ADDIS ABABA UNIVERSITY
COLLEGE OF TECHNOLOGY AND BUILT ENVIRONMENT (CTBE)
SCHOOL OF MECHANICAL AND INDUSTRIAL ENGINEERING (SMIE)
INDUSTRIAL ENGINEERING CHAIR

Optimizing Bottle Filling Efficiency Through a Machine Learning Approach: a Case of National Alcohol and Liquor Factory

This Thesis Submitted to School of Mechanical and Industrial Engineering. Present in Fulfillment of the Requirements for Degree of Master's Science in Industrial Engineering

By: Jemera Melaku

Advisor:

Ameha Mulugeta (PHD)

Addis Ababa, Ethiopia

Feb,2026

Addis Ababa University

College of Technology and Built Environment (CTBE)

School of Mechanical and Industrial Engineering (SMIE)

Industrial Engineering Stream

Optimizing Bottle Filling Efficiency Through A Machine Learning Approach
A Case of National Alcohol and Liquor Factory
By Jemera Melaku

Approved by Board of Examiner

Dr. Ameha Mulugeta ______________ ______________

Advisor Date Signature

Dr. Kassu. J ______________ ______________

Internal Examiner Date Signature

Dr. Gezahegn. T ______________ ______________

External Examiner Date Signature

Dr. Abdulkadir Aman ______________ ______________

Interim Head, SMIE Date Signature

Dr. Shegaw Ahmed ______________ ______________

Interim Vice Executive Dean
for Academic Affairs CTBE Date Signature

DECLARATION

I declare that the work in this thesis entitled 'Optimizing Bottle Filling Efficiency Through A Machine Learning Approach: A Case of National Alcohol and Liquor Factory' is my original work. I have not submitted it to any other degree and have properly cited all references used.

____

____

Jemera Melaku Date

Certified that the above declaration made by the candidate is correct to the best of my knowledge.

____

____

Dr Ameha Mulugeta (Advisor) Date

ACKNOWLEDGES

I would first of all like to thank God that I have been blessed with the gift of life and that I have the strength and courage to do this thesis.

I am also very thankful to my advisor, Dr. Ameha Mulugeta, who guided me, motivated, and provided valuable feedback during this work. He has been a very useful person throughout the research.

I would wish to acknowledge my thanks to the production and technical staff in the National Alcohol and Liquor Factory in allowing me to conduct my observations, interviews, and discussions about the filling inaccuracies. I like the fact that they are patient and ready to offer their time and experience.

I would like to express my sincere gratitude to my family who has been supporting me greatly. My parents have been my strongest supporters and I would not have done this thesis without their support.

Lastly, I want to express my gratitude to my colleagues and friends who assisted, advised, and supported me. Without the help of all these individuals, this thesis would not have become possible, and I am really grateful to all of them.

Table of Contents

Acronyms

ABV- Alcohol by Volume

ANN- Artificial Neural Network

DCS- Distributed control system

DT- Decision Tree

HMI- Human Machine Interface

IOT- Internet of things

ML-Machine Learning

NALF- National alcohol and liquor factory

PID- Proportional integral derivate

PI-Proportional integral

PLC-Program logic controller

RF-Random forest

SCADA- Supervisor control and data acquisition

SOP-Standard Operating Procedure

SPC-Statistical Process Control

SVM-Support vector machine

XGB-Extra-gradient boost

List of Figure

List of Table

List of Appendix

Abstract

The bottle filling process is an important process in beverage production since it has direct influence on the quality of the product, compliance with regulations, and the efficiency of production. Nonetheless, major deviations on the bottle filling process are noticed in the National Alcohol and Liquor Factory (NALF). Records of inspections of 35 production days in January-February 2025 found that 27,972 bottles were underfilled and 18,648 overfilled, with an average of about 1332 bottles in need of reprocessing per day. These variations raise manpower, energy use, equipment depreciation, operation expenses and lowers the total output. The factory used in the 2024 annual performance report has reached an output of 85 percent of the intended production capacity, which indicates the necessity of the better monitoring and control of the process. In this study, the aim was to maximize the efficiency of bottle filling based on a supervised machine learning approach. The study involved both quantitative and qualitative research methods using both exploratory and experimental research designs. Primary data were gathered with the help of structured interviews, structured discussions with experts of the company, and direct observations of the filling process, whereas secondary data contained records of maintenance, production records, and reports on quality inspection. A supervised machine learning approach was used in this investigation to identify filling results. A Random Forest classifier was used to distinguish filling outcomes as underfill, correct fill, or overfill. The model had a training accuracy of 99.6% and a testing accuracy of 99.3%. A macro and weighted average performance and reliability were used to validate model performance and reliability, and out-of-bag error estimation and stratified k-fold cross-validation were used to validate. To facilitate real-time monitoring and dynamic process control, the developed model was coupled with the filling process by use of a three-layer architecture. The conclusions reveal that machine learning can be used to enhance the stability of the process, minimize deviations during filling, and decrease the use of resources, as well as assist in making decisions that are supported by data. This Study adds a viable machine learning framework to enhance filling efficiency and promote intelligent manufacturing behaviors in the production of beverages.

Key words: *Bottle filling process, beverage manufacturing, Machine Learning, Random forest classification, Process optimization*

CHAPTER ONE

INTRODUCTION AND PROBLEM JUSTIFICATION

1.1. Introduction

Optimization has long been recognizing as critical methodology for enhancing manufacturing efficiency and product quality, as demonstrated by numerous studies across various industries. The evolution of optimization techniques has progressed from basic statistical methods to sophisticated Artificial intelligence (AI)-driven approaches, with each advancement building upon previous research findings.

Early Foundation by Montgomery (2009) establishing statistical process control (SPC) as a fundamental of optimization tool, demonstrating its effectiveness in reducing variability of manufacturing outputs. Subsequent studies by Johnson et al. (2018) expanded these principles to automated systems, showing how internet of things (IOT), enable sensor could improve process control by 42% in beverage production environments. These technological advancements enable more precise monitoring of key parameters like fill volume, Pressure and flow rates.

Recent research has focused heavily on machine learning applications for optimization. Chen & Zhang (2020) made significant contributions by developing convolutional neural networks for visual inspection systems, achieving 94,7% accuracy in controlled manufacturing settings. Their work establishing important benchmarks for computer vision application in quality control. Parallel studies by Okafor et al. (2021) adapted these techniques for developing world contexts, though they identified a 15-20% performance gap due to infrastructure challenges, highlighting the need for context specific optimization approaches.

The African Manufacturing consortiums comprehensive 2023 study provided valuable comparative data, showing that adaptive machine learning systems outperformed traditional PID controls by 31% in accuracy across multiple African plants. However, their research also revealed persist challenges, particularly regarding power stability affecting optimization reliability in 78% of surveyed facilities. These finding were corroborated by Muthoni (2023) , whose work specifically addressed implementation barriers in Kenyan manufacturing environments. This study aims to develop machine learning based optimization framework to enhance bottle filling efficiencies at National Alcohol and liquor factory by addressing key challenges such as fill accuracy, production variability and resource inefficiency.

1.2. Background and Justification

The beverage manufacturing industry represents a critical sector for economic development in East Africa, with bottle filling operations serving as a key determinant of product quality and operational efficiency (world bank,2022). National Alcohol and liquor factory, as a regional industry leader, faces significant challenges in maintaining precise fill levels, with internal audits revealing 4-7% annual production loss due to filling inaccuracies National Alcohol and liquor factory, internal report,2024).These operational deficiencies align with findings from the Africa Manufacturing consortiums(2023) regional study, which identified fill level inconsistencies as a persistent challenges across 78% of surveyed bottling plant in East Africa.

Prior research has systematical documented the limitation of conventional filling system in developing market contexts. Johnson, and Roberts (2021) comparative analysis of automation systems demonstrated that traditional programmable logic controllers (PLCs) fail to maintain consistent filling accuracy under condition of power instability and equipment variability challenges particularly prevalent at National Alcohol and liquor factory manufacturing facilities. These findings are reinforced by Zhang & Li, (2022) viscosity studies, which established that mechanical filling systems cannot adequately compensate for the non-liner relationships between filling parameters and liquid properties.

Recent advancement in machine learning (ML) offer compelling solutions to these challenges, as evidenced by multiple peer-reviewed studies. Zhang et al. (2023)controlled trails achieved 92% fill accuracy using convolutional neural networks, while Mkumbo (2023) implementation study at a Tanzanian bottling plant documented a 68% reduction in products waste through machine learning (ML) based optimization.

Machine learning (ML) techniques offer promised avenue to address these challenges by leveraged data to improve decision making and operational process. Various studies have documented how machine learning will transform the practices used in manufacturing operations. The study by Tzeng et al. (2019) showed that equipment performance data analysis through predictive maintenance algorithms leads to decreased downtime and enhanced plant efficiency in bottling plants. Kumar and Kumar (2021) demonstrated the use of machine learning (ML) techniques in food and beverage sector production optimization which lead to better quality management and decreased waste levels.

New possibilities for enhancing filling process was presented by recent technological developments, especially in machine learning. Large volumes of data can be analyzed by machine learning algorithms to find trends and adjust operating parameters, which greatly improves filling accuracy. By putting these leading-edge strategies into practice, National Alcohol and liquor factory can overcome the difficulties presented by conventional filling procedures, which frequently depend on human supervision and mechanical adjustments that might not react rapidly to changes in manufacturing conditions.

Extensive reasons support the implementation of Machine Learning (ML) at National Alcohol and liquor factory. Machine Learning (ML) surpasses traditional statistical process control in terms of predictive accuracy to automatically detect filling issues that might trigger product defects before their manifestation. The ability of Machine learning (ML) models to learn from new production data thanks to their flexibility results in enduring effectiveness during operational changes. The financial advantage through lower product waste together with operational improvement efficiency establishes compelling economic benefits. The adoption of this advanced technology allows National Alcohol and liquor factory to become an industry leader in manufacturing innovation as it meets strict quality standards. The aim of this study is optimization of bottle filling efficiency with Machine Learning (ML) techniques to address filling challenges at National Alcohol and liquor factory.

1.3. Statements of the Problem

The stability of the fill levels is important to ensure product quality in beverage manufacturing process. This study also shows that an improper filling process can create significant waste, non-compliance with regulations, and harm a company’s reputation. The work of Ibrahim et al. (2023), demonstrate that inadequate process monitoring lead to both underfilling and overfilling, with measurable impacts on customer trust (22% higher return rates) and profitability. Although excessive filling of the bottles may lead to product loss and increased cost, the under filled bottles fail to meet the consumer expectations and is a loss-making aspect.

After that, Dubey et al. (2019) highlighted that because producers bear the costs of the surplus goods and potential regulatory fines, overfilling results in avoidable operational costs. First, it has a negative impact on the financial result and, second, it contributes to inefficiency when increasing the production scale. Furthermore, errors associated with filling in may contribute to an inconsistent product portfolio, which lowers customer satisfaction even further. One of the essential operational management tasks that need to be addressed in the context of the beverage manufacturing field is the performance goal, namely the enhancement of the operating effectiveness in the context of the realization of competitive advantages in the global market (Tadesse et al., 2024).

Although it is operationally significant, there is a high fill deviation at the National Alcohol and Liquor Factory during the filling process. Inspected records between early January and the end of February 2025 of a total of 35 production days showed an underfill and overfill rate of 27,972 and 18,648 respectively, which equates to about 1,332 units per day required to be refilled. All nonconformity bottles will have to be re-processed to meet quality and regulations which will result in low productivity, high labor needs and unnecessary costs of operation.

Repeated refilling also leads to acceleration of wear of vital filling parts and the use of more electricity and compressed air, thus increasing operating and maintenance costs. As the 2024 annual performance report of the factory showed, merely 85 percent of the intended production capacity was met, which implies that the filling errors, over-reprocessing, and the lack of efficient resource use are among the most important factors contributing to the output gap.

The main issues of the current filling process are the lack of real time, data-based monitoring and control. The fill deviations are normally identified once they have taken place and therefore do not allow corrective measures to be taken on time, which further exposes the product to losses, rework and equipment degradation.

To overcome these challenges, predictive and optimization-based approaches are needed that can anticipate filling deviations early and enable process-based proactive adjustment. Monitoring and control systems using machine learning have the potential to predict filling outcome, help dynamic process control, reduce reprocessing and material losses, use less energy and prevent mechanical stress on filling equipment. Therefore, the use of machine learning in the filling process is a more adaptable approach to both overfilling and underfilling challenges which ensures increased operation efficiency, longer service life and improved productivity.

1.4. Research Question.

The list of the research questions of the study.

1. What are the proportions of overfill, underfill, and acceptable fill in the bottle filling processes and how it impacts overall filling accuracy in the National Alcohol and Liquor Factory
2. What are the major process parameters affecting filling accuracy in the National Alcohol and Liquor Factory, and how does the variation in process parameters cause overfill and underfill during the filling process?
3. How can a machine learning model be used to accurately predict the outcome of bottle filling, and how can this be integrated with the filling process for dynamic process control?

1.5. Objective of the Study

1.5.1. General Objective

This study aims to optimizing bottle filling efficiency through a supervised machine learning approach to address overfilling and underfilling challenges at the National Alcohol and liquor factory

1.5.2. Specific Objectives

1. To determine the percentage of overfill, underfill, and acceptable fill in the bottle filling processes at the National Alcohol and Liquor Factory and to analyze the effects on filling accuracy.
2. To identify and analyze the key process parameters that influence bottle filling accuracy at the National Alcohol and liquor factory
3. To develop Machine Learning model for predicting filling out comes and Integrate with the filling process for dynamic process control.

1.6. Scope of the Study and Justification

This study focuses on the optimizing filling process efficiencies through machine learning approaches to address filling challenges. The scope was limited to the classification of filling outcomes (correct fill, overfill risk, and underfill risk) based on indirectly measured process variables since there is no direct volume measurement sensor.

This study was considered only bottle filling, as in the production of 40% ABV Ouzo and Vodka, so as to be able to provide a focused and thorough exploration of the topic of filling accuracy and efficiency. This emphasis is reasonable since the filling process is a sensitive activity, which directly influences the quality of products, compliance with regulations, and loss of materials underfill or overfill conditions.

The limitation to 40% ABV products guarantees uniformity in physicochemical characteristics of the materials like their density and viscosity, an aspect that reduces variability in the results due to other factors other than process performance. This enables the research to control the impact of operational parameters in the accuracy of filling.

The study also focuses on major process parameters, including valve timing, pressure, and the length of the cycle, as these parameters directly affect the results of filling and can be measured. The increase of the scope to cover other stages of production or product types would add into the picture the unnecessary complexity and make the analysis less clear. **1.7. Limitation of the Study**

This study is limited to a single production line at the National Alcohol and Liquor Factory and highlights on certain parameters of operation without considering others that may not produce a significant impact on filling accuracy. Second, due to limitation of historical data, the study produces synthetic data based on the data collected from the company.

1.8. Significance of the Study

The significance of this study is that machine learning model was developed and applied to improve the accuracy of the bottle filling, and this is one of the significant progressions in the integration of technology in the manufacturing process. The study shows how data-driven decision-making with the application of predictive analytics can result in a quantifiable increase in critical operation metrics, such as optimize process efficiencies.

The study offers practical applications for the bottle filling industry. The study where help process engineers and plant managers identify critical operation and process parameters, implement predictive controls, and improve quality of production. This can be used as a model for other manufacturing companies looking to adopt machine learning to address similar operations.

Academically, the study offers a good understanding of how machine learning model-based optimization methods can be applied in the bottle filling process, thus contributing to the body of knowledge on the industrial process control and manufacture systems.

1.9 Organization of the study

The study is presented in five chapters aligned with the research problem and objectives. Chapter One lays the foundation for this study by providing the background, problem statement, objectives, questions and significance of the research, the scope and limitations of the study, as well as the structure of the thesis. A detailed theoretical and empirical literature review, spanning studies on the implementation of machine learning for industrial processes, especially in the beverage industry, is provided in Chapter Two. It also identifies techniques for generating synthetic data and highlights research gaps that this study aims to fill. Chapter Three outlines the research design, stating the data sources used and methods for generating synthetic data, data analysis, machine learning model development, developed model integration method. It also describes supervised classifiers, experimental design, and model assessment metrics with references and standards. Chapter Four presents the Analysis and Discussion section, with an analysis that interprets findings in relation to the study's aims and existing scholarship. Chapter Five presents the study's conclusions and recommendations, including main contributions, recommendations for practice, and recommendations for further research.

CHAPTER TWO

LITERATURE REVIEW

Manufacturing has dramatically changed due to technological advancements, particularly in the areas of artificial intelligence (AI) and machine learning (ML). These technologies allow for improved efficiency of processes, quality control, and decision making within industries. The following is a review of literature on the application of machine learning (ML) in optimizing bottle filling operations, particularly at the National Alcohol and liquor factory.

2.1 Process Optimization and Efficiency

In the beverage industry, especially in large-scale and mass production companies such as National Alcohol and liquor factory, the bottle filling needs to be efficient and accurate for quality control and regulatory purposes and to avoid losses in the production process. Two concepts that are fundamental to the overall production performance, namely, process optimization and process effectiveness, are interconnected concepts. Process optimization is concerned with the functioning of the system, while process effectiveness is concerned with achieving the outcomes of the system. Both aspects need to be understood and addressed in order to overcome similar filling challenges that National Alcohol and liquor factory is currently facing. Optimization in beverage manufacturing companies such as the National Alcohol and liquor factory, which has mass production capabilities, is particularly important for maintaining product consistency, avoiding machine downtime, and preventing underfilling and overfilling, which can lead to noncompliance with regulations, customer complaints, and loss of profits (Ghobakhloo, 2018).

I. Concept of Process Optimization

Process optimization in industrial manufacturing refers to a systematic process of improving the performance of a given process by manipulating controllable input variables to achieve a desired output. Based on systems theory and control engineering, process optimization conceptualizes a manufacturing process as a dynamic system in which inputs such as machine speed, product temperature, feed pressure, bowl pressure, and valve timing are closely linked to outputs such as fill volume accuracy, throughput, and product quality conformity.

Optimization has historically been achieved using deterministic approaches such as proportional integral derivative (PID) controllers, Statistical process control (SPC) and Design of experiments (DoE). These techniques are based either on predetermined logic or on statistical assumptions; thus, they cannot perform well under conditions of high variability or nonlinearities in the system (Sullivan et al., 2012). Also, these traditional methods don’t adapt to unplanned disturbances and involve significant human intervention to address issues such as temperature variation, carbonation variation, or mechanical drift, typically found in high-speed beverage filling lines.

As a result of the increasing prevalence of automated and intelligent control in industry, the focus has shifted to developing data-driven optimization strategies. For example, Salah et al. (2022) used a sensor-equipped filling system, working within an Industry 4.0 ecosystem, to improve the bottling of yogurt. Eliminating multiple filling points, and instead filling from one controlled point, resulted in major improvements in accuracy and cycle time, allowing for a clear demonstration of how sensor fusion and real-time feedback can increase operational efficiency.

Likewise, deep learning has emerged to be used as a process optimization tool. Gao et al. (2024) proposed the use of a deep neural network architecture using sensory information, such as temperature, vibration, and acoustic signals, for predictive control of manufacturing lines. This resulted in a significant reduction in downtime and raw material waste as control actions were taken immediately to address anomalies.

Machine vision applications have extended the use of filling level detection for bottle filling operations as well. Miah et al. (2021) developed a CNN-based vision system for monitoring the fill level of beverages in bottles. Overfilling and underfilling were then identified with high accuracy using digital image processing techniques such as segmentation and morphological operations. It transformed quality assurance, enabling immediate correction.

Alternative methods for non-invasive defect detection have been investigated as well, including acoustic analysis. Similarly, Zhou et al. (2021) introduced a system called

Learning ADD that utilized Hilbert-Huang Transform (HHT) and feature extraction to identify faults in beverage bottles prior to filling. The model achieved an F1-score of 98.5% when implemented on an edge device, highlighting the trend of edge AI in industrial quality control. All of these latest developments represent a shift from optimization via inflexible, rule-based decisions to adaptive, intelligent systems that can learn from historical process data and react to real-time fluctuations. The combination of sensor networks, machine learning and edge computing has allowed for even more accurate, predictive and autonomous control of the processes, which is especially applicable in high-throughput and quality-sensitive spaces such as beverage manufacturing.

II. Process Efficiency

In beverage manufacturing, process efficiency can be understood as the capability of manufacturing systems to convert input materials, mainly water, syrup and carbon dioxide, into a final product with minimal waste, downtime and variation. This can be seen as a very important factor that influences the performance of operations and profitability in general. Common measures of efficiency include the throughput rate, cycle time, Overall Equipment Effectiveness (OEE), and resource utilization (Muchiri & Pintelon, 2018).

For bottling operations, process efficiency is affected by equipment downtime, the synchronization and efficiency of conveyor systems and the accuracy of the filling mechanisms. This can cause loss of efficiency through, for example, excessive or insufficient stoppage, foam generation, poor timing of valves, insufficient pressure control, which can result in underfill, overfill, or spillage (Kagermann et al., 2013). These are a waste of resources and lead to rework and poor customer satisfaction.

As a result, modern beverage plants have experienced greater efficiency in their operations. State-of-the-art sensor technologies and process data enable the monitoring of filling pressure, flow rates and product temperature continuously and help decrease process variability (Lee et al., 2015) . Moreover, machine learning-based predictive maintenance plans reduce the chances of machines becoming inoperable and, as a result, optimal operational conditions experience fewer disruptions (Tzeng et al., 2019).

In developing countries such as Ethiopia, process efficiency is negatively impacted by old machines, low levels of automation and unreliable utility supply. It has been noted that efficiency in this regard is best achieved through a combination of improved equipment and data-driven monitoring solution (Tesfaye & Tadesse, 2021) . Improving process efficiency is then a positive strategy since it will also affect production costs, sustainability, and global beverage market competitiveness.

2.2 Bottling Process and Filling Accuracy

The beverage industries bottling process is an essential and major portion of their production line and consists of three main processes: rinsing, filling and capping. The two processes are interrelated and must be well synchronized in order to ensure product quality, regulatory compliance, and efficient production. Ammar et al. (2019) state that the filling stage can be affected by machine adjustment, container's geometry, liquid's nature (e.g., viscosity and density) and environment (e.g., temperature and pressure).

The most relevant stage is filling stage it directly impacts the quality of the product delivered to consumers. Deviations in filling accuracy not only fall outside regulatory specifications, thereby jeopardizing manufacturers, but also have economic and reputational impacts. International standards also keep filling accuracy under strict control. According to the International Organization of Legal Metrology, OIML R87, and European Union Directive 76/211/EEC, tolerances for prepackaged liquids state that underfilling is not permitted, and overfilling may not exceed allowable limits (OIML, 2004; European Commission, 1976).

The challenge is made complicated in the case of Beverage industry, like the ones manufactured by National Alcohol and liquor factory, since the product is pressure sensitive. Ding et al. (2021) note that to keep the pressure at an acceptable level to maintain taste and shelf life, it should be kept on very narrow margins, and even a slight variance during the filling operation can lead to excessive foaming, wrong volume output, or pressure loss that all can potentially ruin the integrity of the product.

The effects of filling inaccuracies are highlighted in the previous studies. According to Zhang et al. (2020) , although overfilling causes fewer complaints in consumers, it still results in unnecessary product giveaway That is why it causes considerable losses in the long term. Conversely, underfilling is dangerous to not only violate, but also erode consumer confidence and brand image, which are detailed in (Wang et al., 2018) . To illustrate, the food safety and standards regulating bodies of most countries insist that bottling companies observe strict volumetric tolerances in pre-packaged drinks like in Ethiopia.

Bottling lines traditionally use PLC-based (Programmable Logic Controller) automation, and, among other parameters, perform the filler valve operation with closed-loop control. These systems are the best in reliability, though in most cases, they lack the adaptability. Conventional PLC logics are not able to explain the actual processes variations, that is, misalignment of the bottles, pressure variations or variations on product viscosity under temperature changes (Grzelak et al., 2022). This will produce rigid parameter settings that might not be consistent to the present production reality.

Some sophisticated bottling solutions are just starting to incorporate sensor-based responds and real-time analytics to deal with this. As an example, Chevtchenko et al. (2023) discovered that 15 % of fill-level controls are enhanced when flow meters are combined with pressure transducers and vision systems as opposed to systems composed entirely of time-based filling systems. Nonetheless, there is still a need in human operators, who should define the interpretation of data and set parameters manually, in these systems.

Recently, there is a study on the improvement of the bottling process by employing machine learning (ML). Researchers used ML algorithms (classification model) and managed to train them on previous filling data and use it to predict and correct possible mistakes like overfill or underfill (Miah et al., 2021) . These systems will learn dynamically all the time based on a new data and thus will not require a person to re-optimize their system in different circumstances.

In addition, Liu et al. (2020) showed the isolating possibility of how filling head pressure, valve timing, conveyor speed, and liquid temperature should be modeled, and indeed how these variables develop non-linear relationships with fill accuracy: this can be used with neural networks. Their own prototype system had over 93 % at correct fill classification when beverage drink bottling was happening at real time.

I. Bottle Filling Technology

a. Gravity Filling

Gravity filling uses hydrostatic pressure to pour water from a reservoir into the bottle. It is often used in low-viscosity non-carbonated spirits and wine. Gravity filling is easy and cost-effective but is limited in high-speed operations. because flow control and accuracy are poor flow control and accuracy leading to fill variability. Lefebvre et al. (2018) found that gravity fillers are suitable at low throughput and for low foam liquids.

b. Pressure Filling

Pressure filling is applied to propel the product through the bottle to increase flow and flow rate. It is particularly suitable for bottled beverages, such as beer or sparkling wine, where internal pressure is manageable, since it prevents CO2 loss. Schöffski and Löwer (2019) pointed out that precision in pressure control is important for minimizing foaming and maintaining consistent fill volumes in pressurized liquids.

c.Vacuum Filling

Vacuum filling has become a standard practice in the liquor industry because it reduces foaming and maintains uniform fill levels. This is achieved by the pressure difference between the vacuum and atmospheric pressure applied to the bottle or filling bowl. Zakhvatayeva et al. (2024) experimentally demonstrated that the fill accuracy in vacuum systems is highly dependent on valve timing, vacuum pressure, and fill time, particularly when using high-speed multi-head rotary configurations for spirits such as vodka, whiskey, and rum.

d. volumetric Filling

Volumetric filling produces a predetermined volume of liquid that is supplied by pistons or measuring chambers, and are highly accurate. Gupta and Singh (2022) argued that volumetric fillers maintain precise fill volumes in multiple runs but are sensitive to fluid properties such as temperature and viscosity, which make them crucial for reproducibility at high production speeds.

e. Flow meters-based filling

Flow-meter-based filling uses continuous flow sensors, such as electromagnetic flow meters, thermal mass flow meters, or Coriolis flow meters, that measure the volume of liquid that flows into each bottle. This method is highly accurate and versatile, making it an ideal choice for modern beverage packaging lines that handle a variety of products and fluid characteristics. Industry research supports the conclusion that flow-meter fillers significantly reduce filling errors, since they are used with automated control systems (Ace Filling Systems, 2023).

2.3 Operational Parameters Affecting Filling Accuracy

The ability to fill accurately and consistently within the alcohol and liquor industry is an essential component of product quality, regulatory compliance and consumer safety. Most factors that affect the filling process, including vacuum pressure, inlet pressure, temperature and filling methods, influence filling accuracy. The literature review summarizes and simulates current studies on these parameters and their impact on filling accuracy.

I. Temperature control

The temperature control, which occurs during filling, is a decisive element of reducing volumetric foam and maintaining solubility. Foam has an influence on optical and pressure sensor, and may create a misread in fill detecting systems. The study by Zhang et al. (2022) revealed that neural network models work in real-time in controlling the volume of foam when trained with historic temperature data and adjusting the cooling rate. To the same extent, Miah et al. (2021) included regression-based machine learning (ML) to forecast idealized temperature setpoints, and they demonstrated that it bears strong correlation with lessened fill differentiation. Huang et al. (2020) optimised an edge-computing structure to intervene continually in the temperature of the product getting a better fill accuracy. Singh et al. (2019) developed correlations between ambient and liquid temperature dynamics and foam behaviour and suggested predictive models of thermal processes. This was reinforced by Yoon and Park (2018) who employed PID-based feedback loops with the learning algorithms to stabilize fill temperatures.

II. Vacuum Pressure

Vacuum pressure is an essential component of the filling process of liquids. The results of Anderson et al. (2021) provide evidence that optimal vacuum pressure is a key factor of the increased fluid flow rate during filling. Their research shows that precise vacuum control reduces air entrapment, an important factor associated with measurement accuracy. For example, Yang and Lee (2020) show that maintaining a consistent vacuum level reduces air bubbles in the liquid during filling. This is especially important in alcohol production because the presence of air may result in oxidation and impact flavor and quality.

III. Product inlet Pressure

Inlet product pressure is another factor that influences filling accuracy. Chen et al. (2019) show that suitable inlet pressure is especially important in order to keep liquid transfer to containers smooth. It is essential that the inlet pressure prevents excessive foaming, especially in alcoholic beverages, where foam can disrupt filling and lead to errors. The interaction between inlet pressure and vacuum pressure, which influences the flow rate and consistency, was also assessed by Garcia & Tran (2021) . Their findings show that most advanced monitoring systems can dynamically adjust these pressures in real-time, ensuring optimal filling conditions.

IV. Filling speed

Research conducted by Lee et al. (2022) emphasized the importance of fill speed and accuracy. In addition to the sensitivity of volume measurements, they investigate the relationship between volume filling speed and volume measurement accuracy in they investigate the nature of fine discrepancies in high-viscosity products. the nature of fine discrepancies in high-viscosity products. In contrast, Johnson et al. (2023) argue that it is advantageous to have a slow fill rate to ensure better control and accuracy. One of the problems with high-speed production environments is that they require systems to meet two key requirements: precise fills and efficient throughput.

V. Valve Timing Accuracy

Consistency volumetric control in automated filling systems is basic in valve timing. The difficulties in observed fill volumes may result when there are variabilities in the opening or closing delays, at the scale of milliseconds, also. Miah et al. (2021) applied predictive controller models with those performance curves to provide greater precision in timing. Park et al. (2020) therefore utilized the concepts of reinforcement learning to find the optimal valve actuation sequences depending on outcomes of past fills. The study by Li and Zhang (2020) was carried out through deep learning networks in order to detect time patterns in the appearance of valve errors. Narayan and Singh (2019) introduced correction factors into the PLC logic to take into consideration valve wear, and hysteresis. To reduce the valve-related fill discrepancy by more than 20 percent, Thomas et al. (2018) indicated that the use of machine learning-driven compensators helped to eliminate the presence of valves.

VI. Sensor Accuracy

Sensor accuracy is a parameter of primary importance for filling operations in automated bottling processes since the controls are based on the sensor measurements (level, flow, pressure, temperature). Considerations regarding sensor selection, positioning and the mitigation of disturbances such as foaming and turbulence have been proven in the literature and industry experience to greatly improve level detection and minimize fill errors. Regular calibration against traceable standards such as ISO/IEC 17025 and NIST as well as automated calibration along with sensor-health monitoring limits drift and uncertainty in measurements, minimizes underfill/overfill and increases reliability in the process. Two best practices for maintaining accuracy with sensors and maintaining filling accuracy in production lines are to use redundancy, for example, using both level and flow data, and regular maintenance.

VII. Container Design

The design and geometry of containers significantly affect filling accuracy. Garcia & Tran (2021) report that particular container shapes, such as narrow-necked bottles, require precise filling. This technique calls for specific filling nozzles and methods to ensure filling accuracy. The way containers are sealed can impact filling. Brown et al. (2021) discuss the benefits of certain closures in filling containers to improve the overall process and prevent spills and errors.

VIII. Filling Method

The filling method is also important in accuracy. In particular, Chen et al. (2019) show that pressure filling methods produce higher accuracy than gravity filling methods. They find the advantages of releasing pressure filling in applications of alcohol products whereby it produces fewer inconsistencies because the ability to control liquid dynamics is higher. In addition, Yang & Lee (2020) point out that filling nozzles must be designed and sized appropriately for accurate fills. The correct adjustment to product features will minimize filling mistakes and improve operational efficiency.

IX. Environmental Factors

The factors such as ambient temperature and humidity may influence filling accuracy. Patel and Raghavan (2018) show the influence of these conditions on liquid behavior. For instance, humidity fluctuation may increase liquid density and influence fill levels. In addition, hygiene procedures during filling are essential, as Brown et al. (2021) report that incorrect cleaning lead to contamination that impacts product quality and overall fill accuracy. As a result, environmental control measures are necessary to minimize variability and ensure filling accuracy.

X. Regulatory Compliance

Another crucial factor that affects the accuracy of filling in the alcohol industry is regulatory compliance. Lee et al. (2022) note that labeling, dosage, and quality of alcoholic drinks are tightly controlled by strict regulations. As a consequence, accurate filling must be applied to meet these requirements, and proper filling techniques must be applied. The completion of filling operations also serves as a source of documentation of compliance and quality assurance for the industry, which provides additional documentation(Patel & Raghavan, 2019).

2.4 Quality Metrics in Beverage Manufacturing

Quality metrics are used in beverage manufacturing as important indicators of whether or not processes are stable, comply with regulations, and deliver a satisfactory product to consumers. Essentially, they help to control quality and safety of products and increase efficiency of operations. Previous research has pointed out that these measures, when managed correctly, can help minimize waste, maximize profit, and protect a brand's reputation (Patterson et al., 2020; Rao & Singh, 2019).

I. Filling Volume Accuracy

Fill volume accuracy guarantees that the volume contained in each bottle is at the correct level and is within the permissible tolerances. While underfilling can create complaints or even regulatory violations, overfilling raises production costs as a result of product loss. López et al. (2021) state that a 1% overfill in the production of large-scale beverages may cause a significant economic loss. These tolerances are set by international standards such as EU directive 76/211/EEC and OIML R87. Meurer and Bergmair (2018) showed that electronic and sensor-controlled filling systems achieved less than ±0.5% variation in the fill volume and therefore much better compliance.

II. Consistency and Repeatability

Consistency and repeatability result in a uniform product in terms of product attributes such as temperature level, density, viscosity, etc., from run to run. According to Chen and Huang (2020), a desired product consistency is fundamental for consumer loyalty. As an example, Nair et al. (2019) used statistical process control (SPC) methods to reduce variability in the filling process, thereby ensuring that the resulting fill levels are reproducible.

III. Microbiological Safety

Microbiological safety is an essential part of beverage quality assurance. Health concerns and product recalls due to spoilage by bacteria, yeast, and molds can be disastrous. According to Singh et al. (2018), microbial contamination ranks at the top of global beverage recalls. Silva and Oliveira (2019) showed a decrease of 30% in the contamination levels in beverage filling plants when microbiological monitoring was combined with automated sterilization systems.

2.5. Machine Learning Application in Manufacturing Process Optimization

Machine Learning (ML) has become a game-changing technology in manufacturing settings where it can be used to do predictive analytics, intelligent process control and adaptive quality control. Over the past years, Machine learning (ML) has been widely deployed in many parts of the industry, especially through predictive applications such as predictive maintenance, quality control and anomaly detection, and optimization. Of all possible varieties of ML methods, supervised learning algorithms (Decision Trees, Random Forests, Support Vector Machines (SVM), K-Nearest Neighbors (KNN), and Artificial Neural Networks (ANNs) become most frequently utilized basing on their performance in classification and regression tasks.

In particular, Li et al. (2022) designed a neural network model to forecast the defect pattern in the high-speed packaging lines. Their model was able to reach the classification accuracy of more than 90 %, which means that ML may be applied in dynamic manufacturing environments. Instead, Bui and Wang (2020) used data in real-time provided by the sensors-temperature, flows rate, and viscosity-as a training dataset, obtaining classification models able to predict and autonomously correct any deviation that may have occurred in the food and beverage processing lines, leading to the minimization of the waste rate and increased product consistency.

Rahman et al. (2021) utilized SVM-based methodology to determine the fill level in the liquid filling processes. The model had high accuracy of predicting overflows and underflows, when an analysis was carried out on fill head pressure, valve timing, and conveyor speed. In their study, they were keen to point out the power of machine learning (ML) in the ability to identify surpassed interactions between many variables that are most of the time not identified by traditional PLC-based control systems.

Choudhary et al. (2021) used Random Forest classifiers to do optimization of injection molding in plastic manufacturing. The report highlighted the fact that material defects were reduced by 23 % on the one hand and the production throughput improved by 17 % on the other. Such findings highlight the wider industrial use of machine learning (ML) in the field of industrial optimization more broadly than only maintenance and failure detection tasks.

Nevertheless, the vast majority of existing applications focus either on anomaly detection, predictive maintenance, or end-product inspection. There is a lack of research regarding volumetrically optimized bottle filling machine learning entrenched. Few studies have concentrated on the direct integration of machine learning to optimize volumetric accuracy in bottle filling process especially that of the carbonated soft drinks which pose a further challenge due to pressure sensitivity and foaming behavior. This underscores a visible research gap in the literature, namely in creating real-time and adaptive filling control mechanisms based on ML, which can be able to learn real-time data, exhibit multivariate sensor data, and act dynamically to change filling parameters.

Notably, Miah et al. (2021) have tried to solve this problem by creating a classification model that could be trained with historical data about the filling line, which would decide whether a fill would be an underfill, an overfill, or a proper fill. Their model reached them more than 90 percent correct and the amount of goods giveaway had diminished vastly, as well as rule breaking. In the same way, Liu et al. (2020) showed that by simulating operational parameters, such as fill head pressure, valve timing, and fluid temperature, it was possible to reach the normal classification rate of 93.2% during the filling of carbonated drinks with the help of neural networks, which testified of the ML potential of highly specific process control.

2.6 Application Machine Learning in The Beverage Manufacturing

Recently, use of Machine Learning (ML) in the beverage industry has increased, particularly for enhancing process efficiency and product quality, as well as predictive maintenance. The high levels of automation in modern beverage production lines, particularly bottling lines, provide a good fit for data solutions. But there is still lots of untapped potential of ML in this field, particularly in terms of real-time filling optimization and developing economies.

The use of machine learning (ML) in beverage manufacturing has been shown to be practical and advantageous in a few studies. For instance, Bui and Wang (2020) used machine learning (ML) models to track real-time data from sensors in beverage lines to allow for fast corrections of operational deviations. They found that the application of supervised learning algorithms could improve the repeatability of the filling process by adjusting to varying operational conditions, including liquid temperature and viscosity, and valve timing.

Li et al. (2022) proposed a defect prediction model for packaging lines based on neural networks with a classification accuracy of more than 90%, highlighting ML’s capacity to identify patterns and anomalies that conventional rule-based systems tend to overlook. In the same line, Miah et al. (2021) used an intelligent classification model for the identification of underfills and overfills based on historical data of the production process, showing the applicability of machine learning (ML) techniques to improve volumetric precision in the filling process as well.

But, despite this progress, the majority of applications remain in large-scale operations in developed countries that have a more advanced incorporation of the industry 4.0 technology, such as sensor fusion, real-time analysis capabilities and cloud-based control systems. These facilities have the infrastructure and trained workforce to integrate and maintain machine learning (ML) optimization. Zhang et al. (2023), for example, used a reinforcement learning model to control foam generation in carbonated beverage filling lines, decreasing waste by 28% while also achieving more consistent filling volumes.

On the other hand, many of the developing countries, such East African nations, struggle to incorporate machine learning (ML). As Tesfaye and Taddese (2021) state, "many manufacturing concerns in Ethiopia are unable to invest in modern equipment and machines; they suffer from a scarcity of a digital infrastructure and face a shortage of employees equipped with skilled training in data science and automation". That poses a huge limitation to being able to implement more complex machine learning (ML) models in actual local production settings like bottling plants such as the National Alcohol and liquor (NALF)

In addition, there is less emphasis on real-time control and optimization of the filling process, as is the case for existing studies on predictive maintenance and anomaly detection, particularly for carbonated soft drinks where high-precision control is necessary due to pressure sensitivity and foaming behavior. Existing systems use PLC logic, incapable of adapting to nonlinear interactions among several working variables.

2.6.1 Machine learning Techniques: Algorithms and their Application

Supervised machine learning (ML) constitutes a data-driven approach in which models are trained on labeled datasets, where both input features and output labels are known (Hastie et al., 2009) . In this manner, predictive models can be used to classify new observations or predict values of continuous outcomes. Supervised ML finds broad applications in the beverage industry for quality assurance, process control, and predictive maintenance. Examples of classification tasks are detection of defective bottles; fill levels of underfilled, correctly filled, and overfilled; and sensory quality predictions. Some of these involve regression to predict shelf life and chemical composition in order to ensure consistency and adherence to stringent regulations such as those laid out in EU 76/211/EEC and OIML R87 (Chandna et al., 2021).

Decision Tree (DT) are among the most interpretable of the supervised Machine learning (ML) methods. They partition the data by making hierarchical decisions on the features. Decision Tree (DT) are often used in beverage manufacturing for fill-level classification and the monitoring of important parameters of the process. Decision trees have the benefit of being transparent; operators and quality managers can comprehend the logic behind the decisions and justify the decisions made, which is useful for many production sectors that are highly regulated. According to Liakos et al. (2018), decision tree-based models are particularly useful for food and beverage quality control because of their easily understandable rules.

Random Forests (RFs) extend decision trees by generating an ensemble of many trees, using bootstrapped samples and averaging the predictions to improve robustness and reduce overfitting (Breiman, 2013). Random Forest (RF) have been used in the beverage industry for wine quality classification, fault detection in bottling lines during processing and batch classification into quality grades. This ability to cope with noisy data renders Random Forests appropriate for on-line process control in filling operations. Cortez et al. (2009) showed how Random Forests performed better than single decision trees while predicting the quality of wines based on physicochemical properties, further demonstrating the applicability of Random Forests for beverage quality classification.

Support Vector Machines (SVMs) construct the optimal hyperplane that separates the classes, and they are capable of doing linear and non-linear classification using kernel functions (Cortes & Vapnik, 1995). SVMs are commonly used in the beverage industry for sensory classification of beverages according to chemical composition or based on consumer taste panel results as well as predicting carbonation levels and detecting anomalies in the production line. SVMs have demonstrated to be a very accurate predictive model in studies related to wine quality, thus can be recommended for research and practice applications (Cortez et al,.2009). In Wine Quality support vector machine (SVM) provide high predictive accuracy making them a reliable choice for both research and practice applications.

Artificial Neural Networks (ANN) are nonlinear, flexible models based on the concept of biological neurons. ANNs are good at capturing complex relationships and are ideal for large datasets. Applications in the beverage industry have included quality prediction for wine and beer, classification of flavor profiles from spectroscopic or chemical data, and estimation of shelf life based on variable storage conditions. Although less interpretable than decision trees, ANNs “can have very good predictive performance, thus being most useful for processes believed to be complex and nonlinear.” According to Zupan and Gasteiger (1999), Artificial Neural Networks (ANN) are suitable for sensory evaluation even in the food and beverage industry, when there are several features at play that interact with one another.

XGBoost, Light GBM, and Cat Boost are Gradient Boosting Algorithms that are used to reduce errors in classification by adding ensembles of weak learners (T. Chen & Guestrin, 2016) . The beverage industry has used these algorithms in prediction of fill-volume, prediction of shelf life, estimation of microbial risk and consumer preference modeling. Gradient boosting algorithms tend to work better on large and complicated data than other tree-based algorithms because of their error correction strategy, which is iterative, and their capacity to detect the more difficult interactions among features. According to Li et al. (2021) , XGBoost is more accurate in predicting quality of beverages, in particular, using a combination of chemical, sensory and production parameters.

2.6.2 Supervised Machine Learning Model Train, Test and Evaluation

The current trend of implementing Machine Learning (ML) within industrial process industries is on the rise as conventional methods of control are unable to ensure the quality of products in dynamic operating environments. Traditional methods of statistical process control and fixed setpoint cannot represent complex and nonlinear dependencies between multiple process variables (Zhang et al., 2020). Conversely, ML-based solutions offer data driven models that can be used to model these interactions and hence improved prediction and control of quality results, especially in cases where the critical variables cannot be directly measured.

A. Machine Learning model train in manufacturing process

The training phase is fundamental to the creation of any ML model, whereby patterns are learned in a procedural way upon a sequence of labeled historical process data. The quality of datasets required to train successfully in the industrial context implies that these data sets need to capture the variety of the actual operating conditions, such as nominal, transient, and faults (Kumar et al., 2019). Preprocessing of data including its cleaning, normalization, and feature scaling is commonly recognized as an essential step to eliminate noise and make sure that the data model is learning strong patterns and not mere chance variation (Wang and Yu, 2021). Recent manufacturing quality prediction research has shown that state-of-the-art supervised learning methods like a Random Forest, Support Vector Machines (SVM), and deep neural networks prove superior to the traditional linear model in nonlinear process prediction (Qin, 2022). Indicatively, Li et al. (2021) demonstrated that ensemble techniques were highly effective at predicting faults in a chemical reactor as compared to the usual regression. In addition, domain-scale feature engineering, in which expert experience guides the design or modification of the input variables, has been shown to enhance the performance of MLs in complicated processes such as injection molding and continuous casting. The fact that algorithms and extensive preprocessing is the focus highlights the importance of prudent training addictions towards the credible model construction in the industrial setting.

B. Test, Evaluation and validation of Machine learning models

Training should be followed by testing and evaluation of the ML model to ensure that it can generalize to unknown data. Studies of industrial ML usage continuously emphasize the issue of overfitting where the model performs well based on the training data, but poorly in practice (Zhang and Cheng, 2020). Such methods as hold-out testing, k-fold cross-validation, and stratified sampling are also adopted to reduce this (Mohamed and El Morshedy, 2020). These strategies make sure that the behavior of the model has been a real predictive capacity in the unknown operating conditions. The regular metrics of quantitative evaluation such as accuracy, recall, precision, F1-score, and confusion matrices are commonly applied to evaluate the classification models in tasks of quality prediction (Zhao et al., 2021). As an example, in their research on predictive maintenance of machining processes, Gao et al. (2022) used cross-validation and various measures and showed that deep learning models outperformed traditional feature-based feature classifiers. In real-time process control, strict model analysis is especially important because making decisions with incorrect predictions might result in a faulty product or excessive operating risks (Jin et al., 2023) . Thus, the literature has a strong spread of cautionary testing and verification of performance according to the deployment of ML models in the industrial controllers or supervisory systems.

c. Classification model evaluation

Classification model evaluation measures how well the constructed model can correctly predict the class labels (e.g. underfill, normal fill, overfill). In the industrial filling process control, it is very important to evaluate the performance of the models as it may have an impact on product loss, non-compliance to regulation, or quality rejection.

1. Confusion matrix

Confusion matrix is a fundamental Metrix of classification model evaluation:

Table 2.1 Confusion matrix

	Predicted positive	Predicted negative
Actual Positive	True Positive (TP)	False Negative (FN)
Actual Negative	False Positive (FP)	True Negative (TN)

Source:

2. Accuracy

Accuracy measure the proportion of correct fill instance:

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN}$$

Although accuracy is intuitive, it may be misleading in imbalanced datasets (He & Garcia, 2009) . In the case of filling control systems, where there may be few defective cases, accuracy alone is not sufficient.

3. Precision

Precision measures how many of the cases identified as positive were actually correct:

$$\text{Precision} = \frac{TP}{TP+TN}$$

High precision is desired when the cost associated with false alarms is to be minimized, for example., the mistakenly identifying normal fill as overfill. Precision, thus, represents the reliability of the predictions (Sokolova & Lapalme, 2009).

4. Recall(sensitivity)

Recall measures how many true positive cases are correctly detected:

$$\text{Precision} = \frac{TP}{TP+FN}$$

In filling control, it is crucial to achieve a high recall value when detecting underfilled products since missing these underfilled products may lead to non-compliance with quality requirements. Recall measures the ability to detect (Powers, 2011).

5. F1-score

The F1-score is the harmonic mean of precision and recall and is calculated as follows:

$$\text{F1-score} = \frac{Precision*Recall}{Precision*Recall}$$

The harmonic mean penalizes extreme values of precision and recall, which are recommended when the costs of false positives and false negatives are both considered to be important (Sokolova & Lapalme, 2009).

6. Macro-Average Metrix

Macro-average calculates the unweighted mean across classes:

$$\text{Mean Precision}_{\text{macro}} = \frac{1}{n}\sum_{i=n}^{n} M_i$$

$$\text{Mean F1-score}_{\text{macro}} = \frac{1}{n}\sum_{i=n}^{n} M_i$$

$$\text{Mean Recall}_{\text{macro}} = \frac{1}{n}\sum_{i=n}^{n} M_i$$

Where n is number of class

Since all classes are considered equally important in the macro-average, it is relevant when the minority classes are of interest (e.g., underfill) (He & Garcia, 2009).

7. Weighted Averaged metrics

Weighted-average considers the imbalance of the classes by weighting each class according to its support (the number of samples):

$$\text{Precision}_{\text{weighted}} = \sum_{i=1}^{c} Wi * Precision_i$$

$$\text{F1}_{\text{weighted}} = \sum_{i=1}^{c} Wi * F1_i$$

$$\text{Recall}_{\text{weighted}} = \sum_{i=1}^{c} Wi * Recall_i$$

$$W_i = \frac{ni}{N}$$

n_i =number of samples in i class

N= Number of samples

The weighted averaging accounts for the distribution of classes in the real world (Powers, 2011).

2.7 Integrations of Machine learning Model with Manufacturing Process

Integration of the machine learning (ML) models in the manufacturing process has received a lot of coverage as industries undergo the transformation to intelligent and adaptable manufacturing systems in the Industry 4.0 paradigm. Traditional manufacturing control approaches are largely founded on fixed logic, statistical process control, and simplified first-principle concepts, which in most instances are unable to address complex nonlinear performance, variability among processes and equipment degradation. As previous research has shown, the ML-based models can address these drawbacks by acquiring data-driven dependencies between the process variables and the quality results and provide a better way to monitor, predict, and control the manufacturing system (Qin, 2022).

The initial studies of the integration of ML into manufacturing were concentrated on the offline uses of the technology, including quality prediction, fault diagnosis, and predictive maintenance. As an example, Kumar et al. (2019)demonstrated that managed ML models were capable of forecasting quality variations with the help of previous production data, which can be used to make useful suggestions on process enhancement. Nonetheless, recent works have placed more focus on online and real-time integration wherein the outputs of the ML models are implemented in production processes to aid in making operations decisions. Zhang & Cheng (2020) emphasized that the concept of ML-integrated systems allows a timelier reaction to disturbances in the processes than the classical approach to control, mainly in a multivariate context.

Another popular integration architecture identified in the literature is the deployment of the ML models on a supervisory or decision-support level, where it interfaces with the existing control systems including Programmable Logic Controllers (PLCs) or Distributed Control Systems (DCS). Here, ML models work with real-time process data and provide predictions or classifications, but control actions execution is still a part of the traditional control framework. Jin et al. (2023) have shown that this system of layered integration enhances the reliability and safety of the system since it does not break the existing control hierarchies, yet makes them better with data-driven intelligence.

Some of the key fields of ML integration in manufacturing processes are quality control and soft sensing applications. Critical quality attributes cannot be practically or economically measured in most industrial operations. ML-based soft sensors derive these attributes based on the indirect process variables to be in a position to continuously monitor the quality of the process and also detect deviations early. The study by Wang and Yu (2021) has shown that the implementation of ML soft sensors on the production lines achieved a major decrease in the variation of the products and higher adherence to the quality specifications. These methods are especially applicable to the processes of filling and packaging, where the quality characteristics of the process (e.g. fill volume, etc.) have to be determined indirectly in many cases.

I. Steps of Integrations of Machine learning Model with Manufacturing Process

The integration of the machine learning (ML) models into the manufacturing environments is normally implemented within a systematic and multi-stage architecture to provide reliability, interpretability, and compatibility with the established industrial control systems. Successful integration is always highlighted through previous researches as being not only model accuracy but also systematic alignment with process knowledge, data infrastructure and control architecture. The feature importance and explain ability analysis is performed before the model is incorporated into the manufacturing process in order to prove the model decision logic. According to Kim and Lee (2021), one of the requirements of industrial acceptance of ML systems is interpretability. The importance of features study makes sure that the human models are guided by variables that have physical significance and that there is a correspondence between the data-driven understanding and the process engineering expertise. The step is also useful in prioritizing the control variables and making the integration framework simpler.

The manufacturing control system is then integrated against the ML model, usually at the level of supervisory or decision-support level. In the literature, a layered architecture where the ML model communicates with the PLCs, DCS or SCADA systems is used instead of directly controlling actuators (Jin et al., 2023). The ML model in this configuration is used to provide predictions, classifications, or recommendations, whereas the implementation of the control actions is done through the prevailing industrial control logic which ensures safety and reliability. In more modern applications, the results of ML are fed into closed loop or semi closed loop control solutions. Li et al. (2021) prove that deviation risks can be identified with the help of ML-based classes and appropriate parameter manipulations are planned, which allows controlling processes on a dynamic basis. An adaptable response of the manufacturing system to process disturbances, equipment wears, and changes in operating conditions is provided by this step. The last stage is the validation of the integrated ML system in the actual production conditions and constant monitoring.

2.8 Synthetic data generation for Machine learning model in industrial process

I. Importance of Synthetic data

Synthetic data generation refers to the process of producing artificial datasets that authentically mimic real data in terms of statistical properties, structure, and dynamics. This has become particularly relevant within an industrial context, in which historical data regarding processes can be unavailable, unbalanced or proprietary (Chakraborty et al., 2020). Specifically, in the beverage industry where bottling and filling operations are ultimately precise and are held to international standards, it is convenient and reliable to use artificial data for the training of models, optimization of the systems and quality control. Synthetic data provides a means of experimentation under realistic scenarios and of zero-cost non-intrusive experimentation that supports machine learning decision systems to be deployed in the industry by allowing for experimentation without stopping the production line.

2.8.1 Method of synthetic data Generation

a. Rule-Based generation

This includes rule-based data generation which is one of the most commonly used methods in industry. Thresholds and operating ranges for process variables are established by using expert knowledge and industry-specific conventions. As an example, the fill volumes in beverage filling systems can be described as underfill (≤989.8 mL), correct fill (990–1010 mL), and overfill (≥1010.3 mL), based on the tolerance standards set by OIML R87 and EU Directive 76/211/EEC. In the same way, feed pressure (0.2 – 0.4 bar), product temperature (15-20 °C) and valve timing can also be translated into operational rules, to be followed in the creation of realistic synthetic records. These procedures make the output data compatible with engineering knowledge and regulations. According to Liakos et al. (2018) ; Cortez et al. (2009) an example of a successful application of rule-driven modeling to the prediction of wine quality based on physicochemical properties. The rule-based classification provided a clear and well-defined boundary between normal and abnormal operations something that is consistent with typical engineering logic and production capabilities (Choudhary et al., 2021).

b. parametric simulation:

Parametric simulation approaches extend realism by creating synthetic data sets in which the behavior of processes is simulated by sampling from statistical distributions that characterize the processes. Fill volume could be, for instance, normally distributed with the mean equal to the nominal target of 1000 mL, feed pressure could vary according to a log-normal distribution and product temperature variability could be modeled by a triangular distribution since we expect more sales during the season than during the off-season. This method accounts for the averages, while also incorporating randomness into the produced synthetic datasets.

Also, Patki et al. (2016) argued that, contrary to the assumptions behind non-parametric simulations, parametric approaches are more realistic in that they maintain natural variability in the data, and Zhang et al. (2021) highlighted the value of parametric approaches for sensitivity analysis of production systems, in order to test the robustness of predictive models across a range of different, changing conditions for model processes. Parameters were assigned a probability distribution to represent their physical nature and inherent uncertainty during operation, in line with previous research on industrial process modeling (Doukas et al., 2019; Fang et al., 2020).

c. Monte Carlo simulation:

Monte Carlo simulation is also widely used in industry. This approach uses repeated sampling from specified probability distributions to create large synthetic datasets that incorporate the uncertainty inherent in the process variables. Monte Carlo simulation, by varying product viscosity, valve timing and the pressure at which products are fed, can be applied, for example, in beverage manufacturing to simulate thousands of filling scenarios. It allows the study of normal and extreme outcomes; thus, enabling considerations of risk assessment and system optimization. Doukas et al. (2019) stated that Monte Carlo simulation is useful in food engineering for handling uncertainty and its application to beverage systems has been beneficial for predicting variation in packaging and optimizing quality control programs.

d. Advance Approach:

The two most recent developments in artificial intelligence for generative machine learning have been the Generative Adversarial Networks (GANs), and the Variational Autoencoders (VAEs). Generative Adversarial Networks (GANs) use a generator-discriminator scheme to generate very realistic samples and can be particularly useful in scenarios for which rare events need to be simulated, such as extreme overfill or leaking valves in a beverage process that are rarely present in real datasets (Goodfellow et al.2014 ; Frid-Adar et al.2018) . Depending on their configuration, VAEs can encode the input features into latent variables, which are later decoded into novel data samples, thus providing a probabilistic framework to help model nonlinear relationships among the process parameters. This has been used to create examples of industrial datasets consisting of variable sensor outputs and predictive maintenance datasets (Kingma & Welling, 2013; Xie et al.2022).

A further important approach is agent-based simulation (ABS), which simulates the interactions of autonomous agents in a system. In a beverage plant, the agents could be filler nozzles, conveyor belts, or bottles passing through a line of production. Agent-Based simulation (ABS) enables the analysis of emergent behaviors such as bottlenecks, misalignments or machine stoppages that result from the interactions of agents. The usefulness of agent-based simulation (ABS) for production and logistics was demonstrated by Macal & North (2010) and Negahban & Smith (2014) respectively for complex industrial supply chains.

Finally, hybrid approaches that combine rule-based constraints with stochastic or generative methods are increasingly advocated for the generation of industrial data. For instance, beverage fill volumes could be limited according to OIML R87 rules (rule-based), sampled from Gaussian distributions (parametric), and further augmented with GANs to simulate a rare anomaly. Analogous hybrid frameworks have been proposed by Chakraborty et al. in 2020, for which the authors highlight the advantages of increased realism and applicability of synthetic datasets in manufacturing when physical models are combined with machine learning.

Thus, in developing and producing models of industrial processes, rule-based simulation, parametric simulation, Monte Carlo simulation, generative model simulations (GANs, VAEs), agent-based modeling, hybrid models and variations of all of the above are crucial. In the beverage industry, these methods deal with the problems of sparsity, class imbalance, non-classification event simulation, and privacy concerns, which, among other applications, allow for effective training of supervised machine learning algorithms for applications including fill classification, defect detection, predictive maintenance, and quality grading. Combining these techniques makes the use of AI in industrial settings more robust and meets regulatory quality specifications.

2.9 Related Work in Bottle Filling Process

More recently, due to advances in data science and machine learning, it has become possible for researchers and manufacturers to look into developing intelligent systems to improve the bottle filling process. With the exception of integrating adaptive control systems, the focus of these studies is on achieving greater accuracy in the fill, minimizing the waste of material, and avoiding quality variations. While the use of ML is gaining popularity more specifically in the context of bottle filling optimization, in particular for Alcoholic drinks, it is still a relatively new research area with some important contributions developed only in the last five years.

Abubakar et al. (2022) investigated improving the bottling process via an experimental data-driven approach. They used machine-learning-based auto-parameter configuration, including regression learning and rule optimization to automatically adjust filling parameters. They concluded that machine learning (ML) based tuning could decrease variability in fill volume and provide more uniformity than manual adjustment. But the analysis was a general experiment in bottling and not specifically based on liquid factors, such as alcohol by volume (ABV) dependent viscosity or vacuum dynamics experienced in filling systems.

Kiangala & Wang (2018) have applied artificial neural networks (ANNs) to model filling behavior in automated beverage bottling lines from a supervised learning approach. To train the ANN to predict deviations in fill volume using short inputs like fill time and flow rate, historical production data were used. Prediction accuracy was higher than that of linear regression models. linear regression models. However, the results were only based on a few parameters and did not consider important mechanical degradation factors such as nozzle wear and valve delay in high-speed alcohol beverage filling.

Lillo Otarola et al. (2025) employed process modeling and optimization techniques in wine bottling to combine hydraulic modeling algorithms. Their work, although not purely machine learning (ML) based, provided data-driven modeling of dependencies for documenting process interactions. The findings indicated that control of multiple dependencies was achieved through coordination. However, the focus was on scheduling and throughput optimization rather than ensuring the accuracy of the filling process and process adjustments. Miah et al. (2021) from historical production data developed an intelligent classification model which classifies status of bottle fill as underfill, overfill, and correct fill. Their supervised machine learning based approach yielded over 90% accuracy and showed that models could be trained to be used for predictive control even when minor values of operational parameters such as valve timing or temperature were different.

Liu et al. (2020) reviewed the use of neural networks in beverage filling, looking at the non-linear components of temperature, pressure, valve actuation and conveyor speed variables. Their prototype system was able to predict fills with correct accuracy greater than 93% in real time tests, which demonstrates the ability of deep learning to identify complicated dynamics of the process. Zhang et al. (2023) also applied a deep-Q-network (DQN) reinforcement learning approach to tackle the issue of inconsistent foam formation and volume during carbonation. This approach resulted in an approach to minimize foam waste of 28% while consistently achieving the desired fill levels by dynamically modifying the nozzle pressure and fill speed from feedback obtained in the process. It highlights how reinforcement learn (RL) could be utilized for control of such processes in a filling operation that is very fast.

Zhang et al. (2021) used machine learning models to predict the accuracy of liquid filling in industrial bottling systems. The model relied on flows, valve opening times, and filling pressure to predict final fill volume. They found that supervise vector regression (SVR) performed better than most statistical models in terms of capturing non-linear relationships. These improvements are limited, although the study is focused on general liquid products and does not specifically address alcoholic beverages because the effects of viscosity and temperature may vary in other liquids. Chen et al. (2020) used random forest algorithms to analyze multivariate filling process data from beverage bottling plants. Using operational data, the random forest (RF) model identified key effects of overfill and underfill events and assigned values for valve timing, pressure change, and filling time based on their importance. The results suggested that ensemble learning techniques can successfully deal with nonlinear interactions. However, this study focused on fault diagnosis rather than real-time optimization or closed-loop control.

Chen and Li (2019) in a similar research used Feedforward Neural Networks (FNNs) on a pharmaceutical filling system that worked primarily on high-viscosity liquids such as syrups. The model offset the flow delay and attained a 15-percent increase in fill accuracy compared with standard gravimetric systems. Though this work is not part of a carbonated beverage, it is an indication of how ML is useful in handling variable filling conditions. Kumar et al. (2022) presented model-free reinforcement learning at an alcohol filling line. Their system had learned what speed a pump needed and when a valve should open so that after 1,000 cycles they cut the incidence of over filling the tank by 19 percent. This suggests the potential of the autonomous learning systems to achieve consistent increase in bottle filling processes.

Ahmad & Yusoff (2022), investigated the effect of feed line pressure adjustment on fill accuracy in distilled spirit bottling using empirical studies and simple statistical analysis to show that pressure regulation may reduce underfill rates; and they did not consider interactions with timing, vacuum pressure, product properties like alcohol content and viscosity. Similarly, Salem and El-Banna (2023) investigated flow control in beer and malt beverage filling lines with feedback control loops and classical proportional-integral-derivative (PID) tuning, which helped reduce fill variability and primarily utilized flow control and did not include other controlling variables such as valve timing, pressure, and mechanical wear. Tao et al. (2022) investigated the development of PI/PID controllers for controlling liquid levels in alcoholic beverage fillers through laboratory-scale experiments and control system simulations to show that level tuning increases the measurement accuracy, and their design focused on actuator response and liquid levels.

Omitting influences on the filling behavior of high-alcohol-content beverages. Other studies, such as Özkan & Şimşek (2019), implemented compensation algorithms for flow variables based on sensor feedback in alcohol beverage filling systems, Cheng & Li (2020) conducted empirical tests of flow variable variability to assess filling head parameters, although they mostly used flow variables as independent variables without adequately modeling their interactions. Furthermore, Statistical process control (SPC) techniques measure filling time variations in beverage lines (Ramos, 2022), while foam detection sensors use simple rule-based flow adjustments (Singh et al., 2021),and they do not consider complex dependencies that drive filling performance.

Such achievements notwithstanding, the consistency among the papers is the use of highly advanced technology including the use of real-time sensors, IoT-based systems, and powerful computation, which is frequently inaccessible in a development country bottling plant. As an example, Wang et al. (2022) compared both conventional PID controllers and a regression-based Machine Learning (ML) model in beverage lines and discovered that the fill variability reduced by 22 percent; nevertheless, the local adaptation is more cumbersome, as the system requires high-resolution sensors, which are not possible in a context such as Ethiopia. Moreover, little research is performed to deal with the scalability of Machine Learning (ML) model to different production environments or its economical implementation into mid-sized plants. It still lacks certain research that will show how it is possible to accommodate Machine Learning (ML) even within the PLC-based environment without the mass reconstruction of the system.

2.10 Summary and Research Gap

The previous literature on bottle-filling optimization from the alcoholic beverage industry has largely incorporated a small set of parameters, even with the acknowledged complexity of industrial filling systems.

Ahmad and Yusoff (2022) found that underfills in distilled spirit bottling lines were reduced by adjusting feed line pressure, via experimental tests and statistical analysis. But, the valve timing and vacuum stability were not used, nor were product properties such as alcohol concentration and viscosity. Similar to Salem et al. (2023), who used classical PID control to control flow rate in beer and malt beverage filling systems to achieve improved volume consistency, flow as an isolated variable and overlooked mechanical variables such as valve response delay and nozzle wear. Tao et al. (2022) evaluated the effect of PI/PID controller design on the control of liquid levels in an alcoholic beverage filler and showed enhanced level attainment accuracy, although the PI/PID controller design emphasized actuator control, with the pressure-temperature-viscosity interactions important to the high-alcohol content liquor filling process.

Other studies, such as Ozkan and Simşek (2019) and Cheng and Li (2020), have used sensors to measure pressure and flow in addition to experimental testing to evaluate filling pressure effects, although the models have used in those studies only considered individual factors and failed to model dynamic interactions. More advanced approaches, including statistical process control of fill-time deviations (Ramos et al., 2018) and rule-based systems with foam detection sensors (Singh et al. 2021), enhanced monitoring, but were reactive and inflexible. Although recent studies have begun to use machine learning algorithms based on artificial neural networks; support vector regression and random forest models to predict deviations in drink bottling (Li et al., 2021; Wang et al., 2022), many of these applications have been standardized for beverage systems and rely on fewer parameters, focused on predictions rather than multi-parameter optimization and dynamic process control.

Table 2.2 Summary of Literature Gap

Study	Parameters Considered	Tools Used	Gap
Ahmad and Yusoff (2022)	Feed line pressure	Experimental tests, statistical analysis	Did not consider valve timing, vacuum stability, and product properties like alcohol concentration and viscosity.
Salem et al. (2023)	Flow rate	Classical PID control	Focused on flow as an isolated variable; overlooked mechanical variables such as valve response delay and nozzle wear.

Tao et al. (2022)	Actuator control (pressure, temperature, viscosity)	PI/PID controller design	Emphasized actuator control while neglecting interactions between pressure, temperature, and viscosity in high-alcohol filling.
Ozkan and Simşek (2019)	Pressure, flow	Sensors, experimental testing	Considered individual factors without modeling dynamic interactions affecting filling.
Cheng and Li (2020)	Pressure, flow	Sensors, experimental testing	Similar to Ozkan and Simşek; failed to model dynamic interactions.
Ramos et al. (2018)	Fill-time deviations	Statistical process control	Reactive and inflexible approach without proactive dynamic control.
Singh et al. (2021)	Foam detection	Rule-based systems	Reactive and inflexible; did not utilize predictive capabilities for dynamic control.
Li et al. (2021)	Various beverage system parameters	Machine learning (ANN)	Applications standardized for beverage systems, relying on fewer parameters and lacking focus on multi-parameter optimization.
Wang et al. (2022)	Various beverage system parameters	Machine learning (SVR, random forest)	Similar to Li et al.; focused on predictions instead of dynamic process control and optimization across multiple parameters.

Source: Literature review

As a result, a major gap is evident in the development of integrated machine learning-based dynamic filling control models that combine mechanical factors, process parameters, and product properties for alleviating filling problems encountered in the bottling industry for liquor and alcohol products. This is addressed through the model of nonlinear interactions within multiple operational domains and the use of Random Forest algorithms that can be applied to the prediction fill outcomes and dynamic optimization of filling accuracy in real industrial conditions.

CHAPTER THREE
RESEARCH METHODOLOGY

The chapter provides the methodology used in achieve the goals of the study, Optimizing Bottle Filling Efficiency with Machine Learning to meet Bottle filling Challenges. This chapter outlines the methodology employed in developing a machine-learning-based classification model to optimize bottle filling efficiency. The process is designed to be systematic and reproducible, as expected in engineering and in existing research studies on industrial processes, machine learning, and quality control. The entire process from identifying the data sources to evaluating the model's performance was structured to ensure that the results are robust, valid and reliable.

3.1 Research Approach

For purpose this study quantitative and qualitative research approach was employed. The quantitative data facilitated the develop, train and testing of machine-learning models as it allowed the structural analysis of process behavior, system performance, and parameter interactions. The qualitative aspect offered the interpretative information on the practice of operations, root cause of filling deviation, decision-making patterns, and the factors that form the background of the fill accuracy.

Combining these approaches, the study managed to make sure that the machine-learning models obtained were not only technically sound but also contextually based, understandable, and able to produce practically relevant and operational recommendations.

3.2 Research Design

The research design adopted in this study is a systematic exploratory and experimental research design to answer the five research questions stated in Chapter One, and the overall goal of improving the efficiency and accuracy of bottle filling using machine learning to control the dynamic processes. The research design is designed in four sequential and interdependent phases with each one having a clear research question.

The first step of the study focused on assessing baseline filling efficiency for the bottle-filling process during normal operation. This assessment relied on historical production data from the previous production run to evaluate how consistently the system performed against the target fill volume. The proportion of correctly filled bottles indicated by underfill and overfilling was analyzed as an indicator of overall process efficiency and compliance.

The second phase was focused on at identifying and analyzing the key operational and mechanical properties that affect filling accuracy. Exploratory data analysis (EDA) was used to investigate statistical analysis, distributions and correlation of key parameters with fill class and Root Cause Analysis (RCA) was used to identify a parameter related to filling inaccuracies. These findings were applied to engineering principles, industry standards, historical process data and expert opinion with technical staff from the case company. This collaboration of data and knowledge allowed the study to develop a model of process behavior that was reliable and provided an understanding of how important parameters contribute to the process of filling errors.

The third stage focused on the developing a supervised machine learning model for classifying fill-level outcomes into underfilled, correctly filled, and overfilled results. The operational insights and key parameters identified in Stage II guided the development of the model. It is possible to construct a synthetic dataset based on physical, historical data and process-recognition constraint. which captures real-world operational variability and parameter interactions. Standard machine learning processes were applied, including feature selection and engineering, algorithm selection, hyperparameter tuning, training and testing, and performance evaluation. Integrated the validated machine learning model into the filling process to assist in real-time monitoring and dynamic process control. During this stage, the model is not just a predictive tool, but also a decision support tool that is capable of identifying fill deviations and directing corrective actions during the filling process. This integration demonstrates the practicality of machine learning in an industrial filling environment and proves its suitability for stability and accuracy.

3.3 Research Framework

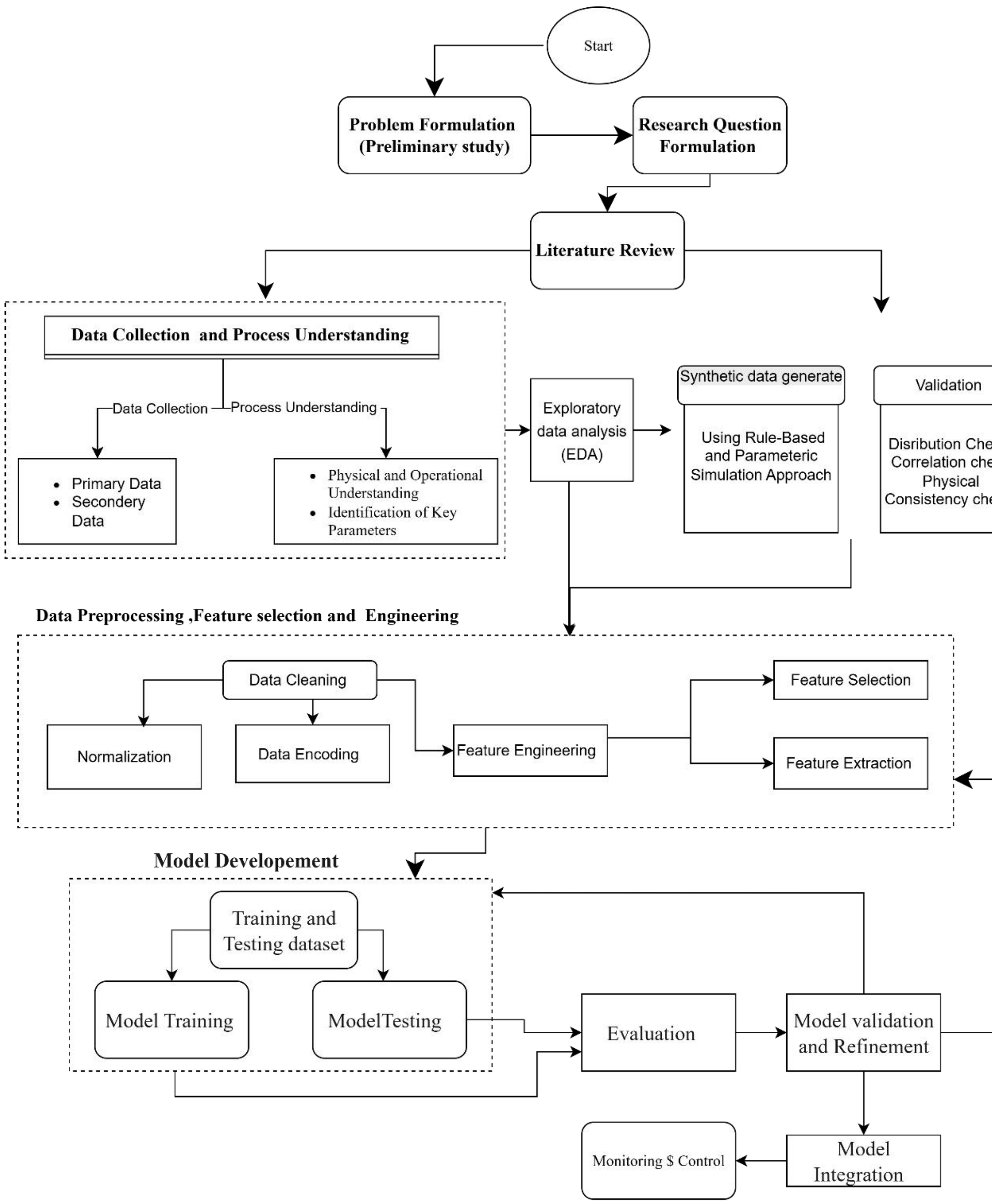
This research framework indicates a systematic way to build a machine learning model that predicts the outcomes of bottle filling. It starts with the problem formulation where the fundamental problem is clearly defined with a clear understanding of its meaning within the operational environment of the National Alcohol and Liquor Factory. The next stage will be the research question formulation phase as it is the formulation of particular questions that will lead the investigation including the orientation to the most important operational parameters and their influence on filling the accuracy.

The literature review is conducted to investigate the available literature, gaps in knowledge, and direct the methodological framework. With this review, researchers are able to base their work on proven results, and identify the way their study can add a new knowledge. The next step is the phase of data collection and process comprehension, comprising of exploratory data analysis (EDA) summarization of the data attributes in a visual form. Primary and secondary data are collected and the emphasis is put on the attributes, which are important both physically and operationally that affect the filling process.

Synthetic data generation is used in cases where genuine data are scarce, thus enabling the generation of representative sets of data which can be used to train the model. This also causes comprehensive data cleaning and preparation, which is an important task that includes eliminating inaccuracies, completeness, and feature engineering to form or reform variables necessary to enhance model performance. Normalization methods can also be used in order to bring the value of data into a similar scale to better learn the model.

The second step is developing the model, which involves a careful selection and extraction of features of data. A portion of the data is used to train the model, and another separate portion is used to test the model, which makes it robust and generalizable. After the development, evaluation is carried out to determine the performance of the model in relation to the measures that were set to provide a researcher with the ability to evaluate the success of the model.

After evaluation, the model is then validated and refined with the help of evaluation feedback which is in turn used to improve the model. Lastly, the tested model can now be applied to operation environments with the certainty of an application being practical, and it will deal with the challenges encountered in the real-world concerning accuracy of filling and efficiency.



Source: own, 2026

Figure 3.1 Research Flow

3.4 Data Collection, Consideration and Assumptions

The data collection was performed on the automated bottle filling station during production. To collect data describing only steady state behaviour of the process, certain conditions were imposed during data collection.

First, data were collected during stable production operation only. Instances of machine start-up, shut-down, product change-over, and maintenance were not considered in order to exclude transient effects that could influence the variability of the process. Second, only valid production recipes running at standard production speeds were considered.

For the purpose of this study, data were collected from both primary and secondary source to ensure complete coverage of the filling process.

I. Assumptions of the study

1.9 Assumptions of The Study

These assumptions are essential to maintaining a consistent analytical baseline, reducing process complexity, and implementing machine learning models using available industrial data.

1. Nominal Process Parameters: It is assumed that correct filling occurs when the filling process operates within the nominal process parameters. This assumption used in this study are that the bottle filling process is run under nominal process parameter conditions and the main variables involved like the timing of opening a valve, pressure and time spent in the cycle fall within their respective operation ranges under normal production. The assumption is that these parameters are constant and reflect typical operating conditions at which it is possible to conduct consistency in filling performance analysis.

2. Product property consistency: Considering that vodka and ouzo products share similar physical and chemical properties affecting the filling process, both products can be used with the same operational models. Also, the study presumes the product properties (40% ABV Ouzo and Vodka) are similar. The physicochemical properties of the products, e.g. density, viscosity is assumed to be constant during production process. This assumption is

important in that variations in filling performance are not considered as caused by change in product properties but on the parameters of the process.

II. Primary data collection

Primary data were collected from the bottle-filling line during the active production cycles at the National Alcohol and Liquor Factory. These this data were collected in order to determine filling current filling efficiency and to identify factors affecting filling accuracy in the normal operating conditions.

Primary data included real-time machine operating parameters and Quality inspection data were also taken to classify filled bottles into correct fill, overfill, and underfill. This data collected throughout 60 production shifts to observe regular variability in the processes. Direct observation of filling processes was conducted to record the behavior of equipment, process stability, and anomalies, including overfilling and underfilling. Structured interviews were conducted with seven domain experts involved in the filling operation at national alcohol and liquor factory. The experts include two filler machine operators, two maintenance controllers, two quality inspector and one production team leader who have a wealth of experience and responsibility in the production line. To minimize unintentional bias and to establish consistency in the line of questioning, a predefined interview protocol was prepared prior to the data collection. The interview protocol contained targeted questions on occurrences of frequent underfill and overfill incidents, valve behavior, pressure stability, control system, and mechanical wear aspects. Individual interviews were conducted to prevent influence from other respondents and their responses were recorded.

Structured discussion was carried out with seven domain experts involved in the filling operations at the national alcohol and liquor factory. The expert panel included two filler machine operators, two maintenance controllers, two quality controllers, and one representative from the production team. They were chosen due to their direct operational responsibility and technical experience with the automated filling system.

This was done in a series of moderated group discussions following a facilitation script. The moderator introduced the problem fill-level deviation and took the experts through each iteration of “why”. At each iteration, the cause identified was discussed among the group and the moderator only proceeded to the next level once consensus was achieved. There was no discussion of corrective actions as the intent was to capture only the causal relationships. All responses were recorded and the chain was verified with production, maintenance, and quality inspection records.

Primary data were collected from HMI by using video record (such as, Vacuum pressure and product inlet pressure during operation), by using Structured interview, structured discussion as well as manual quality checking tools during maintenance and production. The recorded data and interview results were supplemented with direct observation of the filling process.

All the data, both directly obtained and derived, were cross-validated with production records and quality records to provide the accuracy, consistency, and reliability in further analysis, modeling, and identification of root causes.

III. Secondary data

Secondary data were collected from factory records and documented sources to complement the primary data and provide historical and contextual information on filling performance.

The secondary data consisted of past production history, quality inspection reports, overfill and underfill records, maintenance report, standard operating procedures (SOPs), equipment manuals, industry guidelines, and published technical literature about the liquid filling process. These data were useful in comparing the present filling performance with the past trends and normal operating conditions.

Secondary data was collected through document reviews and archival data extraction to check for patterns and make the overall analysis more reliable.

IV. Data classification and labeling of filling outcomes

The process parameter data collected were categorized into three types of filling outcome correct fill, underfill, and overfill, depending on inspection and quality tolerance criteria set by the case company. These categories were informed by the records of documented limits of process parameters that denote tolerable operating ranges in normal filling conditions.

In the case of the correct fill category, 1,000 data points were taken each of the key process parameters (Pressure, valve timing, flow rate, fill time, cycle time and product properties) in the stable production process of the filling system working within the specified tolerance limit. This category was sampled larger since correct fill is the prevailing operating condition in normal production and that enough numbers of observations are required to represent the natural variability of the stable process operation.

In the case of the underfill and overfill conditions, 500 data points were obtained of each category. These samples represent working conditions where the key process parameters (Pressure, valve timing, flow rate, fill time, cycle time and product properties) were not within the nominal range which caused filling errors. Each deviation class was chosen to have 500 observations to ensure that there was a sufficient sample to represent the abnormal process behavior, as well as have a balanced dataset to perform later statistical analysis and development of machine learning models.

Besides production documentation and quality inspection records, structured interviews with domain experts such as filler machine operators, maintenance controllers, quality controllers, and production staff were carried out to determine prevalent working conditions that were related to fill deviations. The information gained through these discussions assisted in validating the parameter ranges that were related to underfill and overfill events and aided in labeling the deviation cases in the dataset. This integrated application of the process data that was recorded and the expert knowledge enhanced the credibility of the filling outcome classification.

3.5 Process Understanding

The filling operation of liquor and alcohol is founded on a vacuum-induced filling concept in this research, whereby, the operation of the transfer of liquid in a supply tank to the individual bottles is dictated by the interplay of vacuum pressure differentials, the dynamics of valve actuation, and fluid properties. The dissimilarity in pressure between the filling bowl and the interior of the bottle serves to be the main force in liquid movement. This driving force is however not autonomous since the start and end of flow is regulated by valve timing and mechanical response so the filling process is in itself dynamic.

An elaborate filling process revealed that there are important process variables that have a direct impact on filling performance. These parameter variations change the flow behaviour within the filling cycle and consequently, have an impact on fill accuracy and repeatability. Besides process variables, the mechanical integrity of the vital filling elements which include the filling nozzle, valve seals, and pumping elements is also an important process stability contributor. These components may undergo mechanical wear or degradation to lead to unintended residual flow following the act of closing the valves, and this can lead to systematic overfilling. These mechanical effects present non-ideal behaviour which cannot be represented with only setpoints.

Filling duration is usually determined to be that slot in time when the filling valve is open under vacuum conditions. Nevertheless, filling time is not the only factor that determines the effective volume to be supplied in this period. Rather, it is highly determined by balance of pressure, fluid nature and valve dynamics. Therefore, the same operating conditions and valve open times can end up delivering different fill volumes depending on the operating conditions.

Moreover, operational conditions including delay in the valve activation, inaccuracies during the indexing of the bottles and mechanical resetting between two consecutive courses influence the total machine through-put and fill consistency. These effects are even greater at higher production speed, where even slight timing variations can be carried to fill-level differences which are measurable.

It can be concluded that a detailed analysis of these interconnected parameters is the key to the creation of reasonable mathematical models and information-based models of the filling process. The process knowledge allows the definition of realistic parameter ranges, the use of physically enforced constraints, and the synthesis of synthetic data that provide an accurate representation of the real-life variability in operations, system constraints and dynamism in industrial filling settings.

3.6 Data Analysis Techniques and Tools

This Study used a statistical analysis, graphical visualization, and structure root cause analysis tools to analyze the performance of filling and to establish the key parameters influencing filling accuracy at the National Alcohol and Liquor Factory. The approaches taken were to guarantee systematic, repetitive, and physically significant analysis data of the filling process.

I. Descriptive data Analysis

To assess the existing performance of filling, statistical and line chart were performed on the data related to production that were gathered at the National Alcohol and Liquor Factory. The analysis was aimed at measuring the accuracy and variability of filling as a measure of operation. The central tendency and dispersion of the filling process as compared to the nominal target volume were determined through the calculation of the mean and standard deviation of the fill volume. The process consistency and identification of the general variability in filling performance were measured using these measures.

In order to determine the operational performance with time, the ratio of underfilled, overfilled, and filled bottles was determined in terms of percentages production per shift. Line charts with linear trend lines were used to compare the filling strength across 60 production shifts and bar charts were used to compare the percentage of overfilled, underfilled, and correctly filled bottles.

II. Root Cause Analysis (RCA)

Root Cause Analysis (RCA) was performed to move beyond performance observation and identify the underlying cause of filling inaccuracies. The Fishbone (Ishikawa) method was used to systematically separate the potential causes into Method, Process, Equipment, Product Properties, Environment, and Pressure, which emphasized factors directly related to flow rate or effective valve open time. Following the cause and effect analysis,5-Why analysis was used to link identified causes to a more meaningful root cause.

III. Parameter Distribution and Relationship Analysis

According to the results of root cause analysis (RCA), the main parameters influencing the precision of filling were chosen to be studied further. Histograms were used as the distribution analysis to investigate variability, skewness, and operation drift of these parameters. Assessment of parameter sensitivity was possible through visualization of parameter behavior at various filling conditions (underfill, normal fill, and overfill).

Then correlation analysis was conducted to measure the association between fill volume and the chosen parameters. The parameters that strongly affect filling accuracy and the ones that have weak connections were identified using correlation coefficients and weak correlations were removed. The joint distribution and correlation analysis were a measure that both physical rationale and statistical evidence were used to substantiate the parameter selection.

IV. Data analysis tools

Python was a powerful programming language for data analysis that provided libraries and tools for manipulating data, statistical analysis, and visualization as discussed in the table below.

Table 3.1 Python Library

	Library	Use
Python	Pandas	Import data, Organize datasets and correlation analysis
	NumPy	Perform mathematical calculation and Descriptive Statistics
	Matplotlib	Constrict Histogram and visualization correlation with scatter plot

As mentioned on table 3.1, Pandas was a library that handles and cleans data efficiently, allowing users to import, organize, and process data on-demand. NumPy consists of fundamental functions which can be used to compute numerical correlation coefficients and other statistical measures. Matplotlib is a visualization tool that makes plots clear and visually informative such as scatter plots to illustrate correlation and histograms to represent data distributions.

3.7. Synthetic Data Generation

In industrial filling systems, full and detailed historical data is often not available due to sensor bias, inconsistency in logging, and operational constraints. This study uses synthetic data generation for process optimization and machine learning models, synthesized samples are constructed based on real process knowledge, historical data, expert opinions, and physical constraints.

I. Parameter Selection

Synthetic data generation was conducted within processes parameters identified through structured interviews, direct observation, equipment specifications, and operation guidelines.

II. Synthetic Data Generation method

A probabilistic rule-based and parametric simulation approach was adopted. Baseline values were sampled within standard operating conditions under constrained probability distributions in order to reflect the natural process variation. Interdependencies between parameters were applied to ensure physical consistency. These include the effects of viscosity on temperature and the effects of pressure on flow rate and valve open time as well as closing delay on delivered volume.

class-specific biasing was used to achieve varied filling outcomes. The correct filling samples were placed within the target setpoint with little difference. Underfilling samples emphasized inefficient flow delivery, such as shorter valve opening times, higher viscosity, restricted venting, and delayed pressure buildup. Overfilling samples exhibited extended valve open time, increased closing delay, higher inlet pressure, residual post-fill flow and nozzle wear rate.

When these relationships are specified, probability distributions are selected on each parameter that gives directions on the statistical behaviour of the parameter in the simulation. Lastly, a parametric simulated representation of the process variables is done, based on the developed relationships and distributions to generate synthetic data which is representative of the dynamics of the underlying processes.

II. Data composition, Validation and acceptancy criteria

1. Data decomposition

The data used for this study were a composed of multisource and multivariate data collected from a National Alcohol and liquor production line. To adequately capture the dynamic nature of the filling process, process variables, mechanical variables, and product properties were included in the dataset, and were represented real-world applications in high-speed liquor filling systems. The data set contains three specific fill conditions, correct fill, overfill, and underfill, enabling regression-based classifications of filling states. To reduce class imbalance and increase model performance, also added historical production data and synthetic data generated based on historical data. Synthetic output reflects real process performance and standard operating ranges in real-world conditions and statistical distributions.

2. Data validation

Validation of the data was carried out before model design and, industrial realism. delivery to ensure reliability, consistency and, industrial realism. industrial realism. First, data validation was performed on synthetic data distributions against historical data in the form of descriptive statistics, distributions and variance analysis. It was verified that the major measures of mean, standard deviation, skewness, and range are within operationally acceptable limits.

3. Acceptancy criteria

The acceptability of the synthetic dataset was evaluated based on statistical similarity tests, correlation structure comparison and physical constraint validation benchmarking. The Go decision was defined as (i) mean deviation less than 5%, (ii) standard deviation less than 10%, (iii) preserved correlation structure within ±0.1 tolerance and (iv) no violation of machine operational limits. The synthetic dataset met all the criteria and was accepted as suitable for model development.

3.8 Data Preprocessing, Feature Selection and Engineering

I. Data preprocessing

In order to maintain data integrity physical feasibility and logical consistency, data level constraints were defined and applied before analysis and modeling.

Physical feasibility:

The following physical constraints were applied:

- Timing parameters (e.g., valve opening time, valve closing delay, filling cycle time) must have positive values. Negative timings were removed as unrealistic.
- The real fill volume needs to be above or equal to zero.

Logical consistency:

Rules were applied to maintain the sequence of a process:

- The valve opening time is typically before the valve closing time in a filling cycle.
- Filling cycle time should be exceeding the valve opening time.
- Parameter sets that disrupted the mechanical process order were considered invalid and removed from the dataset.

Lastly, data labeling was confirmed by identifying fill classes as underfilled, correctly filled and overfilled based upon past data and inspection reports and set tolerance levels. These preprocessing steps were used to make sure that the dataset was a good model of realistic filling behavior, and was to be used later to feature engineer and analyze using machine learning

II. Feature Engineering

The feature engineering was carried out to improve the representation of the filling process by considering relative and normalized characteristics and utilization-based features instead of using raw sensor values. This strategy enhances model interpretability and strength especially when used in different operating conditions. This was based on two important engineered features derived by the primary process variables which are valve delay ratio and cycle time utilization

Valve delay ratio was the ratio of the delay in response of the valve and the valve open duration. These characteristic measures the ratio of the non-productive valve time during the filling window and is a measure of the dynamism of the actuation of the valves. Deviations in this ratio suggests either delays in the mechanical response, wear or imprecision in timing the control and all of which directly affect the accuracy of the fill.

Cycle time utilization was obtained by taking the ratio of effective fill time to total machine cycle time. This characteristic is the utilization of the available cycle time, and it gives an understanding of the throughput constraints together with timing constraints especially during high operating rates.

III. Feature selection

A process of feature selection was conducted in order to determine which parameters involved in the process and product were to be employed to create the machine learning model of outcome classification filling. The original set of candidate parameters was developed using process knowledge about the liquid filling system at the National Alcohol and Liquor Factory (NALF), historical production data and literature available on liquid filling processes. The parameters that were taken into consideration were vacuum pressure, inlet pressure, cycle time, fill time, valve open time, valve delay, nozzle wear, viscosity, product temperature, density, and alcohol by volume (ABV).

Before the selection, data preprocessing was carried out to maintain data quality and consistency. This involved deletion of incomplete records, deletion of abnormal operating periods, including start-up and shutdown, and checking of parameter values by physically plausible operating limits. The selection of features was guided by correlation analysis was used to assess how each parameter relates to the expected fill volume. Parameters with negligible correlation or little significance in filling dynamics were not included in subsequent analysis. The rest of the parameters were kept as candidate features to be used to develop the model.

3.9 Machine Learning Model Development

The machine learning model was developed to classify filling outcomes as underfilled, correctly filled and overfilled based on the engineered and selected features. This classification problem was created as a supervised learning exercise using labeled data from historical quality inspection records.

I. Model Selection

Random Forest classifier was selected for its high predictability and ability to construct nonlinear relationships and complex interactions between process variables. Random

Forest algorithm was robust to noise and less likely to overfit, particularly when dealing with variability and uncertainty in industrial process data.

II. Training and Testing

The dataset was split into 70% training, 30% test data. This model was trained on historical production data, Synthetic data and multiple decision trees were constructed from random subsets of the training data and input features. The ensemble model allowed for complex, non-linear relationships between process variables and filling outcomes. thereby minimizing the risk of overfitting.

After training, the Random Forest model was tested using a different dataset that was not in the training phase. The testing process examined the model's generalization capability to unseen data and whether the predictions were biased against training data. A hold-out test set was used as an accurate measure of model performance in real-world environments.

III. Model evaluation metrics

Model performance was evaluated using standard classification performance measures. Classification accuracy was to evaluate the accuracy of fill outcome predictions. The precision and recall were calculated for each class in order to evaluate the detection of underfilled, correctly filled, and overfilled bottles. The F1-score was used to ensure a balanced assessment of class-wise performance, as the harmonic mean of precision and recall. Additionally, it was used to study misclassification patterns and identify error distributions for fill outcome categories through confusion matrix analysis.

IV. Model Validation

In order to ensure reliability and generalization of the trained Random Forest model for classify filling outcomes (underfill, correct fill, and overfill), a structured validation methodology was implemented. These validations were based on multiple performance metrics, stratified k-fold cross-validation, and out-of-bag (OOB) estimation.

First, macro-averaging and weighted-averaging for precision, recall, and F1 scores were used as scoring devices for the multi-class classification task underfill, correct fill, and overfill. The confusion matrix was used to calculate the first levels of precision, recall, and F1 score for each class. To ensure that all filling outcomes are equal, regardless of their frequency in the dataset, the macro-average was then taken as the unweighted average of the class-specific measures. Meanwhile, weighted averages were computed by combining the measures in terms of class-specific averages with their respective support numbers (number of samples per class) to provide a general measure of performance across classes. Both the averaging and sampling methods contributed to a comprehensive and balanced assessment framework for validation of the model.

Second, a stratified k-fold cross-validation was conducted for the training data to ensure robustness and minimal dependence on a single data split. This procedure divided the dataset into k equal parts with proportional distributions of underfill, correct fill, and overfill classes in each fold. The model was iteratively trained on k–1 fold and validated on the remaining fold until all folds had been used as validation data. All performance measures were averaged across folds to obtain a fair and unbiased estimate. A strategy was needed to balance class sizes and assess fair filling outcomes for minority groups.

Third, a Random Forest algorithm internal validation method of estimating out-of-bag (OOB) error was employed. Because every tree is trained using bootstrap sampling, a subset of observations is not included in the tree's training sample. These excluded observations (OOB samples) were used to estimate prediction error for votes from trees in which observations were not included during training. Additionally, out-of-bag (OOB) estimation provides a second objective assessment of model performance without further data partitioning.

3.10 Model Integration with Filling Process

The machine learning model was validated and then incorporated into the filling process, allowing for process control and decision-making in real-time. It is designed to operate within automated processes in the industrial environment, while also improving accuracy of filling and process stability.

I. Integration Architecture

The integration followed a closed-loop architecture in which the machine learning model acts as a decision support system. The system consists of the physical filling machine, process sensors integrated with the PLC, a machine learning inference module, and control feedback mechanisms.

This architecture allows for continuous process monitoring and a real-time evaluation of filling performance, without altering the existing automation and production processes.

II. Real Time data flow

A set of process variables were collected in real time from the data sensory during each filling cycle. These variables were pre-processed to ensure consistency, and then transmitted to the machine learning inference module. The trained model analyzed the data being received and predicted the fill level outcome of each bottle.

The inference process is designed to operate within the time constraints of high-speed filling, at a rate of 9000 bottles per hour, ensuring that the model's implementation avoiding delays.

III. Decision support and control actions

The machine learning model provided decision support signals to guide process changes based on the predicted fill class. Adjustments such as extending the valve open time or increasing the vacuum pressure were initiated as predictions indicated that the underfilled condition were being addressed. When overfill was predicted, flow reductions were made, and valve timing corrections were implemented. The correct fill prediction enabled the process to continue unimpeded.

3.11 Ethical Consideration

The study was carried out according to ethical rules. The information has been gathered using the express permission of the National Alcohol and liquor factory.

All company data was the production records, maintenance records, and quality reports, which were used exclusively in academic research purposes. The study preserved confidentiality and integrity of data. The data collected was limited to access and only accessed by analysis concerning this study.

In order to collect primary data, such as structured interviews and structured discussion with experts of a company, all participants provided informed consent. The participants were informed of the study purpose and their answers were collected and analyzed without mentioning personal identities. Participation was voluntary, and no individual was pressured to provide information. The study did not entail any manipulation of production processes that could harm product safety or regulatory issues.

CHAPTER FOUR

RESULT AND DISCUSSION

4.1 Background of Case Company

I. Brief description of Company

The National Alcohol and Liquor Factory is a government a state-owned enterprise established by Council of Ministers Regulation No. 74/85 (as amended by Regulation No. 415/2017) The factory that manages two manufacturing branch factories and one service branch. When the factory was established, it was established with a paid-up capital of 147,000 birr, and based on the improvement and expansion project undertaken by the company, the authorized and paid-up capital was increased to 622,673,899 birrs. This state-owned enterprise is over a century old and has branches located in different geographical locations, established at different times:

They are:

- Mekanissa branch establish 1957
- Akaki branch establish 1938
- Sebata branch establish 1906

II. Production capacity and products

National Alcohol & Liquor Factory occupies a pioneering status signaling a major breakthrough in the onset of beverage industrialization in Ethiopia over the last century. The factory improved its manufacturing technologies and production capacity through time and reached in producing 3.2 million liters of Pure Alcohol & 7.3 million liters of alcohol liquors annually during 2015 using its old machineries. But, at this time the factory realized that liquor products supply to the local market is short of supply in most cases.

National Alcohol and liquor factors products

- Denatured Alcohol and Pure Alcohol or Extra Neutral Alcohol (ENA)
- Baros Dry Gin
- Ouzo
- Coffee Liquor

- Super MINT
- Double Ouzo

III. Filling Method and Capacity

This study specifically focused on Mekanissa Branch, whereby, the process of filling is carried out with the help of an advanced vacuum filler. This process forms part of the delivery of the liquids with a high degree of accuracy, high levels of productivity as well as quality.

The vacuum filler can be highly efficient with an amazing average capacity of 9000 bottles per hour (BPH) which is supported by a 50-head design in such a way that a number of containers can be filled at the same time. Regardless of such capabilities, it is common to get fewer than 37,000 bottles per shift in the branch because of a several operational challenges such as machine adjustments, bottle shortages, contamination of bottle and filling error highly affected production efficiency.

4.2 Determine the Percentages of Filling Inaccuracy and Acceptable Fills in The Bottle Filling Processes, and Analyze Their Effects on Overall Filling Accuracy.

The NALF bottle filling line has both strengths and major weaknesses but it's the major weaknesses prevent the system from achieving optimal utilization of installed capacity. Inefficiency measures along with measures of performance such as throughput, time utilization, filling accuracy, rework and scrap, downtime, maintenance, and standard measures like line efficiency, offer a picture of the state of operations.

I. Accuracy Metrics

Filling Accuracy Rate (FAR): which measures the ratio of correct filled bottles to total filled bottles.

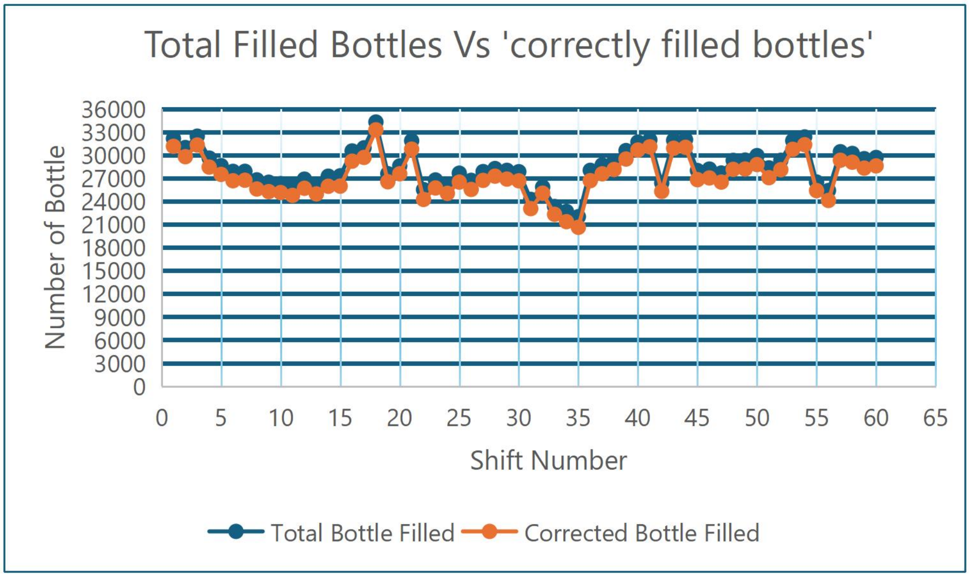
$$\text{FAR} = \frac{\text{Correctly Filled Bottles (CFB)}}{\text{Total Filled Bottles(TFB)}} *100 \text{............. (Equation 4.1)}$$

$$n = \frac{1{,}637{,}241}{1{,}706{,}374} *100$$

$$\underline{\underline{= 95.949\%}}$$

a. Comparison of Total filled bottles vs Correctly filled bottles over shifts

The figure 4.1 shows a high percentage of correctly filled bottles overall, but with this percentage dropping at a number of shifts (for example between 15-22 and 30-36), increasing percent rework for that period. While this indicates that the percent of correct fills is generally in the high 90s, these drops indicate that the percent of correct fills drops by a few percentage points at these shifts (increasing the defect level from around 1-3% to around 3-6%).



Source: Own, based on data from NALF (June-September 2025)

Figure 4.1 Total filled bottles vs 'correctly filled bottles'over shifts

II. Percentage of overfill and underfill Bottles

a. Percentage of overfill bottles (POFB)

$$\text{POFB} = \frac{\text{Overfilled Filled Bottles (POFB)}}{\text{Total Filled Bottles(TFB)}} * 100 \text{…………………. (Equation 4.2)}$$

$$= \frac{28891}{1{,}706{,}374} * 100$$

$$\underline{\underline{\mathbf{=1.693\%}}}$$

This indicates that the percentage of overfill was 1.693%, with information from 60 shifts. This suggests process improvement should also include overfilling as it can cost, cause waste, and impact overall operating efficiency.

b. Percentage of Underfilled bottles (POUB)

$$POUB=\frac{\text{Underfilled Filled Bottles (POFB)}}{\text{Total Filled Bottles(TFB)}}*100\text{............ (Equation 4.3)}$$

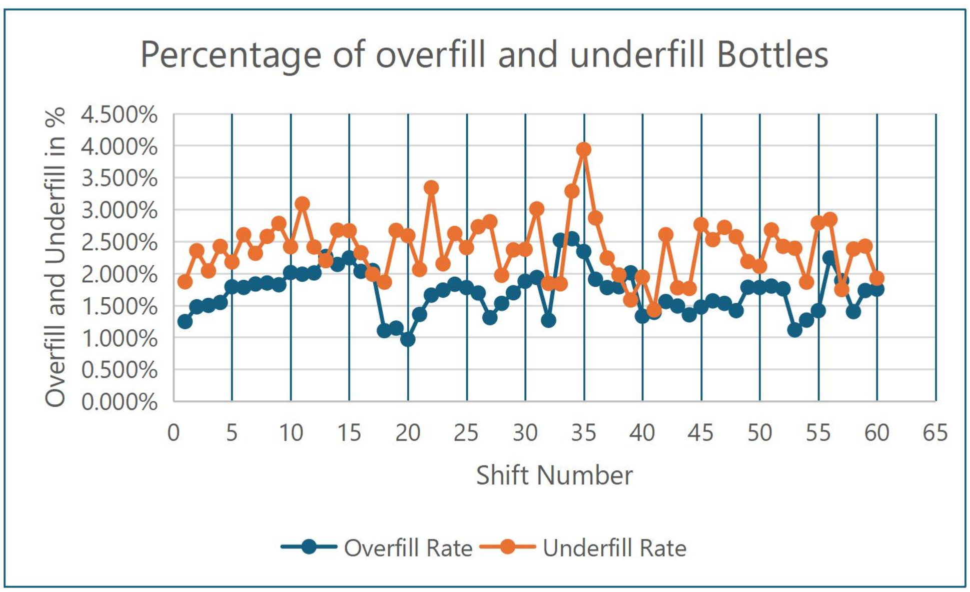
$$=\frac{40242}{1{,}706{,}374}*100$$

$$= \underline{\underline{\mathbf{2.358\%}}}$$

This indicates that the underfill rate over the course of 60 shifts was 2.358%. This suggests that the process requires improvement, as it might affect product quality and overall efficiency.

c. Comparison of percentages of overfill and underfill

The figure 4.2 shows percentage of overfilled and underfilled bottles across shifts. Both are seen to be varying but percentage of underfilled is greater than percentage of overfilled and exhibits sharp rises more often (around shifts ~18-24 and around ~33-36). Percentage of overfilled is usually around 1 to 2.5% whereas percentage of underfilled is around 1.8 to 3.5% reaching close to 4%.



Source: Own, based on data from NALF (June-September 2025)

Figure 4.2 Overfill and underfill rate

III. Descriptive Analysis of Filling Efficiency

Table 4.1 Descriptive Analysis of Filling Efficiency

Descriptive Statistics			
	N	Mean	Std. Deviation
Total filled bottles per Shift	60	28439.567	2544.4302
Underfill bottles	60	670.70	93.249
Overfill bottles	60	481.52	77.474
Correctly filled bottles	60	27287.350	2587.5890
Inaccuracy Filled Bottles	60	4.0976%	0.64753%
Filling Accuracy Rate (FAR)	60	95.90243%	0.647534%
Valid N (listwise)	60		

Source: Own, based on data from NALF (June-September 2025)

The table presents descriptive statistics of a number of variables associated with the bottle filling process over 60 shifts. On average, 28440 bottles were filled per shift, with a standard deviation of 2,544.43 bottles, so the number of bottles can vary greatly. The average underfill rate was 671 bottles per shift; the low standard deviation suggests that the low standard deviation of 93.25 indicates that process improvements are needed to reduce the underfill issue. On the other hand, the average number of bottles overfilled per shift was 482 bottles, with a very low standard deviation of 77.47, indicating consistency in overfilling, which may increase costs and waste.

Among underfilled and overfilled bottles, on average 27,287 bottles were correctly filled per shift, indicating the impact of corrective measures on total filling efficiency. This means that the average error rate in filled bottles was about 4.10% with standard deviations of 0.65%, indicating a persistent issue with inaccuracy across shifts.

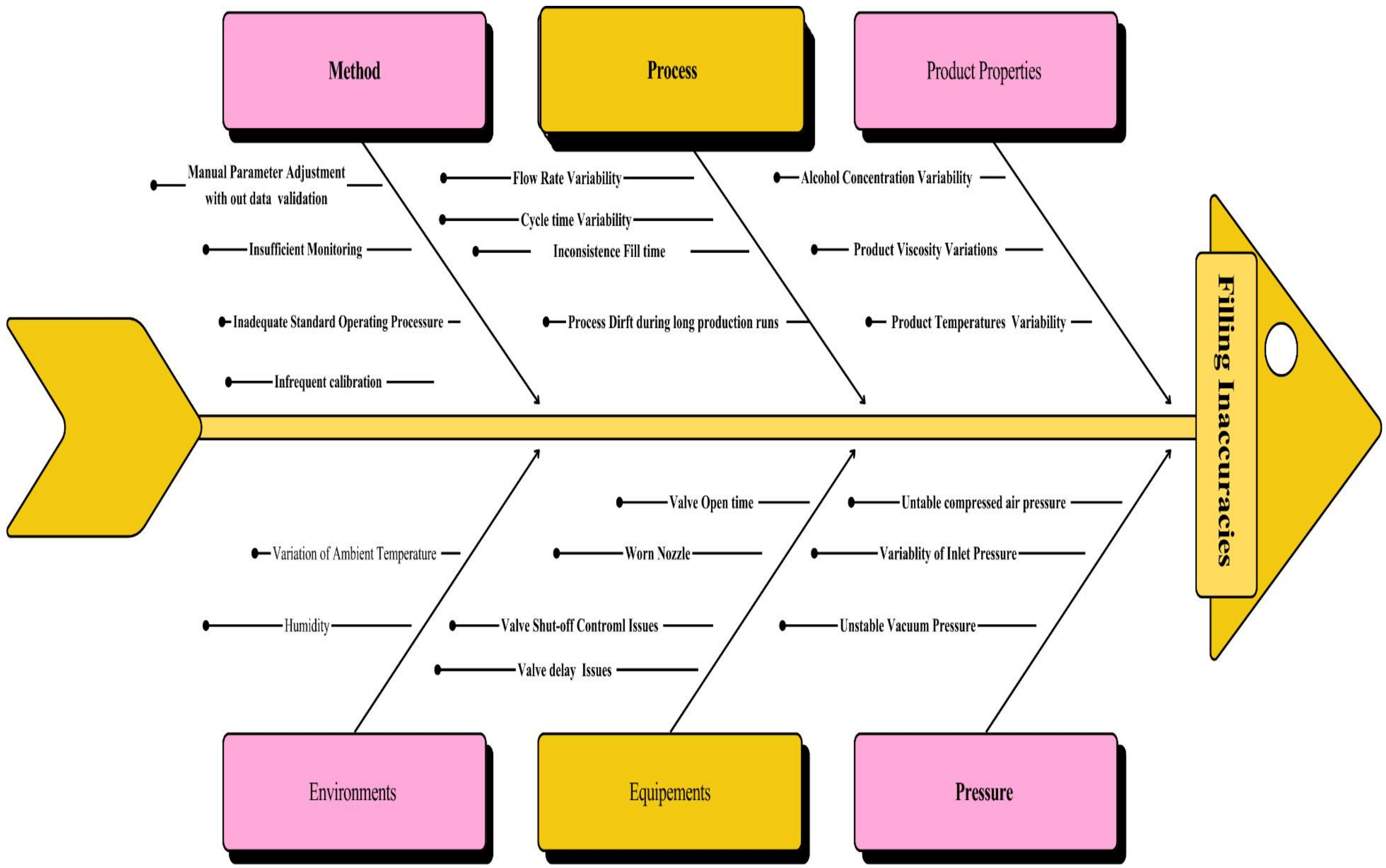
4.3 Identify and Analyze the Key Parameters that Affect Filling Accuracy

4.3.1 Identify the Key Parameters that Affect Filling Accuracy

I. Root Cause Analysis of Filling Inaccuracies

Root Cause Analysis (RCA) was important to this study because such inaccuracies of filling (like overfilling and underfilling) are observable results, not the underlying causes of variation in processes. RCA facilitates the methodical discovery of the underlying factors that affect fill volume and the study was no longer be driven by trial and error parameters manipulation but give a systematic insight into the cause and effect of deviations. This were made sure that the actions that are used to improve will focus on the real causes of filling errors.

The Root Cause Analysis (RCA) tool that can be used in this study is the Fishbone (Ishikawa) diagram since it offers a clear and systematic way of classifying all possible causes of the deviation of the fill volume. The Fishbone diagram can be used to perform a narrow scope of analysis to understand the fill volume as the particular failure mode by categorizing causes under process parameters, equipment condition, material properties, operating conditions, and vacuum-related factors. The structure of visualization is useful in making sure it is complete and at the same time specific enough so that root cause identification is effective



Source: Own, based on data from NALF (June-September 2025)

Figure 4.3 Fishbone diagram

a. Cause -Effect Analysis of Overfilling

A cause and effect analysis were performed using the fishbone diagram to assess the potential contributing factors to the inaccuracy of fill in the bottling process. Process parameters, equipment conditions, material attributes, and environmental parameters that impact delivered fill volume were analyzed.

Table 4.2 Cause -Effect Analysis of Overfilling

Category	Specific Cause	Physical Mechanism (Why it causes overfilling)	Effect on Filling Outcome
Method	Manual parameter adjustment without data validation	Valve open time increased based on assumptions rather than actual flow behavior	Excess liquid delivered per fill cycle
	Insufficient monitoring	Gradual increase in fill volume remains undetected during production	Persistent overfilling across batches
	Inadequate SOP	Lack of standardized timing and pressure adjustment rules	Systematic overfill during product changeover
	Infrequent calibration	Drift in timing and pressure references	Progressive overfilling over time
Process	Flow rate variability	The higher flow rate increases the volume of liquid delivered per unit time.	Increased dispensed volume
	Cycle time variability	Extended effective filling window	Overfill during unstable operation

	Inconsistent fill time	Valve remains open longer than required	Excess liquid transfer
	Process drift in long runs	Thermal and mechanical changes increase flow rate	Gradual shift toward overfilling
Product Properties	Alcohol concentration variability	Higher alcohol content reduces viscosity	Faster flow through filling valve
	Product viscosity variation	Lower viscosity reduces flow resistance	Overfill at constant valve timing
	Product temperature variability	Elevated temperature decreases viscosity	Increased fill volume
Environment	High ambient temperature	Raises product temperature and actuator response speed	Slight but systematic overfilling
	Humidity variation	Affects pneumatic and sensor performance	Indirect increase in overfill risk
Equipment	Valve shut-off control issues	Valve does not seal immediately after closure command	Residual liquid flow
	Valve delay issues	Mechanical inertia prolongs effective fill time	Overfill proportional to delay
	Worn nozzle	Reduced internal resistance increases flow rate	Higher delivered volume
	Incorrect Valve Open time	Extended opening times allow for liquid to continue flowing after reaching the target fill level.	leading to the risk of overfilling.

Pressure	Unstable compressed air pressure	Alters valve actuation and opening duration	Inconsistent but frequent overfill
	Inlet pressure variability	Increases pressure differential across valve	Accelerated liquid flow
	Unstable vacuum pressure	Stronger vacuum increases suction force	Excessive filling beyond target

Source: Own, based on data from NALF (June-September 2025)

The table above illustrates that overfilling is mainly due to increase in effective delivery of liquids during the filling process. In all of the categories, the most common mechanisms that result in overfill are high flow rate or prolonged effective valve open time.

Causes associated with the method include adjusting the manual parameters and inadequate monitoring which lead to overfilling by imparting erroneous valve timing within an open-loop control program. Such methodological concerns are not the direct causes of overfill but facilitate it as they allow detection of increased flow conditions in time and correct them.

Causes that are process-related emphasize the dynamism of the filling operation. The process drift is caused by variability in the flow rate and cycle time especially when long production runs occur. As the flow rate rises with thermal or mechanical action the fixed valve open time causes the over-delivery of volume. This is further increased in product property-related causes. Lower viscosity in liquid due to the presence of an increased alcohol content or increased product temperature lowers the resistance of flow and accelerates the velocity of the flow through the filling nozzle. The control system fails to provide compensation against the variation in viscosity hence overfilling is experienced in other settings when everything remains constant.

Causes that are related to equipment, particularly, the valve shut-off delay, valve open timing and nozzle wear are direct physical contributors towards overfilling. Delayed valve closure will leave behind some liquid flow, which puts in an additional volume relative to flow rate and delay time. The increased inlet or vacuum pressure causes that contribute to these effects are pressure related and escalate the pressure difference across the valve.

Table 4.3 Cause-Effect Analysis of Underfill

Category	Specific Cause	Physical Mechanism (Why it causes underfilling)	Effect on Filling Outcome
Method	Manual parameter adjustment without validation	Valve open time set shorter than required for actual flow conditions	Insufficient liquid delivered per fill cycle
	Insufficient monitoring	Progressive reduction in fill volume not detected promptly	Repeated underfilled bottles
	Inadequate SOP	Lack of defined rules for adjusting timing and pressure	Systematic underfill during product or size changeover
	Infrequent calibration	Timing and pressure reference drift toward lower effective flow	Gradual increase in underfilled units
Process	Flow rate variability	Fluctuations of flow rates can lead to insufficient delivery of liquid during the filling cycle.	Lead to underfill
	Cycle time variability	Shortened effective filling duration	Incomplete bottle filling
	Inconsistent fill time	Valve closes earlier than required	Liquid volume below target
	Process drift in long runs	Increased resistance or cooling reduces flow rate	Progressive shift toward underfilling
Product Properties	Alcohol concentration variability	Lower alcohol content increases viscosity	Slower liquid flow
	Product viscosity variation	Higher viscosity increases flow resistance	Underfill at constant valve timing
	Product temperature variability	Lower temperature increases viscosity	Reduced fill volume
Environment	Low ambient temperature	Reduces product temperature and slows actuator response	Systematic underfilling
	Humidity variation	Affects pneumatic response	Increased risk of

		and sensor accuracy	underfill
Equipment	Valve shut-off instability	Premature or incomplete valve opening	Restricted liquid flow
	Valve response delay	Delayed valve opening shortens effective fill time	Insufficient liquid transfer
	Partially clogged or worn nozzle	Increased internal resistance reduces flow rate	Lower delivered volume
	Incorrect Valve Open time	A shorter opening time for the valve limits liquid entry before the valve closes.	Underfill, Increase process variability
Pressure	Low compressed air pressure	Incomplete valve actuation	Reduced flow into bottle
	Inlet pressure drop	Lower pressure differential across valve	Slower filling
	Insufficient vacuum pressure	Reduced suction force during filling	Incomplete bottle filling

Source: Own, based on data from NALF (June-September 2025)

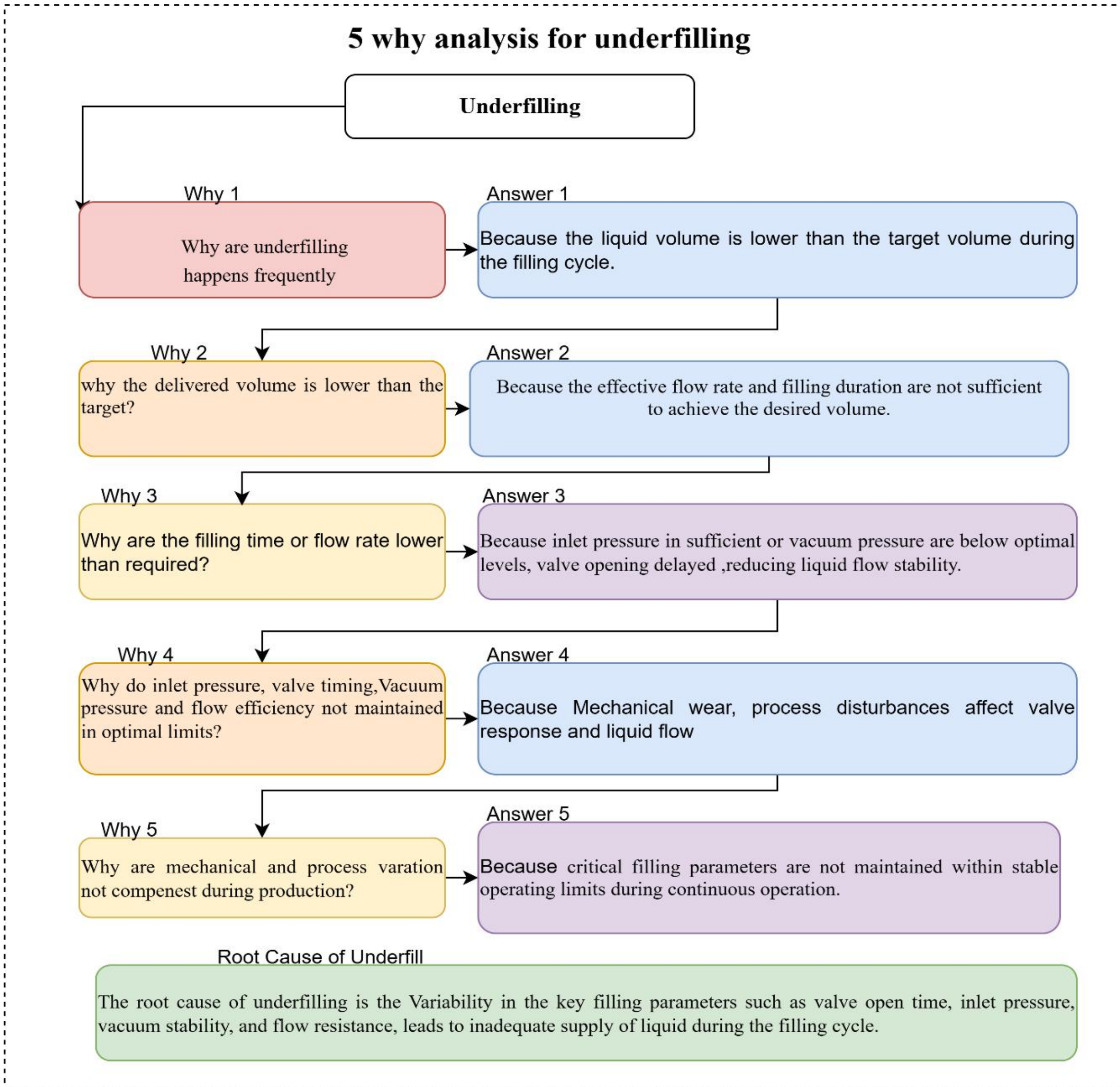
The results presented in the above table indicate that underfilling is caused by the converse physical processes which are slower flow rate or decreased efficient filling time. Just like overfilling, these mechanisms take place at constant valve timing conditions, which are not changed according to real-time conditions. The causes associated with the method include underfilling due to setting of the valve open time shorter than that needed to operate at the actual conditions. Lack of monitoring and calibration enables the progressive attenuations in the volume supplied to continue unnoticed.

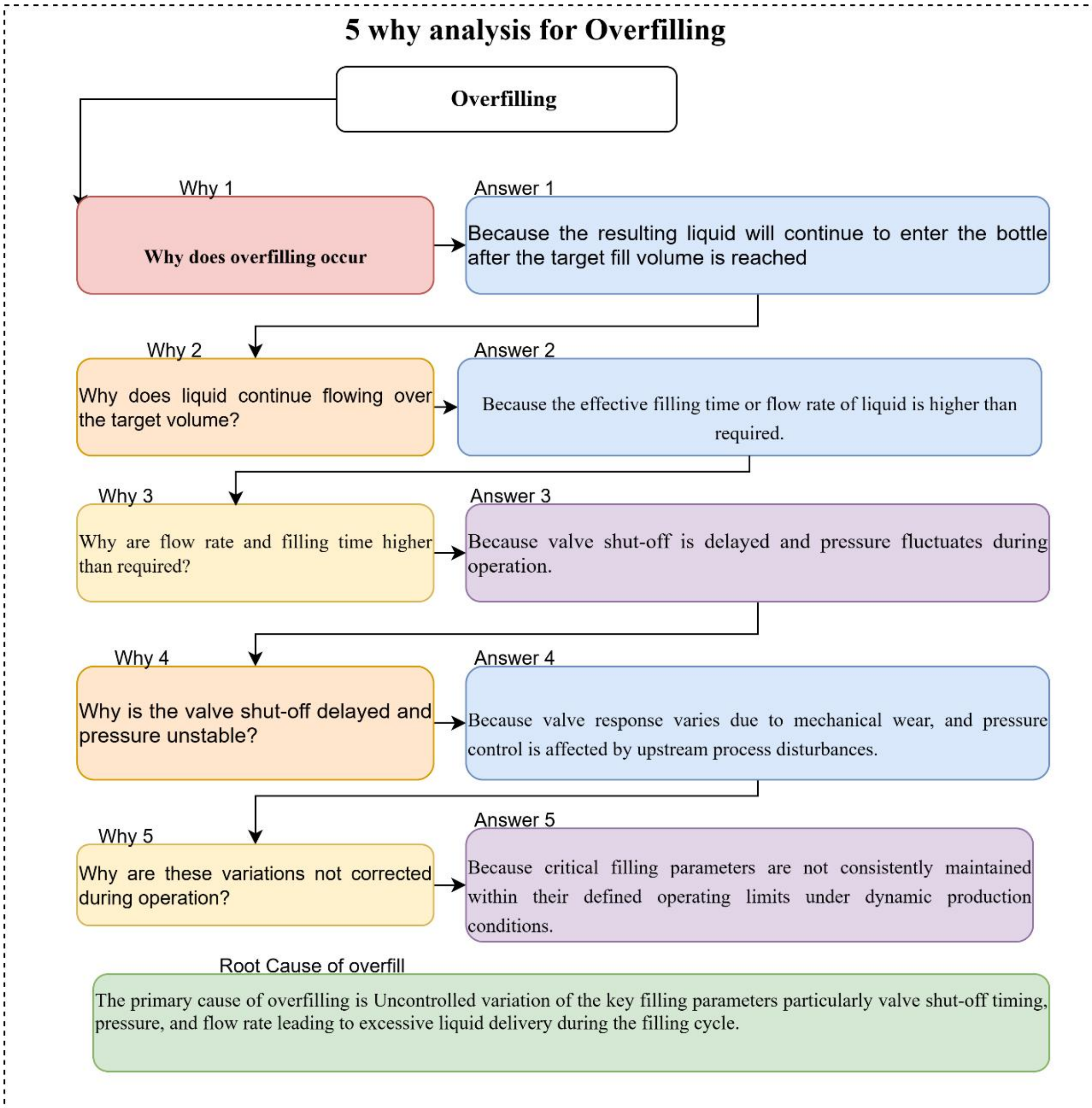
The causes of the process indicate that the variability of the flow rate and the decrease in the cycle time may decrease the duration during which the effective filling occurs. With prolonged production cycles, resistance to flow or cooling may be experienced and slow down the flow rate leaving it to be incomplete. The causes caused by product property are of critical importance in underfilling as more viscosity is caused by low temperature or by decreased alcohol concentration. The increased viscosity raises the flow resistance and decreases the velocity of the liquid and this means that there is inadequate volume transfer within the fixed fill time.

Causes related to equipment like latent opening of valves, partial blockage of the nozzle or incomplete valve actuation are the direct causes that restrict the flow of liquids. The driving force of entry of the liquid in the bottle is also weakened by causes related to pressure such as low vacuum or inlet pressure, which strengthens the underfill conditions.

II. 5 why analysis for underfilling and overfilling

The 5 Whys Analysis is critical in the process of developing a machine learning model that can resolve problems of overfilling and underfill. The main objective is to find out and know the underlying causes of the problems to improve the level of prediction and to prevent mistakes related to the filling process.


5 why analysis for underfilling
Underfilling
Why 1
Why are underfilling happens frequently
Answer 1
Because the liquid volume is lower than the target volume during the filling cycle.
Why 2
why the delivered volume is lower than the target?
Answer 2
Because the effective flow rate and filling duration are not sufficient to achieve the desired volume.
Why 3
Why are the filling time or flow rate lower than required?
Answer 3
Because inlet pressure in sufficient or vacuum pressure are below optimal levels, valve opening delayed ,reducing liquid flow stability.
Why 4
Why do inlet pressure, valve timing,Vacuum pressure and flow efficiency not maintained in optimal limits?
Answer 4
Because Mechanical wear, process disturbances affect valve response and liquid flow
Why 5
Why are mechanical and process varation not compenest during production?
Answer 5
Because critical filling parameters are not maintained within stable operating limits during continuous operation.
Root Cause of Underfill
The root cause of underfilling is the Variability in the key filling parameters such as valve open time, inlet pressure, vacuum stability, and flow resistance, leads to inadequate supply of liquid during the filling cycle.



According to the analysis of the 5 Whys of overfilling and underfilling, it could be observed that the main problem is uncontrolled variation of the key filling parameters particularly valve shut-off delay, valve timing, Inlet pressure, Vacuum pressure and flow rate. This type of system is important to adjust to changes in pressure, fluid characteristics, valve timing, and equipment conditions during the filling process. Without it, the remaining flow was hysterical and hard to control, thus leading to either over overfilling or underfilling.

III. Key Parameters Affecting Filling Accuracy

According to the Root Cause Analysis (RCA) and the further 5-Why analysis, the main parameters by which the filling accuracy depended were found systematically by determining the root causes of observable filling deviations. The fill volume deviation as the failure mode was specifically analyzed, and only those parameters that have a direct physical or operational effect on the filling behavior were taken into consideration. Such a formal procedure allowed isolating the main causes of filling inaccuracy and minor or accidental causes.

The Root Cause Analysis (RCA) indicated that filling accuracy is mainly controlled by parameters of the valve timing, pressure, flow conditions, and mechanical factors of the filling system. Repeated 5-Why questioning revealed that variation in fill volume sources to differences in valve open time, valve response delays, inlet pressure, flow rate, and stability of the vacuum. Mechanical factors like nozzle wear were found to cause long-term drift and head-to-head variability as opposed to the effects of short-term fluctuations. Those parameters have a direct influence on the quantity of liquid supplied at each filling cycle by changing the effective flow duration and flow rate.

Table 4.4 Key Parameters Affecting Filling Accuracy

Parameter Name	Unit	Description / Effect on Filling Accuracy
Product Temperature	°C	Liquid temperature affecting expansion and flow characteristics.
Alcohol by Volume (ABV)	-	Alcohol concentration influencing density and viscosity indirectly.
Viscosity	mPa·s	Resistance to flow; limited impact for low-viscosity products like vodka.
Inlet Pressure	Bar	Pressure driving liquid into the bottle; higher pressure increases flow rate and volume.
Vacuum Pressure	Mbar	Vacuum level during filling; stabilizes flow and affects fill termination.
Flow Rate	L/s	Rate of liquid flow through the nozzle; directly proportional to fill volume.
Valve Open Time	S	Duration for which the filling valve remains open; directly controls delivered fill volume.
Shut-off Delay	S	Delay between shut-off command and actual flow termination; causes post-fill overfilling.
Valve Delay	S	Time lag between valve command and actual opening; affects effective filling duration.
Cycle Time	S	Total duration of one filling cycle; influences synchronization and effective fill time.
Fill Time	S	Total time of the filling operation; longer times generally increase fill volume.
Nozzle Wear	%	Extent of nozzle degradation affecting flow area, sealing, and long-term volume drift.
Density	g/ml	Mass per unit volume; minor influence on volumetric filling under stable conditions.

Source: Own, based on data from NALF (June-September 2025)

The parameters identified, along with the units and functions of each of them summarized in above Table 4.21. These parameters are the most determining parameters in causing volumetric variation of the vacuum filling process especially when no volume measurement sensor is directly used. Time based parameters, like valve open time, fill time, valve delay and shut-off delay, were observed to directly influence the effective filling time, and the properties relating to pressure and flow like inlet pressure, flow rate and vacuum pressure controlled the liquid deliveries phenomenon. Also, mechanical degradation drivers, with nozzle wear percentage being the most prominent one, were found to cause long-term drift in filling accuracy.

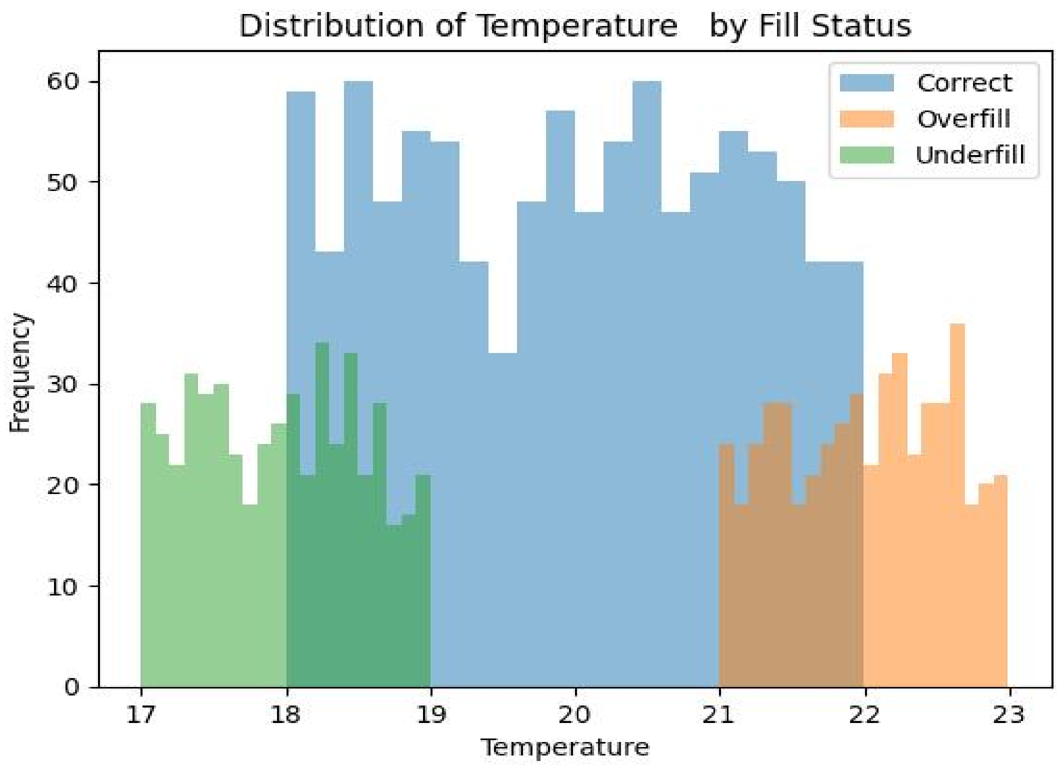
4.3.2 Analyze the Key Parameters that Affect Filling Accuracy

This section assesses the engineering informed synthetic dataset that was developed to the National Alcohol and Liquor Factory (NALF). The objective is to confirm internal consistency, summarize key distributions and elicit preliminary relationships that was used to build and interpret models.

I. Parameters Distributions and Visualizations

Distribution of Product Temperatures

The figure called Distribution of Temperature by Fill Status shows the filling accuracy of bottles that were sorted into three sets each of which had different frequency distributions over the 17℃ to 23℃ temperature ranges. Correct fill class in light blue bars depicts the highest frequencies at particular intervals with the highest frequencies being around 18℃ and 22℃ with 40 to 60 instances. The Overfill category, denoted by orange bars, on the other hand, has lower frequencies on the whole, with significant peaks at 21℃ and 23℃, indicative of problematic temperatures, and frequencies between 20 and 30 incidences. The smallest group, the Underfill class, represented by light green bars, is mainly at the low-temperature end of the spectrum, and most of the time around 17℃ and 19℃, with frequencies of 30 to 40 instances.

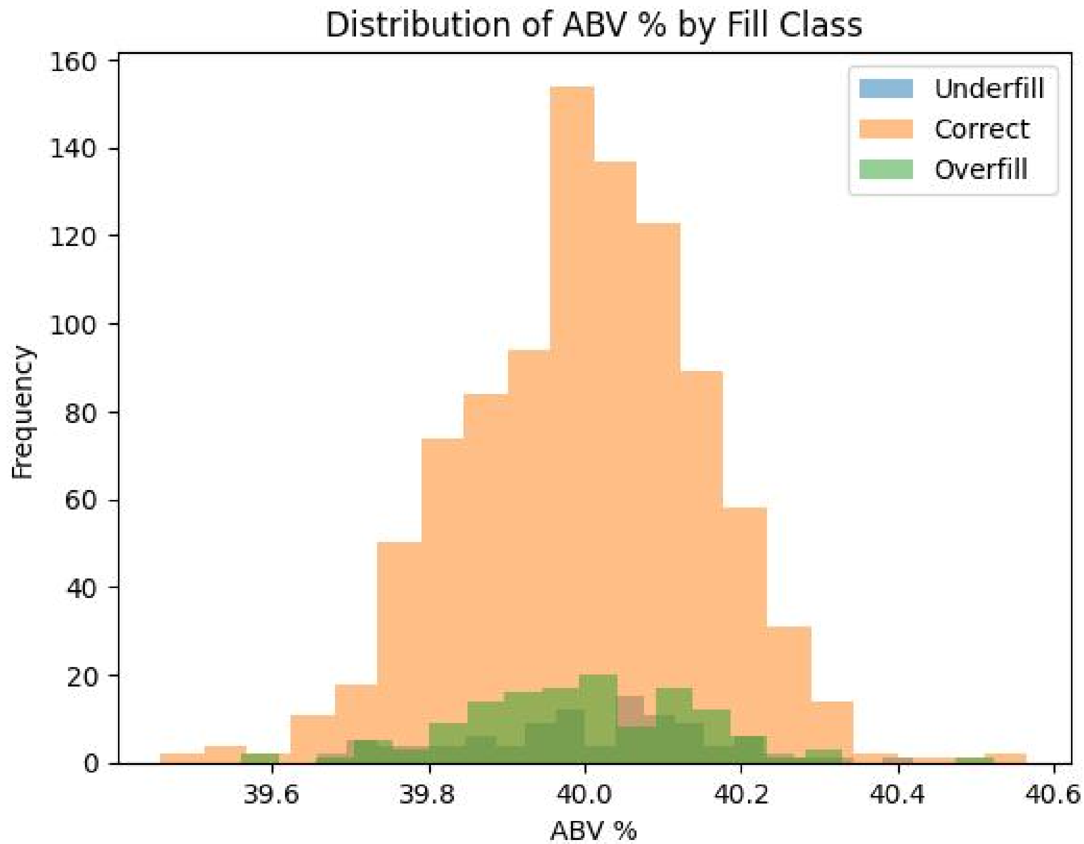


Source**:** Own, based on data from NALF (June-September 2025)

Figure 4.4 Distribution of Temperature.

Distribution Alcohol Concentration:

Figure4.3 Distribution of ABV percent by Fill Class demonstrates how the alcohol by volume (ABV) percentages were distributed in three classes Correct, Overfill and Underfill. The histogram clearly shows that the frequency is the most concentrated in the 40% ABV mark with the highest frequency in the Correct class that reflects the effective fill processes in the 40% ABV mark. On the other hand, the Overfill category indicates a very large number of cases where bottles fill beyond the appropriate ABV and this may indicate a problem either with the filling equipment or with the working procedures. Moreover, it is possible to note that the Underfill category demonstrates significant numbers of bottles that do not reach the required ABV, which can be explained by the following issues: calibration issues or poor settings of the equipment.

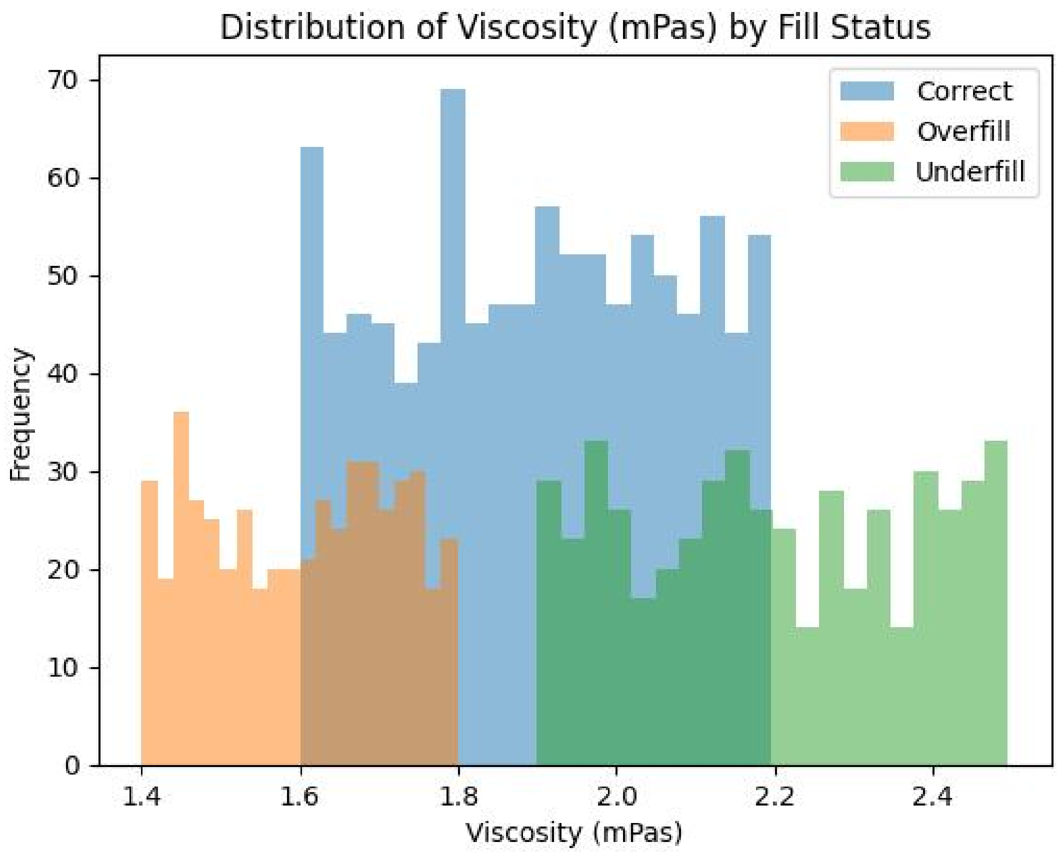


Source**:** Own, based on data from NALF (June-September 2025)

Figure 4.5- Distribution of Alcohol Concentration by fill class

Distribution of Viscosity:

The figure 4.4 show the Distribution of Viscosity (mPas) by Fill Status and it was used to show how the viscosity can influence the filling accuracy between 3 classes; Correct, Overfill, and Underfill. The Correct class with light blue bars indicates that there are greater success rates of successful fills at the viscosity levels between 1.6 and 2.2 mPas which indicates that these two viscosity levels are ideal in filling. Conversely, the Underfill class, represented by orange bars, shows high frequencies in the viscosities of 1.9-2.5 mPas, implying that high viscosity could be a factor in increasing the possibility of Underfilling, presumably because of difficulty in controlling the fill process. The Overfill class which is in light green is the most dominant at the lower levels of viscosity which is more concentrated around the 1.4 to 1.8 mPas range of the viscosity, which may imply that liquids flowing below it may be too fast causing uneven fill levels. All these observations demonstrate the importance of viscosity in the filling process, and the necessity to carefully monitor and make modifications in order to maximize the filling process and assure the quality of the products with the help of various types of liquids.

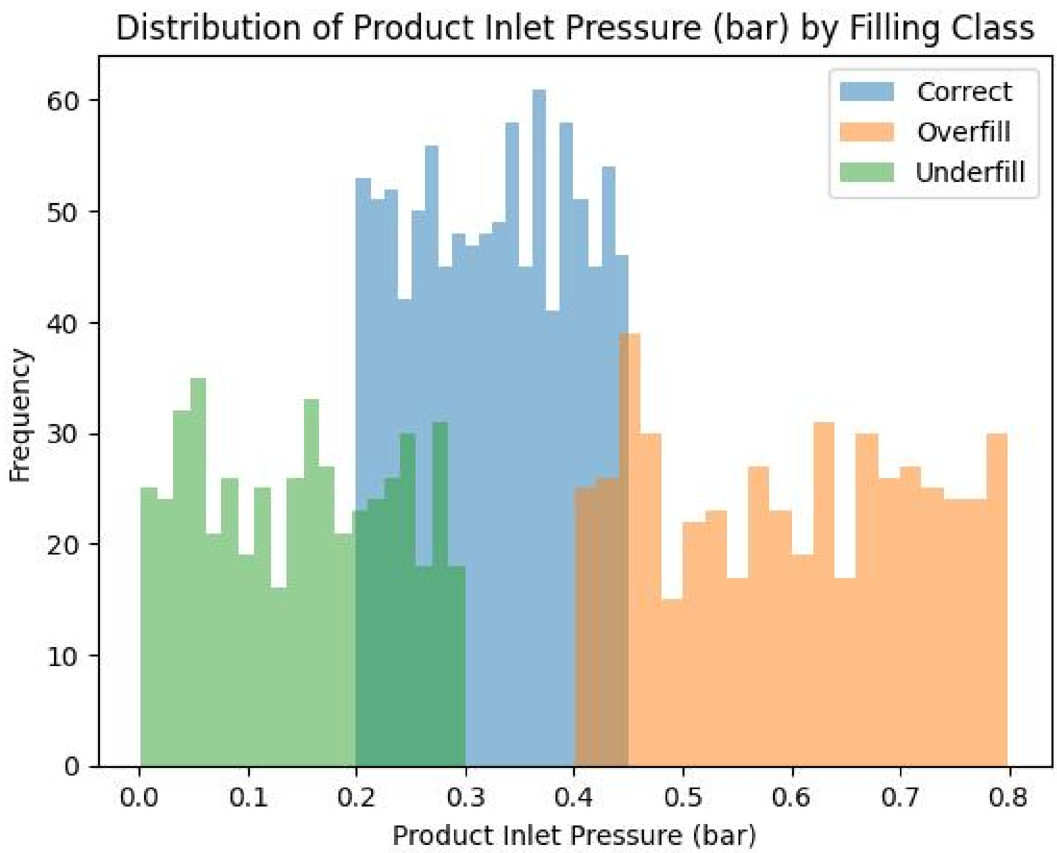


Source**:** Own, based on data from NALF (June-September 2025)

Figure 4.6 Distribution of Viscosity by fill class

Distribution Product inlet Pressure:

The figure 4.5 show Distribution of Product Inlet Pressure (bar) by Filling Class gives significant information on the effect of product inlet pressure on filling accuracy under three groups: Correct, Overfill and Underfill. The Correct class represented by light blue bars shows that the frequency of successful fills is higher with inlet pressure in the range of 0.2 to 0.45 bar showing that this is the range that is best to be used in attaining accurate fills. On the other hand, the Overfill category which is depicted in orange peaking at pressures of about 0.4-0.8 bar, it is evident that high pressures can cause excessive filling which may probably be because the liquid is forced into containers too fast. In the meantime, the Underfill category typified by light green is more common at lower pressures, especially between 0 and 0.3 bar which indicates that insufficient pressure leads to underfilling. These observations underscore the importance of the optimum product inlet pressures that would remain consistent and precise during product filling, and the need to monitor and adjust in the manufacturing processes to help improve the product quality and efficiency in the production process.

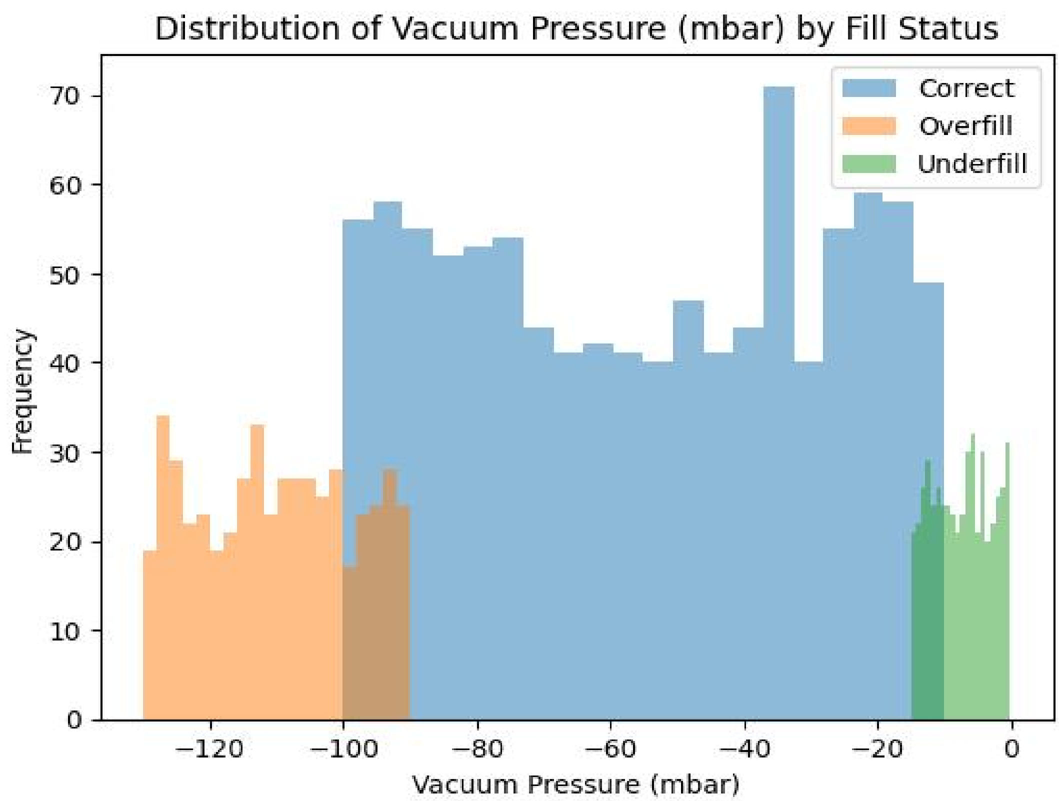


Source: Own, based on data from NALF (June-September 2025)

Figure 4.7 Distribution Product inlet Pressure by fill class

Distribution of Vacuum Pressure:

Figure 4.8 Distribution of Vacuum Pressure (mbar) by Fill Status gives some useful information on the relationship between changes in vacuum pressure and filling accuracy which is classified as either Correct, Overfill, and Underfill. The blue bars indicating Correct fill status exhibit a high percentage of correct fills in the range of -100 to -10 mbar which indicates a good vacuum pressure level of filling to get the desired volume without over filling or under filling. On the other hand, the orange bars indicating Overfill status are the highest between -130 and -90 mbar, indicating that higher vacuum pressure can cause overfilling because of an imbalance in the atmospheric pressure that causes more liquid to be pushed into the container than desired. The Underfill status light green bars are mostly seen below -10 mbar and indicate how insufficient vacuum pressure might lead to insufficient filling whereby the system is unable to pull the appropriate volume of liquid. These observations reiterate the need to pay close attention to vacuum pressure during the filling process to increase accuracy, reduce waste, and increase efficiency in production. Also, this knowledge can be used in quality assurance, whereby companies should introduce the practices of regular calibration, and real-time monitoring technologies. In sum, the histogram has highlighted that the key to attaining the desired fill results is to ensure that the vacuum pressure is ideal and therefore aid in enhancing the quality of products and efficiency in the manufacturing facilities.

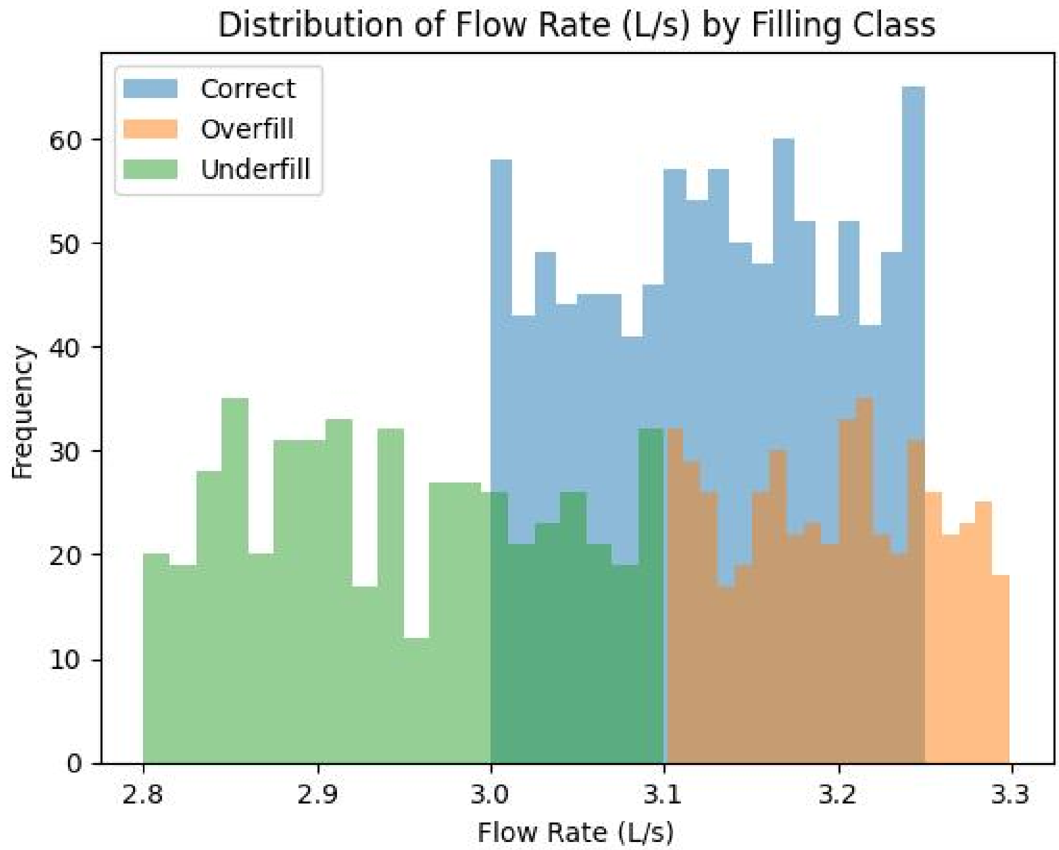


Source**:** Own, based on data from NALF (June-September 2025)

Figure 4.8 Distribution of vacuum pressure by fill class

Distribution of Flow rate:

The histogram, Distribution of Flow Rate (L/s) by Filling Class, gives useful information about the correlation between flow rates and accuracy of filling and they are classified into three statuses namely, Correct, Overfill, and Underfill. The bars of light blue, which depict Correct filling, indicate that there is a high frequency of flow rates being concentrated around 3.0 L/s, which means that this interval is good in attaining the desired volume without any deviations. On the other hand, the orange bars of Overfill indicate that the rate of flow under 2.9 L/s is connected to a significant rise in the number of overfills, and the possibility of excessively slow a flow rate causing inefficiency in a filling process. Conversely, the green bars that denote Underfill mean that underfilling happens in the flow rates that surpass 3.2 L/s since the rapid exit fails to provide the necessary time in which the target amount of the volume can be filled in. On the whole, the histogram highlights the necessity to keep the flow rates within a certain optimum to guarantee the accurate filling, which makes it possible to focus on the necessity of the accurate monitoring and control in the manufacturing operations to improve the quality of the products and reduce the amount of waste.

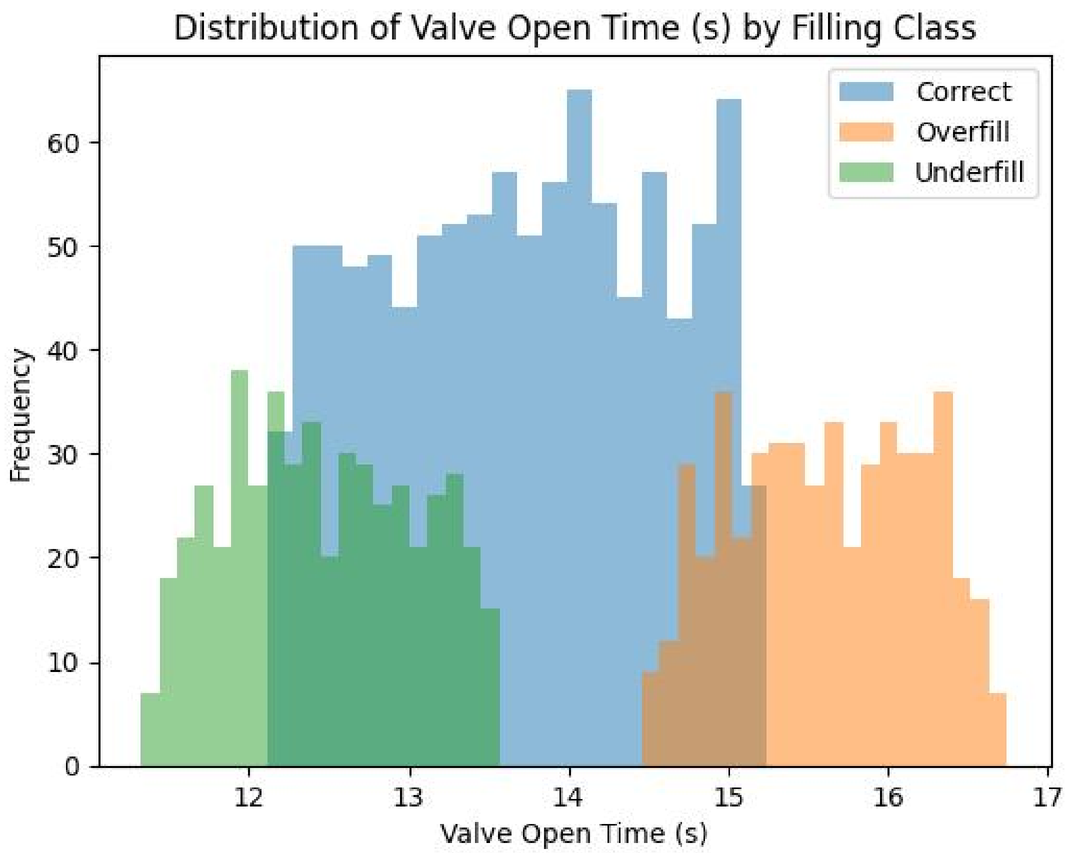


Source: Own, based on data from NALF (June-September 2025)

Figure 4.9 Distribution of Flow rate by fill class

Distribution of Valve Open time:

Figure 4.10"Distribution of Valve Open time (s) by Filling Class" depicts important trends on the relationship between valve open times and fill accuracy in terms of Correct, Overfill and Underfill. The light blue bars showing Correct fill status demonstrate that there is a high rate of valve open times around 12.2-15.3 seconds which appear to be an optimum time which is not too long or too short that results in proper filling. Conversely, the orange bars of the Overfill status align towards the 14.5 seconds mark, which means that the longer the valve open time the higher the chances of overfilling the container, since the longer the valve remains open, the greater the amount of liquid that can pour into the container than intended. In the meantime, the light green bars marking Underfill status mostly have a low valve open time less than 13.5 seconds, and this shows that inadequate open times may lead to inadequate filling, as there is no time to dispense the appropriate volume of liquid. These observations highlight the significance of putting close attention and optimization of the valve open times in order to achieve better accuracy in filling, waste minimization, and performance in the manufacturing environment. Additionally, the analysis can be used in quality control operations so that manufactures used can set certain parameters of how the valves should work in order to achieve improved quality of the products and production results. On the whole, the histogram shows the importance of having a strong connection between the valve open time and filling accuracy, which confirms the need to have tight control in the manufacturing processes.



Source**:** Own, based on data from NALF (June-September 2025)

Figure 4.10 Distribution Valve open time by fill class

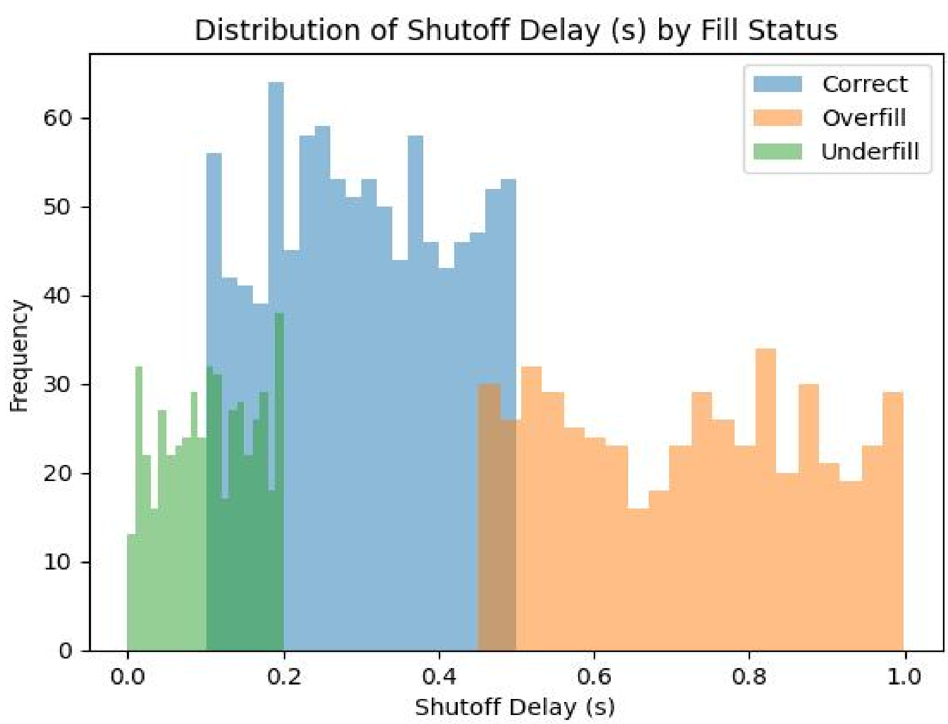
Distribution of Shutoff Delay:

The histogram used shows the distribution of the shutoff delay time (in seconds) of three classes of the fill, which are in seconds, Underfill, Correct fill and Overfill. The x-axis is the shutoff delay times whereas the y-axis is frequency.

The spike in frequency in the underfill category is observed with shorter delay times, and it occurs at higher frequencies mostly at 0.02 seconds. This is an indication that delayed filling can be caused by the little delay. The proper fill class is more centered around a 0.05 seconds value which is the most balanced and it implies the optimal delay that fits well with satisfactory fill levels.

However, the pattern of overfill category is quite different, there is a decreasing tendency in frequencies of the overfill category as the shutoff delay time increases. This implies that the longer the delay, the higher the chances that the process will become overfilled, producing a smoother distribution.

In general, the histogram shows that it is necessary to adjust the shutoff delay times to maximize the results of the filling to reduce the risk of underfilling as well as overfilling and find the best range of 0.05 seconds.



Source: Own, based on data from NALF (June-September 2025)

Figure 4.11 Distribution of Shutoff Delay by fill class

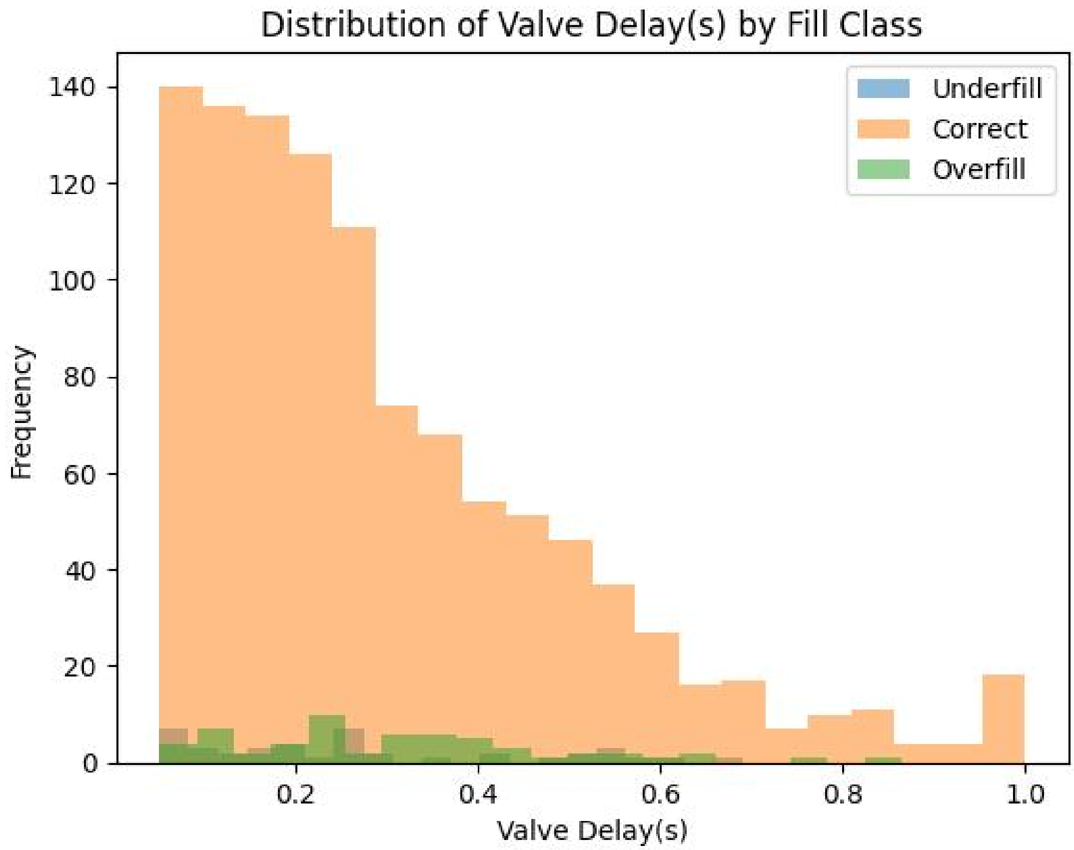
Distribution of Valve Delay:

The histogram shows the arrangement of the histogram of the time taken by valves (the distribution in three classes) to fill: Underfill, Correct fill and Overfill. The x-axis represents the valve delay times whereas the y-axis represents the frequency of the incidents.

Underfill category shows that the data has a sharp peak of about 0.02 seconds indicating that a shorter delay by the valve can cause incomplete filling. The correct fill class (individual of orange color) has a high frequency distribution at 0.04 seconds suggesting that the time interval is best suited to obtain the desired fill levels.

On the other hand, the frequency of the overfill class is lower on the average with a steady decrease in the number of occurrences with an increase in the valve delay time. The implication of this trend is that increasing the valve delays would lead to a lower number of cases of overfilling since fewer cases would be reported under this category.

In general, the histogram shows that the time of valve delay and the fill results are one of the most important factors that can be improved to increase the filling efficiency by ensuring that the probability of filling the valves to the proper level is high and the chances of underfill and overfill are reduced to the minimum possible.



Source: Own, based on data from NALF (June-September 2025)

Figure 4.12 Distribution of Valve Delay by fill class

Distribution of Cycle Time:

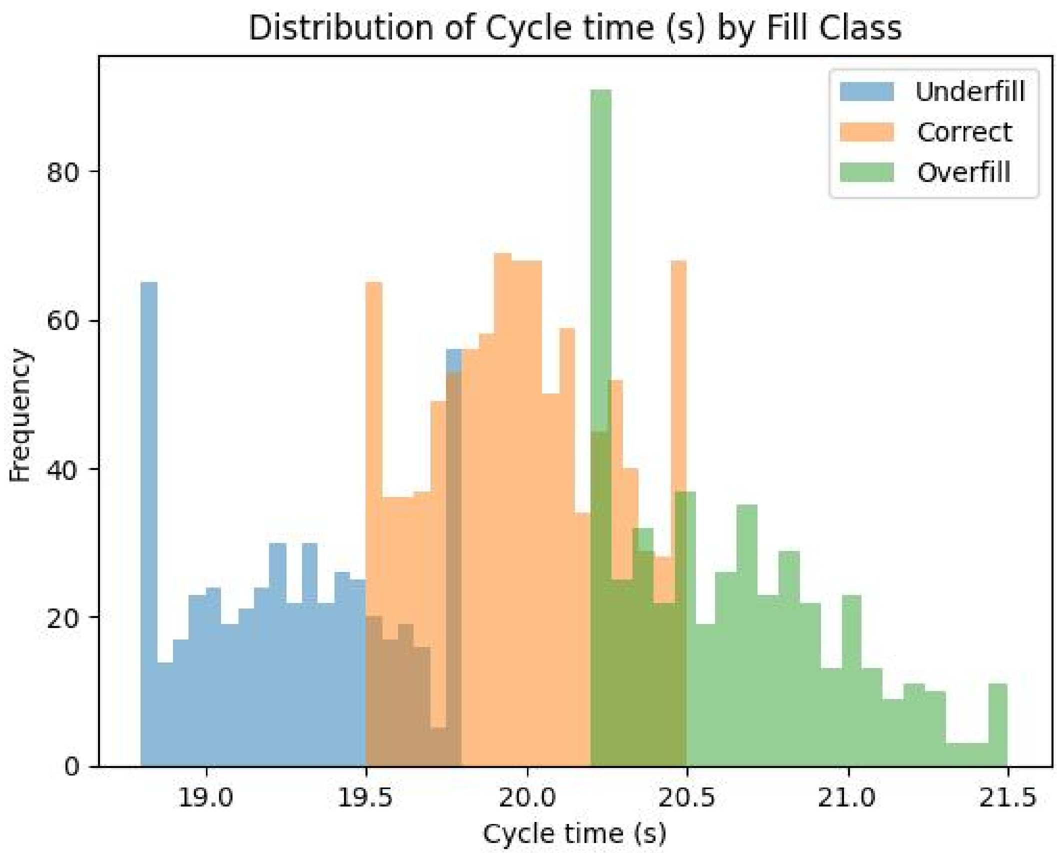
The histogram demonstrates three fill classes: Underfill, Correct fill and Overfill with their respective cycle time distribution in seconds. The x-axis shows the cycle times whereas the y-axis shows the frequency of occurrence in each category.

The relative stability of the frequency across the range in the underfill class (blue) is relatively constant with the highest frequency occurring at approximately 19.5 seconds. This indicates that there are numerous cases of underfilling at this length of time, which means that less cycle times will result in incomplete fills. Conversely, the most accurate category of fill (orange) has a strong peak just around the 20.0-second mark making it the most common cycle time to reach the optimal filling. This implies that cycle times occurring within this range are the best when it comes to meeting target fill levels.

The overfilled category (green) however records significantly lower records than the other two categories with fairly spread frequency peaks and minor concentration at 20.5

seconds. It means that longer cycle times might lead to overfilling, and this situation is not that common in general.

In general, histogram underlines the importance of cycle times optimization. The minimum of about 20.0 seconds seems essential to achieve the highest number of correct fills and the lowest number of cases of underfilling and overfilling. The results indicate that the packing performance of the system under consideration can be effectively increased by the careful algorithmic manipulations with cycle times.



Source**:** Own, based on data from NALF (June-September 2025)

Figure 4.13 Distribution of Cycle Time by fill class

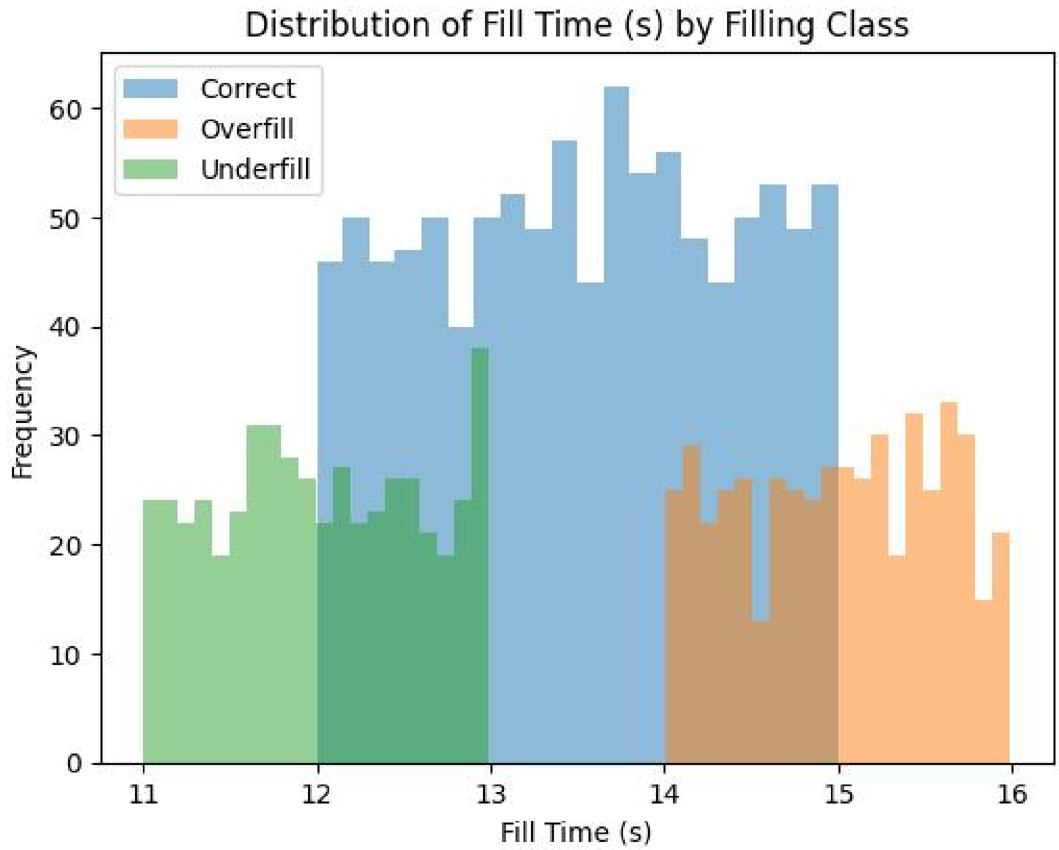
Distribution of Fill time:

The histogram shows distribution of the fill times (in seconds) in three fill classes, namely, the Underfill, Correct fill, and Overfill. The fill times are plotted against the x-axis whereas the frequency of occurrence per category is plotted on the y-axis.

The frequency in the underfill category (blue) has a regular trend but is specifically peaked at 15.5 seconds, which indicates that a significant proportion of instances of underfilling takes place at this period. This means that less time taken to fill is more likely to leave it unfinished. Still, the appropriate fill category (orange) has an overwhelming peak at 16.5 seconds, which is the most frequent fill time to reach optimal filling. This implies that fill time of this magnitude is optimal when it comes to achieving the desired fill levels.

The overfill (green) has another pattern; it has fewer overall occurrences with a peak of approximately 17.5 seconds. This indicates that fill time can cause overfilling, but the cases are rare as compared to the other classes.

In general, the histogram indicates that it is necessary to pay attention to the time when fill processes should be performed. The objective of approximately 16.5 seconds seems important to the achievement of maximization of correct fills as well as reduction of underfilling and over filling. The information confirms the hypothesis that the accuracy in the adjustments made to satisfy the fill times can greatly enhance the performance of the filling in the discussed system.



Source: Own, based on data from NALF (June-September 2025)

Figure 4.14 Distribution of fill time by fill class.

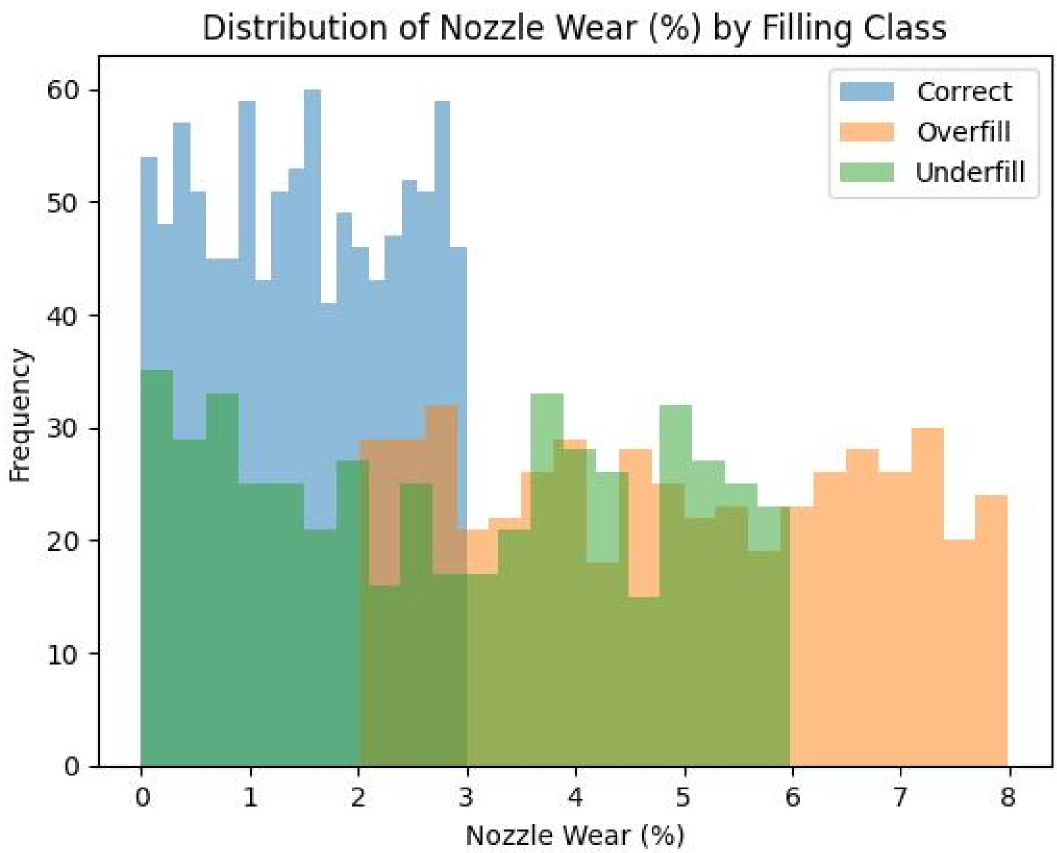
Distribution of Nozzle wear:

The histogram depicts the percentage distribution of the nozzle wear to three fill classes of Underfill, Correct fill and Overfill. The x-axis is a percentage of wear on the nozzle with a range of 0 to 4, but the y-axis indicates the frequency of the number of times it happened in each range.

In the underfill category, the light brown bars display a steady distribution with the top being at 0.5 to 1.0 percent indicating lower percentages of wear at the nozzle are more related to underfilling. This means that less common and less experienced nozzles can result into underfilling. On the other hand, the properly fill class which is marked by orange color also has a relatively homogeneous distribution with a small peak of around 1.5-2.0 which implies that moderate nozzle wear is ideal in ensuring correct fills.

In overfill, the concentrations in the green bars are mostly 1.5 to 3.0 percent, which suggests that when the percentages of wear are greater, it may be a contributing factor to cases of overfilling. This similarity to the correct fill class indicates that though possible wear may still cause proper filling, a larger amount of wear is more probable to cause overfilling.

Generally, the histogram emphasizes the connection between nozzle wear and fill performance so that there was a necessity to ensure that nozzle wear is within a certain range to ensure the optimal filling performance.



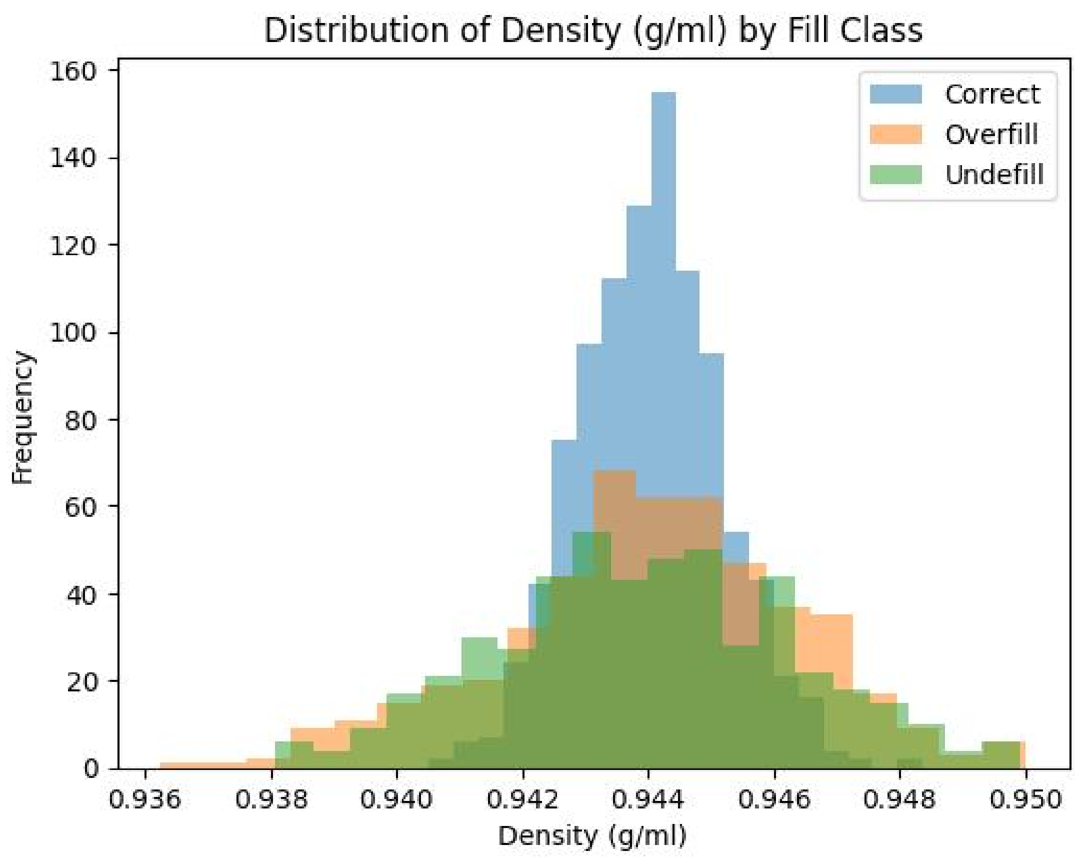
Source**:** Own, based on data from NALF (June-September 2025)

Figure 4.15 Distribution of Nozzle wear by fill class

Distribution of density by fill class:

The histogram presents three fill classes, including Correct, Overfill, and Underfill, each having different implications to quality control in the filling process. The "Correct" category shows examples of the density of the filled material that satisfy the desired requirements and show a sharp concentration around 0.944 g/ml. This indicates that filling is normally under control and this is an indication of good procedures and quality assurance measures.

Conversely, the class of Overfill denotes those cases when the density is beyond acceptable values. Even though the frequency of this class seems to be lower than the frequency of the correct fills, its occurrence will draw attention to the necessity of attention, since the risks of overfilling products and the possible consequences of overfilling should be. Finally, the Underfill category shows the events when the density is lower than the minimum possible, which is a major issue of concern of the integrity of the product and customer satisfaction. Underfilled products may lead to dissatisfaction since customers may be getting less than what was promised thus there should be more adjustments on the calibration of the machines as well as the practice in the operation process.



Source**:** Own, based on data from NALF (June-September 2025)

Figure 4.16 Distribution of density by fill class

Distribution of Fill Volume:

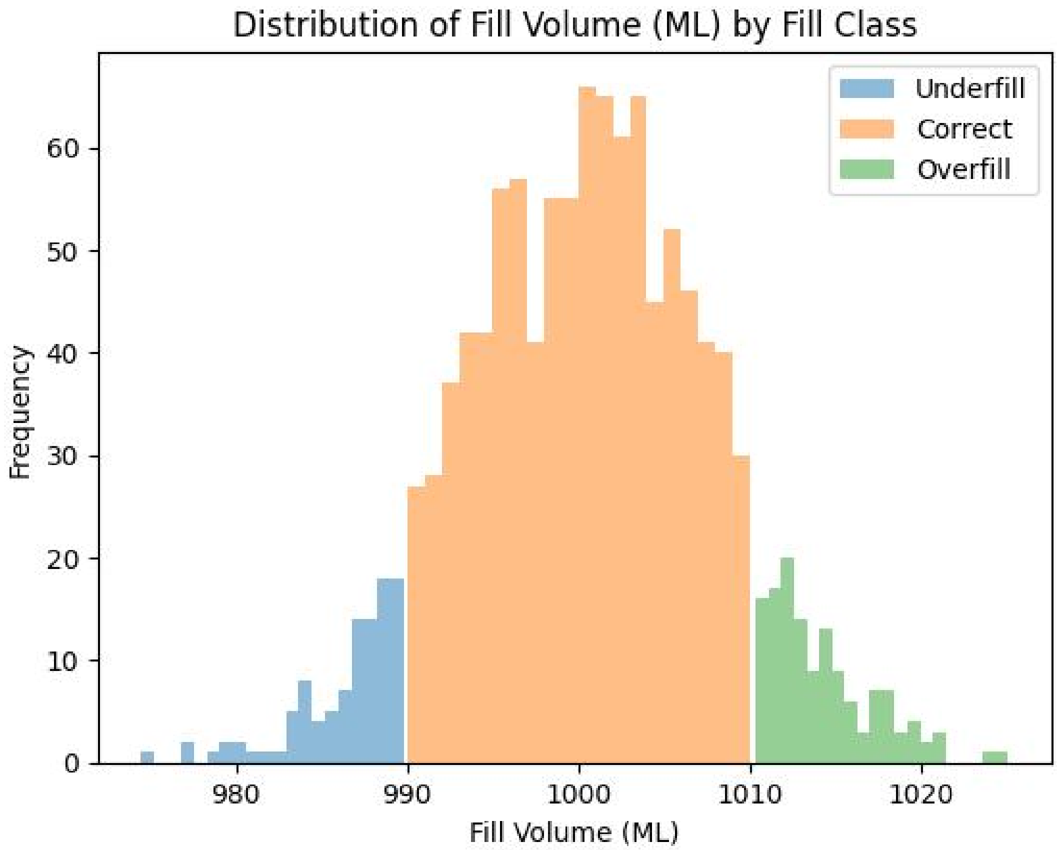
The histogram shows the distribution of the number of milliliters of volume that one should expect to fill by three fill classes, which are Underfill, Correct fill, and Overfill. The x-axis is the expected fill volume, and it is between 970 and 1020 ml, and the y-axis is the frequency of the occurrence of each class of fills.

The blue bars in the underfill category give an observable peak at around 980 ml. This implies that many events are under ideal fill volume and this means that underfilling is a frequent problem where the desired volume is in the range.

The orange bars depict the Correct fill class as the peak is rather large with the value being 1000 ml. This means that the optimum fill volume should be near this value and where the frequency of getting the correct fill is the maximum showing that it is important to follow the anticipated volume of 1000 ml in order to get the desired filling outcome.

The overfill category, represented by the green bars, in contrast, has lower frequency, and has peaks of 1010 ml and 1020 ml. This implies that although there is a certain amount of overfilling, this is lower than the case with underfilling, and the ideal filling.

In general, the histogram shows that there are substantial differences in the performance of the fill over the given volumes. It emphasizes the fact that one ought to target the right fill volume of approximately 1000ml in order to streamline the filling process, as well as discovering the existing rates of underfilling, especially near 980 ml. The tracking of these trends will assist in reducing fill errors and ensure more stable outcomes.



Source: Own, based on data from NALF (June-September 2025)

Figure 4.17 Distribution of Fill Volume.

II. Descriptive Statistics of key parameters

The 2,000 data points were gotten in the National Alcohol and Liquor Factory with the help of the historical PLC logs and quality-inspection records. Among these, 1,000 are the normal conditions, 500 are the underfills cases and 500 are the overfills cases, which are known by the recorded deviation events and confirmed by the expert review. The descriptive statistics give useful background information about the maximum, minimum mean and standard deviation of various parameters utilized in the production process.

Table 4.5 Descriptive Statistics of key parameters

Descriptive Statistics					
Key Parameters	N	Minimum	Maximum	Mean	Std. Deviation
Product Temperature	2000	17.00	22.97	19.9654	1.34502
ABV %	2000	39.80	40.20	40.0006	.05898
Viscosity (mPas)	2000	1.40	2.49	1.9036	.20960
Inlet Pressure (bar)	2000	.00	.80	.3496	.18511
Vacuum Pressure (mbar)	2000	-129.87	-.07	-56.8398	41.51215
Flow rate (L/S)	2000	2.80	3.30	3.1003	.11754
Valve open time (s)	2000	14.81	17.88	16.2718	.75750
Shutoff delay(s)	2000	.01	.10	.0345	.01992
Valve delay (s)	2000	.01	1.00	.2553	.16196
Fill time (s)	2000	11.00	15.99	13.5173	1.28473
Cycle time (s)	2000	18.80	21.50	19.9717	.56824
Density (g/ml)	2000	.94	.95	.9440	.00186
Nozzle Wear%	2000	.00	3.99	1.5153	.96967
Valid N (listwise)	2000				

Source**:** Own, based on data from NALF (June-September 2025)

As stated on table 4. The averages reveal the average level of performance: the average values of the products temperature are 19.97 with a standard deviation of 1.35, which means that temperature is under control. The alcohol content is very consistent as indicated by the standard deviation of the alcohol content is about 0.05898 with a value of 40.0006. Viscosity means 1.90 and a standard deviation of 0.20960, which shows that the thickness varies moderately.

The mean of the Inlet Pressure is 0.35 bar with the standard deviation of 0.185111 implying a tendency of variation in the operational pressure. Conversely, the mean of Vacuum Pressure is -56.84 mbar and the standard deviation is 41.51 showing that there is a high range of the levels of the vacuum. The Flow rate mean is 3.10 L/S and the standard deviation of flow rate is very low and equals 0.11754 which indicates that there are no fluctuating conditions in flow rate. In the case of valve management, the average of the Valve open time is 16.27 seconds and the standard deviation of the predicted value is 0.75750 which indicates that there is some variation in the time of operation. The mean of the "Shutoff delay" is 0.0345 seconds and its standard deviation is 0.01992 which is under tight control and the mean of the Valve delay is 0.2553 seconds with standard deviation of 0.16196.

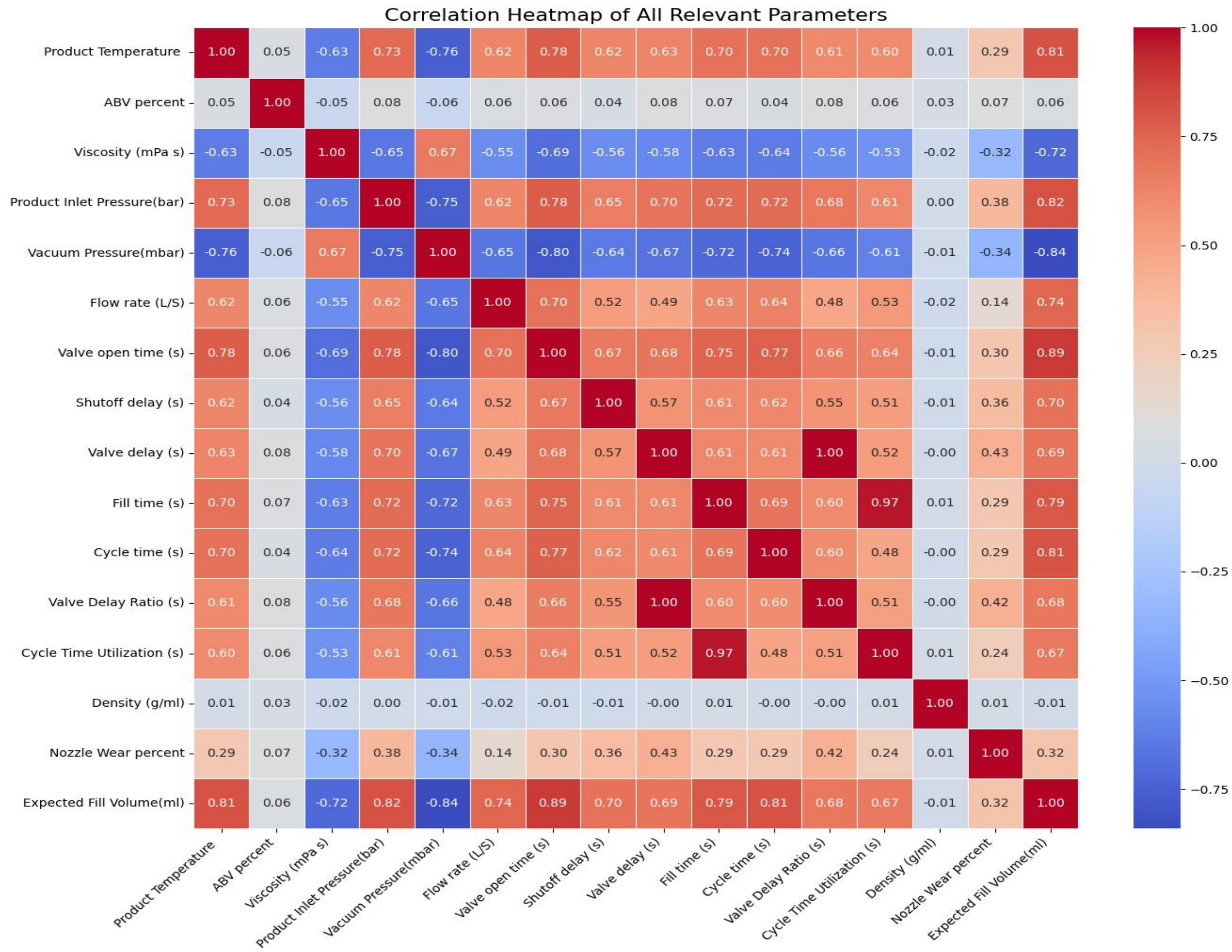
The standard deviation of Fill time equals 1.28 and the mean value of Fill time is 13.52 which implies that fill time operations are more variable. Finally, the Cycle time has an average of 19.97 with standard deviation, 0.56824. In general terms, the standard deviations of most parameters can be seen to be generally smaller, indicating that the process is well controlled, but there are also several parameters, such as the vacuum pressure and fill time, which are more potentially varying.

III. Correlation of Parameters with Fill Volume

In this research, it is important to know the relationship between parameters and fill volume due to a number of reasons. First, it allows making the prediction analysis because we can predict how the variation of one parameter like temperature, machine speed, timing parameter and vacuum pressure can affect the fill volume. Such predictive power is critical in quality control since the products are kept within the standards of consistency.

The relationship between different parameters and fill volume have been correlated to determine the existence of important relationships affecting filling process. The key correlations identified in the analysis are the following:

The correlation analysis shows that process-related timing, flow, pressure, and mechanical parameters are mostly the drivers of the expected fill volume. High positive values are seen between expected fill volume and valve open time ($r = 0.89$), cycle time ($r = 0.81$), product inlet pressure ($r = 0.82$), product temperature ($r = 0.81$), fill time ($r = 0.79$) and flow rate ($r = 0.74$), which means that higher duration of valve actuation, higher flow rate and higher pressure and temperature will lead to higher volumes of product being dispensed. Likewise, valve delay, valve delay ratio, and shutoff delay show medium to high positive relationships (r = 0.68-0.70), indicating that slow valve response is a source of additional product flow after the target cutoff point. The use of cycle time also displays a positive relationship ($r = 0.67$) supporting the effect of the general timing of the process on filling results. Conversely, expected fill volume is strongly negatively correlated with vacuum pressure ($r = -0.84$) and viscosity ($r = -0.72$), meaning that the less negative vacuum situations are, the less suction drives the flow in the bottle, whereas the more viscous the fluid is, the more it resists flow and thus the lower the filled volumes. Nozzle wear has a weak to moderate positive relationship with the expected fill volume ($r = 0.32$), which indicates degradation effect on filling accuracy over the time instead of dominant effect. Parameters of product composition such as the density ($r = 0.01$) and alcohol by volume (ABV) percentage ($r = 0.06$) have negligible correlation with the predicted fill volume, indicating that they are not a significant cause of variability in filling. Altogether, these results show that the expected fill volume is majorly regulated by the dynamic process and mechanical parameters and the material properties have little impact, which supports the significance of tracking and managing the flow, timing, and pressure variables to prevent overfilling and underfilling.



Source**:** Own, based on data from NALF (June-September 2025)

Figure 4.18 Correlation of parameters with fill volume

4.4 Synthetic Data Composition and Validation

This study employs synthetic data derived from past process data collected from the National Alcohol and Liquor Factory (NALF). Synthetic data was chosen to address the need for additional historical data while also maintaining the realism and operational aspects of the actual filling process. But instead of replacing real data, synthetic data was used to augment and generalize historical observations, enabling effective development of machine learning models.

I. Synthetic Data Composition

NALF's production history served as a base line of synthetic data. These data were presented with realistic operating ranges, nominal values, and patterns of variability for key process parameters such as valve timing, vacuum pressure, inlet pressure, flow rate, product properties, cycle time, and mechanical condition. Additionally, the structured interviews with experienced operators were used to interpret the history and to identify abnormalities associated with underfilling and overfilling.

From this historical baseline, synthetic data were generated through a rule-based probabilistic and parametric simulation. In this analysis, each process variable was sampled at various predetermined ranges based on the historical National Alcohol and liquor factory (NALF) data and the probability distribution chosen to allow processes to vary. These data were generated by explicitly specifying physical and operational constraints in historical National Alcohol and liquor factory (NALF) data, including machine cycle limits and valve timing relationships.

The volume of filling for each synthetic record was estimated using a parametric flow model based on the vacuum filling principle. Each data point was divided into three operational categories based on the estimated volume: underfill, correct fill, and overfill. This created a data set that reflects both the daily operating conditions and deviations in the National Alcohol and liquor factory (NALF) production environment.

II. Synthetic data validation

Validity of the synthetic dataset was determined by use of a structured validation system which comprised of physical feasibility validation, statistical distribution validation, and correlation validity. The multi-layered design of this method guaranteed that the data that are produced will reflect the behaviour of the rotary vacuum filling process accurately and can be used later in the analysis of machine learning.

First, physical feasibility was carried out to ensure that all the generated samples matched the basic constraints of the filling system. Parameters like fill time, valve open time, actuation delays, etc were limited to non-negative numbers, and constrained to not exceed the machine cycle time. The flow rate and vacuum pressure numbers were limited to realistic operational ranges within the specifications and knowledge of the operators. Any physical or logical violation was rejected.

Second, statistical distribution validation was performed by comparing the distributions of synthesized variables with historical NALF data. Descriptive statistics and distribution plots indicated that synthetic data closely matched the central tendencies and variability patterns of the real process while providing coverage for a few, but plausible operating conditions.

Third, correlation analysis was employed to ensure that the physical relationship between variables that were known to exist was maintained in the dataset. Anticipated interdependence like the positive association between effective fill time and filled volume was evident as well as the effect of vacuum pressure on the flow rate. The correlations prove that the man-made data do not exhibit arbitrary correlations but still remain causally consistent and therefore reliably train machine learning model

The general outcome of the validation is that the synthetic dataset is physically realistic, statistically consistent, and behavioral. As a result, the data are considered valid and suitable to be used in the production of machine learning models to predict the accuracy of filling and optimize the same (Appendix IV-Synthetic data descriptive statistics, distribution and correlation).

III. Acceptance criteria

The Go / No-Go decision framework was developed to objectively determine whether the generated synthetic dataset adequately represents the statistical and physical characteristics of the filling process data from the company.

This decision was based on quantitative statistical similarity measures, distribution comparison, correlation preservation, and physical feasibility validation.

If predefined acceptance criteria were satisfied then GO decision (synthetic data accepted). If criteria were not satisfied then NO-GO decision (data rejected or regenerated)

Table 4. 6 Acceptance criteria (Go and No-go decision)

No	Validation Dimension	Acceptance threshold	Result	Decision
1	Mean	≤5%	Satisfied	pass
2	Std deviation	≤10%	Satisfied	pass
3	Correlation	≤0.1	Satisfied	pass
4	Physical feasibility	100 valid	100valid	Pass

Source**:** Own, based on data from NALF (June-September 2025)

Based on the above table 4.6, all validation criteria were satisfied, it means the synthetic dataset closely resembles the real company data in terms of distribution, variability, correlation structure, and engineering feasibility. The absence of physical constraint violations confirms that the generated values are realistic and operationally meaningful within the filling process.

4.5 Machine Learning (ML) Model developing for the Classification Filling outcome and Integration with Filling Process

In order to solve the problem of filling in accuracy during the bottling process, Supervised machine learning (ML) models were trained to provide predictions on filling accuracy when a combination of parameters was selected. These models were developed in a systematic manner, where features were appropriately selected, data was prepared, and algorithms were appropriately utilized.

I. Feature Selection and Engineering for model development

After the correlation analysis, only parameters that were significant with expected fill volume were retained for subsequent model development. These parameters included vacuum pressure, inlet pressure, cycle time, fill time, valve open time, valve delay, nozzle wear, viscosity, and product temperature. Density and alcohol by volume (ABV) were excluded because they were poorly correlated with fill volume and they do not have a significant role in filling dynamics within the operating range of the current system.

Features engineering was applied to the retained parameters to increase robustness and interpretability of the model. In particular, valve delay ration drive from valve open time and valve delay to represent the effective valve actuation behavior and the cycle time utilization was calculated from cycle time and fill time to represent the fluid delivery efficiency within each production cycle. These engineered features, along with expected fill volume, were the final set of inputs for the machine learning model. Hence, the model was constructed in terms of statistically significant, physically meaningful, and process relevant features, resulting in better predictability and generalizability.

II. Data preparation and Algorithm Selection

Preprocessing of the dataset was done before it was model trained in order to make it quality and consistent. The steps that were used are as follows:
Data Cleaning: The missing values were filled in with mean imputation (numerical features) and mode imputation (categorical features). Unrealistic values (such as negative flow rates) were eliminated.

Train-Test Split: The data were split into a training (70%) and testing (30%) sets and stratified sampling was maintained to achieve a balanced sample in terms of classes underfill, correct fill, and overfill.

As the aim of the research is to categorize and forecast the precision of the filling process by the essential operational parameters, the algorithms that showed high efficiency in classifying and regression tasks in the framework of optimization of the industrial processes have been taken into consideration.

As in the research, Random forest algorithm was selected due to their strength, interpretability and proven ability to process complex non-linear industrial data.

4.5.1 Model Train, Test, Evaluations and Validation

I. Machine learning model training

Training was done using labelled filling process data from the NALF production line. Preprocessing included synchronization for each bottle, elimination of outliers, and normalization. The decision boundaries were constructed using the training dataset for correct fill, underfill risk, and overfill risk classes. Cross-validation was employed to ensure robustness and avoid overfitting.

1. Training Accuracy

$$\text{Training Accuracy} = \frac{\sum TP}{N} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \text{(Equation 4.4)}$$

Where:

TP is total number of correctly classified sample

N is total number of training samples

$$\text{Training Accuracy} = \frac{2789}{2800}$$

$$\text{Training Accuracy} = \underline{0.9961 = 99.6\%}$$

2. Confusion matrix analysis (training dataset)

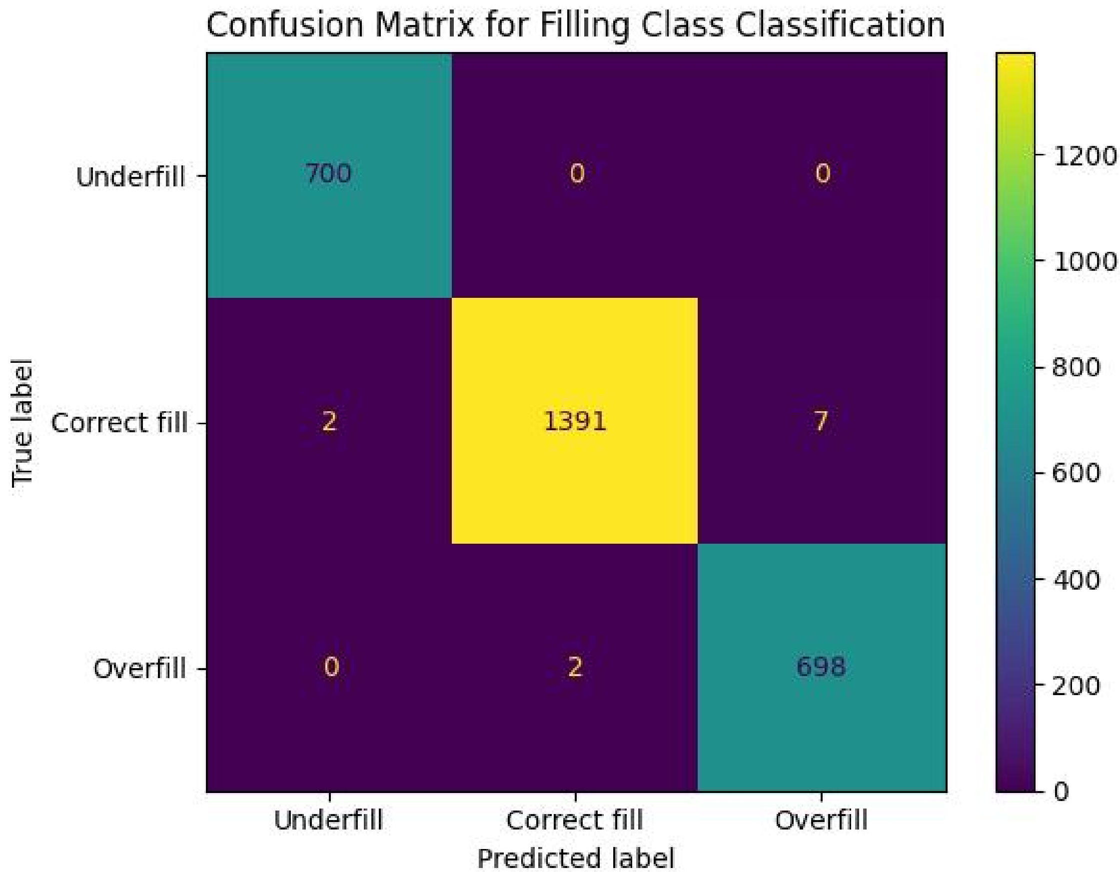
A confusion matrix was useful in fill level classification to determine the performance of a model in terms of differentiating between fill levels (e.g., correct fill, Overfill, underfill). The values in each cell of the matrix was reflect the success of the model in predicting each type of fill level.

The Underfill class is showing perfect performance and this is reflected by the confusion matrix. The model was able to recognize all 700 cases of bottles that were labeled as not being filled perfectly. This training accuracy shows that the model is very efficient in identification of underfilled bottles.

The model was also very effective in the Correct Fill class, with a high number of correctly filled bottles (1391) being classified correctly. Nonetheless, they had seven cases that were mistaken to be overfilled and two bottles misclassified as underfill, which shows that there are some challenges in identifying between the three categories. Although these misclassifications are generally very important in identifying properly filled bottles, such misclassifications are the way to enhance the false negative and false positive.

The Overfill class is also showing a great performance in the confusion matrix as the model correctly predicted 698 cases of the overfilled bottles. Yet, it misclassified two cases as Correct Fill which implies that the system works, but it is not entirely easy to distinguish between correctly filled and overfilled bottles. The misclassifications, even though in minor amounts, underscore the fact that the model should be constantly improved. To enhance the capability of the model to effectively identify the differences between these closely related categories, it is possible that the model should be given diverse and representative training data.

The Random Forest classifier was achieved highly accurate on the training dataset, which indicates that it is possible to understand how the filling process variables interact with filling condition classes. It shows that there is significant diagonal dominance in the training confusion matrix, and that most of the underfill, correct fill, and overfill samples are clearly distinguished. The majority of misclassification occurred within adjacent classes, as fill conditions within tolerance limits intersected.



Source**:** Own, based on data from NALF (June-September 2025)

Figure 4.19 Confusion matrix for training

II. Model Testing and performance analysis

Model testing is used to test the trained machine learning model from an unseen dataset which was not used in the model training. Testing aims to determine whether the model can generalize learned patterns and accurately classify underfill, correctly filled, and overfilled under new and realistic operating conditions.

1**. Testing performance analysis**

a. Testing Accuracy

$$\text{Testing Accuracy} = \frac{\sum TP}{N}$$

Where:

TP is total number of correctly classified sample

N is total number of testing samples

$$\text{Testing Accuracy} = \frac{1192}{1200}$$

$$\text{Testing Accuracy} = \underline{0.9933 = 99.33\%}$$

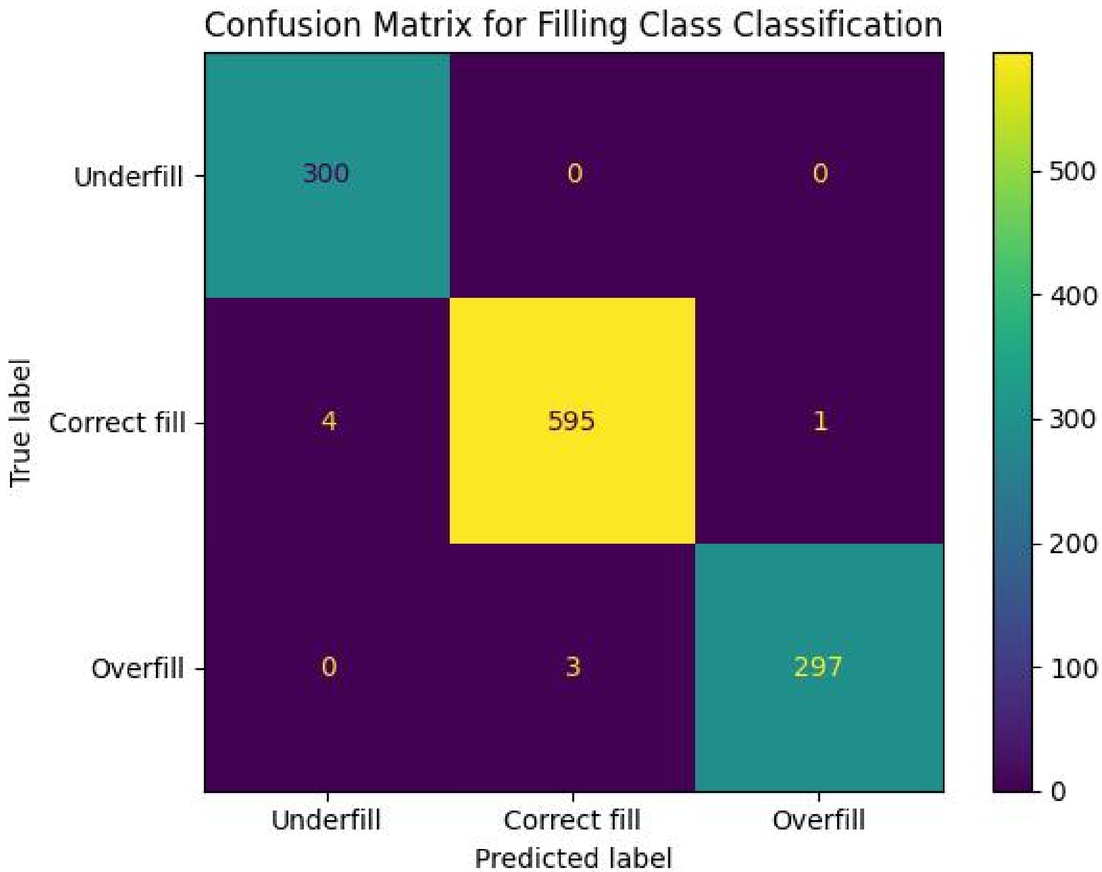
b. Confusion matrix analysis

In machine learning, a confusion matrix is utilized to assess the performance of a classification model; it is a prediction performance measurement tool for accuracy of a classification model. The confusion matrix from the testing data is a detailed class-wise assessment of model performance. The matrices are highly diagonally dominant; thus, a majority of the samples for each filling classes were correctly classified.

The Underfill class performs well with 300 instances correctly classified as Underfill. This means the model effectively detects most cases in this subset. However, 4 instances were misclassified; these actual Underfill cases were misclassified as Correct Fill.

The model scored 595 correct classifications in the Correct Fill category, demonstrating it can correctly identify instances of this type of fill. However, this process did not work out as well: 4 cases of Correct Fill were incorrectly labeled Underfill, and 1 case was incorrectly labeled Overfill. These misclassifications imply some overlap between the Correct fill and the other classes, suggesting that the model may not always be able to distinguish between them.

This shows the model's usefulness when it comes to identifying Overfill instances. But it was subject to 3 misclassifications, where actual Overfill instances were incorrectly assigned to Correct Fill. This suggests the model is successful in recognizing true Overfill cases, as there are no false negatives.



Source: Own, based on data from NALF (June-September 2025)

Figure 4.20 Confusion matrix for test

2. Class Wise Precision recall and F1-score

Class-wise precision, recall, and F1-score are important to assess classification models due to the fact that they offer a detailed insight into the performance of the model, deal with the problem of class imbalance, direct improvement, and ensure reliability in critical areas.

Table 4.7 Class Wise Predicted and True Values

Prediction Vs Actual	Predicted Underfill	Predicted Correct Fill	Predicted Overfill
True Underfill	300	0	0
True Correct Fill	4	595	1
True Overfill	0	3	297

Source: Own, based on data from NALF (June-September 2025)

Class wise calculation Precision, Recall and F1-Score

Precision $(P)_{Class}$ $P=\frac{TP}{TP+FP}$………………… (Equation 4.5)

Recall $(R)_{Class}$ $R=\frac{TP}{TP+FN}$……………………… (Equation 4.6)

F1 Score $(F1)_{Class}$ $F1=2*\frac{P*R}{P+R}$……………………... (Equation 4.7)

Table 4.8 Class Wise True positive, False positive and False negative.

Class	TP	FP	FN
Underfill	300	4	0
Correct Fill	595	3	5
Overfill	297	1	3

Source: Own, based on data from NALF (June-September 2025)

Table 4.9 Class Wise Precision, Recall and F1 score.

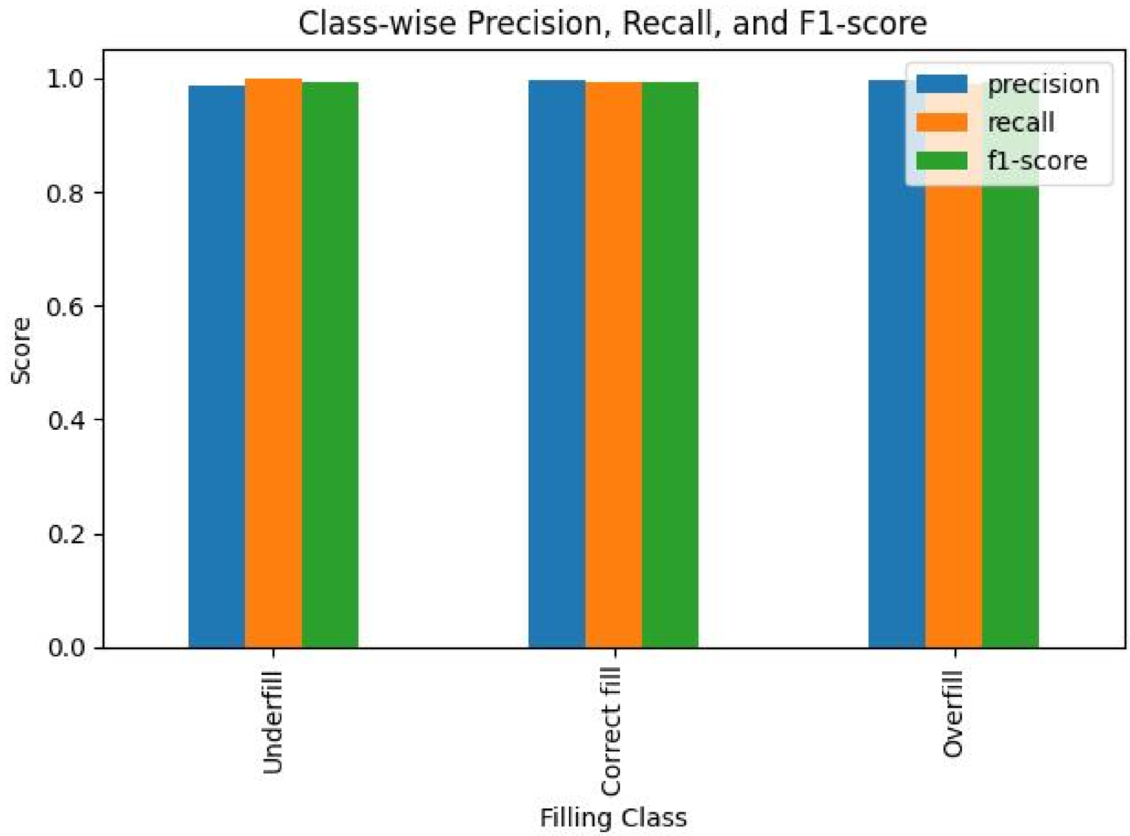
	Precision	Recall	f1-score
Underfill	0.987	1	0.9934
Correct fill	0.995	0.992	0.9933
Overfill	0.997	0.990	0.9933

Source: Own, based on data from NALF (June-September 2025)

Three major metrics: precision, recall, and F1 score, may be used to examine the model performance in terms of its individual classification efficiency to each fill condition, which is useful in the classification of the liquor and alcohol processing filling conditions. The precision measures of the "Underfill" class were very impressive as the rate of accuracy was about 0.987 meaning that almost all the cases classified as "Underfill" were true positives. Also, this model achieved a maximum score of 1 (100%) in the recall, indicating that this model was effective in detecting all real instances of Underfill and no false instances of the same, thus performing zero false negatives. This outstanding performance is further supported by the F1 score of approximately 0.993 which indicates a good balance between precision and recall. These findings suggest that the model is quite useful in the correct identification of the cases of Underfill, thus a very useful element in the classification.

In the case of the correct fill class, the model showed excellent results with a precision of approximately 0.995. This high precision rate means that in cases where the model predicts an instance to be Correct fill, it is correct 99.5 %, which means that there are very few misclassifications. Besides, the recall rate of this model was around 0.992, which proves that the model virtually perfectly hits the true instances of Correct fill with insignificant false negatives. The fact that the resulting F1 score is approximately 0.993 is also an indicator of the effectiveness of the model, having a praiseworthy precision and recall rate. On balance, these measures suggest that the model is effective in differentiating the cases of "Correct fill" and this aspect adds to the accuracy of classifying it.

The class performance analysis of the category Overfill gave impressive results, and its precision is approximately about 0.997. This shows that most of the cases which were predicted to be "Overfill" were truly correct though the model also had some false positives. The recall on this model was also approximately 0.990 and this shows that this model was able to detect a significant number of real cases of Overfill, although the number of false negatives did exist. A high balance between precision and recall is indicated by the F1 score of about 0.993. Although marginally less than the rest of the classes, these metrics confirm that the model is good at correctly classifying the cases of Overfill as well as makes a positive contribution to the overall classification.



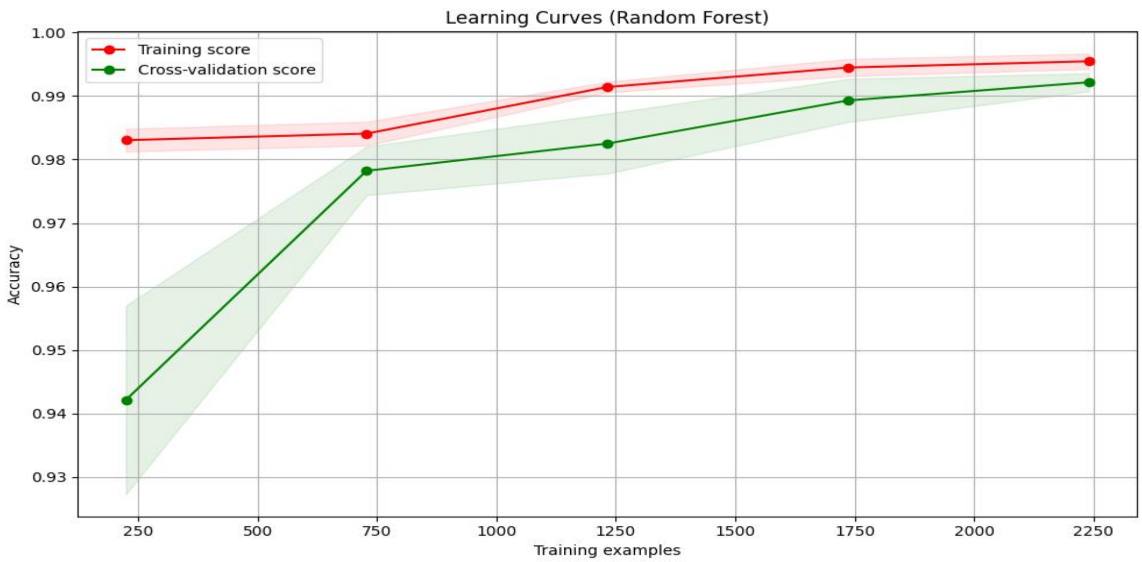
Source: Own, based on data from NALF (June-September 2025)

Figure 4.21 Class wise precision, recall and f1-score

III. Overfit and Underfit test

Learning curve

A learning curve is a graphical representation that characterizes the relationship between the training data size and the model's accuracy. The learning curve shows that there are some valuable features of the performance of a Random Forest model. To begin with, the accuracy of the training is very high and tends to level off with increase in the training examples implying that the model gets to learn on the basis of the training data with no indication of underfitting. Also, the cross-validation score is also large and very close to the training score, which means that the model has generalization to unknown data and is not overfitting. The occurrence of the small gap between the training and cross-validation scores further supports this fact because a bigger gap would normally indicate that the model is doing well on training data, and good on the validation data. Lastly, the fact that the curve levels off when the sample of training data becomes larger indicates that there is a threshold after which the accuracy of the model is unlikely to improve significantly with more data. All in all, the learning curve indicates that the Random Forest model is highly tuned such that it balances well on learning training data and at the same time, successfully generalizing the novel and not seen data.



Source**:** Own, based on data from NALF (June-September 2025)

Figure 4.22 Learning curve (Training Vs cross-validation)

2.Test vs Training Score

The study examines the difference in performance between training and validation data in checking whether a study is overfitting or underfitting. A properly fitted model has a high accuracy on both sets with an insignificant variation, meaning it has learnt significant patterns that can be generalized instead of memorizing the training data.

Table 4.10 Training Vs Testing score

		Score
1	Training	99.61%
2	Testing	99.33%

Source**:** Own, based on data from NALF (June-September 2025)

As mention in the table 4.7, the model shows great performance without any indication of overfitting or underfitting. The narrow margin of just 0.27 % existent between training (99.60) and testing accuracy (99.33) implies that the model has acquired the underlying patterns without memorizing the training data. This implies that the model generalizes a lot and would possibly work on new unseen data in the same manner. The accuracy on both sets being high confirms that the model is not too simple (underfitting) or too complex (overfitting) but instead represents the optimal balance which makes it reliable and ready to be deployed.

IV. Model Validation

Model validation is defined in this study as the systematic verification of the machine learning model's ability to accurately predict the filling statuses such as correct fill, underfill risk, or overfill risk. in real-world production conditions at NALF, which can be used for real-time monitoring and dynamic control of the filling process.

a. Precision, Recall and F1-score validation

In order to validate the predictive capability of the developed model as compared to the general predictive accuracy, there was the use of precision, recall, and F1 score as metrics of class specificity. Such metrics give an accurate evaluation of the capability of the model to categorize correctly the state of underfill, correct fill and overfill, which is critical to effective control of the filling process.

Both weighted-average and macro-averaged measures were examined to have a balance of assessment across all filling classes. Macro-averaging gives the same weight to underfill, correct fill and overfill classes, thus showing the model performance concerning the minority conditions or high-risk conditions. Weighted-averaging allows the distribution of classes and real production distributions, and that gives a realistic evaluation of the projected operational performance.

Macro Average Metrics

Macro-averaging calculates each measured metric individually for each class and takes the unweighted arithmetic mean for each class.

For particular metrics M across n class

Macro Average = $\frac{1}{n}\sum_{i=n}^{n} M_i$……………………………………(Equation 4.8)

Where M_i is the metric score for class i.

Mean Precision $=\frac{1}{n}\sum_{i=n}^{n} M_i$……………………………...... (Equation 4.9)

Mean F1-score$_{macro}$ $=\frac{1}{n}\sum_{i=n}^{n} M_i$…………………………….. (Equation 4.10)

Mean Recall = $\frac{1}{n}\sum_{i=n}^{n} M_i$... (Equation 4.11)

Table 4.11 Macro Average score

Mean Accuracy	Mean Precision	Mean Recall (macro)	Mean F1-score (macro)
0.9944	0.9928	0.9938	0.9933

Source: Own, based on data from NALF (June-September 2025)

Weighted Average metrics

For particular metrics M across n class

Weighted Average = $\frac{n}{\sum_{j=1}^{c} nj}$... (Equation 4.11)

Where M_i is the metric score for class i.

Precision $_{\text{weighted}}$ $= \sum_{i=1}^{c} Wi * Precision_i$............................. (Equation 4.12)

F1 $_{\text{weighted}}$ $= \sum_{i=1}^{c} Wi * F1_i$... (Equation 4.13)

Recall $_{\text{weighted}}$ $= \sum_{i=1}^{c} Wi * Recall_i$.................................... (Equation 4.14)

Table 4.12 Weighted Average score

Mean Precision (Weighted)	Mean Recall (Weighted)	Mean F1-score (Weighted)
0.99336	0.99333	0.99333

Source: Own, based on data from NALF (June-September 2025)

The macro-averaged precision (0.9928), recall (0.9938), and F1-score (0.9933) indicate that the trained Random Forest classifier consistently performs as a superior classifier across all filling classes underfill, correct fill, and overfill, and assigns equal importance to each class. Additionally, since the recall is high with the high macro average, the model is able to identify the filling deviations without discriminating the dominant correct fill class. This indicates that underfill and overfill events, regulatory and cost-critical events, are accurately identified.

Likewise, the weighted average precision of 0.99336, recall of 0.9933, and F1-score of 0.9933 are representative of model performance in actual production where correct filling occurs more frequently. The close score between macro-averaged and weighted average indicates that class imbalance is not particularly sensitive to model behavior. This shows that both the sensitivity to risk and the stability of operation are maintained with the classifier.

Overall, the consistently high and closely well-aligned macro and weighted scores indicate that the model is robust, unbiased, and generalizable. These findings support the model's feasibility for both real-time monitoring and control of the dynamic filling process at the National Alcohol and Liquor Factory.

b. Stratified K-fold cross validation

The k-fold cross-validation technique is a resampling method for model validation to determine the performance of a Random Forest model on unseen data. In a multi-class bottle filling problem, where all filling conditions are underfill (0), correct fill (1), and overfill (2), k-fold validation ensures that the model performance is not biased by any single split in the training and testing set and that all filling conditions are represented.

This stratification also ensures that each fold is consistent with the original class distribution of the dataset and thus maintains a representative distribution of underfilled, correctly filled, and overfilled samples across sets. This is particularly important in the bottling process, where correct fill is observed most frequently and underfill or overfill events are less frequent. Given the class balance across folds, stratified K-Fold validation is an accurate indicator of the model's generalization performance under diverse operating conditions.

Table 4.13 Five-fold (5-Fold) mean and std deviation score

		Score
1	Mean	0.9911
2	Std deviation	0.00467

Source**:** Own, based on data from NALF (June-September 2025)

Random Forest model with stratified k fold cross-validation had a mean score of 0.9911 and a standard deviation of 0.0047 on the training dataset.

The large mean F1-score shows that the model exhibits good overall classification performance in all three filling conditions underfill, correct fill, and overfill because it performs well in balancing precision and recall in a multi-class environment. This implies that the model can detect correctly the two minority fault classes (underfill and overfill) with reasonable consistency in the dominant correct-fill class.

The standard deviation (+-0.0047) is low indicating high stability in performance across the k validation folds. Since stratification maintains a balance of classes in each fold, the lowest variability prove that the predictive ability of the model is not related to data partitioning and is constant under various operating conditions, including changes in flow rate, pressure, the timing of valves and temperature.

c. Out-of- bag (OOB) error Validation

In this study, the Random Forest model is based on bootstrap sampling, where each decision tree is trained on a randomly selected set of training data drawn with replacement. This results in a set of observations that have been repeated in the tree's training set while others are excluded. Out-of-bag (OOB) samples represent the excluded observations and represent roughly 36.8% of all data from all trees. OOB samples were then used as an internal validation set based on cumulative prediction made from trees in which the observation was not used for training. This provides an unbiased estimate of model performance for the multi-class filling classification problem: underfill, correct fill, and overfill, without additional validation data.

Table 4.14: Out-of-bag (OOB) vs Test score

		Score
1	Out-of-bag (OOB)	0.9925 (99.25%)
2	Test accuracy	0.9933 (99.33%)

Source**:** Own, based on data from NALF (June-September 2025)

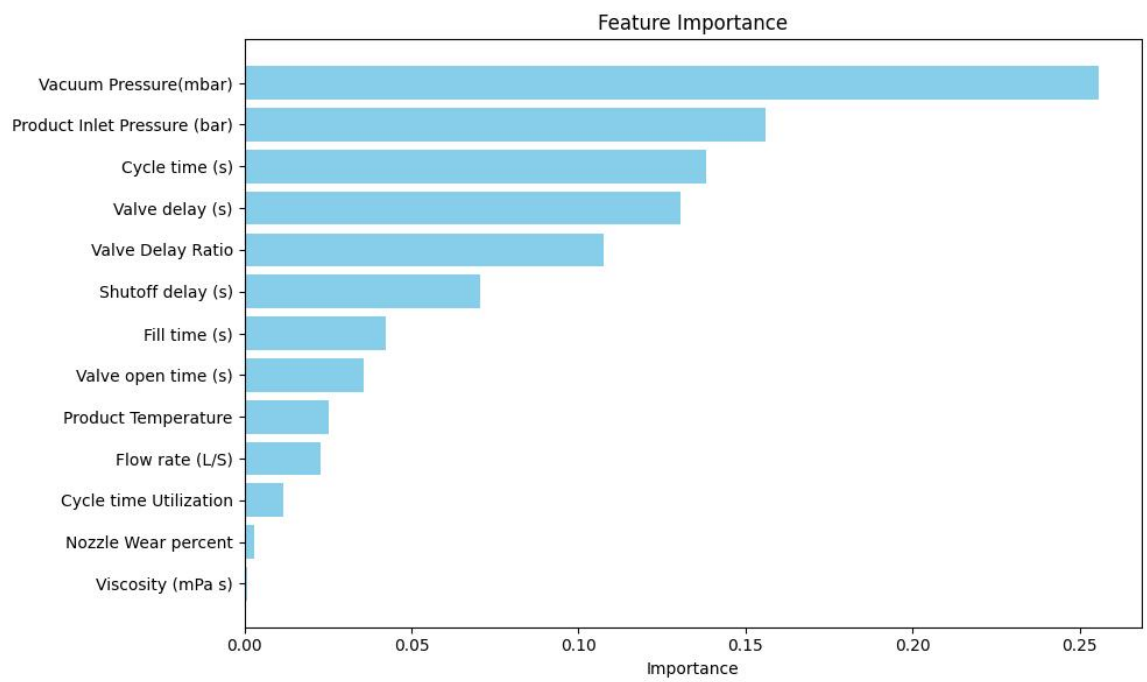
The Random Forest model has an out-of-bag (OOB) score of 99.25%, which indicates that it correctly identifies most observations that were not used to train individual trees in the ensemble. The very small difference between the out-of-bag (OOB) score and independent test accuracy (99.33%) demonstrates a high level of consistency in model performance and suggests robust generalization capabilities. These similarities suggest that the regularization strategy and ensemble structure mitigate overfitting and enable high predictive accuracy across underfill, correct fill, and overfill classes at different operating conditions.

4.5.2 Analysis of Feature importance of the Developed Machine Learning Model

The feature importance analysis is a vital part of the given study to make sure that the developed Machine Learning model to control the dynamic filling process can be reliable, interpretable and physically meaningful, especially when a direct volume measurement sensor is not present. Because the variables of fill accuracy are based on indirect process variables, it is important to know which parameters are most effective in determining the classification of the model of filling conditions. The analysis of feature importance can be seen to determine the important drivers of underfill and overfill behavior which can be checked to ensure that the decisions made by the model are made based on process-relevant variables and not the spurious linkages. This interpretability is important to deploying it in industry since it provides operator confidence, allows integration with PLC based control systems, and allows its use to meet quality and regulatory standards. Moreover, the importance of features guides directly the design of the control strategy since the highest priorities of the corrective actions should be given to the most influential parameters, which enhances the speed of response, stability of control, and resistance to the disturbances of the processes and mechanical wear. In this regard, the role of feature importance analysis becomes a vital point in linking the concept of data-driven intelligence and the set of established process engineering principles to guarantee the effectiveness and industrial feasibility of the ML-based control system.

I. Feature importance

The feature importance analysis shows that vacuum pressure is the most important factor in determining filling conditions with an importance score of 0.2558(25.6%). This confirms that vacuum strength has a direct effect on both the equilibrium draw rate and the volume transfer at the filling stage. The second most significant parameter is product inlet pressure with an important score of 0.1560 (15.6%), which stabilizes flow into the filling system and compensates for upstream pressure changes. This also explains the important role of temporal control in volumetric consistency in terms of cycle time (0.1382) and valve delay (0.1305). The valve delay ratio (0.1076) emphasizes the high sensitivity of the filling process to synchronization of valve actuation and flow initiation. Parameters such as shutoff delay (0.0707), fill time (0.0422), and valve open time (0.0355) are moderately involved, suggesting that small timing differences in the valve closure, and opening can generate measurable fill-level variations. But product temperature (0.0251), flow rate (0.023) and cycle time (0.012) are relatively small effects, indicating they are largely mediated indirectly by dominant pressure and timing conditions. Finally, the use of nozzle wear (0.0031), and product viscosity (0.0009) does not play a significant role in the assessment of their respective operating conditions; therefore, their effect on fill order is secondary and only has a significant impact in the case of extreme or long-term degradation. Overall, these results indicate that fill-level accuracy in the studied filling process is largely controlled by pressure and timing and therefore provides a clear reason for control actions in ML-based dynamic process control.



Source**:** Own, based on data from NALF (June-September 2025)

Figure 4.23 Feature Importance

4.5.3 Implications of the Developed Machine Learning Model for Dynamic Process Control

The results of the feature importance feature make a direct and practical impact on the development and application of the Machine Learning (ML) based strategy of the dynamic process control. The decisive rule of the vacuum pressure, then the flow rate, and the parameters of valve timing, well shows that the effective control measures must be aimed at regulating these three variables in order to quickly and stabilize the underfill and overfill risks. The Machine Learning (ML system can react effectively to disturbances in the process and reduce the number of unwarranted alterations to low-impact variables by placing increased control effort on the parameters that most affect the system, and thus, reducing control oscillations, and maintaining the stability of processes. Moreover, the average effect of mechanical condition pointers, including wear in the nozzle and shutoff delay, underscores the capability of the Machine Learning (ML) based controller, which can be adjusted to gradual deterioration of equipment without the need to recalibrate the controller on a regular basis. The insignificance of the product property variables agrees that the control strategy can largely be the same even among product batches that are standardized thus easier to deploy in industrial settings. Altogether, these results allow building a prioritized, interpretable, and physically consistent closed-loop control system where the decisions made through ML help to increase the accuracy of filling, minimize the loss of products, and contribute to the efficient operation at high speeds when no direct volume measurement sensors are in place.

4.5.4 Integration of developed Machine Learning Model with The Filling Process

Liquid product filling systems used in industries are subject to dynamic disturbances due to variation of product, mechanical wear and variation in the operation. Conventional methods of rule-based control do not have the ability to assure stable fill accuracy, particularly where there are no direct volume measurement sensors. Machine Learning (ML) has emerged as a potential solution with the ability to learn the complex multivariate relationships which cannot be described using the traditional control logic.

A classification-based machine learning (ML) model with the highest accuracy of 99.33% was developed to dynamically control the filling condition based on which each cycle was classified as Correct Fill, Underfill Risk and Overfill Risk, used in combination with the vacuum filler control system that helped to adjust the process closed-loop. Importance of features analysis also offered a sense of control and interpretation of the control priorities.

I. Integration architecture

Three-layer integration architectures presentation layer (HMI), processing layer(intelligence) and data accusation layer

The integration system has a three-layer structure, which divides the concerns and ensures consistency across the system. The Data Acquisition Layer is the data collection layer that gathers process data through basic sensors networks and from existing PLC systems if available. This layer provides data validation and preprocessing for quality before transmission to the upper layers. The Processing Layer is the intelligence layer and contains the 99.31% accuracy trained machine learning (ML) model, feature engineering pipelines, and complex control algorithms that translate prediction into actionable information. The Presentation Layer provides human-machine interfaces via web and mobile applications with real-time visualization, alert systems, and control interfaces that enable operators to interact with the predictive system in real time.

II. Dynamic filling process Monitoring function

The machine learning enabled system proposed, therefore, fundamentally improves the monitoring capacity of the National Alcohol and Liquor Factory (NALF), as it transforms post-process inspection into real-time, in-process intelligence.

First, the filling process provides real-time classification of fill condition for each bottle. Rather than measuring volume downstream or performing manual inspection, the trained multi-class machine learning model classifies every filling event as either correct fill, underfill risk, or overfill risk prior to the bottle exiting the filling zone. This allows corrections to be taken early on in the development of the deviations.

Second, trend-based deviation detection allows the system to detect process drifts that are not identified in individual bottles. It evaluates sequential prediction results in an incremental manner that reveals slow degradation phenomena such as vacuum pressure drift, valve response delay, or progressive nozzle wear. These trends are often obscured in traditional inspection systems until significant losses occur.

Third, there is an intelligence component of confidence-based monitoring and alerting. The confidence score for each prediction indicates that the operators can distinguish between stable operating conditions and uncertain or unusual processing states. High confidence predictions provide confidence in system behavior, while low confidence outputs produce alerts for operator observation or investigation.

In summary, this type of monitoring shifts the quality management function of NALF from reactive detection at the time of filling to more accurate, real-time monitoring at the filling procedure thereby greatly reducing product loss, reprocessing, and operating inefficiency.

III. Dynamic filling process control function

The proposed system builds on the ability to monitor real-time, and facilitates automatic, closed loop control of the filling process at NALF using machine learning based on predictors rather than observed defects.

For immediate corrective action, when the model predicts an underfilling or overfill risk with high confidence, the system makes immediate corrections for the next bottle. These controls are prioritized according to feature importance and include adjusting the vacuum pressure, followed by the flow rate, and timing of the valves. This allows us to quickly correct deviations without delays.

For a more gradual adaptive control, the system makes incremental parameter changes when there are persistent trends across multiple filling cycles. This strategy minimizes aggressive control reactions that may disturb the process but also compensates for slow variations such as pressure decay or timing issues.

Additionally, equipment sensitivity compensates for mechanical degradation. Detections of nozzle wear or valve response degradation trigger predefined compensation rules rather than manual recalibration. It thereby reduces mechanical stress, extends the life of parts, and maintains long-run processes.

This control system dynamically varies its degree of automation depending on the confidence of the predictions, the level of risk and operation conditions. Level 1 (Fully Automated Control) allows autonomous control when the model confidence is more than 95% and the conditions in the process are not changing. Level 2 (Semi-Automated Control) is prompted when medium-confidence predictions are done, and the operator must confirm them, and an automatic execution safety is provided after a 30-second delay.

Level 3 (Advisory mode) offers control recommendations when the conditions are low-confidence and the operator is to make final decisions. Level 4 (Manual Mode) is undocumented and this is a level that is used to disable automated control in case of a safety-critical event or high system uncertainty. The system carries out transitions between levels of control automatically without human intervention to guarantee adaptive and risk aware operation.

4.5.5 Machine learning (ML) based monitoring and control framework

The framework proposed incorporates the process data, the information about the state of equipment, and the properties of the products to allow intelligent monitoring and control of the bottle filling process. The data collection includes several sources, such as: production and quality logs, sensor measurements (pressure, flow and timing) product characteristics (temperature and viscosity) and equipment condition indicators like nozzle wear, which are then synchronized by means of a time-based data collection system.

The synchronized data of the multi-sensors are then subjected to a machine learning-developed prediction module to determine the filling condition and identify anomalous trends. The system also has a mechanism to detect possible deviations using confidence based alert system that prevents major filling errors.

According to these outputs, a decision logic layer decides the suitable control strategy. These involve instantaneous corrective control measures, incremental compensatory measures, and equipment sensitive compensation to consider mechanical changes. Control signals are also used to control the filling operation by applying the signals to the filling machine actuators. The system provides this feedback loop and therefore, the system maintains learning and stability in the process so that filling accuracy and operational efficiency can be achieved.

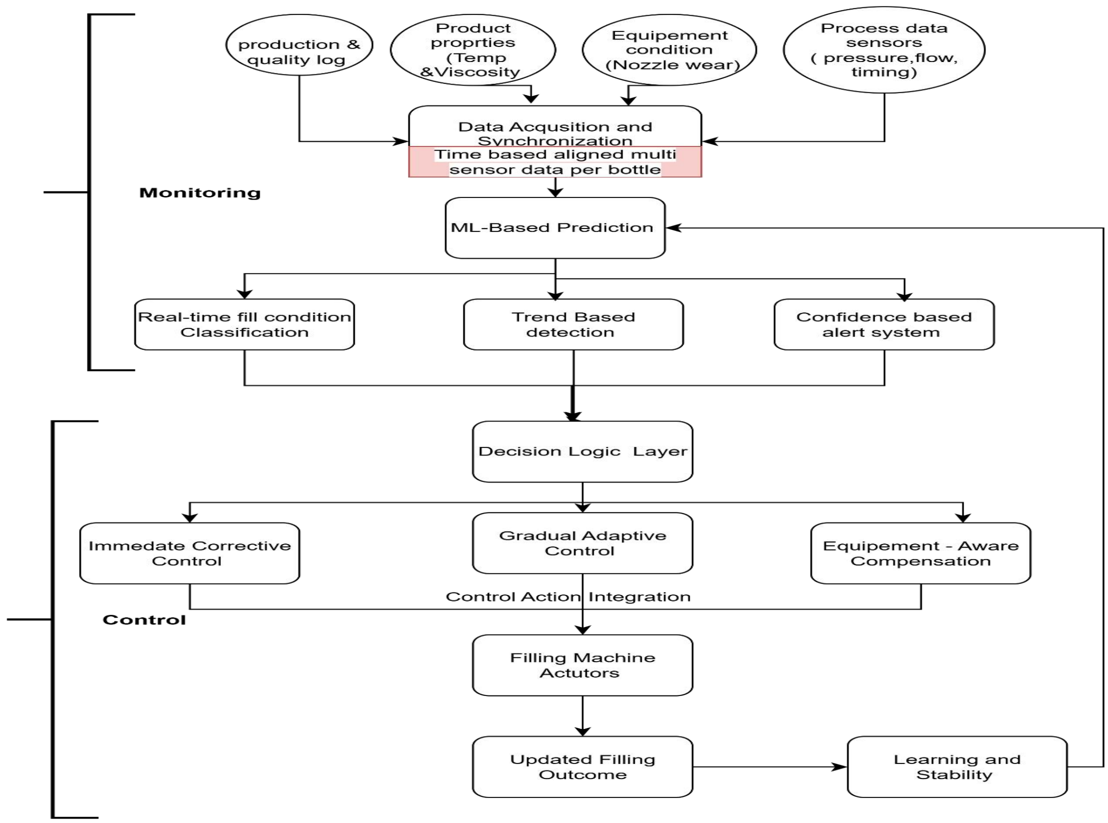


Figure 4.24 Machine learning (ML) based monitoring and control framework

4.6 Discussion

The primary purpose of this study was to examine the efficiency of the filling process in the current system at NALF, focusing on accuracy, overfilling, and underfilling rates. The accuracy of filling was 95.949%, which is acceptable but still less than the industry benchmark of 98% for acceptable filling accuracy. The findings also indicate that the overfilling rate was 1.693% and the underfilling rate was 2.358%, respectively. These figures illustrate major quality concerns as ideal standards recommend less than 1% overfill rate and require a minimum underfill rates for product integrity and customer satisfaction. In addition, an adequate recycling rate of at least 2% is another key part of the process, because of the commitment to resource efficiency and sustainability.

Overfill and Underfill rates were needs improvements. Even the smallest errors can cost more materially, reduce productivity, generate complaints from customers and weaken trust. Recent research such as the one from Doe et al. (2023) shows that even small deviations from optimal filling standards can have a significant impact on overall product perception and operational efficiency.

The second objective of the study to identify key parameters affect filling accuracy, this study was identified multiple parameters which affected filling accuracy such as, vacuum pressure, inlet pressure, cycle time, fill time, valve timing (valve open time, valve delay, shut-off delay), flow rate, product temperature, alcohol by volume (ABV), viscosity, density, and nozzle wear. Overall, the combined findings from root cause analysis, cause-and-effect analysis and the 5 Whys technique indicate that vacuum pressure, inlet pressure, cycle time, fill time, and valve timing parameters strongly affect liquid volume to be pumped per filling cycle.

The findings demonstrated that pressure and time variables directly control the force used to transfer liquid into bottles, and the duration of fluid transfer. The pressure differential between inlet pressure and vacuum pressure determines the flow, initiating and maintaining the process, and fill time and cycle time determine the effective window at which liquid is introduced. It is the valve timing that precisely determines the flow's start and stop point, which makes the filled product sensitive to small variations in timing. When these parameters are out of the appropriate range, the volume delivered can easily exceed the target and result in overfilling or underfilling. Fluid properties such as viscosity, density, temperature, and alcohol by volume (ABV) indirectly influence the accuracy of filling by altering flow resistance and discharge behavior under pressure and timing conditions.

These results are supported by other recent papers on liquid filling systems. Chen et al. (2025) concluded that volumetric accuracy is primarily dependent on pressure, duration of filling, and valve actuation, whereas fluid properties also depend on flow behavior. Similar results were obtained from Song et al. (2024) showing that both inlet pressure and valve opening are important factors in the internal flow dynamics and volumetric delivery of filling valves, that pressure and timing were dominant in controlling fill volume. Together, these data support the conclusion that accurate control of pressure differentials and valve timing is important to maintain consistent filling performance in high-speed liquid bottling operations.

The third purpose this study was to analysis the key parameters that affect filling accuracy by using distribution analysis and correlation analysis. This was necessary for understanding parameter behavior, identifying the most important contributors to filling variability, and providing an efficient means to select features within the machine learning model.

The distribution analysis of historical process data shows that major operating parameters such as valve open time, fill time, cycle time, flow rate, inlet pressure, and vacuum pressure are strikingly variable over the production cycles. This variation directly contributes to deviations in filling accuracy and clearly suggests that overfilling and underfilling are processes that fluctuate, rather than being random events. This has also been observed in recently published literature on liquid filling systems, in which the dispersion of timing and flow-related parameters was identified as a major cause of volumetric inconsistency (Kumar & Singh, 2021;Zhang et al., 2022).

Similarly, parameters of product composition such as density and alcohol by volume (ABV) showed narrow and stable distributions, which suggests minimal variation across batches. This finding agrees with earlier literature suggesting that in controlled production environments, material properties are relatively constant and have very little impact on volumetric filling accuracy relative to dynamic process parameters (Montgomery, 2020).

It also performed correlation analysis to determine the relationship between key parameters and the expected fill volume. The findings showed strong correlations between expected fill volume and valve open time, fill time, cycle time, flow rate, inlet pressure, and product temperature. These relationships indicate higher product flow rates, implying greater driving forces that increase product volume. This result follows recent work on automated filling systems that have identified valve timing and flow dynamics as the most significant impact on fill accuracy (Chen et al. 2021;Rossi & Mancini, 2023).

In contrast, inverse relationships between vacuum pressure and viscosity with the expected fill volume indicated negative correlations. A decreasing vacuum pressure reduces the flow of suction-driven fluid into the bottle, while a higher viscosity reduces product flow and volumetric throughput. These inverse relationships are well-known in the literature of fluid behavior and packaging studies and show the physical existence of these relationships (White, 2019; Li et al. 2022). Density and ABV were not significantly related to fill volume, indicating that product composition does not significantly influence filling variability during stable operation.

synthetic data were validated using compering with statistical distributions and physical consistency. The fact that the synthetic data distribution is close to the real data distribution shows that the synthetic dataset was able to capture the statistical properties of the filling process that were there in the underlying data. Also, physical consistency tests were done to ensure that the most important relationships between fill volume and valve timing, flow rate, vacuum pressure, and viscosity were maintained.

This is in line with current literature on industrial analytics, which has suggested that synthetic data can be effectively used to develop models when distributional similarity and physical constraints are considered (Frid-Adar et al. 2018; Xu et al. 2023).

The fourth objective was to develop, test, and validate a machine learning model that accurately classifies the filling condition as underfill risk, correct fill, and overfill risk. For this purpose, the study focused on systematic feature selection and engineering, rigorous training and testing practices, multiple measures of evaluation, and robust validation strategies.

Model development began with correlation-based feature selection for input variables selected based on their statistical significance with fill volume as described in previous analyses. This design maintained the integrity of the model's inputs as they were data driven and process relevant, consistent with recommendations from the recent manufacturing analytics literature that support the use of domain knowledge with statistics (Zhang et al. 2019; Wang et al. 2021). To further generalize the model, feature engineering techniques were employed, resulting in three features valve delay ratio, cycle time utilization, and expected fill volume. These engineered variables capture normalized process behavior and time efficiency independent from raw measurements, which have been shown to improve robustness in industrial machine learning models (Jia et al., 2020). Cleaning and transforming the data was conducted as part of preprocessing to ensure model stability and precision.

To balance the amount of learning and the independent evaluation of performance, the dataset was divided into a training and a testing split of 70% and 30%, respectively, which is a common practice in machine learning research. The model had a training accuracy of 99.6% signifying a good learning of the underlying dynamics of the filling process. Notably, the accuracy of tests 99.33% was very close to the result of the training accuracy, which indicated that the model can be successfully generalized to unseen data and that it does not overfit substantially. The same result has been demonstrated by recent research whereby models based on Random Forests confirmed minimal performance errors on the training and testing dataset when used in the complex manufacturing processes (Li et al. 2022 ; Ahmed et al. 2023).

A confusion matrix, precision, recall, and F1-score were used to evaluate the model as they are especially appropriate to multi-class industrial classification problems. The confusion matrix shows that there was very low misclassification through all filling outcomes. Underfill events obtained 100 % recall, so no underfilled bottles were missed and this is crucial in compliance with the regulation. High precision, recall, and F1-scores (all over 0.99) were also reported in correct fill and overfill classes. These findings are a sign of balanced performance and the effective control of false positives and false negatives. Similar investigations also find that ensemble-based classifiers attain a better class wise balance when evaluated with precision, recall, and f1 -based metrics as opposed to accuracy alone (He & Garcia, 2009 ; Zhang et al. 2019).

Model validation was conducted using several mechanisms to ensure robustness and reliability. First, the study analyzed macro and weighted average precision, recall, and F1-scores. Although a strong alignment of macro and weighted scores shows consistent modeling performance across minority (underfill, overfill) and majority (correct fill) classes, these findings indicate no class imbalance in the results. Second, stratified k-fold cross-validation was used to maintain class distribution across folds and ensure steady performance estimates that support generalization capability. Finally, the Random Forest algorithm's Out-of -bag (OOB) error validation, provides performance metrics closely aligned with testing accuracy. The convergence of these validation results aligns with recent literature that emphasizes the agreement between k-fold and Out-of -bag (OOB) error estimates are an important indicator for model reliability in industrial applications (Breiman, 2013; Wang et al. 2021).

The integration of the developed Random Forest classification model into the National Alcohol and Liquor Factory (NALF) filling machinery can be viewed as shifting towards real-time intelligent process control. The concept of integrating machine learning into manufacturing is aligned with the values of the Industry 4.0 and cyber-physical production system, in which data-driven models continuously engage in physical processes to provide adaptive decision-making (Lee et al., 2015).

Machine learning techniques have been extensively known in manufacturing systems because of ability to model complex, nonlinear relationships between process variables and quality outcomes. Wuest et al. (2016) point to the fact that the ensemble learning techniques, including Random Forest are especially adapted to the industrial setting because they are robust, interpretability through feature importance, and are resilient to noisy production data. This capability is useful in the NALF filling process because it is used to measure critical parameters such as vacuum pressure and inlet pressure, which provide precise control without making repeated adjustments.

Moreover, predictive analytics combined with production systems will improve operational efficiency by converting the quality control to a proactive one rather than a reactive one. Zhang et al. (2017) have shown that through the application of data-driven architectures, process stability and wastes are minimized when real-time analytics are connected to manufacturing execution systems. On the same note, Zonta et al. (2020) explain that predictive and intelligent systems minimize operational variability and equipment degradation when incorporated into industrial control loops.

In this study, the validated classification model acts as a decision engine in the filling process and allows real-time classification of fill conditions, trend identification, and confidence-driven intervention. The system allows connecting actuator adjustments (vacuum pressure, inlet pressure, flow rate, and valve timing) to prediction outputs, which creates a closed-loop control structure in accordance with the principles of cyber-physical manufacturing (Lee et al., 2015) . this integration does not only reduce overfilling and underfilling but also reduce reprocesses, utility consumptions and mechanical wear, thereby making processes in better efficiency and sustainability.

4.7 Empirical and Theoretical contribution of the study

It is a practical study based on evidence of enhancing bottle filling efficiency in an actual industrial setting in the National Alcohol and Liquor Factory (NALF). It determines the important process parameters that influence filling accuracy in terms of the vacuum pressure, inlet pressure, and timing of the opening of the valve with real production and operational data. The study also shows that filling outcomes can be accurately classified by a supervised machine learning model (Random Forest) and be utilized to make real-time decisions. Moreover, a practical solution exists: the incorporation of the model into the filling process would help to decrease reprocessing and waste production and increase efficiency in the production process. These results can be directly implemented in the manufacturing industries of similar beverages.

The study is also relevant to the current literature and advances the literature on the combination of process engineering and machine learning approach for filling process optimization. It builds upon the use of supervised machine learning in industrial process control by combining statistical analysis (distribution and correlation) with feature selection and model building. Another important feature of the study is the introduction of a structured framework that integrates the root cause analysis, parameter analysis, and machine learning to enhance the accuracy of filling. Moreover, it sheds light on how the importance of models and engineered variables can augment model performance and generalization in manufacturing systems.

CHAPTER FIVE

CONCLUSION, RECOMMENDATION AND FURTHER STUDY

5.1 Conclusion

The current study was carried out to maximize the efficiency of bottle filling at the National Alcohol and Liquor Factory (NALF) through a machine learning-based method aimed at solving the issue of constant filling variances. The evaluation of the existing filling system showed that the total filling accuracy was 95.9, which is lower than the acceptable industrial standard of 98. Moreover, the ratio of underfilling (2.358%) and overfilling (1.693%) was greater than the acceptable recycle limit of 2% which suggests that there were inefficiencies in the current filling procedure and that this should be improved.

This study has determined the major root causes of filling inaccuracies through the 5-Why analysis and the cause-and-effect (fishbone) analysis. The significant causes were divided into process parameters, pressure conditions, equipment-related factors, operational methods, and product properties. Among them, changes in the vacuum pressure, inlet pressure and valve timing were identified as the most significant factors contributing to deviations during filling.

Additional analysis through distribution and correlation method revealed that the process parameters do not exhibit the same trend in different filling situations. The results of the correlation analysis indicated that the variables of vacuum pressure, inlet pressure, cycle time, fill time, and valve timing are strongly associated with fill volume. Conversely, density and alcohol by volume (ABV) had a negligible effect and were not included in model development. The importance analysis of features also proved that the pressure and valve-related parameters are the most powerful factors influencing filling accuracy.

The supervised machine learning model (Random Forest classifier) was created to categorize the results of the filling into underfill, correct fill, and overfill. The model was performing very well in terms of training and testing accuracy of 99.61 and 99.33 respectively, which indicates that the model has high learning capacity and generalization. Evaluation measures such as precision, recall and F1-score had values that were uniformly high across all classes. The confusion matrix ensured that there were minimal misclassifications, and underfill and overfill conditions were well detected.

The validation of the models was made by several methods, such as Out-of-Bag (OOB) validation, 5-fold cross-validation stratified, and training and testing validation. The value of OOB accuracy (99.25) and the cross-validation mean score (99.11) with very low standard deviation proved that the model is stable, reliable, and is not influenced by overfitting.

Lastly, the study revealed that it was practical to integrated the developed model into the filling process with the help of three layers based on the architecture data acquisition layer, machine learning processing layer, and decision-making layer. This integration makes it possible to monitor and control the filling process in real-time and dynamically and identify and address filling deviations in time.

Altogether, the results suggest that the use of machine learning during the filling process was help to reach a higher level of filling accuracy, less rework, less waste of resources, and increase the efficiency of the operations. The study offers a feasible and scalable solution to enhance the efficiency with industrial filling and contributes to the shift towards smart manufacturing systems.

5.2 Recommendation

According to the results of this study, the following recommendations where made:

1. The machine learning model developed must be implemented in an actual production process where it can constantly provide predictive monitoring and allow early identification of the deviation of filling.
2. More attention should be paid to stabilizing vacuum pressure, inlet pressure, and valve timing parameters by increasing calibration and monitoring.
3. There should be a predictive maintenance approach with nozzle wear measurements along with valve performance measurements to minimize mechanical-induced variability and downtime.
4. The constant data acquisition and regular retraining of models should be established to provide flexibility to the equipment aging, environmental changes, and product variation.

5.3 Future research

Although this study has successfully developed and integrated a machine learning-based classification model for enhanced filling accuracy, several issues have been identified. There are several areas still open for further study and refinement.

1. Future studies may extend the classification approach to regression-based prediction for direct fill volume variations instead of categorical classification.
2. Other studies may also explore the use of advanced machine learning methods such as gradient boosting, deep learning architectures, or hybrid models that combine statistical and physical based approaches.
3. Future studies should also consider multi-product and multi-line validation to evaluate the general applicability of the proposed framework across various bottle sizes, viscosity ranges, and product formulations.

Reference

Abubakar, A. N., Kiangala, K. S. , & Basso, F. (2022). *Auto parameter configuration of a bottling process* . *10*(11), 2265–2273.

Ahmad, A., & Yusoff, R. (2022). Feed pressure adjustment for accurate distilled spirit filling. . *Journal of Food Engineering,* , *315*, 110804.

Ahmed, M., Khan, S., & Kim, J. (2023). Machine learning-based fault detection in industrial processes: A comparative study. . *Computers & Industrial Engineering, 176*, 108953.

Ammar, M., El-Afandy, M., & El-Dakhakhni, W. (2019). Process control and automation in beverage filling systems. *A Critical Review*. International Journal of Food Engineering,* , *15*(4), 317–326.

Anderson, J. , Kim, S. , & Lee, H. (2021). Impact of Vacuum Pressure on Fluid Dynamics in Liquid Filling Processes. *Journal of Fluid Mechanics*, *711*, 123–145.

Breiman, L. (2013). Random forests. *Machine Learning*, *45*(1), 5–32.

Brown, J. , Davis, K. , & Thompson, R. (2021). Sealing Mechanisms in Alcohol Packaging: . *A Review. Packaging Technology and Science*, *34*(3), 1–15.

Bui, H., & Wang, X. (2020). Real-Time Monitoring and Machine Learning-Based Correction of Process Deviations in Beverage Manufacturing. *Journal of Food Engineering*, *282*.

Chakraborty, S. , Konar, D. , Bhattacharyya, S. , & Chattopadhyay, S. (2020). A survey on synthetic data generation for machine learning applications. *International Journal of Automation and Computing, 17*(3), 321–338.

Chen, F. , Zhang, L. , & Wang, T. (2019). Effects of Vacuum Conditions on Filling Accuracy in Automated Systems. *International Journal of Engineering Science,* , *132*, 237–245.

Cheng, L., & Li, H. (2020). Effect of filling head pressure on beverage filling accuracy. . *International Journal of Beverage Technology, 15*(2), 45–56.

Chen, L. , & Zhang, W. (2020). Deep vision systems for precision fill-level inspection in beverage production. *Journal of Food Engineering, 285*.

Chen, Q., Wang, C., Wang, T., Lei, B., Wang, J., & Guo, J. (2025). Impact of surface tension, viscosity, pump settings, and nozzle size on filling process capability and accuracy in high-concentration biopharmaceuticals. *Journal of Pharmaceutical Sciences*, *114*(6), 1003565.

Chen, T., & Guestrin, C. (2016). *XGBoost: A Scalable Tree Boosting System*. https://doi.org/10.1145/2939672.2939785

Chen, Y., Wang, L., & Zhou, X. (2021). Analysis and optimization of liquid filling processes using process data analytics. *Journal of Manufacturing Processes,* , *64*, 123–134.

Chen, Y., Zhang, H., & Liu, X. (2020). Multivariate quality analysis of beverage filling processes using random forest algorithms. *Journal of Manufacturing Processes, 56*, 1092–1103.

Chevtchenko, S. F., Rocha, E. D. S., Santos, M. C. M. Dos, Mota, R. L., Vieira, D. M., De Andrade, E. C., & De Araujo, D. R. B. (2023). Anomaly Detection in Industrial Machinery Using IoT Devices and Machine Learning: A Systematic Mapping. *IEEE Access*, *11*, 128288–128305. https://doi.org/10.1109/ACCESS.2023.3333242

Choudhary, A. , Harding, J. A. , & Tiwari, M. K. (2021). Data mining in manufacturing: a review based on the kind of knowledge. *Journal of Intelligent Manufacturing*, *32*, 145–162.

Cortez, P. , Cerdeira, A. , Almeida, F. , Matos, T. , & Reis, J. (2009). Modeling wine preferences by data mining from physicochemical properties. *Decision Support Systems*, *47*(4), 547-553Cortez, P.,.

Ding, H. , Xu, J. , & Li, Y. (2021). Carbonation control and its impact on filling accuracy in soft drink production. *Beverage Industry Science,* , *13*(1), 22–31.

Doukas, D. , Koulouriotis, D. , & Xanthopoulos, A. (2019). Monte Carlo simulation-based analysis in food engineering and risk assessment. . *Journal of Food Engineering*, *240*, 65–72.

Dubey, R. , G. A. , Childe, S. J. , Papadopoulos, T. , & Helo, P. (2019). Supplier relationship management for circular economy: Influence of external pressures and top management commitment. *Journal of Cleaner Production*, *237*, 117782–117793. https://doi.org/https://doi.org/10.1016/j.jclepro.2019.117782

Frid-Adar, M. , Klang, E. , Amitai, M. , Goldberger, J. , & Greenspan, H. (2018). Synthetic data augmentation using GAN for improved liver lesion classification. . *Neurocomputing, 321*, 321–331.

Gao, F., Chen, X., & Li, Y. (2022). Deep learning-based predictive maintenance and quality classification in machining processes. *The International Journal of Advanced Manufacturing Technology.* , *110*(3–4), 567–580.

Gao, L. , Zhang, Y. , & Chen, M. (2024). Sensor fusion-based deep learning for real-time process optimization in manufacturing. . *International Journal of Computational Intelligence Systems*, *17*(2), 1–15.

Garcia, R. , & Tran, P. (2021). Automation and Control in Liquid Filling Systems: The Role of Vacuum Management. *Journal of Manufacturing Processes*, *58*, 512–525.

Goodfellow, I. , Pouget-Abadie, J. , Mirza, M. , Xu, B. , Warde-Farley, D. , Ozair, S. , & Bengio, Y. (2014). *Generative adversarial nets. Advances in Neural Information Processing Systems (NeurIPS), . 27*, 2672–2680.

Grzelak, M., Borucka, A., & Guzanek, P. (2022). Application of Linear Regression for Evaluation of Production Processes Effectiveness. *Lecture Notes in Mechanical Engineering*, 36–47. https://doi.org/10.1007/978-3-030-78170-5_4

Hastie, T. , Tibshirani, R. , & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction* ((2nd ed.)). Springer .

He, H., & Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering,* , *21*(9), 1263–1284.

Ibrahim, A. , Schmidt, K. , & Watanabe, H. (2023). Real-time computer vision for fill-level monitoring in low-resource bottling plants. *Journal of Food Engineering*, *347*, 111492–111505.

Jia, F., Zhang, T., & Chen, L. (2020). Quality prediction and control using machine learning methods in manufacturing systems. *Journal of Manufacturing Systems*, *56*, 1–15.

Jin, Y., Shi, J., & Ding, X. (2023). Machine learning for real-time industrial quality monitoring and control. *IEEE Transactions on Industrial Informatics.*, *19*(2), 1123–1132.

Johnson, R., Smith, K. , & Lee, H. (2018). IoT-enabled real-time viscosity compensation in carbonated beverage filling. *Journal of Manufacturing Systems*, 112–125.

Johnson, T. , Smith, R. , & Lee, J. (2023). Vacuum Pressure Management in Food and Beverage Filling Applications. *Food Engineering Journal*, *87*(3), 455–467.

Kiangala, K. S., & Wang, Z. (2018). Artificial neural network-based modeling and optimization of an automated bottling plant. *IEEE*, *6*, 65837–65847.

Kingma, D. P. , & Welling, M. (2013). *Auto-encoding variational Bayes. arXiv preprint* .

Kumar, A., & Kumar, R. (2021). *Machine learning applications in food and beverage production optimization:A quality and waste reduction perspective. . 14*(8), 1425–1443.

Kumar, A., Shankar, R., & Thakur, L. S. (2019). A big data-driven framework for quality prediction and control in manufacturing systems. *Computers & Industrial Engineering.*, *128*, 109–120.

Kumar, R., & Singh, S. (2021). Statistical analysis of volumetric filling systems in beverage packaging. *International Journal of Advanced Manufacturing Technology,* , *115*, 2457–2480.

Lee, J. , Bagheri, B. , & Kao, H. (2015). A cyber-physical systems architecture for industry 4.0-based manufacturing systems. . *Journal of Manufacturing* , *3*(1), 18–23.

Lee, M. , Brown, J. , & Patel, R. (2022). Impacts of Vacuum Pressure on Filling Performance: . *A Review. Chemical Engineering and Processing: Process Intensification,* , *178*.

Liakos, K. G. , Busato, P. , Moshou, D. , Pearson, S. , & Bochtis, D. (2018). Machine learning in agriculture. *A Review. Sensors,* , *18*(8).

Li, H., Zhao, Y., & Sun, J. (2022). Influence of viscosity and pressure on flow behavior in automated filling systems. . *Chemical Engineering Research and Design, 180*, 35–46.

Lillo Otarola, L., de la Fuente-Mella, H., Peña Domarchi, A., Kundu, A., & Ceroni-Díaz, J. (2025). Optimal scheduling of the wine-bottling process: A multi-dependency model with hydraulic considerations. *Applied Sciences, 15*(9), 4685–4697.

Liu, Q. , Zhang, J. , & Chen, X. (2020). Neural network modeling for accurate beverage filling system design. *Journal of Intelligent Manufacturing*, *31*(2), 455–467.

Liu, Y. , dab, w, & lous, B. (2020). Application of Neural Networks in Carbonated Beverage Filling Process. *Computers and Chemical Engineering,* , *142*.

Li, X. , Wang, Y. , & Xu, H. (2021). Machine learning techniques for beverage quality prediction: A comparative study. *Journal of Food Engineering*, *292*.

Li, Y. , Zhang, W. , & Yang, X. (2022). Predictive maintenance and quality control in beverage filling using ML techniques. *Journal of Manufacturing Processes*, *76*, 413–425.

Li, Z. , Zhang, J. , & Ren, L. (2022). Neural Network Models for Packaging Line Defect Prediction. *IEEE Transactions on Industrial Informatics*, *18*(5), 3112–3120.

Macal, C. M. , & North, M. J. (2010). Tutorial on agent-based modelling and simulation. . *Journal of Simulation*, *4*(3), 151–162.

Miah, M. A. , Rahman, M. , & Chowdhury, T. (2021). Application of machine learning algorithms in filling line quality prediction. *Computers and Industrial Engineering*.

Miah, M. R. , Rahman, A. , & Jahan, N. (2021). Intelligent Classification Model for Bottling Accuracy Using Machine Learning. *Procedia Manufacturing*, *53*, 311–318.

Miah, R. , Islam, M. , & Rahman, M. (2021). Adaptive volumetric control in beverage filling using supervised regression models. *Journal of Food Processing Engineering*, *44*(1).

Mkumbo, A. (2023). Machine learning implementation in African beverage production. *International Journal of Production Economics, 255*(108215).

Montgomery, D. (2009). *Statistical quality control in modern manufacturing.* (7th ed).

Muchiri, P. , & Pintelon, L. (2018). Performance measurement using Overall Equipment Effectiveness (OEE): Literature review and practical application discussion. . *International Journal of Production Research*, *46*(13), 3517–3535.

Muthoni, L. (2023). Infrastructure limitations to Industry 4.0 adoption in Kenyan manufacturing. *Manufacturing Review*, *10*(5).

Negahban, A. , & Smith, J. S. (2014). Simulation for manufacturing system design and operation: Literature review and analysis. . *Journal of Manufacturing Systems,* , *33*(2), 241–261.

Okafor, P., Eze, C. , & Musa, A. (2021). Adapting computer vision systems for variable-condition African factories. *IEEE Transactions on Industrial Informatics, 17*(8), 5432–5441.

Özkan, B., & Şimşek, M. (2019). Flow and pressure compensation in carbonated beverage filling systems. *Journal Food Engineering*, *72*(12), 895–910.

Patel, R. , & Raghavan, R. (2019). The Importance of Filling Accuracy in Pharmaceutical Manufacturing. . *Journal of Pharmaceutical Sciences,* , *108*(10), 3074–3083.

Patki, N. , Wedge, R. , & Veeramachaneni, K. (2016). The synthetic data vault. . *IEEE International Conference on Data Science and Advanced Analytics (DSAA),* , 399–410.

Powers, D. M. W. (2011). Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation. *Journal of Machine Learning Technologies, 2*(1), 37–63.

Rahman, S. M. , Bung, G., & lius.M. (2021). Optimizing Liquid Filling Systems Using Support Vector Machines. *International Journal of Advanced Manufacturing Technology*, *114*, 1023–1035.

Ramos, L. ,. (2022). Filling process optimization through modifications in machine settings. *Journal of Material Process*, *10*(11), 2263–2273.

Rossi, M., & Mancini, F. (2023). Machine learning–based monitoring of industrial filling processes. . *Computers & Industrial Engineering,* , *176*, 108995.

Salah, D. , Bendaoud, M. , & Derrouiche, R. (2022). Process optimization in Industry 4.0-based yogurt filling line. *Transitioning from Multiple to Single-Point Control. *Processes**, *10*(11).

Salem, A. , & El-Banna, A. (2023). low rate control in beer and malt beverage filling using PID tuning. *International Journal of Beverage Engineering, 18*(1), 23–35.

Singh, P., Sharma, R., & Gupta, N. (2011). Foam detection-based flow control in beverage filling systems. . *Journal of Food Science and Technology,* , *58*(12), 4662–4673.

Sokolova, M. , & Lapalme, G. (2009). Systematic Analysis of Performance Measures for Classification Tasks. . *Information Processing & Management.*

Song, Z., Zeng, M., Li, Z., Ma, X., & Li, Z. (2024). Numerical investigation of inlet-pipe configurations on fluid dynamics and flow-field characteristics inside filling valve. . *Flow Measurement and Instrumentation*, *76*, 102612.

Sullivan, R. , Zhao, Z. Y. , & Boyd, M. (2012). A review of control strategies for process optimization in beverage filling lines. *Journal of Manufacturing Systems*, *31*(3), 309–318.

Tadesse, H., Singh, B., Deresso, H., Lemma, S., Singh, G. K., Srivastava, A. K., Dogra, N., & Kumar, A. (2024). Investigation of production bottlenecks and productivity analysis in soft drink industry: a case study of East Africa Bottling Share Company. *International Journal on Interactive Design and Manufacturing*. https://doi.org/10.1007/s12008-023-01715-9

Tao, Y. , Li, J. , & Wang, S. (2022). Level control in carbonated alcoholic beverage filling lines using PI/PID controllers. *Journal of Process Control*, *112*, 50–62.

Tesfaye, B. , & Tadesse. (2021). Industry 4.0 adoption challenges in developing economies: A case study from Ethiopian manufacturing. . *African Journal of Science, Technology, Innovation and Development*, *13*(7), 829–839.

Tzeng, Y.-C., Chen, W.-H., & Lin, C.-Y. (2019). Predictive maintenance algorithms for bottling plant efficiency: A machine learning approach. *Journal of Manufacturing Systems,* , *52*, 231–245.

Wang, J., Ma, Y., Zhang, L., Gao, R. X., & Wu, D. (2021). Deep learning for smart manufacturing: Methods and applications. *Journal of Manufacturing Systems*, *48*, 144–156.

Wang, Y. , Huang, S. , & Zhang, L. (2018). Regulatory compliance in soft drink packaging: A case-based analysis. *Journal of Food Engineering and Safety,* , *12*(3), 198–206.

White, F. M. (2019). *Fluid mechanics* (9th ed.). McGraw-Hill.

Wuest, T., Weimer, D., Irgens, C., & Thoben, K. D. (2016). Machine learning in manufacturing: Advantages, challenges, and applications. *Production & Manufacturing Journal*, *4*(1), 23–45.

Xie, X. , Yu, Y. , & Li, J. (2022). Variational autoencoder-based synthetic data generation for predictive maintenance in manufacturing. . *Journal of Intelligent Manufacturing*, *33*(6), 1601–1616.

Xu, J., Liu, Q., & Zhang, W. (2023). Synthetic data generation for industrial machine learning applications. *Journal of Smart Manufacturing* , *11*, 45612–45623.

Yang, L. , & Lee, J. (2020). Air Removal Techniques in Liquid Filling: Vacuum Pressure Implications. *International Journal of Food Science & Technology*, *55*(1), 45–52.

Zhang, H., & Cheng, R. (2020). Overfitting prevention and validation strategies for industrial machine learning models. *Computers & Chemical Engineering.*, *138*, 106835.

Zhang, H., & Li, Wei. (2022). Viscosity variation studies in beverage production. *Food Engineering Journal*, *45*(3), 210–225.

Zhang, H. , Xu, J. , & Li, Y. (2021). Sensitivity analysis in manufacturing systems using statistical simulation. *Computers & Industrial Engineering,* , *151*.

Zhang, L., Wu, H., Schmidt, K., & Chen, Xiaoping. (2023). Machine learning algorithms for fill accuracy in bottling operations. *Nature Machine Intelligence*, *5*, 456–468. https://doi.org/https://doi.org/10.xxxx/s42256-023-00645-1

Zhang, T. , Li, M. , & Zhao, Y. (2020). Economic implications of overfilling in automated bottling lines. *Food Industry Economics*, *8*(2), 75–83.

Zhang, Y., Liu, Q., & Chen, D. (2021). Prediction of liquid filling accuracy using support vector regression in industrial bottling systems. *Journal of Process Control*, *98*, 30–41.

Zhang, Y., Ren, S., Liu, Y., & Si, S. (2017). A big data analytics architecture for cleaner manufacturing and maintenance processes of complex products. *Journal of Cleaner Production*, *142*, 626–641.

Zhang, Y., Wang, S. , & J. G., & Ji, G. (2019). A comprehensive survey on ensemble learning: Boosting, bagging, and random forests. *IEEE Access*, *7*, 10965–10978.

Zhao, Y., Jin, N., & Gao, Z. (2021). Performance evaluation metrics for multi-class classification in manufacturing quality prediction. *Journal of Intelligent Manufacturing.*, *32*(6), 1589–1602.

Zhou, J. , Li, X. , & Peng, Y. (2021). LearningADD: Acoustic-based defect detection in beverage bottles using edge AI. . *Sensors and Actuators A: Physical*.

Zonta, T., da Costa, C. A., da Rosa Righi, R., de Lima, M. J., da Trindade, E. S., & Li, G. P. (2020). Predictive maintenance in the Industry 4.0: A systematic literature review. *Computers & Industrial Engineering*, *150*, 106889.

Appendix I-Data Collection tool

ADDIS ABABA UNIVERSITY

COLLEGE OF TECHNOLOGY AND BUILT ENVIRONMENT (CTBE)

SCHOOL OF MECHANICAL AND INDUSTRIAL ENGINEERING (SMIE)

INDUSTRIAL ENGINEERING STREAM

I am a Master's student at **Addis Ababa University, College of Technology and Built Environment (CTBE), School of Mechanical and Industrial Engineering (SMIE), Industrial Engineering Stream**. This questionnaire is part of a research study entitled**: "Optimizing Bottle Filling Efficiency with Machine Learning Approach to Overcome Filling Challenges: A Case of National Alcohol and Liquor Factors (NALF)",** conducted as part of the requirements for the fulfillment of a Master's degree in Industrial Engineering at Addis Ababa University, College of Technology and Built Environment (CTBE), School of Mechanical and Industrial Engineering.

Purpose of this checklist:

The purpose of these checklist is to collect relevant information that helps identify operational challenges in the bottle filling process and explore improvement opportunities using data-driven and machine learning approaches. The data collected will be used solely for academic purposes. Your responses will remain strictly confidential, and your identity will not be disclosed.

Your cooperation is highly appreciated and will contribute significantly to the success of this study.

With sincere thanks and pleasure,

Jemera Melaku

For any questions or further information, please feel free to contact me at:

phone; 0710944232 / 0928594093 Email: -jemeramelaku23@gmail.com / jemera.melaku@aait.edu.et

Part I: - Bottle Filling Quality Control & Diagnostic Checklist

	Bottle Filling Quality Control & Diagnostic Checklist				
Parameter Group	**Parameter / Feature**	**Max value (At normal conditions)**	**Tolerance (At normal Condition)**	**Observed (Avg/Range)**	**Deviation Status (OK/Issue)**
Flow and valve control Parameters	Flow rate(L/s)				☐ OK ☐ Issue
	Valve open time(s)				☐ OK ☐ Issue
	Shut off Delay (s)				☐ OK ☐ Issue
	Valve response delay(s)				☐ OK ☐ Issue
Pressure related parameters	Inlet Pressure (bar)				☐ OK ☐ Issue
	Vacuum Pressure (mbar)				☐ OK ☐ Issue
Product Properties	Product Temperature (°C)				☐ OK ☐ Issue
	Alcohol by Volume (Abv, %)				☐ OK ☐ Issue
	Density (g/ml)				☐ OK ☐ Issue
	Viscosity(mpas)				☐ OK ☐ Issue
Mechanical and operational **parameters**	Cycle time				☐ OK ☐ Issue
	Nozzle Wear				☐ OK ☐ Issue

Part II. **Defect tracking checklist**					
Defect tracking checklist (shift number--------)					
Date	**Shift**	**Line No.**	**Defect Type**	**Count (This Shift)**	**Root Cause (Observed)**
		1	Underfill		

		1	Overfill		
		1	Underfill		
		1	Overfill		
		1	Underfill		
		1	Overfill		

Part III- Production Data (Week---)			
Shift	Total produced Bottles	Underfilled Bottles	Overfilled Bottles
1			
2			
3			
4			
5			
6			

Part IV - Historical Sensor Data Collection Form

Date: ____________________

Shift: ____________________

Time	Product Temperature (°C)	Inlet Pressure (bar)	Vacuum (mbar)	Flow rate (L/s)	Valve Timing (s)	Filling Time (s)	Valve delay	Comments Notes
1								
2								
3								
4								
5								

6								

Part V- Structured Interview

Parameter	When Overfilling Happens (Quantitative Range)	When Underfilling Happens (Quantitative Range)
Inlet Product Pressure	What inlet pressure range (minimum–maximum) has been observed during overfilling events?	What inlet pressure range (minimum–maximum) has been observed during underfilling events?
Flow Rate	What flow rate range has been observed when overfilling occurs?	What flow rate range has been observed when underfilling occurs?
Vacuum Pressure	What vacuum pressure range has been observed during overfilling conditions?	What vacuum pressure range has been observed during underfilling conditions?
Valve Open Time	What valve open time range has been observed during overfilling events?	What valve open time range has been observed during underfilling events?
Valve Closing Delay	What valve closing delay range has been observed during overfilling?	What valve closing delay range has been observed during underfilling?
Valve Delay Ratio	What valve delay ratio range has been observed during overfilling conditions?	What valve delay ratio range has been observed during underfilling conditions?
Shut-off Delay	What shut-off delay range has been observed during overfilling?	What shut-off delay range has been observed during underfilling?
Filling Time	What filling time range has been observed during overfilling events?	What filling time range has been observed during underfilling events?
Cycle Time	What cycle time range has been observed during overfilling conditions?	What cycle time range has been observed during underfilling conditions?
Product Temperature	What product temperature range has been observed during overfilling?	What product temperature range has been observed during underfilling?
Product Density	What density range has been observed during overfilling conditions?	What density range has been observed during underfilling conditions?
Product Viscosity	What viscosity range has been observed during overfilling?	What viscosity range has been observed during underfilling?
Alcohol by Volume (ABV)	What ABV range has been observed during overfilling conditions?	What ABV range has been observed during underfilling conditions?
Nozzle Wear Condition	What measurable nozzle	What measurable nozzle

	wear level (e.g., leakage rate or wear stage) has been observed during overfilling?	wear level has been observed during underfilling?
Sensor / Valve Response Delay	What response delay range (ms or s) has been observed during overfilling events?	What response delay range (ms or s) has been observed during underfilling events?

Part VI- Equipment Condition Data Collection Tool

Sheet 1: Inspection Checklist

Parameter	Observations/Measurements	Pass/Fail	Comments
Nozzle Wear Condition			
Visual Inspection	[ ] No wear [ ] Minor wear [ ] Severe wear	[ ] Pass [ ] Fail	
Dimensional Check	Measured Diameter: ____ mm	[ ] Pass [ ] Fail	
Valve Response Condition			
Response Time	Measured Time: ____ seconds	[ ] Pass [ ] Fail	
Operational Check	[ ] Normal [ ] Delayed response		
Leakage Observations			
Leakage Observed	[ ] Yes [ ] No		Specific Location: ____
Shut-off Delay Observations	Measured Delay: ____ seconds	[ ] Pass [ ] Fail	

Part VII- Process Observation Data Collection Tool

Table A.: Process Observation Checklist

Date	Shift	Parameter Observed	Observation Details	Action Taken
	1	Abnormal Filling Behavior		
	2	Operational Deviation		
	3	Environmental Issue		
	4	Abnormal Filling Behavior		
	5	Operational Deviation		
	6	Environmental Issue		
	7	Abnormal Filling Behavior		

Appendix II - Nominal tolerance under normal conditions of key parameters.

Parameter Name	Unit	Nominal tolerance under normal conditions
Valve Open Time	S	12.5- 15.5
Fill Time	S	12 - 15
Inlet Pressure	Bar	0.2 - 0.45
Flow Rate	L/s	3.0 - 3.25
Valve Delay	S	0.005 - 0.45
Shut-off Delay	S	0.01- 0.8
Product Temperature	°C	18 - 22
Vacuum Pressure	Mbar	-100 - (-10)
Cycle Time	S	19.5 - 20.5
Nozzle Wear	%	2% - 6%
Density	g/ml	0.945 - 0.949
Viscosity	mPa·s	1.6 - 2.2
Alcohol by Volume (ABV)	-	39.5 - 40.5

Source**:** Own, based on data from NALF (June-September 2025)

Appendix III - Summary of Average Parameter Ranges Contributing to Overfill and Underfill: Insights from Structured Interviews

Parameter Name	Unit	Overfill	Underfill
Valve Open Time	S	14.5-16.7	11-13.5
Fill Time	S	14-16	11-13
Inlet Pressure	Bar	0.4-0.8	0.0- 0.3
Flow Rate	L/s	3.1 -3.3	2.8- 3.0
Valve Delay	S	0- 0.8	0-0.6
Shut-off Delay	S	0.45-1	0.0-0.2
Product Temperature	°C	21- 23	17-19
Vacuum Pressure	Mbar	-130 –(- 90)	-20-0
Cycle Time	S	20.5 – 21.5	18.5-19.85
Nozzle Wear	%	2% - 8%	0-6%
Density	g/ml	0.945 - 0.949	0.945 - 0.95
Viscosity	mPa·s	1.4 - 1.8	1.9-2.5
Alcohol by Volume (ABV)	-	39.5 - 40.5	39.5 - 40.5

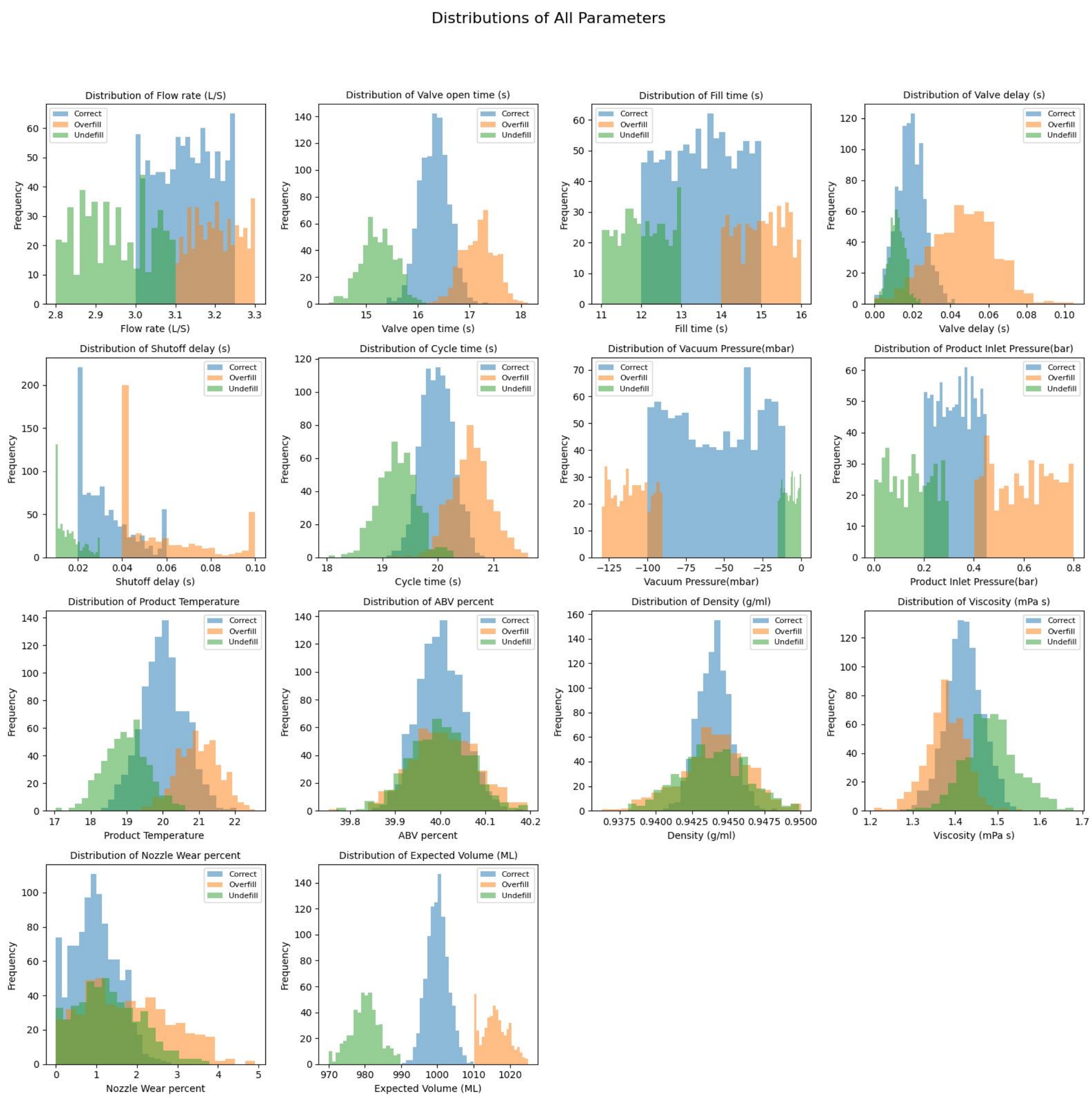
Source**:** Own, based on data from NALF (June-September 2025)

Appendix IV-Synthetic data descriptive statistics, distribution and correlation

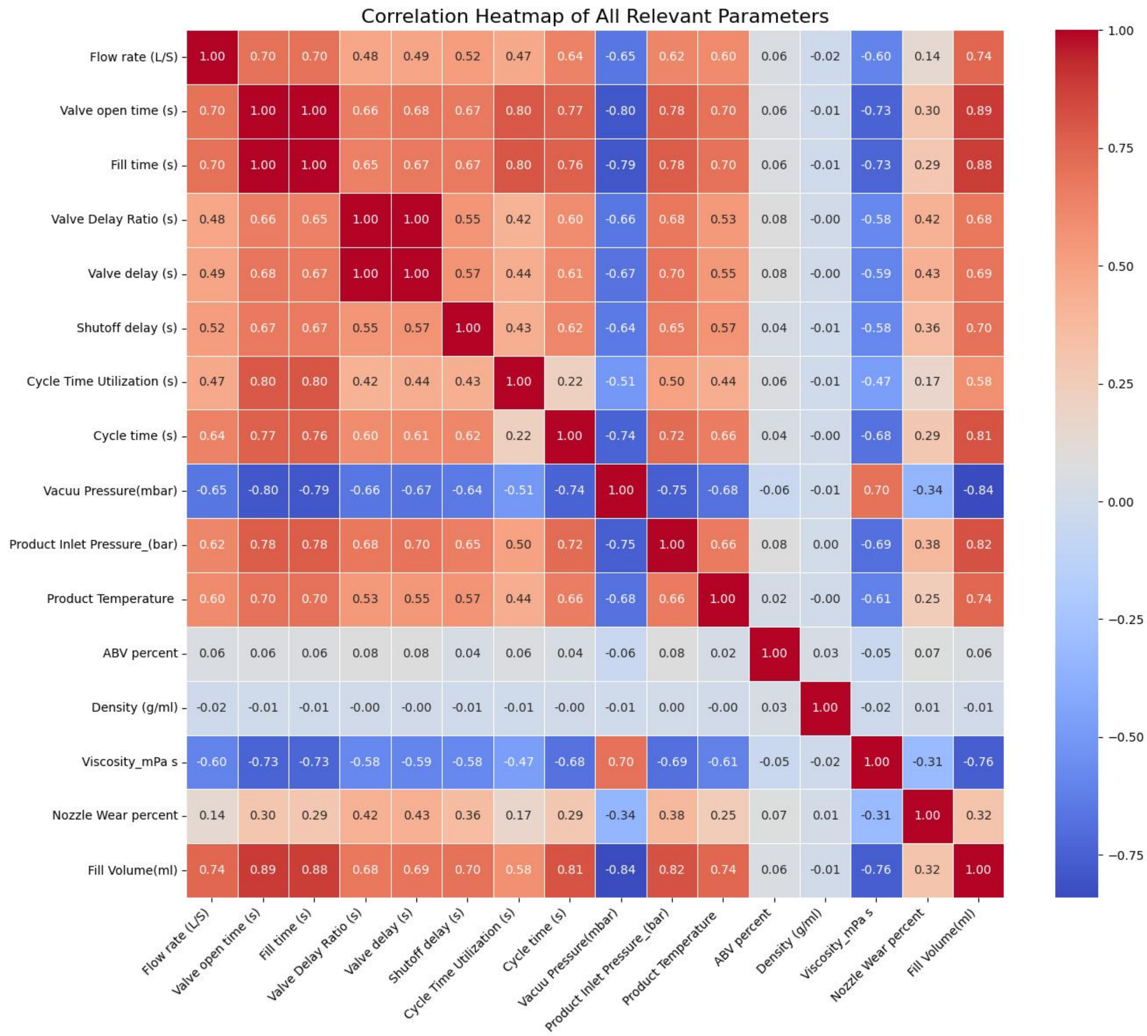
Table: descriptive statistics of Synthetic data

Descriptive Statistics					
Key parameters	N	Minimum	Maximum	Mean	Std. Deviation
Flow rate (L/S)	2000	2.73	3.40	3.1003	.11903
Valve open time(s)	2000	14.28	18.13	16.266	.75388
Fill time (s)	2000	11.00	15.99	13.517	1.28473
Valve delay (s)	2000	.03	1.23	.2575	.16815
Shutoff delay (s)	2000	.01	.10	.0345	.01992
Cycle time(s)	2000	18.03	22.50	19.967	.57389
Vacuum Pressure (mbar)	2000	-134.4	-.07	-56.85	41.54431
Inlet Pressure (bar)	2000	.00	.88	.3504	.18529
Product Temperature	2000	17.04	22.55	19.99	.92643
ABV%	2000	39.75	40.19	40.002	.06104
Density (g/ml)	2000	.94	.95	.9440	.00183
Viscosity (mPas)	2000	1.21	2.84	1.4408	.14099
Nozzle Wear%	2000	.00	4.73	1.2723	.84559
Valid N (listwise)	2000				

Source**:** Own, based on data from NALF (June-September 2025)



Source**:** Own, based on data from NALF (June-September 2025)



Source**:** Own, based on data from NALF (June-September 2025)