# Introduction

## Background

In the modern beverage batch manufacturing industry, maintaining consistent production efficiency is crucial. There are challenges in the current production regime, especially concerning downtimes in the production of mucilage containing materials across different production tanks. This not only disrupts the flow of production but also signifies existing inefficiencies or gaps. Downtime can have a significant impact on production schedules, lead times and overall productivity. The sheer complexity of batch data, coupled with its vastness, has made it difficult to derive actionable insights manually, leading to potential overlooked areas of optimization. Furthermore, without predictive mechanisms in place, anticipating these downtimes for better scheduling remains a challenge. This research intends to address these gaps through a comprehensive exploration of production downtimes, an efficiency-driven analysis using machine learning, and the development of predictive models to enhance scheduling processes.

Downtime refers to the period when a system is non-operational or fails to execute its primary function. The length of this non-operational phase signifies the time span between the onset of a system malfunction and its inability to perform its intended role. Predicting factory downtime serves to pinpoint areas where efficiency can be swiftly enhanced without altering operations, as stated by Kadam et al., 2014. Such predictions empower managers to make judicious choices concerning scheduling, manpower allocation, and production strategizing.

Traditionally, manufacturers depended on historical data and the insights of experts to forecast downtime. However, these methods can be considered subjective and might confine the scope of predictions. In contrast, the integration of machine learning models and big data analytics has significantly improved the precision of downtime forecasts, (Kadam et al, 2014)

## Mucilage Containing Beverage Production Process

### Production Tanks and Instruction Steps

To produce a batch of Mucilage Containing Beverage, it involves on average 27 instruction steps. The production line consists of production tanks divided into systems of varying capacities and each system is denoted as 22, 23, 25 and 25 followed by MT (mobile tank), so in system 22 MT, there is 5 tanks , 1- 5 of capacity 20 tonne. These production tanks have two jobs, one is where batch production occurs and a destination tank where after certain instruction steps, the batch needs to be stored. This type of process can be termed continuous, as the product is an order in volume, (Kang et al, 2020)

The production schedule determines the material batch to be produced, the quantity required, and this decides which production tank is to be used. The production process is mixed between automated and manual where some ingredients such as water is added via a tank delivery and gum addition is manual by a production operator. Fig 1 hows an example of production tank system. Each instruction step has parameters that is logged on a shopfloor system which is a batch data storage database. Instruction steps can be described as a recipe but in this organisation can also be known as phases. Each instruction step is monitored by various metrics, e.g Phase duration, this the length in minutes of how long each phase takes. There is also target phase times for the duration and flowrate of ingredients, these targets have been determined based on historical batch data.

### Raw Material : Mucilage

These beverage materials are so called because they contain an ingredient called mucilage or otherwise known as gum. The purpose of this ingredient is important to these materials as they add stability to the beverage batch but also is known to aid the enhancement of colour ingredient., (Chung et al, 2016, Benech et al , 2008)). But it’s this gum ingredient’s behaviour in the beverage production process which can impact the production process downtimes measured as either phase start delays which is the measure of the delay in the instruction step or phase overrun, which is the overrun time of the phase. For the purposes of this study, phase overrun will be the downtime measure for the mucilage containing beverage materials.



Figure 1Process Production Tank System

If the Gum addition instruction step overruns, creating downtime, it affects the following production process steps such as the mixing via agitation step, processing via deaeration step and the final texture of the beverage batch. The main reason the gum addition step in the production of beverages materials is problematic is due to length of time it takes to add the gum itself which is manual and to dissolve in the production tanks. After the initial addition, the rest of it tends to float on its surface. Therefore, it takes longer to mix into solution. The longer the agitation instruction, step takes, the more gas that is created in the production tank resulting in a longer deaeration step. The agitators are switched off first and then deaeration starts. Fig 2 shows an example of a sample batch process system and fig 3 shows the inside of a production tank. The Production tank environment is not complicated but it is what happens inside during the production of these batches that can create issues such as downtimes. All these are monitored and recorded as phase overrun metric and logged in each batch details on the shopfloor system.

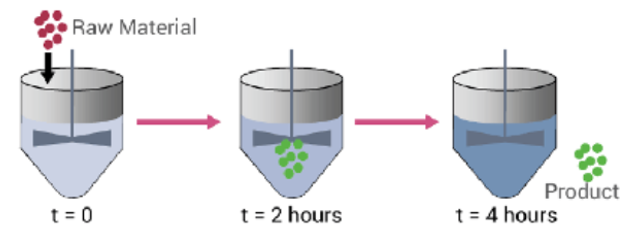


Figure 2 Sample Batch Process System

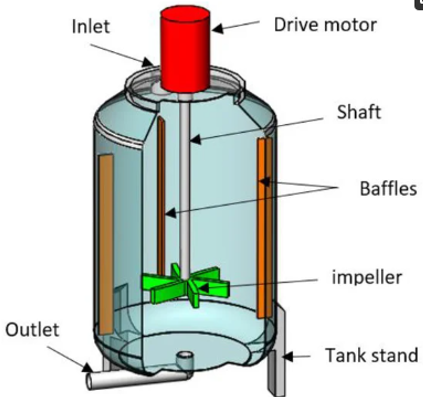


Figure 3 Schematic Diagram of Production Tank

### The Batch Data Process

The current data process from beverage batches production involves collecting from the various production line operations. This data is used to measure the effectiveness of a production process. The standard measurement is called overall Equipment Effectiveness (OEE). This is computed for the whole production line for each production tank at each production step. It looks at the availability, performance, and quality to determine production efficiency.

Lepenioti et al, 2020 states that this data, has the credentials to move beyond these OEE metrics and with the recent advancements of machine learning, predictive and prescriptive analytics using machine learning are possible with the aim of supporting the operator on the shopfloor and enhancing the production process. Another common term for this process is data mining which is essentially extracting valuable information from a vast amount of raw data and transforming into an understandable structure for future use. **(**Ge et al, 2017)

Data sources can be external and internal. In this research, support for data and information communications come from various industrial information systems which is integrated. and it is stored on the relational database, (Min et al, 2019)

The data for this research was originally stored on the manufacturing shopfloor database system. For each batch produced, each production process time step was recorded. By selecting the production time, we were able to request via SQL query all material batch details that had a deaerating step in their process. Therefore, reducing the amount of data was sent and stored as a CSV file. This data will need to be pre-processed before it can be passed to a machine learning model, ( Lee et al ,2019).

The Production batch data in the organisation is acquired from software called FactoryTalk® Batch. It provides a flexible batch control system. It allows you to specify procedures or recipes and enforce their execution in production. The FactoryTalk Batch Server operates FactoryTalk® Batch software, (Rockwell automation, 2006, Kuhar et al m 2015). Figure 4 below shows an example of how this software looks like during a production.

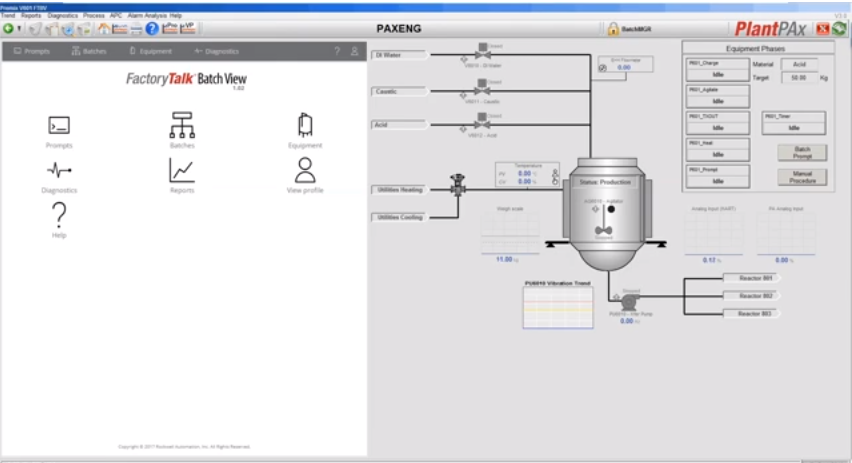


Figure 4 Screenshot of the FactoryTalk Batch View

This FactoryTalk® Batch software uploads batch data from each of the production activity known as phases per batch to a Microsoft SQL database which is managed by SQL Server Management Studios. One of the components of this software is the FactoryTalk Event Archiver which translate the FactoryTalk batch tab-delimited ASCII electronic batch record files to a user-specified file type. This organisation uses Microsoft SQL database which is managed by SQL Server Management Studios to maintain files for each batch created.

Microsoft SQL Server Management Studio (SSMS) is a software application developed by Microsoft that is used for configuring, managing, and administering all components within Microsoft SQL Server. A central feature of SSMS is the Object Explorer, which allows the user to browse, select, and act upon any of the objects within the server, (Hughes et al , 2020).

### Overview

* Research Objective 1: Understanding Production Downtimes
* First, we wanted to understand the current process better. We investigated how often and why there were pauses (downtimes) in the various phase stages in the production manufacturing tanks. This gave us a clear picture of where there were production phase overruns for each tank.
* Research Objective 2: Using Machine Learning to Analyse Data
* Next, we turned to machine learning to analyse the production batch data. By studying the data, we wanted to point out where the process could be made more efficient. This step showed the value of using advanced tools and methods to analyse production data.
* Research Objective 3: Predicting and Planning for Downtimes
* Lastly, we created machine learning models to predict when these downtimes might happen. With these predictions, we aimed to plan production schedules better, reducing the number of pauses and making the whole process faster.
* The comprehensive approach that follows integrates data collection, pre-processing, model selection, training, evaluation, and optimization

### Research Overview

#### Problem Statement

In the realm of beverage production, operational downtimes can significantly hinder productivity, resulting in cost overruns and inefficiencies in the manufacturing process. Despite the availability of abundant batch data, a systematic methodology to understand, analyze, and predict these downtimes has yet to be implemented comprehensively. The overarching goal of this research is to enhance the beverage production process using machine learning techniques.

#### Objectives

* Research Objective 1: Understanding Production Downtimes

First, to understand the current process better by investigating how often and why there were downtimes in the various phase stages in the production manufacturing tanks. TO give us a clear picture of where there were production phase overruns for each tank.

* Research Objective 2: Using Machine Learning to Analyse Data

To apply machine learning to analyse the production batch data. By studying the data, to point out where the process could be made more efficient. This step shows the value of using advanced tools and methods to analyse production data.

* Research Objective 3: Predicting and Planning for Downtimes

To create machine learning models to predict downtimes in particular production instruction steps such as agitation , mucilage addition and deaeration phases .With these predictions, production schedules have the potential to improve, reducing the number of overruns and making the whole process faster.

Following on from here, section 2 is the literature review, delving into existing research related to this this topic to provide an overview of key findings and methodologies used previously. Chapter 3 and 4 is concerned with the research methodology where the research methods used in this study are outlined in detail. Chapter 5 is the results and discussion where data and findings from the machine learning models results are given and discussed. The discussion will connect with findings from the literature review, and highlight any implications, significance, and potential limitations of the results. Chapter 6 is the conclusion which will summarize the main findings, discuss their implications, and give suggestions for future work.

# Literature Review

## Introduction

The literature review will focus on a range of areas relating to production process optimization through the application of machine learning models for the downtime prediction in beverage containing mucilage production.

I will look at the importance and impact of efficient production in industry in general to examine how efficiency is achieved and changes to this process. Next to focus on production downtimes in batch manufacturing, define the different types of downtimes and the factors contributing to downtimes. To review methods and techniques that are used to quantify and monitor downtimes.

Following this, I will examine machine learning models applications in manufacturing process by first defining machine learning and giving a background detailing the advantages and challenges of implementing it in production. Further delve into using machine learning in production downtime predictions and its advantages for production optimization.

## Efficiency Driven Analysis of Batch Data

The main motivation behind the analysis of batch data is to improve the efficiency of the production process through reducing downtimes, maximizing output, minimizing waste, optimizing resource utilization, and shortening the overall production cycle. There is an increasing importance to enhance the effectiveness and efficiency of decision making in a production process, through mining the production data both online and offline using more efficient techniques. (H., XIA et al, 2022). This is a methodical examination and evaluation of batch data and can involve inspecting, cleaning, transforming, and modelling data to discover useful information. This is more often referred to as data mining.

The world of manufacturing has the potential to utilize machine learning to extract patterns from existing datasets, which can serve as a basis for predicting future system behaviour, (Alpaydin et al 2010, Nilsson et al, 2005). Alpaydin also claims “Storage data only becomes useful when it is assessed and translated into knowledge that we may use, for example to develop predictions”. This is true of this research.

In the manufacturing context, "batch production" is a method where items are produced in groups or batches rather than in a continuous stream. "Batch data" would then refer to the data generated during these batch production processes. It could include variables like production start and end times, quantities produced, downtimes, error logs, equipment metrics, and any other relevant data points that can provide insights into the production process.

According to Kovalev et al, 2019, it highlights the importance of batch data as the head of the process of digital transformation part of the industry 4.0 revolution. Digital transformation is the approach used by the production industry undergoing this revolution for the optimization of production data.

Production Batch data is worthless on its own, the manufacturing industry requires efficient processes to be able to derive valuable information from it. The following are examples of processes of examining production batch data to uncover hidden patterns and correlations include data mining, machine learning, natural language processing.

## Traditional Process Optimisation Methods

Prior to machine learning, traditional methods used for improving production efficiency included manual inspection, statistical tools, expert systems, and mathematical modelling, (Wang et al, 2018) Other traditional statistical methods such as statistical control charts (SPC) are now deemed insufficient when it comes to enhancing production processes, (Ismail et al, 2021). The advantages to these control measures were their applicability and simplicity but are now not able to keep up with the increasing complexity of production and volume of data being gathered as a result. (Ismail et al 2022).

Ge et al, 2017 describes the importance of analysing batch production data for patterns and relationships between production variables leading to useful information can be extracted and used by Statistical models such as Operational equipment efficiencies (OEE). OEE can be developed for various applications such as process monitoring or fault diagnosis.

Another important aspect of batch data collection is highlighted in Arif et al ,2023, where it states most existing quality monitoring models only look at one manufacturing state and the batch data gathered is not processed until after the product is made or manufacturing process is over. This has a negative effect on resources, time, and production performance.

Three areas that make traditional methods obsolete and where utilising machine learning shines are:

* Information and communication technologies – mode of production has changed, large-scale tasks, operating performances and environments are more complex, (Wang et al, 2018a)
* Increased demand for real time dynamic self-adaptive and precise production management (Arashpour et al, 2018, Lamon et al, 2010.)
* The completion of various kinds of information systems deployed in manufacturing enterprises. E.g., CAPP, computer-aided process planning, (Papananias et al, 2019.)

Machine learning cannot completely replace traditional methods. One can learn from the other while machine learning can aid the identification and modification of the parameters of the traditional methods to improve processes (H., XIA et al, 2022.)

Also, another consideration is the time and financial cost of developing and programming models of machine learning, this may be out of reach of small-scale manufacturers, however for large scale manufacturers and their various production lines, it is the varying complexities and lack of efficiencies these models can bring. It can be challenging to implement AI – machine learning in an entire organisation with existing processes and systems if the company lacks robust technology infrastructure and collected data. This is often a limitation It’s important that there is clear strategy stemming from top management to achieve goals otherwise AI will fail (Kang et al 2020, Heio et al, 2021).

## What is Machine Learning ?

Machine Learning model systems learn from data, identify trends or patterns from data, make decisions based on structure feedback and then perform tasks on their own, with continued improved performance and problem-solving skills without human intervention (Helo et al, 2019). It models the complex relationship between input and output data, (Wang et al, 2018b).

Machine learning can be divided up into 3 types: unsupervised, supervised and reinforcement types, (Pugliese et al , 2021) as shown in fig. Its also shows the associated machine techniques. For this research, type of machine learning relevant is supervised as we have an input and output from a set of labelled training data and the technique is regression as the input value or our target variable is a numerical continuous valuable which is the attribute phase overrun times. These algorithms are used to optimize the coefficients of each independent variable to achieve a minimum error in the prediction.

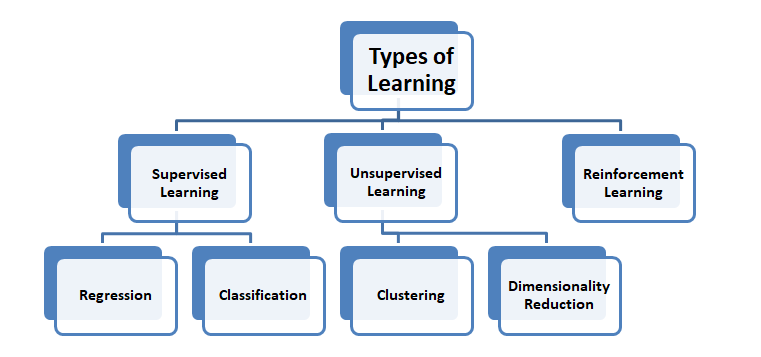


Figure 5 Types of Machine learning

The common methodology in using Machine learning algorithmsfor the improvement in manufacturing process and quality optimization are as follows, (Ismail et al, 2021, Kant et al.,2015, Aksa et all.,2021, Ahmad et al.,2018, Kulkarni et al.,2020, Koksal et al., 2011)

* Data Collection – collect data on the performance of the chosen parameter under different operation conditions, e.g., phase overrun , phase duration ,flow rates, temperatures etc
* Data Prepossessing – removing outliers and missing values ensure the data is clean and accurate
* Feature selection – selecting the relevant features that can help prediction
* Model development – develop the predictive model and evaluate it using metrics such as accuracy, precision, and recall
* Model Optimization – to improve its performance using hyperparameters or ensemble learning techniques
* Model Deployment – deploying the model in a real-world setting and validate its performance.

According to Kang el al, 2020, most machine learning applications in the manufacturing industry were concerned with supervised learning, due to the abundance of production data available. Regression is the main task applied for quality optimization problems while classification and anomaly detection are mainly applied for product failure detection.

Machine learning has successfully been applied in industry, these are summarized below, (Diez-Olivan et al ,2019, Aksa et al, 2021, Menezes et al, 2019):

* **Descriptive** – accounting and analysis of historical data. e.g. Fault detection and diagnosis
* **Predictive** – Considers near past data to predict coming future trends, biases tendencies and behaviours through causation and correlation.
* **Prescriptive** – finds or prescribes the best mode route manner or moves to operate based on given data (output) and models (inputs). help make decisions on what to do and how to minimize failure impact
* **Preventative** – determine the potential of failures happening.
* **Detective analytics** – makes diagnostics of collected data to eliminate and rectify inappropriate values used in predictive analytics.
* **Cognitive analytics** – automated predictions, prescriptions, and detections for smarter decisions over time

## Machine Learning for Predictive Analysis

Through this literature review, there was no direct research on the process optimization using machine learning models by predicting production phases downtime in the beverage manufacturing. However, there was plenty of research reviews on other types of processes in the manufacturing industry using various machine learning (Monostori et al, 1996, Md et al , 2022, Paturi et al, 2021).

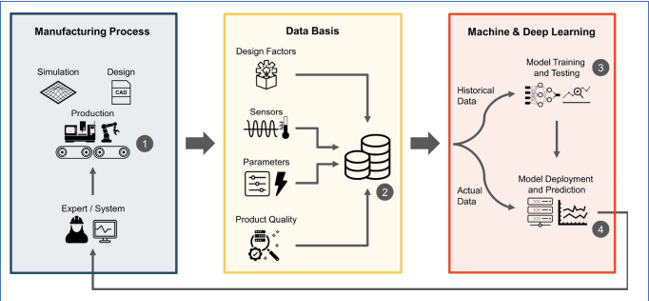


Figure 6 Predictive Quality Approach for a Selected Manufacturing Process

Tercan et al, 2022, reviewed research publications between 2012 and 2022 dealing with predictive quality in manufacturing using machine learning models. The categorization was based on the manufacturing processes and machine models involved. None of the processes involved a beverage production process. Predictive quality uses machine learning methods in production to predict product-related quality base on process and product data. Most research papers reviewed by Tercan et al, 2022, involve having a base machine model and then doing a comparison with other machine models and these mainly are ensemble methods. Ensemble methods involve the combination of multiple learning models, thereby aggregating their decisions to make a prediction.



Figure 7 Baseline Models used in Predictive Quality Process

In this research the initial part is focused on the exploratory data analysis where the goal is to understand the data and identify inefficiencies during the various instruction steps identified across the production tanks used. This is not a machine learning task but it is a crucial preliminary step before applying machine learning models, (Santos M et al,2023, Komorowski et al, 2016).

## Machine Learning Models Application in Production Process Optimization

Most manufacturing processes involve multistage steps to produce accurate products. Papananias et al, 2019 developed a Bayesian linear regression model to estimate part quality of and associated uncertainties given in process monitoring data. The predicted results compared well with the experimental measurements and further a neural network model was developed which also showed similar results. It highlighted challenges to the model of evolution of more complex products, big data, and manufacturing intelligence. It also gave the suggestion to look at self-organising maps (SOM, Lieber et al, 2013), or principal component analysis for the consideration of process variables such as high sampling rates or large measurement uncertainties to reduce them. A study on the injection moding process by Farahani et al, 2021, looked at over 10 different machine learning models. It was determined that Neural Networks is a perfect model for this type of prediction but the computional power, time and cost are factors need to be considered.

A research paper concerned with engine oil aeration process step looked at a gaussian regression model used to correlate the identified features to measure oil aeration. The results were successful in the prediction of oil aeration to an uncertainty of +/-0.02 from the measured oil aeration values. The model was trained using previous oil pressure data. The results also highlighted that importance of looking at sampling measure as the cases used showed overfitting. This was calculated from using the metrics of RMSE, root mean square deviation. (Kulkarni et al 2021)

Under Predictive tool performances, logistic regression as a machine learning model has been used for the reliability estimation for cutting tools using the variable vibration signals. It is using correlation analysis approach to estimate the reliability and failure time of the cutting tools. It concluded that by further looking at the mechanical knowledge and probability density functions of other variables could further enhance the model’s predictions on machine tool performances. (Chen et al, 2010)

IBM Research has developed a process and system regression optimization service for optimising set points for process controls. The Process and System Regression Optimization service consists of three main components -- (1) Regression component and (2) Single Process Optimization component, and (3) System-wide Optimization component. The three components are used to model the problem at hand with the regression component being used to train one or more regression models based on historical data from the process or asset, and one of the optimization components used in combination with the trained model(s) to optimize the control points. Datasets are trained on the API and then it uses various models of regression to make predictions requested. It uses Mean Squared error and R2 as metrics to compare different regression algorithms.

Phan et al, 2021, used this IBM AutoAI toolkit for automated machine learning to search for the right learning algorithm and optimize its hyperparameters. They were able to restrict their models to decision tree, multivariate adaptive regression splines. Using cross validation to pick the best performance model.

Hassani et al, looked at the efficiency of the equipment rather than the product process in manufacturing, the methodology included a case study where the data was based on results from 2 machines. The applied methodology included data preparation, exploratory analysis of the data and creating training sets and applying various models. Including support vector regression, random forest. Model accuracy was calculated with and with cross validation. It is the same methodology for optimization of equipment as is process.

## Deep Learning via Neural Networks Application in Production Process Optimization

Deep learning provides advance analytics tools for processing and analysing big manufacturing data. It is seen as a breakthrough solution to the challenges of multimodal data, the high dimensionality of feature space and multicollinearity among data measurements. It has applications in speech recognition, image recondition, it allows automatically processing of data towards highly nonlinear and complex feature abstraction via a cascade of multiple layers, (Wang et al, 2018, Rivas et al, Manami et al, 2023, Trask, A.W., 2019, Chollet, et al., 2021). The following looks at its application in process optimizations in manufacturing.

Pfrommer et al, 2018, used ANN as a surrogate model to optimize the draping process of textiles. He used the surrogacy process to eliminate the costly process of trial experimentation that is often the approach for process optimization. The study achieved improved results using surrogacy with deep ANN but highlighted the importance of the training data and the sampling.

A combination of a neural networks with a genetic algorithm to predict critical parameters in a particle board and axial flow designs manufacturing process. The Genetic algorithm was then applied to the trained NN to determine the optimum values using the parameters successfully. (Cook et al, 2000, Liu et al 2023)

Fang et al, 2019, learned from Chen and Wang et al, 2013 studies to forecast the cycle time in a wafer fabrication factory using a fuzzy back propagation network. It shows that the NN models lacked generalization and fitting capabilities to deal todays big manufacturing data. It also highlighted the importance of considering expertise for feature extraction to reduce the input dimensionality. Fang et al, 2019 considered all this in their study of shop floor real time job remaining time prediction. They used a Deep stacked sparse Autoencoder to deal with every changing manufacturing floor. In its conclusion it suggested looking at LSTM, long short time models to analyse time series problems and find bottlenecks in the production line. It also suggests reinforcement learning to help with self-learning in the production control and optimization.

A Comparison between ARIMA, LSTM and GRU models was done on time series forecasting for bitcoin technology. ARIMA, Autoregressive integrated Moving average, turned out to be the better model, followed by the GRU model gated recurrent network whereas LSTM allows the tracking the dependencies of new observations with past ones. This study gave details on the methodology on creating and comparing types a recurrent network, with certain parameters and evaluation metrics. It highlighted the importance of data normalization, using Scikit-learn making it easier for the network model to learn. Another highlight from these studies was the importance of the architecture of the neural network itself and optimisation of these methods (Yamack et al, 2019, kuric et al, 2022, Sun et al, 2019).

The time series predictions is used to monitor time changes and monitor trends in the development of the examined parameter. The most used models in research and practice for time series prediction are linear autoregressive models (AR and ARX), LSTM neural networks, moving average mode (MA) and autoregressive moving average (ARMA), (Kuric et al.,2022).

Yamack et al, 2019 conducted a comparison between three different machine learning models in making a time series prediction. The three were ARIMA (Auto regressive integrate moving average), LSTM (Long Short-term Memory) and GRU (Gate recurrent Unit). The topic was predicting the price of Bitcoin. This research compared all three models by accuracy (measured by MAPE and RMSE) and time. It concluded that the outcomes could have been of effected by the parameters that were chosen and the total amount of data used.

Fang el al.,2019 examined the time predictions to complete a job in a discrete manufacturing system using a deep learning-based approach. A deep Stacked sparse autoencoder (S-SAE) model was designed to enable a machine algorithm to learn highly varying status of manufacturing for prediction. In this research the performance was compared to different models, such as linear regression, back propagation neural networks, multi- layer neural networks and deep brief neural networks. To do this, the above models were trained and tested using the same dataset via a fivefold cross-validation. He concludes that future work, includes using LSTM neural networks to analyse time series in production.

A study on a predictive maintenance of an industry machines by Geltz et al , 2021, looked at several machine models using firstly traditional models such as XGBoost , ridge regression and then neural networks using a multi-layer perceptron. The traditional methods produce good accuracies whilst the neural networks did not. They concluded that the lack of enough data points hampered neural networks thus highlight the importance of the data set. Kang et al , 2020 also noted that a larger dataset helps to achieve better results with neural networks , but they’re main advantage is ability to handle complexity which can be found in productions

## Production Process of Manufacturing Mucilage containing containing Beverages

A production line involves a series of factory operations that refine materials into end products (Kang et al., 2020). In this study, the focus is on a tank-based production line involving steps like agitation and deaeration. While machine learning excels at deciphering complex patterns, its effectiveness in this context is uncertain. A manufacturing company may operate multiple production lines, each with distinct challenges. Vast data from some lines can lead to effective machine learning solutions; however, applying the same solution to lines with limited data might not yield the desired results (Kang et al., 2020). The Production lines are classified as continuous where the target variables are continuously measured. According to Kang et al, 2020, most machine learning models applied to Quality optimization problems on production lines are regression type. For this study the target variable for downtime predicting will be the phase overrun times. This attribute is measure of how much longer each instruction step took to complete above the target time predetermined.

### Process Instruction Steps

Three of the instruction steps in the process of producing mucilage containing beverage batches give rise to phase overrun downtimes. Each step is linked to each other.

Once the raw ingredients are added, the agitators are started, and the mixing begins. Agitators are equipment used to homogenise media inside a tank, they work by rotating immersed impellers at a controlled speed, call revolutions per minute. The fig below shows the various components of an agitator, which are present in the current production tanks that produce the mucilage containing beverage batches. From the production process batch data, agitation times are different for each tank which leads to varying production times and can lead to batch downtimes. The reasons for this is the dispersion of mucilage gum ingredient, once viscosity increase has started, agitation of the solution and therefore powder dispersion becomes increasingly difficult.

This results in longed mixing times are required to complete dispersion/hydration and creating more gas that needs to be dispersed during the deaeration phase.

It plays a crucial role in ensuring the quality and shelf life of the final product which was documented by Feilner et al. It involves the removal of dissolved gases such as oxygen from the beverage. The deaeration time must be carefully calibrated to achieve optimal results with the goal of preserving the desired sensory characteristic and stability of the beverage (Paquin et al, 2009).

There are many process parameters that can affect the length of deaeration time a beverage batch needs. Examples of process parameters include the type of final product to be produced, the equipment used, the initial methods of addition and nature of raw material added, temperature and pressure factors. Optimization of these parameters can lead to a lower deaeration time thus an increased efficient process without compromising the final product quality and stability.

## Summary

The literature review underscores the application of machine learning in optimizing production processes, with an emphasis on predicting downtime in beverages containing mucilage. By analyzing batch data, the goal is to enhance efficiency, reduce waste, and streamline the production cycle. Industry 4.0 accentuates the importance of digital transformation and the leverage of batch data, with analytics tools like machine learning yielding invaluable insights from this data, , (Rai et al, 2021).. Traditional methods of optimization, like Statistical Process Control (SPC), are becoming outdated due to evolving technologies and the burgeoning integration of information systems. Machine learning, especially supervised learning, excels in identifying data patterns, continually refining its decision-making, and has found myriad applications in manufacturing, ranging from descriptive to cognitive analytics. In the broader context of manufacturing, machine learning is notably used for predictive analysis. Although the beverage industry hasn't been a focal point, other sectors have tapped into its potential, with various studies showcasing machine learning models, including the emerging deep learning techniques, for prediction and process efficiencies,, (Aksa et al, 2021).

Furthermore, a consistent theme across the literature is the adoption of multiple machine learning models to evaluate data and forecast outcomes. Equally pivotal is the data: its acquisition, refinement, and preliminary understanding of inherent trends and correlations. The nature of the data—be it labelled, structured, or unstructured—directly influences the choice of a machine learning model. There's a shared methodology across diverse research, encompassing distinct, replicable steps, which will significantly benefit this study. The proposed metrics for assessing predictions and trends will further fortify the research foundation. The crux of the present study is to assess the efficacy of machine learning models in predicting phase overrun times for mucilage-containing batches, employing a spectrum of production tanks with diverse capacities.

# Research Methodology

Research methodologies typically fall into two main categories: primary and secondary data collection (Saunders, Lewis, and Thornhill, 2016). For the research questions poised at the start of this paper, a combination of primary and secondary data collection will be utilised.

Within both categories, various methods of data collection can be utilized. This section aims to outline the chosen collection strategy or strategies that, in the author's view, best serve to answer the research questions presented in Section 1 above.

## Primary Research Data Collection

The author determined that employing a qualitative and quantitative approach for collecting primary research data would be the most suitable technique to gather first-hand insights, approaches as observing and interviewing individuals knowledgeable in beverage process production and data analytics.

Factors considered by the author when choosing this primary research method encompassed:

* Research Objectives – the appropriateness of the research method involved.
* Author status – role/access in the organisation if applicable
* Expertise of the Selected Participants – availability and validity
* Timeframe – how long this research method could take.
* Ethical Considerations – gaining consent of the participants to partake and the use the data collected.
* Bias - ensuring a diverse range of inputs by involving individuals with distinct roles in the beverage production process.

### Data Collection Through Observation Qualitative Approach

The author, employed full-time at a beverage production company, undertook participant observation in the production area to deeply understand the process, which inspired the research questions. Gaining access required a candid explanation to the production manager about the study's purpose and the expected insights, like comprehending process terminologies. Noting observations and informal conversations was cleared in advance, ensuring participant anonymity. Potential research challenges included the author's limited process comprehension, possibly leading to observer error, and the "Hawthorne effect" where observation might alter participants' behaviour. These potential pitfalls would be addressed in subsequent in-person interviews. This observational strategy provided vital insights and facilitated future interactions with process experts for in-depth interviews.

### Data Collection through in-person Interviews

In-person interview can take many forms but the one used for this research is an unstructured informal interview, and for practicality and ease of transcription, these interviews were conducted online via the Microsoft Teams platform. By employing an informal interview using open- ended questions as this data collection method, it facilitated a more comfortable scenario allowing for in depth discussions surrounding the beverage production process and the potential use of data analytics. Chosen Participants were able to freely talk about the process and the author was able to observe the participant reactions and adapt the interview flow, accordingly, as suggested by Saunders et al. (2012). This is a more flexible approach.

To address potential respondent bias, the author ensured a broad range of perspectives by engaging individuals from various roles within the beverage production process. Each interviewee contributed a distinct viewpoint: one oversaw the production process, another actively worked in the production area, and the third played a role in data analytics within the organization. Through interviews with this diverse group, the author aimed to achieve a comprehensive understanding of the existing beverage production process, encompassing both common practices and individual variations.".

Transcripts from the interviews can be seen in appendix XXX

## Secondary Research Data Collection

The secondary data source used was internal to the organisation, from various industrial information systems which is integrated. The support systems are FactoryTalk batch software and SQL Server Management Studios.

This data can be unstructured, and need to be prepared, by screening for duplicate data, missing data, irrelevant data records. Extraction of indicators and features by labelling the data that will be needed in the learning/training process.

The data for this research was originally stored on the manufacturing shopfloor database system. For each batch produced, each production process time step was recorded. By selecting the production time, the author was able to request via SQL query all material batch details that had a deaerating step in their process. Therefore, reducing the amount of data was sent and stored as a CSV file.

The author chose a quantitative method for gathering secondary research data. Selection factors included:

* Relevance of secondary data to research goals
* The author's position and access within the organization
* Availability and knowledge of chosen participants
* Estimated research duration
* Ethical aspects, including participant consent
* Bias - ensuring a diverse range of inputs by involving individuals with distinct roles in the beverage production process.

## Research Validity

To determine the extent of which the primary and secondary data collection above accurately assesses the research problem objectives is the basis of validity. We need to validate if the data collected for this research answers the research questions poised and are these answers trustworthy and meaningful.

The research addresses multiple facets of validity to ensure reliable findings in exploring the beverage production process and data analytics. Internal validity is reinforced by diverse participant selection and method triangulation, enhancing data reliability. External validity aims for generalizability by capturing varied roles and data sources like FactoryTalk® Batch and SQL Server Management. The study boosts ecological validity by using real-world settings and expert interactions. Content validity is ensured through on-site observations and expert interviews, offering a comprehensive understanding of the process. Finally, face validity is established through diverse data collection methods and transparent research objectives, ensuring methods seem credible. Rigorous checks for biases and methodological precision are integral to maintaining these validity levels.

## Research Ethics

Ethics in research is important because it ensures that the participants are treated fairly and respectfully. It will also help protect their privacy and well-being. It also ensures that the data gathered is honest and trustworthy.

In the context of the primary data collection, the expert participants will be invited to join the research with the freedom to withdraw consent or exclude their data contributions at any point before the submission date. During the interviews a professional approach will be maintained with confidentiality and anonymity of the participants and the data information given preserved for research purposes. A master file containing the responses will be kept by the author and will not be shared. This master file will be securely stored, and password protected as an enhanced security. Additionally, if the participant has any queries during or after the interview, they will be answered promptly and sensitively.

The secondary data source used was internal to the organisation and it is stored on the relational database system called shopfloor system. The author followed proper protocols when gaining access to the data by engaging in the relevant channels within the organisation. Such channels include the organisations legal counsel and ethics point software where you can determine if a non-disclosure is relevant. However, since the data received will be swiftly anonymized to prevent linkage to the individual participants and organisation, there was no legal or ethical worries. The safeguarding of the data will be ensured through secure storage and the implementation of passwords.

Furthermore, the study acknowledges the importance of adhering to the General Data Protection Regulation (GDPR) and is committed to implementing necessary measures to uphold compliance throughout all stages of the research process. This demonstrates a commitment to ethical research practices and the protection of participants' privacy and rights.

## Sampling Strategy

For effective research, understanding the quantity and type of data required is crucial. While the data gathered constitutes the population, it's not always necessary to analyze it all. Instead, a sampling strategy might be employed for conciseness without sacrificing representativeness. The sampling approach differs for primary and secondary data. Primary data, sourced from in-depth unstructured interviews, will utilize nonprobability sampling due to its qualitative nature. Conversely, secondary data will use probability sampling, ensuring every item in the population has an equal chance of inclusion.

Primary data for this research was gathered using in-depth, non-structured interviews, employing a non-probability, purposive sampling strategy. Participants were chosen based on their expert knowledge and unique perspectives. Purposive sampling, while not representing the broader population, emphasizes in-depth information, making it suitable for exploratory research. While this approach may limit generalizability, it ensures high validity by focusing on specific research questions and drawing from expert insights. The main advantage of purposive sampling is its ability to provide rich, targeted information, especially on complex topics. However, it might carry potential biases and may not always represent wider populations. In conclusion, while its findings might not be broadly generalizable, the depth of insights gained from purposive sampling makes it a crucial tool in research design.

Secondary data was sourced from the company's internal relational database, the "shopfloor system", which houses comprehensive production data. Given the vast volume, categorized as Big Data, a sampling strategy was essential. While probability sampling was initially considered, the sheer size made it unfeasible. Thus, data from the past two years focusing on mucilage-containing materials was chosen, leading to a purposive sampling approach. This method aimed to identify patterns over a specific time by examining production variables. The sampling frame encompassed all production batches, with the sample specifically filtering for mucilage-based materials, considering the various production tanks used. Of the 347 batches produced over two years in 16 tanks, all were included in the sample. Homogenous sampling was the chosen technique, focusing on similar production tanks. The strategy prioritizes depth over broad representation, offering detailed insights but potentially limiting generalizability. Purposive sampling provides rich, targeted insights but may introduce selection biases and challenges in broader applicability.

# Experimental Methodology

In this methodology chapter, the primary intention is to delve deep into understanding the production downtimes. This will involve a systematic examination of the frequency and underlying reasons for downtimes in various phase stages across the production tanks, pinpointing potential areas of production phase overruns. Following this, the plan to employ machine learning techniques to analyze the production batch data, aiming to identify avenues for process optimization and to highlight the profound implications of utilizing advanced analytical methodologies. To culminate, the methodology will focus on the design and development of predictive machine learning models. The goal with these models is to foresee potential downtimes, facilitating a more efficient structuring of production schedules, minimizing interruptions, and thereby optimizing the overall production process.

## Data Collection

Historical batch data related to the production process was requested from the Microsoft SQL Server Management Studio (SSMS). Microsoft SQL Server Management Studio (SSMS) is a software application developed by Microsoft that is used for configuring, managing, and administering all components within Microsoft SQL Server. A central feature of SSMS is the Object Explorer, which allows the user to browse, select, and act upon any of the objects within the server.

The SQL query was structured to encompass a two-year time frame, focusing on batches containing mucilage beverage that undergo a deaeration phase, inclusive of all relevant phase details and time durations as per the research objectives. Once obtained, the data was initially in Excel format but was transitioned to a CSV format. This conversion not only ensured data integrity but also facilitated seamless integration with Python programs, given the format's readability and compatibility with the language.

## Data Preprocessing

### The Dataset

Each entry is complete with granular details, from timestamps to specific attributes related to production phase durations, raw ingredient materials involved, and other relevant metadata. While the vastness ensures a comprehensive representation of the production traces, it necessitates rigorous pre-processing to sift through noise and redundancies.

The original production csv file contained data on beverage batches that contain a mucilage ingredient produced in various production tanks with differing capacities over time of 2 years. In compliance with the organisations data usage agreement, specific data points have been omitted to maintain the confidentiality. Table 1 below provides an overview of the updated dataset’s column data and explanations. The dataset is called. ProductionDateupdated1.

Table 1 ProductionDataupdated 1 dataset table

Over 46 different type of beverage mucilage containing materials with a total of 367 beverage batches were produced across 16 production tanks of varying capacities. Which tank they were produced in; was dependent on the production schedule for that time. Table 2 gives details of all the productions tanks available for making mucilage containing beverage materials. It shows the capacity of each tank and number of materials and batches produced.

For the purposes of the study, the downtime prediction will be based on beverage batch data produced in the production tanks grouped together according to their capacity and not in individual tanks. From Table 2, it shows the count of batches produced in each tank is limiting and may cause issues for some machine models.

Table 2 Production Tank Details

The data was manipulated using python function aggregation and group by to transform the raw data into a summarized format for the purposed of applying a machine model. So in order to examine the data in terms of the production tanks, the tanks were group by their capacities and then the various feature attributes of interest were aggregated. This created a new dataframe whereby we could apply the machine learning model to.

The dataset was transformed again to show the phase overrun data for batches produced in these production tank groups but looking specifically at the three instruction steps/production phases of interest such as agitation, gum addition, and deaeration step. These instruction steps can be seen in the following table 3. Under STEP1\_CONS step, the ingredient details referring to gum addition was selected, and the batches details selected, and attributes were aggregated. The same was completed for STEP 1,2 ,3 AGITATION steps and HP step which is the phase that is related to the deaeration phase (as confirmed by interview participant no.1). Machine learning model were applied to these specific prepared dataframes.



Table 3 Production Instruction Step

### Software, Libraries, Web Applications

For this research, the programming language employed was Python. Its selection stemmed from its user-friendly nature and the ease with which the researcher could learn and utilize it. The vast availability of open-source libraries in Python further enhances its adaptability, allowing customization based on the programmer's requirements, (McKinney,2012). One of the major advantages of using python and its libraries, is that they can be used together, in tandem for data analysis, scientific computing and machine learning. The following are examples of the open-source python libraries that were used:

* Pandas – provides data structures like DataFrames and Series that make it easy to handle structured data. It allows for quick data cleaning, visualisation, and statistical analysis. In this research its main use was in the handling of missing data, filtering rows and aggregation of data.
* NumPy -employed for its basic mathematical and array operations.
* Matplotlib – A plotting library – creates and display graphs and visualisations of the data working alongside other libraries such as pandas.
* TensorFlow – Deep learning frameworks – used for building and training neural network models.
* Seaborn – works along with Matplotlib to create statistical graphics.
* Scikit-Learn - allows a programmer to quickly implement a range of machine learning algorithms in conjunction with other libraries such as pandas and matplotlib.
* Scipy – a advanced scientific computing library built on NumPy for scientific computational tasks. It adds additional functionality.

All the libraries above were used in Jupiter notebook which is another open-source web application that allows the creation and sharing of documentation that contain live code, equations, visualisations, and narrative text.

### Data Cleaning

#### Handling Missing Values

Missing values are entries in the dataset, such as 0 or not a number (NaN). Using the panda’s library in python, the amount and type of missing values should be determined, (S.Xu et al , 2015). What columns in the data contain missing values and will it have an impact on the model used are questions that will be answered. Missing data can cause bias in estimating model parameters and loss of information ( Ismail et al , 2022, Lee et al , 2019).

#### Removing Duplicates

To protect data integrity, the dataset will be checked for duplicate rows of data. Duplicated data can distort the actual data analysis and give inaccurate results and skew distributions. Using panda’s library, duplication can be determined, and before removing by drop function, the data results need to be check in case they are valid repetitions.

#### Data Normalization

This pre-processing technique involves the transformation of all numeric variables in the dataset to a standard scale. This is necessary when the data features in the dataset have different units or varying scales which can give skew interpretations from machine learning models.

An investigation using the Quantile-Quantile (Q-Q) Plot was looked at to determine the effect of the different types of normalisations such as standard scaling and minmax scaling have on the data. Using the data from the production tank 22MT – reference ProductionTank22\_df2, the following graphs shows results of the distribution of the phase overrun data. Using two examples of standard scalar and min-max scaler , there is a S – Shaped curve which indicates that the data may be following a logistic or s curve distribution rather than a normal distribution.

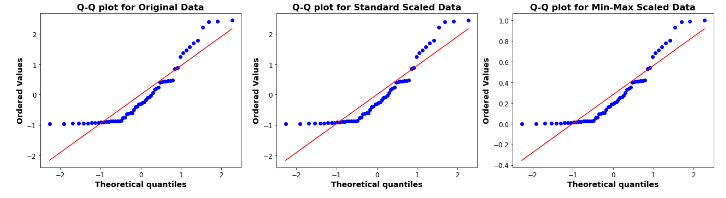


Figure 8 Q-Q plots for the Original and Types of Scaled Data

From the data above, there was no difference between the two scalar methods and below table gives details of the machine models and their sensitivity to scaling.

Table 4 Normalization and Machine Learning Models

The StandardScaler function from Scikit-learn was used to execute Z-score normalization, ensuring each feature contributes equally to model efficiency. This a solid general-purpose choice and had broad applicability,

## Exploratory Data Analysis

Exploratory Data Analysis is important initial step in understanding the nature of production downtimes using phase overrun as the target in the beverage manufacturing process. It will be used to discover trends or patterns, to spot anomalies and check statistical assumptions with the help of summary statistics and graphical representations under the headings of Univariate Analysis and Bivariate and Multivariate Analysis. The following are the purpose of EDA :

* Visualising Downtime Frequencies – using histograms and barplots to visualise the frequency of phase overruns in the production process.
* Identifying Causes for Downtimes – Scatterplots, boxplots or heatmaps/correlations can determine relationships between the different variables in the dataset. (Kumar et al , 2020)
* Spotting anomalies – highlighting of outliers, through boxplots
* Descriptive Statistics – concise summary of the downtimes giving understanding to average downtime durations.

This is aided by the python libraries: Pandas/ NumPy/ Matplotlib/ Scikit-Learn/ Seaborn.

## Application of Machine Learning Models

In this methodology, machine learning models will be utilized to predict production downtimes, a critical factor in optimizing operational efficiency. Given the continuous nature of the target variable – downtime as phase overrun –a regression-based approach ws adopted. The first step involved preprocessing our dataset, ensuring the data was cleaned, normalized, and relevant features were engineered to capture intricate patterns relating to production downtime. The data was then split into training and validation sets, ensuring a representative distribution. Several regression algorithms, from linear regression to more complex ensemble methods, were evaluated based on their ability to accurately predict downtimes. The chosen model was trained on the training set and its performance evaluated on the validation set using metrics accuracy and Root Mean Squared Error (RMSE). To avoid overfitting and ensure our model generalized well, cross-validation techniques were employed. Once satisfactory performance was achieved, the model underwent hyperparameter tuning to further refine its predictions. The ultimate goal of this methodology is to develop a robust machine learning model that can effectively forecast production downtimes, enabling proactive measures to minimize their occurrence and duration.

### Dataset Split

Once the dataset is established, it needs to be split into three subsets: Training, validation, and test sets. There is no optimal proportional for splitting, according to research, the rule of thumb was 70% for training and 15% for both validation and testing. It depends on the size of the dataset, (Barkov, 2019., Rajasekaran et al, 2022) The training set is used to build the model, validation set will be used to choose the machine learning algorithm and find the best values for hyperparameters, and test set will be used to assess the model,

To train the machine learning model, the collected data must be divided up into training and test sets, which are randomly divided by 70:30 ratio of the entire data set. (Cavalcante et al, 2019, Goli et al, 2019).

Using the train\_test\_split function in Scikit-learn, the dataset was partitioned into 80% for training and 20% reserved for testing.

### Evaluation Metrics

This was facilitated using the metrics module in Scikit-learn the predictions from the models are compared to the true values to calculate the Mean Squared Error (MSE) and the R-squared (R2) score for both training and testing datasets. The MSE provides a measure of the prediction error, whereas the R2 score provides a measure of the proportion of the variance in the dependent variable that is predictable from the independent variables.

### Machine Models

### Cross Validation

For each model in the table, a 5- fold cross-validation was performed, (CV=5). The mean and standard deviation of the mean square error was calculated.

A list of the machine models is creating so that a loop can iterate over each model to perform the cross validation. The model is trained and evaluated 5 times (CV=5), each time with a different split of data in training and validation sets.

The scoring parameter is neg\_mean\_squared\_error which means the mean squared error will be calculated for each of the 5 crosses., ( Alpaydin et al , 2014, Kuhn et al , 2013)

### Hyperparameter Tuning

Hyperparameter tuning is performed using GridSearchCV. This ensures that the best parameters are chosen for the model to enhance its performance.

The following steps will be used for hyperparameter tuning:

* A grid of potential hyperparameter values will be defined.
* **GridSearchCV** is used to search over this grid and find the best hyperparameters for the model based on 5-fold cross-validation.
* The best model (with the optimal hyperparameters) is then used for predictions and the same metrics as before are computed.

The hyperparameters for each model examined are outlined in appendix 5.2 and it shows the reasoning behind the choice. Various regression algorithms were evaluated with their corresponding hyperparameters. Linear Regression was implemented without any hyperparameter tuning due to its simplicity. Ridge and Lasso Regression both utilized alpha values ranging from 0.01 to 10, offering different regularization strengths. Ensemble models, like Random Forest and Gradient Boosting, considered multiple configurations in terms of tree depth and estimator count. Techniques such as K-Nearest Neighbors and Support Vector Machines were fine-tuned based on neighborhood count, kernel types, and regularization. Furthermore, neural network-based approaches, from simple dense networks to LSTMs, were also explored, tuning parameters like neuron counts in layers and batch sizes. The LSTM model tuning was particularly computationally intensive, demanding careful configuration. Lastly, an optimized dense neural network was approached using RandomizedSearchCV to identify optimal architecture and training settings. This extensive approach ensures a comprehensive search over potential model configurations to yield the best predictive results.

S. Farahani et al, 2021 suggested the importance to have a framework for the model adaptation and hyperparameter tuning for better model implementation. This is an extensive approach to a comprehensive search over potential model configurations to yield the best predictive results.

### Conclusion

This chapter presented a comprehensive methodology for leveraging machine learning to enhance production parameters. The following chapter will discuss the results obtained from implementing this methodology.

# Results and Discussion

## Introduction

This chapter explains the impact of the primary data and secondary data on the research objectives, through the information received from the three expert participants and the production batch data downloaded from the organizations database system.

The author examines various machine models to determine if the production downtime as phase overrun variable can be predicted for the various production tanks used to produce mucilage containing beverage materials. It leverages historical data and relationships between features and a target variable to determine if predictions about phase downtime in the future is possible. The downtime was examined for three instruction phases of the production: Agitation, Deaeration and the addition of Gum ingredient.

## The Primary Data – Interviews

From the in-depth interviews, responses from the participants with expert knowledge and distinct perspectives in the research area was reviewed and summarised here. The actual interview transcripts are in appendix 5. 1.

### Participant no 1: Data Scientist

Participant number 1, a Data Scientist, played a crucial role in enhancing the existing system's objectives. Initially, the system solely focused on executing batches, with data collection lacking a structured approach for analytics. To rectify this, the participant was involved in refining tracking mechanisms by developing a stored procedure that summarized batch data comprehensively, covering aspects like start times, consumption, problems, weights, agitation times, deaeration times, and homogeneity. They further established targets for different phases through historical data analysis, providing clear objectives for each batch and aiding in issue identification when batches fell short. Resource management was addressed by quantifying departmental needs for shared resources like mobile tanks. Operational Efficiency (OE) calculations, focused on equipment effectiveness rather than usage, were employed to pinpoint areas for improvement. The participant also highlighted the company's emphasis on data analytics, mentioning the creation of a Digital Performance Management (DPM) system for data visualization and future consideration of machine learning where it genuinely adds value. Moreover, they stressed the significance of preventative maintenance, citing examples of early issue detection through thermal balance tracking and the importance of team awareness and training in efficient problem resolution.

The key insights from the Data Scientist's contributions revolved around enhancing batch process tracking, target setting, issue resolution, and emphasizing the role of data analytics and preventative maintenance in optimizing organizational processes.

### Participant no. 2 Production Manager

Participant 2 highlighted the company's data-driven approach, where they utilize operational data to calculate Overall Equipment Effectiveness (OE) figures for each batch, which are then reviewed during daily 9:15 meetings involving cross-functional teams. They stressed the importance of acting based on these figures, with diligent follow-up to ensure completion. OE calculations are performed at the phase level, considering prerequisite times, enabling a granular assessment of phase efficiency. The interviewee discussed various batch phases, such as agitation, deaeration, and ingredient additions, outlining their timing and criteria for proceeding to the next phase. Quality checks before batch finalization were emphasized to meet quality standards. Initiatives to reduce downtime were mentioned, including pre-weighing bulk powders, and optimizing temperature and pump speed. Downtime challenges like liquid filling delays and breakdowns were addressed. Recording downtime reasons and their impact on OE data were explained. Setting targets for parameters like flow rate involved collaboration between different teams. Tank assignments, numbering, and capacity details were provided, with some batches requiring multiple tanks. Different systems, ranging from 20-tonne to medium-to-small systems, were discussed based on production volume needs. An efficiency improvement example involving juice barrel handling was presented. The interviewee reiterated the importance of continuous collaboration, data analysis, and process enhancement to minimize downtime and boost productivity.

### Participant no.3 Production Operative

From the participant's point of view, the conversation primarily revolved around production downtimes and efficiency in a manufacturing process, specifically related to gum mixing. The participant discussed the challenges of interruptions during production, which could result in extended downtimes, impacting operational efficiency (OE). They emphasized the importance of accurately recording extra downtime for batches where gum mixing takes longer due to poor agitation, as this data is crucial for analysis. Additionally, the participant acknowledged that target times for production phases should be more realistic to account for batch-to-batch variations and highlighted potential solutions like improving agitation and pre-mixing highly concentrated gum batches. The conversation also touched on the phases of a typical batch, including ingredient addition, agitation, and transitioning to the high-pressure phase.

The key insight from the participant's point of view is the focus on addressing production downtimes and operational efficiency challenges in gum mixing, emphasizing the need for accurate data recording, realistic target times, and potential solutions to improve the manufacturing process.

# Appendices

## Primary Data: Interview Transcripts

## Machine Learning Models and their Hyperparameters

