Analysis of the state and fault detection of a plastic injection machine

João Costa[1], Rui Silva[2], Gonçalo Martins[2], Jorge Barreiros[1,3], Mateus Mendes[1,3,4,*]

[1] Polytechnic University of Coimbra, Coimbra Institute of Engineering,
Rua Pedro Nunes-Quinta da Nora, Coimbra, 3030-199, Portugal
[2] Sinmetro LDA, Rua dos Costas, Lote 19, Loja 74, R/C 2415-567 Leiria
[3] RCM^{2+} Research Centre for Asset Management and Systems Engineering,
Rua Pedro Nunes, Coimbra, 3030-199, Portugal
[4] Institute of Systems and Robotics,
Department of Electrical and Computer Engineering, University of Coimbra,
3030-290 Coimbra, Portugal
[*] Corresponding author
{rui.silva,gmartins}@sinmetro.pt, {a2022143368,jmsousa,mmendes}@isec.pt

Abstract

Predictive maintenance is essential for minimizing unplanned downtime and optimizing industrial processes. In the case of plastic injection molding machines, failures that lead to downtime, slowing production or manufacturing defects, can cause large financial losses or even endanger people and property. As industrialization advances, proactive equipment management enhances cost efficiency, reliability, and operational continuity. This study aims to detect machine anomalies as early as possible, using sensors, data science techniques, statistical analysis and classification models. A case study was carried out, including machine characterization and data collection from various sources. Clustering methods identified operational patterns and anomalies, classifying the machine's behavior into distinct states. Applying unsupervised learning techniques, namely clustering method DBSCAN, we have developed a methodology for early detection of anomalies in machine operational data. Reducing dimensionality with PCA and identifying distinct operating patterns, we were able to effectively distinguish between normal and abnormal conditions. State classification was carried out using the resulting cluster data. The XGBoost achieved the best performance among the models tested, reaching an accuracy of 83%.

Keywords: Predictive maintenance; Plastic Injection Machine; Fault detection

1 Introduction

Unplanned downtime in industrial machinery poses a significant challenge to manufacturing efficiency. Unexpected machine failures can lead to severe financial losses, production delays, and increased maintenance costs. In plastic injection molding, where precision and continuity are crucial, predicting and preventing such failures is essential to maintain productivity and reduce operational risks. Failures in these machines can stem from various sources, including mechanical wear, electrical faults and human errors, making early detection a complex but necessary task.

Plastic Injection Molding (PIM) machines are heavy industrial equipment that require specialized maintenance interventions. Ideally they operate continuously during many hours, days or even weeks, in order to maximize production and minimize setup time. Nonetheless, they can suffer numerous problems that require fast qualified maintenance interventions. A number of possible problems and solutions are discussed below. The problems include fixed plate deformation, obstructions in the injection system, pressure and temperature variations, mold cooling failures, and other common challenges in this type of equipment.

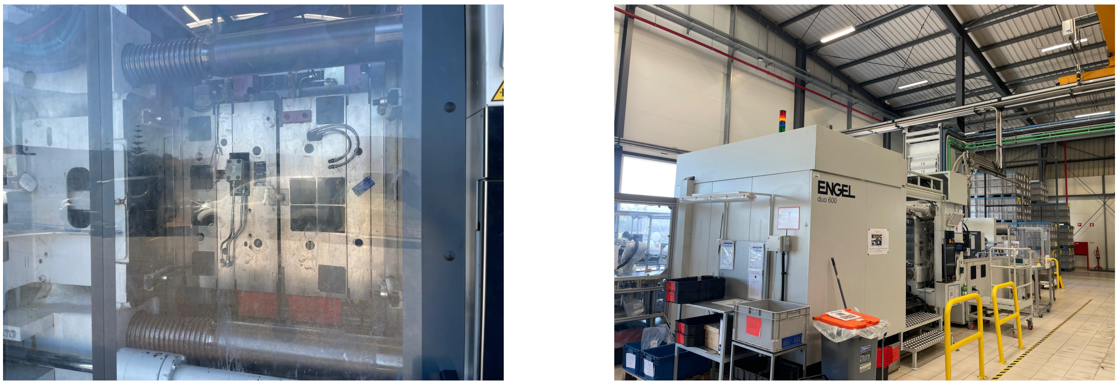
Recent advancements in data science and machine learning have enabled the development of predictive maintenance strategies aimed at reducing unplanned stoppages. Anomaly detection has shown promising results in identifying patterns associated with machine failures, allowing early interventions [1]. However, despite the growing body of research, many industrial applications still rely on reactive maintenance, leading to inefficiencies and high costs. This study aims to contribute to bridge this gap for this particular type of equipment, by applying data-driven clustering and classification methods to detect failures in a plastic injection machine at a very early stage.

The main objectives of this research are: (1) to analyze machine behavior through literature and dataset analysis; (2) to group and identify states through clustering techniques; and (3) to classify the data to be used later in fault detection in a real-world industrial context. The study follows a structured methodology, starting with data collection and machine characterization, followed by clustering analysis and classification.

The remainder of this paper is organized as follows. Section 2 describes the operating principles and main components of the plastic injection machine. Section 3 reviews the current state of the art. Section 4 outlines the data and methodology adopted in this study. Section 5 discusses the results, and Section 7 presents conclusions and future research directions.

2 Plastic Injection Machines

Plastic Injection Machines are complex equipment, heating the plastic and passing molten plastic through more or less complex molds for production. In the this table 1 will be able to see the components that make up the machine, its location and a brief description of what it does.



(a) Plastic Injection Machine Mold

(b) Plastic Injection Machine Structure

Figure 1: Plastic Injection Machine

Table 1: Name, location, function and appearance of the most important parts of a PIM

Name	Location	Function	Appearance
Barrel	In the middle of the machine, surrounds the screw	Captures and mixes the plastic, maintaining uniformity and pressure during injection	Long and cylindrical; pointed at one end; covered by heating bands
Gates	Between channels and mold cavity	Control the flow of plastic going into the cavity	Small openings
Heaters	Around the barrel	Heats the barrel to melt the plastic	Metal bands around the barrel
Hopper	At the top of the machine	The place where plastic is introduced in its solid form. Sometimes contains a dryer to remove moisture	Funnel (Conical Shape)
Mold	Connected to the Fixed and Movable Platens	Gives the final shape to the molten plastic, forming the desired part	Metal block, typically two parts, containing a cavity, cooling channels, and vents
Mold Cavity	Inside the mold	Creates the final shape of the mold and contains cooling cavities	A space in the mold that forms the desired shape
Movable Platen	Connected to one half of the mold	Presses one half of the mold against the other during part manufacturing and releases it once the part is finished and cooled	Flat, rectangular, and metallic
Nozzle	At the end of the barrel, near the mold	Directs the plastic into the mold cavity and prevents it from cooling before entering the mold	Tapered outlet
Pellets	Inside the barrel and nozzle	Plastic material inserted into the machine for molding. Common plastics include ABS, PP, or Nylon, sometimes with additives	Small plastic granules
Reciprocating Screw	Inside the barrel	Mixes and compresses the plastic coming from the hopper	A metal spiral
Runners	Inside the mold	Directs the molten plastic, maintains a uniform flow, and reduces plastic waste	Long and narrow channels
Sprue	Central channel that connects the nozzle to the entry point of the molten plastic	Directs the molten plastic from the nozzle to the runners	Cone-shaped channel

2.1 Steps in Injection Molding

Injection molding is a critical manufacturing process in the plastics industry, capable of producing high-volume, high-precision components for a wide range of applications. Here is a detailed description of the plastic injection molding process.

1. **Feeding and Preparation of Raw Material -** The process begins with loading plastic raw materials, usually in the form of small pellets or granules, into the machine's hopper. These thermoplastic materials, such as polypropylene (PP), polyethylene (PE), polystyrene (PS), polycarbonate (PC), or acrylonitrile butadiene styrene (ABS), can be mixed with additives such as colorants, UV stabilizers, or reinforcing agents to alter the appearance of the part. The hopper feeds the material into a heated cylinder by gravity.

2. **Plasticizing the Material -** Inside the cylinder, a reciprocating screw rotates and moves the plastic forward. As the material advances, it is gradually heated by both external electric heaters surrounding the barrel and the frictional heat generated by the shearing action of the screw. This combined heat melts the pellets into a homogeneous, viscous molten state called a pillow. At the front of the barrel, a check valve prevents the molten plastic from flowing backwards, ensuring that the full volume of material is injected forward when needed.

3. **Injection into the Mold -** Once a sufficient amount of molten plastic has accumulated in front of the screw (a process known as "shot size preparation"), the screw stops rotating and moves forward, acting as a plunger. It forces the molten plastic through the nozzle and into the mold cavity under high pressure. The mold, which is tightly closed under tons of pressure, contains the negative shape of the final part. The high injection pressure ensures that the molten plastic fills every detail of the cavity, including thin walls, small features and complex geometries.

4. **Cooling and Solidification -** Once the mold is filled, the cooling phase begins. The plastic begins to solidify when it comes into contact with the cooler walls of the mold. Most molds have an integrated cooling system, usually channels that circulate water or oil, to control and accelerate the cooling process. Cooling time is crucial and depends on the material, part thickness and mold design. Adequate cooling ensures dimensional stability and prevents problems such as warpage or sink marks.

5. **Mold Opening and Part Ejection -** After sufficient cooling time, the mold opens and ejector pins push the solidified part out of the cavity. In multi-cavity molds, several parts can be ejected simultaneously. Sometimes, robotic arms or conveyors assist in removing and organizing molded parts, especially in automated production lines. Once the part is ejected, the mold closes again and the next cycle begins. A complete injection molding cycle can take anywhere from a few seconds to a few minutes.

6. **Post-Molding Operations -** Although injection molding produces parts that are nearly finished in shape, some post-processing may be required. In this case, the piece is automatically inspected for defects, then transferred to an oven where it is hardened and strengthened before being stored.

2.2 PIM Problems and Failures

PIM are complex devices which require careful optimization and maintenance to operate smoothly and safely. According to industry operators who perform daily maintenance on injection molding machines, a number of issues are common [2].

One of the most common problems in PIM machines is the obstruction of the injection devices due to plastic residue accumulation or contaminants. This issue may arise from improper cleaning procedures, low-quality materials, or incorrect processing conditions. Additionally, incorrect material selection and improper processing parameters can lead to material degradation, overheating, and nozzle blockage. Regular maintenance and cleaning of injection devices are essential to prevent obstructions. Operators should use appropriate cleaning agents to remove accumulated residues. Using compatible materials and optimizing processing parameters also significantly reduces obstruction risks and enhances machine efficiency.

The Mold Cooling System is another sensitive part. Inefficient mold cooling can result from poor cooling channel design and inadequate heat dissipation capacity, leading to uneven temperature distribution, extended cycle times, higher defect risks, and reduced product quality. Optimizing the layout and configuration of cooling channels ensures uniform temperature distribution throughout the mold. Proper positioning, consistent channel diameters, and appropriate spacing improve cooling efficiency, leading to better productivity and higher-quality injected parts.

Pressure and Temperature Variations in Injection can also be a source of problems. Pressure variations in the injection process are influenced by changes in material viscosity and polymer flow behavior. Temperature fluctuations and humidity content can alter material flow properties, causing injection pressure instability. Controlling material properties through proper storage, handling, and humidity monitoring is critical. Injection parameters must be adjusted based on material type, product shape, and mold design. Proper pressure calibration minimizes variations and ensures consistent production quality.

Part Adhesion and Removal Issues are another typical problem. Part adhesion to the mold can result from inadequate use of mold release agents. Insufficient extraction force or poorly designed ejector pins may lead to deformation or incomplete removal of molded parts. Applying mold release agents correctly and optimizing surface finish reduce friction and facilitate part removal. Adjusting extraction force and modifying ejector pin design ensures efficient part ejection, minimizing defects and material waste.

The Hydraulic System can also cause frequent failures. Hydraulic failures, including oil leaks due to worn seals, damaged hoses, or loose connections, can significantly impact machine performance. Insufficient hydraulic pressure caused by pump failures, valve blockages, or fluid contamination also disrupts operation. Regular hydraulic system inspections prevent oil leaks and ensure all connections are secure. Monitoring pressure gauges and performing preventive maintenance on pumps, valves, and filters, help maintain optimal hydraulic performance and avoid machine failures.

The Electrical and Control System must also be monitored to prevent potential failures. Electrical issues, such as power fluctuations or wiring failures, can cause unexpected machine shutdowns and production delays. Malfunctions in the control system, including software or hardware issues, can impact machine performance and process regulation. Installing surge protectors, voltage regulators, and uninterruptible power supplies (UPS) stabilizes power supply and prevents sudden failures. Routine inspections of electrical wiring, terminals, and connectors, help prevent malfunctions. Diagnosing and resolving control system issues in advance ensures stable operation and high-quality production.

By addressing these issues through proper maintenance and optimization strategies, the efficiency, durability, and quality of plastic injection molding machines can be significantly improved.

3 Literature Review

A comprehensive literature review was conducted, searching scientific databases such as Scopus, IEEE Xplore, and ScienceDirect. Keywords used included "plastic injection molding failures," "Predictive Maintenance in plastic injection machine," "mold cooling efficiency," and "hydraulic system failures in injection molding machines." The papers were selected based on their relevance to fault identification, predictive maintenance, and operational optimization in plastic injection molding (PIM) machines.

3.1 Fault Detection

Zhang & Alexander [3] explore the use of pressure signals measured directly in the mold cavity as a source of information for identifying faults in the injection process. The study's main hypothesis is that different types of faults — such as incomplete filling, the presence of bubbles or burn marks — manifest themselves in a characteristic way in the cavity's pressure profile throughout the molding cycle. To process these signals, the authors apply Principal Component Analysis (PCA) as a dimensionality reduction technique, facilitating the extraction of relevant patterns. They then use wavelet transforms to decompose the signals and extract representative dynamic features. These features are then fed into an artificial neural network, which is responsible for classifying the different types of faults based on previously labeled examples. The study demonstrates that cavity pressure signals provide a robust basis for early fault diagnosis, with the potential to significantly improve the quality of the parts produced and the efficiency of the process. In addition, Zhang and Alexander emphasize the importance of approaches based on real process data for continuous and adaptive monitoring, anticipating current trends in predictive maintenance and intelligent manufacturing.

Kozjek *et al.* [4] explore the application of data mining techniques for fault diagnosis in PIM processes. The study examines how algorithms such as J48, Random Forest, and k-Nearest Neighbors can identify operational failure patterns, aiming to improve production efficiency and maintenance planning. This study demonstrates how Data Mining (DM) can uncover defective operational patterns and improve quality and productivity in PIM. DM provides an alternative to traditional methods like statistical process control and experimental design by discovering significant patterns and relationships in industrial data. The dataset consists of six months of operational records from five European PIM machines: Process parameter logs input and output values per cycle, Alarm logs, and Tool change records. Approximately 2.2 million cycles were recorded, with over ten different tools used per machine. Alarms were classified into First-degree alarms (immediate machine stoppage) or Second-degree alarms (non-critical issues). The final dataset contained 62 numerical attributes and a binary classification label. To ensure an unbiased classification baseline of 50%, the dataset included equal instances of normal and faulty cycles. A Python-based system was developed for data processing, including reading, encoding, filtering, transformation, and querying. The results indicate that the J48, Random Forest, JRip, Naïve Bayes, and k-NN algorithms effectively identify patterns related to defective operating conditions. All tested algorithms outperformed the standard accuracy benchmark of 50.0%. J48, Random Forest, JRip, and Naïve Bayes exhibited higher classification accuracy than k-NN, as they leverage target attribute information during model induction. Random Forest offers the advantage of easy parameter tuning and relatively high predictive performance. However, its interpretability is limited, whereas J48 and JRip provide interpretable rule-based models. Key parameters affecting defective operating conditions in Injection Molding Units (UMSs) for a selected tool and machine include temperature, opening time, and cycle time.

Gözde Aslantaş *et al.* [5] investigate the application of classification models to identify various fault types in plastic injection molding machines, utilizing sensor data collected from the equipment. The study focuses on predictive maintenance aimed at anticipating machine failures to optimize operational uptime and reduce maintenance costs. The dataset comprises continuous measurements of parameters such as vibration, temperature, pressure, and other operational characteristics, labeled according to distinct fault categories. To address the issue of class imbalance, the Synthetic Minority Over-sampling Technique (SMOTE) was employed, resulting in a more balanced dataset for training the classification models. The authors evaluated state-of-the-art supervised algorithms, including Random Forest (RF) and Extreme Gradient Boosting (XGBoost), assessing their performance in multi-class fault classification. The combination of XGBoost with SMOTE achieved the highest accuracy, approximately 98%, outperforming other methods. This outcome underscores the effectiveness of integrating advanced algorithms with data balancing techniques to enhance the accurate detection of machine states, thereby facilitating informed predictive maintenance decisions. This study is particularly pertinent to the analysis of plastic injection molding machines, as it demonstrates the practical applicability of classification models within real industrial environments, providing a robust foundation for the development of intelligent condition monitoring systems.

Ke & Huang [6] propose an approach for classifying the quality of injection-molded components through machine learning, focusing on Multilayer Perceptron (MLP) models. The study relies on the extraction of quality indices derived from process data, primarily internal pressure and other variables measured during the production cycle. The research evaluates different classification schemes by varying the number of quality categories to assess their impact on model performance. The dataset includes samples reflecting natural process variability and defects, which are quantitatively transformed into indices to serve as model inputs. Results indicate that the MLP model effectively distinguishes multiple quality levels with high accuracy, although the classification performance is influenced by the granularity of the defined classes. The study emphasizes the utility of classification techniques as efficient tools for process monitoring in injection molding, enabling early detection of states that may compromise the final product's quality. This research complements machine state analysis by linking process monitoring with product quality, demonstrating how classification models can support the maintenance of operational excellence in injection molding processes.

Aslantaş *et al.* [7] propose a machine learning model for identifying and predicting failures in plastic injection molding (PIM) machines. They research on sensor data analysis to anticipate failures before they occur, enabling more efficient maintenance planning and preventing unexpected production downtimes. To achieve this goal, classification algorithms such as Random Forest (RF) and Extreme Gradient Boosting (XGBoost) were applied to data collected from three machines in a home appliance manufacturing plant. Additionally, the SMOTE technique was employed to balance data distribution and enhance model accuracy. This study introduces a machine learning model designed to predict failure types in PIM machines based on sensor data. The model is constructed using classification algorithms to analyze sensor readings and forecast machine failures before they occur. Failure identification in PIM machines follows multiple stages, from raw data collection to classification. The estimation of Remaining Useful Life (RUL) plays a crucial role in identifying machines with a higher likelihood of failure. RUL, defined as the time interval between the present moment and the point of failure or maintenance requirement, is computed using historical maintenance and failure data. Several factors, such as clamping force, cycle time, and oil temperature, influence this interval. These data are obtained through sensors, process parameter logs, alarms, and maintenance records. Two classification algorithms were employed to develop predictive maintenance models: Random Forest and XGBoost. Raw sensor data alone do not sufficiently describe failure types, making feature extraction a crucial step. Extracted features include minimum, maximum, mean, skewness, kurtosis, and entropy. In this study, raw data were aggregated into 60-minute intervals, and features were automatically extracted for each interval. Data were collected from three plastic injection molding machines in a Turkish home appliance factory. The dataset sizes range from 205,000 to 913,000 records, covering the period from 2018 to 2021. A separate predictive model was constructed for each machine using variables such as clamping force, cycle time, and oil temperature. Experimental results indicate that RF and XGBoost, with and without SMOTE, demonstrated strong performances. XGBoost with SMOTE achieved the best performance, with an average accuracy of 98%. The highest F-score was also obtained using XGBoost with SMOTE (0.98), compared to RF (0.88) and other variations. This study estimated the types of failures in plastic injection molding machines, addressing the issue as a multi-class classification problem for predictive maintenance. Handling missing values and extracting relevant features were crucial steps in the process. The XGBoost and Random Forest (RF) algorithms were evaluated, with XGBoost demonstrating superior performance, especially when combined with the SMOTE technique. The achieved accuracy was 98% with XGBoost + SMOTE.

3.2 Maintenance and process optimization

Pierleoni *et al.* [8] discuss the evolution of maintenance strategies in industrial settings, emphasizing the importance of Predictive Maintenance (PdM) within Industry 4.0. Effective maintenance management is critical to avoid unexpected failures, which can result in high costs and negatively impact product quality and system reliability. Despite its importance, many companies have yet to adopt advanced strategies to optimize their maintenance budgets. The study focuses on the application of PdM in four electric plastic injection molding machines equipped for Industry 4.0. Analyzing failure patterns and monitoring process variables, a methodology was developed to predict component wear and prevent complete breakdowns. Data were collected from machine sensors, measuring variables such as injection pressure, plasticization volume, cycle time, temperature, and motor force. Since most companies follow preventive maintenance strategies that avoid obvious failures, obtaining real defect data was challenging. Instead, failures were inferred based on qualitative analysis of historical data and expert interviews. The collected data were labeled into two categories: ‘optimal’ operation and ‘functional limit,’ allowing the development of predictive models for twelve different machine-product combinations. A key limitation of the study is the scarcity of real failure data, which affects the accuracy of adverse condition classifications.

Pérez-Mora *et al.* [9] propose a study of “Plastic Injection Molding Process Analysis: Data Integration and Modeling for Improved Production Efficiency.” They present a data-driven framework aimed at improving the efficiency and quality control of plastic injection molding through the application of classification models. The study integrates real-time data acquisition from injection molding machines using a custom IoT-based DAQ system, capturing key variables such as injection time, cycle time, and mold pressure. A thorough exploratory data analysis was conducted to identify patterns and correlations that influence product quality and process stability. The core of the methodology involves the application of supervised classification algorithm, specifically logistic regression and random forest—to detect anomalies and predict process behavior. These models were trained to classify whether a production cycle met optimal operational conditions, with the random forest model achieving a remarkably high accuracy of 99.5%, and logistic regression reaching 97%. Feature importance analysis revealed that cycle time was the most critical variable influencing classification performance. The models were embedded into an intelligent agent system capable of real-time monitoring and adaptive decision-making. A graphical interface was developed to facilitate operator interaction, providing visual analytics and actionable recommendations based on model outputs. The integration of classification

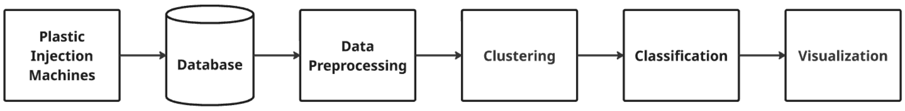


Figure 2: Diagram illustrating the steps of the process used followed.

modeling with real-time feedback mechanisms demonstrated tangible improvements in production reliability, energy efficiency, and defect reduction. This study underscores the potential of machine learning classification in enabling predictive, responsive, and intelligent manufacturing systems.

4 Data and Methodology

In this study, an unsupervised learning approach was used. This is a type of machine learning that analyzes data without human supervision. Unlike supervised learning, it works with unlabeled data, allowing models to discover patterns and insights independently. This approach was employed to analyze the dataset, using Principal Component Analysis (PCA) for dimensionality reduction and Density-Based Spatial Clustering of Applications with Noise (DBSCAN) for clustering. This method is particularly useful for organizing large volumes of information into clusters and identifying previously unknown patterns. The methodology followed a structured workflow involving data preprocessing, feature extraction, and clustering to uncover meaningful patterns in the data. Figure 2 illustrates the steps of the process that was followed.

4.1 Dataset Description and Data Preprocessing

The dataset consists of sensor values recorded by the machine over time. It contains failure data, although it doesn't show what kind of failure it was. Due to its high-dimensional nature, a preprocessing step was necessary to reduce the dataset size while preserving crucial information to improve clustering performance.

This dataset contains records from April 2024 to January 2025, covering a sampling period of 274 days. The records were sampled at a frequency of seconds, based on a *best effort* approach. There is a total of 48 million records across 129 variables and 7 machines. The subset for analysis was subsequently reduced to 19 critical variables, containing 3,242,214 records from a single machine, Machine 76. Table 2, as well as figures 3 and 4, show the statistics of the variables used for clustering and the plots of the records for two of the variables utilized.

As we can see from these two graphs 3 & 4, the data sampling period together with where the data has the highest frequency, and the areas where black stands out are due to the high number of points. The dataset had areas of missing records, in all the critical variables used. As can be seen in Figure 3, it is possible to observe the absence of records for some time intervals. The values were replaced using the forward-fill (ffill) method in python pandas library, which drags the last valid value known until a new valid value is found. This approach was chosen because sensor values often remain constant over during operation making the previous value a reasonable estimator.

4.2 Dimensionality Reduction Using PCA

Principal Component Analysis (PCA) is a statistical technique that transforms high-dimensional data into a lower-dimensional space while retaining most of its variance. This process is partic-

Table 2: Description of variables with their respective statistical values.

Variable	Number of values	Mean	Max	Min	Std. Dev.
Specific_injection_pressure_peak_value	170645	1710.18	2071.20	0.0	363.77
Switchover_volume_actual_value	170642	159.31	502.76	0.0	29.71
Specific_pressure_at_switch_over	170644	1631.39	2000.60	0.0	338.75
Specific_holding_pressure_peak_value	170642	1629.60	1997.90	0.0	337.83
Material_cushion_smallest_value	170642	35.22	495.08	0.0	23.65
Material_cushion_after_holding_pressure	170643	55.38	697.64	0.0	25.63
Material_cushion_end_holding_pressure	170642	36.23	538.11	0.0	28.06
Shot_volume	170644	991.61	1096.76	0.0	161.45
Injection_time	170644	3.72	20.50	0.0	0.42
Speed_peak_value	170642	0.43	0.65	0.0	0.06
Specific_back_pressure_peak_value	170644	78.77	301.30	0.0	10.78
Plasticizing_volume	170642	1049.19	1132.39	0.0	161.97
Plasticizing_time	170644	14.83	330.21	0.0	2.76
Clamping_force_peak_value	170644	5077.88	5255.10	0.0	465.12
Mold_opening_stroke_peak_value	170642	644.68	652.20	0.0	57.50
Cycle_time	170644	59.11	2335.19	0.0	25.31
Cooling_time	170642	24.48	333.04	0.0	3.57
Cycle_time_holding_pressure	170641	14.41	20.00	0.0	2.13
Barrel_Temperature_Zone_Actual_Temperatures_INJ1_Z01	170641	200.39	280.20	22.90	27.33

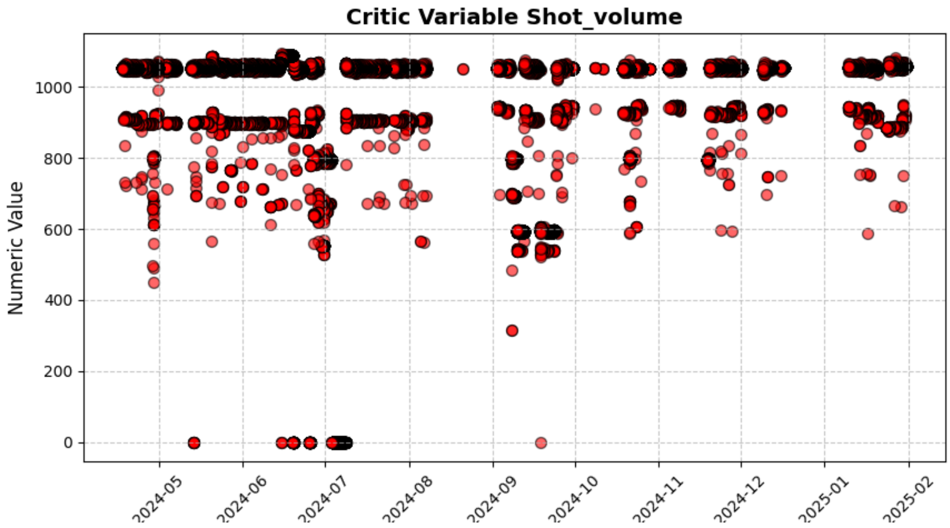


Figure 3: Critical variable analysis of shot volume for Machine 76. It is evident from the chart that there are discrepant values, different states and also missing values

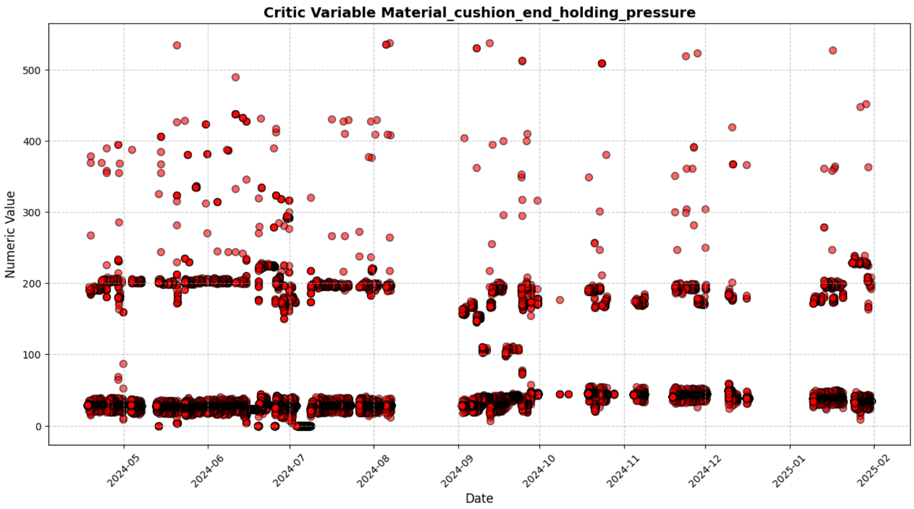


Figure 4: Critical variable analysis of material cushion end holding pressure for Machine 76.

ularly useful in clustering, as it removes noise and redundant information, leading to improved performance. PCA identifies directions, known as principal components, along which data exhibits the most variance, and projects the data onto these components.

Since PCA is sensitive to scale, all numerical features were standardized using z-score normalization, ensuring a mean of zero and a standard deviation of one. This step prevented features with large magnitudes from dominating the analysis.

As the dataset contains a large number of records and the DBSCAN clustering model uses a large amount of memory, it was necessary to reduce the dataset preserving as much as possible of the variance. For this reason, PCA was tested with *variance threshold* values between 0.8 and 0.995, which meant to have between 2 and 4 components.

4.3 Rationale and Objectives for Clustering

Since the dataset does not contain any labels indicating the machine's operational state, it is necessary to apply an unsupervised learning approach to infer these states from the data. Specifically, we use clustering to group similar observations together based on the values of the available features.

The rationale for employing clustering lies in its ability to uncover natural groupings within the data without prior knowledge of the categories. By doing so, we aim to identify distinct operational states of the machine. Each cluster is expected to correspond to a different state, such as normal operation, early signs of degradation, or imminent failure.

Although this technique is quite interesting, it presents certain challenges. Most notably, the fact that the groups identified in the data may not reflect real-world groupings. To address this, several tests needed to be conducted using different parameters, followed by a meeting with company stakeholders to determine the number of clusters that best fit the context. This gave greater reliability to the study.

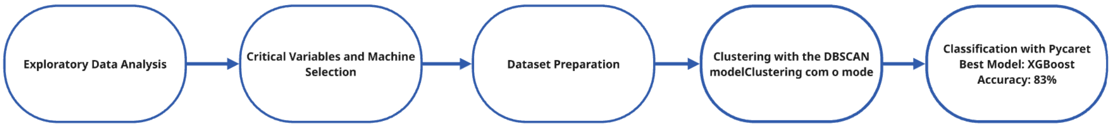

Exploratory Data Analysis
Critical Variables and Machine Selection
Dataset Preparation
Clustering with the DBSCAN modelClustering com o mode
Classification with Pycaret
Best Model: XGBoost
Accuracy: 83%

Figure 5: Diagram showing the developed work process.

4.4 Clustering with DBSCAN

DBSCAN is a density-based clustering algorithm that groups data points based on their density, making it well-suited for datasets with irregular cluster shapes. Unlike K-Means, DBSCAN does not require specifying the number of clusters beforehand. Instead, it identifies clusters as dense regions separated by areas of lower density.

DBSCAN assigns each data point to a cluster or labels it as noise if it doesn't meet the density criteria — that is, if the number of neighboring points within a given radius is less than a predefined minimum. A core point is one that has at least this minimum number of neighbors within the radius. Points that do not have enough neighbors are classified as noise.

This model contains two parameters that can be changed to obtain different results. These parameters, min_samples and eps, are responsible for defining the minimum number of points within the radius to form a cluster and for defining that radius, respectively. The values used for this study for min_samples are between 50 and 300 and for eps between 0 and 2. These parameter values were chosen based on the dataset values so that no single cluster or several irrelevant clusters are created.

4.5 Classification with PyCaret

PyCaret is an open-source, low-code machine learning library in Python that automates machine learning workflows. It was designed and implemented to simplify the work of data scientists. With just a few lines of code, it automates the entire machine learning workflow, from data preparation and model comparison to hyperparameter tuning, cross-validation, and final model selection.

In this study, PyCaret was also used for easy and balanced splitting of the dataset across all clusters, ensuring the same proportion of data from each cluster was used for both training and testing. Additionally, PyCaret trains multiple machine learning models and tests various hyperparameter configurations for each one. The best-performing model is then selected, saved, and used in the application. It also offers a variety of visualizations to assess classification performance and detect potential issues such as overfitting or underfitting.

Before carrying out the classification tests to determine the most effective classifier, the data were divided equally, percentage-wise, among all the clusters. This approach aims to ensure that, regardless of the number of samples per cluster (for example, one cluster with 1000 samples and another with only 100), the proportion of data allocated for training and testing remains uniform. The data_split_stratify parameter, provided by PyCaret, ensures that.

Figure 5 illustrates the steps followed during the data analysis process.

5 Clustering and Classification Results

Clustering and classification results are presented and discussed below.

Table 3: Description of clusters, their respective colors, and meanings. ID is the label of the cluster. Number -1 refers to noise points, which are isolated and do not form any cluster. Sil is the silhouette score.

ID	Color	Description	Sil.
-1	Light Blue	*Outliers*, possible noise points or very severe faults.	79%
0	Dark Blue	Reduced or anomalous operating condition.	79%
1	Light Green	Normal operation with a 4-cavity mold.	79%
2	Dark Green	Normal operation with a mold different from the light green one.	79%
3	Pink	Machine operating with fewer than 4 cavities.	79%
4	Red	Machine stopped.	79%
5	Yellow	Anomalous operation, capturing anomalies from the pink *cluster* (3 cavities).	79%
6	Orange	Severe anomaly, but with low occurrence, possibly irrelevant.	79%

5.1 DBSCAN Results

After optimization, DBSCAN identified for the machine a total six clusters plus noise. Noise in the present case can be discrepant samples caused by electromagnetic noise, spurious phenomena or a machine malfunction. In the present study, after careful analysis, the noisy samples were considered not relevant for further analysis.

Figure 6 shows a visualization of the results of the clustering algorithm. The figure illustrates how the algorithm identifies different clusters over the time, for variable Shot Volume. Figure 7 shows the same for variable Injection Pressure. Those clusters were formed using four PCA components, which explain up to 99.5 % of the variance in the dataset. Table 3 shows a description of each cluster. Points assigned label -1 are not included in any cluster, so they are considered noise. The description of the clusters and the respective assignment were discussed together with the provider of the dataset, and validated by the customer.

As the figures and the table show, DBSCAN was able to successfully separate the working states of the machine, with a silhouette score of 0.79. Points assigned labels -1 (noise), or 6, may require urgent attention because they refer to possible anomalous states. Cluster 5 refers to a state which also requires attention, because the machine is not operating in good condition.

5.2 Detailed Cluster Analysis

Since the machine operates with a four-cavity mold, a specific cluster immediately forms, represented in light green in Figure [6]. If one of the cavities exhibits a defect or malfunction, the machine operates with three cavities. This operating condition forms a separate cluster. Another example is the replacement of the mold with one that has only a single cavity, represented in the graph by the dark green color.

The clusters with the highest potential to represent an anomalous machine state were identified as dark blue and yellow clusters. As shown in Figure [7], this cluster corresponds to values that fall above and below the ranges considered normal, thus indicating a potential irregularity in the machine's operation. Additionally, the light blue points, considered outliers, may be classified as anomalous points.

Through detailed analysis of the clusters with the dataset provider, it was possible to identify that one of them is directly related to the machine's idle state. This cluster, represented in red, corresponds to the moment when the machine is in standby mode. In this state, the machine is not performing any active operations, awaiting its next task or activation. Distinguishing this cluster is essential for understanding machine inactivity periods and effectively monitoring its

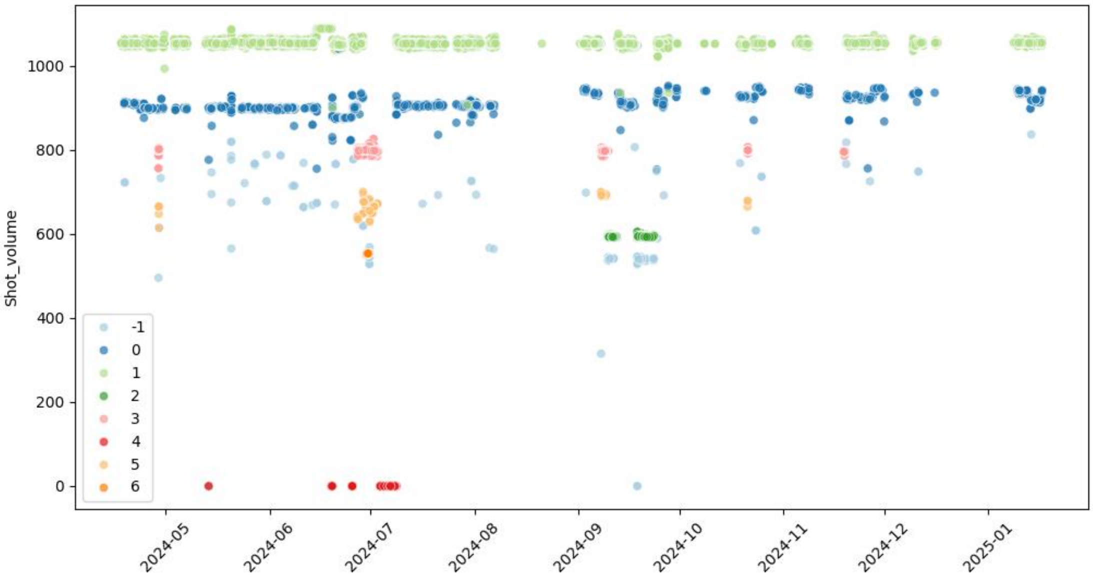


Figure 6: Illustration of DBSCAN clustering results for variable Shot Volume. There are 7 clusters with labels 0-6, plus noise with label -1. Cluster meanings and colors are described in Table 3

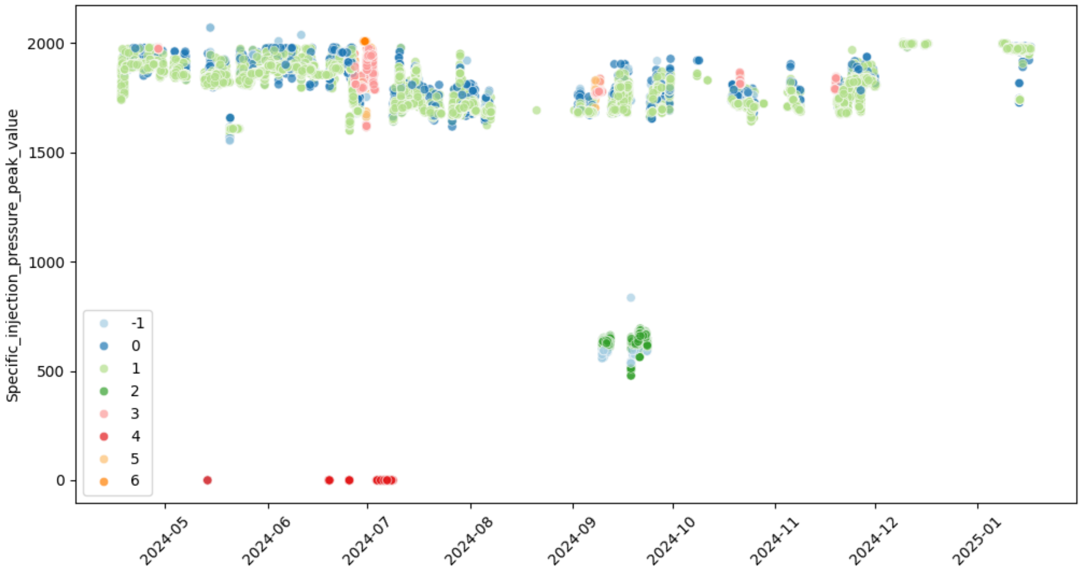


Figure 7: Illustration of DBSCAN clustering results for variable Specific Injection Pressure Peak Value. Cluster meanings and colors are described in Table 3

Table 4: Statistical data of the principal components of each cluster and noise. ID is point label assigned by DBSCAN

ID	**PCA1**				**PCA2**				**Description**
	min	max	mean	std	min	max	mean	std	
-1	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	Noise. Possible Severe Faults
0	-105.31	506.25	217.51	135.43	-300.29	238.49	-43.62	77.99	Reduced or Anomalous Operation
1	-37.88	540.91	229.67	137.53	-312.02	173.47	-48.17	89.29	Normal Operation with 4-Cavity Mold
2	-1880.73	-1614.96	-1688.68	24.43	688.59	860.92	735.20	16.29	Normal Operation with 1-Cavity Mold
3	-167.88	370.22	134.89	97.70	-154.71	241.31	-20.66	52.02	Operation with Less than 4 Cavities
4	-5411.40	-5409.07	-5410.74	0.72	-2640.40	-2632.49	-2635.88	2.67	Machine Stopped
5	-210.95	332.50	107.44	131.40	-149.69	250.29	-20.42	76.00	Reduced or Anomalous Operation
6	428.60	435.49	429.91	1.03	-218.35	-203.47	-213.67	3.25	Possible Anomaly

operational cycles.

Table 4 shows some statistical properties of each cluster. The statistics are shown for the two first components of the Principal Component Analysis (PCA1 and PCA2): min is the minimum, max is the maximum, mean is the average value and std is the standard deviation within the cluster. For noisy samples the values are not shown, since noise does not form a cluster. A more detailed description of the clusters is given below.

- ***Points labeled* -1: *Outliers* / Possible Severe Faults**. Those points are extremely dispersed. Considered *outliers*, they potentially represent noise in the data or severe machine failures.
- ***Cluster* 0: Reduced or Anomalous Operation**. There is lower variability compared to *outliers*, with PCA1 values from -105.30 to 506.25. Indicates reduced or anomalous machine operation, possibly associated with an alert condition or low operational efficiency.
- ***Cluster* 1: Normal Operation with 4-Cavity Mold**. Well-concentrated values, with a PCA1 mean of 229.67 and controlled dispersion. Represents normal machine operation with its most common 4-cavity mold, indicating smooth process running.
- **Cluster 2: Normal Operation with 1-Cavity Mold**, PCA1 values concentrated in a negative region (-1880.73 to -1614.95) with low standard deviation. Indicates normal machine operation with a different mold compared to Cluster 1. Useful for classifying different operational modes of the machine.
- ***Cluster* 3: Operation with less than 4 Cavities**. Moderate variability, PCA1 values from -167.87 to 370.21. Represents an operational state where the machine runs with fewer than 4 cavities, expected in certain production cycles. This occurs when one of the cavities has a defect and is covered up, keeping the machine working in a normal state but with only 3 cavities, reducing the number of products produced.
- ***Cluster* 4: Machine Stopped**. Extremely concentrated values in PCA1 (-5411.39 to -5409.07) and PCA2 (-2640.39 to -2632.49), with negligible standard deviation. Indicates that the machine is stopped, with no significant operational variations.
- ***Cluster* 5: Reduced or Anomalous Operation**. Intermediate values, PCA1 ranging from -210.95 to 332.50, with controlled dispersion. Suggests anomalous machine operation, representing low production or failure state associated with specific conditions.
- ***Cluster* 6: Possible Anomaly**. Values concentrated in PCA1 (428.60 to 435.49), slightly negative variation in PCA2 (-218.35 to -203.47). Low presence in data, considered irrelevant for overall analysis. May represent a rare anomalous situation.

Table 5: Classification metrics by class

ID	Precision	Recall	F1-score	Support
0	0.599	0.566	0.582	14394
1	0.831	0.981	0.900	629724
2	0.956	0.412	0.575	187125
3	0.695	0.617	0.654	50564
4	0.865	0.957	0.908	7427
5	0.720	0.512	0.598	2119
6	0.915	0.789	0.847	1217
Accuracy	0.833 (Total support: 892570)			
Macro avg	0.797	0.690	0.724	892570
Weighted avg	0.846	0.833	0.812	892570

5.3 Classification Results

Once the clustering process was completed and the clustering-labeled dataset prepared, PyCaret was applied to proceeded to automatic classification model selection with the best performance based on predefined metrics. Initially, several models were automatically tested using PyCaret; however, none demonstrated true robustness, with the highest accuracy reaching approximately 69%. Consequently, the XGBoost model was implemented due to its well-known robustness in similar cases. Using tree-learning algorithms, extreme gradient boosting (XGBoost) is one of the most efficient classification models available. However it is still not available in PyCaret, so it was applied separately.

For XGBoost training and validation was used a dataset with about 890,000 records, included only data from 7 clusters.

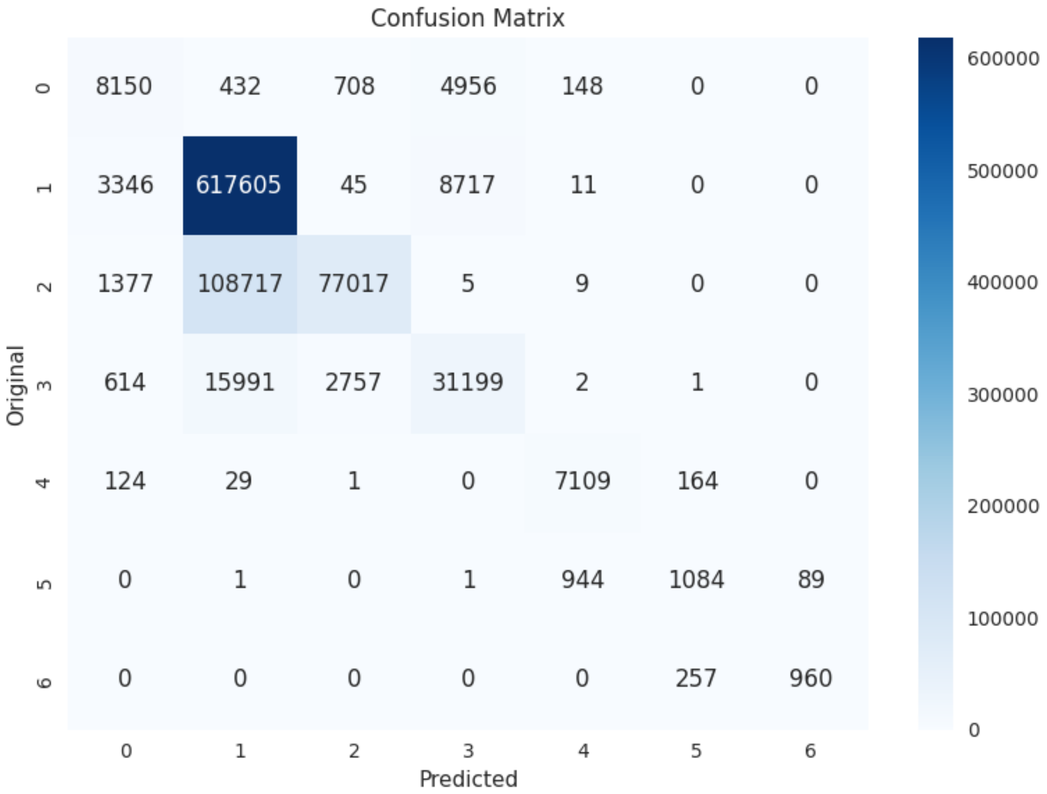
A detailed graphical analysis of the time series shows that there are some instantaneous transitions, which are but noise, for it is not possible for the machine to transition from one state to another and then immediately back to the previous state. Hence, the time series was filtered using a median filter with a window of variable width. This approach replaces each data point with the median value of its neighbors within a window of a given size. A window of size 5, applied specifically to the cluster labels, showed the best performance. The median filter reduces the impact of noise and outliers by smoothing the data, thus enhancing the quality of the training set.

Applying the median technique resulted in an improved accuracy of 83.26%, a value considered acceptable for reliable classification.

Figures 9 and 10 shows the time series of variable Shot_Volume along a period of time. The points are marked with the respective cluster color. On the left is the actual value, on the right is the predicted cluster.

6 Discussion

The results obtained from the DBSCAN clustering model provide valuable insights into the operational states of the plastic injection molding machine. A comparison with state-of-the-art methods, such as Predictive Maintenance by Pierleoni *et al.* [8] and Aslantaş *et al.* [7], represents an important first step. This is because, with the classification model, it becomes possible to determine when machines require maintenance, detect anomalies, or identify human errors. For example, if a mold is changed but an operator forgets to adjust the settings, this method can



Original \ Predicted	0	1	2	3	4	5	6
0	8150	432	708	4956	148	0	0
1	3346	617605	45	8717	11	0	0
2	1377	108717	77017	5	9	0	0
3	614	15991	2757	31199	2	1	0
4	124	29	1	0	7109	164	0
5	0	1	0	1	944	1084	89
6	0	0	0	0	0	257	960

Figure 8: Confusion Matrix

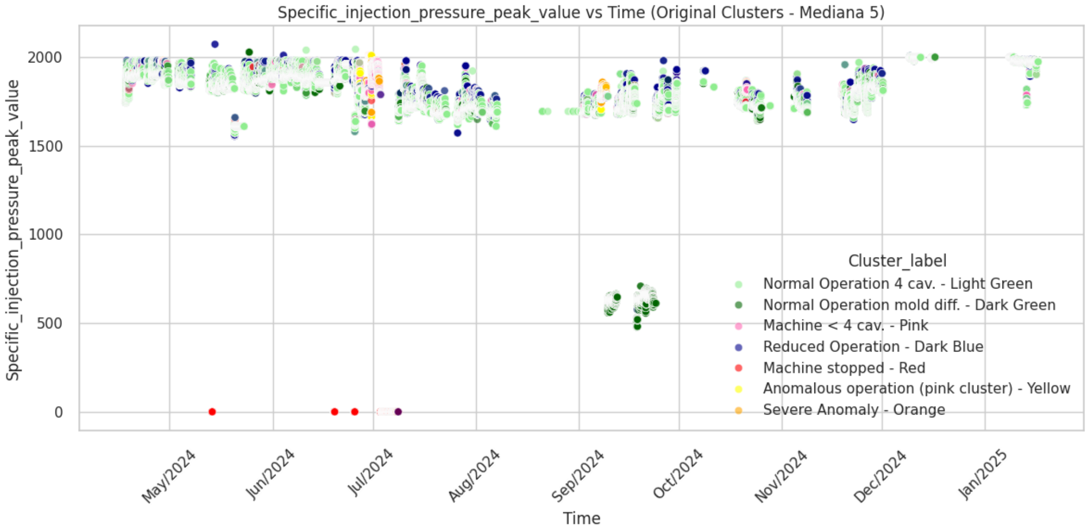


Figure 9: Classification Model: Comparison between Label and Prediction Label

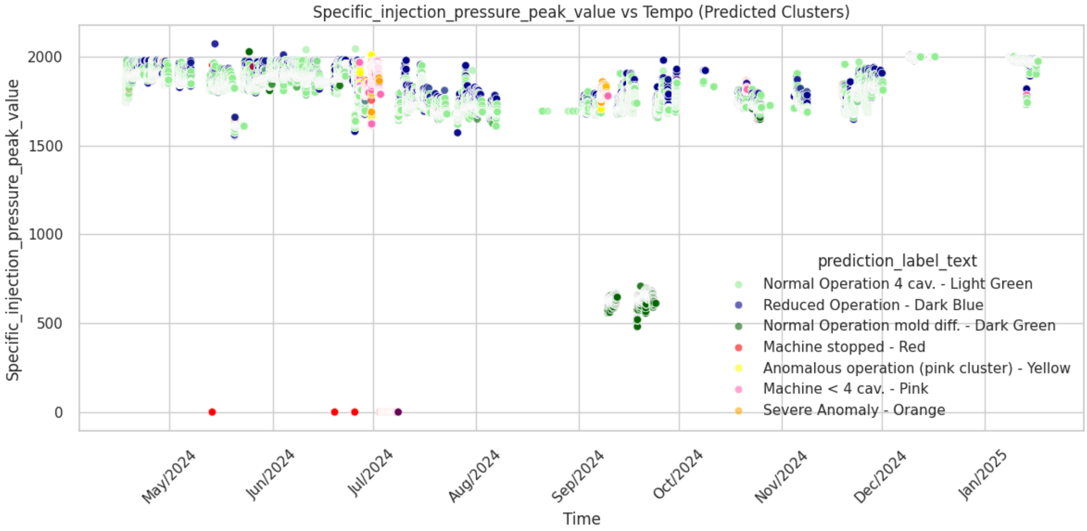


Figure 10: Classification Model Clusters Predicted

detect, and an alarm can be generated.

One of the main contributions of this study is the identification and differentiation of machine states, which will assist industrial companies operating plastic injection machines to detect anomalies more easily. Unlike conventional analyses that rely on artificially generated data or data not originated from real contexts, this study successfully identified distinct machine states using a real dataset that covers some months of operation. This distinction is crucial, because it increases the likelihood that the clustering step will be more accurate and effective in real-world applications. Utilizing Principal Component Analysis to reduce data dimensionality and integrating the Silhouette Score metric to assess clustering quality, our methodology ensures robust segmentation of machine states, thereby supporting improved monitoring and predictive maintenance strategies.

A key insight from our findings is the effective detection of anomalous states. The identification of outliers (label -1), or anomalous operation (labels 0, 5, and 6), as potential severe malfunctions aligns with industry concerns regarding common failures in plastic injection molding machines [2]. These include issues such as injection device obstruction, mold cooling inefficiencies, and variations in pressure and temperature. By clustering machine states based on real operational data, our method offers a data-driven approach to early and straightforward fault detection, complementing conventional preventive maintenance.

Segmenting machine states into seven distinct clusters provides a more granular understanding of operational behaviors. For instance, the differentiation between normal operation with a four-cavity mold (light green cluster) and operation with the same mold but less operational cavities (pink clusters) enables more precise monitoring of mold performance. Similarly, identifying idle states (red cluster) allows better assessment of machine downtime and potential maintenance periods.

In our study, different parameters for DBSCAN were tested, and variance adjustments in PCA significantly improved clustering runtime and efficiency. Similar to the work of Zhang and Alexander *et al.* [3], Principal Component Analysis was instrumental in retaining the most relevant features, thereby substantially enhancing the clustering performance in this study. The achieved Silhouette Score of approximately 80% indicates high-quality clustering performance, validating our model's effectiveness in distinguishing machine states. Moreover, the approval of the dataset supplier adds credibility and reflects the real-world applicability of the results.

However, as this is the first time the supplier made the dataset available, some variables lacked relevance or meaning for this study. Additionally, due to the project's novelty, there were data shortages and inconsistencies, complicating external validation. After internal analysis, the most critical variables were identified and focused upon in this study.

Compared to the study by Pérez-Mora *et al.* [9], although clustering was primarily employed here as a strategy to generate labeled data for subsequent supervised classification, the achieved accuracy values were comparable, demonstrating the effectiveness of the proposed methodology even when relying on unsupervised techniques for initial data labeling.

One conclusion reached is that using more molds in the machine could lead to additional clusters, potentially overlapping with the malfunction of other molds. In such cases—although not observed here—different approaches would be necessary, such as performing clustering separately for each mold. However, this would undermine one main advantage of the current approach: the ability to detect human errors.

Another challenge stems from the early stage of the company, where collected data may contain inconsistencies and the variables used for analysis may be redefined as their relevance becomes clearer. Consequently, the classification model requires regular retraining to remain aligned with the evolving dataset, incorporating any added or removed variables.

This research advances understanding of plastic injection molding machine behavior by introducing a clustering-based approach for state identification and classification to facilitate anomaly detection. Future work may involve retraining the classification model with new data and variables, integrating additional sensor inputs, and refining clustering techniques to further enhance predictive capabilities and operational efficiency.

In summary, the main objectives outlined in the introduction were fully achieved. The behavior of the plastic injection molding machine was thoroughly analyzed using both literature review and real-world dataset examination. Through the application of unsupervised clustering techniques, specifically DBSCAN combined with PCA, distinct machine operational states were successfully identified and characterized. Furthermore, these clusters served as a foundation for training a classification model capable of detecting anomalies and potential human errors in an industrial setting. The alignment of the results with real operational scenarios confirms the practical value of the proposed methodology and its potential for implementation in predictive maintenance systems.

7 Conclusion

This study addressed the challenge of accurately identifying and monitoring malfunctions in plastic injection molding machines. Unsupervised learning DBSCAN clustering algorithm was applied in order to determine machine states automatically, and then use them to train the classification model.

The results demonstrated that the more molds are used in the machine, the larger the number of clusters that are formed during the clustering process. Identifying seven distinct clusters, including normal operation, idle periods, and potential anomalies, our approach provides a more detailed and data-driven method for assessing machine behavior. The ability to detect abnormal states further enhances machine diagnostics and predictive maintenance strategies.

This research contributes to the state of the art by moving beyond structural optimizations and incorporating machine learning techniques to analyze machine performance dynamically.

Despite its advantages, this study has certain limitations. The clustering model's effectiveness depends on the quality and quantity of available data. Additionally, external factors such as material variations and environmental conditions may influence clustering outcomes, necessitating further refinements.

In addition, classification models must be retrained whenever new variables are introduced or existing ones are removed, as well as when new data becomes available, in order to ensure that the model remains accurate and up to date.

Future work should explore integrating additional sensor data, refining clustering algorithms, and incorporating real-time adaptive models to enhance predictive capabilities. Expanding this approach to other industrial machinery could further validate its applicability and effectiveness in improving manufacturing efficiency.

References

[1] Nurkamilya Daurenbayeva, Almas Nurlanuly, Lyazzat Atymtayeva, and Mateus Mendes. Survey of applications of machine learning for fault detection, diagnosis and prediction in microclimate control systems. *Energies*, 16(8), 2023.

[2] TopStar Machine. Common operating problems and solutions for several plastic injection machine, 2025. Plastic injection machine play a vital role in modern injection molding manufacturing processes. Available at https://www.topstarmachine.com/common-operating-problems-and-solutions-for-several-plastic-injection-machine/. Last accessed on: Feb. 13, 2025.

[3] Jin Zhang and Suraj M. Alexander. Fault diagnosis in injection molding via cavity pressure signals. In *IISE Annual Conference. Proceedings*, pages 1–6. Institute of Industrial Engineers, 2004.

[4] Dominik Kozjek, Rok Vrabič, David Kralj, Peter Butala, and Nada Lavrač. Data mining for fault diagnostics: A case for plastic injection molding. *Procedia CIRP*, 81:809–814, 2019. 52^{nd} CIRP Conference on Manufacturing Systems (CMS), Ljubljana, Slovenia, June 12-14, 2019.

[5] Gözde Aslantaş et al. Estimating types of faults on plastic injection molding machines from sensor data for predictive maintenance. *Journal of Manufacturing Systems*, 2020.

[6] Kun-Cheng Ke and Ming-Shyan Huang. Quality classification of injection-molded components by using quality indices, grading, and machine learning. *Polymers*, 13(3):353, 2021.

[7] Gözde Aslantaş, Tuna Alaygut, Merve Rumelli, Mustafa Özsaraç, Gözde Bakırlı, and Derya Bırant. Estimating types of faults on plastic injection molding machines from sensor data for predictive maintenance. *Artificial Intelligence Theory and Applications*, 3(1):1–11, 2023.

[8] Paola Pierleoni, Lorenzo Palma, Alberto Belli, and Luisiana Sabbatini. Using plastic injection moulding machine process parameters for predictive maintenance purposes. In *2020 International Conference on Intelligent Engineering and Management (ICIEM)*, pages 115–120, 2020.

[9] Jose Isidro Hernández-Vega, Luis Alejandro Reynoso-Guajardo, Mario Carlos Gallardo-Morales, María Ernestina Macias-Arias, Amadeo Hernández, Nain de la Cruz, Jesús E. Soto-Soto, and Carlos Hernández-Santos. Plastic injection molding process analysis: Data integration and modeling for improved production efficiency. *Applied Sciences*, 14(22), 2024.