Internship Report

OES-PDA Data Analysis for Inclusions Characterisation of High Quality Steel Grades

Alireza Behjoee

August, 2024

University of Groningen, Campus Fryslân BSc Data Science & Society

Internal Supervisor: Noman Haleem

External Supervisor: Aida Abbasalizadeh

Internship Report

University of Groningen, Campus Fryslân, BSc Data Science & Society

5th August, 2024

Student

Alireza Behjoee

S5395372

a.a.behjoee@student.rug.nl

Ritsumastraat 294, 8911 KM, Leuuwarden, The Netherlands

+31 681022619

Host Company

Tata Steel

Wenckebachstraat 1, 1951 JZ Velsen-Noord

Company Supervisor: Aida Abbasalizadeh

aida.alizadeh@tatasteeleurope.com

University mentor: Dr. Noman Haleem

n.haleem@rug.nl

Preface

During the minor in my Bachelor's program in Data Science and Society (DSS) at the University of Groningen, I chose to undertake an internship at Tata Steel IJmuiden in the Netherlands. From August 5 to November 5, I worked as a data scientist in their Center of Expertise department. My main responsibilities included data collection, cleaning, preprocessing, and visualization, as well as identifying trends and generating insights to support operational improvements.

During my Bachelor, I developed a strong foundation in data analysis, visualization, and computational methods. I chose to undertake this internship at Tata Steel to apply these skills in a real-world industrial setting. I wanted to experience how data science tools and techniques could contribute to actual industrial processes, where data-driven insights have significant operational impacts. I found this internship position at Tata Steel through LinkedIn. After initial discussions with the team, I was assigned a project centered around the analysis of Optical Emission Spectroscopy (OES) and Pulse Discrimination Analysis (PDA) data. This assignment addressed Tata Steel's ongoing challenge with SEN (Submerged Entry Nozzle) clogging and surface defects in Non-Grain Oriented (NGO) electrical steels, where accurate control of chemical composition is critical.

In the following sections, I will elaborate on Tata Steel's production processes, the specific challenges the company aimed to address, the tasks I was assigned, a summary of the results of my work, and the key learning outcomes. I will also discuss how these experiences align with the objectives of the DSS program.

Table of Contents

1. Introduction

1.1. Company Overview

Tata Steel IJmuiden, located in the Netherlands, is a leading steel production facility in Europe and an integral part of the global Tata Steel group. Tata Steel provides high-grade steel products to industries worldwide, including the automotive, construction, and energy sectors. The plant produces a diverse range of steel grades, each tailored to meet the specific needs of its clients, with a strong focus on innovation and sustainable practices throughout its production processes. In addition to its large-scale production capabilities, Tata Steel is home to a number of specialized departments that focus on advancing steel technologies and optimizing processes. Approximately 9,000 employees work at Tata Steel IJmuiden. However, I worked in the Center of Expertise department, a specialized team of 30 experts dedicated to driving innovation and sustainability by improving steel products, optimizing production processes, and developing environmentally friendly steel making techniques.

1.2. Technical Context and Challanges

Among the specialized steel products manufactured at Tata Steel, Non-Grain Oriented (NGO) electrical steels are critical for electric vehicles, where high surface quality, magnetic properties, and electrical efficiency are essential. The UTAM grade, which is a special type of NGO steel, is specifically engineered for this purpose. It is characterized by high silicon (Si) and aluminum (Al) content, along with low carbon (C) levels. Despite the precision required in its production, Tata Steel has faced challanges over the past few years in ensuring the quality of UTAM grades,

with two major challenges affecting their suitability for high-performance applications: SEN (Submerged Entry Nozzle) clogging and surface defects.

SEN clogging occurs in the casting stage, where the SEN nozzle directs the flow of molten steel into the caster mold. When clogging happens, the flow of steel is disrupted, causing inconsistencies in casting and, in severe cases, it can stop or slow down production, leading to economic losses and operational inefficiencies. Surface defects, another major problem, compromise the final steel quality, reducing its suitability for high-performance applications in electric vehicles. Both SEN clogging and surface defects are largely caused by the formation of solid inclusions, such as calcium sulfide (CaS), within the molten steel during the steelmaking process. These inclusions either accumulate within the SEN, causing blockages, or remain dispersed in the steel, leading to surface defects that undermine product quality.

1.3. Internship Project and Objectives

In this context, my internship centered on using my data science skills to analyze Optical Emission Spectroscopy with Pulse Discrimination Analysis (OES-PDA) data and assess its potential to predict clogging and surface defects. The primary goal of my work was to examine inclusion patterns, with a focus on identifying heats more prone to these issues, particularly those with high concentrations of CaS. This involved addressing key research questions:

1. Can OES-PDA data reveal differences in CaS content between problematic and non-problematic heats?
2. Which samples are most effective in predicting whether a heat will be problematic?

3. Are there other types of inclusions, aside from CaS, that might also contribute to clogging and surface defects?

By identifying patterns in the OES-PDA data, my work aimed to provide Tata Steel with actionable insights to improve quality control measures. These insights could potentially inform future adjustments in production processes, ultimately enhancing the reliability and quality of UTAM steel production.

2. Overview of OES-PDA data

OES refers to Optical Emission Spectroscopy, while PDA stands for Pulse Discrimination Analysis, which is applied within OES to measure the chemical composition of inclusions in steel. The process involves creating a rapid series of high-energy sparks in the argon-filled gap between an electrode (the cathode) and the steel sample's surface (the anode). These sparks ionize the argon, forming a plasma, and simultaneously melt, evaporate, and excite the elements within the sample. As the excited atoms return to a lower energy state, they emit light at wavelengths specific to each element. These emissions are detected and quantified against known standards, providing precise measurements of the sample’s composition. This sparking process lasts only a few milliseconds. The OES-PDA dataset encompasses two distinct groups: the old process and the new process. In the traditional (old) process, the steel is fully 'killed' by adding aluminum after the decarburization, significantly reducing oxygen levels. In contrast, the new process semi-kills the steel, maintaining higher oxygen levels, which alters the formation of inclusions compared to the traditional approach.

The OES-PDA raw data used in this study includes a total of 43 heats, categorized as follows with 5 to 9 burns collected per sample:

- 33 heats from the old process, with:
    - 7 identified as problematic
    - 26 identified as non-problematic
- 10 heats from the new process, with:
    - 1 identified as problematic
    - 9 identified as non-problematic

Samples are taken at different stages during the ladle furnace process:

- Ladle Furnace 60: Sample before the start of ladle furnace treatment
- Ladle Furnace 61: Intermediate sample taken during killed phase, mainly for sulfur (S) and trimming adjustments.
- Ladle Furnace 62: Another intermediate sample taken during killed phase, used for fine trimming.
- Ladle Furnace 63/64: Sample taken during killed phase, usually due to unreliable results from samples 61 or 62.
- Ladle Furnace 65: Sample taken before calcium treatment, rarely taken and mostly only a slag sample.
- Ladle Furnace 66: Typically the final sample from the ladle furnace process.

In order to classify heats as problematic or non-problematic, certain criteria—established by Tata Steel and the client prior to the start of my internship—were used. For a detailed explanation of these criteria, please refer to Appendix A. Primarily, heats are categorized based on surface

defect evaluations carried out by Tata Steel and the client, with defect measurements split into two main groups: defects extending above 40 cm and those above 25 cm. Another key criterion involves the analysis of SEN images taken after each casting sequence. These images are closely inspected to identify clogging. Additionally, other factors, such as the stopper slope, are considered to assess consistency in the steel flow rate, as variations in stopper positioning can signal potential issues during casting.

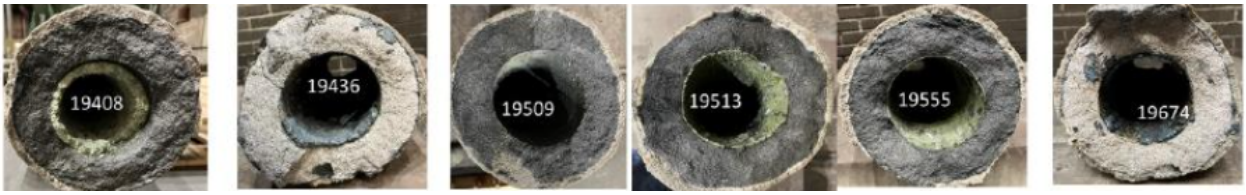

Figure 1. Images taken from SEN, showing its conditions after each sequence.

3. Methodology

Since it was established that cloggings in UTAM grades were associated with CaS and elevated sulfur (S) levels can contribute to clogging, the analysis began with defining a function to detect common upper outliers in both calcium (Ca) and sulfur (S) levels using the three-sigma rule. This rule identifies values that deviate significantly from the mean by considering the distribution of the data.

The three-sigma rule classifies any data point lying more than three standard deviations away from the mean as an outlier. To enhance the reliability of our results, the three-sigma rule was applied iteratively over 10 iterations. During each iteration, the function recalculated the mean

and standard deviation after removing the previously detected outliers, progressively improving the accuracy of the analysis. For all the heats, no common upper outliers were found after the fifth iteration, indicating that the process effectively removed all significant deviations.

This analysis was further extended to include additional combinations such as CaMg, CaAl, MgS, AlS, MgAl, CaMgS, CaAlS, CaMgAl, and MgAlS to explore whether there are other combinations that could potentially contribute to clogging and subsequently problematic heats.

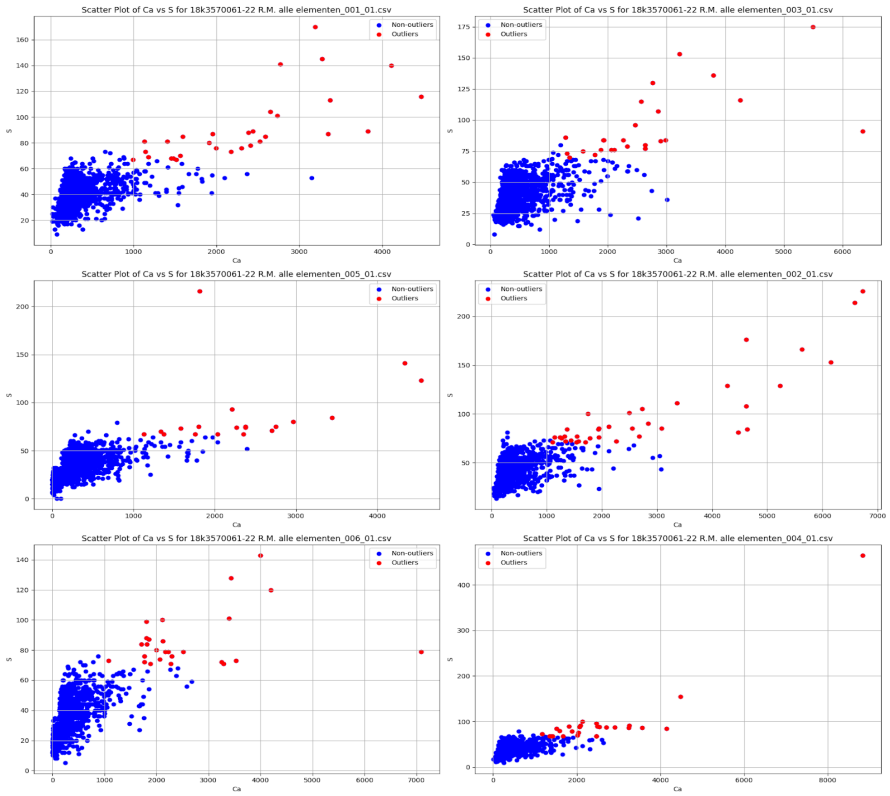

Figure 2. Detected common upper outliers for CaS heat K3570, sample 61, for each burn.

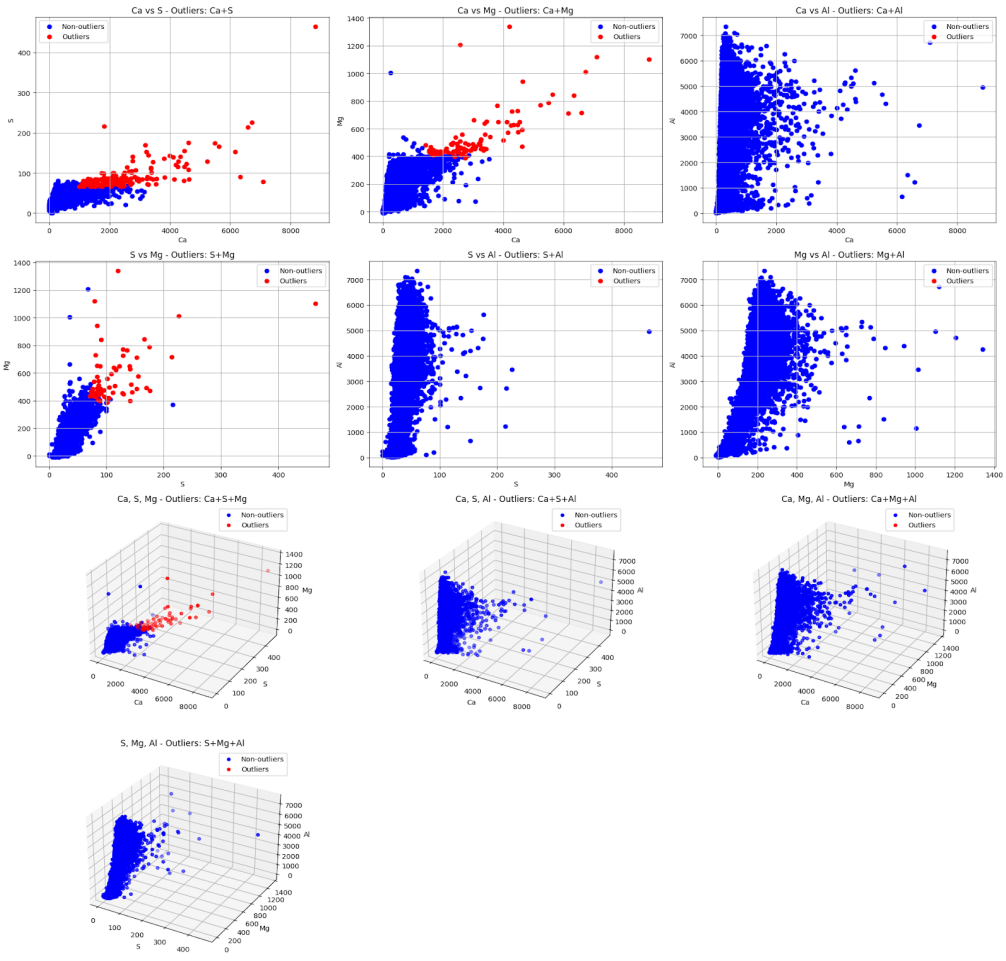

Figure 3. Detected common upper outliers of binary and ternary combinations of Ca, Al, Mg, and S for heat K3570, sample 61

After identifying the common upper outliers for each burn, the average was taken to account for the varying number of burns across different samples. This method was applied to all samples, and as an example, the results for heat K3570 (Sample 61) are shown in Figure 4a, and for heat K3783 (Sample 61) in Figure 4b.

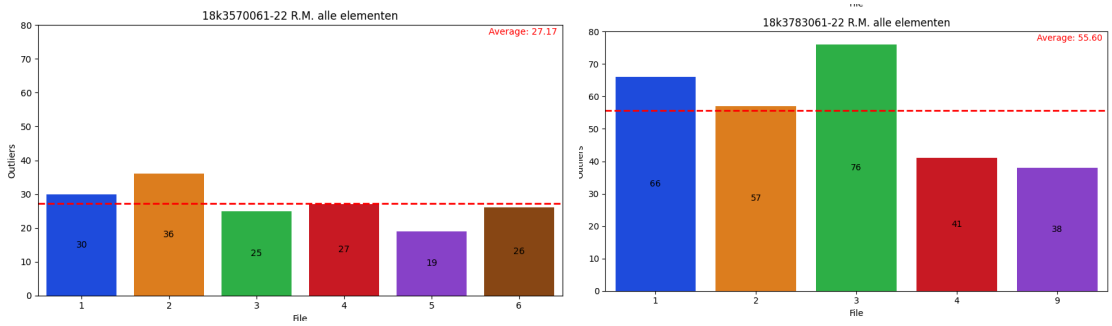

Figure 4. (a) The CaS average for heat K3570, Sample 61 (Old Process Non Problematic), (b) the CaS average for heat K3783m Sample 61 (Old process Problematic).

4. Results and Discussion

For the first part of the analysis, sample 66 was used from each heat, as this sample was available for almost all heats and provided a larger dataset for analysis. The total number of heat samples was 37, categorized as follows:

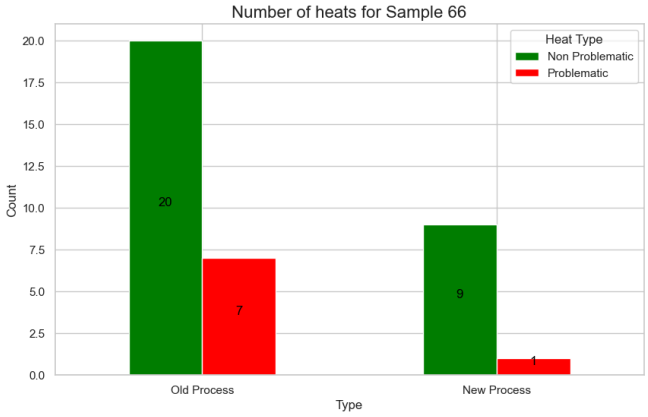

Figure 5. Number of heats for Sample 66

- Old Process: 27 samples
    - Non-problematic: 20

    - Problematic: 7
- New Process: 10 samples
    - Non-problematic: 9
    - Problematic: 1

After identifying the common upper outliers for each heat, a boxplot for the average number of CaS was plotted to assess potential differences in the CaS averages across different heat categories. It is important to note that the new process problematic heat was not included, as there was only one sample for this category.

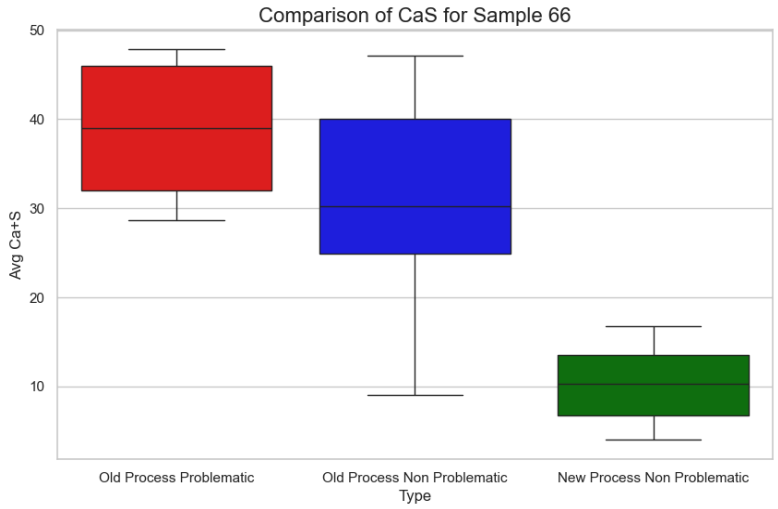

Figure 6. Comparison of CaS for sample 66

Figure 6 reveals several important insights. First, there is a noticeable overlap between the old process problematic and non-problematic heats. While the median CaS value for the old process problematic heats is higher, the ranges of both categories overlap significantly. This overlap suggests that CaS levels alone may not be useful for differentiating between problematic and non-problematic heats within the old process. The wide spread in CaS values for the old process

non-problematic heats, including the presence of higher outliers, further complicates the distinction. In contrast, the new process non-problematic heats show much lower average CaS values, indicating that these heats are less likely to experience CaS-related clogging issues. Therefore, sample 66 does not provide a clear distinction between problematic and non-problematic heats in the old process.

After finding that sample 66 was not a useful indicator, the focus shifted to samples taken after the addition of aluminum (Al) and silicon (Si). For some heats, this corresponded to samples 61, 62, or 63. However, the issue with these samples was that they were unavailable for almost half of the heats, which significantly reduced the data available for our analysis. In total, there were 20 samples: 10 from the old process (5 problematic and 5 non-problematic) and 10 from the new process (9 non-problematic and 1 problematic).

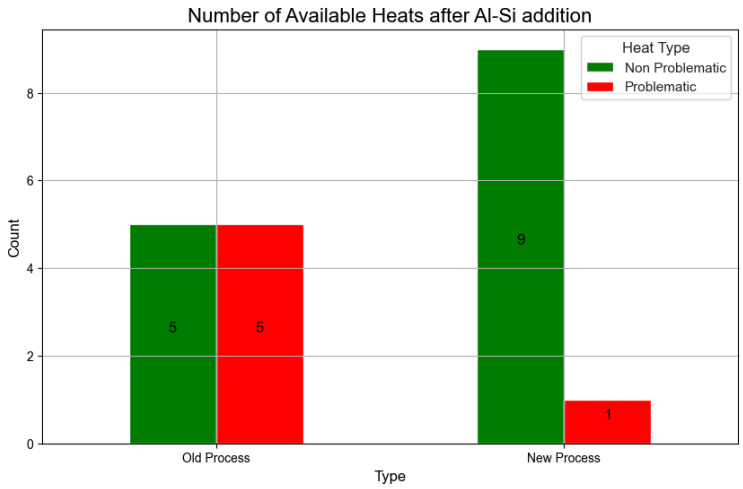

Figure 7. Number of available heats after Al and Si addition

The same boxplot was plotted for samples taken after the addition of aluminum and silicon shown in Figure 8. This time, the boxplot revealed a significant difference between the different

categories: new process non-problematic heats, old process non-problematic heats, and old process problematic heats. It should be noted that, similar to the sample 66 analysis, new process problematic heats were excluded from this analysis due to the limited availability of data. The boxplot for CaS content after the addition of aluminum (Al) and silicon (Si) reveals a much clearer distinction between the heat categories compared to the previous analysis of Sample 66. The new process non-problematic heats show significantly lower CaS values, tightly clustered within the 5-10 range. While there is still some overlap between the old process problematic and old process non-problematic heats, the separation between these two groups is more pronounced in this analysis compared to Sample 66, where the overlap was more extensive. This indicates that samples after Al and Si additions are a more effective differentiator between the heat categories, especially between the old and new processes.

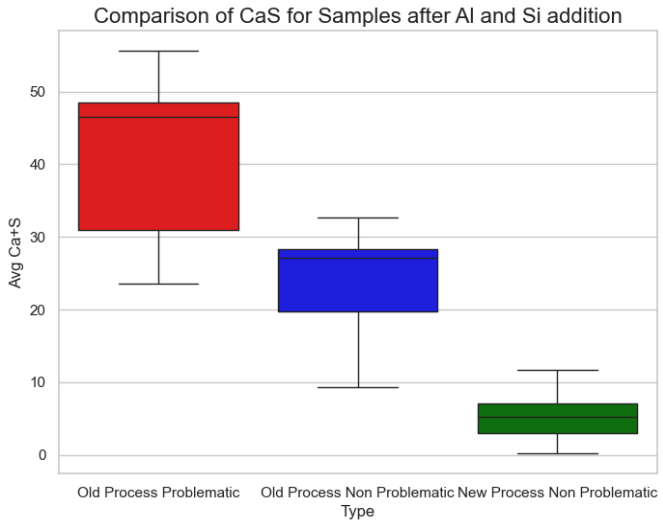

Figure 8. Comparison of CaS for Samples after Al and Si addition

Since CaS outliers are more clearly visible in samples taken after aluminum and silicon additions compared to Sample 66, further investigation was conducted into CaS outlier trends throughout

the entire process, as shown in Figure 9 for the old process and Figure 10 for the new process. The line graphs are divided into old and new processes, with problematic heats represented by dashed lines and non-problematic heats by solid lines. As illustrated in Figure 9, the old process shows more variability in CaS levels during the process, particularly between problematic and non-problematic heats. The problematic heats (dashed lines) display significantly higher CaS levels, especially in burn 61 and burn 66, compared to the non-problematic heats, which tend to maintain lower and more stable CaS values. In contrast, Figure 10 reveals that in the new process, CaS levels generally start low at the beginning of the process and gradually increase throughout the process, with a noticeable rise in later burns (65 and 66). Notably, K8959, the only problematic heat in the new process, exhibits the highest average CaS levels among the new process heats. However, since it is the only problematic sample from the new process, it is insufficient for reliably distinguishing between problematic and non-problematic heats within this process using Sample 66.

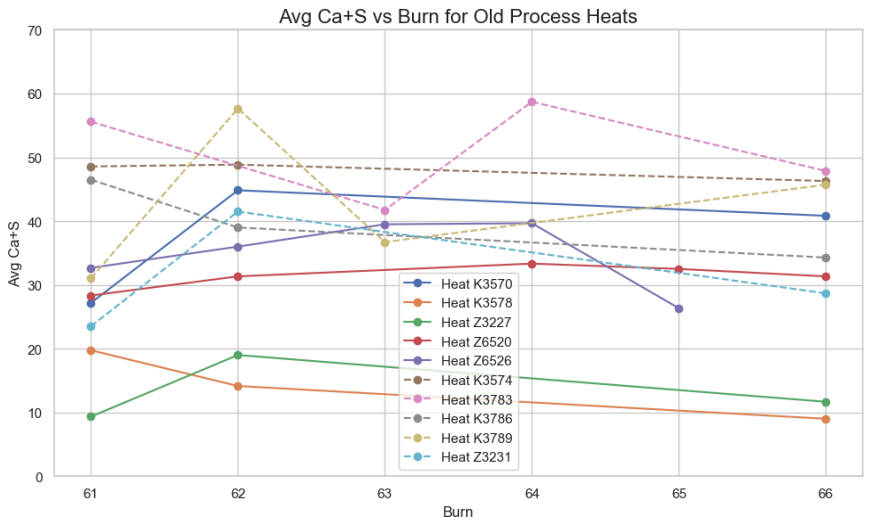

Figure 9. Avg CaS at each sample for Old Process heats

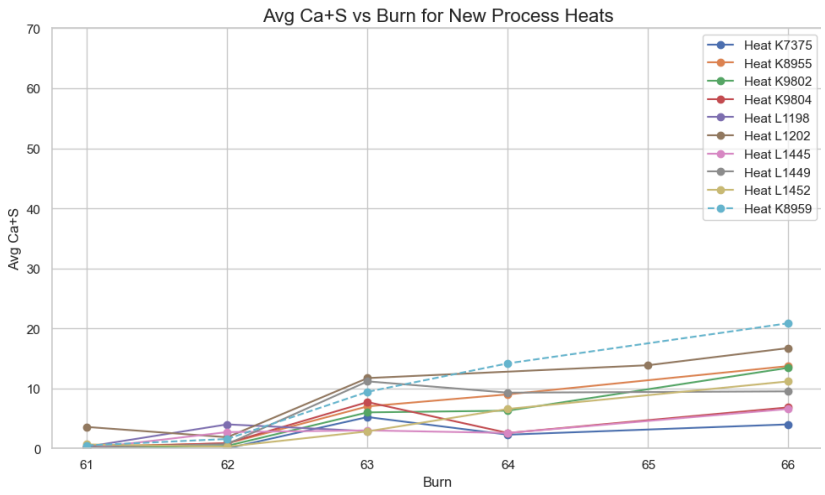

Figure 10. Avg CaS at each sample for New Process heats.

To assess whether other combinations contribute to problematic heats and the formation of CaS inclusions, a correlation analysis was conducted, as illustrated in Figure 11. As expected, Ca+S showed the strongest positive correlation (0.67) with problematic heats. Similarly, Ca+Mg+S and Mg+S exhibited notable positive correlations (0.59 and 0.56, respectively), indicating that sulfur combined with either magnesium or both calcium and magnesium is strongly associated with problematic heats. In contrast, combinations involving aluminum displayed negative correlations, reflecting the presence of oxide inclusions, which are less associated with problematic heats. An exception to this trend was Ca+Mg, which also indicated the presence of aluminum oxide inclusions. These findings highlight the importance of inclusion types, as oxide inclusions generally exhibit a negative correlation with problematic heats. Overall, MgS and CaMgS appear to significantly contribute to problematic heats, emphasizing the need to monitor and control these inclusion types.

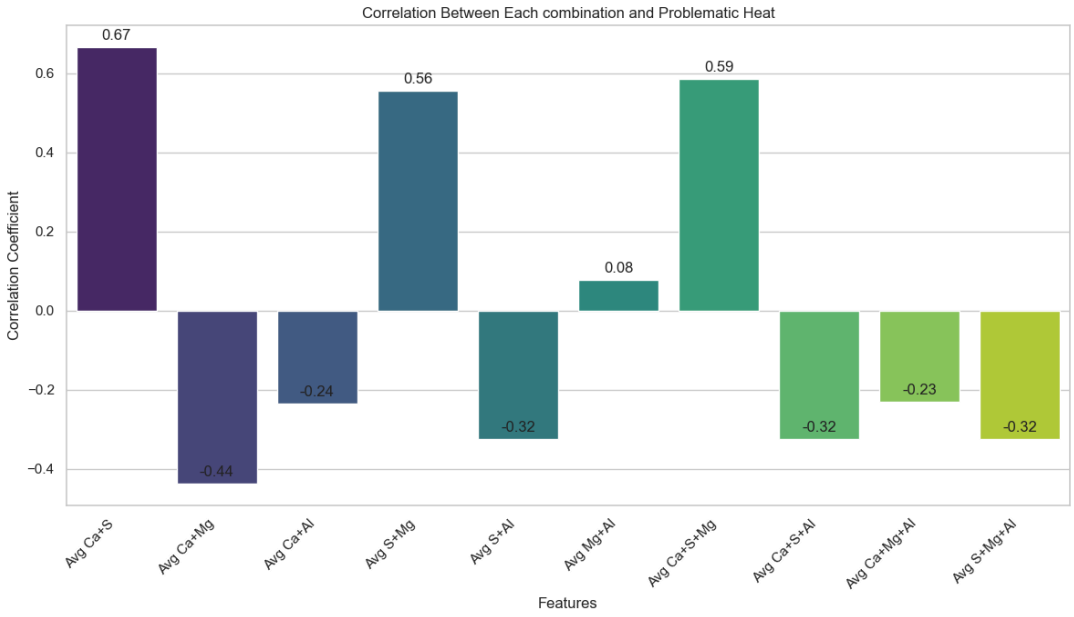

Figure 11. Correlation between each combination and Problematic Heat

5. Limitation

A key limitation of this study is the lack of available data for samples taken after the addition of aluminum (Al) and silicon (Si). While it has been shown that these samples are more effective in distinguishing between problematic and non-problematic heats compared to earlier samples like Sample 66, the limited availability of data after Al-Si addition restricts the depth of analysis. This makes it difficult to fully explore and validate the trends seen in these later-stage samples. Additionally, the scarcity of data for problematic heats in the new process further complicates efforts to identify reliable indicators that can help differentiate problematic and non-problematic heats.

6. Evaluation

I gained valuable insights and skills during this internship, aligning closely with the learning outcomes I initially outlined in my internship plan. In the following section, I will discuss these key outcomes in detail.

6.1. Data Handling and Preprocessing

One of the most amazing aspects of my internship was the experience of handling and preprocessing large datasets. Given the extensive volume of OES-PDA data, automation became essential to efficiently process and prepare the data for analysis. Without automation, managing such a large dataset would have been extremely time-consuming and prone to error. I developed automated methods for data cleaning and organization, applying various preprocessing techniques to handle missing values, inconsistent formats, and outliers. This experience reinforced my understanding of the critical role automation plays in data science, especially when working with large-scale industrial datasets, and allowed me to develop a more structured, efficient approach to data handling. These skills that directly contribute to my overall learning outcomes in data management and preprocessing.

6.2. Statistical Analysis

Another learning outcome from my internship was the application of statistical analysis techniques, particularly in identifying common upper outliers within the OES-PDA dataset.

Initially, I employed the interquartile range (IQR) method to detect outliers. However, I found that this approach did not capture all common upper outliers in the dataset. To address this, I applied the three-sigma rule, which allowed for a more refined identification of extreme upper outliers. Additionally, the choice of correlation analysis methods was based on the characteristics of the dataset, ensuring that relationships between variables were accurately assessed. This experience directly contributed to my learning outcomes in statistical analysis, highlighting the importance of selecting appropriate methods based on data characteristics to achieve reliable results.

6.3. Data Analysis and Pattern Recognition

The third key learning outcome from my internship was developing the ability to analyze data effectively and interpret patterns and trends within the OES-PDA dataset. By examining different samples, I identified trends that offered insights into the behavior of inclusions across various stages of the steelmaking process. This experience significantly enhanced my skills in data analysis, pattern recognition, and the interpretation of complex datasets.

6.4 Programming and data visualzation

The fourth learning outcome of my internship was the significant improvement in my data visualization and programming skills. I performed extensive data visualization to communicate findings effectively, creating clear visuals such as box plots, line charts, and bar charts to highlight trends and differences across process categories. This work enhanced my ability to

translate raw data into meaningful insights. Additionally, my programming skills improved significantly during this internship. I became more proficient in automating data processing workflows, optimizing scripts for handling large datasets, and implementing statistical and visualization libraries to streamline analysis. These enhancements have provided me with a stronger foundation to tackle complex data challenges in the future.

However, I should mention that due to limited data for the new processes and samples taken after aluminum and silicon additions, as well as time constraints, I was unable to develop a machine learning model as initially outlined in my internship plan. Nevertheless, there is still an opportunity to pursue this aspect of the project. If Tata Steel can provide a more comprehensive and analysis-ready dataset, I would be eager to continue this work by developing a machine learning model that could offer deeper insights into inclusion behaviors and quality outcomes. This extension could allow me to further contribute to the project and potentially use it as the foundation for my bachelor's thesis.

7. Contribution Reflection

The internship has given me a lot of valuable experiences. However, I believe I was a valuable intern at Tata Steel as well. During my time there, I was able to provide a fresh perspective from the data side of the project. Although my knowledge of materials and chemistry was limited, my skills in data science were highly valuable to the team. By identifying patterns in the data, I helped the team, who are experts in materials and chemistry, to conduct research and better understand the reasons behind these patterns. I believe my contribution filled a significant gap

within the team, as they required someone proficient in data techniques and programming to enhance their understanding of production processes.

8. Usefullness of knowledge and skills connected to DSS

The knowledge I gained during my internship has been highly relevant and valuable to the core principles of the Data Science program I have been studying. Through handling and analyzing the extensive OES-PDA dataset, I applied essential data science concepts from my coursework, including data cleaning, preprocessing, and statistical analysis, in a real-world industrial context. Working with large datasets, applying automation, and identifying patterns and trends reinforced my understanding of data-driven decision-making and analytics. Additionally, using statistical methods allowed me to enhance my practical skills in statistical analysis. All in all, this internship provided an amazing opportunity to bridge theory and practice, deepening my understanding of how data science techniques can address complex, real-world challenges.

Furthermore, as I plan to pursue a master's degree in data science, this experience has provided invaluable preparation. The practical foundation I built during this internship will be instrumental as I continue my studies and move into a more specialized area within the field of data science.

9. Appendix A: Classification Criteria for Heats

This appendix provides an overview of the criteria used to classify heats as problematic or non-problematic in NGO electrical steel production. For detailed standards and examples, please refer to the *202040416 UTAM data for OES-PDA study_AA* document.

Heats were labeled as problematic or non-problematic based on five criteria:

1. Canoe-shaped defects > 40 cm
2. Canoe-shaped defects > 25 cm
3. Stopper slope
4. Surface defect count by the client
5. SEN images taken after each sequence