Data-Driven Oxide Problem Characterization and Optimization in Float-Zone Silicon Crystal Growth Production

Chen, Tingting

Publication date:
2023

Document Version
Publisher's PDF, also known as Version of record

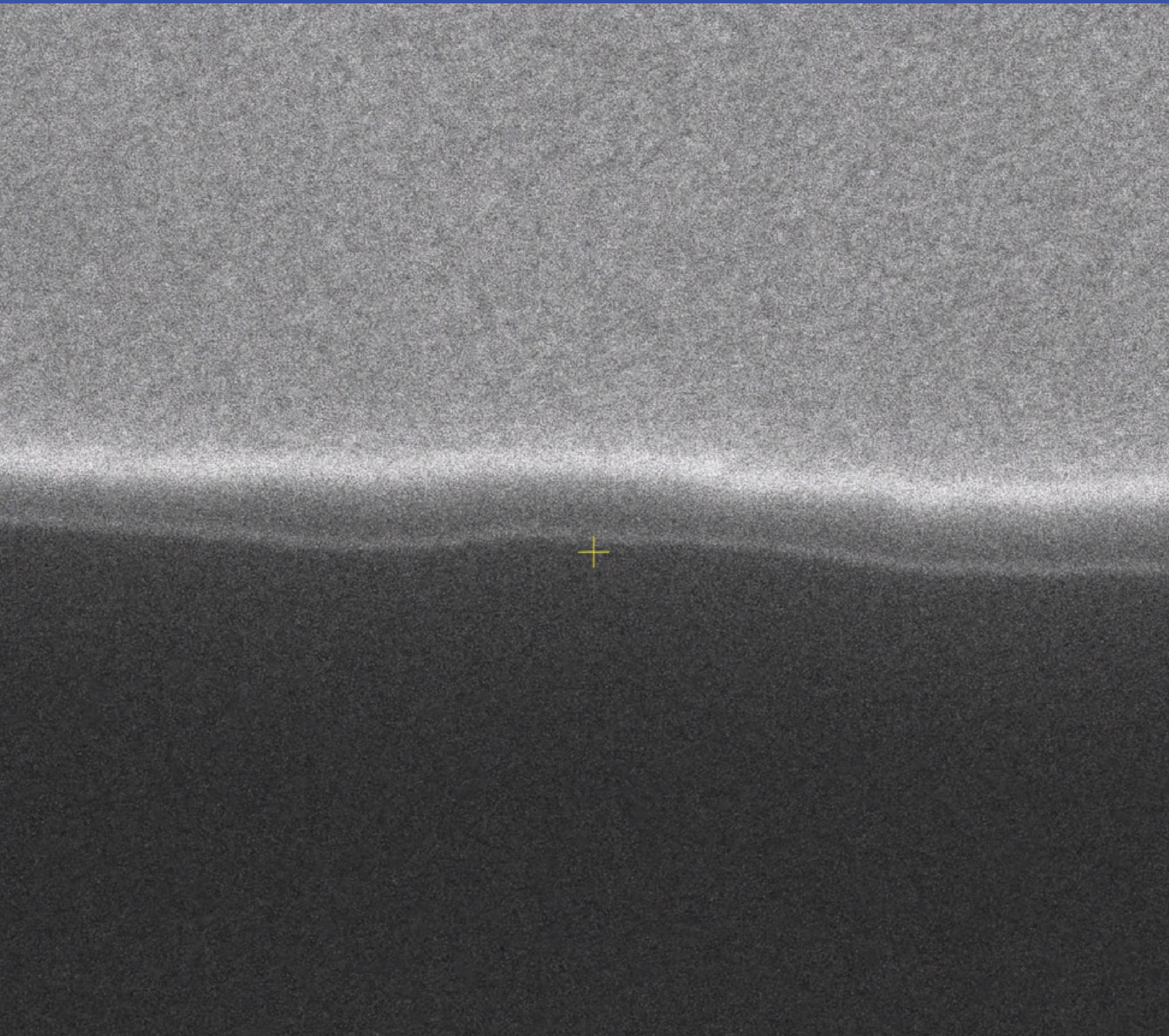
Link back to DTU Orbit

Citation (APA):
Chen, T. (2023). *Data-Driven Oxide Problem Characterization and Optimization in Float-Zone Silicon Crystal Growth Production*. Technical University of Denmark.

Department of Civil and Mechanical Engineering
PhD Thesis

Data-Driven Oxide Problem Characterization and Optimization in Float-Zone Silicon Crystal Growth Production

Tingting Chen



Data-Driven Oxide Problem Characterization and Optimization in Float-Zone Silicon Crystal Growth Production

PhD thesis
Octorber, 2023

By
Tingting Chen
Supervised by
Matteo Calaon, Senior Researcher
Guido Tosello, Associate Professor

Cover photo: Tingting Chen, 2023
Published by: DTU, Department of Civil and Mechanical Engineering, Produktionstorvet, Building 427, 2800 Kgs. Lygnby, Denmark
www.construct.dtu.dk
ISBN: [978-87-7475-782-5]

Preface

This dissertation has been prepared as a requirement to fulfill the Ph.D. degree at the Technical University of Denmark (DTU), Kongens Lyngby, Denmark, in the Department of Civil and Mechanical Engineering (DTU Construct). The work described in this dissertation was carried out from November 2020 to November 2023 under the supervision of Senior Researcher Matteo Calaon and Associate Professor Guido Tosello.

This Ph.D. project is undertaken in the context of the Horizon2020 European Training Network DIGIMAN4.0 (DIGItal MANufacturing Technologies for Zero-defect Industry 4.0 Production, http://www.digiman4-0.mek.dtu.dk/, Project ID: 814225

Tingting Chen

Tingting Chen
Department of Civil and Mechanical Engineering (DTU Construct)
Technical University of Denmark
31 October 2023

Abstract

It is undisputed that silicon wafers have become crucial to our modern lives and world's commercial and military applications. As the key process for the fabrication of silicon wafers, the single-crystal growth process has been driven to increase good-for-order single-crystal silicon yield while keeping costs low. To this end, this dissertation, conducted as part of DIGIMAN4.0 funded by Horizon 2020, the EU Framework Programme for Research and Innovation, collaborating with Topsil Global Wafers, delves into the process optimization in Float-Zone (FZ) crystal growth production, with the aim of increasing yields and reducing costs. The core of the project introduces a unique challenge, an oxide contamination problem happening in the FZ process which is strongly associated with crystal yields, and a systematic investigation of the oxide problem around three topics: what the problem is (know-what), why it happened (Know-Why) and how to address it (Know-How) is conducted.

Firstly, an investigation was carried out regarding the nature of the surface anomaly by material characterization and visual characterization. Specifically, material characterization was achieved by the characterization of the polysilicon surface using Focus Ion Beam Scanning Electron Microscopy (FIB-SEM) and Energy Dispersive Spectroscopy (EDS), which clearly defined that the surface anomaly is associated with enrichment of oxygen and oxygen loss by evaporation. Visual characterization was performed on the FZ images captured from the FZ vision system, which showed that the surface typically appears at the beginning of the cone phase, and the surface anomaly can present in three categories, including the spot, shadow, and ghost curtain and their characteristics and potential impacts were discussed.

Secondly, to enable an efficient recognition of the oxide layer (surface anomaly) for Know-What, an oxide identification based on Deep Learning was developed, which has been shown to be effective in capturing the occurrence of oxide without involving any human being, thus laying a foundation for automatic responses of oxide. In addition, in order to build the trust on the developed oxide identification, Grad-CAM was employed to increase the transparency on the model and to explain why the model makes such a prediction.

Targeted at Know-Why, a thorough investigation of the relationship between oxide and other data sources was carried out by means of Association Rule Mining. The results demonstrated that the oxide is strongly associated with a high moisture level in the FZ chamber during the FZ process.

In order to uncover the source of the high moisture level, a Deep Learning-based multi-modal moisture predictor was established, and a model explainability analysis was conducted on the prediction model to examine its decision-making process. This can provide us with insights into corrective measures.

Finally, an automatic response conceptual framework for mitigating oxide formation was proposed for Know-How to provide constant monitoring and dynamic adaption. The framework integrates the findings from Know-What and Know-Why in the diagnostic strategy for responding to the oxide problem. In addition, the framework considered prognostic strategy, which relies on the capability of Know-When in forecasting the occurrence of the oxide problem, thus allowing the decision maker to take preventive measures to decrease the probability of the oxide occurrence.

In conclusion, with the motivation of increasing crystal yields and reducing costs in the FZ process, this dissertation focuses on process optimization by undertaking a systematic investigation regarding an oxide problem that can affect crystal yields. The investigation offers a comprehensive understanding of the oxide problem from three perspectives: Know-What, Know-Why and Know-How, which sets the stage for future advancement for the field.

Keywords: Float-Zone crystal growth; problem solving; deep learning; root cause analysis; process optimization;

Resumé

Det er ubestridt, at siliciumskiver er blevet afgørende for vores moderne liv og verdens kommercielle og militære anvendelser. Som nøgleprocessen til fremstilling af siliciumwafers, enkeltkrystalvækstprocessen, er derfor blevet drevet til at øge udbyttet af et-krystal silicium, der er godt for bestilling, samtidig med at omkostningerne holdes lave. Til dette formål dykker denne afhandling, udført som en del af DIGIMAN4.0 finansieret af Horizon 2020, EU's rammeprogram for forskning og innovation, i samarbejde med Topsil Global Wafers, ind i procesoptimering i Float-Zone (FZ) krystalvækstproduktion , med det formål at øge udbyttet og reducere omkostningerne. Kernen i projektet introducerer en unik udfordring, et problem med oxidforurening, der sker i FZ-processen, som er stærkt forbundet med krystaludbytte, og en systematisk undersøgelse af oxidproblemet omkring tre emner: hvad problemet er (ved-hvad), hvorfor det skete (Know-Why), og hvordan man adresserer det (Know-How) udføres.

For det første blev der udført en undersøgelse vedrørende arten af overfladeanomali ved materialekarakterisering og visuel karakterisering. Specifikt blev materialekarakterisering opnået ved karakterisering af polysiliciumoverfladen ved hjælp af Focus Ion Beam Scanning Electron Microscopy (FIB-SEM) og Energy Dispersive Spectroscopy (EDS), som klart definerede, at overfladeanomalien er forbundet med berigelse af ilt og opløsning af ilt. Visuel karakterisering blev udført på FZ-billederne taget fra FZ vision-systemet, som viste, at overfladen typisk optræder i begyndelsen af keglefasen, og overfladeanomalien kan forekomme i tre kategorier, herunder plet-, skygge- og spøgelsesgardin og spøgelsesgardin. deres karakteristika og potentielle virkninger blev diskuteret.

For det andet, for at muliggøre en effektiv genkendelse af oxidlaget (overfladeanomali) for Know-What, blev der udviklet en oxididentifikation baseret på Deep Learning, som har vist sig at være effektiv til at fange forekomsten af oxid uden at involvere noget menneske, således lægger et grundlag for automatiske reaktioner af oxid. For at opbygge tilliden til den udviklede oxididentifikation blev Grad-CAM desuden brugt til at øge gennemsigtigheden på modellen og forklare, hvorfor modellen foretager en sådan forudsigelse.

Målrettet Know-Why blev der gennemført en grundig undersøgelse af forholdet mellem oxid og andre datakilder ved hjælp af Association Rule Mining. Resultaterne viste, at oxidet er stærkt forbundet med et højt fugtniveau i FZ-kammeret under FZ-processen.

For at afdække kilden til det høje fugtniveau blev der etableret en Deep Learning-baseret multimodal fugtprædiktor, og en modelforklarlighedsanalyse blev udført på forudsigelsesmodellen for at undersøge dens beslutningsproces. Dette kan give os indsigt i korrigerende foranstaltninger.

Endelig blev der foreslået en begrebsramme for automatisk respons til afbødning af oxiddannelse for Know-How for at give konstant overvågning og dynamisk tilpasning. Rammen integrerer resultaterne fra Know-What og Know-Why i den diagnostiske strategi for at reagere på oxidproblemet. Derudover overvejede rammen en prognostisk strategi, som er afhængig af Know-When's evne til at forudsige forekomsten af oxidproblemet, hvilket giver beslutningstageren mulighed for at træffe forebyggende foranstaltninger for at mindske sandsynligheden for oxidforekomsten.

Afslutningsvis fokuserer denne afhandling med motivationen for at øge krystaludbyttet og reducere omkostningerne i FZ-processen på procesoptimering ved at foretage en systematisk undersøgelse vedrørende et oxidproblem, der kan påvirke krystaludbyttet. Undersøgelsen giver en omfattende forståelse af oxidproblemet fra tre perspektiver: Know-What, Know-Why og Know-How, hvilket sætter scenen for fremtidig avancement for feltet.

Nøgleord: Float-Zone krystalvækst; problemløsning; dyb læring; rodårsagsanalyse; procesoptimering;

Acknowledgement

I would like to express my deepest gratitude to my supervisors, Matteo Calaon and Guido Tosello, for their unwavering support, invaluable guidance, and continuous encouragement throughout this research journey. Their expertise, insightful feedback, and dedication to my academic and personal growth have been instrumental in shaping the direction of this dissertation.

I would like to extend special thanks to my industrial partner Topsil Globalwafers for their invaluable contributions, especially Christian Hindrichsen, Nico Werner, Lars Conrad-Hansen, Michael Jensen, Monica A. Lund and Andreas T. Hofer. Their provision of data and necessary information as well as their support have been pivotal to the success of this research. Their practical insights and real-world perspectives have enriched the context of my work.

I am grateful to my colleagues and friends for their support during both challenging and joyful times. Their encouragement and shared experiences have made this journey memorable and meaningful.

My heartfelt thanks go to my family for their unwavering belief in me and their unconditional love. Your constant encouragement and sacrifices have been my driving force.

Lastly, I am thankful to all those whose names may not appear here but whose contributions have been equally important in shaping this work. Your collective efforts have left an indelible mark on my academic pursuit.

This dissertation would not have been possible without the support, guidance, and collective efforts of all those mentioned and unmentioned above. Thank you.

Tingting Chen
October, 2023

Contents

List of Figures

List of Tables

List of Acronyms

Acronym	Description
AI	Artificial Intelligence
ICs	Integrated Circuits
CZ	Czochralski crystal growth
FZ	Float-Zone Crystal growth
IGBTs	Insulated gate bipolar transistors
ML	Machine Learning
DL	Deep Learning
SEM	Scanning Electron Microscope
EDX	Energy Dispersive X-ray Spectroscopy
FIB-SEM	Focus Ion Beam Scanning Electron Microscopy
MLC	Multi-label classification
CNNs	Convolutional Neural Networks
RNNs	Recurrent Neural Networks
GCNs	Graph Convolutional Networks
DNNs	Deep neural networks
CAM	Class Activation Map
BCE	Binary Cross Entropy
FL	Focal loss
ASL	Asymmetric loss
PS	Powerset
BR	Binary Relevance
ARM	Association Rule Mining
XAI	Explainable Artificial Intelligence
SE	queeze-and-excitation
GAF	Gramian Angular Field
MTF	Markov Transition Field
LSTM	Long Short-Term Memory
SHAP	SHapley Additive exPlanations
MSE	Mean squared error
RMSE	Root mean squared error

CHAPTER 1

Background and objectives

"In the middle of every difficulty lies opportunity." - Albert Einstein

1.1 Introduction

It is undisputed that semiconductors have become crucial to our modern life and to the commercial and military applications of the world. Semiconductors have been instrumental in the rapid evolution of enormous modern technologies such as communication devices, Internet of Things (IoTs), robotics, and Artificial Intelligence (AI), to name a few. These small but powerful components are the cornerstones of modern electronics and have had a profound impact on virtually every aspect of our lives. From communication and computing to transportation and healthcare, semiconductors have enabled the manipulation and control of electrical signals, forming the basis for integrated circuits (ICs), microprocessors, memory devices, and a wide range of other electronic components that power our interconnected world. As reported in the Historical Billings Report of the Worldwide Semiconductor Trade Statistics group shown in Figure 1.1, worldwide semiconductor sales revenue has been rapidly increasing since 1986. In 2022, worldwide semiconductor sales topped $574 billion dollars [1].

The semiconductor is often referred to as a material with a crystalline structure and its electrical conductivity falls between a conductor and an insulator. A semiconductor is distinct from a conductor in that it only allows electrons to flow freely when they have attained a certain energy level [2]. Examples of semiconductor materials include silicon, germanium and gallium arsendie [3]. Among all semiconductor materials, silicon is the most popular choice of semiconductor material, making up more than 90% of all semiconductor and solar cell wafer production due to its abundance and the extensive knowledge of its processing [4].

Before the silicon becomes the heart of the modern electronics, it needs to go through several steps, as shown in Figure 1.2. Modern electronics such as ICs and microelectronic devices are mainly built on a canvas, a silicon wafer, which is a thin slice of crystalline silicon with high purity. Figure. 1.3 shows a silicon ingot and cut silicon wafers. The key process for the fabrication of silicon wafers is the single-crystal growth process, which produces a single-crystal silicon ingot that determines the properties of the wafer [7]. Recently, silicon wafers have gained increasing

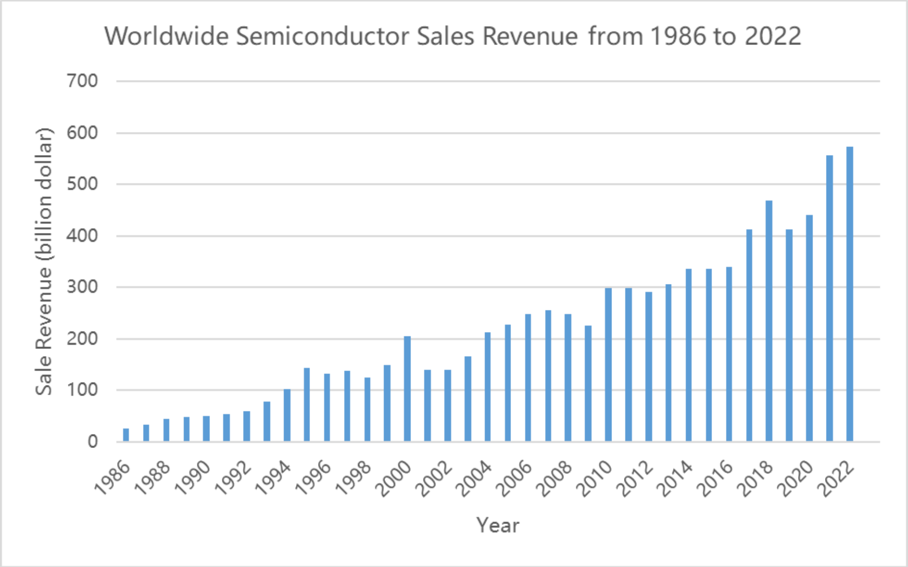


Figure 1.1: Worldwide Semiconductor Sales Revenue from 1986 to 2022 (in billion US dollar), as reported in [1].

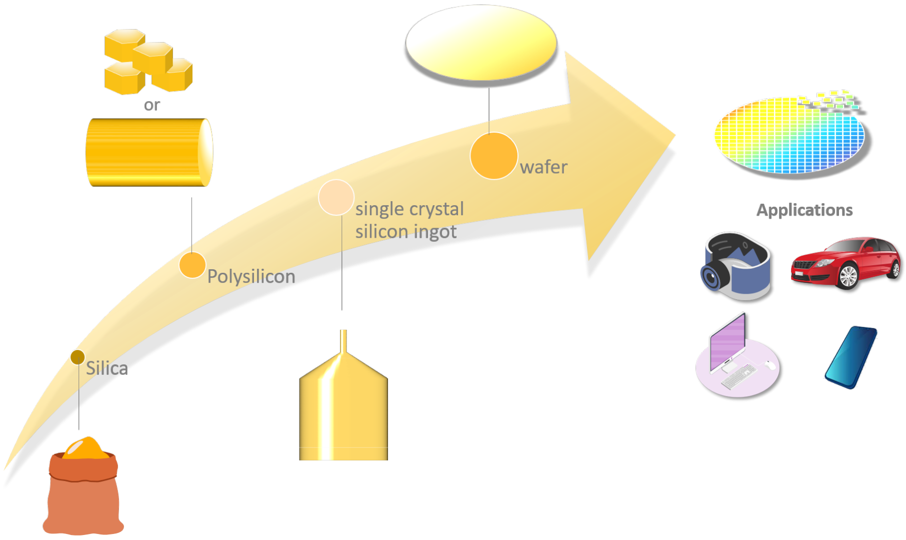
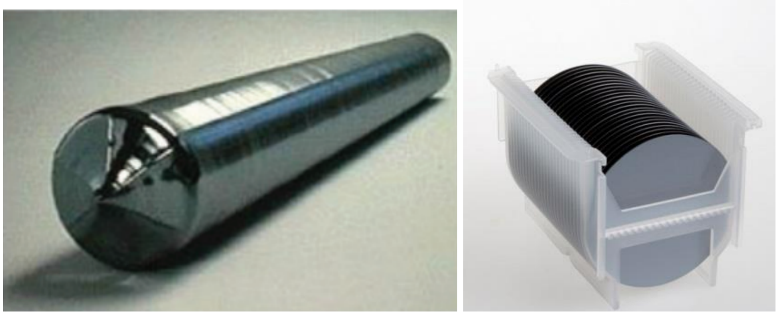


Figure 1.2: The process flow of silicon from raw material to semiconductor applications, adapted from [5, 6].

attention due to the US-China tech war [8, 9], with semiconductors being a key battleground. The United States has imposed export control on high-end semiconductor fabrication equipment in order to obstruct China's attempts to become a major player in high-tech industries (robotics, artificial intelligence, etc.). [10]. This has caused a ripple effect throughout the global supply chain [9]. Furthermore, the Covid-19 pandemic has caused a surge in demand and sales, leading to a shortage of the main semiconductor wafers [11]. The lack of wafer supply has had an impact on a variety of sectors, including those involved in the production of advanced computer chips, military and transportation [12]. Consequently, the current market trend necessitates an increase of good-for-order single-crystal silicon yield while keeping costs low [13, 7], in order to address supply chain vulnerabilities and improve competitiveness.



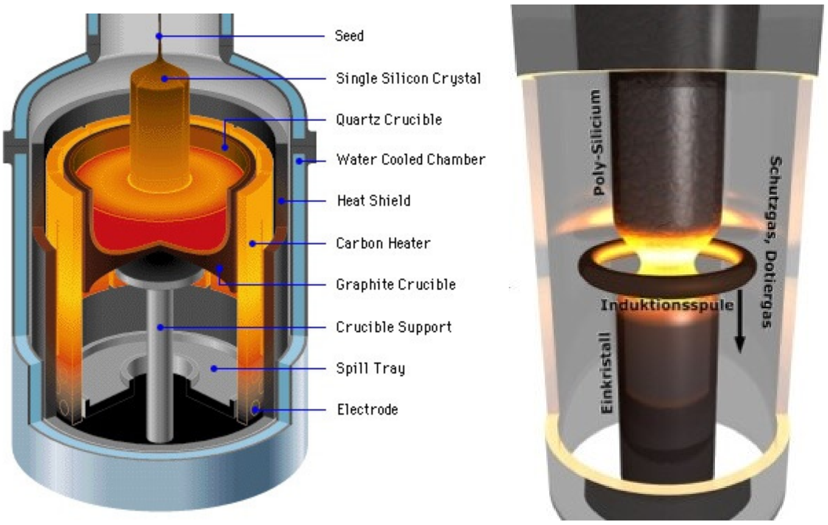
(a) A silicon ingot

(b) The cut silicon wafers

Figure 1.3: Pictures of silicon ingot and wafers. (a) A silicon ingot [14]. (b) A box of cut silicon wafers [15].

There are two main techniques for producing single-crystal silicon from polysilicon: Czochralski (CZ) crystal growth and Float-Zone (FZ) crystal growth. Both crystal growth methods involve dipping a silicon seed crystal mounted on a movable puller into a polysilicon melt in an inert gas atmosphere [16, 17]. Then an oriented seed crystal begins to grow from the melt due to the crystallization of silicon atoms [16]. The difference between the two methods lies in the process of generating polysilicon melt and crystal growth. In the CZ process, a single-crystal ingot is pulled from the polysilicon melt kept in a quartz crucible, while in the FZ process, the polysilicon rod is melted at one end using an induction heater, resulting in a melt zone where the seed crystal grows. Figure 1.4 demonstrates the schematic setup of these two processes. The difference in the growing crystal between the CZ process and the FZ process can be seen in Figure 1.5.

The CZ process is used predominately in the crystal growth of silicon production, as it allows for large crystal diameters (up to 450 mm [17]), resulting in lower production cost per wafer [17]. However, it has a major disadvantage: the high incorporation of oxygen from the melt in a silica crucible, leading to the generation of oxide precipitates and thermal donors [18]. The typical



(a) Czochralski crystal growth

(b) Float-Zone crystal growth

Figure 1.4: The schematic setup of the CZ and the FZ machine [22].

oxygen concentrations are in the range of $3 \times 10^{17} \sim 9 \times 10^{17}$ $atoms/cm^{-3}$ [16]. This limits the application of CZ silicon in some important electronic devices where the oxygen requirement is significant [17]. In contrast, the FZ process produces a higher purity silicon crystal with much lower concentrations of impurities, particularly a lower oxygen content (below 5×10^{16} $atoms/cm^{-3}$ [19]) due to the absence of the crucible. Therefore, FZ silicon crystals are usually used in high-efficiency solar cells and high-power devices where purity is essential. For instance, insulated gate bipolar transistors (IGBTs) that are vital components in electric vehicles for their high-speed switching characteristics [18], require a minimum oxygen oxygen concentration to avoid as-grown defects, oxide precipitates and thermal donors [20].

Nevertheless, the FZ process appears to be less competitive in market acceptance compared to the CZ process because of its several drawbacks. First, the diameter of the FZ silicon crystal is currently restricted to 200 mm due to high-voltage breakthroughs at the "pancake inductor" [21, 17] and the availability of suitable polysilicon rods [13]. Furthermore, the production cost of the FZ process is much higher than that of the CZ process, which is attributed to the higher cost of the feedstock material used in the FZ process [13]. The FZ process necessitates polysilicon to be in a precise cylindrical shape with a smooth surface and no cracks, which causes the costs associated with polysilicon feed rods to be more than half of the cost of ownership [13]. Therefore, increasing the crystal yield is considerably important for the FZ process.

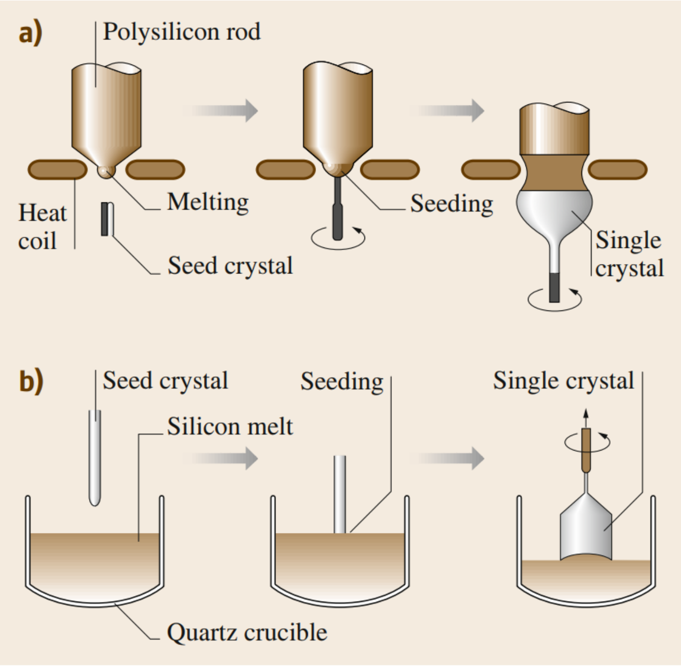


Figure 1.5: Principles of (a) FZ process and (b) CZ process [23].

1.2 Problem Definition

This research project is based on the FZ process with the collaboration of Topsil GlobalWafers A/S, a global supplier of ultra-pure silicon to the semiconductor industry. As mentioned above, the recent market trend has forced silicon suppliers to accelerate their efforts to increase good-for-order yield and reduce costs. The cost of ownership of crystal growth processes is predominantly determined by yield, as the majority of the production cost is associated with the raw materials, polysilicon [13]. This is more evident in the FZ process since the cost of the polysilicon rod accounts for more than 50% of the production costs, which is much more expensive than the chunk polysilicon used for the CZ process [13]. Consequently, compared to the CZ process, the crystal yield is even more critical for the FZ process [13]. Therefore, this project is initiated with the motivation of increasing the FZ crystal yield by identifying and correcting the factors that lead to lower yields, such as impurities or other process variables.

As shown in Figure 1.6, an FZ machine consists of a pressurized chamber, which is equipped with an inductive coil and two movable holders for grasping polysilicon and seed crystal, respectively. There is an optical access to the FZ process given by a quartz window through which a FZ vision system is able to acquire FZ images and carry out image analysis, for instance geometrical measurements for dynamically controlling the FZ process. In the FZ process, the inductive coil begins with heating the bottom of the polysilicon rod into a molten zone, where the seed crystal begins crystal growth, and then the molten zone would slowly move along the polysilicon rod by shifting the polysilicon rod downward, resulting in silicon ingots. A FZ process cycle mainly consists of six phases: 1) melt drop phase, 2) feed tip phase, 3) neck phase, 4) cone phase, 5) cylinder phase, and 6) closing phase [24], as seen in Figure 1.7.

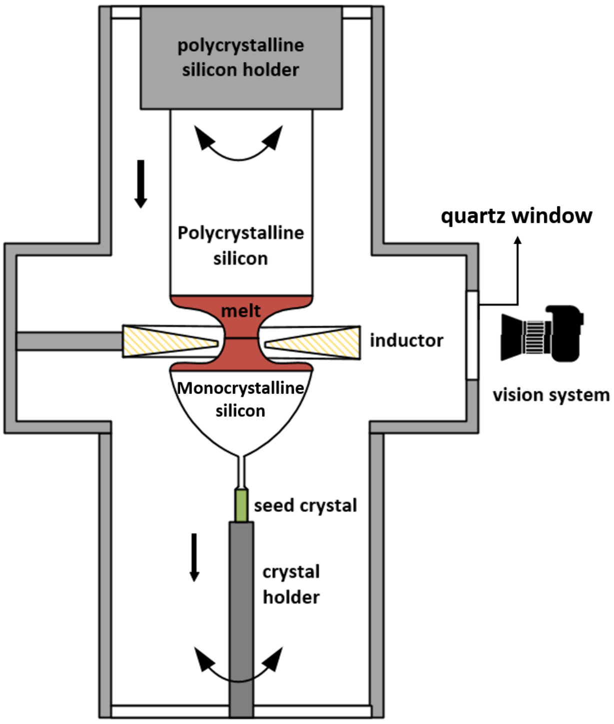


Figure 1.6: The detailed schematic of the FZ Machine [25]. Within the FZ chamber, the polysilicon and seed crystal are held by the polysilicon holder and the crystal holder, respectively. During the FZ process, the polysilicon and seed crystal are moved downwards while they are rotated in opposite directions to stabilize the molten flow. At the same time, a vision system is used to observe the FZ process through a quartz window.

The polysilicon rod used for FZ is required to have a smooth surface and homogeneous thermal and electrical properties during the FZ process, ensuring a stable melting process [17, 13]. In addition, the polysilicon rod is supposed to be heated homogeneously, and its surface shown in the vision system is supposed to be homogeneous as well. However, during the process, anomalies might occur on the polysilicon surface, which is reflected in FZ images captured from the vision system: an appearance of a relatively high-contrast region on the polysilicon surface. The difference between normal polysilicon surface and abnormal surface can be seen in Figure 1.8. The anomaly on the surface of the polysilicon is suspected to be caused by oxidation. The presence of the anomaly may disrupt the stable melting behavior, which being a disturbance may affect the performance of the FZ automatic control system which assumes the identical processing condi-

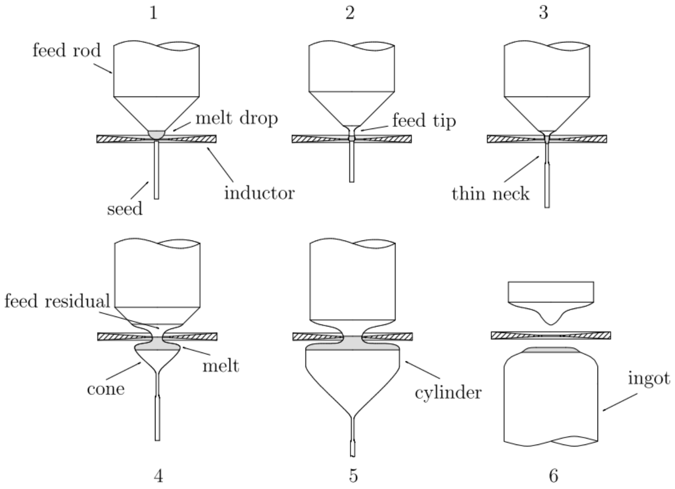
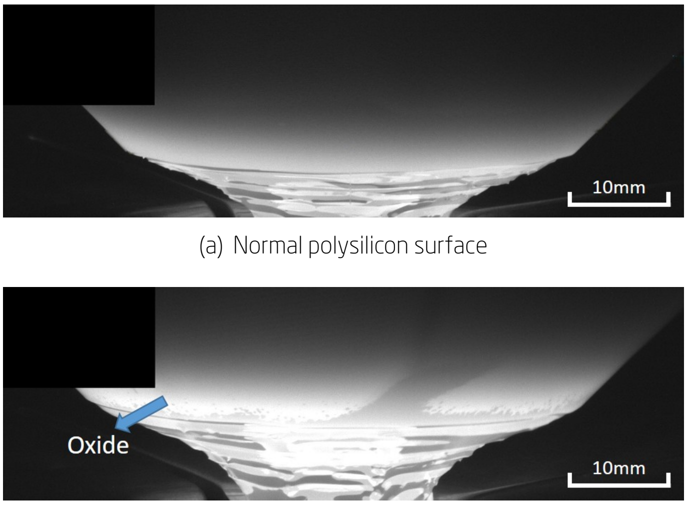


Figure 1.7: Phases of the Floating Zone process. Producing the melt drop (1), forming the feed tip (2), creating the thin neck (3), making the cone (4), growing the cylinder (5) and closing the crystal (6). Solid material in white, melt in gray [24].

tion. Besides, the presence of the anomaly indicates that the ability of the FZ machine to provide a protection atmosphere for the process has been degrading. Last but not least, if the anomaly is associated with the oxygen, it would act as a contamination to threaten the product quality, which would consequently affect good-for-order yield, since the product would become scrap if it does not meet the customer's requirements. Considering that the FZ process is a batch process lasting for 10-20 hours, the anomaly may potentially lead to a waste of production resources, including time, energy and, in particular, the expensive polysilicon rod. Therefore, considering the above reasons, it is necessary to investigate the anomaly in order to enhance the process efficiency and reliability.

1.3 Research Questions

The FZ process enables the production of ultra-pure silicon with low oxygen content due to the preclusion of contamination from the crucible. However, in return, customers pose a strict requirement on oxygen concentration in the final products for the FZ process. The product that does not meet the customer's requirements would become a scrap, which would further affect crystal yield. Therefore, the FZ crystal suppliers need to carefully control the oxygen content during the FZ process. However, the anomaly that appears on the polysilicon surface might raise alarms. On the one hand, the anomaly, possibly being a contamination to the FZ process, may pose a risk of negatively affecting the crystal field, thus further increasing the production cost. On the other hand, the presence of the anomaly indicates the degradation of FZ machine performance


[redacted]
[redacted]

(a) Normal polysilicon surface

(b) Polysilicon surface with the oxide

Figure 1.8: The polysilicon images in different FZ crystal growth productions (the confidential information in the upper left corner is concealed using a black rectangle): (a) polysilicon surface with homogeneous intensity under normal condition (b) polysilicon surface with a high-contrast appearance under abnormal condition [25].

in keeping FZ chamber sealed, thus a maintenance might be needed.

This Ph.D. project was undertaken in the context of DIGIMAN4.0 project ("DIGItal MANufacturing Technologies for Zero-defect Industry 4.0 Production", https://www.digiman4-0.mek.dtu.dk/). DIGIMAN4.0 is a European Training Network supported by Horizon 2020, the EU Framework Programme for Research and Innovation (Project ID: 814225). The DIGIMAN4.0 project consists of 15 ESRs, 6 academic partners and 20 industrial parners. The vision of DIGIMAN4.0 is to achieve zero-defect industry 4.0 production with the aid of digital manufacturing technologies. The 3 main technological cornerstones of DIGIMAN4.0 are:

- Precision manufacturing processes development.
- Digital manufacturing technologies and their integration into process chains.
- Industry 4.0 pilot production for zero-defect manufacturing.

Based on the background and motivation described above, the focus of this dissertation will tackle the following research questions:

1. What are the nature and characteristics of the surface anomaly and its potential impacts on the product quality and the associated industrial challenges?
2. What are the key factors that influence the formation of the surface anomaly in the FZ process and the practical actions to mitigate the surface anomaly?

3. How can the knowledge gained be translated into practical guidelines and control strategies to proactively manage the surface anomaly?

1.4 State-of-the-Art

1.4.1 Oxygen in Crystal Growth

Despite the detrimental effects of oxygen on the FZ process, the investigation of oxygen within the FZ process has been relatively underexplored in the existing literature. Currently, there is only one study that looks into the FZ process monitoring of potential oxygen in relation to the surface characteristics of polysilicon [26]. This research, however, only focuses on recognizing the anomaly on the surface of polysilicon without thoroughly defining the anomaly and its cause and effect for the FZ process. On the contrary, there are a handful of research efforts dedicated to explaining the potential sources and impacts of oxygen in the CZ process. Considering the commonalities of the FZ process and the CZ process, the insights on the oxygen problem in the CZ process may shed light on the FZ process.

In the case of oxygen in the CZ process, oxygen is introduced mainly from the silica-made crucible (SiO_2). As soon as the polysilicon becomes molten in the silica crucible, oxygen is transferred to the polysilicon melt from the crucible [16]. It is yet to be determined whether the oxygen present in the molten polysilicon should be considered as dissolved SiO or simply as dissolved O [16]. Nevetheless, in any case, it is known that oxygen would evaporate from the silicon melt through the reaction of silica and the silicon melt:

$$\mathrm{SiO_2 + Si \longrightarrow 2\,SiO}$$

resulting in SiO vapor with the vapour pressure of 12 mbar at the melting point of silicon [27].

More than 99% of the dissolved oxygen is removed from the melt in the form of SiO vapor [27], which would be handled by the gas flow inside the growth chamber [13]. The rest of the oxygen is incorporated into the growing crystal. As the growing crystal cools down, the solubility of oxygen in the silicon lattice decreases quickly. This leads to an increase in oxygen supersaturation, which in turn causes the formation of oxygen precipitates and thermal donors [16]. The level of oxygen is typically within the range of $3 \times 10^{17} \sim 9 \times 10^{17}\ atoms/cm^{-3}$ in CZ silicon wafers depending on the specific application [16], which is considered acceptable for the fabrication of low-power devices such as integrated circuits. However, oxygen levels of this magnitude are usually seen as an undesirable impurity in high-power devices. It is estimated that oxygen precipitation begins when the oxygen concentration reaches $3 \times 10^{17}\ atoms/cm^{-3}$ [28], and this can have a negative effect on the minority carrier lifetime [29], which is a key factor in the performance of high-power devices such as solar cells. The minority carrier lifetime represents the average time that a minority carrier remains mobile before recombining with a majority carrier [30, 31]. A shortened minority carrier lifetime indicating a higher rate of recombination would impede efficient carrier transport and energy conversion, resulting in performance degradation of high-power devices. This explains why these high-power devices are preferably produced by the oxygen-lean FZ process instead of the CZ process. However, this on the other hand implies that reducing the oxygen level is much more critical in the FZ process than in the CZ process.

1.4.2 Problem-Solving in Manufacturing

When an issue arises in production that leads to the failure to reach the production target, it is necessary to have a process to thoroughly examine the problem and generate a solution in a systematic way. Several systematic methods have been developed, such as Kepner-Tregoe [32], Plan-DO-Check-Act (PDCA, seen in Figure 1.9) [33], Eight Disciplines (8D) [34], to name a few. Kepner-Tregoe method is a practical and structured methodology used for identifying problems, making decisions, and evaluating potential risks [32]. In the Kepner-Tregoe method, there are four distinct steps including situation appraisal, problem analysis, decision analysis, and potential problem analysis [35]. The 8D method, originating in Ford Motor Company in the early 1990's [36], provides a structured process to discover solutions and improve continuously. The 8D method comprises eight steps: 1) forming a team, (2) description of the problem, (3) definition of interim containment actions, (4) root cause analysis, (5) definition of potential corrective actions and effectiveness verification, (6) implementation of corrective actions, (7) definition of preventive measures to prevent recurrence and (8) congratulations to the team [37]. These systematic methods offer a well-defined structure to facilitate the improvement and discovery of solutions.

However, these methods usually require extensive expert knowledge and experience of the team members. Besides, as digitization advances in industries, there is a large abundance of available data and information, making it impossible to analyze using traditional methods where the ability to solve problems mainly depends on human beings. Camarillo et. al [36] tried to address the first problem of lacking enough knowledge and experience of some team members, by proposing a software system combining classic problem-solving methods with Case-Based Reasoning (CBR) with a Product Lifecycle Management (PLM) system. However, the problem of big data produced in production still remains.

The widespread emergence of digital technologies and the advance of computing power have been gradually leading to a new generation of networked, information-based technologies, data analytics and predictive modeling [39]. The growth in the amount and size of data generated

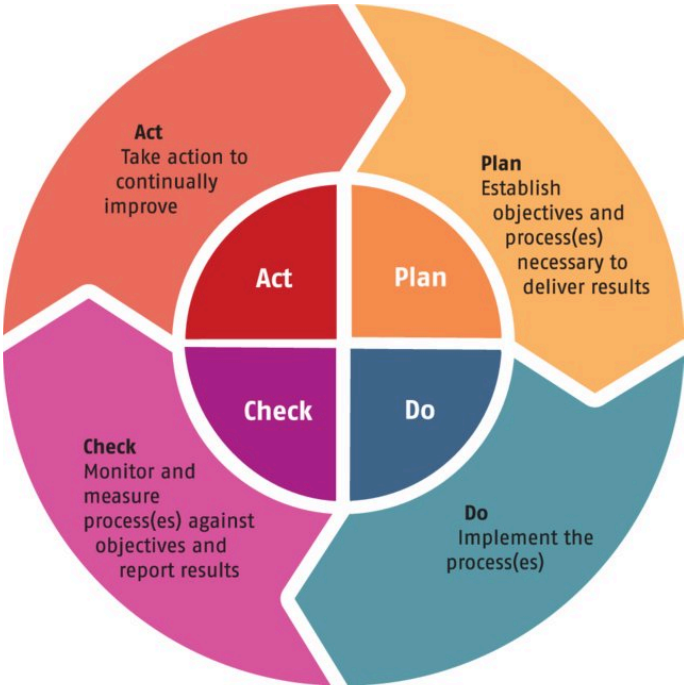

Act
Take action to continually improve
Plan
Establish objectives and process(es) necessary to deliver results
Act
Plan
Check
Do
Check
Monitor and measure process(es) against objectives and report results
Do
Implement the process(es)

Figure 1.9: The PDCA Cycle for continuous improvement. [38]

in the manufacturing sector has highlighted the importance of data analytics in the problem-solving process. Consequently, Machine Learning (ML) which possesses the capacity to process large amounts of data automatically and derive meaningful insights from data, is gaining more and more attention in the manufacturing industries. These advances in ML are shifting the traditional manufacturing era into the smart manufacturing era of Industry 4.0 [40].

The analytical capabilities in manufacturing enabled by digitalization and machine learning can be categorized into four stages for different types of analytic problem: descriptive analytics (Know-What), diagnostic analytics (Know-Why), predictive analytics (Know-When) and prescriptive analytics (Know-How) [41, 42, 43]. Know-What aims to summarize the current states of machines, processes or production systems, which can help in rapid decision-making. For instance, typical examples of Know-What in manufacturing are defect detection in quality control [44, 45], fault detection in process/machine monitoring [46, 47], or soft sensor modelling [48, 49]. Know-Why is generally supported by the outputs from Know-What. Know-Why aims to identify inner patterns from historical data and thus discovering the reasons why something is happening [50]. On the one hand, Know-Why can indicate most important factors for understanding Know-What. On the other hand, Know-Why is the prerequisite for Know-When since the reliability of predictions is heavily dependent on the quality of casual inference. Know-When, built on Know-Why, involves timely predictions of the happening of events or predictions of key variables based on historical data, so that decision makers can take actions in early stages. For instance, Know-When in manufacturing includes quality prediction based on relevant variables [51, 52], or predictive maintenance by detecting incipient anomalies before breakdown [53, 54] or predicting the Remaining Useful Life (RUL) [55]. Know-How, on the foundation of Know-When, can provide recommendations for adapting the expected disturbances or achieving self-optimization. Examples in manufacturing include prediction-based process control [52, 56], scheduling of predictive maintenance tasks [57, 58], dynamic scheduling in flexible production [59, 60], inventory control [58]. Four-know analytics can be seen in Figure 1.10. The four stages of advanced analytics are actually in line with the concept of conventional problem-solving methods, with the difference lying in how information is derived from the data. Rather than relying on human expertise, Four-Know analytics emphasizes data-driven insights and predictive capabilities.

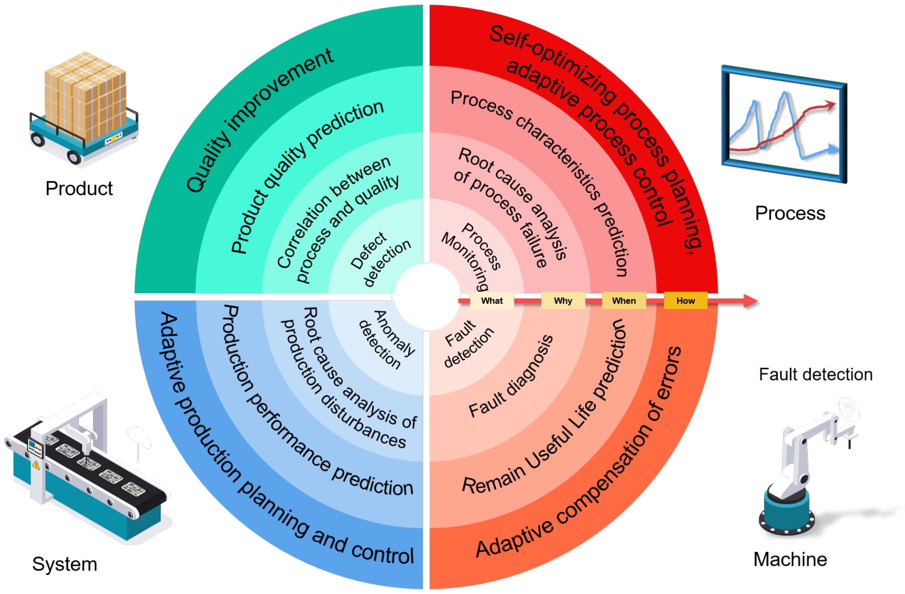

Quality improvement
Product quality prediction
Correlation between process and quality
Defect detection
Self-optimizing process planning, adaptive process control
Process characteristics prediction
Root cause analysis of process failure
Process Monitoring
What
Why
When
How
Anomaly detection
Root cause analysis of production disturbances
Production performance prediction
Adaptive production planning and control
Fault detection
Fault diagnosis
Remain Useful Life prediction
Adaptive compensation of errors
Product
Process
Fault detection
System
Machine

Figure 1.10: ML applications in Four-Level and Four-Know [43]. Four-Know from Know-What to Know-How are demonstrated in four concentric circles from inner circle to outer circle, respectively. Each circle is divided into four quarters according to Four-Level.

Inspired by classic problem solving methods and four stages of data analysis, the study of the surface anomaly in the FZ process can comprise the following steps as seen in Figure 1.11:

- Defining the surface anomaly on the polysilicon surface. The first step is to investigate the nature of the surface anomaly and its potential impacts, which can help assess the significance of the problem.

- Identifying the surface anomaly through data analysis. Following the definition of the problem is the identification of the problem, in order to understand what is happening inside the chamber across different process conditions. In this step, the associated analysis in Know-What can be employed to take advantage of the abundance of data present in the FZ process.

- Performing root cause analysis and consequent decision analysis. Root cause analysis is necessary to understand the fundamental triggers of the problem, which can shed light on effective solutions to mitigate or address the problem. This can be turned to Know-Why analysis enabled by ML to discover the associated factors to the problem.
- Responding automatically to the surface anomaly for continuous improvement. To continuously improve the process and to efficiently prevent recurrence of the problem, it is essential to proactively manage the problem, which requires the integration of the insights gained in previous steps.

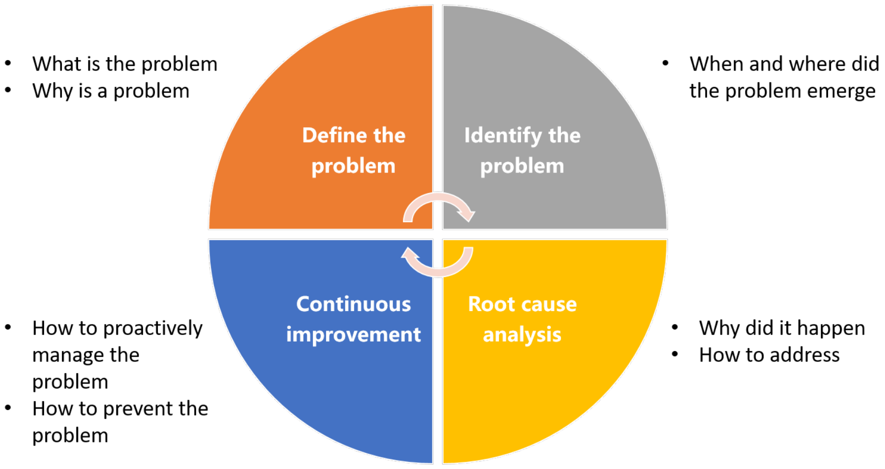


Figure 1.11: Problem-solving pipeline for the surface anomaly on the polysilicon surface in the FZ process.

1.5 Thesis Outline

Figure 7.1 provides an overview of the structure and organization of this thesis following the problem solving framework: Know-What, Know-Why and Know-How. The research presented in this thesis is divided into seven chapters, five of which address specific aspects of the research problem. The following is a brief outline of the thesis.

1. **Chapter 1:** This chapter serves as an introduction to the research problem, providing the background, motivation, and context for the study. It outlines the research problem, objec-

tives, and significance. In addition, this chapter presents an overview of the state-of-the-art methodology, laying the foundation for the research problems addressed in subsequent chapters.

2. **Chapter 2:** In this chapter, a clear definition of the surface anomaly on polysilicon from a physical point of view is established. This is achieved by the characterization of the polysilicon surface using Focus Ion Beam Scanning Electron Microscopy (FIB-SEM) and Energy Dispersive Spectroscopy (EDS). The information uncovered would be used for the subsequent analysis.

3. **Chapter 3:** This chapter presents the research methodology of oxide classification based on images in order to address the Know-What, thus laying a foundation for the subsequent research. It explains the definition of the problem in oxide classification, data sources, and data analysis techniques employed with the aid of machine learning. In the end of the chapter, the experiment results are presented and the findings interpreted from the results are discussed. This chapter also addresses the simplest root cause level: location-time.

4. **Chapter 4:** In this chapter, a thorough investigation of the relationship between oxide and other data sources is carried out by analyzing the merged data from the oxide classes and other data sources with the aid of association rule mining. This is also equivalent to the solution of Factors => Problem, and the root cause level presented here is physical cause. The uncovered the cause would be further investigated in the following chapter by the solution of Factors => Root cause.

5. **Chapter 5:** Based on the finding from the previous root cause analysis, this chapter aims to discover further root cause by building a regression predictor with using process data and control data as the input and the physical cause from the previous findings as the output. In addition, Explainable Artificial Intelligence (XAI) is used to indicate the important features that contribute the most to the predictions, which can offer us insights into the corrective actions to mitigate the oxide problem.

6. **Chapter 6:** To proactively manage the oxide problem, an automatic response conceptual framework is developed for continuously improvement of the process. This framework considers both diagnostic and prognostic approaches, incorporating oxide detection and the discoveries from Know-Why.

7. **Chapter 7:** This final chapter summarizes the primary discoveries and contributions of the study. It also outlines the practical implications of the research and proposes potential paths for further exploration in the field.

CHAPTER 2

SURFACE ANOMALY CHARACTERIZATION

""There are no eternal facts, as there are no absolute truths." - Friedrich Nietzsche

2.1 Motivation

Though it has been revealed before the anomaly shown on the polysilicon surface is possibly due to oxidation, there is still no clear definition of its characteristics. It is still necessary to conduct surface anomaly characterization to gain insight into the subsequent root-cause analysis and the problem solving process. Therefore, in this chapter, we provide a thorough characterization of the surface anomaly shown on the polysilicon surface from the aspects of appearance and element analysis.

The growth of a dislocation-free FZ silicon crystal is a complex process, particularly when it comes to large-sized crystals [17]. FZ runs are sensitive to the growth environment, ground vibrations, or structural anomalies of the polysilicon feed rod, which may potentially lead to FZ runs failure [17]. Hence, a lot of efforts have been put in the polysilicon feed rods and FZ machines in order to ensure the production of dislocation-free FZ ingots. There are strict requirements for the quality of polysilicon rods, as the generation of dislocations during growth can be heavily influenced by the quality of polysilicon rods [13]. To ensure homogeneous heat dissipation in the axial direction of the polysilicon rod and stable cylindrical growth of the crystal, the polysilicon rod must have homogeneous thermal and electrical properties, free of anomalies such as cavities and cracks [17, 13]. Therefore, polysilicon rods have to be ground to obtain a precise cylindrical shape with a smooth surface [13]. Additionally, in order to avoid the recrystallization of the FZ process due to the oxidation of the silicon, the polysilicon rods are normally required to be carefully cleaned with a mixture of nitric and hydrofluoric acid and then rinsed with ultra-pure water [17, 61]. Besides, significant effort has also been put into the FZ machine to guarantee consistent process performance. During the FZ process, the growth chamber is pressurized and filled with high purity argon as a protective gas, thus reducing the risk of oxidation and contamination [17].

Nevertheless, though with high-quality control on the polysilicon rod and the control on the chamber ambient, anomalies may still occasionally occur. During the FZ process, polysilicon is supposed to be heated evenly [17]. Consequently, its surface shown in the image captured from the FZ vision system is also supposed to be with homogeneous intensity. However, during the process, anomalies may occur on the polysilicon surface, which is reflected in the FZ images taken from the FZ vision system: an appearance of a relatively high-contrast region on the polysilicon surface, as shown in Figure 2.1.

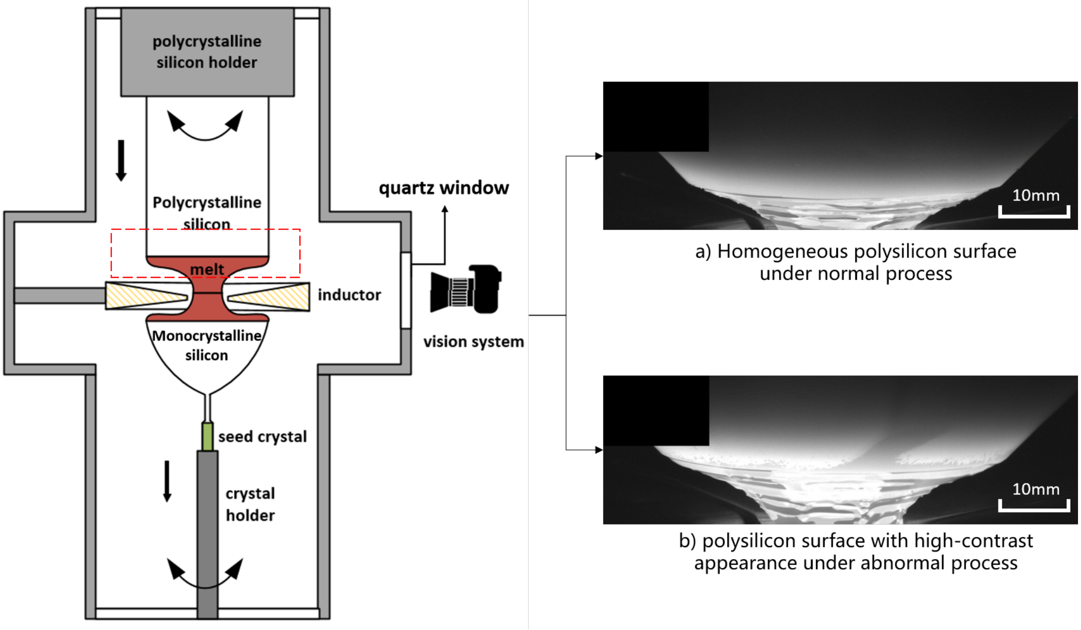

polycrystalline silicon holder
Polycrystalline silicon
melt
inductor
Monocrystalline silicon
seed crystal
crystal holder
quartz window
vision system
10mm
a) Homogeneous polysilicon surface under normal process
10mm
b) polysilicon surface with high-contrast appearance under abnormal process

Figure 2.1: Images of polysilicon taken from the FZ vision system during different FZ crystal growth processes are shown.

As a potential disturbance, the surface anomaly raises alarms to the FZ runs in the aspects of product, process, and machine.

- Product. Surface anomalies may introduce contamination, thus threatening product quality. Substandard parts that fail to meet customer expectation are directly considered as scraps. Considering that the FZ process is a batch process lasting for 10-20 hours, this poses the risk of wasting production resources, including time, energy, etc., which would add to production costs and cause delays in meeting client deadlines.

- Process. The FZ process is normally under automatic control with a model-based control system, which relies on a mathematical model that describes the relationship between input variables and output variables [17]. The mathematical model is developed based on the assumption of consistent and identical processing conditions [17]. Therefore, any unmodeled deviation from these conditions can inherently make the model less reliable and robust [62]. The presence of surface anomaly has altered the thermal properties of the polysilicon, which may pose a deviation in melting behavior, further affecting the control system performance.

- Machine. The presence of a surface anomaly indicates the degradation of the FZ machine capability in providing a protection atmosphere for the process, and machine maintenance may be necessary. In some cases, some surface anomalies may flake off, potentially touching the inductive coil and causing harm to it.

Therefore, it is of great significance to understand the nature of the surface anomaly and its impacts on the process. Consequently, in this chapter, a thorough characterization of the surface anomaly shown on the polysilicon surface is carried out from the aspects of microstructure and appearance.

2.2 Material Characterization of Surface Anomaly

In order to further investigate the nature of the surface anomaly, two polysilicon parts with different degrees of surface anomaly along with their FZ images acquired from the vision system were collected from the terminated FZ process. Figure 2.2 presents the image with a high contrast of the surface anomaly and the corresponding polysilicon part. As seen, the surface anomaly in the polysilicon part appears in three colors: white, blue, and black from the bottom (near the melt edge) to the top. Figure 2.3 presents the image with a slight surface anomaly and the corresponding polysilicon part, where white and gray colors are observed from the bottom to the top. The multicolored anomaly is suspected to be a SiO_2 film on the polysilicon, with the various hues signifying different thicknesses of the film due to the interference of light reflecting off the silicon. The color chart of silicon dioxide thickness when viewed from a 0 degree angle is shown in Figure 2.4 [63].

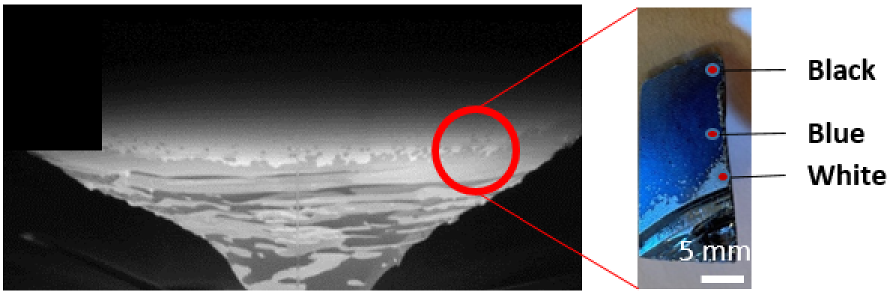


Figure 2.2: Comparison between an FZ image with a visible surface anomaly on polysilicon and the corresponding polysilicon part obtained from a terminated process.

To further our understanding of the surface anomaly from a physical point of view, this section provides a detailed microstructural investigation of the surface anomaly. Cross-sectional observation is an effective method for inspecting the internal structures of materials. In this study,

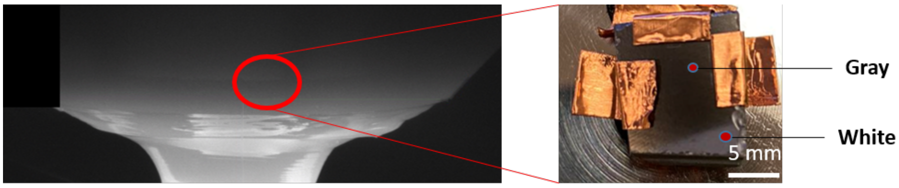


Figure 2.3: Comparison between an FZ image with a slight surface anomaly on polysilicon and the corresponding polysilicon part obtained from a terminated process.

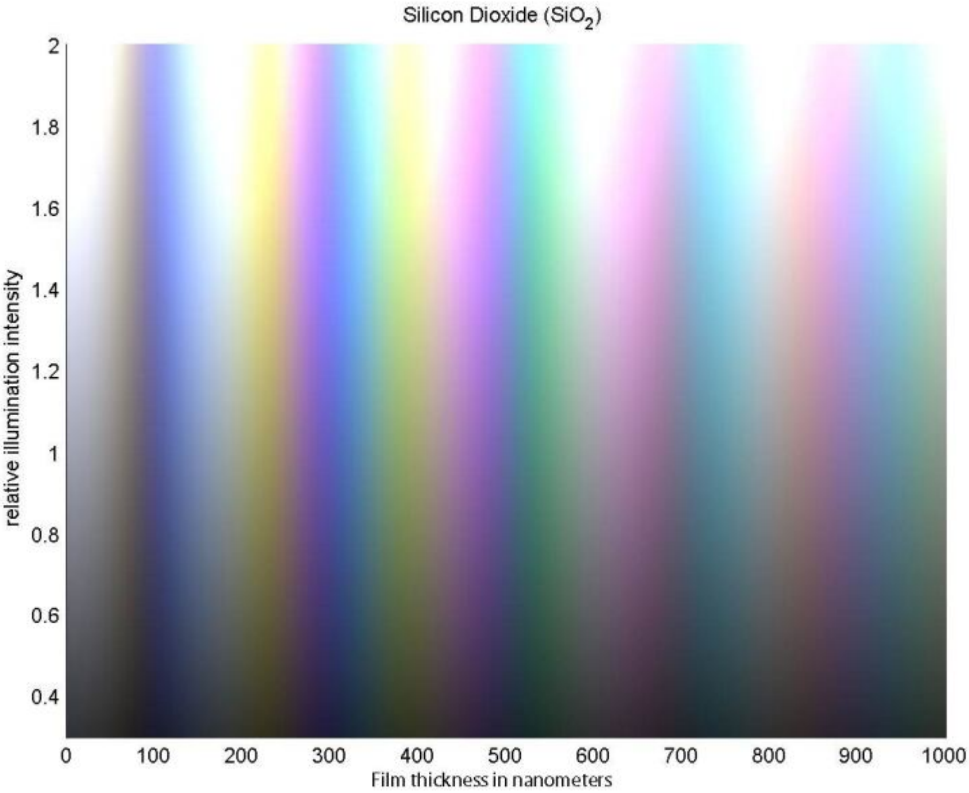


Figure 2.4: The color generated by thin films of silicon dioxide on a silicon substrate at 0 degree angle, or looking straight down onto the oxide [63].

Focus Ion Beam Scanning Electron Microscopy (FIB-SEM) would be employed for both imaging and preparing cross-sectional test specimens used for the subsequent element analysis. Following the sample preparation, Energy Dispersive X-ray Spectroscopy (EDX) is used for the elemental analysis of the cross-sectional specimens. Figure 2.5 presents an overview of the methods used for the material characterization of the film layer on polysilicon.

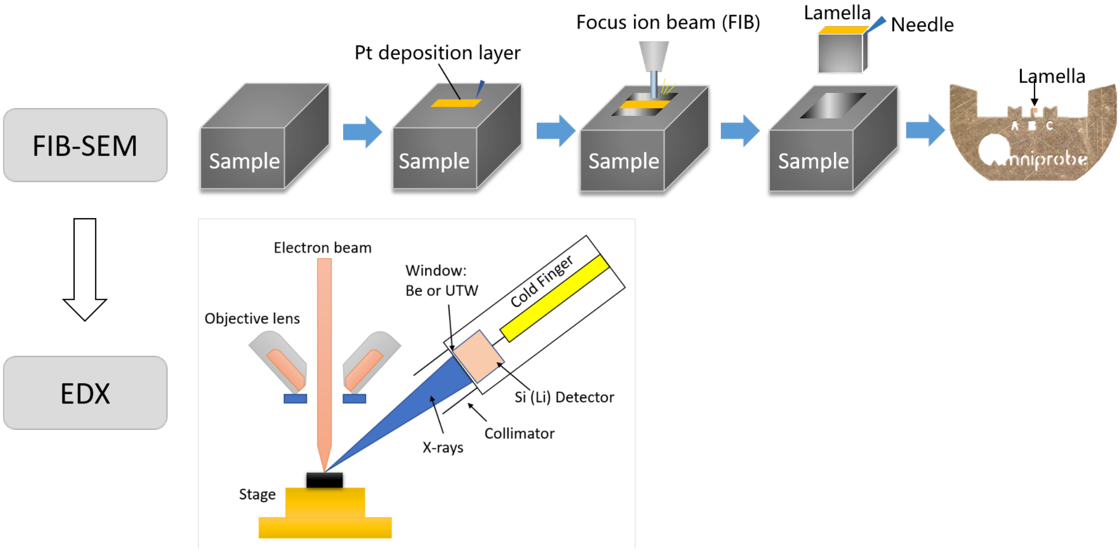


Figure 2.5: Overview of the methods used for material characterization of the film layer on polysilicon. The EDX figure is from [64].

2.2.1 Methods

Focus Ion Beam Scanning Electron Microscopy (FIB-SEM)

Scanning Electron Microscope (SEM) is the most commonly used for the characterization of organic and inorganic materials on a nanometer to micrometer scale [65]. EDX, a nondestructive analytical technique, normally works together with SEM to provide qualitative and semi-quantitative results, thus providing fundamental information on the material composition of samples that cannot be obtained by common laboratory tests [65]. However, when using EDX to study a sample with intermediate layers, one would encounter the penetration depth problem given by the EDX [66]. When attempting to analyze a thin film with EDX, signals from the substrate beneath the film may interfere with the signals from the film itself [66]. This can make it challenging to obtain accurate compositional information specific to the film. In particular, in this case, it is not clear how thick the surface anomaly layer is. Despite we assume that it is SiO_2 film, the same color band of SiO_2 film may have different ranges of thickness. Therefore, if directly conducting surface characterization with EDX, the quantification results of the elemental composition of the film may not be reliable since the penetration depth of X-rays analysis may encompass the substrate beneath the film.

FIB-SEM is a new approach to investigate the internal structures of materials by examining cross-sections at specific sites without the need for coating or mounting, thus avoiding any obstruction of microstructural features or sample contamination [67, 68]. FIB-SEM combines two powerful tools: a focused ion beam and a scanning electron microscope, offering unparalleled capabilities for microstructural analysis and specimen preparation. The focused ion beam, typically composed of gallium ions, enables precise material milling and sculpting at the nanoscale, while the scanning electron microscope provides high-resolution imaging of the sample surface. A schematic of the FIB-SEM system is illustrated in Figure 2.6. The ion beam column is positioned at an angle ranging from 52° to 55° in relation to the vertical axis of the electron beam. FIB-SEM begins with the generation of a highly focused beam of ions, commonly gallium ions (Ga+), within an ion source [69]. Then the gallium ions are accelerated to high energies and pass through the condenser and objective lenses to produce a fine beam [69]. The focused ion beam scans the sample surface, where it interacts with the material.

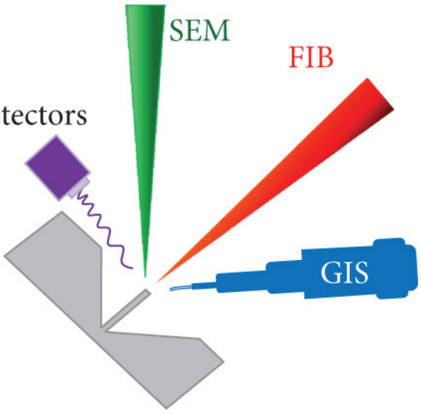


Figure 2.6: Schematic illustration of the FIB-SEM system from [69]. GIS: gas injection system.

Energy Dispersive Spectroscopy (EDX)

EDX is a method for both identifying the types of elements present in the materials and determining the elemental composition of the materials. It is commonly used in conjunction with Scanning electron microscope (SEM) or Transmission electron microscope (TEM). EDX can identify elements with an atomic number greater than boron when present at concentrations of at least 0.1% [70].

The principle of EDX is based on the interaction of high-energy electrons with a sample, leading to the emission of characteristic X-rays [70]. The electron beam strikes the innermost layer of an atom, causing an electron to be ejected from the shell, leaving a positively charged electron hole [71]. To fill the gap, another electron is drawn from an outer shell. As the electron transitions from the outer, higher-energy shell to the inner, lower-energy shell of the atom, the energy difference is released in the form of an X-ray [71]. The X-ray emission spectra of each element are unique, allowing them to be distinguished and their concentrations in the sample to be determined [72].

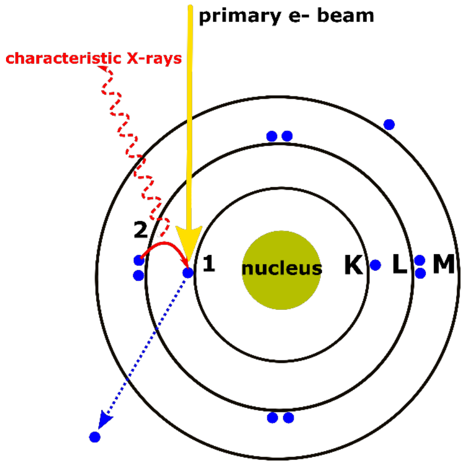
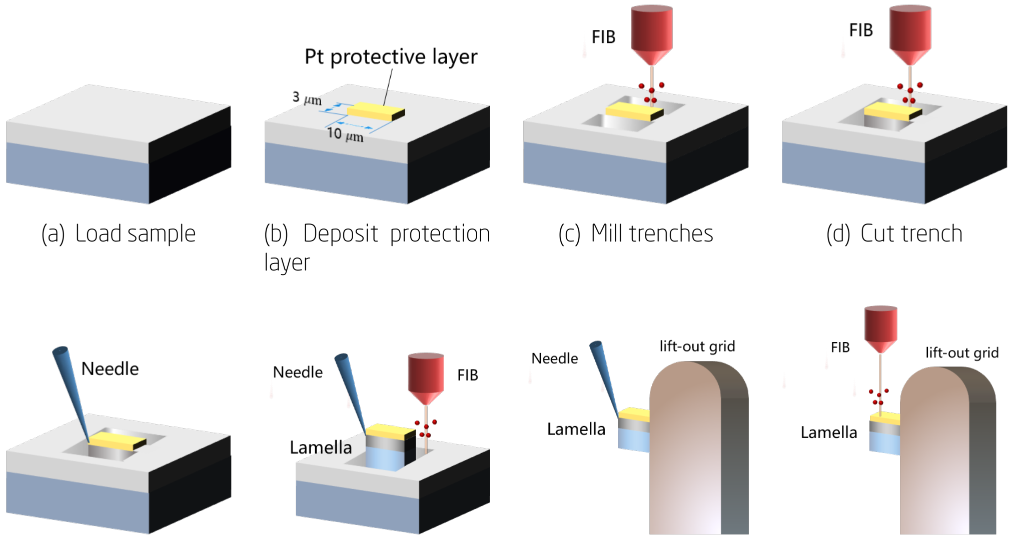


Figure 2.7: The principle of EDX [71]. First, the energy transferred to the atomic electron causes it to be ejected, leaving a vacancy; second, another electron from a higher energy level takes its place, resulting in the emission of a characteristic X-ray [71].

2.2.2 Experimental procedures

In this study, a Helios 5 Hydra UX PFIB was used to fabricate cross sections of the polysilicon sample, as seen in Figure B.1. And a Quanta FEG 200 ESEM along with the Oxford X Max EDS detector, as presented in Figure B.2, was used to conduct element identification and quantification analysis on the cross-sectional specimens.

Prior to the fabrication of cross-sectional samples, an examination of the surface morphology would be conducted. The sample with a slight surface anomaly on its surface can be used as a benchmark for analyzing the surface morphology of the test sample with severe surface anomaly. Then three marked sites shown in Figure 2.2 were chosen as the sites to prepare cross-sectional specimens. The test sample was then introduced into the FIB-SEM instrument. The procedure to prepare the FIB cross sections can be seen in Figure 2.8. The initial imaging of the sample was conducted in the SEM mode to locate the region of interest (ROI). After locating the site for FIB cross sectioning, a protective layer of platinum was deposited over a $10\,\mu\mathrm{m} \times 3\,\mu\mathrm{m}$ area with its height of $1\,\mu\mathrm{m}$, to prevent damage from the Ga ion or the e^- beam to the underlying film [68]. The sample stage is adjusted to an angle of 52 °so that the ion beam is perpendicular to the sample surface during the Pt deposition process. Then FIB mode was employed for precision milling to expose a cross-section of the sample. Two trenches along the front and back surfaces of the lamella were first milled and rough polished with a 30 kV ion beam, followed by another trench for needle on one side of the lamella. The required milling depth depends on the depth of the lamella. Here, the desired depth of the lamella was set to $8\,\mu\mathrm{m}$. Then, an in-situ needle approached the lamella and was welded using platinum at the fixed position. Thereafter, focus ion beam continued the trench cutting until the lamella was free. The lamella was transferred to the pre-defined position on a lift-out grid and fixed using welding. Once the ion beam had cut the lamella from the needle, the needle was retracted. Then, the ion beam performed final polish process to obtain a flat surface, using a 30 kV 0.3nA-thinning, a 30kV 30 pA-polishing and a 12 kV 30 pA-polishing. Figure 2.9 presents some SE images showing the procedures for preparing cross sections using FIB.



(a) Load sample

(b) Deposit protection layer

(c) Mill trenches

(d) Cut trench

(e) Load needle

(f) Cut lamella free and lift it out

(g) Attach lamella to the lift-out grid

(h) Polish the lamella

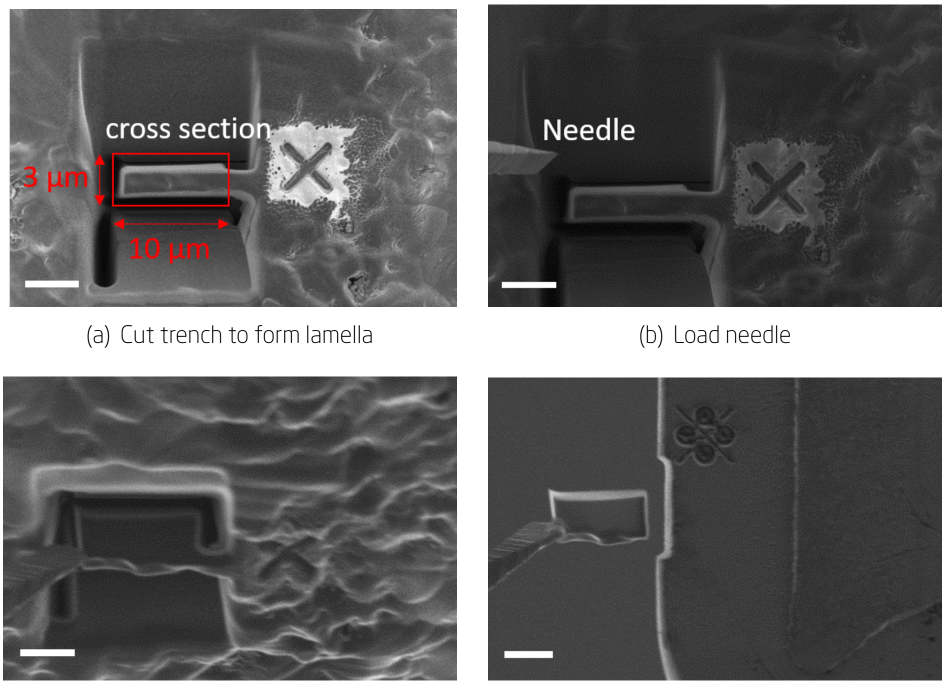
Figure 2.8: The procedure of using FIB-SEM for preparing a cross section. (a) Introduce the sample to the FIB-SEM instrument. (b) Deposit protection layer of Pt. (c) Mill process. (d) Cut trench to form lamella. (e) Load needle and attach the needle to the lamella. (f) Cut lamella free and lift it out. (g) Transfer lamella to a lift-out grid and fix it at the position. (h) Cut the needle and polish the lamella.

Then mounted cross sections were analyzed using a Quanta FEG 200 ESEM along with Oxford X Max EDS detector. Accelerating voltage of 5 kV of the electron beam was selected. Then the cross-sectional specimens were analyzed by EDS line scan for element quantitative analysis. The Oxford Instruments Aztec nanoanalysis software suite was utilized to analyze all EDX micrographs and spectroscopic data.

2.2.3 Results and Discussion

Surface Morphology

Before preparing cross-sectional specimens, the surface morphology of the samples was analyzed by SEM mode in FIB-SEM equipment. Figure 2.10 and Figure 2.11 reveal the surface mor-



(a) Cut trench to form lamella

(b) Load needle

(c) Attach the needle to the lamella

(d) Transfer lamella to a lift-out grid

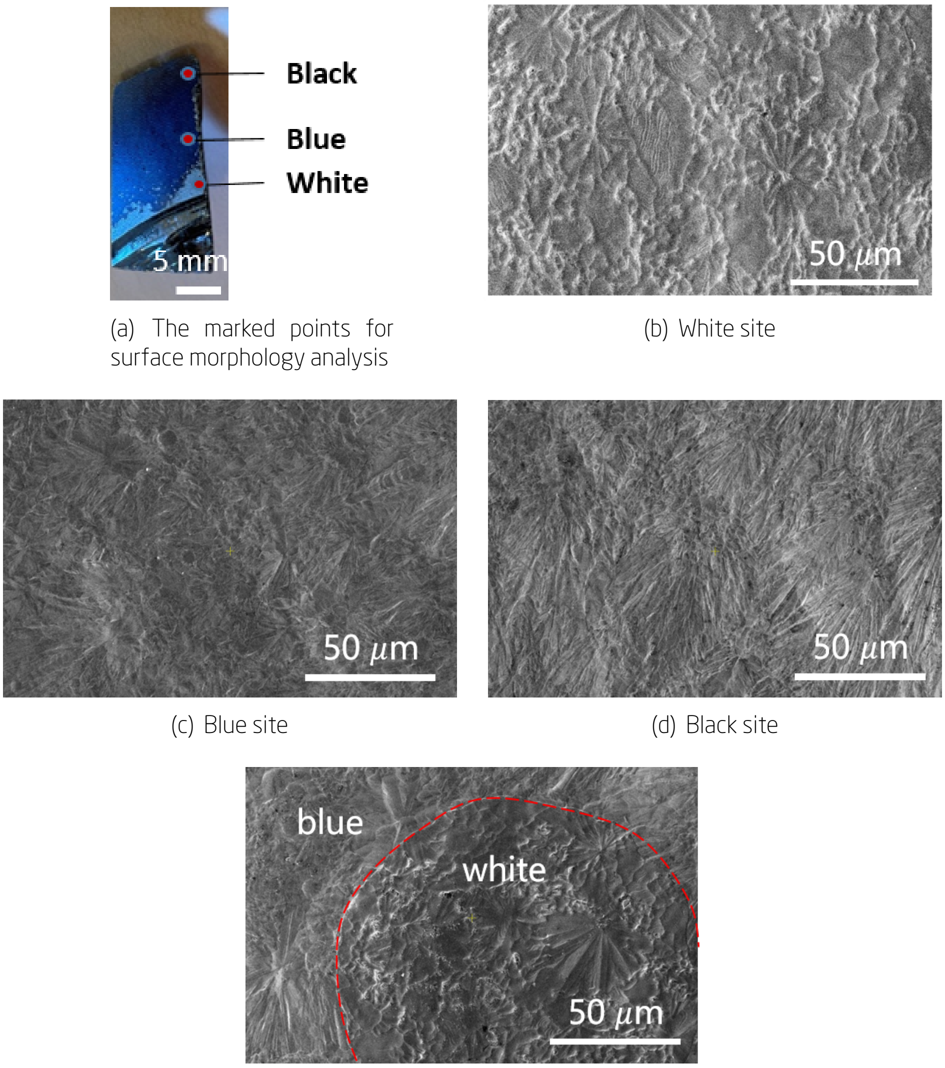
Figure 2.9: Images produced by FIB-SEM that demonstrate the steps of the FIB lift-out technique. The scale bars shown in the images are around 5 $\mu\mathrm{m}$.

phology of the first sample with three colors (white, blue and black) and the second sample with two colors (white and gray), respectively. As seen, both white sites in two samples demonstrate similar uneven texture, while the surface in the second sample with the slighter anomaly has a denser texture than that of the first sample. It is noticed that except for the white region the colored regions share a feather-like texture.

Cross-sections Characterization

Figure 2.12 presents the cross-sectional SEM images of three colored sites on the first sample. As seen in Figure 2.12 (b), no intermediate layer appears between the Pt deposited layer and the original substrate in the white region. Differently, thicknesses of 85 nm and 56 nm are measured in the intermediate layers of the blue region and the black region, respectively (see Figure 2.12 (c) and (d)).

EDX line analysis results of white, blue and black sites can be seen in Figure 2.13, Figure 2.14 and Figure 2.15, respectively. Elements of C, Si, O, and Pt were identified during the EDX analysis. Therefore, the line scanning profiles would present the intensity distribution of C, O, Si, and Pt along the scanning line. The line scan profiles reveal that the intermediate layers at both the blue



(a) The marked points for surface morphology analysis

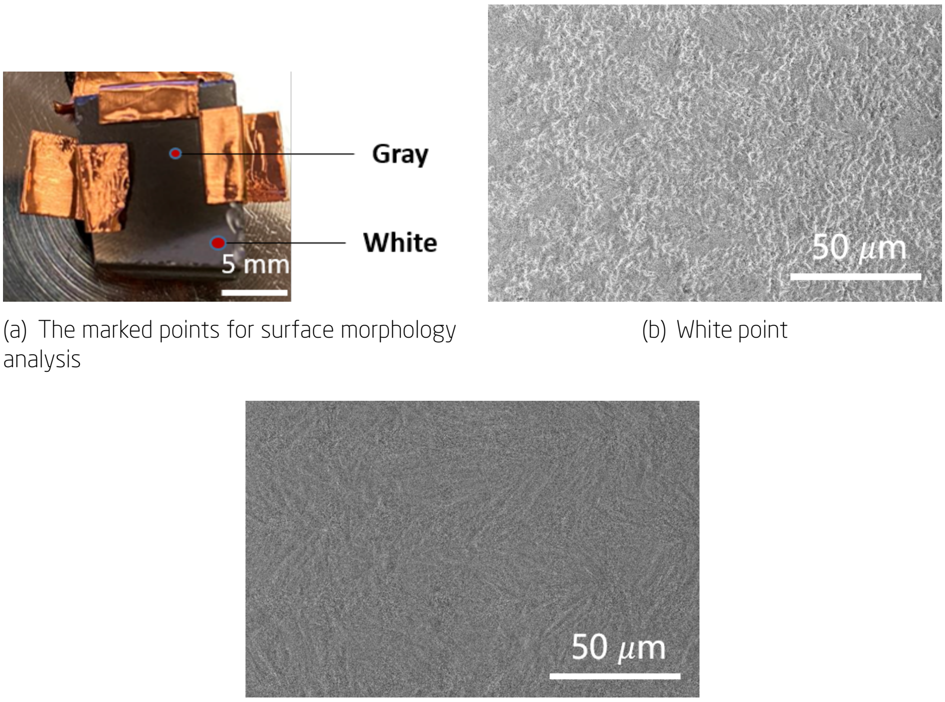
(b) White site

(c) Blue site

(d) Black site

(e) A junction of blue region and white region

Figure 2.10: SEM images showing the surface morphology analysis of the first sample with apparent surface anomaly.



(a) The marked points for surface morphology analysis

(b) White point

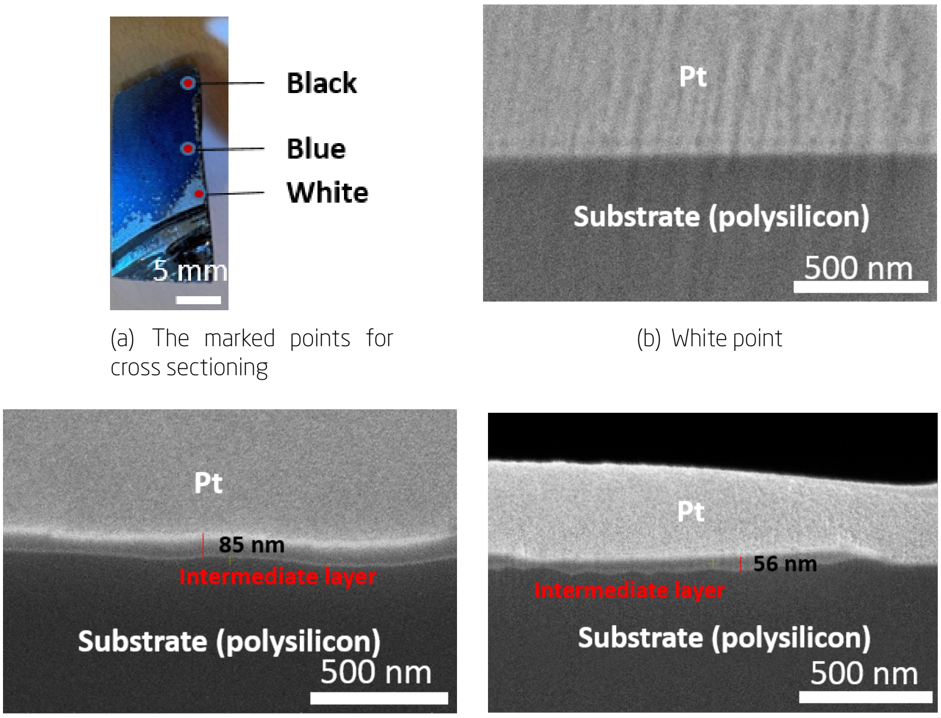
(c) Gray point

Figure 2.11: SEM images showing the surface morphology analysis of the second sample with slight surface anomaly.

and black sites are enriched with oxygen atoms, indicating the presence of the oxide layer on top of these regions.

What is rather surprising is that there is little difference in the C intensity distribution along the EDX line on the cross section of the white site, while a rise was observed in both the blue and black sites. Quantitative measurements were carried out by EDX point analysis on the bottom of each cross-sectional sample, where only the substrate polysilicon locates, as seen in Figure 2.16. Five measurement points were probed for analysis on each cross-sectional sample, and the average and standard deviation were recorded. Figure 2.17 presents the results of EDX point analysis on the substrates of the white, blue and black sites. As seen, as the position on the polysilicon moves from the top to near the melt edge (the same as the position from black site to white site), the intensity of C atom in the substrate polysilicon increases, along with the decrease of the Si atom. This might be due to the fact that as it approaches the melt edge, the closer contact with the high-temperature molten zone might promote the chemical reactions or decomposition processes that release carbon atoms and increase their intensity on the surface.

To sum up, the results suggest that the high contrast observed in the FZ images with surface anomaly is due to oxidation. Specifically, the dark area near the melt edge in the high-contrast FZ



(a) The marked points for cross sectioning

(b) White point

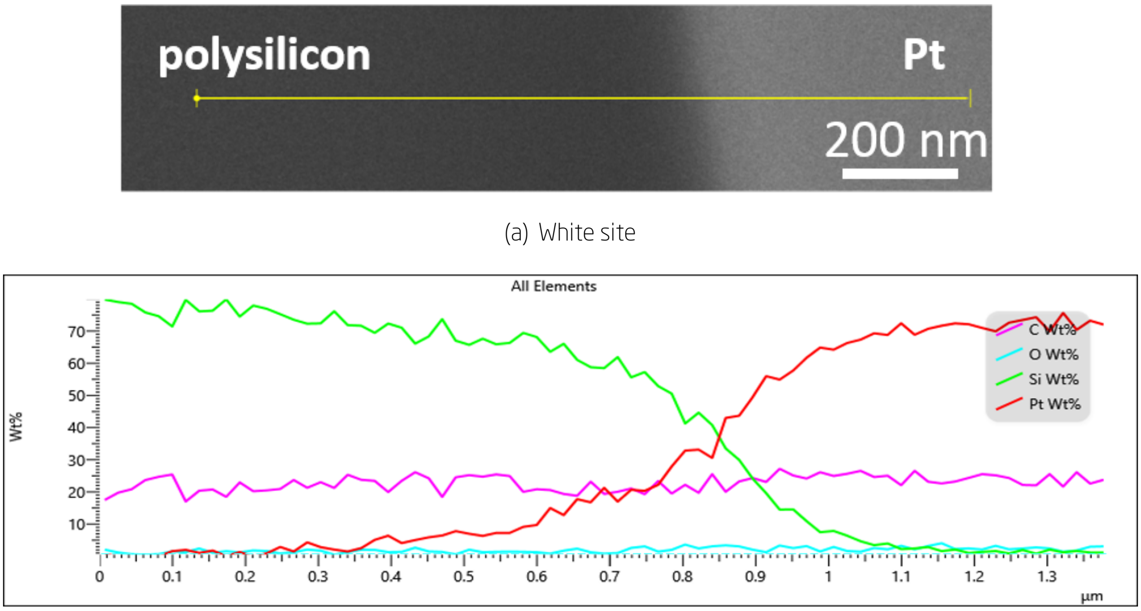
(c) Blue site

(d) Black site

Figure 2.12: Cross-sectional SEM images of three sites in the first sample. (a) Marked points for cross section. (b) The cross-sectional SEM image of the white site. No intermediate layer was observed. (c) The cross-sectional SEM image of the blue site with intermediate layer thicknesses of 85 nm. (d) The cross-sectional SEM image of the black site with intermediate layer thicknesses of 56 nm.

image shown in Figure 2.2, which corresponds to the white surface in the polysilicon part, is the silicon itself without observing oxygen. On the contrary, the bright area at the top is enriched with oxygen. The color presented in the oxidation region of the polysilicon part indicates its thickness of the oxide layer.

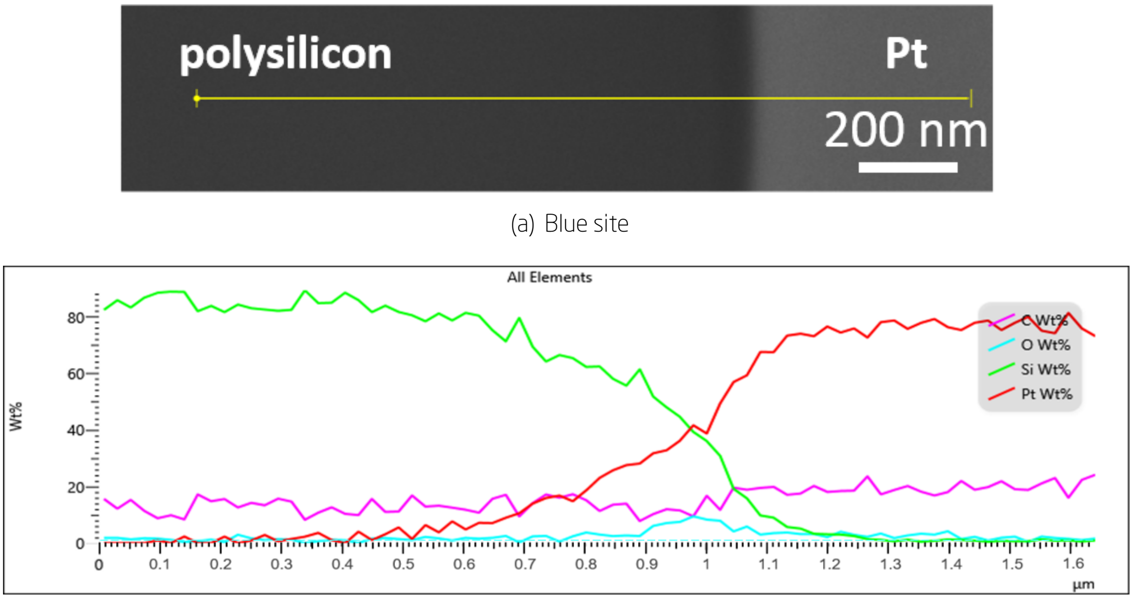
Considering that the white region is closer to the heater at a higher temperature, the absence of detected oxides in that region could be explained by the oxygen loss by evaporation. There may have been oxides present in that region, whereas they were dissolved via the reaction $SiO_2 + Si \longrightarrow 2\,SiO$, where SiO would easily evaporate at temperature over 1100 °C [21]. With evaporation going on, the substrate gradually becomes exposed, while a high-contrast appearance becomes visually evident, as shown in the FZ images. This also implies that the oxidation occurs earlier than when we observe the high-contrast anomaly in the FZ images.



(a) White site

(b) Intensity distributions

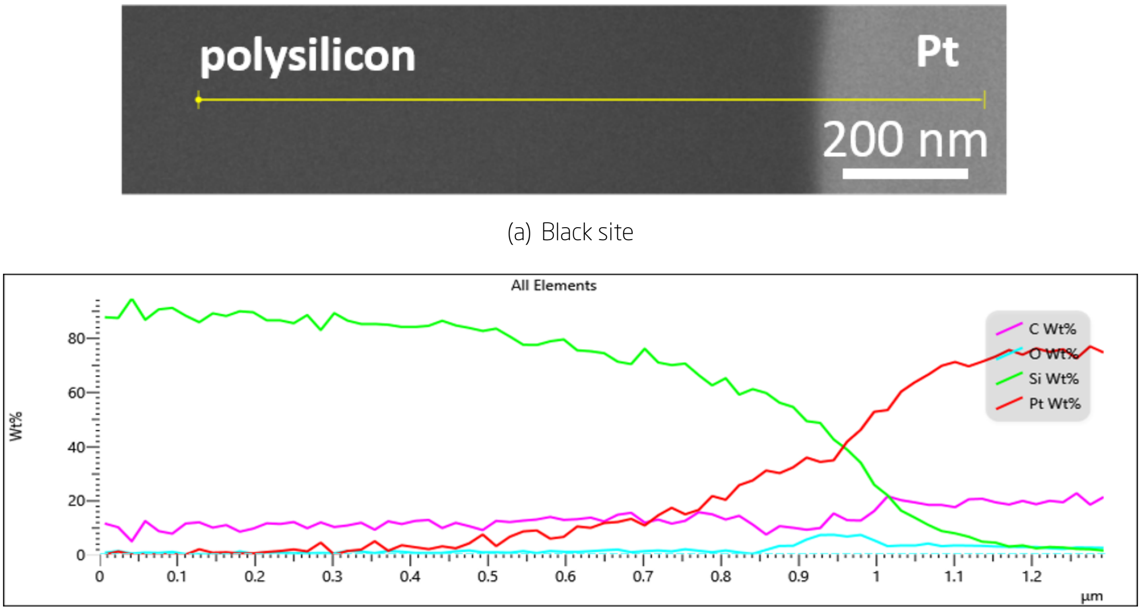
Figure 2.13: EDX line scanning analysis along the white site on the first sample. (a) Cross-sectional SEM image. The yellow line indicates the EDX scanning line. (b) Intensity distributions of C,O,Si and Pt along the scanning line.



(a) Blue site

(b) Intensity distributions

Figure 2.14: EDX line scanning analysis along the blue site on the first sample. (a) Cross-sectional SEM image. The yellow line indicates the EDX scanning line. (b) Intensity distributions of C, O, Si and Pt along the scanning line.



(a) Black site

(b) Intensity distributions

Figure 2.15: EDX line scanning analysis along the black site on the first sample. (a) Cross-sectional SEM image. The yellow line indicates the EDX scanning line. (b) Intensity distributions of C,O,Si and Pt along the scanning line.

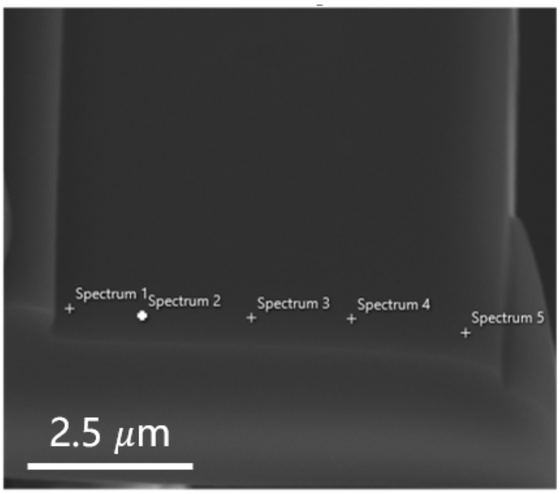


Figure 2.16: Conduct five times of EDX point analysis on the bottom of each cross-section specimens, where only the substrate polysilicon is presented.

2.3 Visual Characterization on Surface Anomaly

The sophistication of the FZ crystal growth process is considerable, yet it is advantageous that it can be well visually observed [17]. As mentioned in Section 1.2, the FZ process is being monitored by the FZ vision system. The optical camera equipped in the FZ vision system films in

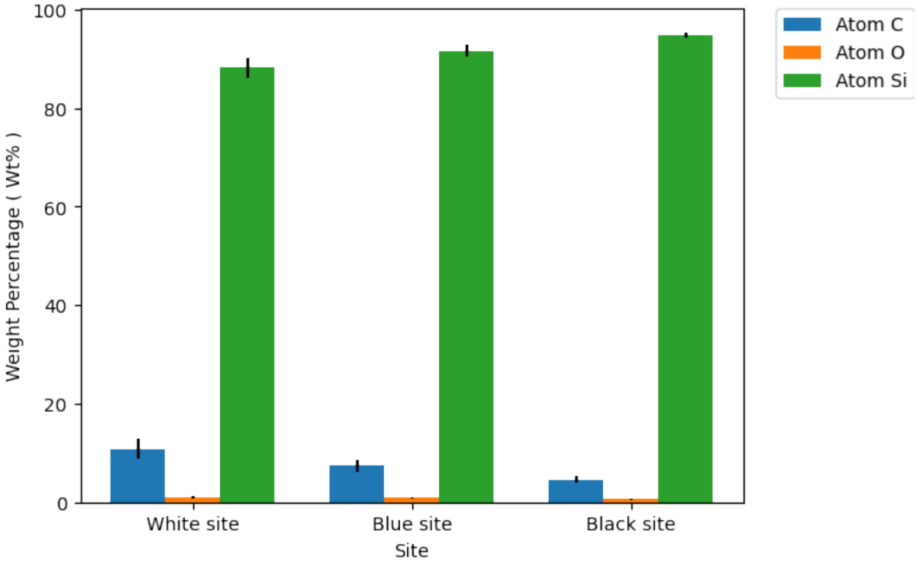


Figure 2.17: The point analysis results of the substrates of white, blue, and black sites.

monochrome at approximately 20 frames per second, capturing grayscale images with dimensions of 1600x1200 pixels. The images stored in the current database were already processed by the FZ automation system to control the FZ process, thus containing some embedded information in the upper left corner such as timestamp, machine serial number, temperature, etc. Since the embedded information contains confidential information, it was marked with a black rectangle. A sketch of the FZ image from the FZ vision system and an example of an FZ image after processing are shown in Figure 2.18.

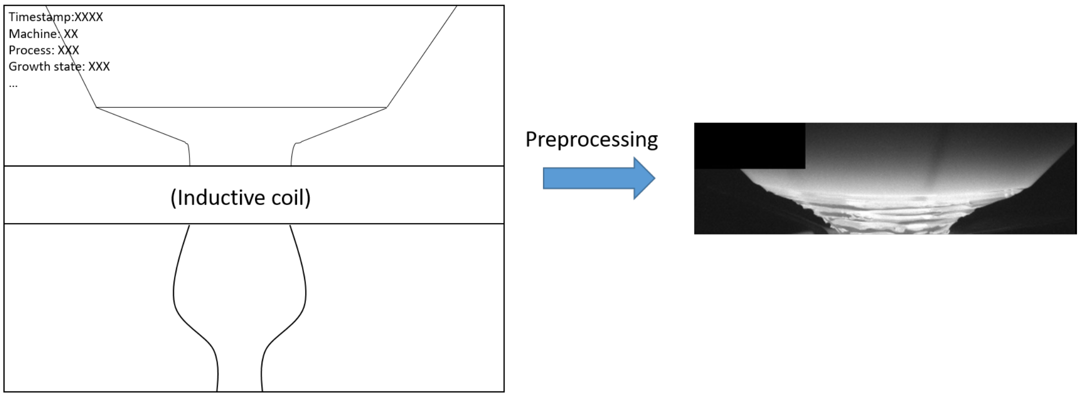


Figure 2.18: An example image before and after processing. The original image is simplified here since it involves the confidential information.

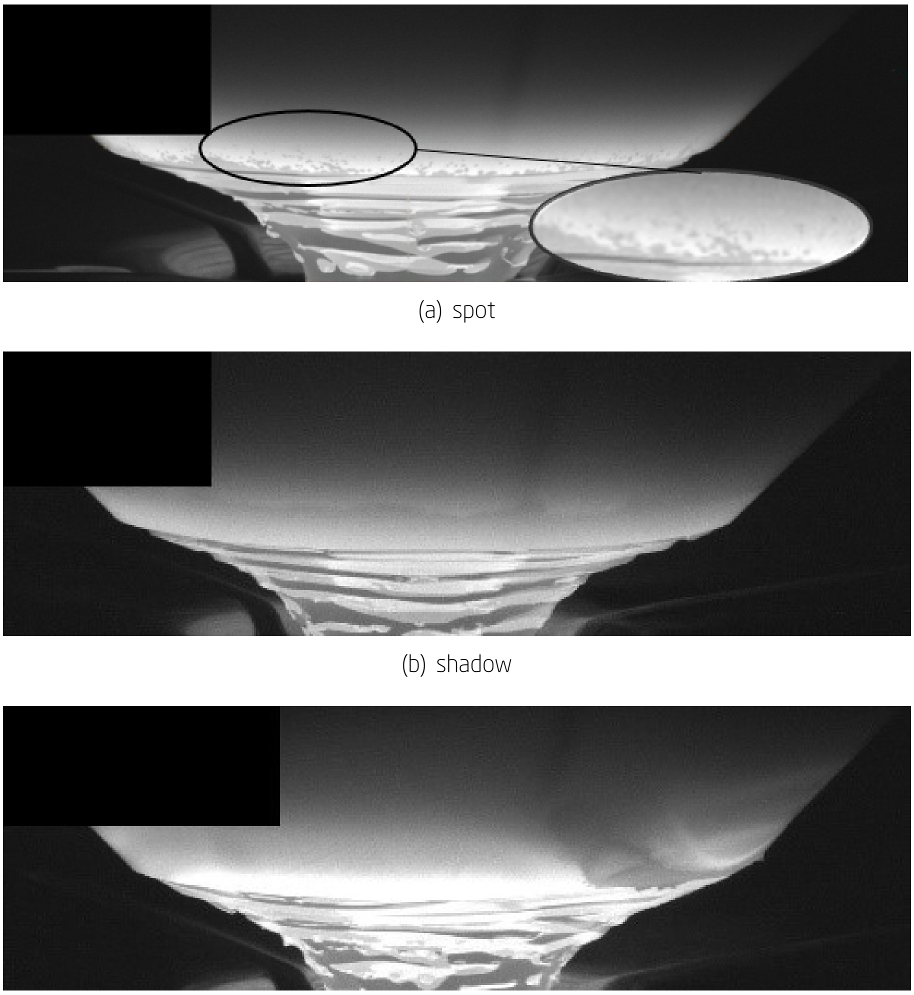
During the FZ crystal growth process, the polysilicon is supposed to be heated homogeneously [17] and its surface is also supposed to be with homogeneous intensity in gray-scale images from the vision system. However, during the process, anomalies might occur on the polysilicon surface, which is reflected in the FZ images captured from the vision system. Typically, the surface anomaly is often visually evident in the FZ image from the beginning of the cone phase and gradually fades before moving to the cylinder phase. This might be due to the fact that as the FZ process goes from the cone phase to the cylinder phase, the volume of heated polysilicon increases, thus accelerating the dissolution and evaporation of oxygen. Assuming there is a uniform oxide film with a thickness of 85 nm and an oxygen content of around 10 Wt% on the cone shape of polysilicon with the diameter ranging from 10 mm to 170 mm, and assuming that all oxygen are dissolved and diffused into the melt, the oxygen level in the melt is around $1.2 \times 10^{16} atoms/cm^3$, which is already at a low level. In addition, a valued crystal ingot counts from the cylinder phase. Hence, the surface anomaly on the polysilicon has a minor effect on the product quality.

However, since every polysilicon feed rod is carefully cleaned to prevent oxidation, the presence of the oxide layer implies that oxygen could be coming from the chamber ambient or from the machine itself. This indicates that the environment of the FZ chamber has been altered as a result of the introduction of oxygen. While oxygen has the opportunity to react with the polysilicon surface, it is also possible for oxygen to directly react with the growing crystal, thus directly threatening the product quality and crystal yield. The introduction of oxygen into the FZ chamber can promote the formation of dislocations within the crystal [73]. In addition, if oxygen is incorporated into the crystal structure, it can result in oxygen-related defects, which might have negative impacts on the subsequent crystal properties. On the other hand, the introduction of oxygen indicates that the FZ machine may be deviating from optimal condition, and maintenance might be needed. Therefore, it is still necessary to investigate the root causes behind the symptom of surface anomaly in order to improve both the FZ process and the FZ machine.

The surface anomaly can present with different characteristics. Based on the experimental observation of the FZ process, the surface anomaly can be mainly categorized into three classes, including the spot, shadow, and ghost curtain, as seen in Figure 2.19. It is hypothesized that the various types of oxides could signify different process conditions.

Spot-type anomalies are often seen near the melting edge of polysilicon surfaces with many dark spots. These spots tend to grow into more spots as they move closer to the inductive heater. Shadow is a common occurrence in the FZ process, and it is characterized by a large, smooth-edged area with relatively low contrast. The difference between the appearance of spot and shadow may be due to the thickness of the oxide layer. The thickness of the spot may be relativley thicker that that of the shadow, making it difficult to dissolve. Ghost curtain is another appearance of the oxide that is partially detached from the polysilicon substrate and appears darker than other oxides.

Among all oxide types, ghost curtain is the most serious case since ghost curtain is likely to



(a) spot

(b) shadow

(c) ghost curtain

Figure 2.19: Oxide types: (a) spot, point-like melted oxide (b) shadow, regional melted oxide (c) ghost curtain,the oxide that are separated but not entirely detached from the substrate polysilicon.

flake off, which has the potential to cause machine breakdown or product contamination. Furthermore, since ghost curtain is a detached layer that can distort the shape of the polysilicon in the FZ images, as seen in Figure 2.20, it can introduce noise in geometrical measurements, such as polysilicon diameter, which is an important component for modeling the behavior of the FZ process [17]. This can lead to errors in the feedback signal given to the controller, resulting in making incorrect decisions and failing to maintain the desired setpoint. In terms of spot and shadow anomalies, they would sometimes finally disappear and become homogeneous surface again as they move downward to be heated.

Aside from the impacts of various oxides on the machine and process, the frequent appear-

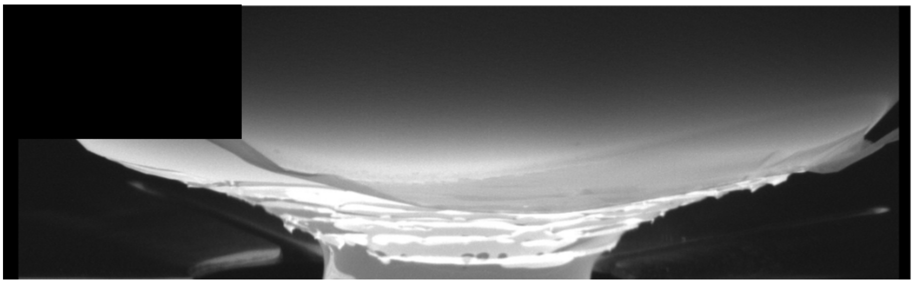


Figure 2.20: An example image with ghost curtain that pose noises in geometrical measurements of the polysilicon diameter. This can lead to less accurate performance of the automatic control system.

ance of the oxide also indicates that the performance of FZ machines is in fact declining and that machine maintenance may be needed.

2.4 Conclusion

Targeted for the surface anomaly emerged on polysilicon surface on FZ images during the FZ process, a thorough investigation regarding the nature of the surface anomaly was carried out by material characterization and visual characterization.

First, to further our understanding of the surface anomaly from a physical point of view, a detailed microstructural material characterization of the surface anomaly was conducted. First, surface morphology analysis was conducted on a sample with apparent high-contrast appearance on the FZ gray-scale image and a sample with slight anomaly on the FZ image, using SEM mode of FIB-SEM. It was observed that the white region near the melt edge in both samples shares a similar uneven texture, whereas other colored regions appear rougher with a feather-like texture compared with the relatively smoother surface of the gray region in the sample with slight anomaly. Then, three sites in white, blue and black color on the first polysilicon sample with apparent surface anomaly were selected for cross sectioning. The findings demonstrate that no layer was present between the platinum layer and the polysilicon substrate at the white site, while an intermediate layer with a thickness of 85 nm and 56 nm was detected at the blue and black sites, respectively. EDX line scanning profiles on these three cross-sectional specimens suggest that the high contrast observed in the FZ images with surface anomaly is due to the oxidation as well as the dissolution and evaporation of oxygen. Typically, the dark part near the melt edge in the high-contrast image shown in Figure 2.2, corresponding to the white surface in the polysilicon part, is the silicon itself without observing oxygen. On the contrary, the colored region beyond the white surface on the polysilicon sample is enriched with oxygen atoms on its surface. The presence of its color indicates the thickness of the oxide layer. These results reveal that oxidation occurs in the early phases. As the oxide layer approaches the heater, part of the oxide layer is dissolved at high temperature and finally evaporates in the form of SiO, resulting in high-contrast appearance in the FZ images.

Second, visual characterization was performed on FZ images gathered from the FZ vision system. Typically, the surface anomaly is often visually evident in the FZ image from the beginning of the cone phase and gradually fades before moving to the cylinder phase. In addition, the oxygen level introduced by the oxide is at a relatively low level. However, this does not clear the alarm of threatening crystal yield, since the presence of an oxide layer indicates the abnormal FZ chamber condition with the introduction of oxygen. The introduction of oxygen may directly impact product quality by promoting the formation of dislocations within the crystal [73] and resulting in oxygen-related defects. Based on the experimental observation of the FZ process, the surface anomaly can present in three categories, including the spot, shadow, and ghost curtain and their characteristics and potential impacts were discussed.

To sum up, the nature and characteristics of surface anomaly were analyzed and discussed in this chapter. The findings reveal that the presence of oxide indicates the problem in both the FZ process and the FZ machine, and it could potentially have negative impacts on the crystal yield. Therefore, it is necessary to optimize both the FZ process and the FZ machine by investigating around the surface anomaly in the following steps: identification of the oxide, root cause analysis of the oxide, and automatic responses to the oxide, which would be detailed in Chapter 3, 4, 6, respectively.

CHAPTER 3
OXIDE IDENTIFICATION BY MULTI-LABEL CLASSIFICATION

"The journey of a thousand miles begins with one step." - Lao Tzu

3.1 Motivation

With the characterization of the surface anomaly on the polysilicon surface, it is confirmed that the surface anomaly is due to the enrichment of oxygen atom. Though the oxygen introduced by the surface anomaly is at a low level, its presence indicates an abnormal FZ chamber condition. On the one hand, the introduction of oxygen may impact product quality by promoting the formation of dislocations within the crystal [73] and resulting in oxygen-related defects. This deviation in product quality could be costly as the product that does not meet customers' requirements would be turned into a scrap. On the other hand, the abnormal ambient indicates that the FZ machine may be deviating from optimal condition and maintenance might be necessary. Therefore, in order to improve the FZ process and FZ machine, it is necessary to identify the oxide problem in a timely manner to allow the decision maker to track deviations and to take corrective measures in order to reduce costs. Hence, this chapter aims to address 'Know-What' in oxide investigation by identifying the oxide. Currently, human identification of surface anomalies is still the norm, which is time-consuming and inefficient. To identify the oxide in a timely manner, this study would take advantage of the FZ images captured from the vision system and attempt to develop an automated oxide identification solution.

The oxide can present different characteristics, including spot, shadow as well as ghost curtain. In addition, the oxide may occur in combinations of these characteristics. The reasons for the occurrence of different oxides remain currently unknown. It is hypothesized that the various appearance of the oxides could signify different process conditions. Therefore, this study extends beyond oxide detection and aims at oxide classification.

The results of this chapter are from an article submitted to the journal of Intelligent Manufacturing [25].

3.2 Problem Definition

Taking into account the potential coexistence of various oxides, the categorization challenge evolves into a multi-label oxide classification dilemma. It should be noted that the multi-label classification problem is different from binary and multi-class classification problem where only a single label is associated with each instance [74]. In multi-label classification, a set of relevant labels could be assigned to an instance simultaneously.

The purpose of this study is to predict a set of oxide types given a FZ image $X \in \mathbb{R}^{m \times n}$. The prediction for X is a set of k binary labels $\{y_1, y_2, ..., y_k\}$, where $y_i \in \{0, 1\}$. The class i is relevant to X when $y_i = 1$. The normal case is generally represented as a vector filled with zeros, which implies that none of the oxides is activated. Thus, a classifier f is needed to map an unlabeled image to a set of labels: $\hat{\mathbf{y}} = f(X)$.

There are variabilities in the FZ machines and their produced products, such as those with different sizes or distinct crystal structures. However, the FZ vision system found little difference in the appearance of the polysilicon under normal or oxidation conditions among the different situations. Thus, these variabilities were not taken into account in this study. Since the oxide typically becomes evident from the start of the cone phase, the image at this time point is selected. A total of 2143 images were collected from six different FZ machines over a two-year period. All images were labeled with oxide types based on their characteristics.

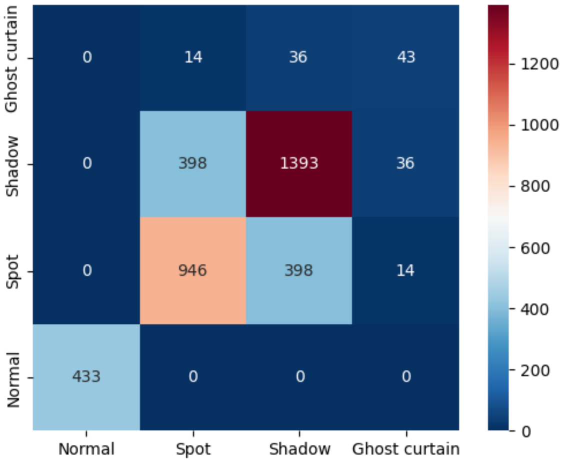


Figure 3.1: Co-occurrence matrix of the oxide dataset [25]. Data imbalance is observed in the dataset in two aspects: within a class and between classes.

Figure 3.1 illustrates the co-occurrence matrix of oxide types. It is evident that the oxide-free process does not dominate the entire dataset, which motivates this study. Additionally, it is clear to observe data imbalance in two aspects: within a class and between classes. Data imbalance within a class means the imbalance between positive and negative samples for a specific class (a sample is defined as positive for a certain class if the sample belongs to the specific class and vice versa). A typical example is the ghost curtain class, where negative samples dominate the class. Data imbalance between classes means the imbalance between the dominant class and the rare class. Typical data imbalance can be observed between shadow class and ghost curtain class. The data imbalance would pose difficulties for model training since the rare class such as the ghost curtain will be 'neglected' since the total loss is mostly dominated by easy samples, for instance, negative samples of the rare class or samples with distinct features from the dominant class. However, ghost curtain as a rare class cannot be neglected as it has great potential to harm the FZ machine and the FZ process. Although the probability of the occurrence of a ghost curtain is quite low compared with other oxides, the economic loss caused cannot be ignored. Therefore, in this study, the data imbalance problem should be addressed.

3.3 Literature Review

3.3.1 Multi-label image classification

In the area of multi-label image classification, there are two key issues involved: Multi-label classification (MLC) and image classification.

MLC is an extension of single-label classification, in which more than one class can be assigned to a single instance [75]. Recently, it has been acknowledged that the utilization of a single label is not typically suitable for real-world applications [76], thus making MLC increasingly important. The primary difficulty of MLC is the non-uniform distribution of samples and their labels across the data space [74, 77]. Conventionally, there were two prevailing approaches for addressing the MLC problem: problem transformation and algorithm adaptation [75]. Problem transformation aims at fitting data to existing algorithms by transforming a multi-label problem to a single-label problem or regression problem. Examples include Binary Relevance (BR) which decomposes the task into a set of binary classification problems, one for each class label [78], and Label Powerset (LP) where MLC is tranformed to multi-class classification by encoding each unique set of labels as a distinct class [79]. Algorithm adaption instead fits the algorithm to multi-label data by adapting learning techniques to deal with the MLC problem, for instance multi-label K-nearest neighbor (ML-KNN) [80]. However, problem transformation and algorithm adaptation approaches are not effective in tackling the imbalance problem in MLC [74]. In addition, neither of these two approaches considers feature extraction, as they are mainly designed for low-dimensional data, such as tabular data. Conversely, image data are high-dimensional, and the relationships between pixels are complex, which may require additional feature engineering to capture the relationships effectively.

Although image classification can be considered second nature for humans, it is much more difficult for an automated classifier [81]. The challenges include the characteristics of high dimensionality of images and viewpoint-dependent object variability and the high in-class variability of having many object types [82]. Traditionally, image classification begins with feature extraction, which involves discovering an ideal image descriptor to convert an array-like image to a vector, thus allowing for further analysis by a classifier [81]. However, the accuracy of this approach depends on the feature extraction step, which usually proved to be a formidable task [81]. Recently, Convolutional Neural Networks (CNNs) [83] have been receiving a great deal of attention and have become the leading architecture in image classification due to their powerful capability for feature extraction and pattern learning [81]. CNN is a type of artificial neural networks, consisting of multiple convolutional layers and pooling layers for downsampling, which empowers CNNs to autmatically capture patterns from grid-like data. Furthermore, the development of GPU implementation and the eagerness to gain better model performance have been major contributors to the resurgence of deep learning, which has also enabled the exploration of deep CNNs [81]. Popular deep CNNs architectures include AlexNet [84], VGG [85], ResNet [86], etc. However, due to the large number of model parameters to be optimized, deep CNNs usually require massive training data, This can be a challenge in the case of multi-label dataset because of high burden of annotations [87] as well as its inherent long-tailed distribution [88].

Recently, transfer learning that exploits the knowledge from a large dataset to boost model performance in customed small dataset has been shown to overcome this challenge. It has been shown that the features learned from large-scale data in many deep neural networks appear to be general, so they are applicable to similar or even dissimilar tasks [89, 90]. Hence, the core concept of transfer learning is to take advantage of the features learned from a large-scale dataset and apply them to a different dataset by copying the first n layers of a pre-trained network to the first n layers of the target network [89]. Recent studies have shown that using transferred features to initialize a new task can improve its generalization performance [89].

On top of the pre-trained CNNs, current research in the field of multi-label classification can be divided into three categories: (1) object localization assisted MLC, (2) label correlation optimization based MLC, (3) loss optimization based MLC. Object localization-assisted MLC aims to improve multi-label classification performance by incorporating object localization information [91]. This approach is often applied when the goal is not only to categorize items in an image but also to precisely pinpoint their locations within the image. Hence, this method acquires not only category labels but also localization information [92, 93] or segmentation information [94], further increasing the annotation burden. In addition, this method usually involves two steps, including localization and classification, which usually suffers from high computational cost. Label correlation learning-based MLC, is another alternative to improve multi-label learning performance by modelling label correlations, for instance, Graph Convolutional Networks (GCNs) [95] or Recurrent Neural Networks (RNNs) [96]. However, GCNs usually require manually defined adjacency matrices [97], while the effectiveness of RNNs in multi-label tasks has yet to be proved [97]. Loss optimization-based MLC attempts to design a loss function for adapting multi-label task and addressing the data imbalanced problem in multi-label dataset [98]. This method is a straightforward solution without the need for a carefully designed architecture and additional external information [98].

3.3.2 Approaches for Addressing Data Imbalance in MLC

As mentioned above, the current multi-label dataset suffers from data imbalance problem, which necessitates an approach to address this problem in MLC. Addressing this problem can be either from data perspective or from algorithm perspective [99]. A commonly used method in a data perspective is data re-sampling, which aims at creating a balanced dataset by over-sampling minority class or under-sampling majority class [100]. However, the main disadvantage of re-sampling methods is that they can lead to the loss of important information or the introduction of noisy data, since they alter the original data distribution [99]. In particular, for multi-label data with complex label correlation, re-sampling may be less effective. Another alternative is to address data imbalance from algorithm perspective, by modifying the classifier or modifying the loss function. For instance, ensemble techniques that combine several base models in order to produce optimal predictions can be used to handle imbalanced data [101, 102]. However, building an ensemble usually requires training a set of base models, which can be computationally costly and time-consuming [74]. Another solution is to adopt a designed loss function to mitigate the imbalanced problem. Most commonly, a modified loss function is used in imbalanced learning, where the minority classes are assigned a higher penalty for incorrect classification [74]. This could be achieved by weighted loss to compensate for the imbalanced distribution, or focal loss [103], or asymmetric loss [98] to address the imbalance by adjusting the loss contribution between hard samples and easy samples.

3.3.3 Model Explainability

Despite the impressive results that have been achieved by CNNs, they are still seen as a ”black box” due to their lack of decomposability [104]. The inability to explain why they are successful or why they fail in certain situations creates a trust issue with CNN-based intelligent systems, particularly in industrial fields where accuracy with minimal error allowance is essential [105]. In order to utilize CNNs-based models in industry, it is essential to increase the transparency of these models and explain their predictions. Several studies have been conducted to visualize CNN decisions, such as [106, 107, 108, 104], which emphasize the pixels that have the most influence on the predictions, thus helping to better comprehend the model's decisions. These approaches can be mainly divided into perturbation-based methods and gradient-based methods [109]. Perturbation-based methods involve making small changes to the input data and measuring how these changes affect the model output. This technique was applied in CNNs by Zeiler and Fergus [106], who proposed Deconvnet to project the feature activations back to the input space by reconstructing the input from its output. However, these methods tend to be computationally intensive as the number of features increases [110] due to the need to predict in perturbed data. Compared to the perturbation-based method, gradient-based methods are computationally efficient as they only require backpropagation for computing gradients [109]. Simonyan et al. [107] visualized CNNs by extracting saliency maps using gradients of the output over the input. However, this method is not class-discriminative [104]. Zhou et al. [108] addresses this limitation by using global average pooling to produce a class Activation Mapping (CAM). It improves the accuracy of heat maps, however it requires the neural network architecture to have a specific structure, typically consisting of a series of convolutional and pooling layers followed by a global average pooling layer and a final fully connected layer for classification [108]. Grad-CAM [104] is an extension of CAM that addresses the limitation of architecture compatibility, by using gradient information directly to calculate importance scores. In general, graident-based methods can provide visual explanation, thus allowing us for fast checking. Besides, compared with perturbation-based methods, gradient-based methods provide high computational efficiency.

3.4 Methodology

Due to the limited size of the dataset and the data imbalance problem, a multi-label oxide classification that takes advantage of transfer learning and the promising performance of asymmetric loss is proposed, as illustrated in Figure 3.2. This will be discussed in more detail in the following subsections.

3.4.1 Transfer learning

Neural networks have been receiving increasing attention in recent years due to their ability to automatically uncover representations from raw data without the need for meticulous feature engineering [111]. A basic neural network consists of an input layer, a hidden layer with many interconnected neurons, and an output layer [43]. It was proven that a single hidden layer can solve any continuous problem [112], but its capacity to represent data was not as strong as the deep neural networks (DNNs) that we are familiar with today, however, there was no effective

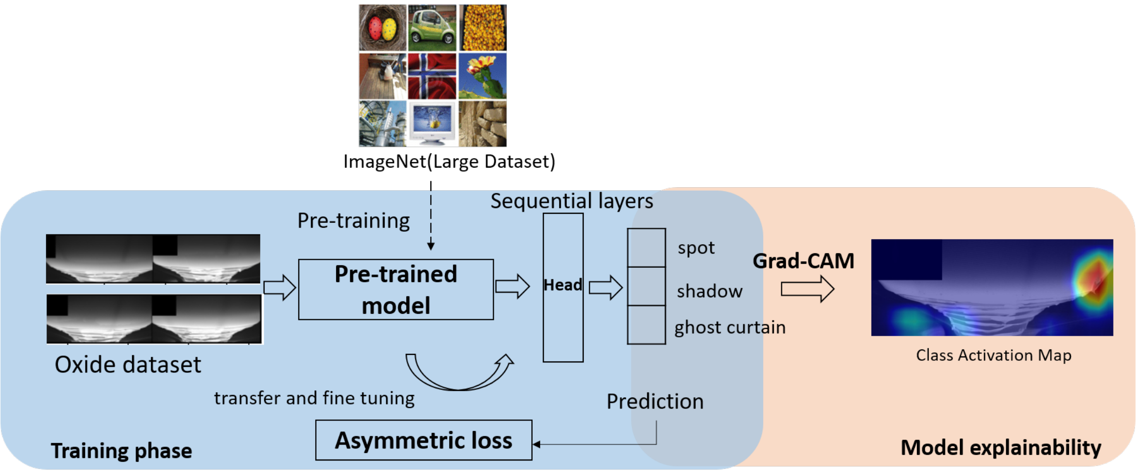


Figure 3.2: Overall architecture of multi-label oxide classification [25]. The features learned on large-scale dataset are transferred to our network by transfer learning while asymmetric loss is employed to address the data imbalance problem. To build the trust on the proposed model for its application in the industry, Grad-CAM is used to enhance the transparency of model decision process.

way to train a DNN back then [113]. The emergence of back-propagation has been a pivotal moment in the shift from shallow to Deep Learning (DL) with multiple hidden layers [113]. This, combined with the greater computing power, has enabled DL to learn complex data representations without the need for manual feature engineering [111], leading to its prevalence. However, with increasing model complexity, the large number of model parameters need to be optimized, making the performance of deep networks highly dependent on the size and quality of the training dataset. Therefore, a substantial amount of data is necessary for each specific task to train a data-demanding model from scratch, which is inefficient and costly in terms of resources. Transfer learning has emerged as an alternative to address this issue. It was demonstrated that the features that many deep neural networks learn from large datasets appear to be universal, thus they can be applicable for similar or even dissimilar tasks [89, 90]. Hence, the core idea of transfer learning is to take advantage of the features and low-level statistics that have been learned from a large-scale dataset and apply them to a different dataset by copying the first n layers of a pre-trained network to the first n layers of the target network [89, 90]. One can choose to either freeze these layers by keeping their weights fixed or fine-tune these layers by optimizing their weights with back-propagation during training on the target dataset. Transfer learning can expedite the training process for a different task with limited data. Additionally, it can help to prevent overfitting since the initial layers have already been trained to recognize general features. In common practice, Transfer learning starts by selecting a base model that has been pre-trained on a large-scale dataset, and then adapting the model to make accurate predictions on the target dataset.

At present, many of the most advanced CNN models can provide excellent performance in computer vision applications. In view of the potential for real-time detection, a model that is both high-performing and computationally inexpensive is desired. To this end, we have compared several popular CNN models [114] and selected ResNet50 and InceptionV3 as our base models.

ResNet50 is part of the ResNet family of residual neural networks, which are designed to tackle the issue of vanishing gradients in extremely deep neural networks [86]. As deeper networks are considered to obtain better model performance, a challenge that hampers exploration is the problem of vanishing or exploding gradients [86]. This is because of repeated multiplication by weights and activation derivatives during the back-propagation of the gradients to earlier layers [115]. ResNet tackled the issue by introducing residual connections, also known as skip connections or shortcut connections (as seen in Figure 3.3), which enabled gradients to flow more easily through the network during training [86]. ResNet50 is a variant of the ResNet architecture that has 50 layers in total, with every two layers connected, forming a residual block.

InceptionV3 is part of the Inception family of deep neural networks [116]. The Inception network was a major breakthrough in the advancement of CNNs, addressing the challenges and complexities that arose due to the data having a high degree of variability. The Inception networks capture multi-scale features by multiple parallel convolutional and pooling layers of various kernel sizes, enabling the network to process information at various spatial scales simultaneously [117], as seen in Figure 3.4. InceptionV3 is an extension of Inception network that further improves the computational efficiency by factorizing convolutions with a larger kernel size into smaller convolutions and asymmetric convolutions [116].

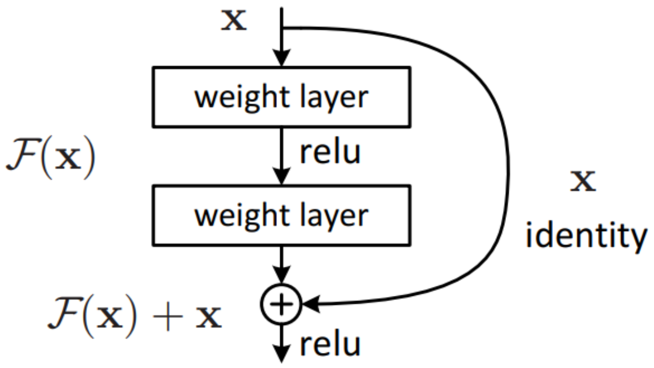


Figure 3.3: A building block with a shortcut connection [86]. The output of layers is added to the original input. By perserving the gradient information to the output, the shortcut connection can ensure the gradients can flow effectively from the output to the input of the block, thus facilitating learning.

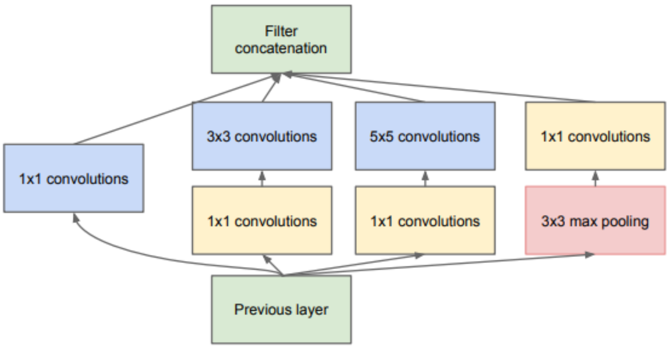


Figure 3.4: Inception module with dimension reductions [117]. Multiple parallel convolutional layers with various kernel size enable the network to capture features at different spatial resolutions.

In this study, the parameters learned from natural images of ImageNet database would be utilized for knowledge transfer. Although the images in the oxide dataset appear to be different from the source domain, it is shown in [90] that it is still possible to enjoy the benefits of pre-trained weights by fine-tuning due to the reuse of features and low-level statistics. Therefore, fine-tune technique would be employed for the transferred models. The base model initially trained on ImageNet had 1000 output units in its output layer to classify images into 1000 classes. To make the model suitable for our task, the original output layer was replaced with the sequential layers outlined in Table 3.1. To prevent the over-fitting on the oxide dataset, the dropout layer is used for regularization. Besides, in this study, adaptive learning rate strategy would be utilized during fine-tuning to achieve better convergence and prevent negative transfer. The experiments would utilize the one-cycle learning rate policy [118], which is a learning rate schedule that rapidly increases the learning rate to a peak to rapidly adjust to the target dataset and then gradually decreases to stabilize the learning process. Furthermore, to mitigate the overfitting problem, image augmentation would be considered to increase the amount and diversity of the training data by adding random changes over different batches.

Table 3.1: Sequential layers are employed to substitute the output layer of pre-trained models. The backbone features refer to the intermediate representation prior to the output layer, which is extracted by the base model [25].

Layer	Input features	Output features	Activation	Dropout
Linear	backbone features (depends on the base model)	256	ReLU	0.2
Linear	256	3 (the number of oxide classes)	-	-

3.4.2 Loss function

Binary cross entropy (BCE) loss is commonly used in multi-label classification, where each logit of the output vector p_i from the network represents probability of the corresponding class i. The total loss is computed by aggregating the BCE loss from all labels, as seen in Eq 3.1:

$$L_{bce} = -\sum_{i=1}^{M} y_i \log{(p_i)} - \sum_{i=1}^{M} (1 - y_i) \log{(1 - p_i)} \tag{3.1}$$

The ground-truth of $i - th$ class, is represented by y_i, with the total number of classes being denoted by M.

However, BCE treats each class equally, which makes it less effective in dealing with data imbalanced problem. To mitigate the influence of imbalances on model training, Focal Loss (FL) was proposed to reduce the error from easy samples and applied to object detection [103]. Specifically, Focal Loss includes a focusing parameter γ to reduce the contribution of easy samples to the error. In the field of classification, the terms 'easy' and 'hard' are often used to denote the level of difficulty for a model to accurately predict a sample. Generally, hard samples can include images that are poor quality or are from a rare class, making the decision-making process more difficult. On the other hand, easy samples are those with distinct features or from a predominant class, allowing the model to make a confident prediction.

$$L_{fl} = -\sum_{i=1}^{M} y_i (1 - p_i)^{\gamma} \log{(p_i)} - \sum_{i=1}^{M} (1 - y_i)\, p_i^{\gamma} \log{(1 - p_i)} \tag{3.2}$$

where γ represents the focusing parameter, designed to reduce the loss contribution from easy sample and thus enable the network concentrate on the misclassified hard samples.

Nevertheless, since the same focusing parameter γ is applied to both positive and negative samples, the contribution of positive samples to the loss is also reduced. This can be a problem in the case of imbalanced datasets, where the positive samples of the rare classes only make up a small part of the data [98]. To address this problem, asymmetric loss (ASL) [98] was developed to decouple the focusing parameters of the positive and negative samples.

$$L_{asl} = -\sum_{i=1}^{M} y_i (1 - p_i)^{\gamma_+} \log{(p_i)} - \sum_{i=1}^{M} (1 - y_i)\, p_{i-m}^{\gamma_-} \log{(1 - p_{i-m})} \tag{3.3}$$

$$p_{i-m} = max\,(p_i - m, 0) \tag{3.4}$$

Where γ_+ and γ_- denote the focusing parameters for positive and negative samples, respectively. The probability margin m is employed to carry out hard thresholding of extremely simple negative samples.

In this case, considering the promising performance of asymmetric loss in dealing with imbalanced data, asymmetric loss would be employed.

3.4.3 Visualization by Grad-CAM

Despite the potential of CNNs to achieve impressive results, the lack of transparency in their internal decision making makes it difficult to comprehend the outcomes. This issue poses a trust dilemma in intelligent systems based on CNNs, particularly in industrial fields where accuracy with a low error allowance is essential [105]. In order to adopt CNNs-based models in industry, it is essential to increase the transparency of the models and explain why they make such predictions. Class Activation Map (CAM) [108] is a technique used to uncover the reasoning process by pinpointing the distinguishing regions with a global average pool. However, due to the need for feature maps to be directly connected to softmax layers, CAM is only applicable to CNNs without any fully connected layer [104]. To address this issue, the Gradient-weight Class Activation Map (Grad-CAM) was developed as a generalization of CAM to generate visual explanations for any CNN-based models using gradient information [104]. Hence, Grad-CAM would be used to provide an interpretive explanation.

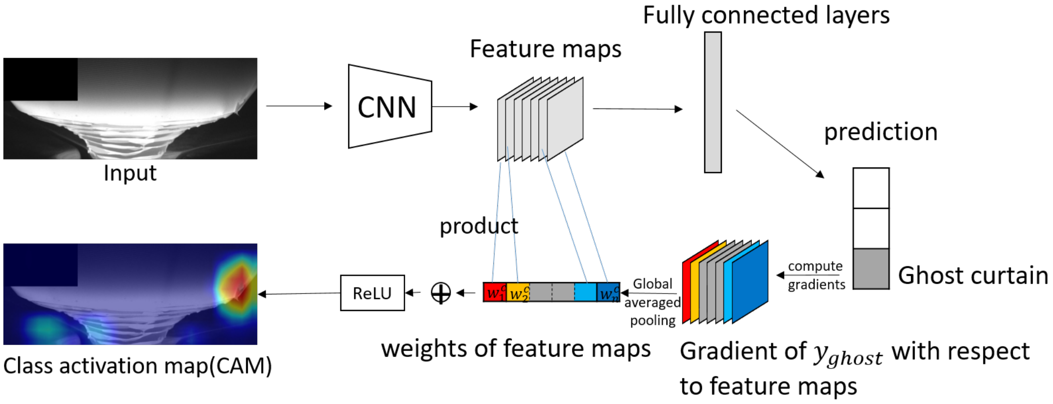


Figure 3.5: The principle of Grad-CAM with taking ghost curtain class as an example. Gradients of the ghost curtain class with respect to the feature maps of the last convolutional layer are used to generate a heatmap that highlights the important regions of the image [25].

The illustration in Figure 3.5 demonstrates the concept of Grad-CAM with the ghost curtain class as an example. To begin, the gradient for the ghost curtain class, $\frac{\partial y_{ghost}}{\partial A^i}$ for the i^{th} feature map is calculated with respect to the feature map, denoted as A^i. The global average pooling is then used to obtain the weights of the feature maps.

$$\alpha_i^{ghost} = \frac{1}{Z} \sum_{w=1}^{W} \sum_{h=1}^{H} \frac{\partial y_{ghost}}{\partial A_{wh}^i} \tag{3.5}$$

The quantity of pixels in the feature map is denoted by Z. The activation at the coordinates (w, h) in the feature map A^i is represented by A_{wh}^i.

The coarse heatmap in Eq 3.6 is generated by combining the feature maps with a weighted sum and then applying ReLU. The final Grad-CAM heatmap is obtained by resizing it to the size of the original image. This heatmap provides an explanation of the decision-making process of the model. The regions with higher values, which appear red in the heatmap, indicate that those features are the most influential in the model prediction. By visualizing these regions, we can understand where the model is looking when making the prediction. This can increase the trustworthiness of the model by verifying the model predictions and the regions that are activated.

$$L^{ghost}_{Grad-CAM} = ReLU\left(\sum_{i=1}^{N} \alpha_i^{ghost} A^i\right) \tag{3.6}$$

3.5 Experiments

3.5.1 Data gathering and preprocessing

The optical camera equipped in the FZ vision system films in monochrome at a rate of about 20 frames per second, capturing grayscale images with a resolution of 1600x1200 pixels. However, in this study, the images captured by the camera at the beginning of the cone phase were chosen for analysis, where an oxide typically becomes evidently. The images in the database had already been processed by the FZ automation system, which included some confidential information in the upper left corner such as the timestamp and machine serial number. To protect this information, a black rectangle was placed on top of it. On the other hand, this information could be potentially used to differentiate oxides during model training, for example to identify which oxides are more common in recent production or which are more frequently seen in a particular machine. To ensure that the model focuses only on the oxide region itself and not on other details, a black box was also used to prevent the model from learning from this information. The dataset was preprocessed by eliminating noises, cropping, and resizing the images. Subsequently, the images were cut into a fixed-sized region of interest and adjusted to the shape of (320x800), as illustrated in Figure 2.18. We did not resize to a square size, which is commonly used in image classification, as it could lead to distortion of spatial information or even the loss of features of small oxides.

The dataset was divided randomly but in a stratified manner into three parts: 70% for the training split, 10% for the validation split and 20% for the test split. All images were first scaled to the range [0,1] by dividing them by 255, and then normalized to the range [-1, 1] by subtracting the mean of 0.5 and the standard deviation of 0.5. These values were not the actual values of the dataset, but rather estimated values to center the data distribution around 0. During training, the training split data was augmented by randomly changing the brightness and contrast across different batches. This was done to increase the amount and diversity of the training dataset, thus reducing the risk of model overfitting.

3.5.2 Baselines and implementation details

ResNet50 and InceptionV3 were selected to assess the suggested approach, taking into account the need for lightweight models for real-time implementation. The weights of the models were taken from the trained weights on the ImageNet database. The output layer of each model was then replaced with the sequential layers listed in Table 3.1. To compare the results, experiments were conducted with the following baselines: models that were trained from random initialization with Binary Cross Entropy (BCE) loss (denoted as BCE-S), pre-trained models trained from transferred weights with BCE loss (denoted as BCE), pre-trained models with Focal Loss (FL) and Asymmetric loss (ASL), as well as pre-trained models using Powerset (PS) and Binary Relevance (BR). It is worth mentioning that all pre-trained models were fine-tuned. To ensure fairness, FL and ASL had the same focusing parameter γ of 2 for negative samples. As suggested in [98], $\gamma_+ = 0$ was set in ASL, while $\gamma_+ = \gamma_- = 2$ in focal loss. Additionally, the probability margin m was set to 0.05 in ASL.

Considering the models trained from scratch may need more iterations to achieve convergence, the pre-trained models and the models trained from random initialization were trained for 50 epochs and 100 epochs, respectively. Adam optimizer with (β_1, β_2) = (0.9, 0.999) and a one-cycle policy, with a maximum learning rate of 1e-4 were used. The batch size was set to 32. During training, the performance of each model was monitored by evaluating the loss on the validation split.

3.5.3 Evaluation metrics

In comparison to single-label classification, which usually uses confusion matrix-based metrics, MLC necessitates more intricate metrics since multiple classes can be present at the same time. The evaluation metrics for MLC can be divided into two categories: instance-wise metrics and label-wise metrics [75]. In this research, we opted for subset accuracy and hamming score, both of which are instance-wise metrics, as well as macro average F1 score, a label-based metric. Since these metrics are calculated using a fixed threshold, they may be sensitive to the threshold selection. Therefore, we also selected a threshold-independent metric, mean average precision.

- Subset accuracy (SA). SA refers to the proportion of samples that are correctly classified, or in other words, predictions that are exactly the same as the actual labels [75].

$$SA = \frac{1}{N}\sum_{i=1}^{N} I\left(Y_i = \hat{Y}_i\right) \tag{3.7}$$

- Hamming score (HS). HS is calculated based on hamming loss (HL). The Hamming score is a popular metric for evaluating the accuracy of multi-label classification tasks. It is calculated by taking the symmetric difference between the predicted labels and the ground truth labels, using the XOR operation in Boolean logic [100].

$$HL = \frac{1}{N}\sum_{i=1}^{N} \frac{\left|Y_i \Delta \hat{Y}_i\right|}{M} \tag{3.8}$$

$$HS = 1 - HL \tag{3.9}$$

Where Δ denotes XOR operation. The greater the HS value, the better the outcomes.

- Micro average F1 score (F_{mi}). F_{mi}, is given by the balanced measure of micro average precision and micro average recall [119]. This metric considers each sample equally, however, this can also make it vulnerable to being heavily impacted by the most prominent classes.

$$F_{mi} \ = 2\frac{P_{mi}R_{mi}}{P_{mi} + R_{mi}} \tag{3.10}$$

$$P_{mi} = \frac{\sum\limits_{i=1}^{M} TP_i}{\sum\limits_{i=1}^{M} (TP_i + FP_i)} \tag{3.11}$$

$$R_{mi} = \frac{\sum\limits_{i=1}^{M} TP_i}{\sum\limits_{i=1}^{M} (TP_i + FN_i)} \tag{3.12}$$

- Macro average F1 score (F_{ma}). F_{ma} refers to the harmonic mean of macro average precision and macro average recall. Unlike F_{mi}, F_{ma} gives equal weight to each class [119].

$$F_{ma} = 2\frac{P_{ma}R_{ma}}{P_{ma} + R_{ma}} \tag{3.13}$$

$$P_{ma} = \frac{1}{M}\sum_{i=1}^{M} P_i, \ where \ P_i = \frac{TP_i}{TP_i + FP_i} \tag{3.14}$$

$$R_{ma} = \frac{1}{M}\sum_{i=1}^{M} R_i, \ where \ R_i = \frac{TP_i}{TP_i + FN_i} \tag{3.15}$$

- Mean average precision (mAP). The metric mAP stands for the mean of average precisions (AP) across all categories.

$$mAP = \frac{1}{M}\sum_{i=1}^{M} AP_i,\ where\ AP_i = \int_0^1 p_i\left(r_i\right) dr_i \tag{3.16}$$

  p_i and r_i represents the precision and recall for class i, respectively. As demonstrated in Equation 3.16, the area under the precision-recall curve is also represented by AP_i. Since mAP takes into account all possible thresholds, it is usually the most important metric for assessing the overall performance of a model. The higher the mAP, the better the performance of the model.

Among these metrics, mAP and F_{ma} would be emphasized considering their comprehensive nature.

Due to the nondeterministic nature of neural network optimization, which leads to uncertainty in model performance, a fixed random seed is necessary to ensure repeatable results. To evaluate the model's resistance to randomness, especially when dealing with small datasets, each experiment was repeated three times using different random seeds (0, 42 and 2023).

3.6 Results and Discussion

3.6.1 Comparison results

Table C.1 compares the performance of two model architectures with different methods on the test set. The results are also depicted in bar charts, as seen in Figures 3.6 and 3.7. The highest and second highest metrics among the different methods are highlighted in red and orange, respectively. All results were obtained with a global threshold of 0.5 for each class, except for mAP. It should be noted that, for the baseline PS, the prediction for each class is inferred from the predictions of class combinations. For instance, if the prediction of "spot and shadow" is above the threshold of 0.5, then both the spot class and the shadow class are activated and denoted as "1", while the rest of classes are noted as "0". Since there is no threshold to consider, mAP is not available for the baseline PS.

As is demonstrated in Figures 3.6 and 3.7, the models with pre-trained weights outperformed those trained from random initialization. The highest mAP of the models trained from random initialization was 79.39%, while the highest mAP of the pre-trained models was 94.12%, as seen in Table C.1. Additionally, a considerable variance in terms of F_{mi} and F_{ma} can be observed in the models from random initialization, indicating their susceptibility to randomness. These observations suggest that random weight initialization can make it difficult to find optimal values.

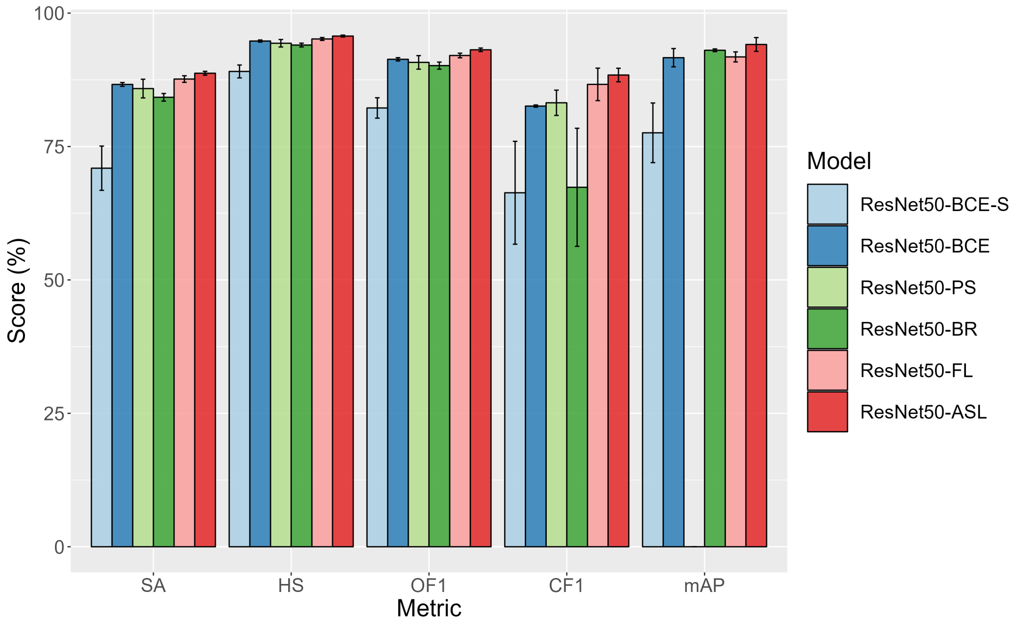


Figure 3.6: Performance of ResNet50 models with their average and standard deviation [25]. The x axis shows the five different metrics and y axis represents the corresponding score in percentage. Different colors in the legend corresponds to different models.

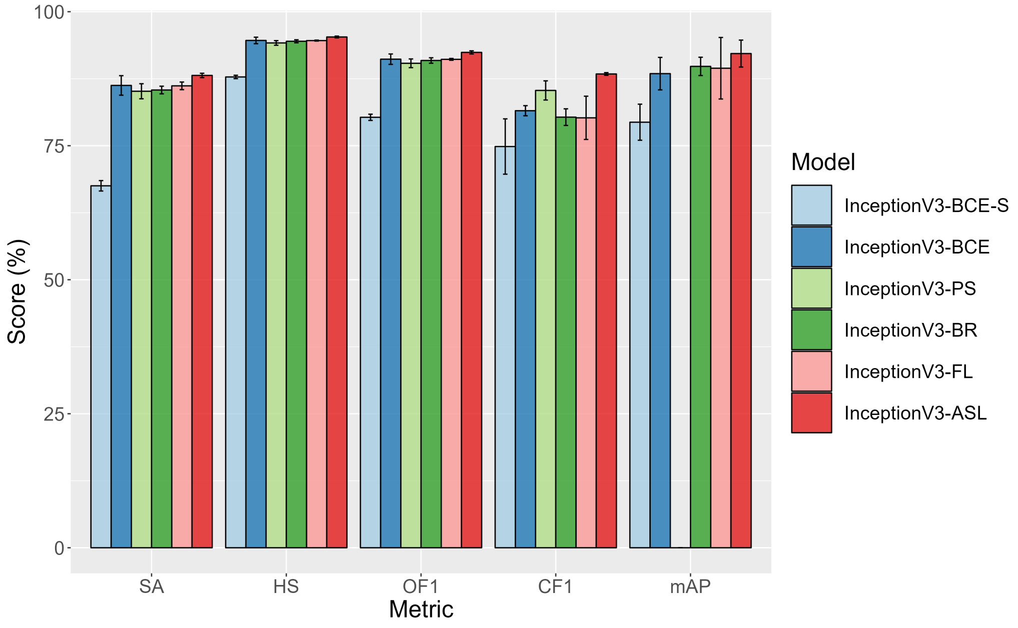


Figure 3.7: Performance of InceptionV3 models with their average and standard deviation [25]. The x axis shows the five different metrics and y axis represents the corresponding score in percentage. Different colors in the legend corresponds to different models.

Among all pre-trained models, ASL outperformed the others in terms of the five averaged metrics. Notably, the highest mAPs achieved by ASL demonstrated its robustness against different thresholds. Furthermore, compared to FL, the highest F1 scores obtained by ASL showed the effectiveness of its asymmetric focus on emphasizing the learning features from positive samples. BR achieved the second highest mAP, but the other metrics did not perform well, suggesting that a specific threshold function might be necessary due to the lack of label consistency.

3.6.2 Effects of various components in the asymmetric loss

In this section, the effects of the asymmetric focusing parameter γ and the probability margin m on asymmetric loss were deeply investigated in this section. Since the oxide dataset size was relatively small, it was not possible to determine the universally optimal hyperparameters based on the experiments. Therefore, the comparison between the conventional BCE loss and ASL under different levels of hyperparameters was mainly focused on, and the conditions of effectiveness of ASL were investigated. Following the recommendation in [98], γ_+ was fixed at 0. Figure 3.8 shows the performance obtained by different values of the probability margin m with six levels of the asymmetric focus parameter γ_- ranging from 1 to 6.

Effect of probability margin

Probability margin m is a tunable hyperparameter in ASL that can attenuate easy samples and reject those that are likely to be incorrectly labeled.

As can be seen in Figure 3.8, the average performance in mAP of the most asymmetric levels was higher than the reference mAPs from BCE loss, along with overlapping uncertainty. However, a decrease in each of the other threshold-dependent metrics was observed with an increasing probability margin, which was particularly noticeable in large asymmetric levels. Specifically, γ_- ranging from 3 to 6 reached their highest points in the probability margin of less than 0.2 in all threshold-dependent metrics, while there was no significant change observed in γ_- of 1 and 2. Generally, the threshold-dependent performances obtained by γ_- below 3 were always above the reference values of BCE loss in the probability margin of less than 0.2.

This might be due to the loss gradients of the negative samples. The loss gradient of negative samples with respect to the input logits z can be seen in Eq 3.18 [98]. Figure 3.9 shows the loss gradient of negative samples at different values of m (with $\gamma_- = 2$). As the probability margin increases, the peak of the loss gradient decreases, resulting in more loss gradients from positive samples and less from negative samples. Additionally, when using a larger γ_-, the loss contribution of negative samples is further reduced, making the loss imbalance from large probability margin more noticeable. Therefore, to achieve better performance than BCE loss, it is

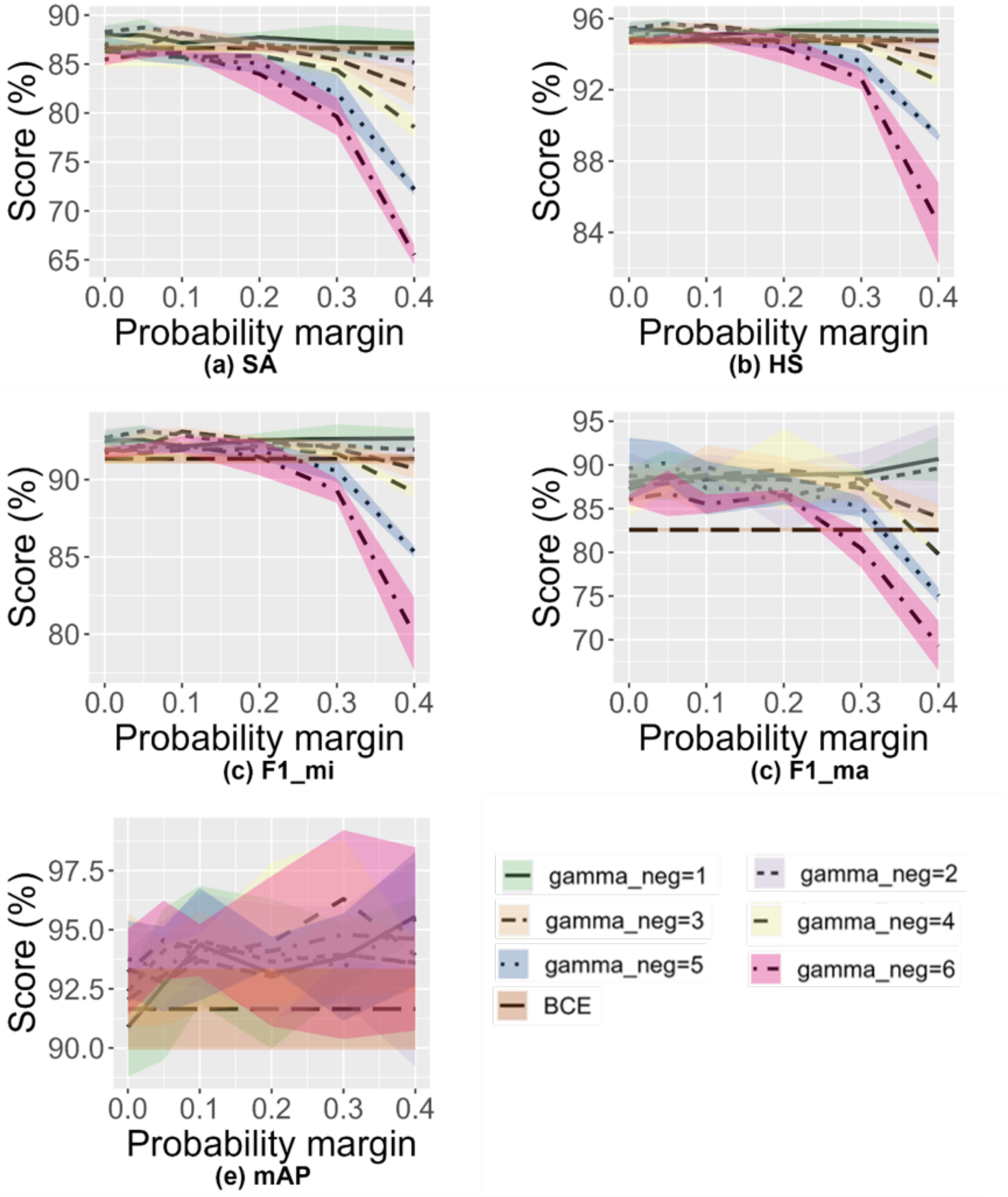


Figure 3.8: The ResNet50 model performance obtained by different values of probability margin m in terms of SA, HS, $F1_{mi}$, $F1_{ma}$, and mAP [25]. The shaded region represents the standard deviation.

recommended to select a smaller probability margin.

$$\frac{dL_-}{dz} = \frac{\partial L_-}{\partial p}\frac{\partial p}{\partial z} \tag{3.17}$$

$$= (p_m)^{\gamma_-}\left[\frac{1}{1-p_m} - \frac{\gamma_- \log(1-p_m)}{p_m}\right] p(1-p) \tag{3.18}$$

Where $p = 1/(1+e^{-z})$ (sigmoid).

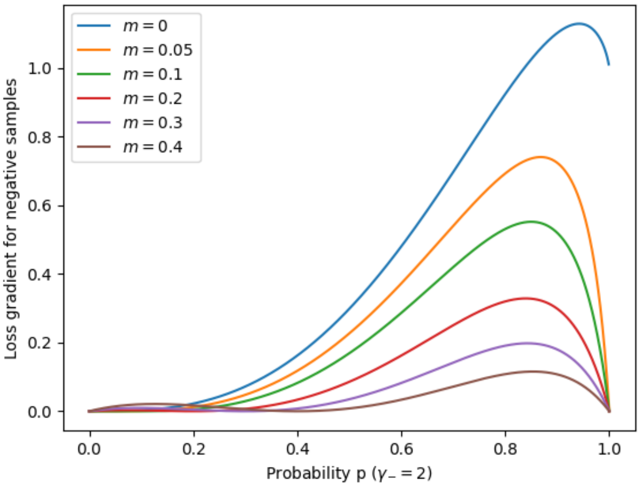


Figure 3.9: Probability Vs. Loss gradient of negative samples under γ_- of 2 [25].

Effect of focusing parameter for negative samples

The focusing parameters γ_- and γ_+ are essential for asymmetric loss. If γ_- is set too low, the downweighting of the easy negative samples is inadequate [98], leading to accumulation of loss gradients from the negative samples. If γ_- is set too high, the contribution of easy negative samples is significantly reduced, and the model may focus excessively on hard samples. Furthermore, model performance can be adversely affected if there is noise in the hard samples.

As can be seen in Figure 3.8, the lower the value of γ_-, the better the performance of the threshold-dependent metrics at the same probability margin, and the more stable the performance at different levels of probability margins. On the other hand, as γ_- increases, the deterioration of threshold-dependent performance with increasing probability margin becomes more evident. Additionally, when γ_- is set to 6, all metrics tend to be more uncertain. This could be related to the fact that the model is more sensitive to the loss contribution from positive samples. Nevertheless, the average mAP of using asymmetric focus was always higher than the reference mAP in BCE loss. This implies that the threshold for larger γ_- may need to be adjusted.

3.6.3 Visual analysis by Grad-CAM

Given the well-trained models, Grad-CAM [104] was employed to assess the relationships between images and labels, thus building trust for application in industry. The heat maps in Table 3.2 were generated from ResNet50-ASL and InceptionV3-ASL, with a random seed of 42. The ground truths and the predictions from the models are displayed at the bottom of the pictures.

As can be seen in Table 3.2, the models using ASL were able to accurately identify the oxide types. Moreover, the heat maps generated from Grad-CAM were able to effectively highlight the features for the activated class. This implies that the heat maps can be used to locate the oxides. In comparison to the Resnet50 model, the InceptionV3 model was able to capture more comprehensive and precise features. This could be attributed to the InceptionV3's strong ability to extract features with different scales of filters. We also showed a misclassified image where the shadow class was incorrectly detected. However, the heat maps still indicated that the interesting regions activated by the shadow class were close to the shadow regions. This could be due to the lower confidence in the shadow class than the threshold of 0.5. This can be improved by optimizing the threshold or exploring the label correlation.

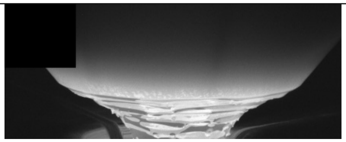
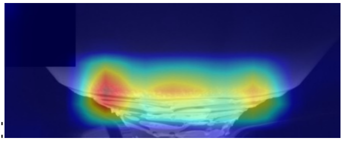
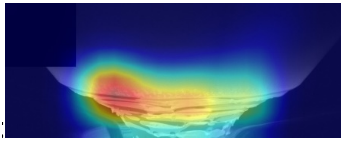
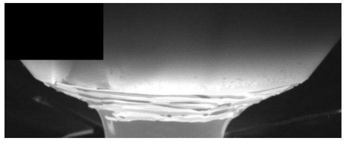
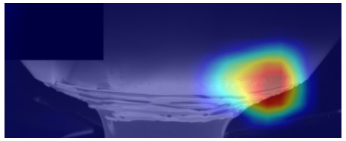
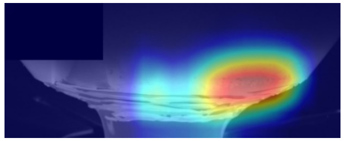
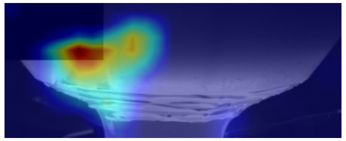
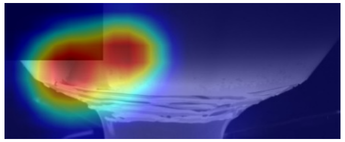
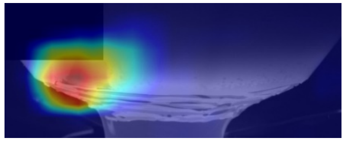
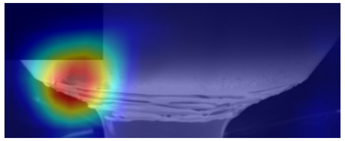
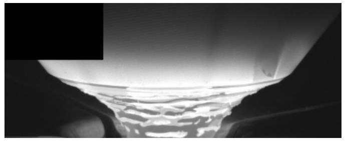
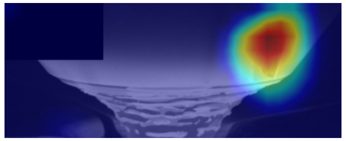
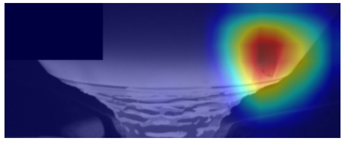
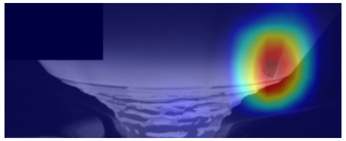
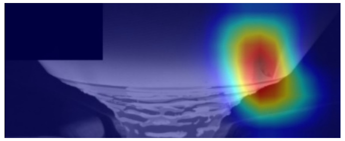
3.7 Deployment

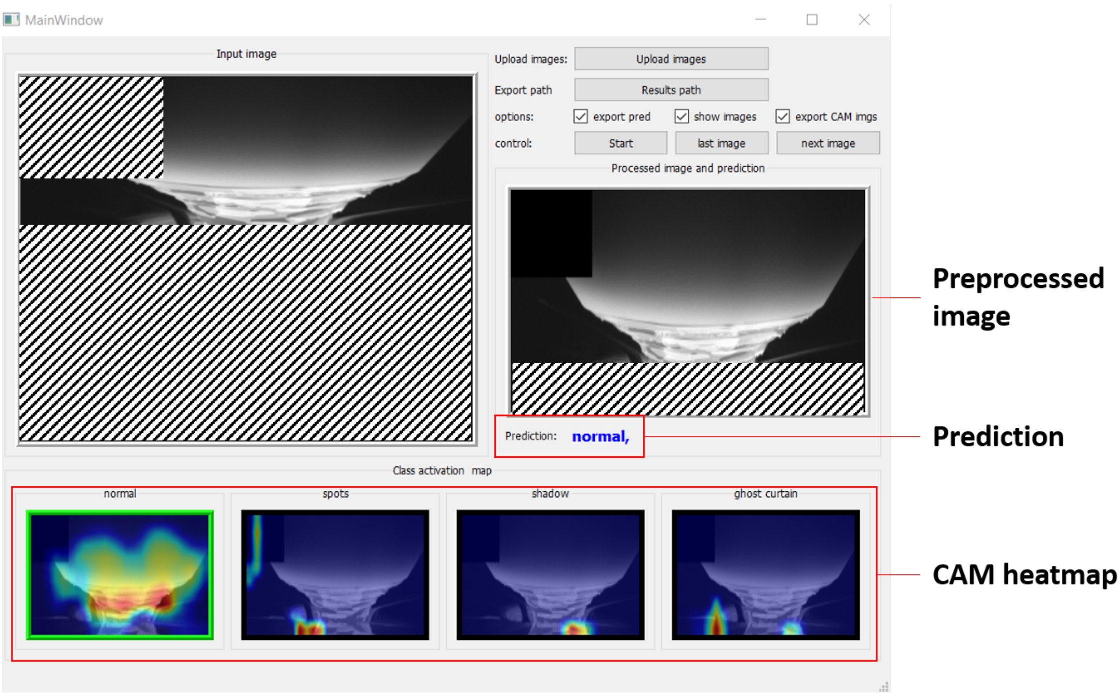
To facilitate the application of the oxide identification in the FZ production, the auther implemented a software that integrates the multi-label oxide classification model and the Grad-CAM for the visualization of the model's decision making. The software currently works offline and can process one image or a batch of images when provided with a well-trained model checkpoint. After inference of the model on the input, the interface would generate the prediction of the model and the associated CAM heatmaps that explain the decision-making process of the model, as seen in Figure 3.10.

3.8 Conclusions

This research examined the categorization of oxides in Float-zone silicon crystal growth production. Three varieties of oxides were identified: spot, shadow and ghost curtain. An oxide dataset was created using FZ images from a vision system integrated on an industrial FZ machine. To address data imbalance and limited dataset size, a method based on transfer learning and asymmetric loss was proposed. The model trained with the pre-trained weights and the asymmetric loss achieved an average subset accuracy of 88.73% and an average mAP of 94.12%, outperforming other baselines. The effectiveness of the asymmetric loss compared to the conventional

Table 3.2: Comparison of Grad-CAM heat maps generated from Resnet50-ASL and InceptionV3-ASL [25]. Predicted classes indicated in green mean they are correctly predicted, while the predictions in red show they should have been found but they were missed by the algorithm.

Input image	ResNet50-ASL		InceptionV3-ASL	
 Ground truth: spot ResNet50-ASL : spot InceptionV3-ASL : spot	spot:		spot:	
 Ground truth: spot, shadow, ghost curtain ResNet50-ASL : spot, shadow, ghost curtain InceptionV3-ASL : spot, shadow, ghost curtain	spot:		spot:	
	shadow:		shadow:	
	ghost curtain:		ghost curtain:	
 Ground truth: shadow, ghost curtain ResNet50-ASL : ghost curtain, shadow (missed) InceptionV3-ASL : ghost curtain, shadow (missed)	shadow:		shadow:	
	ghost curtain:		ghost curtain:	



options: ☑ export pred ☑ show images ☑ export CAM imgs

Prediction: normal,

Figure 3.10: The main interface of the developed software, which enables the analysis of one image or a batch of images (some areas have been obscured due to confidential content). After the model inference, the model prediction and the associated heatmaps for explaining the model decision-making can be obtained on the interface.

BCE loss was also studied. To build trust for the model's integration into industry, Grad-CAM was used to visualize the correlation between inputs and labels.

This study is the initial step in the exploration of oxidation, which provides a basis for further root cause analysis and the potential for automated responses to reduce oxides. In the future, the single-frame oxide classification method presented here could be extended to real-time oxide classification for early detection of oxides, for instance by analyzing in-process video footage captured by the integrated vision system. This would enable the evaluation of oxide formation, growth, and spread, as more images can be obtained from each production run, thus providing a comprehensive understanding of the process and tracking the development of oxides.

CHAPTER 4

ROOT CAUSE ANALYSIS FOR THE OXIDE PROBLEM

"Invisible threads are the strongest ties." - Friedrich Nietzsche

4.1 Motivation

It is not sufficient to merely identify the problem. What is more essential is to discover the fundamental triggers of a problem, thus leading to more effective solutions to address and rectify the persistent oxide problem within the FZ crystal growth production. Section 2.1 and 3.1 have already elaborated on the detrimental impacts of the oxide problem. In summary, the motivation for undertaking this root cause analysis is deeply rooted in the pursuit of excellence in product quality, machine state, and cost reduction. For the FZ process with relatively high production costs, even a marginal improvement in the production efficiency and the crystal yield would result in considerable financial gains. Hence, with these motivations underscoring the significance of root cause analysis of the oxide problem, this chapter focuses on introducing the current state-of-the-art methods for root cause analysis and the methodology applied in this oxide problem case. Finally, the discovered patterns from experimental results can be used for further in-depth exploration of deeper root causes and targeted intervention plans.

4.2 Problem Definition

The aim of this study is to identify the minimum set of root causes $R = R_1, R_2, \ldots, R_m$ that represent the fundamental factors that lead to the occurrence of the oxide problem, given n observations of data from different sources from n FZ production runs, $O = \{O_1, O_2, \ldots, O_n\}$. O_i is the i-th production run data consisting of data from relevant data sources. Addressing the root causes of the problem will result in successful alleviation or resolution of the problem.

4.3 Literature Review

Root cause analysis is a process through which we can understand the fundamental triggers of a problem, thus leading to more effective solutions. Knowledge-driven approaches are widely used in conventional root cause analysis. Knowledge-driven approaches involve using domain-specific expertise, domain knowledge, and human intuition to identify and understand the underlying causes of issues. In general, knowledge-driven approaches are mainly qualitative and semi-quantitative techniques [120]. For instance, examples of knowledge-driven approaches include the Ishikawa diagram [121] which is also called fishbone or cause-and-effect diagram, 5 Why analysis [122] involving repeatedly asking "Why?" to drill down into the underlying causes of a problem, and Failure Mode and Effect Analysis (FMEA) [123] (see Figure 4.1). However, the results from these methods are largely dependent on expert knowledge. Furthermore, these techniques are not equipped to handle the huge and massive amount of data generated by digitalization in manufacturing, making it hard for analysts to identify the underlying causes [124]. Therefore, it is desired to have automatic root cause analysis, which can help analysts perform root cause analysis more efficiently or even reduce the need for domain experts. As introduced before, ML is a good tool for automatically handling a large amount of data and discovering inner patterns from data. Therefore, in this section, we would introduce how ML is applied for root cause analysis by looking into recent relevant literature.

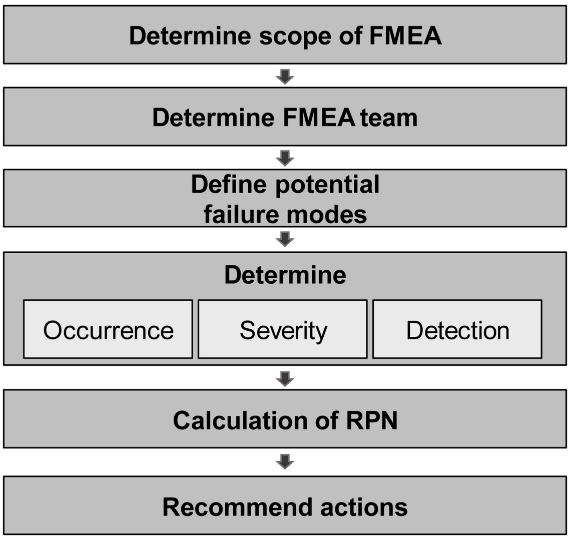


Figure 4.1: Main steps of FMEA [125].

The root cause level can be categorized into three types: location-time, physical causes and log-action [124], as shown in Figure 4.2. Location-time is the simplest level of the root cause, as it only needs to define the where and when the problem is. This is widely used in a continuous process where many sub-processes and machines are involved. Although it cannot point out exactly what the root cause is, location-time information can make it efficient to discover the root cause [124]. The second level of the root cause is equivalent to physical causes, which reflect the physical characteristics of the root cause (what physically caused the problem) [124].

Although the physical causes of a problem have been identified, the root cause analysis is not complete. The reasons why these physical factors occurred are still unknown, making it difficult to begin making improvements. Therefore, the next level of root cause is to find out both what are physical causes and why they happened [124]. In this case, this dissertation attempts to discover the root causes at the highest level, log-action, as it is necessary to understand both the physical causes and the reasons of their emergence, thus allowing corrective plans to mitigate the problem.

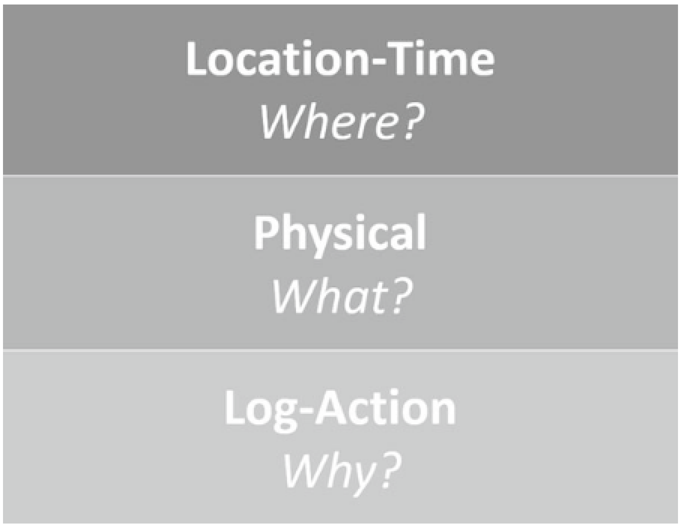


Figure 4.2: Root cause levels in root cause analysis: location-time, physical causes and log-action [125].

Depending on the level of root cause, the solution of root cause analysis can also be categorized into three categories: *Factors* ⇒ *Problem*, *Factors* ⇒ *Root causes* and *Root causes* ⇒ *Root causes* [124], as shown in Figure 4.3.

Factors ⇒ Problem. Most root cause analysis ends at *Factors* ⇒ *Problem*. *Factors* ⇒ *Problem* more or less corresponds to the level of root cause of location-time or physical causes, which aims at finding the factors associated with the problem. The assumption is that if there is a strong association between the discovered factors and the problem, the factors are likely the root causes (physical causes). This solution requires a dataset combining the factors that could potentially contribute to the problem and the problem labels. The most used techniques in the *Factors* ⇒ *Problem* are Association Rule Mining (ARM) [126] and building a classifier or regression model with interpretable structures.

ARM, a type of unsupervised learning, can rapidly uncover relationships between different attributes by examining their co-occurrence [43]. The fundamental concept behind ARM is to identify the associations between different items in a dataset using some statistical measures

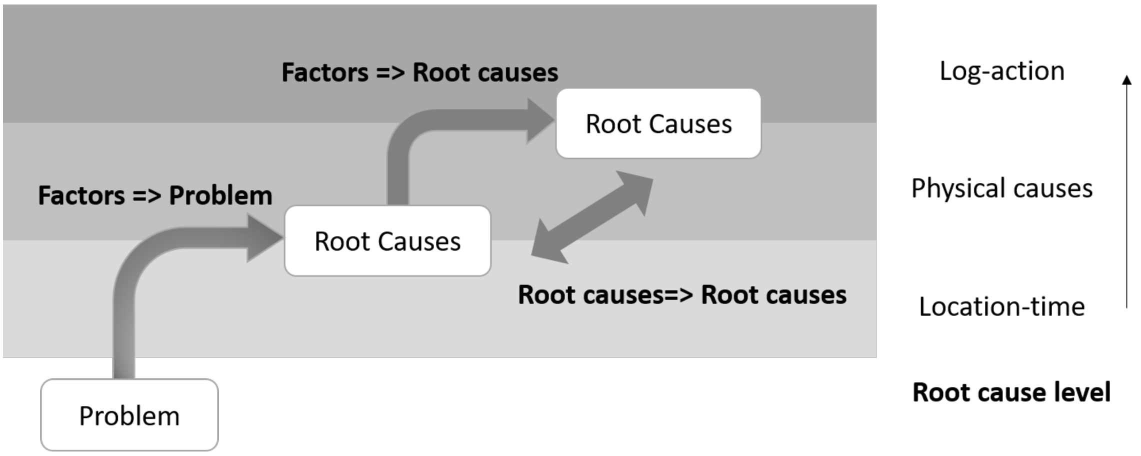


Figure 4.3: Three solutions of root cause analysis: *Factors* ⇒ *Problem*, *Factors* ⇒ *Root causes* and *Root causes* ⇒ *Root causes*, adapted from [124].

[43]. The frequent patterns extracted by ARM are in the form of $X \to Y$, which corresponds to antecedent and consequent, relatively. These frequent patterns can provide valuable insights for decision-making.

Another technique is to build a classifier or regression model and then interpret the decisions of the model. Common classifiers with interpretable structures are tree-based classifiers, such as decision tree or random forest [124]. The results from these interpretable classifiers are intuitively understandable to humans. Furthermore, discriminant factors from these classifiers indicate their strong impacts on the problem, thus implying that they could possibly be root causes. However, better interpretability does not always come with better accuracy. Figure 4.4 shows the relationship between performance and the complexity of different machine learning methods. As seen, the better intrepretability of the model for humans to interpret, the worse the model performance, and vice versa. Given this, it is difficult to trust decision-making by tree-based methods. Recently, Explainable Artificial Intelligence (XAI) [127] is being developed to address this issue, and it has been gaining attractions in recent years. XAI is a branch of Artificial Intelligence that concentrates on making Artificial Intelligence models understandable and clear to humans. The emergence of XAI has enabled us to create intricate models with excellent performance, while allowing us to trust the model by examining the explanations of its decisions, providing us with some insights into the root causes.

Factors ⇒ Root causes. Instead of focusing on the problem, *Factors* ⇒ *Root cause* aims to identify the conditions that correspond to the root causes that were already identified (such as location-time or physical causes) [124]. Hence, it is actually a deeper level of analysis. *Factors* ⇒ *Root cause* corresponds to the root cause level of physical causes (what physical factors caused the problem) or log-action (finding out why they happened). By establishing a link between the

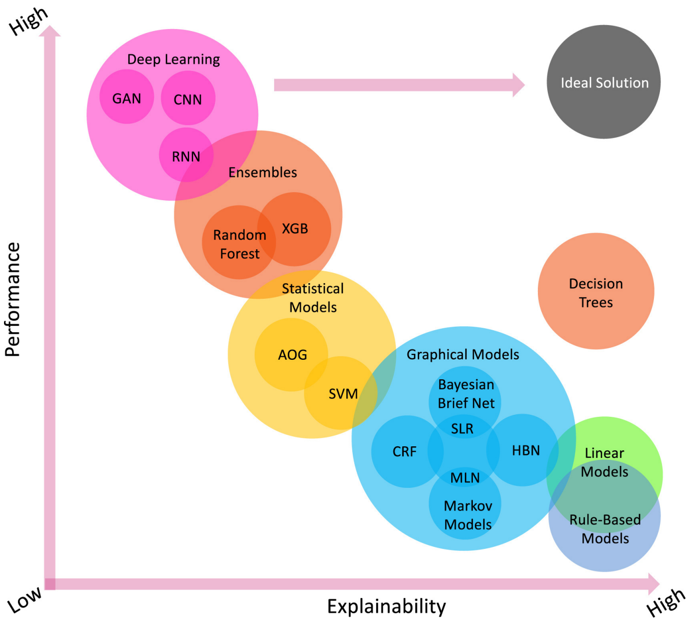


Figure 4.4: Trade-off between accuracy and interpretability of machine learning algorithms [128].

factors and the underlying causes, it is possible to trace back from the visible factors to the root of the problem that needs to be tackled in order to prevent the problem from happening again. This approach is more comprehensive as it attempts to get to the heart of the matter rather than just dealing with its surface symptom.

The difference between *Factors* ⇒ *Problem* and *Factors* ⇒ *Root cause* lies in the objective to be studied. Therefore, the techniques mentioned in *Factors* ⇒ *Problem* are also applicable here. Since the objectives are root causes rather than the problem, the focus in *Factors* ⇒ *Root cause* lies more on accuracy than on interpretability [124]. Besides, the data used for this solution are usually high-dimensional and noisy [124]. Hence, Neural Network is widely used in this solution due to its high model performance and its ability to deal with high-dimensional data [124]. In terms of its low interpretability problem resulting from the black-box model, XAI can mitigate this problem by providing understandable and interpretable explanations for their decisions and actions. Additionally, looking at the interpretable explanations from XAI, it is possible to identify which factors may have a greater influence on the problem or its underlying causes.

Root causes ⇒ Root causes. Based on the findings from *Factors* ⇒ *Root cause*, *Root causes* ⇒ *Root causes* aims to uncover the causal sequence among the relevant root causes, thus offering us a better understanding of which root cause should be prioritized in order to make an informed decision. Compared with the previous two solutions, *Root causes* ⇒ *Root causes* emphasizes more causality than interpretability and accuracy. However, since it is difficult and time consuming to identify the underlying causes, there is a lack of research in this category [124].

Reflecting on our situation, the most effective and practical solutions for our oxide problem are *Factors* ⇒ *Problem* and *Factors* ⇒ *Root cause*. This enables us to gain a comprehensive understanding of the oxide problem and develop a long-term solution while preventing recurrence of the problem. This chapter would mainly focus on *Factors* ⇒ *Problem*.

4.4 Methodology

Though traditional knowledge-driven root cause analysis methods have been effective in many scenarios, the results of these methods are largely dependent on expert knowledge. In addition, as the amount of relevant available data increases, it would be much more time consuming and difficult for experts to analyze them. Given these challenges, we turn to an integration of knowledge-driven root cause analysis and data-driven approaches of root cause analysis that rarely rely on empirical knowledge. This chapter focuses mainly on the solution of **Factors** ⇒ **Problem** which aims at determining the fundamental factors associated with the problem. To this end, the Ishikawa diagram is utilized to identify the data associated with the oxide problem, on which the ARM is chosen to conduct root cause analysis on the data.

The first step of root cause analysis is to gather the relevant data that are associated with the problem. Ishikawa diagram [129] is employed to identify potentially relevant variables causing the oxide effect. Ishikawa diagram, also known as the fishbone diagram or cause and effect diagram, was introduced by Professor Kaoru Ishikawa of Tokyo University in the 1940s [129]. Ishikawa diagram is widely used to identify potential causes of defects or failures of products in the engineering industry [129]. To create the Ishikawa diagram, the problem or effect needs to be clearly defined on the right hand side, as the head of the diagram, as seen in Figure 4.5. It is essential to recognize the key categories of factors that could be causing the issue, and these key categories would be set as the branches that contribute to the problem. The "4M and 1E" approach (Man, Method, Machine, Materials and Environment) is commonly used in the engineering industry [129]. Then the people involved would brainstorm all the potential causes that fall under each major category and add them to complete the diagram. Here, expert knowledge would be utilized to identify potential factors by establishing an Ishikawa diagram for the oxide problem. These factors would then be collected for subsequent ARM analysis.

ARM originated in the retail sector as a way to recognize purchasing habits and enhance sales strategy [126]. Since then, its applications have broadened significantly to other areas such as medical diagnosis [131], finance [132], recommendation systems [133], etc. ARM is a data-driven approach that can provide quantitative evidence of relationships between variables, allowing for

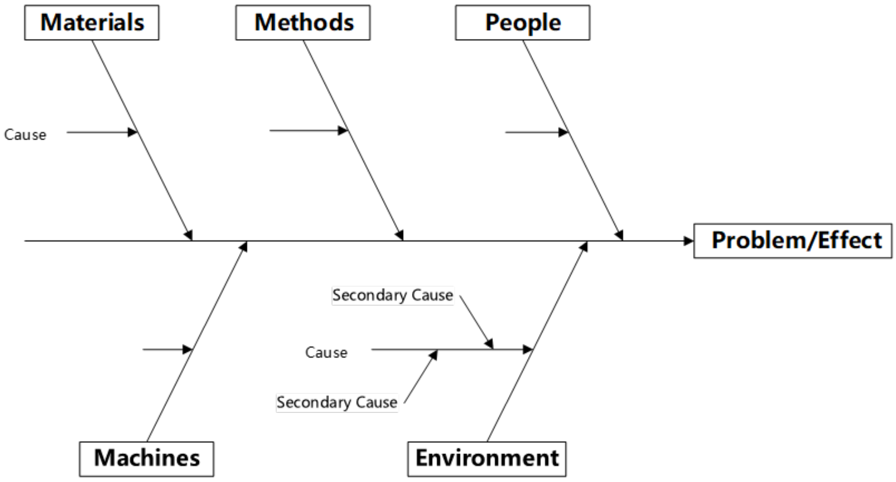


Figure 4.5: A generic representation of the Ishikawa diagram, adapted from [130].

discovering hidden relationships within the data that might have been overlooked. The frequent patterns extracted by ARM are in the form of $X \Rightarrow Y$, which corresponds to antecedent and consequent, also known as Left Hand Side (LHS) and Right Hand Side (RHS), relatively. The frequent patterns are then examined by a minimum threshold of statistical measures, such as Support and Confidence and Lift [134]. The larger these measures, the more robust the rule is. Therefore, one only needs to look into the strong rules extracted from ARM, and examine if they are related to the source of the problem using expert knowledge. However, it should be noted that ARM can only handle binary or categorical attributes, which is not common in manufacturing data. Therefore, if ARM is applied, the manufacturing data should be processed and converted into binary or categorical data.

The entire pipeline of ARM application can be seen in Figure 4.6. In this study, the application of ARM consists of three steps: data collection, data pre-processing, ARM and rules visualization.

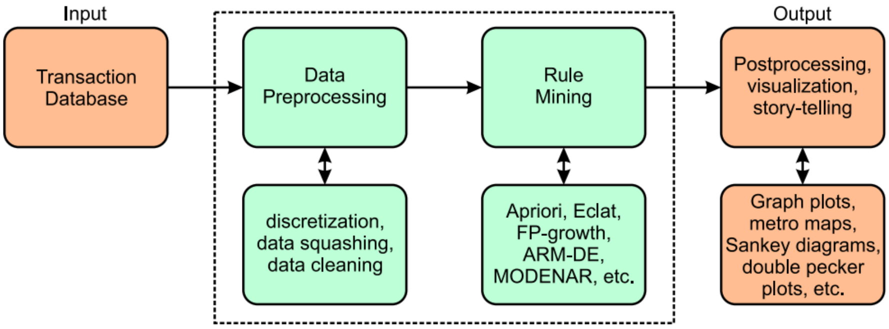

Input
Transaction Database
Data Preprocessing
Rule Mining
Output
Postprocessing, visualization, story-telling
discretization, data squashing, data cleaning
Apriori, Eclat, FP-growth, ARM-DE, MODENAR, etc.
Graph plots, metro maps, Sankey diagrams, double pecker plots, etc.

Figure 4.6: The pipeline of applying ARM [135].

4.4.1 Data Collection

The Ishikawa diagram applied to the oxide problem can be seen in Figure 4.7. Here, the problem is defined as the oxide types at the beginning of the cone phase, as it is hypothesized that the oxide types might reflect on various process conditions. The potential factors were factorized into five categories: ambient factor, machine factor, human factor, material factor, process factor. Since the chamber involves the introduction of oxygen, the ambient factors of the chamber such as moisture level, oxygen level, and pressure were considered. In addition, season factor was also considered to examine whether the process or machine is sensitive to the season weather. Besides, machine properties such as leak rate were also considered. Since the FZ process is fully automatic except the preparation of the machine and post-production work, human factors here only include factors related with the preparation of the machine, such as pumping to remove residual gases and to prepare a pressurized system. Furthermore, factors related to heat dissipation such as feed rod dimensions, crystal dimensions, generator voltage and current were also considered in material and process factors. These identified potentially relevant data would be collected.

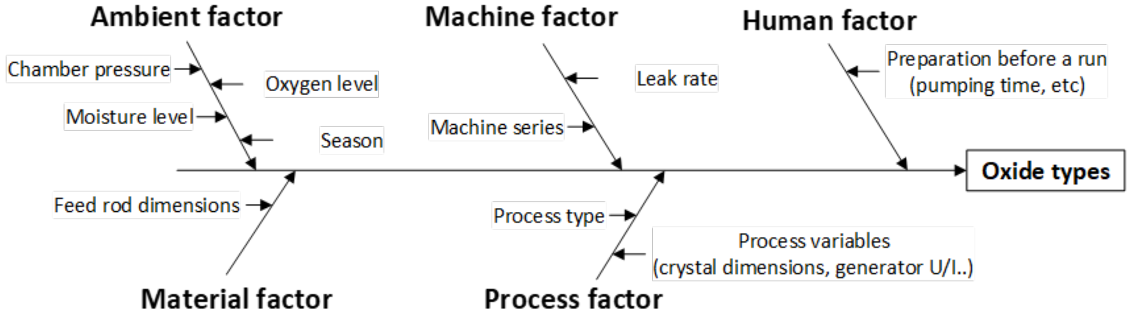


Figure 4.7: Ishikawa diagram for the presence of oxide problem in FZ process.

4.4.2 Data Preprocessing

Following data collection is data processing which is important for obtaining good outcomes from ARM on data. Data preprocessing mainly includes data integration, data cleaning and data transformation.

Data integration

In our case, the identified data are of heterogeneous types involving categorical attributes, numeric attributes as well as time series attributes (such as controlled variables and measured variables in process data). Hence, in order to integrate them into a database for subsequent analysis by ARM, it is necessary to convert them into homogeneous data type. Categorical attributes and numeric attributes are already structured data, thus it is easy to integrate them into an array-like table. The difficulty lies in the time series attributes. There are mainly two methods that can convert time series data into multidimensional data while preserving important characteristics of the data: symbolic aggregate approximation and feature extraction [136]. Symbolic aggregate approximation aims at dividing time series data into several windows, then summary statistics (typically the average or other statistical measures) are computed over each window. Symbolic Aggregate Approximation can offer a better interpretable representation and easy accessibility, while the representations from summary statistics are limited. An alternate strategy, feature extraction, seeks to draw out features from the entire time series data. A common technique in feature extraction is the discrete wavelet transform (DWT), which produces a set of coefficients that represent the characteristics of the time series [136]. Feature extraction retains more information while its implementation is more complex and the outcomes are lack of interpretability.

In this case, symbolic aggregate approximation is employed to reduce the complexity of time series data for easy implementation and better interpretability. Since some FZ process runs may be stopped and restarted due to growing crystal dislocations or other defects [13], this could lead to discrepancies in the time series data. We have chosen to omit data related to production interruptions due to the complexity of the factors and circumstances involved. This decision has resulted in a smaller sample size, but it has allowed us to keep our analysis straightforward. Consequently, any time series data with interruptions before the cone phase are excluded during data integration. Each time series sequence is divided into a set of segments according to the FZ phases, which are encoded into Phase 1-4. This encoding may vary slightly from what is shown in Figure 1.7, but is generally the same, representing the progress of a production run. Then, the average and standard deviation over each segment are computed.

Data Cleaning

The integated data may have missing or inconsistent values due to hardware failure or other errors in data collection. To ensure the high quality and reliability of the ARM results, it is essential to clean the data. In terms of missing values, one can choose to remove the corresponding observation. However, when it comes to a small data set with a limited number of observations, it may be better to estimate the missing value, which is also referred to as *imputation*. For incorrect values, one can apply domain knowledge or statistical methods such as Interquartile Range (IQR) or Z-score to detect outliers and remove them. It should also be noted that some inconsistent data that might be caused by the problem should not be removed [136].

Data Transformation

The ARM for root cause analysis is mainly used for a dataset with binary or categorical attributes [136]. Therefore, the numeric data should be transformed into the form of categorical data type. The procedure of breaking down the numerical attribute into intervals is known as discretization [136]. There are mainly six ways to discretize numeric attributes: Equi-width ranges, Equi-log ranges, Equi-depth ranges, Clustering, Fuzzifying, and Partitioning and Combining [136, 137].

- *Equi-width ranges.* The range of values of the numeric attributes is uniformly divided into equal-width intervals. However, this method may not be effective for a skewed data distribution.

- *Equi-log ranges.* The range of values is divided into intervals in which the difference between the logarithm of the upper limit and the logarithm of the lower limit is the same. This approach is beneficial for data that span across multiple magnitudes.

- *Equi-depth ranges.* This approach partitions the data into intervals with a roughly equivalent amount of data points in each bin. Compared to equi-width ranges, this method can deal with the data distributed non-uniformly better.

- *Clustering.* The partition of the values is learned from unsupervised Clustering methods such as Bayesian Information Criterion (BIC) [138], Density Based Sub-space Miner (DB-SMiner) [139].

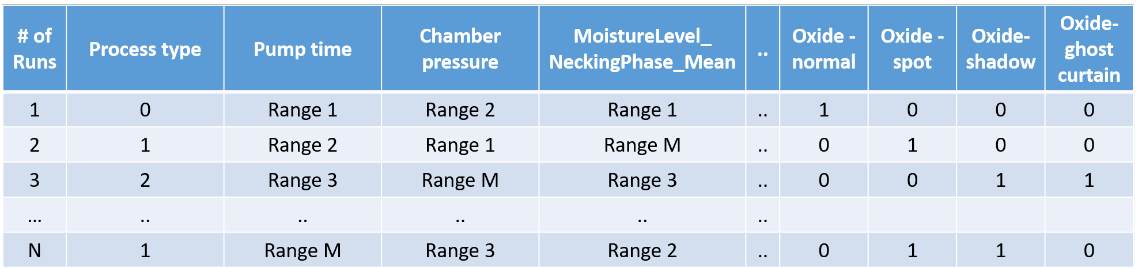
- *Fuzzifying.* The previous partition methods except Clustering share a sharp boundary problem which makes the extracted rules not precise [140, 141]. Therefore, fuzzifying is introduced to divide the data into fuzzy sets, thus reducing the influence of crisp values.

- *Partitioning and Combining.* This approach divides the numerical values into sections and then merges adjacent intervals into larger ones until they have an adequate support value [141].

Considering the data used in this study are not distributed uniformly as different product types have their own growth behavior, equi-depth ranges method is applied for discretizing the numeric attributes into categorical attributes. An example of the preprocessed data after data transformation can be seen in Figure 4.8.

4.4.3 ARM

Suppose there is a set of items $I = i_1, i_2, .., i_m$ representing the attributes and a dataset of observations $T = t_1, t_2, .., t_n$. An association rule can be extracted in the form of $X \Rightarrow Y$ (IF X



# of Runs	Process type	Pump time	Chamber pressure	MoistureLevel_ NeckingPhase_Mean	..	Oxide - normal	Oxide - spot	Oxide-shadow	Oxide-ghost curtain
1	0	Range 1	Range 2	Range 1	..	1	0	0	0
2	1	Range 2	Range 1	Range M	..	0	1	0	0
3	2	Range 3	Range M	Range 3	..	0	0	1	1
…	..	..	..	..	..				
N	1	Range M	Range 3	Range 2	..	0	1	1	0

Figure 4.8: An example of prepossessed data. Time series data are segmented with symbolic aggregate approximation by phases into discrete sequence data, which are further discretized into equi-depth intervals combined with numeric attributes.

THEN Y), where $X \in I$ and $Y \in I$. The association rule is evaluated and filtered by a user-defined threshold of the measures: support, confidence, and lift, which represent the quality of the rule [126].

- **Support**. Support represents the probability of observing X and Y together.

$$Support(X \Rightarrow Y) = P(X, Y) \tag{4.1}$$

- **Confidence**. Confidence represents the accuracy of the rule, is the probability of the occurrence of Y, given that X is observed.

$$Confidence(X \Rightarrow Y) = \frac{Support(X \Rightarrow Y)}{Support(X)} \tag{4.2}$$

- **Lift**. Lift is a balanced measure between support and confidence, and it can represent the enhanced performance in predicting Y given the observation of X. Lift = 1 means that X is independent of Y. Lift >1 indicates X is positively correlated with Y and vice versa.

$$Lift(X \Rightarrow Y) = \frac{Support(X \Rightarrow Y)}{Support(X) \times Support(Y)} \tag{4.3}$$

To extract meaningful association rules efficiently from the data, several algorithms have been developed over the years. Current algorithms work mainly in two steps: first is to find all frequent item sets, then the rules are generated and prunned on these frequent sets with the pre-determined interestingness measures such as support and confidence. The Aprior algorithm [142] is the earlist method that uses the breadth-first strategy to search for all the frequent combinations, while the depth-first search strategy was applied in Eclat [143, 144]. However, these methods face an intensely computationally demanding issue, which is notable in large datasets [145]. To address this problem, the FP-growth algorithm is introduced by transforming the data into a tree-based structure, which avoids the need for costly, repeated database scans [146]. some researchers turn to heuristic methods to optimize the process of generating rules. Examples include GENetic Association Rules (GENAR) [147], Wolf Search Algorithm (WSA) [148]. However, these optimization-based methods may require careful parameter tuning. In addition, the methods are mainly stochastic. The stochastic nature implies that it cannot provide deterministic results, as different runs may produce different results.

In this study, the FP-growth method is employed to extract the rules from the dataset, considering its high computational efficiency and deterministic outcomes. To accelerate the processing time while allowing low support thresholds for rare classes, FP-growth is carried out step by step according to the oxide types, following [149]. The overall procedures of applying FP-Growth for root cause analysis of the oxide problem can be seen in Figure 4.9.

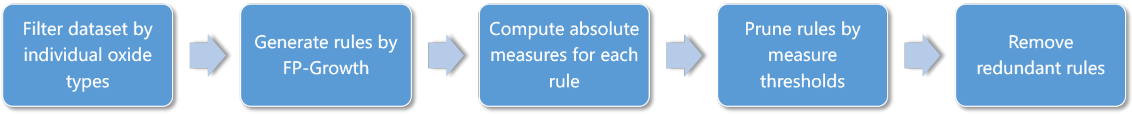

Filter dataset by individual oxide types
Generate rules by FP-Growth
Compute absolute measures for each rule
Prune rules by measure thresholds
Remove redundant rules

Figure 4.9: The overall procedures of applying FP-Growth for root cause analysis of the oxide problem.

Since some oxides appear rarely in the entire dataset, the support threshold needs to be set very low in order to identify as many frequent patterns related to all oxides as possible. This would be a computationally demanding task for FP-Growth. Therefore, to increase computational efficiency, in this research, the entire dataset is divided into multiple sub-datasets that are specific to each oxide. Then the FP-Growth algorithm is applied on each sub-dataset with constraints of the class-wise support threshold and the maximum length of the itemsets generated. The definition of class-wise support can be seen in Eq 4.4.

$$\text{class-wise } support(X \Rightarrow Y|Y) = P(X, Y|Y) \tag{4.4}$$

It is important to be aware that the measures produced by the FP-Growth algorithm are all class-specific, resulting in a fixed confidence and lift of 1. Hence, to prune insignificant rules, it is necessary to compute the absolute measures of these rules over the entire dataset, according to the Eq 4.1,4.2,4.3. Then a threshold-based prunning is carried out by filtering out the rules that do not meet the minimum threshold of the interestingness measures. Finally, redundant rules that have negative or zero improvement as defined by [150], as seen in Eq 4.5, are removed from the pruned rules.

$$\text{a rule} X \Rightarrow Y \text{is redundant if } \exists X' \subset X, measure(X' \Rightarrow Y) > measure(X \Rightarrow Y) \qquad (4.5)$$

The pseudocode for applying the FP-Growth algorithm on the oxide problem can be seen in Algorithm 1.

4.4.4 Rules Visualization

There are various ways to visualize the extracted rules with the focus on visualizing ranging from the interestingnesss measures to the LHS and RHS of the rules. Examples of visualizing association rules include scatter plot, network graph, and matrix [135]. Here we mainly introduce the scatter plot and network graph which would be employed for rules visualization in this study.

- *Scatter plot*. The scatter plot utilizes mainly two interestingness measures, one on each of the axes of the plot. Then each rule is represented as a dot in certain color representing the third interestingness measure. Such that one can easily identify the strong rules with high ability in both measures. However, this also raises the difficulty of interpreting rules. An example of a scatter plot can be seen in Figure 4.10.

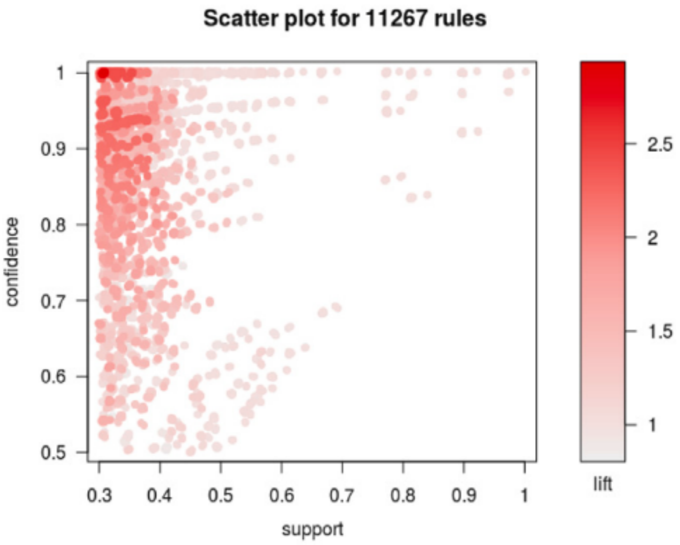


Figure 4.10: An example of a scatter plot for rules visualization [135].

- *Network graph*. Network graph mainly focuses on providing clear relationships among the indivisual itemsets in strong rules.The graph visualizes rules with nodes and edges. Each item is presented as a square node, while each rule is presented as a round node. Then the edges represent the relationship between items and rules. The strength of the rules can be visualized by the color and radius of the rule nodes. This method can easily identify the shared item that contributes to several rules, indicating the importance of the item. However, it can be difficult to explore the network graph when it comes to large sets of rules. An example of a scatter plot can be seen in Figure 4.11.

4.5 Experiments

Through the Ishikawa diagram, 28 potential factors in total were identified that may contribute to oxide. The group of these factors can be seen in Table 4.1. For reasons that include confidential data, we are not listing all potential factors here. Unfortunately, due to the high cost of a moisture sensor capable of tracing moisture in inert purity gases, it is unrealistic to install expensive moisture sensors on all FZ machines. Therefore, currently there is only one moisture hardware sensor which is used by multiple FZ machines in turn, thus limiting the sample size.

Table 4.1: Potential factors that may contribute to the oxide problem.

Category	Factors	Number	Data type
Human	Preparation time, pump times	3	Numerical attribute
Material	Polysilicon weight	1	Numerical attributes
Process	Process type, crystal diameter, generator U/I etc.	14	Categorical attributes and time series attributes.
Machine	Machine series, leak rate	4	Numerical and categorical attributes
Ambient	Chamber pressure, oxygen level, moisture level, Month, etc	6	Numerical, categorical attributes and time series data

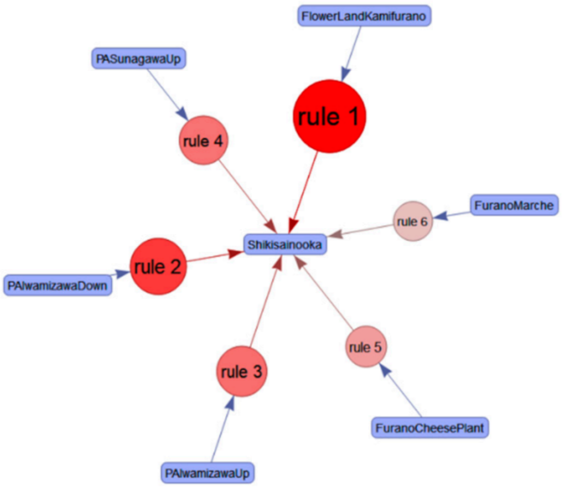


Figure 4.11: An example of network graph for rules visualization [151].

A total of 387 observations of these potential factors along with FZ images were collected from 387 production runs. In the first step of data integration, all time series attributes were preprocessed by the symbolic aggregate approximation method, resulting in multidimensional data consisting of the average and standard deviation over each phase. Considering the significance of the moisture level and the oxygen level to the process, their values at the beginning of the cone phase were also recorded. The integrated dataset was then cleaned from errors and missing data. To enable FP-Growth analysis that can handle only binary or categorical data, the numerical data in the dataset were discretized into three intervals: Low, Medium and High using Equi-depth ranges, such that each interval contains approximately 33.33% of the observations. After data-preprocessing, the dataset consists of 387 samples and 135 features along with three oxide types: normal, spot and shadow (unfortunately, no observation data associated with ghost curtain was found in the gathered data).

The data set was then divided into multiple sub-datasets according to individual oxide types. On each sub-dataset, FP-Growth from rCBA package in R was applied with a minimum class-wise support threshold of 30% and a maximum length of the itemsets of 3 which is equivalent to considering at most two features. The generated rules were pruned with the absolute confidence threshold of 50% and the absolute lift threshold of 1.2, followed by removing redundant rules that have no positive improvement on confidence and lift measures. Finally, a total of 68 rules were identified, with lift values ranging from 1.2 to 11.88. Table 4.2 demonstrates the distribution of the rules after post-processing.

Table 4.2: The distribution of the rules after post-processing.

Oxide type o	$Support(o)$	Number of rules
Normal	0.05	5
Spot	0.66	46
Shadow	0.75	17

4.6 Results and Discussion

The top-20 rules ranked by lift associated with normal, spot and shadow can be seen in Table C.2, Table C.3 and Table C.4, respectively. Scatter plot and network graph offered by the $arulesViz$ package in R were employed for rules visualization. The scatter plot was employed to gain a general understanding of the association rules, while the network graph was utilized to help comprehend the links between the rules. Since the network graph cannot clearly visualize dense rules, here we only visualize top-20 rules ranked by lift for spot type.

Figure 4.12 demonstrates the scatter plot of the rules. As seen, the rules with the highest

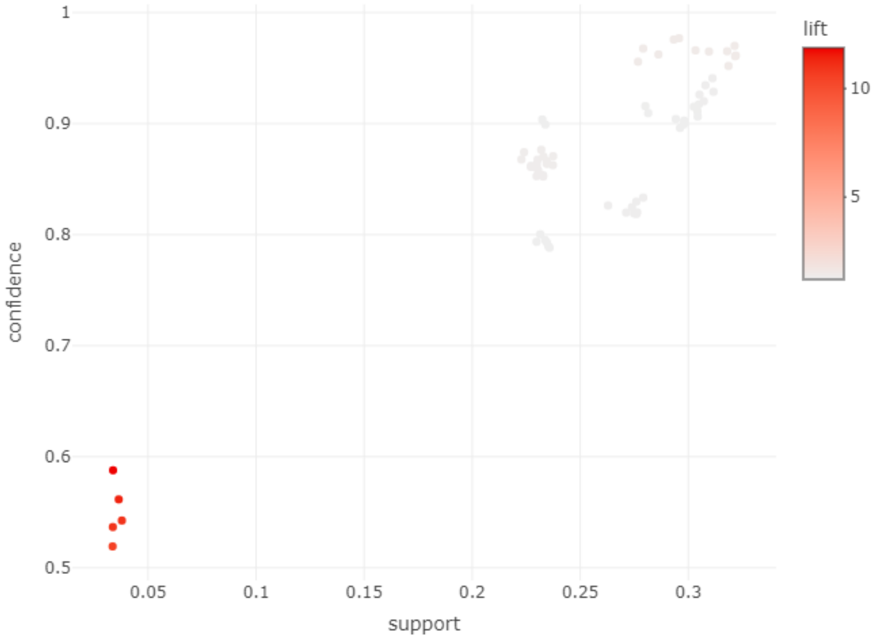


Figure 4.12: The scatter plot of the rules after post-processing. Each dot on the plot represents a specific rule. X and Y axes correspond support and confidence, respectively while the color of the dots represents the lift value.

lift values are concentrated in the bottom left corner of the plot, indicating low support and confidence. In fact, the majority of these rules are linked to the normal type. While these rules may not be very actionable in practice, they offer valuable insights into the optimal conditions for a normal process. Another cluster stands in the upper-right corner of the plot. These rules have both a high level of support and confidence, making them particularly interesting. These rules are associated with spot and shadow oxide types. Focusing on these high-impact rules can drive actionable decisions to address the problem with high efficiency and reliability. Therefore, the rules with high support and confidence are prime candidates for addressing the oxide problem, while the rules in the lower-left corner with high lift values could be an opportunity to assure the optimal conditions for a normal process, thus preventing the recurrence of the oxide problem.

The network graphs of the rules associated with normal, spot and shadow type are demonstrated in Figure 4.13, Figure 4.14 and Figure 4.15, respectively. The itemset in the center of each graph is the shared RHS of the rules. The highlighted itemsets of each graph are the most commonly occurring itemsets that contribute to multiple rules, serving as central hubs. These hubs play a significant role in bridging different item clusters. As demonstrated in Figure 4.13, it is observed that the preparation time at high level is the hub for the rules associated with the normal process, which indicates that allocating resources to keep the preparation time at high level

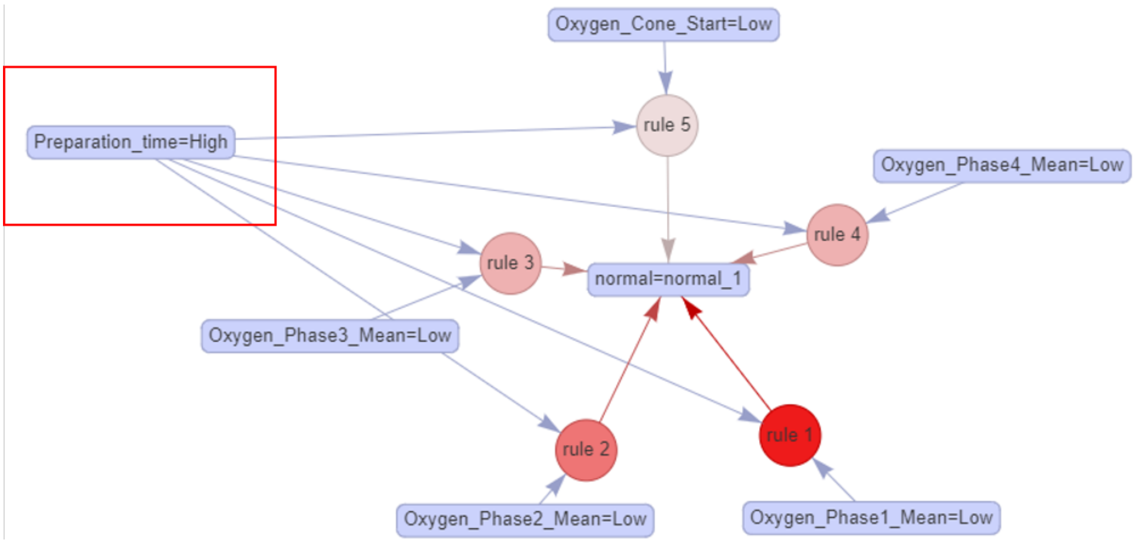


Figure 4.13: The network graph of the rules associated with normal type. The itemset in the center is the shared RHS of the rules. The highlighted itemsets are the most commonly occurring itemsets that contribute to multiple rules.

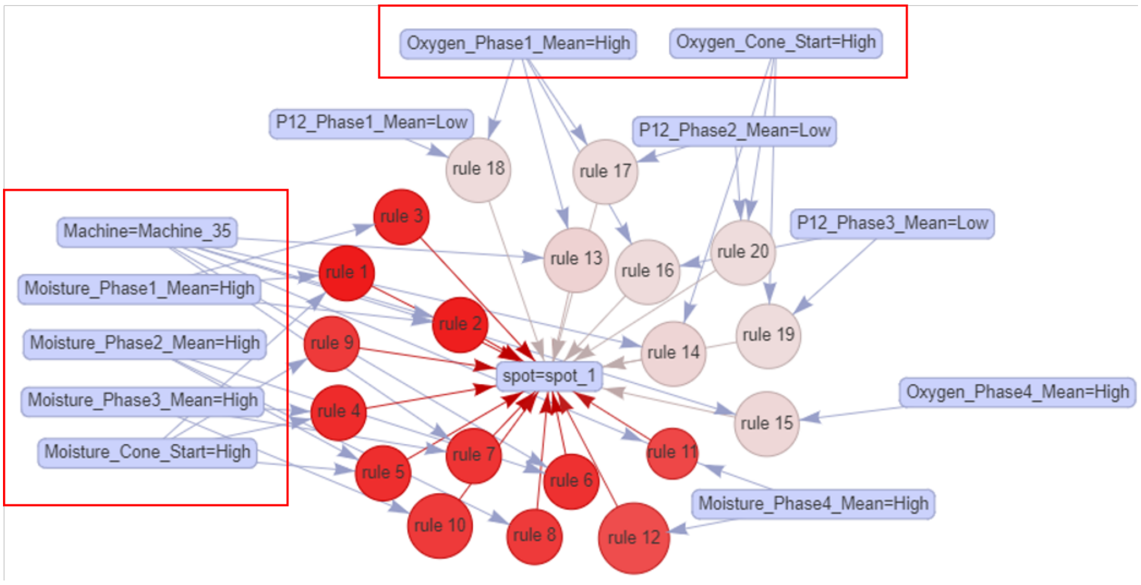


Figure 4.14: The network graph of the rules associated with spot type. The itemset in the center is the shared RHS of the rules. The highlighted itemsets are the most commonly occurring itemsets that contribute to multiple rules.

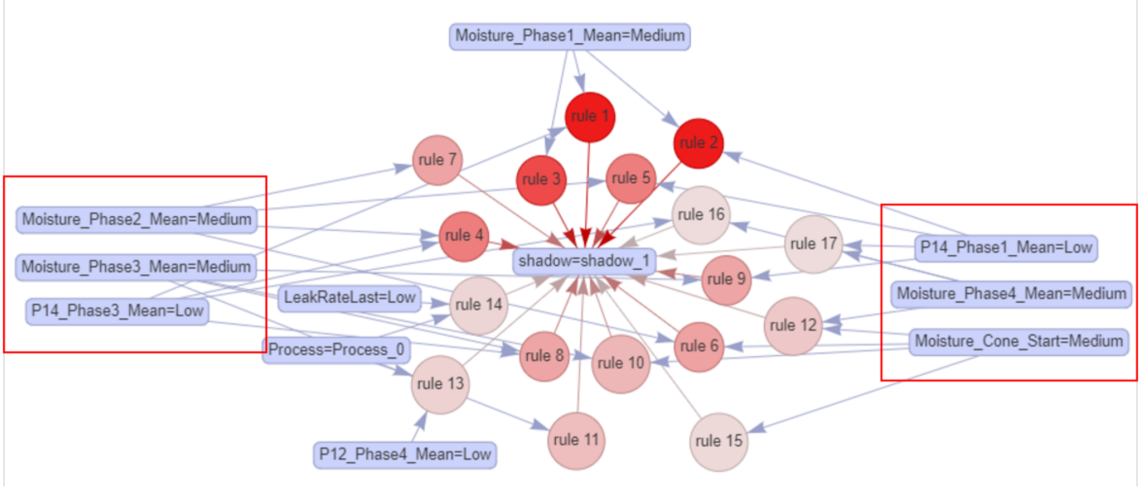


Figure 4.15: The network graph of the rules associated with shadow type. The itemset in the center is the shared RHS of the rules. The highlighted itemsets are the most commonly occurring itemsets that contribute to multiple rules.

may help ensure the occurrence of the normal process. Besides, the itemsets involved are mainly related to oxygen at a low level from Phase 1 to Phase 4, which conforms to the mechanism of oxide layer formation that requires the introduction of oxygen atoms. However, as discussed in the scatter plot, these rules with a weak confidence of less than 0.6 may not be very actionable from the perspective of the company, while they could instead provide insights to prevent the recurrence of the oxide layer. Figures 4.14 and 4.15 demonstrate that the primary hubs for spot and shadow type are mainly situated in the medium to high moisture level from Phase 1. This implies that the oxide might have formed from phase 1 in a relatively humid chamber environment. In comparison to the medium level of water concentration in shadow oxide, the higher moisture level in spot oxide confirms the possibility of a greater thickness of the oxide layer, making it harder to fully transform into SiO at the beginning of the cone phase, thus leaving a spot appearance. Likewise, shadow oxide with less thickness can be easily melted over a large area as it approaches the heater, leaving a shadow appearance. Considering the high support and confidence values of these rules, addressing the problem of a high moisture level in the FZ process is a significant step towards addressing the oxide problem.

4.7 Conclusion

In order to identify the fundamental triggers of the oxide problem and to enable corrective plans to pursue excellence in product quality, machine state, and cost reduction, an integration solution of the knowledge-driven method and the data-driven method was proposed for root cause analysis. Specifically, the Ishikawa diagram was used along with expert knowledge to identify relevant factors, which served as the key data sources for the subsequent data analysis. Then data-driven ARM was employed for conducting root cause analysis of the oxide problem. The gathered data were then preprocessed in order to fit the data to the ARM framework as well as to guarantee the good quality of the outcomes. All relevant time series attributes were converted to multi-dimensional data by symbolic aggregate approximation. Subsequently, the numerical data were cleaned and discretized into intervals using equi-depth ranges method.

Then FP-Growth algorithm in ARM was selected for root cause analysis. To extract meaningful rules efficiently while allowing low support thresholds for rare classes, FP-growth was carried out step by step on multiple sub-dataset divided by the oxide types using a class-wise support threshold of 0.3. The generated rules were further pruned by an absolute confidence threshold of 0.5 and an absolute lift threshold of 1.2, and by removing redundant rules that have no positive improvement on confidence and lift measures. Finally, a total of 68 rules were identified, with lift values ranging from 1 to 11.88. Looking into the rules with high impacts, the results show that the spot type and shadow type are strongly associated with relatively high moisture level inside the chamber from the early phases. This key finding confirms that the oxide layer has formed before it appears apparently in the FZ images. In addition, the higher water concentration occurring in spot than in shadow type explains the greater thickness of spot oxide, making it harder to fully transform into SiO at the beginning of the cone phase, thus leaving a spot appearance. Besides, it is observed that the normal process is associated with high preparation time and low level of oxygen inside the chamber. However, due to weak support and confidence, these would be considered as the opportunities for preventing the recurrence of the oxide problem.

In conclusion, the application of ARM for root cause analysis of the oxide problem has proven to be a valuable and insightful approach for identifying underlying factors contributing to the oxide formation. The insights gained through this methodology include:

- The prime candidate for addressing the oxide problem is to find out the reasons for high moisture level, and to make corrective plans accordingly.
- Ensuring high preparation time and low level of oxygen inside the chamber could be a strategy for preventing the recurrence of the oxide problem.

Therefore, the next Chapter 5 will focus on discovering the reasons for the high moisture level inside the FZ chamber by predicting the moisture level of the FZ process and explainability analysis.

CHAPTER 5
MOISTURE PREDICTION AND EXPLAINABILITY ANALYSIS

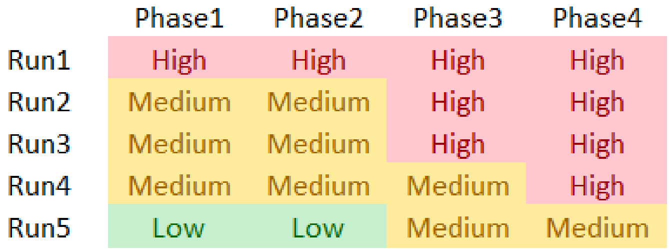
"The most important thing is to never stop questioning." - Albert Einstein

5.1 Motivation

The root cause analysis has revealed that a high moisture level is a major physical factor that could be responsible for the oxide problem. Nevertheless, the reason for the high moisture level inside the FZ chamber remains unclear. In fact, the diagnosis of a high moisture level is as complex as the diagnosis of the oxide problem, as both the FZ machine and the FZ process are rather sophisticated. In order to get to the heart of the matter, it is necessary to trace back from this physical cause to a higher level of root cause, log-action, thus necessitating the solution of **Factors** ⇒ **Root causes**.

Therefore, this chapter is devoted to finding the log-action root cause for a high moisture level. Considering that the solution of **Factors** ⇒ **Root causes** emphasizes more on the accuracy than interpretability, this study turns to establish a regression model for the moisture level using neural networks and try to identify root causes by explaining the model decisions. To do this, a moisture level predictor is established and an explainability analysis is conducted on the developed predictor in order to measure the contribution of each input to model decision, which may shed light on finding the log-action root causes. The moisture level predictor would be developed to estimate the moisture level at the beginning phase of the cone phase, based on the potential factors identified in Chapter 4 together with the FZ images. It may make more sense to estimate the moisture level in the early phase, as suggested by the root cause analysis in Chapter 4. However, if the detection time point is set, for instance, in phase 1, it is still possible to miss the detection of the high moisture level when the moisture level shifts later on, as seen in Figure 5.1. While the moisture level could have undergone changes at any time before the beginning phase of the cone phase, all of these changes will always reflect on this time point when the oxide can be well visually observed. This is also the reason why we include the image as one of the input data, as we believe that this additional input can enhance the accuracy of the model.

Investigating the explanations of the developed predictor can give us useful information about the main sources of influence that can affect the moisture level in the FZ chamber. Additionally, the developed moisture level predictor can be used as a soft sensor for the FZ chamber, thus replacing costly moisture hardware sensors. Currently there is only one moisture hardware sensor that is shared among multiple FZ machines due to its high cost, as it is not feasible to install expensive moisture sensors on all FZ machines. If a moisture predictor is developed with considerable accuracy, it can be quickly and inexpensively implemented on all FZ machines, allowing for the monitoring of the moisture content.



	Phase1	Phase2	Phase3	Phase4
Run1	High	High	High	High
Run2	Medium	Medium	High	High
Run3	Medium	Medium	High	High
Run4	Medium	Medium	Medium	High
Run5	Low	Low	Medium	Medium

Figure 5.1: A portion of the data set reveals the trend of moisture levels across the phases. Different colors signify different moisture levels. As seen, some runs have a shift in moisture levels and the transition point may vary.

In summary, this chapter seeks to develop a moisture level predictor based on the potential factors identified in Chapter 4 along with the FZ images. This will enable us to conduct root cause analysis by understanding the inner workings of the predictor from explainability analysis, as well as to save costs by deploying the predictor as a soft sensor in all FZ machines.

5.2 Problem Definition

The aim of this study is to predict moisture level $y \in \mathbb{R}$ at the beginning of the cone phase given the data of potential factors prior to the cone phase which include table data $X^{table} \in \mathbb{R}^{M_{table}}$, time series data $X^{time} = \{T_1, T_2, .., T_{M_{time}}\}, T_i \in \mathbb{R}^{d_i}$ and image data $X^{img} \in \mathbb{R}^{M_{img} \times m \times n \times 3}$, where $M_{table}, M_{time}, M_{img}$ are the number of table attributes, time series attributes and images, respectively. d denotes the sequence length of a time series attribute, and m, n represent the height and width of a FZ image, respectively.

Since input data involve multiple modalities, with each modality associated with a specific sensor output, the research problem is therefore characterized as multimodal regression $y = f(X^{table}, X^{time}, X^{img})$. Besides, this study also requires an explainability analysis of model behavior, to gain an in-depth understanding of the causes of the high moisture level.

5.3 Literature Review

5.3.1 Multimodal Learning

Generally, multimodal data involve multiple modalities of information, each of which is associated with a particular sensor output, such as images, language, etc [152]. In reality, many real-world issues in various areas involve more than one modality, necessitating the analysis of multimodal data. As a result, there has been a rise in research efforts in different disciplines focusing on multimodal learning, driven by the potential to capture correlations between different modalities. The main core of multimodal learning is to learn multimodal representations that integrate meaningful information from different modalities. However, the presence of heterogeneous multimodal data poses certain challenges [153]. The general pipeline of multimodal representation is "unimodal representation+fusion" where unimodal representation could be obtained independently with several modality-specific learners [154, 155] or could be trained end-to-end (learning multimodal representation while performing downstream task) [156]. Then fusion involves integrating unimodal representations into a compact multimodal representation in a joint manner or in a coordinated manner [152]. Multimodal learning presents a variety of challenges, such as representation, fusion, translation, and alignment, which depend on the applications of multimodal learning, from regression and translation to captioning [153]. Here, we mainly focus on the challenges that are associated with the multimodal regression task: representation and fusion.

The main challenge lies in representation. Bringing together data from various sources can be a difficult task as they often have different structures and representations. Therefore, a key to multimodal learning is to learn how to represent multimodal data in such a way that they have compatible and meaningful representation. The term "representation" is synonymous with "feature" and usually refers to a vector or tensor that represents an entity, such as an image or a sentence [153]. Having good representations can improve the efficiency of data processing, making it easier to analyze the data. On the other hand, good representations can improve the outcomes from machine learning models which heavily rely on the quality of representations, as demonstrated by recent advances in visual object detection [157] and speech recognition [158]. In recent decades, there has been a shift in unimodal representation methods from handcrafted-based representation to data-driven representation [153]. This is due to the undeniable success in various domains achieved by deep learning, which is a type of artificial neural network with multiple hidden layers. Examples of the shift in representation can be seen in the following.

- *Visual representation*. Conventionally, visual representations are designed manually based on prior knowledge in the aspects of gradient [159], texture [160], and so on. The most representative handcrafted visual representation is the Scale Invariant Feature Transform

(SIFT) [161]. Despite the early success of these handcrafted representations, they are not as effective when it comes to more complicated tasks. With the advent of neural networks, neural architectures such as CNNs [84] gradually substitute handcrafted methods for visual representations due to their powerful capability in handling array-like data.

- *Time series representation*. Time series representations can be derived from two perspectives: temporal and spectral [162]. Temporal representations are based on the temporal domain and are typically expressed as a sequence of values. A typical conventional method for obtaining temporal representations is the symbolic aggregate approximation method (as introduced in Chapter 4), where time series data are converted to symbolic sequences. Spectral representations, on the other hand, are derived by transforming time series data into the frequency domain and are usually represented by a collection of frequency components. Example methods of obtaining spectral representations include Discrete Fourier Transform (DFT) and Discrete Wavelet Transform (DWT). However, these representations have been gradually superseded by deep neural network representations [153], such as RNNs with the ability to capture long-range dependencies [163], or CNNs with the ability to extract spatial correlations from univariate time series [164] or multivariate time series [165]. Despite their stong capability to extract patterns in time series, these networks applied to time series have some drawbacks. RNNs are prone to the vanishing gradient problem [166] when processing long sequences, which can make it challenging for them to capture long-range dependencies in the data. CNNs can be adversely affected by noise or outliers in time series, which may negatively impact the learned features [167]. Therefore, some researchers turn to using feature extraction prior to CNNs, in order to enhance the performance of CNNs [167, 168]. A typical example is to encode the time series data into image-like data using some feature extraction methods such as Gramian Angular Field (GAF) [168] or Markov Transition Field (MTF) [169], potentially making CNNs more effective in modeling complex temporal patterns.

The second challenge, fusion, is to jointly leverage the complementary information from multiple modalities to obtain a more comprehensive and accurate multimodal representation of the underlying data [170]. While unimodal representations have been extensively studied, most fusions involve only a simple concatenation or weighted sums of representations in the data level, intermediate level, or decision level [171, 172]. Since unimodal representations often contain data specific to the modality, fusing them in such a straightforward way will reduce discriminative information and restrict the expression of the multimodal representation [173]. Recently, such a fusion paradigm has been rapidly changing due to the prevalence of the attention mechanism [170] which can emphasize discriminative information while suppressing irrelevant information

[173]. Attention mechanisms are typically characterized by the weighted sum of a collection of vectors, with dynamically generated attention weights, thus allowing the model to selectively focus on the most important parts of the input representation [174]. Figure 5.2 represents the unified attention architecture. Squeeze-and-excitation (SE) module [175], a typical channer-wise attention, has been shown to have potential prospects in multimodal fusion [173] by explicitly modeling channel-wise dependencies. Transformer is notable for its self-attention mechanism to model long-range dependencies in the data [176]. Benefiting from its self-attention mechanism, Transformer can combine the learned knowledge from each tokenized input from any modalities in the topological geometry space, which allows Transformer to have a more flexible modeling space [152]. Therefore, the great success of Transformer as well as its flexible modeling motivate researchers to extend Transformer to other modalities and finally to the multimodal context. VideoBERT [177] is the first work to extend Transformer to multimodal tasks with visual and language modalities, which implies a great potential of Transformer in the multimodal context [152].

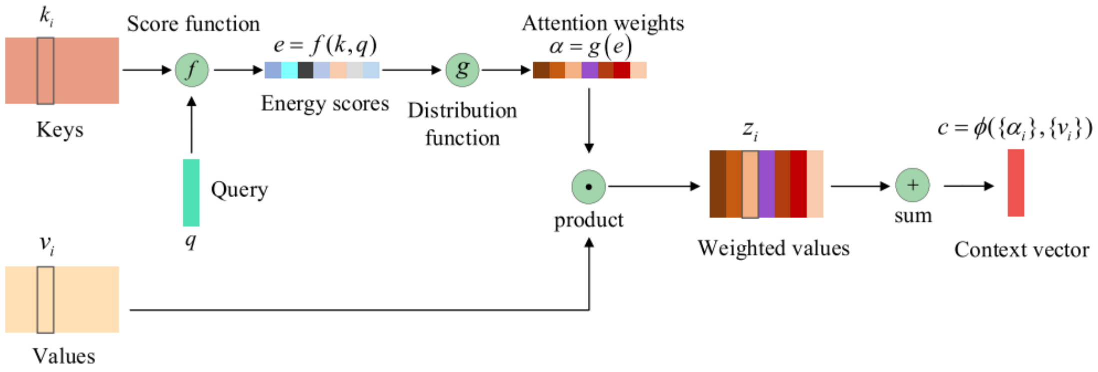


Figure 5.2: The architecture of the unified attention model [174].

5.3.2 Explainable Artificial Intelligence

In order to allow knowledge transferring thus to gain a in-depth understanding of high moisture level, it is necessary to increase the transparency of the built moisture level predictor. To this end, Explainable Artificial Intelligence (XAI) is introduced in this section with the attempt to interpret the machine learning models.

Machine learning has seen a tremendous increase in both research and industrial applications. Many activity sectors have adopted new information technologies, with machine learning at the center, due to its power capability in automatically handling data. The very first machine learning methods, such as decision trees or bayesian models, were easily interpretable [178], allowing humans to examine how input variables affect output. As the amount of data available grows and its complexity increases, we are increasingly relying on more complex models to ensure their effectiveness, making the model a complex black-box. As black-box ML models are increasingly being employed to make significant predictions in critical contexts, there is a growing demand to understand how these models made decisions from various stakeholders [179], in order to build trust in their predictions. This ability to explain their predictions is a crucial evaluation at three stages of artificial intelligence development: Early Development Stage, AI on Par with Human Abilities and AI Surpassing Human Capabilities [104], as shown in Figure 5.3.

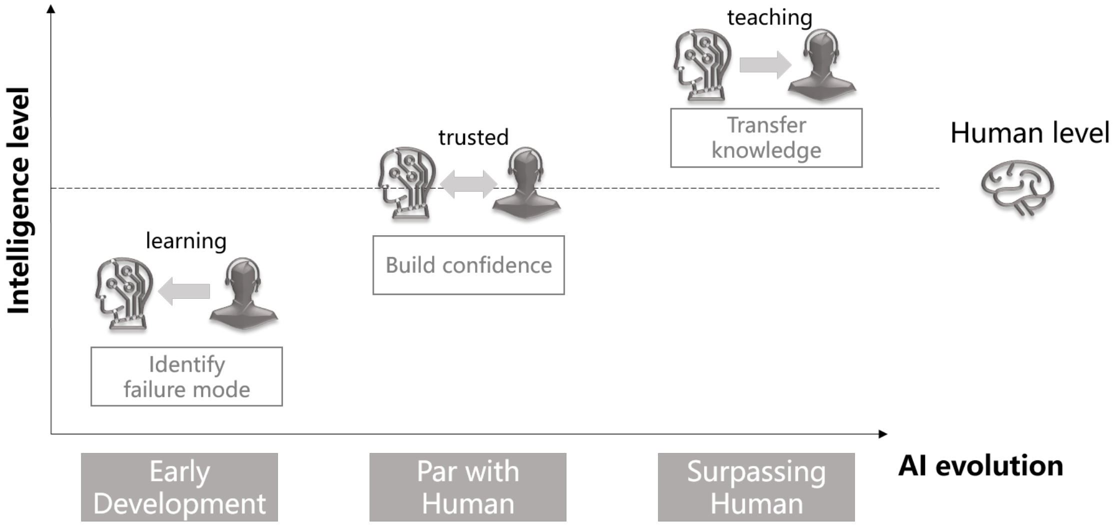


Figure 5.3: Three AI evolution stages: Early Development Stage, AI on Par with Human Abilities, AI Surpassing Human Capabilities based on [104].

Early Development Stage. At the beginning of AI development when AI is not as powerful as humans and is not yet ready for deployment, transparency and explanations are of great value since they help to identify and understand the reasons for the failures of the AI system [104]. By providing insight into the causes and logic behind AI-driven decisions, researchers can recognize the system's weaknesses and direct their efforts towards the most promising areas for improvement. In short, explanations act as diagnostic tools, helping researchers advance AI capabilities by addressing and fixing any issues.

AI on Par with Human Abilities. As AI advances to a point where its capabilities are comparable to those of human intelligence, the main goal shifts to creating trust and confidence among users [104]. To reach this goal, it is essential to provide clear and understandable explanations for AI-generated predictions. These explanations help users understand the rationale behind AI-driven decisions, reducing the perceived obscurity or "black box" nature of the technology. When users can recognize why AI suggests certain courses of action or predictions, they are more likely to trust the technology, using it for a variety of tasks. Therefore, explanations are a fundamental element in inspiring user confidence in AI as a reliable tool.

AI Surpassing Human Capabilities. In the situation where AI surpasses human cognitive abilities, the purpose of explanations changes once more. AI takes on the role of an educator, providing humans with instructions on how to make better decisions [104]. The explanations given by AI are not only about its own actions, but also about the best ways to make decisions. This transfer of knowledge allows people to use AI's remarkable capabilities to their advantage. In essence, AI becomes a mentor, giving people valuable information and wisdom to help them make informed and better decisions.

Hence, under no circumstances should black-box ML models provide reasons to make its functioning understandable to humans. Explainable AI (XAI) is thus proposed to address pressing issues of ML models in terms of interpretability and explainability. Explainable AI (XAI) refers to a suite of ML techniques designed to be more transparent and understandable to humans [180]. It aims to bridge the gap between the inherent complexity of AI models and human understanding, thereby making AI systems more transparent, accountable and trustworthy. In particular for our oxide problem where we have no clue of its emergence, it is significant to gain some insights from the decision making mechanism of the ML models that perform much better than humans.

XAI encompasses mainly two categories: transparent model architecture and post-hoc explainability [178], as shown in Figure 5.4. The difference between transparent model architecture and post-hoc explainability lies in the nature of interpretability. Transparent models are designed with the intention of being understandable to humans. Therefore, high level of transparency allows us to rapidly understand their decision-making process. Post-hoc explainability, on the other hand, is used to explain the decisions made by non-interpretable models. Post-hoc explainability covers the techniques used to convert a non-interpretable model into an explainable one [178] to enhance their interpretability, such as visual explanations, local explanations, global explanation, etc [178]. Compared to transparent model architecture, post-hoc explainability techniques are employed after complex models have been trained to give explanations for their predictions, increasing transparency without changing the complex model itself.

Transparent model architecture. As mentioned above, the very first ML models were used with transparent model architecture. Typical examples include the linear regression model and decision trees. For a linear regression model, such as logistic regression, one can easily identify the influence of variables on the output by examining the corresponding weights. In terms of decision trees, the built-in if-else decision making can be easily understood [43]. However,these methods often sacrifice model performance to maintain transparency. The relatively poor model accuracy limits the application of these interpretable models in recent complex tasks. Nevertheless, their transparent model architectures can compensate for the issue of interpretability shown in black-box ML models. They can act as surrogate models to approximate the behavior of complex models.

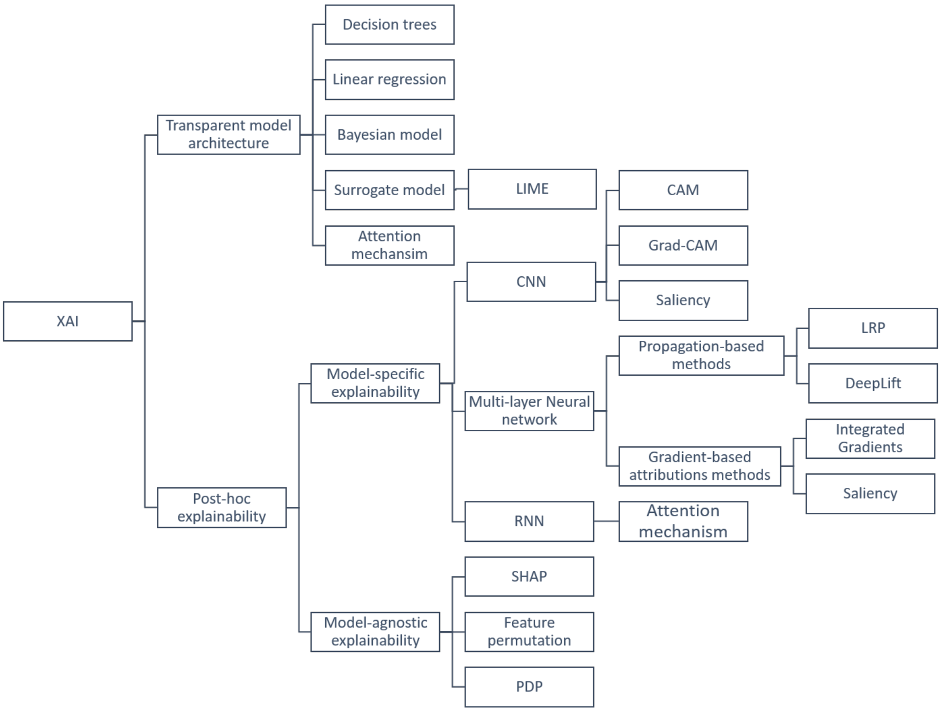


Figure 5.4: Taxonomy of XAI techniques.

They are trained on the predictions of black-box models and only used for explaining their model decisions for black-box models. Local Interpretable Model-agnostic Explanations (LIME) is a typical example that builds local surrogate models for individual predictions of a complex model [181]. However, due to the lack of theoretical evidence, it remains uncertain whether the straightforward surrogate model accurately reflects the more complex model [182]. Furthermore, since the attention mechanism is designed to be transparent and embedded in the structure of networks, the attention mechanism is often considered in this category [183]. However, it should be noted that there is still an argument about whether attention weights in the attention mechanism can provide meaningful explanations for predictions [184, 185].

Post-hoc explainability. Post-hoc explainability methods are not models themselves. Instead, they are explanation techniques applied after a black-box model has been trained. These methods are used to shed light on the internal workings of such complex models. Before moving on to the taxonomy of post-hoc explainability, it is necessary to declare the difference between *local explanation* and *global explanation*.

- *local explanation*: Local explanation seeks to explain a single prediction of a model for a specific instance or data point [186, 187]. Local explanation focuses on comprehending why a model produced a certain prediction for a single input or a small group of inputs. This is especially beneficial when attempting to recognize failure modes or build confidence in individual predictions.

- *global explanation*: Global explanation is concerned with providing an overall understanding of how an ML model works on an entire data set or population [186, 187]. It aims to uncover the general model behavior. This can help us gain an understanding of which features are consistently influential for the model's predictions.

Post-hoc explainability can be further categorized into model-specific explainability and model-agnostic explainability. Model-specific explainability techniques are tailored to a single type of machine learning model, while model-agnostic explainability methods are designed to be compatible with a wide range of machine learning models, regardless of their type.

1) Model-agnostic explainability

Typical examples of model-agnostic explainability are Feature Permutation [188] and SHapley Additive exPlanations (SHAP) [189] which are mainly based on perturbation, and Partial Dependence Plot (PDP) [190]. Feature permutation is a global explanation method that can assess the importance of input features by randomly shuffling their values [188]. This approach is based on the assumption that a feature is considered "significant" if permuting its values leads to an increase in model error, as this implies that the model was dependent on the feature for its prediction and vice versa [187]. PDP is a graphical representation, which is also for global explanation. PDP can visualize relationship between a feature and the predictions of a black-box model [182], thus offering insights into how changes in a feature affect the model output. SHAP is a method for both global explanation and local explanation by computing an additive feature importance for each prediction. These methods allow for high flexibility on model types. However, they are normally computationally intensive as the size and dimensionality of input samples increase.

2) Model-specific explainability

Model-specific explainability is tailored to explain mostly neural networks with local explanation. It often involves leveraging knowledge about the model's design. The neural works applied with model-specific explainability can be categorized into three types: CNN, Multi-layer neural network and RNN [178].

CNNs. In fact, model-specific explainability for CNNs has already been introduced in Chapter 3, specifically in Section 3.3.3, where techniques for visualizing CNN decisions were discussed for building trust on CNN-based intelligent systems. For more details, please refer to the preceding chapter.

Multi-layer Neural Network. Multi-layer neural network is a common and simple type of neural network, which contains multiple layers of interconnected neurons. Model-specific explainability for multi-layer neural network includes propagation-based explainability, and gradient-based explainability. Propagation-based explainability techniques consider prediction as the output of a computational graph (the neural network), and generate explanation by progressively redistributing the output in the lower layers [191]. This technique involves analyzing the forward and backward passes of the network to trace the influence of input features on the model's predictions. Typical examples include layer-wise relevance propagation (LRP) [192] and DeepLift [193]. These methods often align well with the structural components of neural networks, which makes it easier to explain how different layers and neurons contribute to the final prediction. Gradient-based explainability techniques leverage the gradients of the model prediction with respect to its inputs to understand the influence of input features on the predictions. These techniques focus on identifying which input features or data points had the most significant impact on the model's decisions. Examples of gradient-based explainability for multi-layer neural network include Integrated Gradient [194], saliency maps [107], etc. However, gradient-based explainability for multi-layer neural network might not provide as intuitive visual explanations as gradient-based explainability for CNN, which can be a difficult to check whether the explanation is correct.

RNNs. RNNs are well known for their ability to capture long-term dependencies, thus widely used for natural language processing and time series analysis [178]. Fortunately, model-specific explainability techniques for multi-layer neural network like LRP and Integrated Gradients can also be adapted for RNNs with some adjustments. Then here we only introduce explainability techniques only specialized for RNNs. An explainability technique commonly used for RNNs is the attention mechanism. This method assigns different attention weights to each element of the input, allowing the model to concentrate on the most important parts of the input sequence [183].

5.4 Methodology

In the case of our task, the input data involve three modalities: images, tabular data, and time series data, which poses many difficulties: how to address the problem of heterogeneous data; how to capture meaningful multimodal representation; how to provide an interpretation of these modalities. Targeted at such problems, Multimodal GAF-SE-Transformer is proposed. In this section, we first revisit GAF, SE and Transformer as a warm up, then introduce the framework of the proposed model in details including the elaboration of the two major steps: unimodal representation and multimodal fusion. Next, three explainability techniques, including attention mechanism, SHAP, feature permutation, were compared to evaluate the importance of features. By examining feature importance rankings from these explainability techniques, we can identify influential features, which might possibly be root causes. On the other hand, it enables us to prune the input features, thus decreasing computational cost.

5.4.1 Revisiting

Revisiting GAF

Exploring temporal data for hidden patterns, relationships, and structures has been a long-standing challenge in time series analysis and feature engineering. In the context of Deep Learning, it is nature to choose RNNs, including variants like LSTM, to extract representation of time series data by capturing temporal dependency. However, these methods are often difficult to train since they are sensitive to noise, which affects the reliability of the network [195]. In this study, to take advantage of the powerful feature extraction capability of CNN, we turn to 'visually' analyze time series data, encoding time series into images before passing time series to the model. One powerful approach that can encode time series data in images is GAF [196]. GAF is a novel representation technique that transforms time series data into images that can be both visually interpreted and analyzed, thus offering an alternative perspective on time series analysis. The concept of GAF is to represent time series data in a polar coordinate system instead of the traditional Catesian coordinate system by encoding the value as the angular cosine and the time stamp as the radius. Given a univariate time series $X = \{x_1, x_2, .., x_n\}$ with N observations. X is first scaled in the interval [-1, 1] or [0, 1] by MinMaxScaler. The rescaled time series $\tilde{X}$ is then encoded in a polar coordinate system, as seen below.

$$\begin{cases} \phi = \arccos{(\tilde{x_i})}, \tilde{x_i} \in \tilde{X} \\ r = \dfrac{t_i}{N}, t_i \in N \end{cases} \tag{5.1}$$

Where ϕ and r represent the angle and distance in the polar coordinate system, respectively, and t_i is the timestamp. The temporal dependency can be identified by computing the trigonometric sum/difference between each point, which are Gramian Summation Angular Field (GASF) and Gramian Difference Angular Field (GADF), as defined in Eq 5.4 and Eq 5.3.

$$GASF(i, j) = \cos(\phi_i + \phi_j) \tag{5.2}$$

$$GADF(i, j) = \sin(\phi_i - \phi_j) \tag{5.3}$$

Where $GASF(i, j)$ and $GADF(i, j)$ denote the element $(i, j) - th$ in the GASF matrix and the GADF matrix, respectively. GASF emphasizes cumulative patterns and periodicity, whereas GADF concentrates on variations, irregularities, and non-periodic patterns. Given that the length of the original time series is n, the Gramian matrix is of size $n \times n$, which makes the GAFs large. Therefore, Piecewise Aggregation Approximation (PAA) is normally used to reduce the size of data before transformation with GAFs as well as to smooth the time series [196].

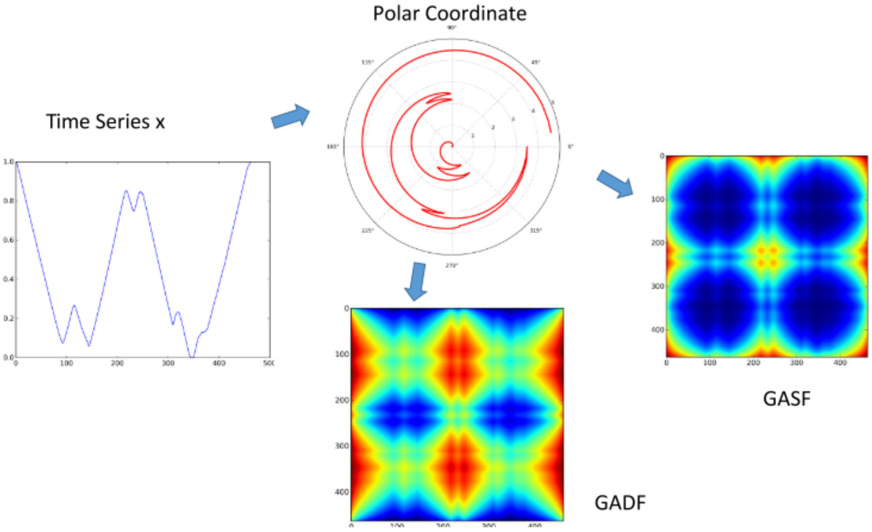


Figure 5.5: Illustration of the encoding map of GASF and GADF [196].

By transforming time series into image-based representations with GAF, the strengths of CNNs and other image processing techniques can be leveraged to extract meaningful patterns and relationships. Hence, in this work, we employ GAF to extract the representations from time series data.

Revisiting SE

SE, an attention mechanism, was firstly proposed by [175] to improve the quality of CNN representations by explicitly capturing the channel-wise dependencies between convolutional features. The channel-wise dependencies are quantified by the nonlinear attention assigned to each channel, representing the channel-wise importance weights. The nonlinear attention is then used to rescale the input representations, emphasizing important channels and suppressing less important ones [175].

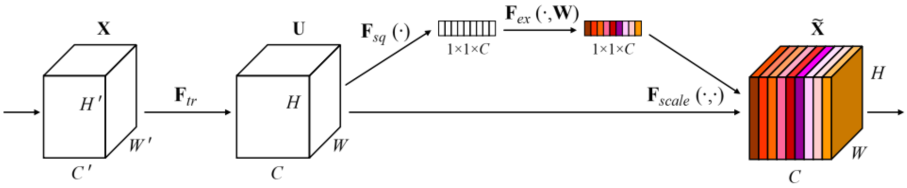


Figure 5.6: Illustration of the architecture of the SE module [175].

The structure of SE module is represented in Figure 5.6. SE module consists of two operations: squeeze and excitation. Suppose that $\mathbf{X} \in \mathbb{R}^{C \times H \times W}$ denotes the input representations having

C feature maps with each spatial dimension of (H, W). The squeeze operation is performed by shrinking the input representations $\mathbf{X}$ through its feature map dimension with average pooling, such that the $c-th$ output representation Y_c from the squeeze step is computed by:

$$Y_c = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} \mathbf{X}_c(i, j) \tag{5.4}$$

In excitation step, $\mathbf{Y} \in \mathbb{R}^C$ fed to a bottleneck with two fully-connected (FC) layers with the non-linearity. Then the channel-wise attention weights $\mathbf{S}$ are given by Eq 5.5:

$$\mathbf{S} = \sigma(F_2(ReLU(F_1(\mathbf{Y})))) \in \mathbb{R}^C \tag{5.5}$$

Where F_1 is a dimensionality-reduction FC layer with reduction ratio 2. F_2 is a dimensionality-increasing FC layer, which can return the intermediate output to the channel dimension. σ is an activation function.

Finally, the enhanced output representation can be obtained by reweighting the input representations $\mathbf{X}$ with the channel-wise attention $\mathbf{S}$:

$$\hat{\mathbf{X}} = \mathbf{S} \odot \mathbf{X} \tag{5.6}$$

Where $\odot$ denotes channel-wise multiplication.

Compared with the input representations, the final output of the SE block has adaptively recalibrated representations. Hence, in this work, we consider utitilizing SE module to well fuse the unimodal representation. In addition, its ability to amplify informative features can shed light on the interpretability analysis on revealing feature importance. Therefore, the explainability analysis of SE will also be discussed.

Revisiting Transformer

Transformer was first proposed in 2017 to address the limitations of RNNs in sequatial modeling tasks [197]. Benefiting from a self-attention mechanism, Transformer achieves state-of-the-art on a variety of NLP tasks and is currently the leading force in NLP domains [152].

The model architecture can be seen in Figure 5.7. The Transformer consists of two main components: the encoder and the decoder. The encoder processes the tokenized input sequence, while the decoder generates the output sequence [197]. Both components are composed of multiple layers, with each containing a multi-head self-attention mechanism and feedforward neural networks [197]. The self-attention mechanism, also named "Scaled Dot-Product Attention" [197], is the heart of the Transformer, allowing for the modeling of complex relationships

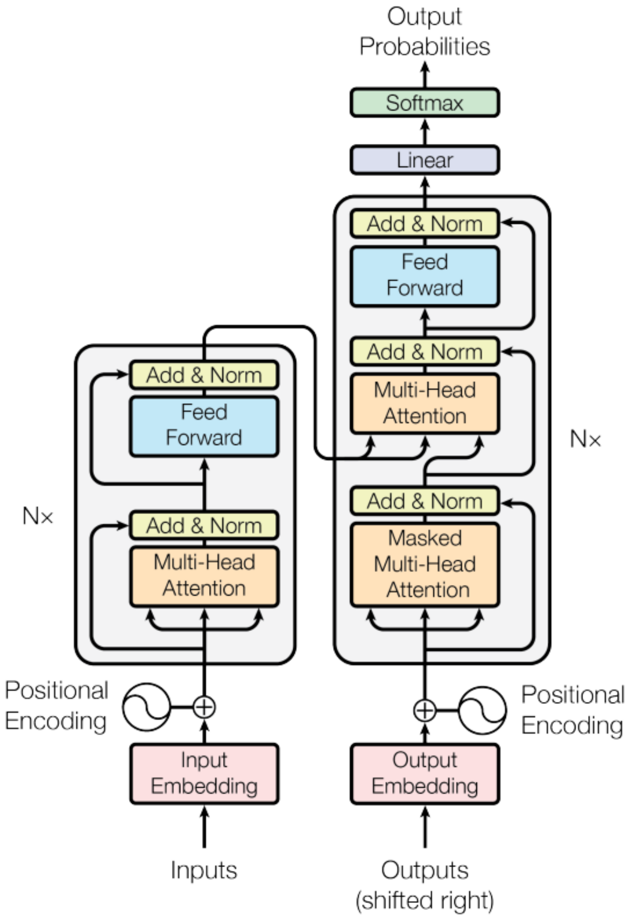


Figure 5.7: The architecture of the Transformer [197].

and dependencies within sequences. Assume that $\mathbf{X} = (x_1, x_2, .., x_N) \in \mathbb{R}^{N\times d}$ is an input sequence with N tokens, each with a d-dimensional vector. In order to take advantage of the order of the sequence, the input sequence is usually processed with position encoding, which can be done through either summation or concatenation [152], as seen in Eq 5.7.

$$\hat{\mathbf{X}} = PE(\mathbf{X}, PositionEmbedding) \in \mathbb{R}^{N\times d} \tag{5.7}$$

Where PE denotes the positional encoding.

After processing, $\hat{\mathbf{X}}$ will go through three projection matrices in the self-attention mechanism, resulting in three embeddings $\mathbf{Q} \in \mathbb{R}^{N\times d_q}$, $\mathbf{K} \in \mathbb{R}^{N\times d_k}$, $d_q = d_k$ and $\mathbf{V} \in \mathbb{R}^{N\times d_v}$, namely Query, Key and Value, respectively.

$$\mathbf{Q} = \hat{\mathbf{X}}\mathbf{W}^Q \tag{5.8}$$

$$\mathbf{K} = \hat{\mathbf{X}}\mathbf{W}^K \tag{5.9}$$

$$\mathbf{V} = \hat{\mathbf{X}}\mathbf{W}^V \tag{5.10}$$

Then self attention $\mathbf{Z}$ can be computed by Eq 5.12:

$$\mathbf{Z} = softmax(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}}\mathbf{V}) \tag{5.11}$$

Self-attention enables each element to take into account all the other elements, thus encoding each input as a fully-connected graph in topological geometry space [152]. In practice, self-attention is performed with multiple self-attention sub-layers arranged in a parallel way and their outputs would be concatenated and projected using a linear projection matrix W^O.

$$\mathbf{Z} = concat(\mathbf{Z_1}, .., \mathbf{Z_h})\mathbf{W}^O \tag{5.12}$$

Where $\mathbf{Z}_i = softmax(\frac{\mathbf{Q}_i\mathbf{K}_i^T}{\sqrt{d_{k_i}}}\mathbf{V}_i)$. With each head focusing on different aspects of the input sequence independently, it can allow the Transformer to generate richer representations of the input sequence. In this work, considering its powerful capability in modeling complex relationships and dependencies within the input, Transformer will be used for fusing unimodal representations from different modalities.

5.4.2 Model overview

For this task, we attempt to extract the representations from these three modalities: images, table data and time series data and then aggregate the discriminative information from the unimodal representation with a SE module and Transformer. Specifically, for the extraction of time series representation, GAF is employed for encoding time series attributes to images, which would be subsequently processed by a CNN. Finally, the proposed model would be analyzed in terms of interpretability based on SHAP. The overall framework of the proposed Multimodal GAF SE-transformer can be seen in Figure 5.8, which consists mainly of two parts: representation module, multimodal fusion module.

Assume that $\mathbf{X} = \{\mathbf{X}_i^{img}, \mathbf{X}_i^{tab}, \mathbf{X}_i^{time}, |i = 1, 2, ..., N\}$ is the input data in the dataset for the moisture level predictor with a size of N. Then the input data are fed into representation module to extract unimodal representations $\mathbf{Z}_i^{img} \in \mathbb{R}^{M_{img} \times d}$, $\mathbf{Z}_i^{tab} \in \mathbb{R}^{M_{tab} \times d}$, $\mathbf{Z}_i^{time} \in \mathbb{R}^{M_{time} \times d}$.

$$\mathbf{Z}_i^{img} = F_{img}(\mathbf{X}_i^{img}) \tag{5.13}$$

$$\mathbf{Z}_i^{tab} = F_{table}(\mathbf{X}_i^{tab}) \tag{5.14}$$

$$\mathbf{Z}_i^{time} = F_{time}(\mathbf{X}_i^{time}) \tag{5.15}$$

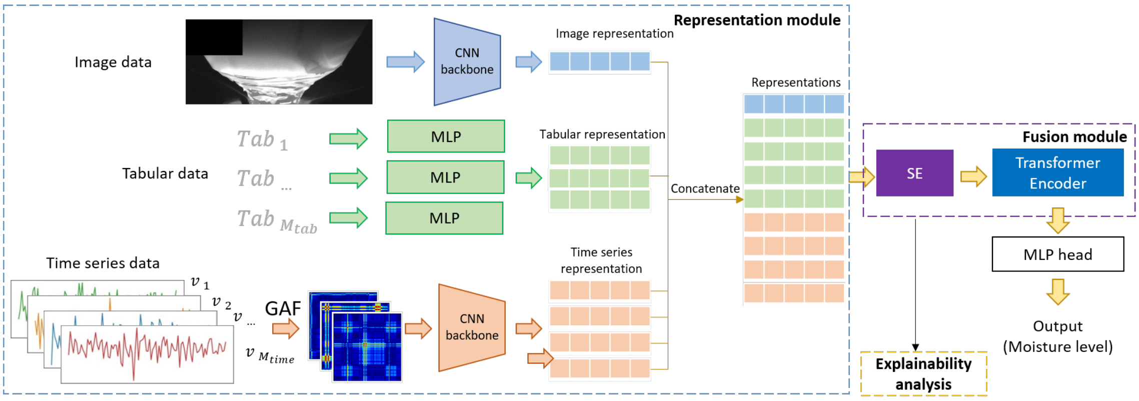


Figure 5.8: The overall architecture of the proposed Multimodal GAF SE-transformer.

Then the unimodal representations were concatenated into $\mathbf{Z} \in \mathbb{R}^{(M_{img}+M_{tab}+M_{time})\times d}$, as seen below:

$$\mathbf{Z}_i = F_{concat}(\mathbf{Z}_i^{img}, \mathbf{Z}_i^{tab}, \mathbf{Z}_i^{time}) \tag{5.16}$$

Then the concatenated representations would be processed with a SE module to fuse the representations while aggregating the discriminative information.

$$\mathbf{Z}_{SE_i} = F_{SE}(\mathbf{Z}_i) \in \mathbb{R}^{(M_{img}+M_{tab}+M_{time})\times d} \tag{5.17}$$

Then the fused representations would be further fused with a Transformer encoder to learn cross-modal interactions.

$$\mathbf{Z}_{TransE_i} = F_{TransE}(\mathbf{Z}_i) \in \mathbb{R}^{(M_{img}+M_{tab}+M_{time})\times d} \tag{5.18}$$

Finally, the output representations from the Transformer encoder would be fed into a MLP to generate a prediction of the moisture level. The the parameter set of the proposed model would be optimized using a loss function $\mathcal{L}$:

$$\hat{y}_i = F_{MLP}(\mathbf{Z}_{TransE_i}) \tag{5.19}$$

$$\mathbf{\Theta} = argmin\ \mathcal{L}(\hat{\mathbf{y}}, \mathbf{y}) \tag{5.20}$$

Where $\mathbf{\Theta}$ are the parameters of the proposed model that need to be optimized.

There are various losses for tackling regression problems, such as Mean Square Error (MSE) loss, Mean Absolute Error (MAE) and Huber loss [198]. MSE loss is the most common loss for regression. MSE is differentiable, making it suitable for a fast convergence of the model. However, due to the property of squaring, MSE penalizes the outliers heavily, which, on the other hand, makes it sensitive to the outliers. Compared to MSE, MAE is more robust to outliers, since it measures the absolute value of the error. However, due to non-differentiable nature, it will affect the efficiency of solving the model [198]. Huber loss [199] is a piecewise function of MSE loss and MAE loss, thus combining the benefits of both losses [198]. In this work, MSE loss and Huber loss are considered to optimize the parameters of the proposed model. Their definitions can be seen below.

$$\mathcal{L}_{MSE}(\hat{\mathbf{y}}, \mathbf{y}) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{5.21}$$

$$\mathcal{L}_{Huber}(\hat{\mathbf{y}}, \mathbf{y}) = \frac{1}{n}\sum_{i=1}^{n}\mathcal{L}_{Huber}(\hat{y}_i, y_i) \tag{5.22}$$

$$\text{Where} \quad \mathcal{L}_{Huber}(\hat{y}_i, y_i) = \begin{cases} \frac{1}{2}(y - \hat{y})^2 & \text{if } |y - \hat{y}| \leq \delta \\ \delta(|y - \hat{y}| - \frac{1}{2}\delta) & \text{otherwise} \end{cases} \tag{5.23}$$

The details of the representation module and the fusion module will be discussed in the following.

5.4.3 Representation Module

Since the input data involve diverse modalities with the problem of heterogeneous data, it is necessary to learn from each individual input and form representations in the same representation space $E \in \mathbb{R}^d$, in order to facilitate the subsequent process of multimodal fusion.

Image Representation

In the previous work on oxide classification, we have uilitized a ResNet 50 model to train a multi-label oxide classifier. Since the trained model has proven that it can effectively classify the oxide types, this implies that the model has well learned the representation of the images with oxides. Hence, the trained oxide classification model is employed to extract the image representations. The hidden layer of the model before the classification head is chosen as the representation layer, yielding a vector $\mathbf{r} \in \mathbb{R}^{2048}$ for each image. Subsequently, a trainable linear layer is used to project the image features to the representation space. The image representations for each image is given by Eq 5.24:

$$\mathbf{v}^{img} = \mathbf{W}^{img}\mathbf{r} + \mathbf{b}^{img} \tag{5.24}$$

Where $\mathbf{W}^{img} \in \mathbb{R}^{d\times 2048}$ and $\mathbf{b}^{img} \in \mathbb{R}^d$ are weight and bias in the trainable linear layer.

If $M_{img} > 1$, then a concatenation of the image representations would be needed. The final image representation would be $\mathbf{V}^{img} \in \mathbb{R}^{M_{img}\times d}$.

Tabular representation

Note that the tabular data contain multiple categorical and numerical features. Different features may originate from distinct distribution, thereby requiring a heterogeneous representation approach [200, 201]. For each categorical feature, we use a simple look table, which is commonly used for word embedding, to represent categorical features. Each numerical feature is independently projected to representation space $E \in \mathbb{R}^d$ through a separate fully connected layer. The tabular representation for a tabular attribute t is given by:

$$\mathbf{v}_t^{tab} = \begin{cases} \mathbf{W}_t^{cat}\mathbf{e}_t + \mathbf{b}_t^{cat} & \text{if t is a categorical attribute} \\ \mathbf{W}_t^{num}t + \mathbf{b}_t^{num} & \text{otherwise} \end{cases} \tag{5.25}$$

Where $\mathbf{e}_t \in \mathbb{R}^{n_t}$ is one-hot vector for t categorical feature with n_t categorical options. $\mathbf{W}_t^{cat} \in \mathbb{R}^{d\times n_t}$ and $\mathbf{b}_t^{cat} \in \mathbb{R}^d$ are weight and bias in the trainable embedding layer for t categorical feature. $\mathbf{W}_t^{num} \in \mathbb{R}^d$ and $\mathbf{b}_t^{num} \in \mathbb{R}^d$ are weight and bias in the trainable fully connected layer for t numerical feature.

The final tabular representation would be the concatenation of categorical representations and numerical representations.

$$\mathbf{V}^{tab} = [\mathbf{v}_1^{cat}, ..., \mathbf{v}_{M_{cat}}^{cat}, \mathbf{v}_1^{num}, ..., \mathbf{v}_{M_{num}}^{num}] \in \mathbb{R}^{M_{tab}\times d} \tag{5.26}$$

Where $M_{tab} = M_{cat} + M_{num}$.

Time series representation

In this work, both GAF and CNN are employed for representing time series data. Assume that a time series attribute in $i-th$ production run $\mathbf{t}^{time} = \{t_1, t_2, ..., t_{n_i}\}$ with the length of n_i. Different production runs may have various time spans, making n_i a variable. Though PPA in GAF can help reduce the size of time series data into the same GAF size, the smoothed GAF encoding maps of the time-series with varying length might have different resolutions, which may potentially the representation learning of the subsequent model. Therefore, in this work, we first unify the length of all time series attributes among different runs. If the time series length is longer than the predefined length, then the extra part of the time series will be truncated off. Otherwise, if the time series length is shorter than the predefined length, then the times series will be padded with the time series from a previous preparation phase and the padding would be inserted in the beginning of the time series. After the truncation and padding, each time series is scaled in the interval [-1, 1] by MinMaxScaler, as seen in Eq 5.27.

$$\mathbf{t}^{time}_{scaled} = \frac{\mathbf{t}^{time} - min(\mathbf{T}^{time})}{max(\mathbf{T}^{time}) - min(\mathbf{T}^{time})} \times 2 - 1 \tag{5.27}$$

Where $\mathbf{T}^{time}$ is the concatenation of the time series attribute $\mathbf{t}$ from N production runs. $\mathbf{T}^{time} = \{\mathbf{t}^{time}_i, \mathbf{t}^{time}_i, .., \mathbf{t}^{time}_N\}$ with a total of N observations of production runs.

After the pre-processing, each time series attribute has the same length $\hat{\mathbf{t}}^{time} = \{\hat{t}_1, \hat{t}_2, ..., \hat{t}_m\}$ where m is the predefined time series length. Since the length of time series data determines the size of generated GAF image, PPA is empolyed to downsample the time series. PPA divides the time series into S_{GAF} bins, corresponding to the desired size of GAF image and extracts the average of each bin [196], resulting in the downsampled time series attribute $\hat{\mathbf{t}}^{time}_{PPA} = \{\tilde{t}_1, \tilde{t}_2, ..., \tilde{t}_{S_{GAF}}\}$. Then each time series attribute is encoded into a polar coordinate system. GASF, which is more commonly used, is adopted to map each time series attribute into a two-dimensional image $\mathbf{I} \in \mathbb{R}^{S_{GAF} \times S_{GAF}}$.

$$\begin{aligned} \phi_i &= \arccos{(\tilde{t}_i)}, \tilde{t}_i \in \hat{\mathbf{t}}^{time}_{PPA} \\ r_i &= \frac{i}{S_{GAF}}, i \in S_{GAF} \\ GASF &= [cos(\phi_i + \phi_j)] \in \mathbb{R}^{S_{GAF} \times S_{GAF}} \end{aligned} \tag{5.28}$$

Thereafter, the subsequent process follows the procedure to obtain the image representation. Considering the presence of multiple time series attributes and small sizes of the dataset which is prone to overfitting, a pre-trained ResNet18 instead of ResNet50 is employed as the shared backbone for extracting image representation, which yields a vector $\mathbf{r}^{GAF} \in \mathbb{R}^{512}$ for each image input. Time series attributes may come from distinct distributions, necessitating a heterogeneous representation strategy. Therefore, following the step in obtaining tabular representation, the image representations of each time series attribute are independently projected into the same representation space $E \in \mathbb{R}^b$ as other modalities using a separate fully connected layer.

$$\mathbf{v}^{time} = \mathbf{W}^{time}\mathbf{r}^{GAF} + \mathbf{b}^{time} \tag{5.29}$$

Where $\mathbf{W}^{time} \in \mathbb{R}^{d \times 512}$ and $\mathbf{b}^{time} \in \mathbb{R}^d$ are weight and bias in the trainable linear layer.

The final time series representation would be:

$$\mathbf{V}^{time} = [\mathbf{v}^{time}_1, \mathbf{v}^{time}_1, ..., \mathbf{v}^{time}_{M_{time}}] \in \mathbb{R}^{M_{time} \times d} \tag{5.30}$$

5.4.4 Multimodal Fusion Module

After the extraction of representations from each modality, the fusion step is necessary to aggregate discriminative information from individual representations to create more comprehensive representations for regression analysis. In this work, considering the fusion capability of the SE module in channels and the cross-modal interaction capability of the Transformer, multimodal fusion is done by an SE module and a Transformer encoder. Besides, while the complex architecture of Transformer makes it challenging to interpret, the SE module is also considered to increase the transparency of the model by inspecting the attention weights. Their details can be seen below.

Fusion by SE Module

Before multimodal fusion, a simple concatenation is performed on image representations $\mathbf{V}^{img}$, tabular representations $\mathbf{V}^{tab}$ and time series representations $\mathbf{V}^{time}$, resulting in the initial multimodal representation $\mathbf{V} = [\mathbf{V}^{img}, \mathbf{V}^{tab}, \mathbf{V}^{time}] \in \mathbb{R}^{M \times d}, M = M_{img}+M_{tab}+M_{time}$. In order to adapt the input shape requirement of the SE module, the multimodal representation $\mathbf{V} \in \mathbb{R}^{M \times d}$ is extended to $\mathbf{V} \in \mathbb{R}^{M \times d \times 1}$. Then SE module is applied on the multimodal representation with the following squeeze step and excitation step.

- **Squeeze step.** Average pooling is utilized to compute the channel-wise representation from global view.

$$Y_c = \frac{1}{M \times 1} \sum_{i=1}^{M} \sum_{j=1}^{1} \mathbf{V}_c(i, j) \tag{5.31}$$

- **Excitation step.** To capture channel-wise dependencies of the multimodal representation, a bottleneck with two FC layers are employed to compute the channel-wise attention, which would then be used for recalibrating the original multimodal representation.

$$\begin{aligned} \mathbf{S} &= \sigma(F_2(ReLU(F_1(\mathbf{Y})))) \\ \mathbf{V}_{SE} &= \mathbf{S} \odot \mathbf{V} \end{aligned} \tag{5.32}$$

In this way, each attribute is viewed as a feature map. By modeling the dependencies between channels in the multimodal representation and adjusting the importance of different attributes, the subsequent network can focus more on informative attributes while suppressing those that are less relevant.

Fusion by Transformer Encoder

Transformer is well-known for its strong capability in capturing the intricate denpendencies among the input due to the self-attention mechanism. In addition, Transformer can represent each tokenized input as a fully-connected graph in a topological geometry space [202]. This trait offers the Transformer high flexibility and compatibility with various modalities [152]. Hence, in this work, a Transformer encoder is employed to capture cross-modal dependencies and to create a fused representation that combines the strengths of each attribute while respecting their unique characteristics.

The representation from the SE module is first processed with a learnable position embedding PE by summation to retain the positional information of the input.

$$\mathbf{Z} = \mathbf{V}_{SE} + PE \tag{5.33}$$

The resulting representation from postition embedding is then served as the input to the Transformer encoder. In this study, we employ the Transformer architecture based on the one utilized in Vision Transformer [203]. This architecture is composed of alternating layers of multihead self-attention followed by MLP blocks[203]. Each MLP block contains two layers with a GELU non-linearity. The fused representation is finally obtained by extracting the averages on each channel of the final layer of the Transformer representation.

$$\mathbf{Z}_l^{'} = MHSA(LN(\mathbf{Z}_{l-1})) + \mathbf{Z}_{l-1} \tag{5.34}$$

$$\mathbf{Z}_l = MLP(LN(\mathbf{Z}_l^{'})) + \mathbf{Z}_l^{'} \tag{5.35}$$

$$\mathbf{Z}_{TransE} = \frac{1}{M}\sum_{m=1}^{M} \mathbf{Z}_L^m \tag{5.36}$$

Where $l = 1, 2, ..., L$. LN denotes the Layernorm layer. $\mathbf{Z}_{TransE}$ is the output representation of the Transformer encoder.

Then the representation from the Transformer encoder would be fed to the final MLP head to generate the estimate of moisture level. The MLP head consists of an LN layer and a linear layer.

$$y_{pred} = FC(LN(\tilde{\mathbf{Z}}_{TransE})) \in \mathbb{R} \tag{5.37}$$

5.4.5 Explainability Analysis

This work is more interested in what the model has learned, which can potentially provide us with insights into the root causes of the high moisture level in the FZ chamber. Therefore, this section aims to measure how much a given attribute matters for model predictions, supposing that the model has achieved satisfactory performance. To this end, XAI techniques are employed to increase the transparency of the model. Specifically, attention-based method supported by SE module, feature permutation and SHAP are compared.

Attention-based Explainability

Since SE module can directly output the importance scores for each channel for each sample, the averaged important scores for each attribute over all data samples are extracted. Supposing A is the attention weight extracted from the squeeze step in the SE module. The global importance score for $i-th$ P_i can be computed by Eq 5.38:

$$P_i = \frac{1}{N}\sum_{j}^{N} A_{ij} \tag{5.38}$$

Feature Permutation

Feature Permutation is a module-agnostic explainability technique that can break the connection between the feature and the true label [187]. In feature permutation, the importance of a feature is determined by the rise in the prediction error of the model when the values of the single feature are randomly shuffled [188]. Assume that the original model error is e_{base}. By shuffling the values of the $i-th$ feature/attribute and feeding the new dataset to the model, the model error e_i can be obtained. The importance of the $i-th$ feature P_i can be determined by comparing e_{base} and e_i, as expressed in Eq 5.39.

$$P_i = e_{base} - e_i \tag{5.39}$$

SHAP explainability

SHAP is a model-agnostic method which is solely based on the inputs and outputs of the model instead of the model performance. SHAP is based on the idea of cooperative game theory, specifically the concept of Shapley values, to address fair rewards among players [204]. SHAP extends this concept to the machine learning field by viewing the inputs as the players and the outputs as the total payoff. The core idea of SHAP is to compute the average contribution of each feature across all potential combinations of features, considering their interactions and dependencies. In this work, considering the heterogeneous properties of the input data, we apply SHAP on the concatenation of unimodal representations $\mathbf{V}$ with M input tokens. Then the contribution of the attribute i, equivalent to the Shapley value ϕ_i, can be given by Eq 5.41:

$$\phi_i = \sum_{S \backslash i} \frac{|S|!(M - |S| - 1)!}{M!} \left[v(S \cup \{i\}) - v(S) \right] \tag{5.40}$$

Where S denotes a subset of the input tokens, $v(S)$ represents the model output given by S. $|S|$ is the number of tokens in S.

By averaging the Shapley values over the dataset with size of N, global importance scores for the input attributes cna be obtained. 5.41:

$$\Phi_i = \frac{1}{N} \sum_{j}^{N} \phi_i^j \tag{5.41}$$

5.5 Experiments

5.5.1 Dataset and Preprocessing

The data set used is based on the dataset used for root cause analysis in Section 3.8, which contains 387 samples, with each sample having 27 relevant factors (oxygen and moisture level excluded but image added in input factors). The dataset was then split in a random but stratified fashion until 80% and 20% of the dataset were in the training split and test split, respectively. The reason why we did not set a validation split is that the dataset size is extremely small, and thus a cross-validation would be employed during training. Following the preprocessing procedures in oxide classification, all images were first scaled into [0,1] by dividing by 255, and then normalized into the range [-1, 1] by utilizing the mean of 0.5 and the standard deviation of 0.5. The continuous tabular data were also normalized by the global average and the global standard deviation in the overall dataset. In terms of time series data, they were scaled into [-1,1] using MinMaxScaler.

5.5.2 Baselines and implementation details

The time series representation with GAF, SE module and Transformer are relatively new components for the multimodal regression task. Therefore, in this study, we established several baselines in order to evaluate their effectiveness, as seen follows. The summary of baseline setup can be seen in Table 5.1.

- **B1: Multimodal LSTM-SE-Transformer**. In this baseline, the GAF module is substituted with a LSTM module in order to evaluate the performance of GAF.

- **B2: Multimodal LSTM-MLP**. Compared with B1, in this baseline, the multimodal fusion module is replaced by a MLP. Both the SE module and the Transformer are removed. The representations of individual attributes from representation module are first flatten and then projected into the embedding space E using a MLP block with two hidden layers.

$$\mathbf{V}_{MLP} = MLP(LN(\mathbf{V}_{flatten})) \tag{5.42}$$

  Where $\mathbf{V}_{flatten} \in \mathbb{R}^{M*d}, \mathbf{V}_{MLP} \in \mathbb{R}^{d}$.

- **B3: Multimodal GAF-MLP**. Compared to the baseline of B2 LSTM-MLP, the LSTM in the time series representation module is replaced by GAF.

- **B4: Multimodal GAF-SE**. Compared to the proposed model, the Transformer module is removed in this baseline to test the effectiveness of the Transformer module in fusing the representation ability. The multimodal representation from SE module reduces their size by average pooling before feeding into the MLP head.

- **B5: Multimodal GAF-Transformer**. Compared to the proposed model, the SE module is removed in this baseline to test the effectiveness of the SE module in enhancing the representation ability.

- **B7: Multimodal GAF-Transformer-SE**. The relative order of the SE module to the Transformer is altered in this baseline. This aims to investigate the influence of the order of the SE module.

- **B7: Multimodal GAF-SE-Transformer (proposed)**.

Considering the dataset size is extremely small, training hyperparameters may greatly affect the model performance. Hence, for a fair comparison, two hyperparameters: learning rate l and training batch size b are chosen by 5-fold cross validation over $l \in \{1e^{-4}, 1e^{-3}, 1e^{-2}\}$ and $b \in \{16, 32\}$ on the training split. Cross-validation is a method that can help to avoid overfitting and make the most of the available data. The principle of 5-fold cross-validation can be seen in Figure 5.9. The training split is randomly divided into five subsets, namely five folds. Then the model is trained with given a hyperparamter set for 5 rounds. With each round four folds would be selected as train split for training and the remaining fold would be used for evaluation. Then the predictions would be averaged over 5 validation splits, resulting in the final performance of the given hyperparameter set. The optimal set of hyperparameters for the model can be determined by identifying the combination that yields the best performance during cross-validation.

The predefined uniform length for the time series data was set to 450. The GAF image size S_{GAF} was set to 256. The number of training epochs was set to 200 while the early stopping

Table 5.1: Baselines setup.

Model	Time Series Representation		Fusion Module			
	GAF	LSTM	SE(←)	TransE	SE(←)	MLP
B1		✓	✓	✓		
B2		✓				✓
B3	✓					✓
B4	✓		✓			
B5	✓			✓		
B6	✓			✓	✓	
B7	✓		✓	✓		

Note: ← and

⇒

represent the order the SE module compared with Transformer module. ← denotes that SE module is before the transformer module and vice versa. ↓ means that the lower the metric value, the better the performance.

regularization method [205] was employed to avoid overfitting. We used the Adam optimizer with (β_1, β_2) = (0.9, 0.999) and one-cycle policy, with a maximum learning rate of l. Besides, as mentioned in Section 2.4, the nondeterministic nature of network optimization necessitates a fixed random seed to obtain repeatable results. To assess the model's resistance to randomness, especially when the dataset is small, each experiment was replicated 5 times using 5 different random seeds (0, 42, 100, 200, 300) and the average and standard deviation of the results were recorded. The training procedure can be seen in Figure 5.10.

In terms of evaluation metrics, Root Mean Squre Error (RMSE) was employed to evaluate the model performance. Their details in the computation can be seen in 5.43.

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2} \tag{5.43}$$

With regard to explainability analysis, three explainability techniques including attention-based method, feature permutation and SHAP method were compared. The explainability analysis is performed based on the proposed model trained with MSE loss, batch size of 32 and maximum learning rate of 1e-4. All important scores are computed on the test dataset. Fuzzy Jaccard Index (FUJI) [206] was employed to evaluate the similarity of the feature rankings yielded by these three explainability techniques. FUJI is a scale-invariant similarity measure used to compare two ranked/ordered lists. Specifically, we used the area under the FUJI curve (FUJI AUC) as the similarity measure. Besides, it should be noted that there is no ground truth in the explainability analysis,

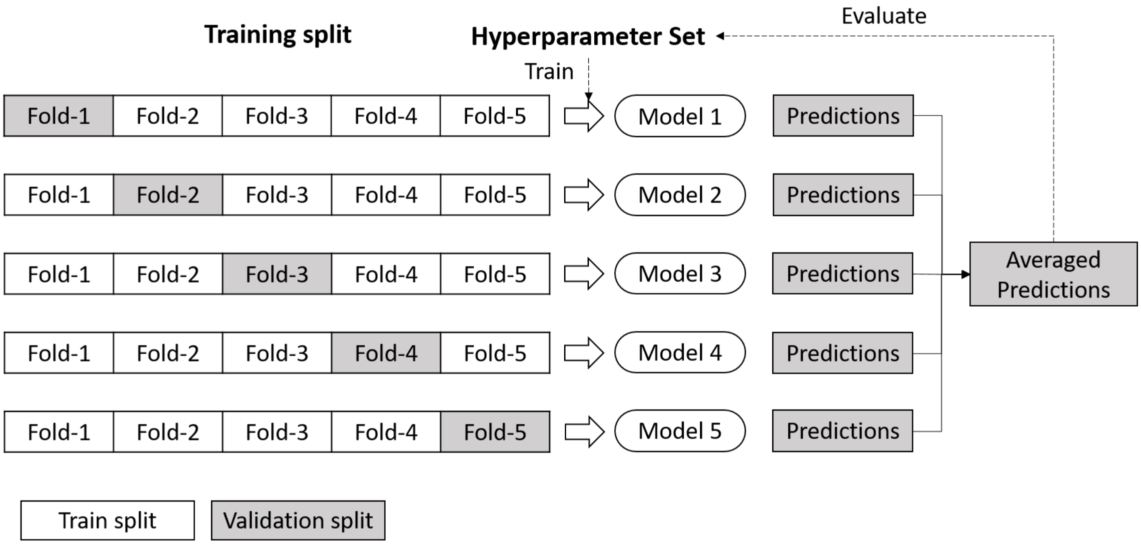


Figure 5.9: The principle of 5-Fold cross validation.

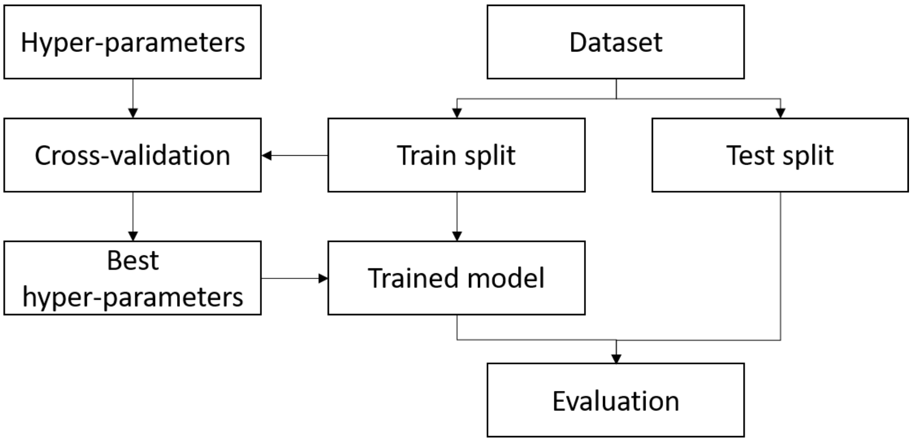


Figure 5.10: Training procedure with hyperparameter tuning using cross validation.

as even for us humans, it remains unclear which features are significant. This poses a difficulty in evaluating their effectiveness. In this work, the model performance yielded by tok-k features from the feature ranking generated by these explainability techniques (the rest of the features were excluded during training) were recorded. It is anticipated that with the inclusion of more influential features the model performance would be enhanced.

5.6 Results and Discussion

5.6.1 Comparison Results

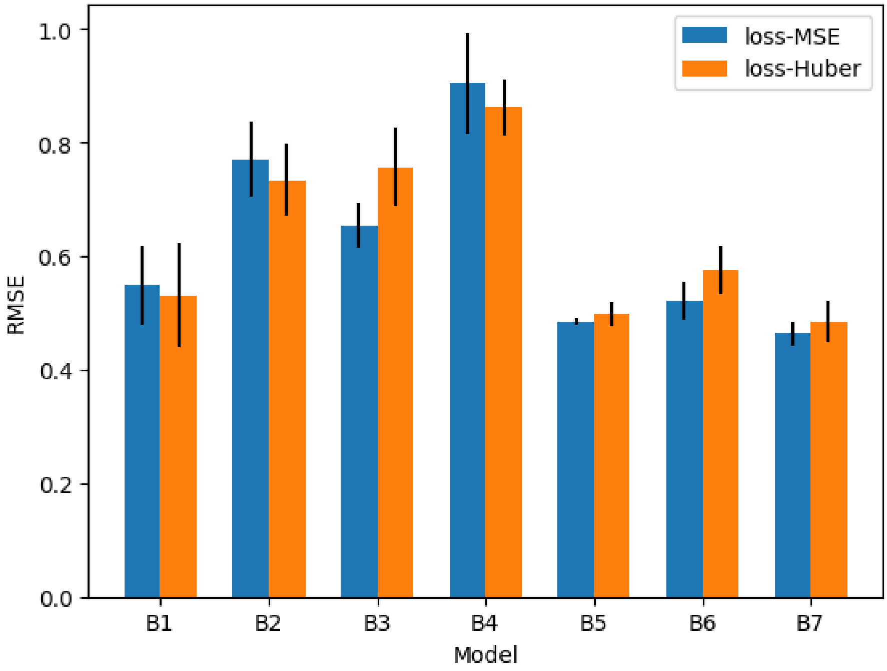


Figure 5.11: The comparison of baselines results with average and standard deviation.

The comparison results between all the baselines are shown in Tables C.19 and C.20. These results are also demonstrated in the form of bar charts, as seen in Figure 5.11. In addition, considering the relatively large uncertainty due to the small dataset and models, one-way Analysis of Variance (ANOVA) test was also conducted to determine if there were any significant differences among the performance of the baselines. Considering that there are only five samples in each group, the Kruskal-Wallis test, a non-parametric test, followed by Dunn's test for post-hoc comparison analysis, was employed. Figure 5.12 shows the significance differences between other baselines with B7, given by the Kruskal-Wallis test followed by a multiple comparison test (Dunn's test).

As seen in Figure 5.11, for the baselines trained with MSE loss, the average performance of B7 (proposed) outperforms other baselines. However, Figure 5.12 (a) shows that not all baselines are significantly different from B7. Specifically, the pairs of B5-B7 and B6-B7 did not show significant differences, suggesting that they have similar capabilities. This may indicate that the addition of the SE module and the order of the SE module may not contribute significantly to the model performance. Nevertheless, the significantly higher performance of B7 than B1 suggests that the effectiveness of the GAF module while the significantly higher performance of B5 than B3 and B4 indicates the effectiveness of the Transformer encoder module in fusing representations compared with solely using the SE module or solely using the MLP module.

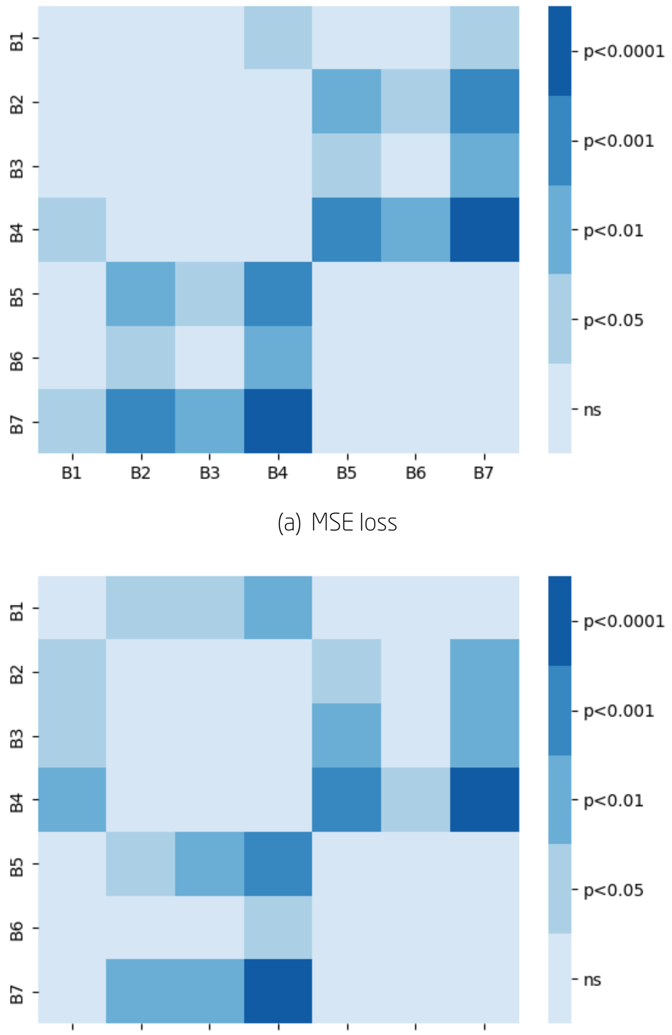
In terms of those baselines trained with Huber loss, B7 (proposed) outperforms other baselines in average performance (see Figure 5.11). However, Figure 5.12 (b) shows that B7 did not significantly outperform B1, B5 and B6. While this confirms the weak contribution of the SE module and the order of the SE module, the presence of no significant difference between B1 and B7 is rather surprising, which might be explained by a larger standard deviation of B1 trained with Huber loss than trained with MSE loss.

Notably, B1-B4 tend to have a large variance in model performance, suggesting that the use of LSTM in time series representation and the use of solely SE or MLP in representation fusion might contribute to such variability. Furthermore, the cross-validation results during hyperparameters presented in Table C.5, C.6, C.8, C.10 also confirm that LSTM is vulnerable to vanishing and exploding gradient problems.

The predictions from the proposed model (Multimodal GAF-SE-Transformer) on the test dataset are presented and compared with the labels in Figure 5.13. Figure 5.14 also illustrates the scatter plot of prediction values vs. labels. Figure 5.13 shows that the predicted moisture level generally follows a trend similar to the actual moisture level over production runs. However, there are instances where the model underestimates or overestimates the moisture level. Examining Figure 5.14 reveals this more clearly. In particular, the model tends to overestimate the moisture level when the measured moisture level is at a low level, while it is likely to underestimate the moisture level when the measured moisture level is at a high level. Overall, the model provides a reasonable estimate of moisture level, but it may require further refinement to improve accuracy, for instance, increasing the amount of data or involving more relevant potential factors, etc.

5.6.2 Explainability Analysis

Despite the lack of significant difference in the incorporation of the SE module, B7 was still chosen to be the object of explainability analysis, considering the transparency of the SE model. Figure 5.15 compares the similarity of the feature rankings from SHAP values, Feature Permutation



(a) MSE loss

(b) Huber loss

Figure 5.12: A heatmap showing the significant differences in the performance of the baselines, as determined by the Kruskal-Wallis test and Dunn's test, is presented. No significant difference is indicated by 'ns'.

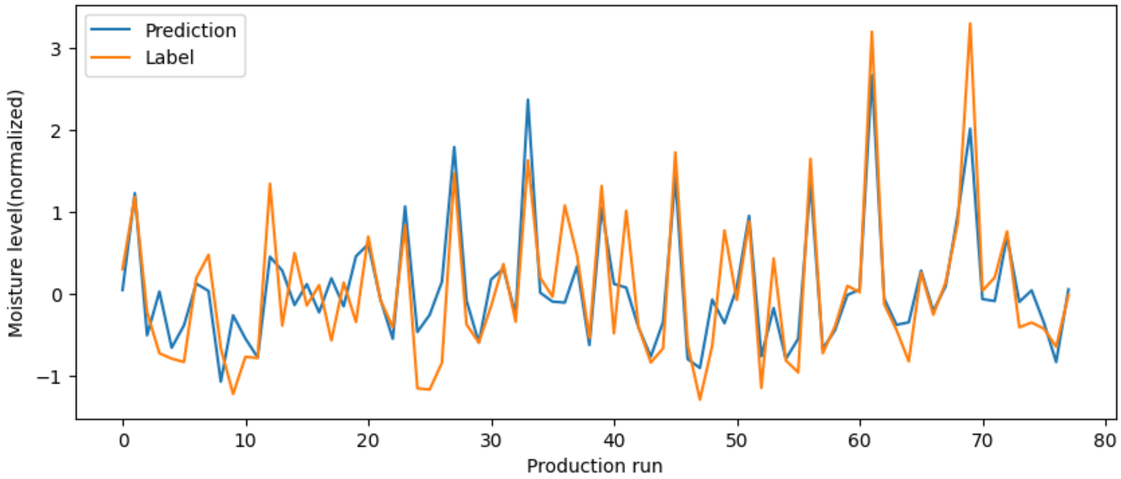


Figure 5.13: A line plot compares the moisture level labels and moisture level predictions from the Multimodal GAF-SE-Transformer trained by MSE loss under a train seed of 42. The x axis of the line plot represents the production run and the y axis shows the mositure level.

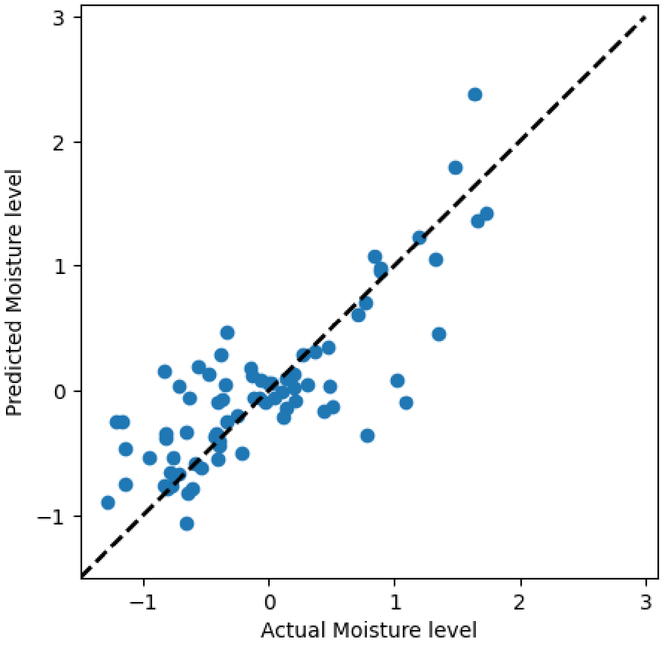


Figure 5.14: The scatter plot of predictions vs. labels, from the proposed model trained with MSE loss, batch size of 32 and learning rate of 1e-4.

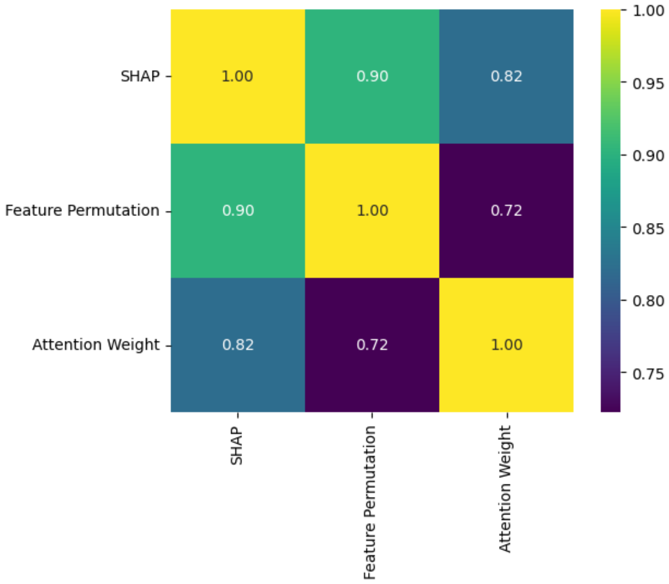


Figure 5.15: The similarity matrix of the important scores among three explainability techniques.

values and attention weights, respectively, using FUJI AUC. As seen, the Feature Permutation important scores appear to be the most similar with SHAP scores, while the similarities between attention scores and the other two methods are relatively low. We also report the model performance yielded by tok-k features (the rest of the features were removed during training) in Figure 5.16. The dashed line represents the baseline model performance using all available features. It is shown that the model generally performs better with additional top features from the SHAP method and Feature Permutation, which indicates that both methods can identify the informative features that contribute the most to the moisture level. These two methods arrive at the baseline performance at the top-6 features. This information can guide decisions about feature selection. However, in terms of attention-based method, only slight improvements are observed in model performance until the model reaches the top-10 features, where a remarkable increase in performance occurs. This suggests that the additional features except for the 10-th top feature may not be contributing significantly to the model prediction. On the other hand, this implies that the explanation ability from attention based method is limited compared to model-agnostic methods.

The top-10 features ranked by these three explainability techniques are demonstrated in Figure 5.17, Figure 5.18 and Figure 5.19. Image factor appears in these all rankings. However, since the image is actually the symptom of high moisture level rather than a cause, it would not be discussed here. Instead, the image is only used to improve the ability of the model to predict the moisture level. It is apparent that there is a high degree of convergence in high-ranked

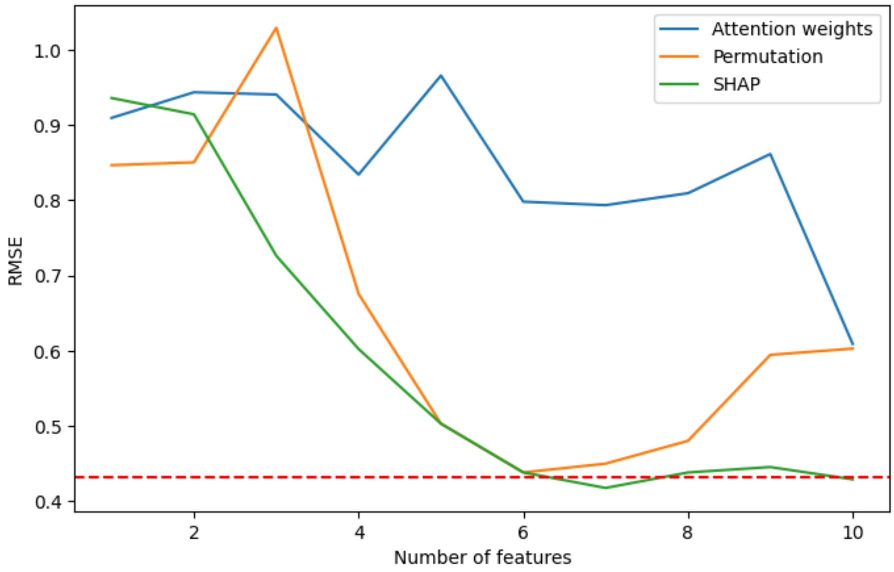


Figure 5.16: Model performance (RMSE in vertical axis) yielded by top-k features from feature rankings generated by attention based method, Feature Permutation and SHAP.

features between SHAP method and Feature Permutation method, though with a different in the feature order. This shared subset of features is particularly informative and represents a stable set of influential variables for predicting the moisture level. For instance, LeakRateLast, SPC_PUMP1TIME, SPC_PUMP2TIME consistently appear as highly important in both SHAP and Feature Permutation analysis. Figure 5.20 visualizes the association among normalized LeakRateLast, SPC_PUMP1TIME, SPC_PUMP2TIME and moisture level by using a 3D scatter plot. It is observed that a high moisture level tends to occur under conditions of high LeakRate in the last pump and low pump time. This information guides us to increase the pump time and carefully monitor the leak rate after the last pump during the preparation of the FZ machine, in order to avoid high moisture level in the FZ chamber.

The association between Month and moisture level is represented in Figure 5.21, where a seasonal trend can be observed. In particular, high moisture level tends to frequently occur in July while during winter the moisture level is relatively low. This implies that the state of the FZ machine might be sensitive to environmental conditions. Besides, this could also be explained by maintenance schedules for FZ machines tied to seasonal changes. Machines may be serviced more frequently when they are less intensively used. However, more investigations are still needed to determine whether the high moisture level is due to machine components or due to maintenance schedules.

As P8 time series attribute appears in both SHAP and Feature Permutation ranking, Figure 5.22 reports the GASF images of the P8 attribute under different moisture levels. The main diag-

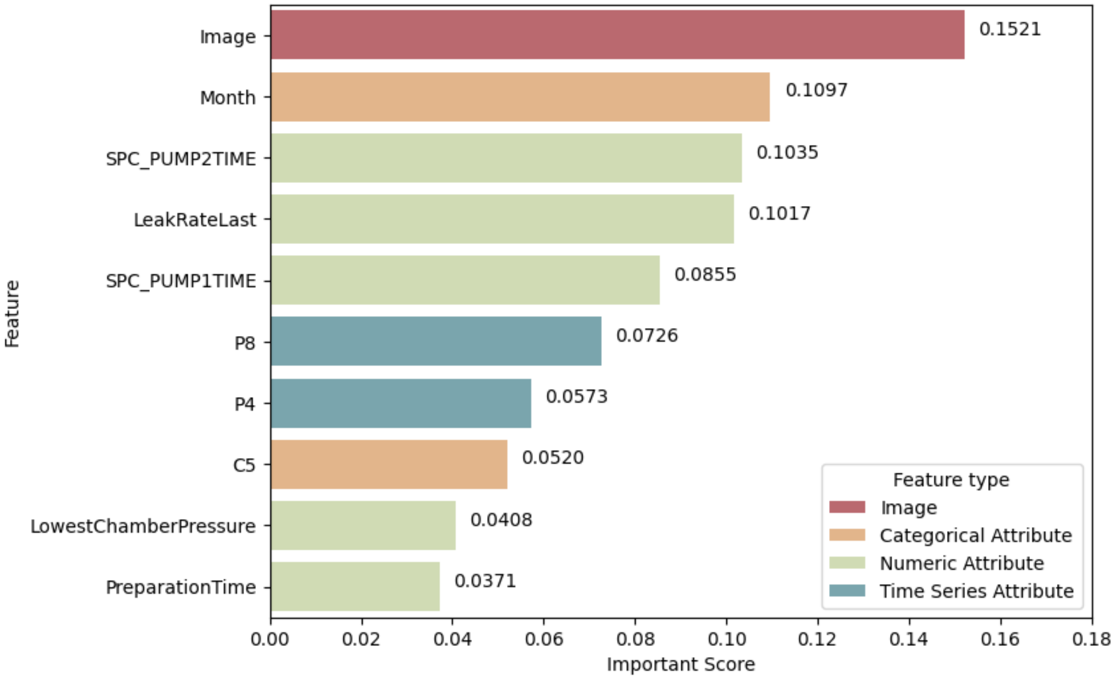


Figure 5.17: Top-10 important features ranked by SHAP values.

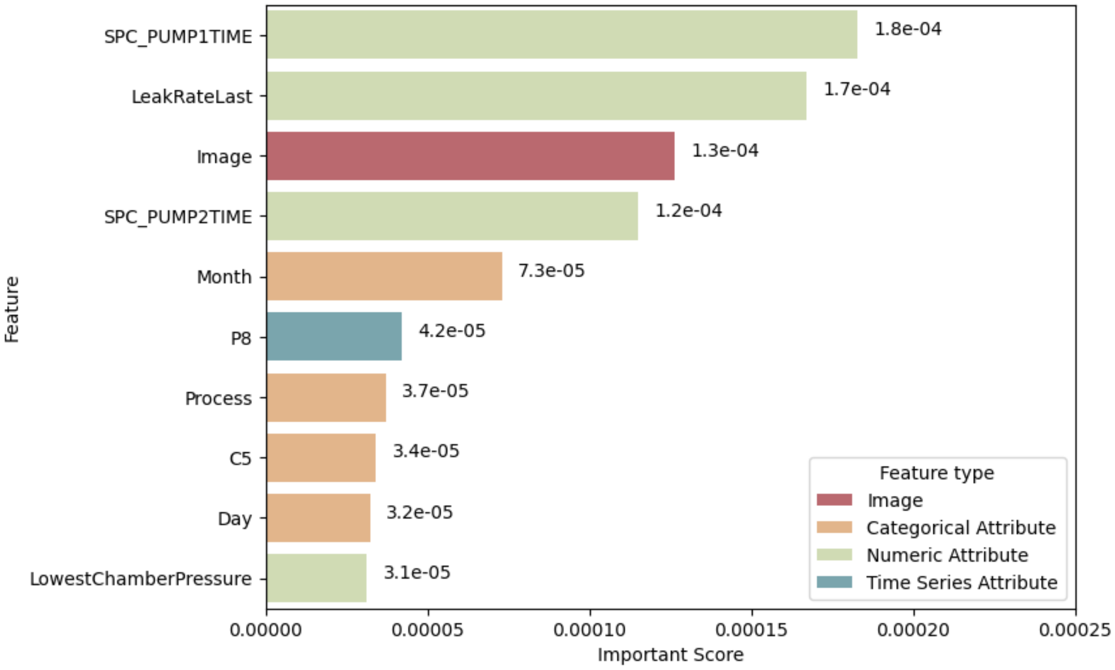


Figure 5.18: Top-10 important features ranked by Feature Permutation values.

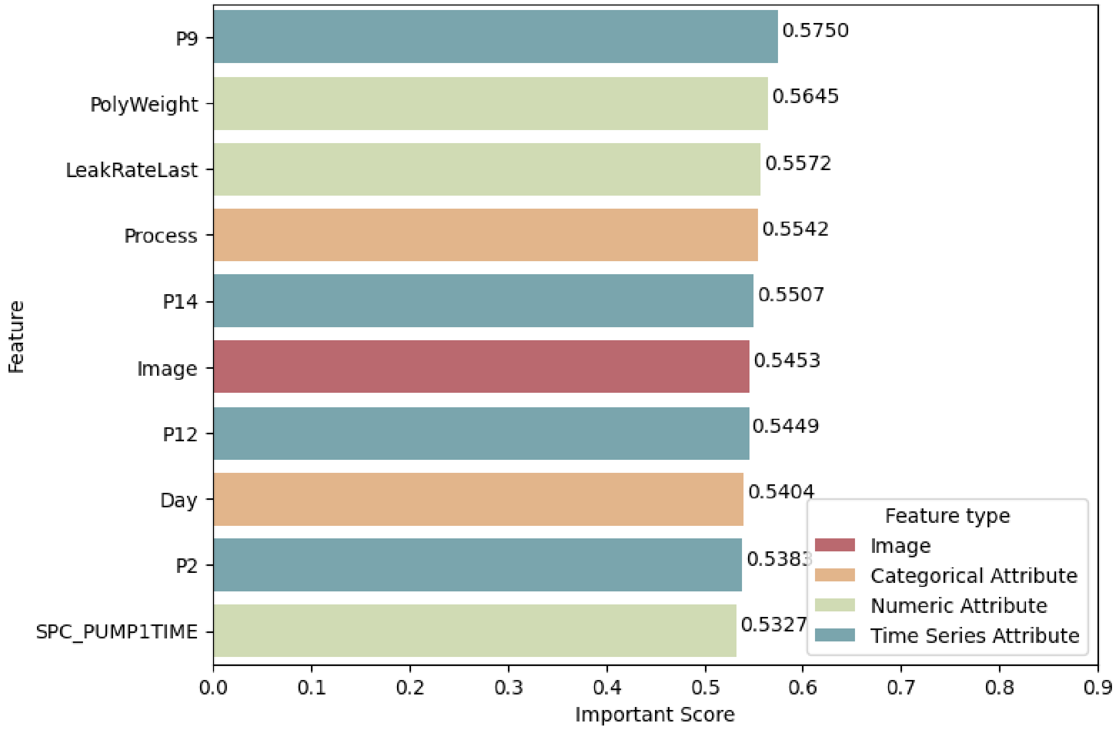


Figure 5.19: Top-10 important features ranked by Attention weights.

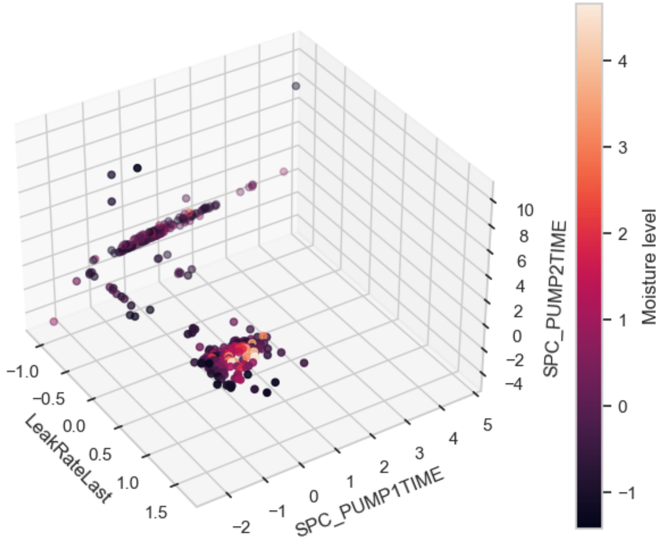


Figure 5.20: Data visualization among LeakRateLast (normalized), SPC_PUMP1TIME (normalized), SPC_PUMP2TIME and moisture level (normalized).

onal of the GASF image illustrates the alteration in time [207]. The pixels that are highlighted in the diagonal line suggest that they could be near the peak or the lowest point of the time series data (because $GASF = cos(2 * arccos(1)) = 1$ or $GASF = cos(2 * arccos(-1)) = 1$). Similarly, if a continuous horizontal line and a continuous vertical line around the highlighted pixel is observed, this correlation indicates that the entire time series is at a relatively high (beyond the average) or relatively low level (below the average). However, if the highlighted lines are disrupted, as seen in Figure 5.22 (e), (f) and (g), the time series may experience a phase shift from the peak to the lower point than the average, or vice versa, which can be inferred by $GASF = cos(arccos(1) + arccos(-1)) = -1$. This indicates that to achieve a moisture level, a careful focus on the phase shift in the P8 attribute might be necessary.

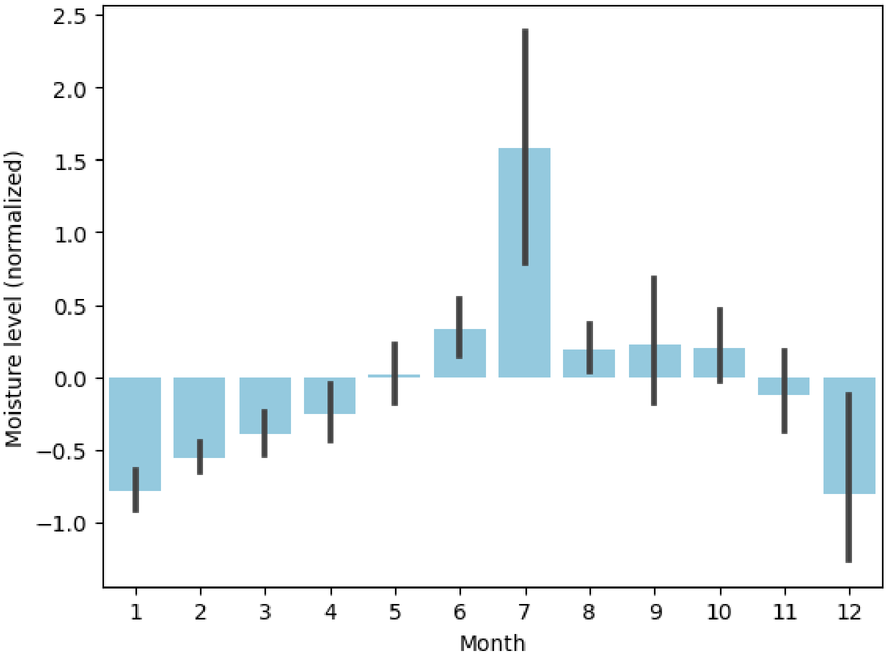
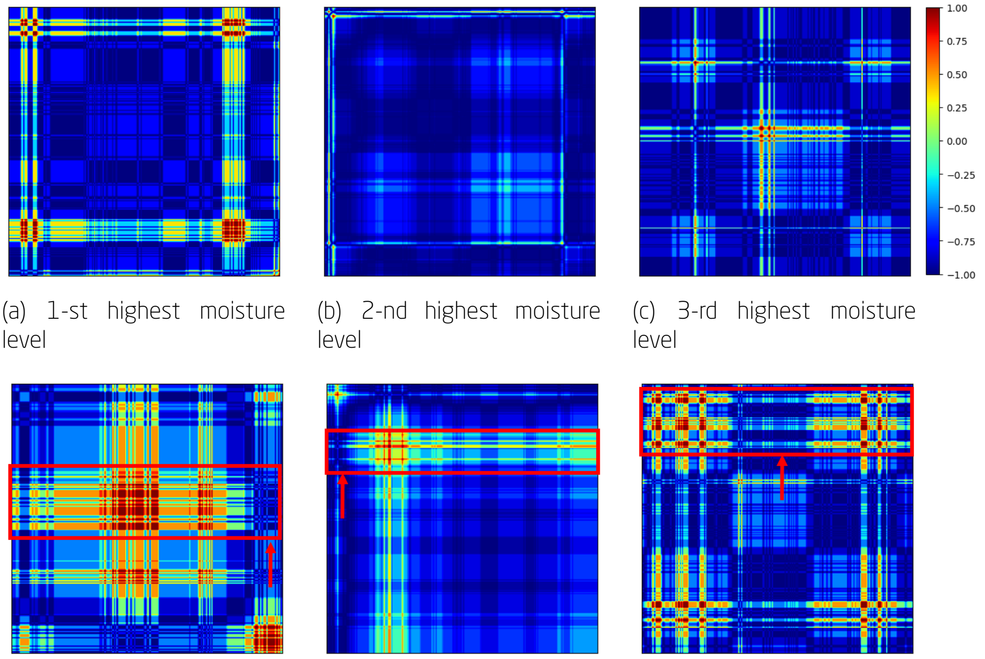


Figure 5.21: The distribution of moisture level across all months.

However, it should be noted that the results of the explainability analysis reveal only the association relationship, which does not necessarily imply casual effects. Association indicates that there is a relationship between two variables, while causation implies a direct cause-and-effect relationship between variables [208]. Causation implies association; however, the reverse is not necessarily true [208]. Therefore, one cannot determine the cause-effect merely on the basis of the explainability analysis. A deeper level of evidence including experimental designs and theoretical understanding is required to uncover a causal relationship.



(a) 1-st highest moisture level

(b) 2-nd highest moisture level

(c) 3-rd highest moisture level

(e) 1-st lowest moisture level

(f) 2-nd lowest moisture level

(g) 3-rd lowest moisture level

Figure 5.22: Comparision of GAF images of P8 time series data at different moisture levels. P8 attributes at low moisture level share a common point: they have highlighted but not continuous lines, with the disruption indicated by arrows.

5.7 Conclusion

The root cause analysis has revealed that a high moisture level is a major physical factor that caused the oxide problem. However, the reason why this high moisture level was present in the FZ chamber is still unclear. Both the FZ machine and the FZ process are complex, making it difficult to pinpoint the exact cause. To get to the bottom of the problem, it is necessary to trace back from this physical cause to a higher level of root cause, such as a log-action. Hence, this chapter mainly focuses on the solution of **Factors** ⇒ **Root causes**, by establishing a moisture level predictor and performing an explainability analysis. The moisture level predictor aims to estimate the moisture level at the beginning of the cone phase, based on the potential factors identified in Chapter 4 along with the FZ images.

Since the input data involve multiple modalities, with each modality associated with a specific sensor output, the research problem is therefore characterized as a multimodal regression problem. In this study, a Multimodal GAF SE-Transformer model was proposed for such a multimodal problem involving image data, tabular data and time series data to address the problem of heterogeneous data as well as to jointly learn meaningful representation from these modalities. Specifically, GAF was employed for the extraction of time series representation by encoding time series attributes to images, which would be subsequently processed by a CNN. Then each attribute in the input data would be transformed to abstract representation by means of CNN or MLP. The concatenation of these representations would be fed into a fusion module composed of an SE module and a Transformer encoder module. Finally, the multimodal representation from the transformer encoder would be fed to the final MLP head to generate the estimate of the moisture level. In addition, since we are more interested in what the model has learned for the further root causes of the high moisture level, explainability analysis was carried out by means of XAI techniques. Specifically, both an attention-based method supported by the SE module and a model-agnostic method SHAP were employed to measure how much a given attribute matters for model predictions.

The experimental results showed that the proposed model trained with both MSE loss and Huber loss can outperform other baselines in average performance, achieving the lowest RMSE error of 0.4642. However, the statistical test only revealed the effectiveness of the GAF module and the Transformer module while no significant difference was observed in the addition of the SE module and the order of the SE module. In terms of explainability analysis, the similarity test was performed on SHAP method, Feature Permutation and attention based method. It was found that there is a high similarity measure between SHAP method and Feature Permutation, while attention-based method appears with relatively similarity measures with other two methods. This was confirmed by looking into the top-10 features ranked by these three explainability techniques. The effectiveness of these three methods was examined by model performance yielded by top-k features. The results showed that SHAP method and Feature Permutation method can generally identify the most important features. Additionally, they reached the baseline model performance using all available features using top-6 features, which can provide us insights of feature selection. However, no significant improvement was observed in the attention-based method with additional top ranking features, which implies the limited explanation ability of attention weights. Nevertheless, the important features ranked by SHAP values revealed that factors such as month, pump time and leak rate contributed the most to the moisture level, which can guide us to focus on these factors when investigating moisture level.

CHAPTER 6

A FRAMEWORK FOR AUTOMATIC RESPONSE OF OXIDE

"Do not wait to strike till the iron is hot; but make it hot by striking." - William Butler Yeats

6.1 Motivation

The mitigation of the oxide problem is not a one-time task but an ongoing process. As mentioned above, the oxide problem within the FZ process is a challenge, as both the FZ process and the FZ machine are sophisticated. The challenge often results from a combination of intricate factors, including material properties, environmental conditions, and numerous production parameters and intricate machine settings. A one-time solution or intervention may be effective in the short term, but it is unlikely to be sufficient in the long term. Addressing the oxide problem requires constant monitoring and dynamic adaption. Central to this notion is the idea of continuous learning and improvement. Each cycle of detection, analysis, and action contributes to a growing body of knowledge about the oxide problem. Over time, the industry would become more adept at recognizing early warning signs, understanding root causes, and fine-tuning their mitigation strategies. It is a dynamic process where insights gained from past experiences inform and enhance future actions, thus fostering an ever-evolving and improving environment.

Therefore, with the acknowledgment of the complexities of the oxide problem and the aspiration to pursue excellence in the process and product quality, a conceptual framework regarding the dynamic and adaptive strategy to respond to the oxide problem is proposed in this work. The framework aims at establishing a systematic and methodical approach to proactively address the oxide issue, with the ultimate goal of assuring product quality, increasing crystal yield, and reducing cost.

6.2 Related Works

Problem-solving in manufacturing has long been a critical focus to enhance product quality, machine efficiency, and overall competitiveness. Many manufacturing companies employ Continuous Improvement. When production fails to meet its target, it is essential to have a process to investigate the issue and come up with a solution in an organized manner. The normal entire process of addressing a problem involves recognizing the problem, determining the root causes, and then making the necessary managerial decisions and actions to bring the process to a normal state [209].

Several problem-solving methodologies have been studied, including Kepner-Tregoe, Plan-DO-Check-Act (PDCA), and Eight Disciplines (8D). These systematic methods provide a clear structure to help improve and discover solutions for responses to a problem. However, these methods require extensive expert knowledge and experience of team members for the timely detection of a problem and diagnosis of the problem. Moreover, most traditional problem-solving methods are reactive. When a similar problem occurs, these problem-solving processes are often repeated and a new investigation is conducted [210], resulting in higher costs in human resources and production inefficiency [124]. Although among these traditional methods there are some proactive methods that can anticipate potential problems and allow proactive measures to prevent failures from occurring [211], they tend to be subjective as they rely on the expertise and experience of team members. With the advancement of digitalization and the growing amount of associated data, these challenges can be overcome. The abundance of data and the availability of computational resources enable the paradigm shift from knowledge-driven to data-driven solutions, such as using data mining and machine learning to make the problem-solving process efficient [124]. Furthermore, data-driven solutions can bridge the gap between problem identification and diagnosis by establishing automation of problem detection and diagnosis, allowing rapid responses to problems as they arise [212, 213]. The analytical capabilities in data-driven solutions particularly enabled by machine learning can be structured into four stages: Know-What (Knowing what is happening), Know-Why (Knowing why it happened), Know-When (Knowing when it will occur) and Know-How (Knowing how to respond) [43]. The normal diagnostic paradigm for problem-solving empowered by data-driven solutions involves Know-What, Know-Why, with optional Know-How.

With the advancements of data analytics and computational capabilities, another paradigm of responding to a problem has been made possible: prognosis, with the aim of predicting when an abnormal behavior is likely to arise [214], which is equivalent to the analytical capability of Know-When. Instead of focusing on the past and present states, prognosis emphasizes the forward-looking way to cope with a problem. Such a concept has been an emerging field applied in the levels of product, process, machine, and system [43]. Examples include predictive quality control [52], process characteristics prediction [215], predictive maintenance [216] and predictive production disturbances [53]. With the timely prediction of abnormal events or variables, prognostics can allow the decision maker to take action in the early stages to minimize the probability of failure at lower costs [217].

In this work, the objective is to offer a well-structured and methodological framework that not only provides guidance for discovering problem-solving solutions, but also emphasizes the significance of feedback loop enabling continuous improvement. A conceptual data-driven framework is proposed to provide a quick and efficient response to the oxide issue, taking into account both diagnostic and prognostic strategies to respond to the oxide problem. As such, our research seeks to offer a holistic solution for proactively managing the oxide problem, which will be elaborated upon in the subsequent methodology section.

6.3 Methodology

The proposed conceptual framework for proactively managing the oxide problem comprises two levels of response: diagnostics and prognostic. Each level of response is structured in 'Four-Know' stages: Know-What, Know-Why, Know-When, Know-How. The overview of the framework can be seen in Figure 6.1.

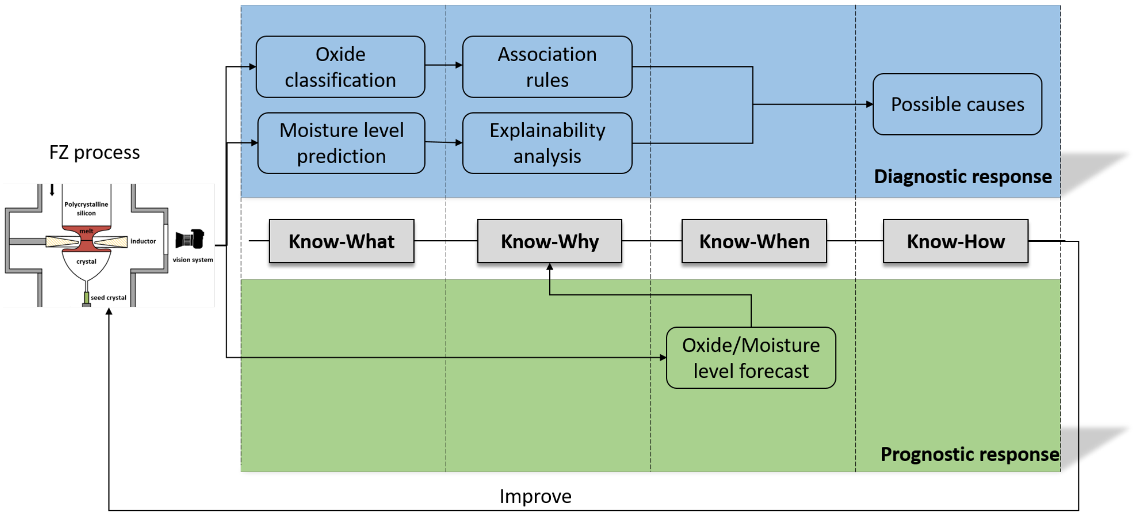


Figure 6.1: The conceptual framework for continuous improvement in mitigating the oxide problem.

Diagnostics response mainly involves the commonly used systematic analysis: identification of the problem (Know-What) followed by their underlying causes (know-why) and subsequent guidance for improvement (Know-How). Know-What in the framework integrates oxide classification and moisture level prediction, serving as the initial screening mechanism for identifying the oxide anomalies or high moisture level. It should be noted that Know-What is not limited solely to the two identification methods proposed in this thesis. It can also be extended to use other variables relevant to the occurrence of the oxide as detection targets or be operated in near real-time for ensuring swift and accurate identification. Subsequently, when the oxide problem is detected, the framework walks into Know-Why for retrieving potential root causes. The potential root causes can be either retrieved from the "IF THEN" knowledge database established from

Association Rule Mining, or retrieved from the features with high local-explainability scores given by explainability analysis on the moisture level predictor. For instance, the identified oxide types as well as the estimated moisture level, can be used to search for rules that meet the specific conditions within the association rules database. To prioritize and allocate resources efficiently, one can evaluate the lift values in the identified rules to identify where intervention is most urgently required. Similarly, one can also turn to explainability analysis to identify the urgent interventions by choosing the features with high local-explainability scores given by explainability analysis on the moisture level predictor. The potential root causes are then used to generate recommendations for process improvements. These recommendations are grounded in a deep knowledge of the causative factors and the factors influencing predictive model decisions. The aim of the diagnostics response is to foster ongoing enhancement. Each cycle of detection, root cause analysis, and action yields knowledge that is incorporated into the response procedure itself, ensuring the process is continuously improved to mitigate or even preclude the oxide problem.

With the accumulation of knowledge yielded from diagnostic response, the industry would become more adept at recognizing early warning signs and the underlying causes. As such, the industry can move on to the next stage of the response, prognostic response. The difference between prognostic and diagnostic response lies in the identification of the problem. Rather than concentrating on monitoring the current and past conditions, prognosis puts emphasis on a forward-looking approach to address an issue. This relies on the capability of Know-When in forecasting the occurrence of the oxide problem. Know-When can be performed by establishing predictive models that forecast the future behavior of the oxide or moisture level. Then the decision-maker can either take preventive measures, for instance process termination or maintenance scheduling for FZ machines to reduce costs. The decision-maker can also move to Know-Why to identify the long-term improvement actions to decrease the probability of oxide occurrence. In this way, the management of the oxide problem can be transformed into the data-driven anticipatory approach, allowing the taking of preventive measures in the very early stages.

6.4 Conclusion

The oxide problem is a critical issue in the FZ process. Addressing it is not a one-time task but an ongoing process. In this work, a conceptual data-driven framework was proposed to provide a quick and efficient response to the oxide problem and feedback loop for process improvement, with the ultimate goal of assuring product quality, increasing crystal yield, and reducing cost. The framework took into account both diagnostic and prognostic strategies to respond to the problem. Each strategy of response is structured in the 'Four-Know' stages: Know-What, Know-Why, Know-When, Know-How. The diagnostic strategy integrated oxide classification and moisture level prediction in the cores of Know-What. If an anomaly is identified, the findings of the root cause analysis in Know-Why would be used to determine potential causes. These potential root causes can be obtained from the "IF THEN" knowledge base from Association Rule Mining or from the features with high local-explainability scores provided by the explainability analysis of the moisture level predictor. Then recommendations can be drawn from these potential causes to improve the process. An alternative solution for responding to the oxide problem was also introduced, a prognostic strategy. The prognostic strategy relies on the ability of Know-When to forecast the occurrence of the oxide problem, thus allowing the decision-maker to take preventive measures or identify improvement actions to decrease the probability of the oxide occurrence. In summary, the developed conceptual framework seeks to offer a well-structured and methodological framework that not only provides guidance for discovering problem-solving solutions, but also emphasizes the significance of feedback loop enabling for continuous improvement.

CHAPTER 7

SUMMARY AND OUTLOOK

"The best way to predict the future is to create it." - Peter Drucker

7.1 Summary

With the ambition of increasing crystal yield and reducing costs, the research presented in this thesis aimed to investigate a surface anomaly emerged on polysilicon during Float-Zone crystal growth process, which turns out to be oxide contamination. The investigation was conducted in the following steps: Know-What, Know-Why and Know-How. First, the related works for each technological area were examined to create a starting point for understanding the limitations and opportunities. Following the establishment of the state-of-the-art, the investigation addressed a diverse aspects of the oxide problem. The nature and characteristics of the surface anomaly were analyzed by microstructural material characterization and visual characterization. Then data-driven Know-What investigation, oxide identification, was conducted by establishing a multi-label classifier based on the images. To identify the fundamental triggers of the oxide problem, Know-Why was investigated by conducting a root cause analysis of the oxide problem. In order to get closer to the heart matter of the oxide problem, a further Know-Why was conducted. A contributor for the oxide problem identified from root cause analysis findings, moisture level, was selected as the responsible variable of a predictor. After the training of the predictor, explainability analysis was performed on the predictor for further root cause analysis. Lastly, a Know-How framework was presented for continuous improvement by integrating the findings from Know-What and Know-Why. In general, Figure 7.1 presents the overview of the structure and organization of this thesis following the problem solving framework: Know-What, Know-Why and Know-How. This PhD project offers an understanding that can be utilized to improve the Float-Zone crystal growth process with a particular focus on a surface anomaly emerged on polysilicon during the FZ process, which turned out to be oxygen contamination. The investigation in this Ph.D. thesis has led to significant insights and advancements in the field of FZ crystal growth, paving the way for enhanced yield and reliability in this critical industrial process. The research was organized based on the research objectives outlined in Section 1.3.

What are the nature and characteristics of the surface anomaly and its potential

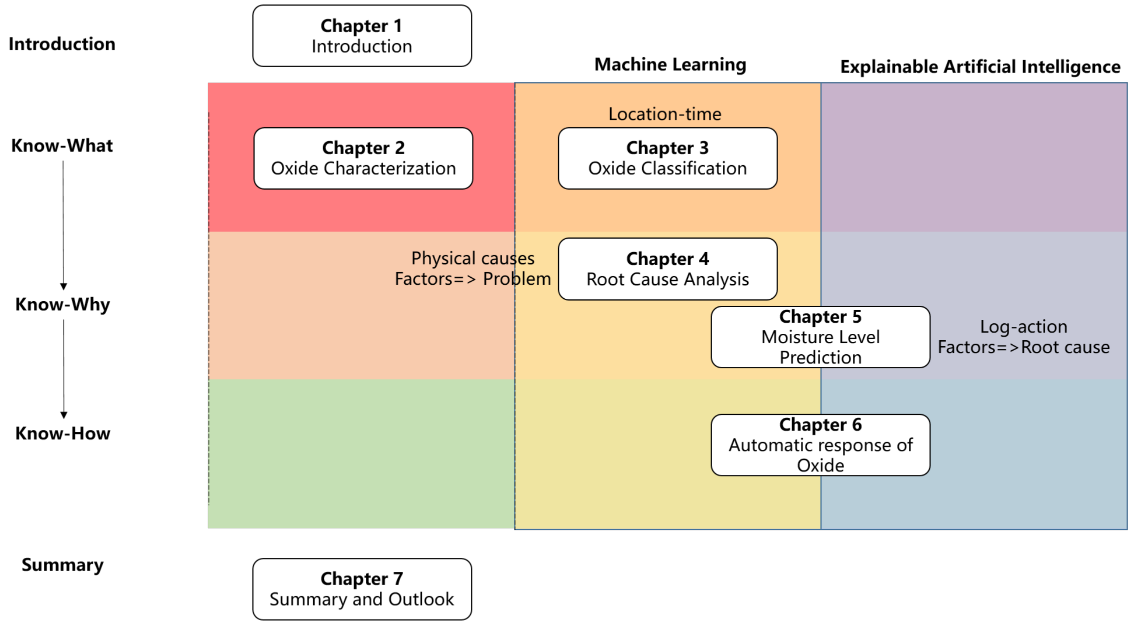


Figure 7.1: The overview of the structure and organization of this thesis.

impacts on the product quality or the associated industrial challenges?

The question was tackled by conducting material characterization and visual characterization on the surface anomaly.

The material characterization suggests that the high contrast observed in the FZ images with surface anomaly is due to oxidation, as well as the dissolution and evaporation of the oxygen. Typically, the dark part near the melt edge in the high-contrast image shown in Figure 2.2, corresponding to the white surface in the polysilicon part, is the silicon itself without observing oxygen. On the contrary, the colored region beyond the white surface on the polysilicon sample is enriched with oxygen atoms on its surface. The presence of its color indicates the thickness of the oxide layer. These results reveal that the oxidation might have occurred in the very early phase. As the oxide layer approaches the heater, part of the oxide layer is dissolved at high temperature and finally evaporates in the form of SiO, resulting in high-contrast appearance in the FZ images.

Visual characterization was performed on FZ images gathered from the FZ vision system. Typically, the surface anomaly is often visually evident in the FZ image from the beginning of the cone phase and gradually fades before moving to the cylinder phase. This might be due to the fact that as the FZ process goes from the cone phase to the cylinder phase, the volume of heated polysilicon increases, thus accelerating the dissolution and evaporation of oxygen. Based on the experimental observation of the FZ process, the surface anomaly can present in three categories, including the spot, shadow, and ghost curtain and their characteristics were discussed.

Though the oxide may disappear before the cylinder phase where the crystal product for sale are sourced, this does not clear the alarm of threatening crystal yield as the presence of an oxide layer indicates the abnormal FZ chamber condition with the introduction of oxygen. On the one hand, the introduction of oxygen may impact product quality by promoting the formation of dislocations within the crystal [73] and resulting in oxygen-related defects. On the other hand, the abnormal ambient indicates that the FZ machine may be deviating from optimal condition and maintenance might be needed. Therefore, it is necessary to optimize both FZ process and FZ machine by investigating around the surface anomaly or the oxide problem.

What are the key factors that influence the formation of the surface anomaly in the FZ process, and the practical actions to mitigate the surface anomaly?

In the pursuit of unraveling the complex dynamics of the oxide problem in the crystal growth process, the root cause analysis of the oxide problem was performed at two different levels, **Factors** ⇒ **Problem**, **Factors** ⇒ **Root cause**.

In terms of the level **Factors** ⇒ **Problem**, data-driven association rule mining was employed to associate the relevant factors with the oxide problem. Specifically, the oxide class is set to be the consequence of the rules. The results showed that the moisture level inside the FZ chamber from the early phases is the main key factor that influences the formation of the oxide layer on polysilicon. It was found that the spot type and the shadow type are strongly associated with a relatively high moisture level from the early phases compared to the normal process without oxide. This key finding confirms that the oxide layer has formed before it appears apparently in FZ images. In addition, the higher water concentration occurring in the spot than in the shadow type explains the greater thickness of the spot oxide, making it harder to fully transform into SiO in the beginning of the cone phase, thus leaving a spot appearance. Other factors that might influence the formation of the oxide include preparation time and oxygen level inside the chamber. However, due to the weak support and confidence of these rules, these would be considered as opportunities for preventing the recurrence of the oxide problem.

Although a high moisture level has been revealed to be a major physical factor that caused the oxide issue, the reasons for the high moisture level present in the FZ chamber remain unclear. On the other hand, taking into account the cost of a moisture sensor, it is not realistic to install one moisture sensor on each FZ machine. Therefore, to gain a complete understanding of the root causes of the high moisture level, and to make it possible to monitor moisture level in an economical way, a root cause analysis at the level of **Factors** ⇒ **Root cause** was conducted by establishing a moisture level predictor based on the potential factors identified previously along with the FZ image. A Multimodal GAF SE-Transformer model was proposed for such a multimodal problem involving image data, table data and time series data to address the problem of heterogeneous data as well as to jointly learn meaningful representation from these different modalities. Following the development of the moisture level predictor, explainability analysis based on XAI was carried out on the predictor, aiming at inspecting what the model has learned, thus providing insight into the causes of the high moisture level.

The experimental results indicated that the proposed model, which was trained with both MSE loss and Huber loss, was superior to other baselines in terms of average performance, with the lowest RMSE error of 0.4642. Nevertheless, the statistical test only demonstrated the effectiveness of the GAF module and the Transformer module, while no significant difference was seen in the incorporation of the SE module and the sequence of the SE module. In terms of explainability analysis, SHAP method, Feature Permutation and attention-based method supported by the proposed model were employed for evaluating the importance of each attribute. It was found that there is a high similarity measure between the feature ranking from the SHAP method and that from the feature permutation, while the feature ranking from the attention-based method appears with relatively similarity measures with the other two methods. The effectiveness of these three methods was examined by the model performance yielded by the top-k features. The findings indicated that SHAP and Feature Permutation techniques were generally successful in pinpointing the most significant features, whereas the features identified by the attention-based approach did not appear to be the most influential. Additionally, SHAP and Feature Permutation reached the baseline model performance using all available features using top-6 features, including Image, Month, SPC_PUMP1TIME, SPC_PUMP2TIME, LeakRateLast and P8 time series attribute. By looking into the association between these features and moisture level, it was found that high moisture levels typically occur in summer and machine preparation under short pump time and high leak rate after the last pump. Therefore, these could be transformed into guidance to lower the moisture level.

How can the knowledge gained be translated into practical guidelines and control strategies to proactively manage the surface anomaly?

Since the mitigation of the oxide problem is not a one-time task but an ongoing process, continuous improvement is still needed for preventing the reoccurrence of the oxide problem. Hence, a framework regarding continuous management of the oxide problem was proposed. The framework embraces the insights from Know-What and Know-Why, which unraveled the characteristics of the oxide, the contributing factors and their interplay, to provide practical strategies for mitigating the oxide problem.

This study proposed a framework that takes into account both diagnostic and prognostic strategies to address the problem. The framework is structured in four stages: Know-What, Know-Why, Know-When, and Know-How. The diagnostic strategy includes oxide classification and moisture level prediction in the Know-What stage. If an anomaly is identified, the root cause analysis in Know-Why is used to determine potential causes. These potential causes can be obtained from the "IF THEN" knowledge base from Association Rule Mining or from the features with high local-explainability scores from the explainability analysis of the moisture level predictor. Recommendations can then be drawn from these potential causes for improving the process. The prognostic strategy, an alternative solution for responding to the oxide problem, relies on the capability of Know-When in forecasting the occurrence of the oxide problem, allowing for preventive measures or identifying improvement actions to decrease the probability of the oxide occurrence. This conceptual framework seeks to offer a structured and methodological approach to discovering problem-solving solutions, while emphasizing the importance of feedback loops for continuous improvement.

7.2 Outlook

This research has yielded a wealth of discoveries and insights that offer exciting prospects for the future of oxygen contamination prevention and FZ process optimization. As we conclude this thesis, we look ahead with eagerness, recognizing several avenues for further investigation and advancement.

- Image extension. The current root cause analysis was based on a single image per data sample. However, this is not sufficient to get a complete view of the polysilicon because the backside of the polysilicon is invisible on the image. For instance, it could happen that the polysilicon we observed on the image is normal without any oxide, but the backside of the polysilicon has a spot or even a ghost curtain. Therefore, with a single image, it is difficult to conclude the surface state of the polysilicon. Further development that needs to be done is to involve more frames into the analysis.

- Enhanced moisture level predictor. First, there is still room for improvement in predictor performance. The high performance of the moisture level predictor can enhance our trust in the important features identified from the explainability analysis. Gathering more data for training is an alternative to improve generalization ability of the model. Besides, future research needs to make progress in integrating comprehensive uncertainty into predictions. At present, the only uncertainty factor taken into account when assessing the model performance is the randomness associated with the model. However, it is also essential to take into account other uncertainty factors that need to be considered, for instance, the errors and variability in the input or label data, and the uncertainty that arises from the assumptions or approximations made in the modeling process. Embracing uncertainty as an integral aspect of the predictions can improve the reliability and robustness of the model outputs.

- Further root cause analysis. Though the research has unveiled that the the oxide problem is associated with high moisture level and high moisture level may be associated with high leak rate and month and pump time. Investigating the underlying reasons for their occurrence could be alternatives that shed light on practical actions to mitigate or even prevent the oxide problem.

- Discover the causal relationship between the features and the moisture level. Uncovering causal relationships requires more than just explainability analysis results, which only show an association between two variables. To determine a cause-effect relationship, a more thorough level of evidence, such as experimental designs and theoretical understanding, is needed [208].

- Industrial implementation. The transition from research to practical application in industrial settings is an integral next step. Industrial implementation is still necessary to validate the usefulness and practicality of the suggested data-driven solutions. For instance, to experimentally validate if the most significant features identified do have an effect on moisture levels or to see whether the suggested actions can help mitigate the oxide issue.

- Deep investigation of the influence of the oxide on the properties. It is still necessary to delve deeper into the influence of the formation and dissolution of the oxide on the FZ process. For instance, an investigation of the impact of the oxide on thermal properties of the polysilicon and the subsequent impact of the melting behavior, or an investigation of the influence of the oxide that appears on the polysilicon on the growing crystal would be an alternative. This can serve as a solid foundation for the development of in-process interventions to dynamically respond to the oxide problem.

CHAPTER A

APPENDIX

Peer-reviewed journal and conference papers and internal technical reports

Following is a list of the publications and technical reports authored or co-authored by the Ph.D. student within the duration of the Ph.D. program.

Paper 1: Chen, T., Sampath, V., May, M. C., Shan, S., Jorg, O. J., Aguilar Martín, J. J., Stamer, F., Fantoni, G., Tosello, G., Calaon, M. (2023). Machine Learning in Manufacturing towards Industry 4.0: From 'For Now' to 'Four-Know'. *Applied Sciences*, 13(3), 1903. https://doi.org/10.3390/app13031903.

Paper 2: Chen, T., Tosello, G., Calaon, M. (2023). Multi-Label Oxide Classification in Float-Zone Silicon Crystal Growth using Transfer Learning and Asymmetric Loss. (submitted to *Journal of Intelligent Manufacturing*)

Paper 3: Chen, T., Tosello, G., Calaon, M. (2023). Closed-Loop Optimization of Ultra-high-quality Float-Zone Single Crystal Silicon Growth using Deep Learning and In-Line Vision System for Automated Surface Anomaly Detection. (Ready for submission)

Paper 4: Chen, T., Tosello, G., Calaon, M. (2023). Multi-Modal Moisture Level Prediction in Float-Zone Silicon Crystal Growth. (Ready for submission)

Paper 5: Chen, T., Tosello, G., Werner, N., Calaon, M. (2022). Anomaly Detection in Float-Zone Crystal Growth of Silicon. Procedia CIRP, 55th CIRP CMS, Lugano, Switzerland, 29 June- 1 July, 107, pp. 1515-1519. https://doi.org/10.1016/j.procir.2022.05.184.

Paper 6: Chen, T., Tosello, G., Calaon, M. Vision based diameter estimation for continuous float-zone silicon crystal growth production. Euspen's 22nd International Conference & Exhibition. 2022, Geneva, Switzerland, 30 May- 3 June, pp. 419-422.

Paper 7: Chen, T., Tosello, G., Conrad-Hansen, L., Calaon. Uncertainty evaluation of diameter measurement in float-zone crystal growth production. Euspen's 23rd International Conference & Exhibition. 2023, Copenhagen, Denmark, 12-16 June, pp. 179-180.

List of Co-Supervised Projects

Following is a list of the projects co-supervised by the Ph.D. student within the duration of the Ph.D. program.

Project 1: Nicolai Skytte Mikkelsen, Detection of Dimsel in Float-Zone Silicon Manufacturing [Bachelor's thesis]. Department of Mechanical Engineering, Technical University of Denmark, 2021, 48 p.

Project 2: Francesco Tumminello, Deep Learning Monitoring of Float-Zone Crystal Growth Silicon Production [Master's thesis]. Department of Mechanical Engineering, Politecnico di Milano, 2022, 79 p.

CHAPTER B

APPENDIX - EQUIPMENT

B.1 Equiment for material characterization of the surface anomaly

Figure B.1: Helios 5 Hydra UX PFIB [218].

Figure B.2: Quanta FEG 200 ESEM [218].

CHAPTER C

APPENDIX - SUPPLEMENTARY RESULTS

C.1 Oxide Identification: Multi-Label Oxide classification

The detailed comparison results of all baselines can seen in Table C.1.

C.2 Root Cause Analysis of the Oxide Problem

The pseudo code of applying FP-Growth for the root cause analysis of the oxide problem can be seen in Algorithm 1.

The generated Top-20 rules for normal, spot and shadow can be seen in Table C.2, C.3 and C.4, respectively.

C.3 Moisture Level Prediction and explainability analysis

Tthe detailed cross-validation results of all baselines and the optimial hyperparameters selected can be seen in Table C.5, C.6, C.7, C.8, C.9, C.10, C.11, C.12, C.13, C.14, C.15, C.16, C.17, C.18.

The comparison of baseline performance with MSE loss and Huber loss can be seen in Table C.19 and Table C.20, respectively.

The Top-10 feature rankings from SHAP, Attention Weights and Feature Permutation can be seen in Table C.21, C.22 and C.23, respectively.

Table C.1: Comparison of different methods. The highest metric and the second highest metric among different methods are denoted in red and orange, respectively.

Model	Method	Pre-trained	Loss	SA	HL	$F1_{mi}$	$F1_{ma}$	mAP
ResNet50	PS	✓	CE	85.86 ± 1.75	94.35 ± 0.70	90.77 ± 1.27	83.20 ± 2.37	-
	BR			84.23 ± 0.71	94.02 ± 0.36	90.16 ± 0.66	67.36 ± 11.06	**93.04** ± 0.26
	Algorithm adaption		BCE-S	70.94 ± 4.16	89.07 ± 1.21	82.23 ± 1.91	66.33 ± 9.64	77.57 ± 5.59
		✓	BCE	86.64 ± 0.36	94.77 ± 0.20	91.35 ± 0.31	82.58 ± 0.21	91.65 ± 1.71
			FL	**87.65** ± 0.62	**95.16** ± 0.27	**92.06** ± 0.43	**86.64** ± 3.04	91.79 ± 0.94
			ASL	**88.73** ± 0.36	**95.70** ± 0.18	**93.13** ± 0.33	**88.40** ± 1.27	**94.12** ± 1.29
Inception V3	PS	✓	CE	85.16 ± 1.41	94.17 ± 0.43	90.37 ± 0.81	**85.31** ± 1.79	-
	BR			85.39 ± 0.71	94.48 ± 0.28	90.90 ± 0.51	80.33 ± 1.55	**89.80** ± 1.71
	Algorithm adaption		BCE-S	67.52 ± 0.97	87.83 ± 0.31	80.30 ± 0.58	74.85 ± 5.16	79.39 ± 3.36
		✓	BCE	**86.25** ± 1.82	**94.64** ± 0.61	**91.15** ± 0.97	81.53 ± 0.94	88.45 ± 3.03
			FL	86.17 ± 0.71	94.61 ± 0.09	91.11 ± 0.17	80.20 ± 4.04	89.45 ± 5.74
			ASL	**88.11** ± 0.40	**95.29** ± 0.16	**92.41** ± 0.29	**88.39** ± 0.25	**92.19** ± 2.51

Algorithm 1 FP-Growth application for root cause analysis of the oxide problem

```
Input: Preprocessed dataset $D \in \mathbb{R}^{N \times M}$. The class-wise support threshold $min\_s_{c}lass$. The absolute confidence threshold $min\_c$. The absolute lift threshold $min\_l$. The maximum length of rules $min\_len$. List of the oxide types $O$.
Output: Association rules $R$ with their global support, confidence and lift.
function FilterDataset($D$, $target\_oxide$, $oxide\_list$)
    Input: dataset $D$, target oxide $target\_oxide$, oxide list $oxide\_list$
    Output: Filtered dataset $D'$
    Initialize $D' \leftarrow D$
    for $o$ in $oxide\_list$ do
        if $o$ = $target\_oxide$ then
            $D' \leftarrow \{r \in D' \mid r.oxide\_o = 1\}$    ▷ extract the data with the presence of target oxide
        else
            $D' \leftarrow D' \setminus \{D'.oxide\_o\}$    ▷ remove the column $oxide\_o$ that is not target oxide
        end if
    end for
    return $D'$
end function
function PruneRules($R$, $measure$)
    Input: Association rules $R$ with $LHS$, $RHS$, $support$, $confidence$, $lift$, the measure us
    Output: Pruned rules $R'$.
    Initialize $R' \leftarrow R$
    for $rule$ in $R$ do
        for subset $LHS'$ of $rule.LHS$ do
            if $lift(LHS' \Rightarrow RHS) > rule.lift$ then
                $R' \leftarrow R' \setminus rule$
            end if
        end for
    end for
    return $R'$
end function
Initialize Association rules $R \leftarrow \emptyset$.
for $o$ in $O$ do
    $D' \leftarrow FilterDataset(D, o, O)$
    $FP\_rules \leftarrow FPGrowth(D', support = min\_s_{c}lass, maxLength = min\_len)$
    for $rule$ in $FP\_rules$ do
        $support \leftarrow support(rule)$ (according to Eq. 4.1)
        $confidence \leftarrow confidence(rule)$ (according to Eq. 4.2)
        $lift \leftarrow lift(rule)$ (according to Eq. 4.3)
        if $lift > min\_landconfidence >$ then
            $R.LHS \leftarrow R.LHS \cup rule.LHS$
            $R.RHS \leftarrow R.RHS \cup rule.RHS$
            $R.S \leftarrow R.S \cup support$
            $R.C \leftarrow R.C \cup confidence$
            $R.L \leftarrow R.L \cup lift$
        end if
    end for
end for
Pruned rules $R' \leftarrow PruneRules(R)$
return $R'$
```

Table C.2: Rules associated with normal type ranked by lift.

LHS	RHS	Support	Confidence	Lift
Preparation_time=High, Oxygen_Phase1_Mean=Low	normal=1	0.04	0.58	11.88
Preparation_time=High, Oxygen_Phase2_Mean=Low	normal=1	0.04	0.56	11.41
Preparation_time=High, Oxygen_Phase3_Mean=Low	normal=1	0.04	0.54	10.97
Preparation_time=High, Oxygen_Phase4_Mean=Low	normal=1	0.04	0.54	10.97
Preparation_time=High, Oxygen_Cone_Start=Low	normal=1	0.04	0.54	10.97

Table C.3: Top-20 Rules associated with spot type ranked by lift.

LHS	RHS	Support	Confidence	Lift
Moisture_Phase1_Mean=High, Moisture_Cone_Start=High	spot=1	0.30	0.97	1.48
Moisture_Phase1_Mean=High, Machine=35	spot=1	0.30	0.97	1.48
Moisture_Cone_Start=High	spot=1	0.32	0.97	1.47
Moisture_Phase3_Mean=High, Moisture_Cone_Start=High	spot=1	0.31	0.97	1.47
Moisture_Phase2_Mean=High, Moisture_Cone_Start=High	spot=1	0.30	0.97	1.47
Moisture_Phase2_Mean=High, Machine=Machine_35	spot=1	0.29	0.97	1.46
Moisture_Phase3_Mean=High, Machine=Machine_35	spot=1	0.28	0.96	1.46
Moisture_Phase2_Mean=High	spot=1	0.32	0.96	1.46
Moisture_Cone_Start=High	spot=1	0.32	0.96	1.46
Moisture_Phase3_Mean=High	spot=1	0.32	0.96	1.46
Moisture_Phase4_Mean=High, Machine=Machine_35	spot=1	0.28	0.96	1.45
Moisture_Phase4_Mean=High	spot=1	0.32	0.95	1.45
Oxygen_Phase1_Mean=High, Machine=Machine_35	spot=1	0.23	0.87	1.32
Oxygen_Cone_Start=High, Machine=Machine_35	spot=1	0.23	0.87	1.32
Oxygen_Phase4_Mean=High, Machine=Machine_35	spot=1	0.22	0.87	1.32
P11_Phase3_Mean=Low, Oxygen_Phase1_Mean=High	spot=1	0.24	0.87	1.32
P11_Phase2_Mean=Low, Oxygen_Phase1_Mean=High	spot=1	0.24	0.87	1.32
P11_Phase1_Mean=Low, Oxygen_Phase1_Mean=High	spot=1	0.24	0.87	1.32
P11_Phase3_Mean=Low, Oxygen_Cone_Start=High	spot=1	0.23	0.87	1.31
P11_Phase3_Mean=Low, Oxygen_Cone_Start=High	spot=1	0.23	0.87	1.31

Table C.4: Rules associated with shadow type ranked by lift.

LHS	RHS	Support	Confidence	Lift
P14_Phase3_Mean=Low, Moisture_Phase1_Mean=Medium	shadow=1	0.31	0.94	1.26
P14_Phase1_Mean=Low, Moisture_Phase1_Mean=Medium	shadow=1	0.31	0.94	1.26
Moisture_Phase1_Mean=Medium	shadow=1	0.31	0.93	1.25
P14_Phase3_Mean=Low, Moisture_Phase2_Mean=Medium	shadow=1	0.30	0.92	1.23
P14_Phase1_Mean=Low, Moisture_Phase2_Mean=Medium	shadow=1	0.30	0.92	1.23
Moisture_Cone_Start=Medium, Moisture_Phase2_Mean=Medium	shadow=1	0.28	0.92	1.22
Moisture_Phase2_Mean=Medium	shadow=1	0.30	0.91	1.22
P14_Phase3_Mean=Low, Moisture_Phase3_Mean=Medium	shadow=1	0.30	0.91	1.22
P14_Phase1_Mean=Low, Moisture_Phase3_Mean=Medium	shadow=1	0.30	0.91	1.22
Moisture_Cone_Start=Medium, Moisture_Phase3_Mean=Medium	shadow=1	0.28	0.91	1.22
Moisture_Phase3_Mean=Medium	shadow=1	0.30	0.91	1.21
Moisture_Cone_Start=Medium, Moisture_Phase4_Mean=Medium	shadow=1	0.29	0.90	1.21
Process=Process_0, P11_Phase4_Mean=Low	shadow=1	0.24	0.90	1.21
Process=Process_0, LeakRateLast=Low	shadow=1	0.23	0.90	1.21
Moisture_Cone_Start=Medium	shadow=1	0.30	0.90	1.20
P14_Phase3_Mean=Low, Moisture_Phase4_Mean=Medium	shadow=1	0.30	0.90	1.20
P14_Phase1_Mean=Low, Moisture_Phase4_Mean=Medium	shadow=1	0.30	0.90	1.2

Table C.5: Cross-validation results of baseline B1 (Multimodal LSTM-SE-Transformer) trained with MSE loss. The optimal hyperparameters with the lowest averaged validation loss are highlighted.

Loss	Seed	Learning rate	Batch size	Averaged validation loss
MSE	0	**0.0001**	**16**	**0.3726**
		0.0001	32	0.3820
		0.001	16	0.4876
		0.001	32	0.5725
		0.01	16	0.9763
		0.01	32	Nan
	42	0.0001	16	0.3931
		0.0001	**32**	**0.3888**
		0.001	16	0.6260
		0.001	32	0.7492
		0.01	16	0.8787
		0.01	32	1.0138
	100	0.0001	16	0.3047
		0.0001	**32**	**0.2957**
		0.001	16	0.5576
		0.001	32	0.5232
		0.01	16	0.8444
		0.01	32	0.8135
	200	0.0001	16	0.3668
		0.0001	**32**	**0.3547**
		0.001	16	0.5290
		0.001	32	0.4434
		0.01	16	1.0237
		0.01	32	0.9237
	300	0.0001	16	0.3646
		0.0001	**32**	**0.3569**
		0.001	16	0.5472
		0.001	32	0.4477
		0.01	16	0.7214
		0.01	32	1.1857

Table C.6: Cross-validation results of baseline B1 (Multimodal LSTM-SE-Transformer) trained with Huber loss. The optimal hyperparameters with the lowest validation loss are highlighted.

Loss	Seed	Learning rate	Batch size	Averaged validation loss
Huber	0	0.0001	16	0.1460
		0.0001	**32**	**0.1395**
		0.001	16	0.2040
		0.001	32	0.1921
		0.01	16	0.3339
		0.01	32	Nan
	42	0.0001	16	0.1560
		0.0001	**32**	**0.1468**
		0.001	16	0.2259
		0.001	32	0.2277
		0.01	16	Nan
		0.01	32	0.3435
	100	**0.0001**	**16**	**0.1244**
		0.0001	32	0.1327
		0.001	16	0.2197
		0.001	32	0.2419
		0.01	16	0.3158
		0.01	32	0.3188
	200	**0.0001**	**16**	**0.1324**
		0.0001	32	0.1575
		0.001	16	0.2252
		0.001	32	0.1566
		0.01	16	0.3480
		0.01	32	0.3286
	300	**0.0001**	**16**	**0.1295**
		0.0001	32	0.1456
		0.001	16	0.1897
		0.001	32	0.2235
		0.01	16	0.3347
		0.01	32	0.2859

Table C.7: Cross-validation results of baseline B2 (Multimodal LSTM-MLP) trained with MSE loss. The optimal hyperparameters with the lowest validation loss are highlighted.

Loss	Seed	Learning rate	Batch size	Averaged validation loss
MSE	0	0.0001	16	0.4270
		0.0001	32	0.4589
		0.001	**16**	**0.3714**
		0.001	32	0.4132
		0.01	16	0.4122
		0.01	32	0.3974
	42	0.0001	16	0.4248
		0.0001	32	0.4418
		0.001	16	Nan
		0.001	**32**	**0.3899**
		0.01	16	Nan
		0.01	32	Nan
	100	0.0001	16	0.4259
		0.0001	32	0.4376
		0.001	**16**	**0.3556**
		0.001	32	0.3699
		0.01	16	Nan
		0.01	32	0.4712
	200	0.0001	16	0.4520
		0.0001	32	0.4461
		0.001	**16**	0.3575
		0.001	32	0.4033
		0.01	16	0.4267
		0.01	32	0.3871
	300	0.0001	16	0.4400
		0.0001	32	0.4684
		0.001	**16**	**0.3625**
		0.001	32	0.3743
		0.01	16	Nan
		0.01	32	Nan

Table C.8: Cross-validation results of baseline B2 (Multimodal LSTM-MLP) trained with Huber loss. The optimal hyperparameters with the lowest validation loss are highlighted.

Loss	Seed	Learning rate	Batch size	Averaged validation loss
Huber	0	0.0001	16	0.1765
		0.0001	32	0.1938
		0.001	**16**	**0.1484**
		0.001	32	0.1569
		0.01	16	0.1533
		0.01	32	Nan
	42	0.0001	16	0.1733
		0.0001	32	0.1877
		0.001	**16**	**0.1507**
		0.001	32	0.1602
		0.01	16	Nan
		0.01	32	0.1642
	100	0.0001	16	0.1719
		0.0001	32	0.1878
		0.001	**16**	**0.1506**
		0.001	32	0.1630
		0.01	16	Nan
		0.01	32	0.1519
	200	0.0001	16	0.1739
		0.0001	32	0.1832
		0.001	**16**	**0.1505**
		0.001	32	0.1580
		0.01	16	0.1605
		0.01	32	Nan
	300	0.0001	16	0.1757
		0.0001	32	0.1847
		0.001	16	0.1568
		0.001	32	0.1677
		0.01	16	Nan
		0.01	**32**	**0.1501**

Table C.9: Cross-validation results of baseline B3 (Multimodal GAF-MLP) trained with MSE loss. The optimal hyperparameters with the lowest validation loss are highlighted.

Loss	Seed	Learning rate	Batch size	Averaged validation loss
MSE	0	0.0001	16	0.4196
		0.0001	32	0.4378
		0.001	**16**	**0.3820**
		0.001	32	0.4139
		0.01	16	0.4181
		0.01	32	0.4305
	42	0.0001	16	0.4257
		0.0001	32	0.4531
		0.001	16	0.3718
		0.001	**32**	**0.3708**
		0.01	16	0.4194
		0.01	32	0.4085
	100	0.0001	16	0.4053
		0.0001	32	0.4172
		0.001	16	0.3759
		0.001	**32**	**0.3639**
		0.01	16	0.4020
		0.01	32	0.4146
	200	0.0001	16	0.4178
		0.0001	32	0.4509
		0.001	16	0.3888
		0.001	**32**	**0.3791**
		0.01	16	0.3903
		0.01	32	0.7115
	300	0.0001	16	0.4245
		0.0001	32	0.4338
		0.001	**16**	**0.3687**
		0.001	32	0.4028
		0.01	16	0.4090
		0.01	32	0.4128

Table C.10: Cross-validation results of baseline B3 (Multimodal GAF-MLP) trained with Huber loss. The optimal hyperparameters with the lowest validation loss are highlighted.

Loss	Seed	Learning rate	Batch size	Averaged validation loss
Huber	0	0.0001	16	0.1754
		0.0001	32	0.1813
		0.001	16	0.1637
		0.001	32	0.1716
		0.01	**16**	**0.1607**
		0.01	32	0.1677
	42	0.0001	16	0.1747
		0.0001	32	0.1855
		0.001	**16**	**0.1561**
		0.001	32	0.1631
		0.01	16	0.1677
		0.01	32	0.1659
	100	0.0001	16	0.1701
		0.0001	32	0.1807
		0.001	16	0.1536
		0.001	**32**	**0.1506**
		0.01	16	0.1715
		0.01	32	0.1708
	200	0.0001	16	0.1731
		0.0001	32	0.1798
		0.001	16	0.1630
		0.001	**32**	**0.1615**
		0.01	16	0.1737
		0.01	32	0.1658
	300	0.0001	16	0.1760
		0.0001	32	0.1902
		0.001	16	0.1725
		0.001	**32**	**0.1597**
		0.01	16	0.1731
		0.01	32	0.1641

Table C.11: Cross-validation results of baseline B4 (Multimodal GAF-SE) trained with MSE loss. The optimal hyperparameters with the lowest validation loss are highlighted.

Loss	Seed	Learning rate	Batch size	Averaged validation loss
MSE	0	**0.0001**	**16**	**0.7661**
		0.0001	32	0.8187
		0.001	16	0.8280
		0.001	32	**0.8447**
		0.01	16	0.8154
		0.01	32	0.8097
	42	0.0001	16	0.8030
		0.0001	**32**	**0.7924**
		0.001	16	0.8153
		0.001	32	0.8646
		0.01	16	1.0149
		0.01	32	0.8957
	100	0.0001	16	0.8223
		0.0001	**32**	**0.7589**
		0.001	16	0.8179
		0.001	32	0.8657
		0.01	16	0.8079
		0.01	32	0.8866
	200	**0.0001**	**16**	**0.7426**
		0.0001	32	0.8377
		0.001	16	0.8250
		0.001	32	0.8234
		0.01	16	0.7933
		0.01	32	0.9330
	300	0.0001	16	0.7853
		0.0001	**32**	**0.7738**
		0.001	16	0.8032
		0.001	32	0.9138
		0.01	16	0.8729
		0.01	32	0.9615

Table C.12: Cross-validation results of baseline B4 (Multimodal GAF-SE) trained with Huber loss. The optimal hyperparameters with the lowest validation loss are highlighted.

Loss	Seed	Learning rate	Batch size	Averaged validation loss
Huber	0	0.0001	16	0.2470
		0.0001	32	0.2257
		0.001	16	0.2468
		0.001	**32**	**0.2580**
		0.01	16	0.2372
		0.01	32	0.2616
	42	**0.0001**	**16**	**0.2389**
		0.0001	32	0.2537
		0.001	16	0.2588
		0.001	32	0.2537
		0.01	16	0.2418
		0.01	32	0.2474
	100	0.0001	16	0.2505
		0.0001	32	0.2579
		0.001	16	0.2416
		0.001	32	0.2393
		0.01	**16**	**0.2362**
		0.01	32	0.2450
	200	0.0001	16	0.2316
		0.0001	32	0.2347
		0.001	16	0.2504
		0.001	**32**	**0.2302**
		0.01	16	0.2490
		0.01	32	0.2644
	300	0.0001	16	0.2446
		0.0001	32	0.2370
		0.001	**16**	**0.2286**
		0.001	32	0.2500
		0.01	16	0.2430
		0.01	32	0.2462

Table C.13: Cross-validation results of baseline B5 (Multimodal GAF-Transformer) trained with MSE loss. The optimal hyperparameters with the lowest validation loss are highlighted.

Loss	Seed	Learning rate	Batch size	Averaged validation loss
MSE	0	**0.0001**	**16**	**0.4102**
		0.0001	32	0.4509
		0.001	16	0.7281
		0.001	32	0.5616
		0.01	16	1.1497
		0.01	32	0.9058
	42	**0.0001**	**16**	**0.3756**
		0.0001	32	0.3914
		0.001	16	0.7405
		0.001	32	0.6656
		0.01	16	0.9909
		0.01	32	1.1648
	100	**0.0001**	**16**	**0.3535**
		0.0001	32	0.3571
		0.001	16	0.8212
		0.001	32	0.8835
		0.01	16	1.0544
		0.01	32	1.2016
	200	**0.0001**	**16**	**0.3787**
		0.0001	32	0.4051
		0.001	16	0.8659
		0.001	32	0.6761
		0.01	16	0.9752
		0.01	32	1.0198
	300	**0.0001**	**16**	**0.4322**
		0.0001	32	0.4354
		0.001	16	0.5501
		0.001	32	0.6484
		0.01	16	1.0994
		0.01	32	1.0199

Table C.14: Cross-validation results of baseline B5 (Multimodal GAF-Transformer) trained with Huber loss. The optimal hyperparameters with the lowest validation loss are highlighted.

Loss	Seed	Learning rate	Batch size	Averaged validation loss
Huber	0	**0.0001**	**16**	**0.1542**
		0.0001	32	0.1736
		0.001	16	0.2331
		0.001	32	0.2793
		0.01	16	0.3348
		0.01	32	0.3069
	42	**0.0001**	**16**	**0.1528**
		0.0001	32	0.1566
		0.001	16	0.2596
		0.001	32	0.2124
		0.01	16	0.3345
		0.01	32	0.3126
	100	0.0001	16	0.1591
		0.0001	**32**	**0.1556**
		0.001	16	0.3285
		0.001	32	0.2955
		0.01	16	0.3285
		0.01	32	0.3692
	200	**0.0001**	**16**	**0.1617**
		0.0001	32	0.1707
		0.001	16	0.3337
		0.001	32	0.2218
		0.01	16	0.3270
		0.01	32	0.3320
	300	0.0001	16	0.1722
		0.0001	**32**	**0.1681**
		0.001	16	0.2838
		0.001	32	0.1877
		0.01	16	0.3279
		0.01	32	0.3209

Table C.15: Cross-validation results of baseline B6 (Multimodal GAF-Transformer-SE) trained with MSE loss. The optimal hyperparameters with the lowest validation loss are highlighted.

Loss	Seed	Learning rate	Batch size	Averaged validation loss
MSE	0	**0.0001**	**16**	**0.4037**
		0.0001	32	0.4937
		0.001	16	0.9143
		0.001	32	0.7764
		0.01	16	0.9580
		0.01	32	0.9582
	42	0.0001	16	0.4491
		0.0001	**32**	**0.3993**
		0.001	16	0.6891
		0.001	32	0.6557
		0.01	16	1.0727
		0.01	32	1.0470
	100	**0.0001**	**16**	**0.3976**
		0.0001	32	0.4080
		0.001	16	0.9623
		0.001	32	0.5898
		0.01	16	1.0305
		0.01	32	1.0447
	200	**0.0001**	**16**	**0.4146**
		0.0001	32	0.4190
		0.001	16	0.8807
		0.001	32	0.7124
		0.01	16	1.0238
		0.01	32	0.9074
	300	0.0001	16	0.3975
		0.0001	**32**	**0.3774**
		0.001	16	0.8500
		0.001	32	0.6942
		0.01	16	1.0768
		0.01	32	1.0330

Table C.16: Cross-validation results of baseline B6 (Multimodal GAF-Transformer-SE) trained with Huber loss. The optimal hyperparameters with the lowest validation loss are highlighted.

Loss	Seed	Learning rate	Batch size	Averaged validation loss
Huber	0	**0.0001**	**16**	**0.1665**
		0.0001	32	0.1766
		0.001	16	0.3032
		0.001	32	0.3176
		0.01	16	0.3282
		0.01	32	0.3223
	42	0.0001	16	0.1644
		0.0001	**32**	**0.1573**
		0.001	16	0.2307
		0.001	32	0.3143
		0.01	16	0.3411
		0.01	32	0.3168
	100	**0.0001**	**16**	**0.1565**
		0.0001	32	0.1651
		0.001	16	0.3894
		0.001	32	0.2500
		0.01	16	0.3273
		0.01	32	0.3223
	200	0.0001	16	0.1757
		0.0001	**32**	**0.1680**
		0.001	16	0.2435
		0.001	32	0.2874
		0.01	16	0.3199
		0.01	32	0.3237
	300	0.0001	16	0.1639
		0.0001	**32**	**0.1635**
		0.001	16	0.2778
		0.001	32	0.2246
		0.01	16	0.3478
		0.01	32	0.3083

Table C.17: Cross-validation results of baseline B7 (Multimodal GAF-SE-Transformer) trained with MSE loss. The optimal hyperparameters with the lowest validation loss are highlighted.

Loss	Seed	Learning rate	Batch size	Averaged validation loss
MSE	0	**0.0001**	**16**	**0.4303**
		0.0001	32	0.4519
		0.001	16	0.5418
		0.001	32	0.6209
		0.01	16	0.9281
		0.01	32	0.9386
	42	0.0001	16	0.4429
		0.0001	**32**	**0.4400**
		0.001	16	0.5219
		0.001	32	0.5100
		0.01	16	0.6659
		0.01	32	0.7808
	100	**0.0001**	**16**	**0.3986**
		0.0001	32	0.4856
		0.001	16	0.6433
		0.001	32	0.6282
		0.01	16	1.0056
		0.01	32	1.0551
	200	0.0001	16	0.4402
		0.0001	**32**	**0.4391**
		0.001	16	0.5570
		0.001	32	0.7938
		0.01	16	1.1114
		0.01	32	1.0405
	300	0.0001	16	0.4122
		0.0001	**32**	**0.4483**
		0.001	16	0.7105
		0.001	32	0.5296
		0.01	16	1.0449
		0.01	32	0.9980

Table C.18: Cross-validation results of baseline B7 (Multimodal GAF-SE-Transformer) trained with Huber loss. The optimal hyperparameters with the lowest validation loss are highlighted.

Loss	Seed	Learning rate	Batch size	Averaged validation loss
Huber	0	**0.0001**	**16**	**0.1763**
		0.0001	32	0.1844
		0.001	16	0.2145
		0.001	32	0.1784
		0.01	16	0.3380
		0.01	32	0.3217
	42	**0.0001**	**16**	**0.1774**
		0.0001	32	0.1811
		0.001	16	0.3409
		0.001	32	0.2858
		0.01	16	0.3135
		0.01	32	0.3556
	100	**0.0001**	**16**	**0.1593**
		0.0001	32	0.1683
		0.001	16	0.2688
		0.001	32	0.2679
		0.01	16	0.3347
		0.01	32	0.3578
	200	0.0001	16	0.1803
		0.0001	32	0.1779
		0.001	16	0.2499
		0.001	**32**	**0.1765**
		0.01	16	0.3297
		0.01	32	0.3281
	300	**0.0001**	**16**	**0.1623**
		0.0001	32	0.1790
		0.001	16	0.2509
		0.001	32	0.2429
		0.01	16	0.3234
		0.01	32	0.3529

Table C.19: Comparison of the performance of all baselines using MSE loss with mean and standard deviation.

Model	Time Series Representation		Fusion Module				RMSE↓
	GAF	LSTM	SE(←)	TransE	SE(→)	MLP	
B1		✓	✓	✓			0.5482 ± 0.0680
B2		✓				✓	0.7709 ± 0.0652
B3	✓					✓	0.6545 ± 0.0402
B4	✓		✓				0.9040 ± 0.0893
B5	✓			✓			0.4858 ± 0.0057
B6	✓			✓	✓		0.5210 ± 0.0342
B7	✓		✓	✓			$\mathbf{0.4642 \pm 0.0213}$

Note: ← and → represent the order the SE module compared with Transformer module. ← denotes that the SE module is before the Transformer module and vice versa. ↓ means that the lower the metric value, the better the performance.

Table C.20: Comparison of the performance of all baselines using Huber loss with mean and standard deviation.

Model	Time Series Representation		Fusion Module				RMSE↓
	GAF	LSTM	SE(←)	TransE	SE(←)	MLP	
B1		✓	✓	✓			0.5307 ± 0.0924
B2		✓				✓	0.7338 ± 0.0643
B3	✓					✓	0.7566 ± 0.0701
B4	✓		✓				0.8621 ± 0.0497
B5	✓			✓			0.4984 ± 0.0210
B6	✓			✓	✓		0.5745 ± 0.0415
B7	✓		✓	✓			$\mathbf{0.4840 \pm 0.0370}$

Note: ← and → represent the order the SE module compared with Transformer module. ← denotes that the SE module is before the Transformer module and vice versa. ↓ means that the lower the metric value, the better the performance.

Table C.21: Top-10 important features ranked by SHAP values.

No.	Feature	Importance Score	Feature type
1	Image	0.1521	Image
2	Month	0.1097	Categorical attribute
3	SPC_PUMP2TIME	0.1035	Numeric attribute
4	LeakRateLast	0.1017	Numeric attribute
5	SPC_PUMP1TIME	0.0855	Numeric attribute
6	P8	0.0726	Time series attribute
7	P4	0.0567	Time series attribute
8	C5	0.0520	Categorical attribute
9	LowestChamberPressure	0.0408	Numeric attribute
10	PreparationTime	0.0371	Numeric attribute

Notes: Some feature names are encoded with indexes since they involve confidential information.

Table C.22: Top-10 important features ranked by Attention Weights from the SE module.

No.	Feature	Importance Score	Feature type
1	P9	0.5750	Time series attribute
2	PolyWeight	0.5645	Numeric attribute
3	LeakRateLast	0.5572	Numeric attribute
4	Process	0.5542	Categorical attribute
5	P14	0.5507	Time series attribute
6	Image	0.5453	Image
7	P12	0.5449	Time series attribute
8	Day	0.5403	Categorical attribute
9	P2	0.5383	Time series attribute
10	SPC_PUMP1TIME	0.5327	Numeric attribute

Notes: Some feature names are encoded with indexes since they involve confidential information.

Table C.23: Top-10 important features ranked by Feature Permutation values.

No.	Feature	Importance Score	Feature type
1	SPC_PUMP1TIME	1.8e-04	Time series attribute
2	LeakRateLast	1.7e-04	Numeric attribute
3	Image	1.3e-04	Image
4	SPC_PUMP2TIME	1.2e-04	Numeric attribute
5	Month	7.3e-05	Categorical attribute
6	P8	4.2e-05	Time series attribute
7	Process	3.7e-05	Categorical attribute
8	C5	3.4e-05	Categorical attribute
9	Day	3.2e-05	Categorical attribute
10	LowestChamberPressure	3.1e-05	Numeric attribute

Notes: Some feature names are encoded with indexes since they involve confidential information.

CHAPTER D

APPENDIX - SUPPLEMENTARY RESEARCH

Apart from investigating the oxide problem, the author also worked on improving the precision of the diameter measurements of the growing crystal to reduce costs. The diameter control of the FZ process is essential for the growth of crystals with good shape and quality [219]. Additionally, keeping the diameter under control limit can help reduce additional waste and reduce scrap [220]. Consequently, it is necessary to measure the crystal diameter with high efficiency and precision in order to obtain homogeneous crystal [221]. Unlike other crystal growth techniques where only indirect measures such as weight-based method or ellipse fitting method [221] are feasible, the crystal in the FZ process is well observed by a vision system, making the optical-based method preferred for diameter measurement [219]. With a vision system pointing at a crystal, the crystal diameter is defined as the distance between two edges of the crystal in the image, as shown in Figure D.1. The process of diameter measurement typically involves two stages. Initially, the camera is calibrated to link 2D image coordinates with 3D world coordinates [220]. Subsequently, image processing techniques are used to detect the edge of the crystal, allowing the diameter to be calculated from the calibration [220]. To ensure reliable estimates of the measured diameter, this research outlines an approach to estimate the uncertainty of diameter measurement in accordance with the Guide to the Expression of Uncertainty in Measurement (GUM) [222].

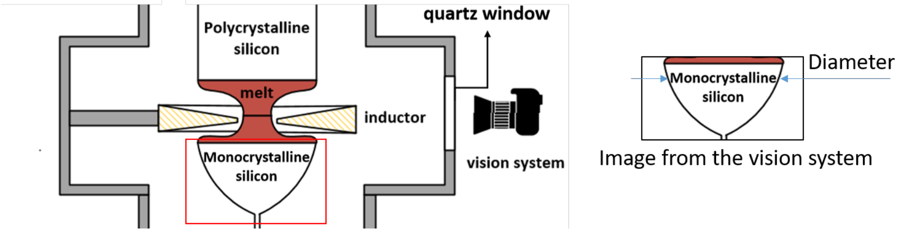


Figure D.1: Diameter measurement on digital images in the FZ process [220].

D.1 Problem Statement

Considering the crystal shape, the measurement of the diameter of the crystal can be simplified as a cylinder measurement [220]. When the optical axis is set to be perpendicular to the cylinder axis, and the focus and aperture of the camera are kept unchanged during the measurement [220], the crystal diameter in the image can be calculated using the pinhole camera model. Due to the property of the pinhole model, the edges of the cylinder are actually generated by the light ray that is tangent to the cylinder, resulting in the incorrect measured diameter $A'B'$ instead of the true diameter, as seen in Figure D.2.

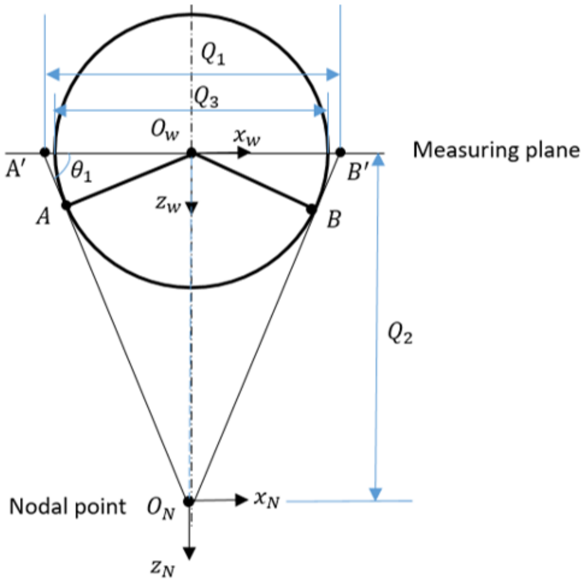


Figure D.2: Incorrect measured diameter due to the property of pinhole camera model [220].

Therefore, a correction based on the trigonometric principle is necessary to transform the measured length into the actual diameter, as seen in Eq D.1. The objective of this research is to measure the diameter of the cylinder with uncertainty evaluation.

$$Diameter\, Q_3 = Q_1 sin(\theta_1) = Q_1 \frac{Q_2}{\sqrt{Q_2^2 + (\frac{Q_1}{2})^2}} \tag{D.1}$$

Where: Q_1 is the distance between two edge points (in 3D world coordinates); Q_2 is the distance between the nodal point of the lens and the origin of world coordinates (in 3D world coordinates).

D.2 Uncertainty Evaluation

The thermal expansion coefficient of silicon is approximately $2.56 \times 10^{-6}K^{-1}$ [223], which is quite significant. Nevertheless, due to the intricate thermal behavior of monocrystalline silicon and the difficulty in measuring the crystal temperature at temperatures higher than 1400 °C, the uncertainty contribution of thermal expansion is not taken into account in this study. Instead, a test bench was set up to simulate the crystal growth in a temperature-controlled room. The measurement procedure starts with calibrating the camera to get the camera-specific parameters. The measuring object is then placed in front of the camera at a fixed distance of 1200 mm. Feature point detection is performed to identify the edge points of the measuring object and the distance between two edge points (Q_1) is calculated. Finally, the diameter is adjusted using Eq D.1. In this study, a test cylinder with a diameter of approximately 203.85 mm (which is similar to the size of a 8 inch silicon ingot) was chosen as the test sample to evaluate the uncertainty. The uncertainty contributors in this study were investigated as follows:

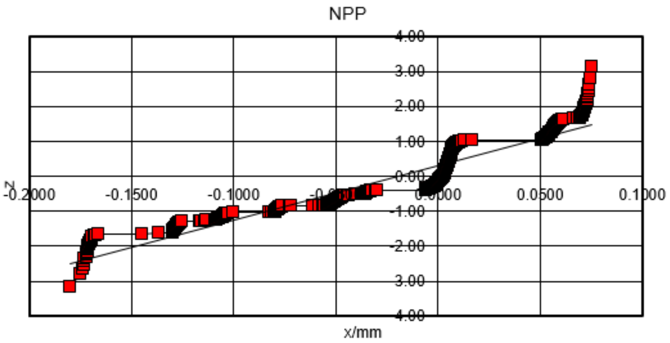


Figure D.3: Normal Probability Plot of experimental data.

D.2.1 Bias of the diameter output from the model (Q_3)

In order to obtain the bias of the model, we measured a multiple cylinder gauge with diameters ranging from 5 mm to 150 mm. The deviation error for each diameter of the cylinder gauge is shown in Figure . The bias of the model is defined as the difference of the averaged 30 measurements from 30 replicated images, against the reference diameter. The χ^2 test on deviation errors gave a value of $\chi^2 = 139.8$ against a confidence interval of 80% from 3.49 to 13.36, which means that the experimental distribution is too different from the normal. The Normal Probability Plot (NPP G) (as seen in Figure D.3) also confirms the presence of systematic effects. Therefore, the presence of a systematic effect is evident, and the null hypothesis must be rejected. There are several reasons could account for the systematic errors:

- Camera model. The pinhole model is a simplified model for single lens, however the industrial camera used is actually composed by several lenses. It is unavoidable to have errors when measuring Q_1 with the simplified linear model.

- Diameter correction model. As seen in Figure D.1, the correction is based on the assumption of a triangle given by the pinhole model. Besides, it can be observed from the correction model that the correction would greatly depend on the diameter. In particular, the corrected diameter will 'shrink' in larger testing diameters. Therefore, the errors of the correction model would apparently reflect on the errors in large diameters, as seen in Figure D.2.

Therefore, regression is used to deduce the pattern of the systematic factor versus the reference diameter. It is assumed that the trend can be represented by a 3-segment function: a linear line model for medium-scale diameters, and two parabolic models for smaller and larger diameters. These regression models were used to adjust the systematic biases, leading to residual errors. The residual errors would be tested by comparing them to the normal distribution to determine if the chosen regression models are acceptable. If the regression models are accepted, then the bias of the test sample can be determined from the regression models. The regression results as well as the residuals after correction are reported in Figure D.4 and Figure D.6.

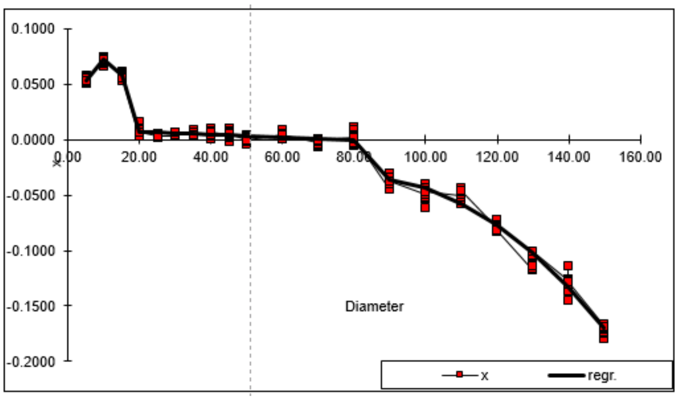


Figure D.4: Regression of the experimental data.

The new test gives a value of $\chi^2 = 11.40$ falling in a confidence interval of 80% from 3.49 to 13.36, which means that the null hypothesis can be accepted. The Normal Probability Plot (NPP G) for the corrected data also confirms this in Figure D.6, which shows that it is alike a straight line.

Finally, given the regression models, a bias of 2.3×10^{-1} mm for the test cylinder with the diameter of 203.85 mm was computed. When a rectangular distribution is taken into account, a bias standard uncertainty contribution of 1.3×10^{-1} mm was obtained for the test cylinder.

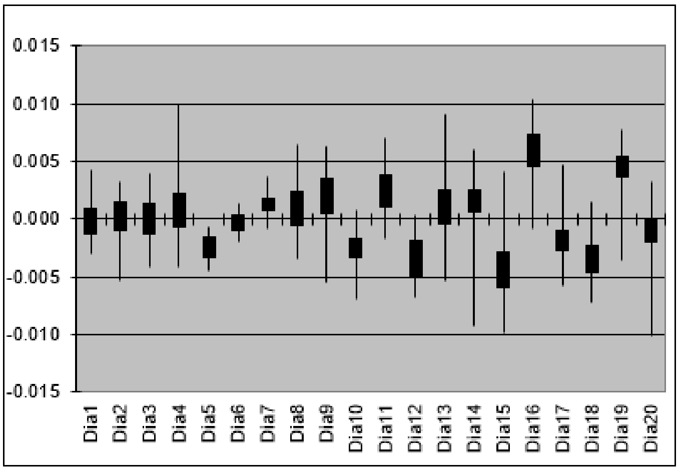


Figure D.5: Residuals of the experimental data using the regression models.

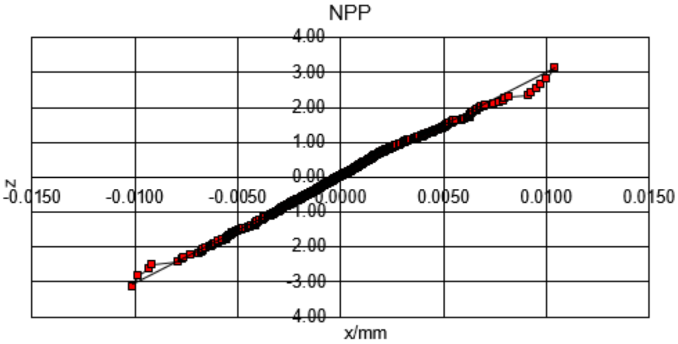


Figure D.6: Normal Probability Plot of the corrected experimental data.

D.2.2 Resolution of edge points detection of the camera (Q_1)

The feature point detection was carried out on a small Region of Interest (ROI) with a height of approximately 34 pixels. The measured distances of pairs of points in ROI were averaged as the final distance of Q_1. Therefore, the resolution was tested by measuring Q_1 at various heights in the ROI. The standard deviation of the 34 pairs of points' distances of 2.7×10^{-4} mm was used as $u_{Q_{1_res}}$.

D.2.3 Reproducibility of Q_1

The reproducibility of the test cylinder is evaluated by taking 30 replicated images and measuring the standard deviation of the Q_1 measurement, which is 6.5×10^{-4} mm, referred to as $u_{Q_{1_repr}}$.

D.2.4 Bias of the measuring distance Q_2

The camera calibration provided a distance with an uncertainty of 1 mm. Considering a rectangular distribution, a bias contribution of 5.8×10^{-1} mm was calculated.

Table D.1 summarizes the contribution of the above uncertainty sources. A standard uncertainty of the test cylinder diameter can be evaluated based on Eq D.2. Finally, an expanded uncertainty U with $k = 2$ was evaluated to be 0.26 mm.

$$u_c = \sqrt{u_{Q^2_{3-bias}} + (\frac{\partial Q_3}{\partial Q_1})^2 (u^2_{Q_{1-res}} + u^2_{Q_{1-repr}}) + (\frac{\partial Q_3}{\partial Q_2})^2 u^2_{Q_{2-repr}}} \tag{D.2}$$

Table D.1: Uncertainty budget of the diameter measurement [220].

Source of uncertainty	Symbol	Standard uncertainty u	Sensitivity coefficient c	c^2u^2
Bias of Q_3	$u_{Q_{3-bias}}$	$1.3 \times 10^{-1} mm$	1	$1.7 \times 10^{-5} mm$
Resolution of Q_1	$u_{Q_{1-res}}$	$2.7 \times 10^{-4} mm$	9.9×10^{-1}	$7.3 \times 10^{-11} mm$
Reproducibility of Q_1	$u_{Q_{1-repr}}$	$6.5 \times 10^{-4} mm$	9.9×10^{-1}	$3.6 \times 10^{-10} mm$
Bias of Q_2	$u_{Q_{2-bias}}$	$5.8 \times 10^{-1} mm$	1.1×10^{-3}	$4.1 \times 10^{-10} mm$
Combined uncertainty	u_c			$1.3 \times 10^{-1} mm$
Expanded uncertainty	$U(k = 2)$			$2.6 \times 10^{-1} mm$

D.3 Conclusion

Diameter measurement and its uncertainty evaluation is highly important in the FZ crystal growth production. This study discussed the uncertainty of such measurements, following the Guide to the Expression of Uncertainty in Measurement (GUM). An experiment was carried out on a test bench of the FZ process to evaluate the uncertainty of measuring a test cylinder with diameter of 203.85 mm, using a multiple cylinder gauge. The uncertainty of the measured diameter was found to be 0.26 mm, without taking into account the uncertainty caused by thermal expansion.

Bibliography

[1] Worldwide Semiconductor Trade Statistics Inc. Historical billings report. https://www.wsts.org/67/Historical-Billings-Report. Accessed August 22, 2023.

[2] Omar Z Sharaf and Mehmet F Orhan. Concentrated photovoltaic thermal (cpvt) solar collector systems: Part i–fundamentals, design considerations and current technologies. *Renewable and Sustainable Energy Reviews*, 50:1500–1565, 2015.

[3] Ying Fu and Ying Fu. Semiconductor materials. *Physical Models of Semiconductor Quantum Devices*, pages 1–66, 2014.

[4] Graham Fisher, Michael R Seacrist, and Robert W Standley. Silicon crystal growth and wafer technologies. *Proceedings of the IEEE*, 100(Special Centennial Issue):1454–1474, 2012.

[5] USJC. How a semiconductor wafer is made. https://www.usjpc.com/en/tech-intro-e/process-e. Accessed Oct 16, 2023.

[6] Solor panel manufacturing process. https://battlebornbatteries.com/how-are-solar-panels-made/. Accessed Oct 16, 2023.

[7] Kuo-tsan Liu and Chia-Ho Chen. Formulation of research and development strategy by analysing patent portfolios of key players the semiconductor industry according to patent strength and technical function. *World Patent Information*, 70:102125, 2022.

[8] Eun Kyo Cho. China's semiconductor strategy and its implications for responding to the us-china technology conflict. *Korea Institute for Industrial Economics and Trade Research Paper No*, 22:1–2, 2022.

[9] Alex Capri. Semiconductors at the heart of the us-china tech war. *Hinrich Foundation*, 22, 2020.

[10] Larisa Kapustina, L'udmila Lipková, Yakov Silin, and Andrei Drevalev. Us-china trade war: Causes and outcomes. In *SHS Web of Conferences*, volume 73, page 01012. EDP Sciences, 2020.

[11] Benjamin Frieske and Sylvia Stieler. The "semiconductor crisis" as a result of the covid-19 pandemic and impacts on the automotive industry and its supply chains. *World Electric Vehicle Journal*, 13(10):189, 2022.

[12] Amir Reza Ansari Dezfoli. Czochralski (cz) process modification with cooling tube in the response to market global silicon shortage. *Journal of Crystal Growth*, 610:127170, 2023.

[13] Wilfried von Ammon. Fz and cz crystal growth: Cost driving factors and new perspectives. *physica status solidi (a)*, 211(11):2461–2470, 2014.

[14] Guilherme Manuel Morais Gaspar, Antoine Autruffe, and José Mário Pó. Silicon growth technologies for pv applications. In *New Research on Silicon-Structure, Properties, Technology*. IntechOpen, 2017.

[15] Topsil GlobalWafers. Preferred float zone (pfz) silicon for power electronics. http://www.topsil.com/media/56273/pfz_application_notelong_version_september_2014.pdf. Accessed August 24, 2023.

[16] Jochen Friedrich, Wilfried von Ammon, and Georg Müller. Czochralski growth of silicon crystals. In *Handbook of Crystal Growth*, pages 45–104. Elsevier, 2015.

[17] Andris Muiznieks, Janis Virbulis, Anke Lüdge, Helge Riemann, and Nico Werner. *Floating Zone Growth of Silicon*, volume 2. Elsevier B.V., second edi edition, 2015.

[18] Kaoru Kajiwara, Kazuhiro Harada, Kazuhisa Torigoe, and Masataka Hourai. Oxygen Precipitation Properties of Nitrogen-Doped Czochralski Silicon Single Crystals with Low Oxygen Concentration. *Physica Status Solidi (A) Applications and Materials Science*, 216(17):1–6, 2019.

[19] Akira Kiyoi, Naoyuki Kawabata, Katsumi Nakamura, and Yasufumi Fujiwara. Influence of oxygen on trap-limited diffusion of hydrogen in proton-irradiated n-type silicon for power devices. *Journal of Applied Physics*, 129(2), 2021.

[20] Masataka Hourai, Toru Nagashima, Hideshi Nishikawa, Wataru Sugimura, Toshiaki Ono, and Shigeru Umeno. Review and Comments for the Development of Point Defect-Controlled CZ-Si Crystals and Their Application to Future Power Devices. *physica status solidi (a)*, 216(10):1800664, 2019.

[21] G Müller, JJ Métois, and P Rudolph. Silicon crystal growth. *Crystal Growth-From Fundamentals to Technology*, page 239, 2004.

[22] Stefano Meroli. Czochralski process vs float zone: two growth techniques for monocrystalline silicon. https://meroli.web.cern.ch/lecture_silicon_floatzone_czochralski.html. Accessed August 30, 2023.

[23] Fumio Shimura. Single-crystal silicon: growth and properties. *Springer Handbook of Electronic and Photonic Materials*, pages 1–1, 2017.

[24] Nico Werner. Analysis and automation of the crucible-free Floating Zone (FZ) growth of silicon crystals, 2014.

[25] Guido Tosello Tingting Chen and Matteo Calaon. Multi-label oxide classification in float-zone silicon crystal growth using transfer learning and asymmetric loss. 2023.

[26] Tingting Chen, Guido Tosello, Nico Werner, and Matteo Calaon. Anomaly detection in float-zone crystal growth of silicon. *Procedia CIRP*, 107:1515–1519, 2022.

[27] W. Zulehner. Czochralski growth of silicon. *Journal of Crystal Growth*, 65(1):189–213, 1983.

[28] Kaoru Kajiwara, Kazuhiro Harada, Kazuhisa Torigoe, and Masataka Hourai. Oxygen precipitation properties of nitrogen-doped czochralski silicon single crystals with low oxygen concentration. *physica status solidi (a)*, 216(17):1900272, 2019.

[29] G Kissinger, Jan Vanhellemont, Eddy Simoen, Cor Claeys, and H Richter. Investigation of oxygen precipitation related crystal defects in processed silicon wafers by infrared light scattering tomography. *Materials Science and Engineering: B*, 36(1-3):225–229, 1996.

[30] N.S. Bajaj and R.A. Joshi. Chapter 3 - energy materials: synthesis and characterization techniques. In S.J. Dhoble, N.Thejo Kalyani, B. Vengadaesvaran, and Abdul Kariem Arof, editors, *Energy Materials*, pages 61–82. Elsevier, 2021.

[31] John D Murphy, RE McGuire, Karsten Bothe, VV Voronkov, and Robert J Falster. Minority carrier lifetime in silicon photovoltaics: The effect of oxygen precipitation. *Solar Energy Materials and Solar Cells*, 120:402–411, 2014.

[32] Edward Lumsdaine and Monika Lumsdaine. Creative problem solving. *IEEE Potentials*, 13(5):4–9, 1994.

[33] Mirko Sokovic, Dusko Pavletic, and K Kern Pipan. Quality improvement methodologies–pdca cycle, radar matrix, dmaic and dfss. *Journal of achievements in materials and manufacturing engineering*, 43(1):476–483, 2010.

[34] Pavol Kaplík, Miroslav Prístavka, Marián Bujna, and Ján Viderňan. Use of 8d method to solve problems. *Advanced Materials Research*, 801:95–101, 2013.

[35] Kepner-Tregoe. What is the kepner-tregoe method? https://kepner-tregoe.com/faqs/. Accessed August 27, 2023.

[36] Alvaro Camarillo, José Ríos, and Klaus-Dieter Althoff. Knowledge-based multi-agent system for manufacturing problem solving process in production plants. *Journal of manufacturing systems*, 47:115–127, 2018.

[37] Carlos A Riesenberger and Sérgio D Sousa. The 8d methodology: an effective way to reduce recurrence of customer complaints. In *Proceedings of the world congress on engineering*, volume 3, 2010.

[38] Arthur Poh. *The Learning Enterprise Innovative Practices for Organisational Transformation by Arthur Poh et al.* 11 2020.

[39] Q Peter He and Jin Wang. Statistical process monitoring as a big data analytics tool for smart manufacturing. *Journal of Process Control*, 67:35–43, 2018.

[40] Rahul Rai, Manoj Kumar Tiwari, Dmitry Ivanov, and Alexandre Dolgui. Machine learning in manufacturing and industry 4.0 applications, 2021.

[41] Jinjiang Wang, Yulin Ma, Laibin Zhang, Robert X Gao, and Dazhong Wu. Deep learning for smart manufacturing: Methods and applications. *Journal of manufacturing systems*, 48:144–156, 2018.

[42] Amine Belhadi, Karim Zkik, Anass Cherrafi, M Yusof Sha'ri, et al. Understanding big data analytics for manufacturing processes: insights from literature review and multiple case studies. *Computers & Industrial Engineering*, 137:106099, 2019.

[43] Tingting Chen, Vignesh Sampath, Marvin Carl May, Shuo Shan, Oliver Jonas Jorg, Juan José Aguilar Martín, Florian Stamer, Gualtiero Fantoni, Guido Tosello, and Matteo Calaon. Machine learning in manufacturing towards industry 4.0: From 'for now'to 'four-know'. *Applied Sciences*, 13(3):1903, 2023.

[44] Yucheng Wang, Liang Gao, Yiping Gao, and Xinyu Li. A new graph-based semi-supervised method for surface defect classification. *Robotics and Computer-Integrated Manufacturing*, 68:102083, 2021.

[45] Lei Yang, Junfeng Fan, Benyan Huo, En Li, and Yanhong Liu. A nondestructive automatic defect detection method with pixelwise segmentation. *Knowledge-Based Systems*, 242:108338, 2022.

[46] Sun Ho Kim, Chan Young Kim, Da Hoon Seol, Jeong Eun Choi, and Sang Jeen Hong. Machine learning-based process-level fault detection and part-level fault classification in semiconductor etch equipment. *IEEE Transactions on Semiconductor Manufacturing*, 35(2):174–185, 2022.

[47] Shenglin Peng et al. Reinforcement learning with gaussian processes for condition-based maintenance. *Computers & Industrial Engineering*, 158:107321, 2021.

[48] Pilsung Kang, Dongil Kim, and Sungzoon Cho. Semi-supervised support vector regression based on self-training with label uncertainty: An application to virtual metrology in semiconductor manufacturing. *Expert Systems with Applications*, 51:85–106, 2016.

[49] Ashok K Srivastava, Pradip K Patra, and Rajesh Jha. Ahss applications in industry 4.0: Determination of optimum processing parameters during coiling process through unsupervised machine learning approach. *Materials Today Communications*, 31:103625, 2022.

[50] Sara Antomarioni, Filippo Emanuele Ciarapica, and Maurizio Bevilacqua. Association rules and social network analysis for supporting failure mode effects and criticality analysis: Framework development and insights from an onshore platform. *Safety science*, 150:105711, 2022.

[51] Ruimin Chen, Yan Lu, Paul Witherell, Timothy W Simpson, Soundar Kumara, and Hui Yang. Ontology-driven learning of bayesian network for causal inference and quality assurance in additive manufacturing. *IEEE Robotics and Automation Letters*, 6(3):6032–6038, 2021.

[52] Sagar Sikder, Indrajit Mukherjee, and Subhash Chandra Panja. A synergistic mahalanobis–taguchi system and support vector regression based predictive multivariate manufacturing process quality control approach. *Journal of Manufacturing Systems*, 57:323–337, 2020.

[53] Tania Cerquitelli, Francesco Ventura, Daniele Apiletti, Elena Baralis, Enrico Macii, and Massimo Poncino. Enhancing manufacturing intelligence through an unsupervised data-driven methodology for cyclic industrial processes. *Expert Systems with Applications*, 182:115269, 2021.

[54] Nikolaos Kolokas, Thanasis Vafeiadis, Dimosthenis Ioannidis, and Dimitrios Tzovaras. A generic fault prognostics algorithm for manufacturing industries using unsupervised machine learning classifiers. *Simulation Modelling Practice and Theory*, 103:102109, 2020.

[55] Dazhong Wu, Connor Jennings, Janis Terpenny, Robert X Gao, and Soundar Kumara. A comparative study on machine learning algorithms for smart manufacturing: tool wear prediction using random forests. *Journal of Manufacturing Science and Engineering*, 139(7):071018, 2017.

[56] Zsolt J Viharos and Richard Jakab. Reinforcement learning for statistical process control in manufacturing. *Measurement*, 182:109616, 2021.

[57] Marcelo Luis Ruiz Rodríguez, Sylvain Kubler, Andrea de Giorgio, Maxime Cordy, Jérémy Robert, and Yves Le Traon. Multi-agent deep reinforcement learning based predictive maintenance on parallel machines. *Robotics and Computer-Integrated Manufacturing*, 78:102406, 2022.

[58] Panagiotis D Paraschos, Georgios K Koulinas, and Dimitrios E Koulouriotis. Reinforcement learning for combined production-maintenance and quality control of a manufacturing system with deterioration failures. *Journal of Manufacturing Systems*, 56:470–483, 2020.

[59] Yi-Hung Liu, Han-Pang Huang, and Yu-Sheng Lin. Dynamic scheduling of flexible manufacturing system using support vector machines. In *IEEE International Conference on Automation Science and Engineering, 2005.*, pages 387–392. IEEE, 2005.

[60] Guanghui Zhou, Zhenghao Chen, Chao Zhang, and Fengtian Chang. An adaptive ensemble deep forest based dynamic scheduling strategy for low carbon flexible job shop under recessive disturbance. *Journal of Cleaner Production*, 337:130541, 2022.

[61] Azmah Hanim Mohamed Ariff, M Saleem J Hashmi, and Dermot Brabazon. Monocrystalline silicon grown using floating zone technique. 2018.

[62] Zhong-Sheng Hou and Zhuo Wang. From model-based control to data-driven control: Survey, classification and perspective. *Information Sciences*, 235:3–35, 2013.

[63] BYU. Silicon dioxide/nitride color vs. film thickness and viewing angle calculator. `https://cleanroom.byu.edu/color_chart`. Accessed September 12th, 2023.

[64] Jihong Yim. Energy-dispersive x-ray spectroscopy. `https://wiki.aalto.fi/display/SSC/Energy-dispersive+X-ray+spectroscopy#cite-summary-5-2`. Accessed September 12th, 2023.

[65] Azad Mohammed and Avin Abdullah. Scanning electron microscopy (sem): A review. In *Proceedings of the 2018 International Conference on Hydraulics and Pneumatics—HERVEX, Băile Govora, Romania*, volume 2018, pages 7–9, 2018.

[66] Hua Younan, Liu Binghai, Mo Zhiqiang, and Jennifer Teong. Studies and applications of standardless edx quantification method in failure analysis of wafer fabrication. In *2008 15th International Symposium on the Physical and Failure Analysis of Integrated Circuits*, pages 1–6. IEEE, 2008.

[67] Nan Nan and Jingxin Wang. Fib-sem three-dimensional tomography for characterization of carbon-based materials. *Advances in Materials Science and Engineering*, 2019, 2019.

[68] MT Lapington, DJ Crudden, RC Reed, MP Moody, and PAJ Bagot. Characterization of oxidation mechanisms in a family of polycrystalline chromia-forming nickel-base superalloys. *Acta Materialia*, 206:116626, 2021.

[69] Lixin Gu, Nian Wang, Xu Tang, HG Changela, et al. Application of fib-sem techniques for the advanced characterization of earth and planetary materials. *Scanning*, 2020, 2020.

[70] M Abd Mutalib, MA Rahman, MHD Othman, AF Ismail, and J Jaafar. Scanning electron microscopy (sem) and energy-dispersive x-ray (edx) spectroscopy. In *Membrane characterization*, pages 161–179. Elsevier, 2017.

[71] Antonis Nanakoudis. Edx analysis with sem: How does it work? https://www.thermofisher.com/blog/materials/edx-analysis-with-sem-how-does-it-work/#:~:text=The%20way%20EDX%20analysis%20works,shell%20to%20fill%20the%20vacancy. Accessed September 12th, 2023.

[72] Joseph I Goldstein, Dale E Newbury, Joseph R Michael, Nicholas WM Ritchie, John Henry J Scott, and David C Joy. *Scanning electron microscopy and X-ray microanalysis*. springer, 2017.

[73] Susanne Richter, Martina Werner, Michael Schley, Friedrich Schaaff, Helge Riemann, Hansjoachim Rost, Frank Zobel, Roland Kunert, Peter Dold, and Christian Hagendorf. Influence of slim rod material properties to the Siemens feed rod and the float zone process. *Energy Procedia*, 55:596–601, 2014.

[74] Adane Nega Tarekegn, Mario Giacobini, and Krzysztof Michalak. A review of methods for imbalanced multi-label classification. *Pattern Recognition*, 118:107965, 2021.

[75] Min Ling Zhang and Zhi Hua Zhou. A review on multi-label learning algorithms. *IEEE Transactions on Knowledge and Data Engineering*, 26(8):1819–1837, 2014.

[76] Hakan Cevikalp, Burak Benligiray, and Omer Nezih Gerek. Semi-supervised robust deep neural networks for multi-label image classification. *Pattern Recognition*, 100:107164, 2020.

[77] Ali Braytee, Wei Liu, Ali Anaissi, and Paul J Kennedy. Correlated multi-label classification with incomplete label space and class imbalance. *ACM Transactions on Intelligent Systems and Technology*, 10(5), 2019.

[78] Matthew R. Boutell, Jiebo Luo, Xipeng Shen, and Christopher M. Brown. Learning multi-label scene classification. *Pattern Recognition*, 37(9):1757–1771, 2004.

[79] Grigorios Tsoumakas, Ioannis Katakis, and Ioannis Vlahavas. Random k-labelsets for multi-label classification. *IEEE Transactions on Knowledge and Data Engineering*, 23(7):1079–1089, 2011.

[80] Min Ling Zhang and Zhi Hua Zhou. ML-KNN: A lazy learning approach to multi-label learning. *Pattern Recognition*, 40(7):2038–2048, 2007.

[81] Waseem Rawat and Zenghui Wang. Deep convolutional neural networks for image classification: A comprehensive review. *Neural computation*, 29(9):2352–2449, 2017.

[82] Dan Claudiu Ciresan, Ueli Meier, Jonathan Masci, Luca Maria Gambardella, and Jürgen Schmidhuber. Flexible, high performance convolutional neural networks for image classification. In *Twenty-second international joint conference on artificial intelligence*. Citeseer, 2011.

[83] Yann LeCun, Bernhard Boser, John Denker, Donnie Henderson, Richard Howard, Wayne Hubbard, and Lawrence Jackel. Handwritten digit recognition with a back-propagation network. *Advances in neural information processing systems*, 2, 1989.

[84] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012.

[85] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[86] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 2016-Decem, pages 770–778, 2016.

[87] Yunchao Wei, Wei Xia, Junshi Huang, Bingbing Ni, Jian Dong, Yao Zhao, and Shuicheng Yan. CNN: Single-label to multi-label. *arXiv preprint arXiv:1406.5726*, 2014.

[88] Hao Guo and Song Wang. Long-tailed multi-label visual recognition by collaborative training on uniform and re-balanced samplings. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15089–15098, 2021.

[89] Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks? *Advances in neural information processing systems*, 27, 2014.

[90] Behnam Neyshabur, Hanie Sedghi, and Chiyuan Zhang. What is being transferred in transfer learning? *Advances in neural information processing systems*, 33:512–523, 2020.

[91] Weifeng Ge, Sibei Yang, and Yizhou Yu. Multi-evidence filtering and fusion for multi-label classification, object detection and semantic segmentation based on weakly supervised learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1277–1286, 2018.

[92] Jiajia Li, Jie Chen, Bin Sheng, Ping Li, Po Yang, David Dagan Feng, and Jun Qi. Automatic Detection and Classification System of Domestic Waste via Multimodel Cascaded Convolutional Neural Network. *IEEE Transactions on Industrial Informatics*, 18(1):163–173, 2022.

[93] Yang Zhang, Moyun Liu, Yang Yang, Yanwen Guo, and Huiming Zhang. A Unified Light Framework for Real-Time Fault Detection of Freight Train Images. *IEEE Transactions on Industrial Informatics*, 17(11):7423–7432, 2021.

[94] Vignesh Sampath, Iñaki Maurtua, Juan José Aguilar Martín, Andoni Rivera, Jorge Molina, and Altor Gutierrez. Attention Guided Multi-Task Learning for Surface defect identification. *IEEE Transactions on Industrial Informatics*, 2023.

[95] Zhao-Min Chen, Xiu-Shen Wei, Peng Wang, and Yanwen Guo. Multi-label image recognition with graph convolutional networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5177–5186, 2019.

[96] Jiang Wang, Yi Yang, Junhua Mao, Zhiheng Huang, Chang Huang, and Wei Xu. CNN-RNN: A Unified Framework for Multi-label Image Classification. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 2016-Decem, pages 2285–2294, 2016.

[97] Ke Zhu and Jianxin Wu. Residual Attention: A Simple but Effective Method for Multi-Label Recognition. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 184–193, 2021.

[98] Tal Ridnik, Emanuel Ben-Baruch, Nadav Zamir, Asaf Noy, Itamar Friedman, Matan Protter, and Lihi Zelnik-Manor. Asymmetric loss for multi-label classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 82–91, 2021.

[99] Peng Cao, Xiaoli Liu, Dazhe Zhao, and Osmar Zaiane. Cost sensitive ranking support vector machine for multi-label data learning. In *Proceedings of the 16th International Conference on Hybrid Intelligent Systems (HIS 2016)*, pages 244–255. Springer, 2017.

[100] Adane Nega Tarekegn, Mario Giacobini, and Krzysztof Michalak. A review of methods for imbalanced multi-label classification. *Pattern Recognition*, 118:107965, 2021.

[101] Muhammad Atif Tahir, Josef Kittler, and Ahmed Bouridane. Multilabel classification using heterogeneous ensemble of multi-label classifiers. *Pattern Recognition Letters*, 33(5):513–523, 2012.

[102] Grigorios Tsoumakas, Ioannis Katakis, and Ioannis Vlahavas. Random k-labelsets for multilabel classification. *IEEE transactions on knowledge and data engineering*, 23(7):1079–1089, 2010.

[103] Tsung Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollar. Focal Loss for Dense Object Detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(2):318–327, 2020.

[104] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. *International Journal of Computer Vision*, 128(2):336–359, 2020.

[105] Elhassan Mohamed, Konstantinos Sirlantzis, and Gareth Howells. A review of visualisation-as-explanation techniques for convolutional neural networks and their evaluation. *Displays*, 73:102239, 2022.

[106] Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I 13*, pages 818–833. Springer, 2014.

[107] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013.

[108] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning Deep Features for Discriminative Localization. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2016-Decem:2921–2929, 2016.

[109] Maksims Ivanovs, Roberts Kadikis, and Kaspars Ozols. Perturbation-based methods for explaining deep neural networks: A survey. *Pattern Recognition Letters*, 150:228–234, 2021.

[110] Marco Ancona, Enea Ceolini, Cengiz Öztireli, and Markus Gross. Towards better understanding of gradient-based attribution methods for deep neural networks. *arXiv preprint arXiv:1711.06104*, 2017.

[111] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436–444, 2015.

[112] Kurt Hornik. Approximation capabilities of multilayer feedforward networks. *Neural networks*, 4(2):251–257, 1991.

[113] Ajay Shrestha and Ausif Mahmood. Review of deep learning algorithms and architectures. *IEEE access*, 7:53040–53065, 2019.

[114] Alfredo Canziani, Adam Paszke, and Eugenio Culurciello. An analysis of deep neural network models for practical applications. *arXiv preprint arXiv:1605.07678*, 2016.

[115] Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 249–256. JMLR Workshop and Conference Proceedings, 2010.

[116] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the Inception Architecture for Computer Vision. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2016-Decem:2818–2826, 2016.

[117] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015.

[118] Leslie N Smith and Nicholay Topin. Super-convergence: Very fast training of neural networks using large learning rates. In *Artificial intelligence and machine learning for multi-domain operations applications*, volume 11006, pages 369–386. SPIE, 2019.

[119] Adam Kubany, Shimon Ben Ishay, Ruben Sacha Ohayon, Armin Shmilovici, Lior Rokach, and Tomer Doitshman. Comparison of state-of-the-art deep learning APIs for image multi-label classification using semantic metrics. *Expert Systems with Applications*, 161:113656, 2020.

[120] Peter Chemweno, Liliane Pintelon, Peter Muchiri, et al. i-rcam: Intelligent expert system for root cause analysis in maintenance decision making. In *2016 IEEE international conference on prognostics and health management (ICPHM)*, pages 1–7. IEEE, 2016.

[121] Luca Liliana. A new model of ishikawa diagram for quality assessment. In *Iop conference series: Materials science and engineering*, volume 161, page 012099. IOP Publishing, 2016.

[122] Samuel Jebaraj Benjamin, M Srikamaladevi Marathamuthu, and Uthiyakumar Murugaiah. The use of 5-whys technique to eliminate oee's speed loss in a manufacturing firm. *Journal of Quality in Maintenance Engineering*, 21(4):419–435, 2015.

[123] Christian Spreafico, Davide Russo, and Caterina Rizzi. A state-of-the-art review of fmea/fmeca including patents. *computer science review*, 25:19–28, 2017.

[124] Eduardo e Oliveira, Vera L Miguéis, and José L Borges. Automatic root cause analysis in manufacturing: an overview & conceptualization. *Journal of Intelligent Manufacturing*, 34(5):2061–2078, 2023.

[125] Marc-André Filz, Jonas Ernst Bernhard Langner, Christoph Herrmann, and Sebastian Thiede. Data-driven failure mode and effect analysis (fmea) to enhance maintenance planning. *Computers in Industry*, 129:103451, 2021.

[126] Rakesh Agrawal, Tomasz Imieliński, and Arun Swami. Mining association rules between sets of items in large databases. In *Proceedings of the 1993 ACM SIGMOD international conference on Management of data*, pages 207–216, 1993.

[127] Filip Karlo Došilović, Mario Brčić, and Nikica Hlupić. Explainable artificial intelligence: A survey. In *2018 41st International convention on information and communication technology, electronics and microelectronics (MIPRO)*, pages 0210–0215. IEEE, 2018.

[128] Jae-Yoon Jung and Donghyun Park. Are AI models explainable, interpretable, and understandable? In *Human-Centered Artificial Intelligence*, pages 3–16. Elsevier, 2022.

[129] Kam Cheong Wong, Kai Zhi Woo, and Kai Hui Woo. Ishikawa diagram. *Quality Improvement in Behavioral Health*, pages 119–132, 2016.

[130] Gerald Gartlehner, Marie-Therese Schultes, Viktoria Titscher, Laura C Morgan, Georgiy V Bobashev, Peyton Williams, and Suzanne L West. User testing of an adaptation of fishbone diagrams to depict results of systematic reviews. *BMC medical research methodology*, 17(1):1–9, 2017.

[131] Marcela X Ribeiro, Agma JM Traina, Caetano Traina, and Paulo M Azevedo-Marques. An association rule-based method to support medical image diagnosis with efficiency. *IEEE transactions on multimedia*, 10(2):277–285, 2008.

[132] George TS Ho, WH Ip, Chun-Ho Wu, and Ying Kei Tse. Using a fuzzy association rule mining approach to identify the financial data association. *Expert Systems with Applications*, 39(10):9054–9063, 2012.

[133] Preeti Paranjape-Voditel and Umesh Deshpande. A stock market portfolio recommender system based on association rule mining. *Applied Soft Computing*, 13(2):1055–1063, 2013.

[134] M Forghani and F Karimipour. Extracting human behavioral patterns by mining geo-social networks. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 40:115–120, 2014.

[135] Iztok Fister Jr, Iztok Fister, Dušan Fister, Vili Podgorelec, and Sancho Salcedo-Sanz. A comprehensive review of visualization methods for association rule mining: Taxonomy, challenges, open problems and future ideas. *arXiv preprint arXiv:2302.12594*, 2023.

[136] Charu C Aggarwal and Charu C Aggarwal. Data preparation. *Data Mining: The Textbook*, pages 27–62, 2015.

[137] Elif Varol Altay and Bilal Alatas. Performance analysis of multi-objective artificial intelligence optimization algorithms in numerical association rule mining. *Journal of Ambient Intelligence and Humanized Computing*, 11:3449–3469, 2020.

[138] Swee Chuan Tan. Improving association rule mining using clustering-based discretization of numerical data. In *2018 International Conference on Intelligent and Innovative Computing Applications (ICONIC)*, pages 1–5. IEEE, 2018.

[139] Junrui Yang and Zhang Feng. An effective algorithm for mining quantitative associations based on subspace clustering. In *2010 International Conference on Networking and Digital Society*, volume 1, pages 175–178. IEEE, 2010.

[140] Keivan Kianmehr, Mohammed Alshalalfa, and Reda Alhajj. Fuzzy clustering-based discretization for gene expression classification. *Knowledge and Information Systems*, 24:441–465, 2010.

[141] Minakshi Kaushik, Rahul Sharma, Sijo Arakkal Peious, Mahtab Shahin, Sadok Ben Yahia, and Dirk Draheim. A systematic assessment of numerical association rule mining methods. *SN Computer Science*, 2(5):348, 2021.

[142] Rakesh Agrawal, Ramakrishnan Srikant, et al. Fast algorithms for mining association rules. In *Proc. 20th int. conf. very large data bases, VLDB*, volume 1215, pages 487–499. Santiago, Chile, 1994.

[143] Mohammed Javeed Zaki. Scalable algorithms for association mining. *IEEE transactions on knowledge and data engineering*, 12(3):372–390, 2000.

[144] Akbar Telikani, Amir H Gandomi, and Asadollah Shahbahrami. A survey of evolutionary computation for association rule mining. *Information Sciences*, 524:318–352, 2020.

[145] Mehrdad Almasi and Mohammad Saniee Abadeh. Rare-pears: A new multi objective evolutionary algorithm to mine rare and non-redundant quantitative association rules. *Knowledge-Based Systems*, 89:366–384, 2015.

[146] Jiawei Han, Jian Pei, and Yiwen Yin. Mining frequent patterns without candidate generation. *ACM sigmod record*, 29(2):1–12, 2000.

[147] Jacinto Mata, JL Alvarez, and JC Riquelme. Mining numeric association rules with genetic algorithms. In *Artificial Neural Nets and Genetic Algorithms: Proceedings of the International Conference in Prague, Czech Republic, 2001*, pages 264–267. Springer, 2001.

[148] Israel Edem Agbehadji, Simon Fong, and Richard Millham. Wolf search algorithm for numeric association rule mining. In *2016 IEEE International Conference on Cloud Computing and Big Data Analysis (ICCCBDA)*, pages 146–151. IEEE, 2016.

[149] Florian Pohlmeyer, Ruben Kins, Frederik Cloppenburg, and Thomas Gries. Interpretable failure risk assessment for continuous production processes based on association rule mining. *Advances in Industrial and Manufacturing Engineering*, 5:100095, 2022.

[150] Roberto J Bayardo, Rakesh Agrawal, and Dimitrios Gunopulos. Constraint-based rule mining in large, dense databases. *Data mining and knowledge discovery*, 4:217–240, 2000.

[151] Tosporn Arreeras, Mikiharu Arimura, Takumi Asada, and Saharat Arreeras. Association rule mining tourist-attractive destinations for the sustainable development of a large tourism area in hokkaido using wi-fi tracking data. *Sustainability*, 11(14):3967, 2019.

[152] Peng Xu, Xiatian Zhu, and David A Clifton. Multimodal learning with transformers: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

[153] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence*, 41(2):423–443, 2018.

[154] Nasir Hayat, Krzysztof J Geras, and Farah E Shamout. Medfuse: Multi-modal fusion with clinical time-series data and chest x-ray images. In *Machine Learning for Healthcare Conference*, pages 479–503. PMLR, 2022.

[155] Yidong Chai, Yonghang Zhou, Weifeng Li, and Yuanchun Jiang. An explainable multi-modal hierarchical attention model for developing phishing threat intelligence. *IEEE Transactions on Dependable and Secure Computing*, 19(2):790–803, 2021.

[156] Shaker El-Sappagh, Tamer Abuhmed, SM Riazul Islam, and Kyung Sup Kwak. Multimodal multitask deep learning model for alzheimer's disease progression detection based on time series data. *Neurocomputing*, 412:197–215, 2020.

[157] Chuyi Li, Lulu Li, Yifei Geng, Hongliang Jiang, Meng Cheng, Bo Zhang, Zaidan Ke, Xiaoming Xu, and Xiangxiang Chu. Yolov6 v3. 0: A full-scale reloading. *arXiv preprint arXiv:2301.05586*, 2023.

[158] Yu Zhang, James Qin, Daniel S Park, Wei Han, Chung-Cheng Chiu, Ruoming Pang, Quoc V Le, and Yonghui Wu. Pushing the limits of semi-supervised learning for automatic speech recognition. *arXiv preprint arXiv:2010.10504*, 2020.

[159] M Chandrakala and P Durga Devi. Two-stage classifier for face recognition using hog features. *Materials Today: Proceedings*, 47:5771–5775, 2021.

[160] Peizhong Liu, Jing-Ming Guo, Kosin Chamnongthai, and Heri Prasetyo. Fusion of color histogram and lbp-based features for texture image retrieval and classification. *Information Sciences*, 390:95–111, 2017.

[161] G Lowe. Sift-the scale invariant feature transform. *Int. J*, 2(91-110):2, 2004.

[162] Ling Yang and Shenda Hong. Unsupervised time-series representation learning with iterative bilinear temporal-spectral fusion. In *International Conference on Machine Learning*, pages 25038–25054. PMLR, 2022.

[163] Mehak Khan, Hongzhi Wang, Adnan Riaz, Aya Elfatyany, and Sajida Karim. Bidirectional lstm-rnn-based hybrid deep learning frameworks for univariate time series classification. *The Journal of Supercomputing*, 77:7021–7045, 2021.

[164] Chaur-Heh Hsieh, Yan-Shuo Li, Bor-Jiunn Hwang, and Ching-Hua Hsiao. Detection of atrial fibrillation using 1d convolutional neural network. *Sensors*, 20(7):2136, 2020.

[165] Tingting Chen, Xueping Liu, Bizhong Xia, Wei Wang, and Yongzhi Lai. Unsupervised anomaly detection of industrial robots using sliding-window convolutional variational autoencoder. *IEEE Access*, 8:47072–47081, 2020.

[166] Razvan Pascanu, Tomas Mikolov, and Yoshua Bengio. On the difficulty of training recurrent neural networks. In *International conference on machine learning*, pages 1310–1318. Pmlr, 2013.

[167] Dandan Peng, Huan Wang, Zhiliang Liu, Wei Zhang, Ming J Zuo, and Jian Chen. Multibranch and multiscale cnn for fault diagnosis of wheelset bearings under strong noise and variable load condition. *IEEE Transactions on Industrial Informatics*, 16(7):4949–4960, 2020.

[168] Kahiomba Sonia Kiangala and Zenghui Wang. An effective predictive maintenance framework for conveyor motors using dual time-series imaging and convolutional neural network in an industry 4.0 environment. *Ieee Access*, 8:121033–121049, 2020.

[169] Mengjiao Wang, Wenjie Wang, Xinan Zhang, and Herbert Ho-Ching Iu. A new fault diagnosis of rolling bearing based on markov transition field and cnn. *Entropy*, 24(6):751, 2022.

[170] Chao Zhang, Zichao Yang, Xiaodong He, and Li Deng. Multimodal intelligence: Representation learning, information fusion, and applications. *IEEE Journal of Selected Topics in Signal Processing*, 14(3):478–493, 2020.

[171] Sidney K D'mello and Jacqueline Kory. A review and meta-analysis of multimodal affect detection systems. *ACM computing surveys (CSUR)*, 47(3):1–36, 2015.

[172] Dhanesh Ramachandram and Graham W Taylor. Deep multimodal learning: A survey on recent advances and trends. *IEEE signal processing magazine*, 34(6):96–108, 2017.

[173] Xiangbo Shu, Jiawen Yang, Rui Yan, and Yan Song. Expansion-squeeze-excitation fusion network for elderly activity recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(8):5281–5292, 2022.

[174] Zhaoyang Niu, Guoqiang Zhong, and Hui Yu. A review on the attention mechanism of deep learning. *Neurocomputing*, 452:48–62, 2021.

[175] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018.

[176] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021.

[177] Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. Videobert: A joint model for video and language representation learning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7464–7473, 2019.

[178] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-López, Daniel Molina, Richard Benjamins, et al. Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information fusion*, 58:82–115, 2020.

[179] Bryce Goodman and Seth Flaxman. European union regulations on algorithmic decision-making and a “right to explanation”. *AI magazine*, 38(3):50–57, 2017.

[180] David Gunning. Explainable artificial intelligence (XAI). *Defense advanced research projects agency (DARPA), nd Web*, 2(2):1, 2017.

[181] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. ” why should i trust you?” explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144, 2016.

[182] Amina Adadi and Mohammed Berrada. Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). *IEEE access*, 6:52138–52160, 2018.

[183] Thomas Rojat, Raphaël Puget, David Filliat, Javier Del Ser, Rodolphe Gelin, and Natalia Díaz-Rodríguez. Explainable artificial intelligence (XAI) on timeseries data: A survey. *arXiv preprint arXiv:2104.00950*, 2021.

[184] Sarthak Jain and Byron C Wallace. Attention is not explanation. *arXiv preprint arXiv:1902.10186*, 2019.

[185] Sarah Wiegreffe and Yuval Pinter. Attention is not not explanation. *arXiv preprint arXiv:1908.04626*, 2019.

[186] Mengnan Du, Ninghao Liu, and Xia Hu. Techniques for interpretable machine learning. *Communications of the ACM*, 63(1):68–77, 2019.

[187] Christoph Molnar. *Interpretable machine learning*. Lulu. com, 2020.

[188] Aaron Fisher, Cynthia Rudin, and Francesca Dominici. All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. *J. Mach. Learn. Res.*, 20(177):1–81, 2019.

[189] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017.

[190] Brandon M Greenwell, Bradley C Boehmke, and Andrew J McCarthy. A simple and effective model-based variable importance measure. *arXiv preprint arXiv:1805.04755*, 2018.

[191] Grégoire Montavon. Gradient-based vs. propagation-based explanations: An axiomatic comparison. *Explainable AI: Interpreting, explaining and visualizing deep learning*, pages 253–265, 2019.

[192] Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one*, 10(7):e0130140, 2015.

[193] Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. Learning important features through propagating activation differences. In *International conference on machine learning*, pages 3145–3153. PMLR, 2017.

[194] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *International conference on machine learning*, pages 3319–3328. PMLR, 2017.

[195] K Palani Thanaraj, B Parvathavarthini, U John Tanik, V Rajinikanth, Seifedine Kadry, and Krishnamurthy Kamalanand. Implementation of deep neural networks to classify eeg signals using gramian angular summation field for epilepsy diagnosis. *arXiv preprint arXiv:2003.04534*, 2020.

[196] Zhiguang Wang and Tim Oates. Imaging time-series to improve classification and imputation. *arXiv preprint arXiv:1506.00327*, 2015.

[197] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[198] Qi Wang, Yue Ma, Kun Zhao, and Yingjie Tian. A comprehensive survey of loss functions in machine learning. *Annals of Data Science*, pages 1–26, 2020.

[199] Peter J Huber. Robust estimation of a location parameter. In *Breakthroughs in statistics: Methodology and distribution*, pages 492–518. Springer, 1992.

[200] Yury Gorishniy, Ivan Rubachev, and Artem Babenko. On embeddings for numerical features in tabular deep learning. *Advances in Neural Information Processing Systems*, 35:24991–25004, 2022.

[201] Gowthami Somepalli, Micah Goldblum, Avi Schwarzschild, C Bayan Bruss, and Tom Goldstein. Saint: Improved neural networks for tabular data via row attention and contrastive pre-training. *arXiv preprint arXiv:2106.01342*, 2021.

[202] Vijay Prakash Dwivedi and Xavier Bresson. A generalization of transformer networks to graphs. *arXiv preprint arXiv:2012.09699*, 2020.

[203] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[204] Lloyd S Shapley et al. A value for n-person games. 1953.

[205] Lutz Prechelt. Early stopping-but when? In *Neural Networks: Tricks of the trade*, pages 55–69. Springer, 2002.

[206] Matej Petković, Blaž Škrlj, Dragi Kocev, and Nikola Simidjievski. Fuzzy jaccard index: a robust comparison of ordered lists. *Applied Soft Computing*, 113:107849, 2021.

[207] Jialiang Cui, Qianwen Zhong, Shubin Zheng, Lele Peng, and Jing Wen. A lightweight model for bearing fault diagnosis based on gramian angular field and coordinate attention. *Machines*, 10(4):282, 2022.

[208] Naomi Altman and Martin Krzywinski. Points of significance: Association, correlation and causation. *Nature methods*, 12(10), 2015.

[209] Venkat Venkatasubramanian, Raghunathan Rengaswamy, Kewen Yin, and Surya N Kavuri. A review of process fault detection and diagnosis: Part i: Quantitative model-based methods. *Computers & chemical engineering*, 27(3):293–311, 2003.

[210] Adriana Ito, Malin Hagström, Jon Bokrantz, Anders Skoogh, Mario Nawcki, Kanika Gandhi, Dag Bergsjö, and Maja Bärring. Improved root cause analysis supporting resilient production systems. *Journal of Manufacturing Systems*, 64:468–478, 2022.

[211] Hu-Chen Liu, Jian-Xin You, Ping Li, and Qiang Su. Failure mode and effect analysis under uncertainty: An integrated multiple criteria decision making approach. *IEEE Transactions on Reliability*, 65(3):1380–1392, 2016.

[212] Yunosuke Maki and Kenneth A Loparo. A neural-network approach to fault detection and diagnosis in industrial processes. *IEEE Transactions on Control Systems Technology*, 5(6):529–541, 1997.

[213] Pangun Park, Piergiuseppe Di Marco, Hyejeon Shin, and Junseong Bang. Fault detection and diagnosis using combined autoencoder and long short-term memory network. *Sensors*, 19(21):4612, 2019.

[214] Alberto Diez-Olivan, Javier Del Ser, Diego Galar, and Basilio Sierra. Data fusion and machine learning for industrial prognosis: Trends and perspectives towards industry 4.0. *Information Fusion*, 50:92–111, 2019.

[215] Parand Akbari, Francis Ogoke, Ning-Yu Kao, Kazem Meidani, Chun-Yu Yeh, William Lee, and Amir Barati Farimani. Meltpoolnet: Melt pool characteristic prediction in metal additive manufacturing using machine learning. *Additive Manufacturing*, 55:102817, 2022.

[216] Xingwei Xu, Xiang Li, Weiwei Ming, and Ming Chen. A novel multi-scale cnn and attention mechanism method with multi-sensor signal for remaining useful life prediction. *Computers & Industrial Engineering*, 169:108204, 2022.

[217] Gregory W Vogl, Brian A Weiss, and Moneer Helu. A review of diagnostic and prognostic capabilities and best practices for manufacturing. *Journal of Intelligent Manufacturing*, 30:79–95, 2019.

[218] Thermo Scientific. Helios 5 laser pfib. `https://www.thermofisher.com/order/catalog/product/HELIOS-5-LASER-PFIB`. Accessed September 12th, 2023.

[219] Tingting Chen, Guido Tosello, Nico Werner, and Matteo Calaon. Vision based diameter estimation for continuous float-zone silicon crystal growth production. In *22nd International Conference of the European Society for Precision Engineering and Nanotechnology (euspen 22)*, pages 419–422. euspen, 2022.

[220] Tingting Chen, Guido Tosello, Lars Conrad-Hansen, and Matteo Calaon. Uncertainty evaluation of diameter measurement in float-zone crystal growth production. In *23nd International Conference of the European Society for Precision Engineering and Nanotechnology (euspen 22)*. euspen, 2023.

[221] Xuanyin Wang, Senwei Xiang, Ke Xiang, and Feng Pan. A novel method for diameter measurement of silicon single crystal. *Measurement*, 121:286–293, 2018.

[222] IEC BIPM, ILAC IFCC, IUPAC ISO, and OIML IUPAP. Guide to the expression of uncertainty in measurement jcgm 100: 2008. *JCGM*, 101:2008, 2008.

[223] M Okaji. Absolute thermal expansion measurements of single-crystal silicon in the range 300–1300 k with an interferometric dilatometer. *International Journal of Thermophysics*, 9:1101–1109, 1988.

DTU Construct

Section of Manufacturing Engineering
Koppels Allé 404
2800 Kongens Lyngby

www.construct.dtu.dk

ISBN: 978-87-7475-782-5