OPTIMIZING SEMICONDUCTOR YIELD PREDICTION: A COMPARATIVE STUDY OF TRADITIONAL MACHINE LEARNING APPROACHES

MUHAMMAD IQBAL BIN NAZERI

BACHELOR OF SCIENCE WITH HONORS IN INDUSTRIAL PHYSICS

FAKULTI SAINS DAN SUMBER ALAM

UNIVERSITI MALAYSIA SABAH

2026

UNIVERSITI MALAYSIA SABAH

BORANG PENGESAHAN STATUS TESIS

Judul Tesis	OPTIMIZING SEMICONDUCTOR YIELD PREDICTION: A COMPARATIVE STUDY OF TRADITIONAL MACHINE LEARNING APPROACHES

Saya MUHAMMAD IQBAL BIN NAZERI

mengaku membenarkan laporan ini disimpan di Perpustakaan Universiti Malaysia Sabah dengan syarat-syarat kegunaan seperti berikut:

1. Tesis ini adalah hak milik Universiti Malaysia Sabah.

2. Perpustakaan Universiti Malaysia Sabah dibenarkan membuat salinan untuk tujuan pengajian sahaja.

3. Perpustakaan dibenarkan membuat salinan tesis ini sebagai bahan pertukaran antara institusi pengajian tinggi.

4. Sila tandakan (/)

☐	Sulit	Mengandungi maklumat yang berdarjah keselamatan atau kepentingan Malaysia seperti yang termaktub di dalam AKTA RAHSIA 1972
☐	Terhad	Mengandungi maklumat TERHAD yang telah ditentukan oleh organisasi/badan di mana penyelidikan dijalankan

☐ Tidak Terhad

Disahkan Oleh:

MUHAMMAD IQBAL BIN NAZERI

Matric No: BS22110042

(Tandatangan Pustakawan)

Tarikh: 16/2/2026

(DR. RODNEY PETRUS BALANDONG)
Penyelia Utama

DECLARATION OF THESIS

Title of thesis	OPTIMIZING SEMICONDUCTOR YIELD PREDICTION: A COMPARATIVE STUDY OF TRADITIONAL MACHINE LEARNING APPROACHES

I MUHAMMAD IQBAL BIN NAZERI

I hereby declare that the content of this thesis is my own work, except for quotations, equations, summaries, and references, which have been appropriately acknowledged. Additionally, if Generative Artificial Intelligence has been used, its application and purpose have been transparently disclosed in the disclosure section below

I acknowledge that during the preparation of this thesis, I have utilized Large Language Models (LLMs) as a supplementary tool to assist in various aspects of my research and writing process. Specifically, I have used LLMs for paraphrasing and refining my writing to improve clarity, coherence, and readability while ensuring the original meaning remains intact. Additionally, I have leveraged LLMs to generate general ideas and structure my thoughts more effectively, using them as a brainstorming aid rather than as a replacement for original critical thinking. LLMs have also been helpful in summarizing complex topics, identifying key points from lengthy articles, and suggesting alternative ways to present information. Furthermore, I have used LLMs to check grammar and syntax, ensuring that my writing adheres to academic standards. In some instances, LLMs provided guidance on citation formats and referencing best practices, although all sources have been manually verified to maintain academic integrity. Importantly, I have critically evaluated all AI-generated content and ensured that it aligns with my own understanding and research findings.

Monday 16th February, 2026

MUHAMMAD IQBAL BIN NAZERI
Matric No: BS22110042

ABSTRACT

Accurate semiconductor yield screening is difficult because manufacturing measurements are high dimensional, noisy, and heavily imbalanced toward non-failure outcomes. This thesis proposes a leakage-resistant machine learning pipeline for yield prediction on the SECOM dataset, integrating data cleaning, missing-value handling, scaling, class-imbalance treatment, and feature selection within a consistent evaluation protocol. Two feature selection strategies, Recursive Feature Elimination (RFE) and mRMR, are compared, and traditional classifiers including Support Vector Machines, Random Forest, XGBoost, Decision Tree, and a linear baseline using Linear Regression are benchmarked using a locked hold-out test set. Macro F1-score is used as the primary metric to reflect minority-class detection under class imbalance. The results indicate that RFE with 50 selected variables provides more reliable screening performance than mRMR at the same dimensionality. While several models achieve identical high test accuracy (93.42%), their Macro F1-scores remain low, confirming that accuracy alone can be misleading for yield screening. The Decision Tree model achieves the highest Macro F1-score (0.1311) with lower accuracy (88.75%), highlighting the trade-off between overall correctness and failure-case recovery. These findings emphasize the importance of imbalance-aware evaluation and compact, leakage-safe pipelines when selecting models for practical deployment in semiconductor manufacturing.

Keywords: semiconductor manufacturing; yield prediction; imbalanced classification; feature selection; Recursive Feature Elimination; SECOM dataset; Macro F1-score.

ABSTRAK

Saringan hasil pengeluaran semikonduktor yang tepat adalah sukar kerana ukuran pembuatan berdimensi tinggi, bising, dan sangat tidak seimbang ke arah sampel tidak gagal. Tesis ini mencadangkan satu saluran paip pembelajaran mesin yang tahan kebocoran data untuk ramalan hasil menggunakan dataset SECOM, dengan menggabungkan pembersihan data, pengendalian nilai hilang, penskalaan, rawatan ketidakseimbangan kelas, dan pemilihan ciri dalam protokol penilaian yang konsisten. Dua strategi pemilihan ciri, Recursive Feature Elimination (RFE) dan mRMR, dibandingkan, dan pengelas tradisional termasuk Support Vector Machines, Random Forest, XGBoost, Decision Tree, serta garis asas linear menggunakan Linear Regression ditanda aras menggunakan set ujian terkunci. Macro F1-score digunakan sebagai metrik utama untuk mencerminkan pengesanan kelas minoriti di bawah ketidakseimbangan kelas. Keputusan menunjukkan bahawa RFE dengan 50 pemboleh ubah terpilih memberikan prestasi saringan yang lebih boleh dipercayai berbanding mRMR pada dimensi yang sama. Walaupun beberapa model mencapai ketepatan ujian yang tinggi dan sama (93.42%), nilai Macro F1-score masih rendah, mengesahkan bahawa ketepatan sahaja boleh mengelirukan bagi saringan hasil. Model Decision Tree memperoleh Macro F1-score tertinggi (0.1311) dengan ketepatan yang lebih rendah (88.75%), sekali gus menunjukkan pertukaran antara ketepatan keseluruhan dan pemulihan kes kegagalan. Dapatan ini menekankan kepentingan penilaian yang mengambil kira ketidakseimbangan kelas serta saluran paip yang padat dan selamat daripada kebocoran data apabila memilih model untuk penggunaan sebenar dalam pembuatan semikonduktor.

Kata Kunci: pembuatan semikonduktor; ramalan hasil; pengelasan tidak seimbang; pemilihan ciri; Recursive Feature Elimination; dataset SECOM; Macro F1-score.

ACKNOWLEDGMENT

I would like to express my sincere gratitude to my supervisor, Dr. Rodney Petrus Balandong, for his invaluable guidance, constructive feedback, and consistent support throughout the completion of this thesis. His expertise and encouragement were essential in shaping the direction of this research.

I also extend my appreciation to my examiner, Madam Fouziah, for her time, careful evaluation, and constructive comments, which helped to improve the quality of this work.

I am grateful to the lecturers and staff of the Faculty of Science and Technology, Universiti Malaysia Sabah, for providing the knowledge, resources, and a conducive environment that supported my academic development.

I would also like to acknowledge the researchers and institutions whose published works provided an important foundation for this study and inspired further exploration in semiconductor manufacturing and machine learning.

On a personal note, I am deeply thankful to my family for their continuous encouragement, patience, and unconditional support, especially my father, Nazeri bin Ribut; my mother, Syuhada bini Abdul Rahman; and my two brothers, Muhammad Ikhwan bin Nazeri and Muhammad Irfan Shah bin Nazeri.

I am especially grateful to my partner, Sofea Nasha binti Mohamad Khalid, for her steadfast support throughout this journey. Her patience, understanding, and consistent encouragement, particularly during challenging periods, helped me remain focused and motivated to complete this thesis.

Alhamdulillah, with His blessings, this work has been completed.

TABLE OF CONTENTS

DECLARATION OF THESIS vii
ABSTRACT viii
ABSTRAK ix
ACKNOWLEDGMENT x
LIST OF FIGURES xiii
LIST OF TABLES xv
LIST OF ABBREVIATIONS xvii
CHAPTER 1 INTRODUCTION 1
1.1 Background 1
1.2 Importance and Relevance of Yield Analysis 2
1.2.1 Local Perspective 3
1.2.2 Market Size and Investment Trends in Semiconductor Yield Analytics 3
1.2.3 Economic Impact of Yield-Loss Events and Manufacturing Failures 4
1.2.4 Government Funding and Policies Supporting Smart Manufacturing Analytics 5
1.2.5 Industry Initiatives and Data-Driven Yield Monitoring Programs 6
1.3 Existing Approaches and Challenges 6
1.3.1 Limitations of Traditional Methods 7
1.3.2 Challenges with Automated Methods 9
1.4 Problem Formulation 10
1.4.1 Research Gap 11
1.4.2 Problem Statement 12
1.5 Study Objectives 13

1.6 Scope of Work 13
1.7 Organization of the Thesis 14
CHAPTER 2 LITERATURE REVIEW 16
2.1 Machine Learning in Semiconductor Yield Prediction and Yield Analytics 16
2.2 Overview of Traditional Machine Learning Algorithms in the Proposed Yield Prediction Workflow 17
2.2.1 Support Vector Machine (SVM) 18
2.2.2 Random Forest (RF) 19
2.2.3 Linear Regression (LR) 21
2.2.4 Extreme Gradient Boosting (XGBoost) 22
2.2.5 Decision Trees 25
2.3 Feature Engineering in Semiconductor Yield Prediction and Analytics 26
2.3.1 Common Feature Types in Semiconductor Yield Analysis 28
2.3.2 Feature Selection Approaches in Semiconductor Yield Analysis 28
2.4 Dataset Diversity and Model Generalizability 30
2.5 Comparative Studies and Benchmarking in Machine Learning for Semiconductor Yield Prediction 31
2.5.1 Analysis of Traditional Machine Learning Approaches 34
2.6 Synthesis of Research Gaps and Implications for This Study 35
CHAPTER 3 METHODOLOGY 37
3.1 Dataset description 37
3.2 Overall Pipeline Structure 38
3.3 Data Ingestion and Initial Cleaning 41
3.4 Data Splitting and Class-Imbalance Handling 41
3.4.1 Implementation details for leakage prevention 42
3.5 Feature Extraction Framework 43
3.6 Feature Expansion and Selection 44
3.6.1 Feature Expansion 44
3.6.2 Feature Selection Strategy 45

3.7 Model Training and Evaluation 46
CHAPTER 4 RESULTS AND DISCUSSION 48
4.1 Overview 48
4.2 Feature selection Analysis (Objective 1) 49
4.2.1 Feature selection methods and subset sizing 50
4.2.2 Comparative results 53
4.2.3 Study-by-study feature comparison with prior SECOM work 55
4.3 Performance of Traditional Machine Learning Algorithms (Objective 2) 57
4.3.1 Performance leaderboard 57
4.3.2 Performance of Linear Models (SVM and Linear Regression) 59
4.3.3 Generalization behavior of non-linear models 62
4.4 Model Generalizability (Objective 3) 63
4.5 Comparison with Previous Studies 66
4.6 Limitations 68
4.7 Chapter Summary 69
CHAPTER 5 CONCLUSION AND FUTURE WORK 72
5.1 Overview 72
5.2 Conclusion of the Study 72
5.3 Contributions of the Study 73
5.4 Future Work 74
5.5 Chapter Summary 75
CHAPTER A APPENDIX 76
A.1 RFE-selected feature list 76
REFERENCES 77

LIST OF FIGURES

Figure 1.1	Illustrative market and investment interest in semiconductor yield analytics tools.	4
Figure 2.1	Conceptual illustration of a non-linear decision boundary produced by a kernel Support Vector Machine (SVM).	20
Figure 2.2	Conceptual illustration of the Random Forest algorithm, where multiple decision trees are trained on random subsets of data and features and aggregated to form the final prediction.	21
Figure 2.3	Conceptual illustration of linear regression, where a linear function is fitted to capture the relationship between an input variable and the target output.	23
Figure 2.4	Illustration of the XGBoost gradient boosting process where multiple decision trees are trained sequentially to correct prediction errors and model complex non-linear relationships (Friedman, J. H., 2001).	24
Figure 2.5	Schematic illustration of a Decision Tree model, where internal nodes represent feature-based splitting rules and terminal leaves provide the predicted class or continuous output.	26
Figure 3.1	Pipeline stages for the proposed SECOM yield prediction framework, illustrating the end-to-end workflow and the training-only constraint used to prevent data leakage.	40
Figure 4.1	Accuracy comparison across feature selection methods.	51
Figure 4.2	F1-score comparison across feature selection methods.	52

Figure 4.3 Confusion matrix for Linear Regression on the locked SECOM test set using RFE-selected features (50 variables). 61
Figure 4.4 Confusion matrix for XGBoost on the locked SECOM test set using RFE-selected features (50 variables). 62

LIST OF TABLES

Table 2.1 Comparative Studies and Benchmarking of Traditional Machine Learning in Semiconductor Yield Prediction 33

Table 4.1 Comparison of evaluated feature selection methods and subset sizing logic. The table distinguishes between each method's selection mechanism and how the final feature count was determined. 53

Table 4.2 Comparative analysis of feature selection strategies based on test-set Macro F1-score (primary metric) and accuracy (secondary metric). "Best model" denotes the classifier with the highest Macro F1-score under each feature selection method. 54

Table 4.3 Study-by-study feature comparison illustrating how feature representations proposed in prior SECOM-related work correspond to, or are extended by, the base, interaction, and selected feature spaces used in this project. 56

Table 4.4 Final model rankings using RFE-selected features (50 variables), sorted by locked test accuracy (tie-broken by locked test Macro F1-score). The Overfit Gap is computed as $\text{Acc}_{train} - \text{Acc}_{test}$. 58

Table 4.5 SVM confusion matrix on the locked SECOM test set using RFE-selected features (50 variables). The test set contains 440 non-failure samples and 31 failure samples. 60

Table 4.6 Generalizability analysis quantified using the Overfit Gap under the RFE+SMOTE training configuration. A lower gap indicates a more robust and deployment-ready model. 65

Table 4.7 Benchmarking this study against selected prior works that report classification accuracy on the SECOM dataset. Metrics are copied as stated in each reference; differences in task definition, preprocessing, imbalance handling, and evaluation protocol may limit direct numerical comparability. 67

Table A.1 List of the 50 features selected by RFE. 76

LIST OF ABBREVIATIONS

AVT	Auditory Vigilance Task

CHAPTER 1

INTRODUCTION

1.1 Background

Semiconductor manufacturing underpins contemporary technological systems, and continued device scaling has increased process complexity while tightening tolerances across fabrication steps. In this context, maintaining high yield and consistent product quality is essential for sustaining profitability and industrial competitiveness (Adly et al., 2015; Chen et al., 2000). Because wafer fabrication involves large material volumes and tightly coupled process stages, even modest yield degradation can translate into substantial costs through scrap, rework, delayed delivery, and reduced effective capacity (Ahmed et al., 2025).

In semiconductor manufacturing, yield denotes the proportion of functional dies relative to the total number of dies produced on a wafer. Yield analysis therefore involves identifying yield-loss events and understanding the process conditions and defect mechanisms that cause them (Ashby et al., 2024). Modern fabrication lines are instrumented with extensive in-line sensors and metrology, producing high-dimensional and heterogeneous data streams that reflect complex interactions among equipment, materials, and environmental factors (Akpabio et al., 2021; Aridas et al., 2025). However, conventional manufacturing oversight approaches, including manual inspections and static statistical process control, may be insufficient to capture nonlinear relationships, subtle interactions, and time-varying behavior in these data (Adipraja et al., 2024; Amuru et

al., 2023).

Machine learning methods provide a data-driven means to model such relationships, enabling predictive analytics for early identification of yield risk and supporting proactive process control (Chang et al., 2024; Chen & Toly, 2018). When integrated with sensor-based monitoring, these techniques can improve the timeliness and consistency of decision-making by learning patterns associated with yield loss and by reducing variability that obscures weak signals in traditional analyses (Ahmed et al., 2025; Amuru et al., 2023).

1.2 Importance and Relevance of Yield Analysis

Early and accurate identification of yield-loss drivers is essential for sustaining the reliability and economic performance of semiconductor fabrication lines. Robust yield analysis supports timely interventions, such as tool tuning, recipe adjustment, or targeted maintenance, thereby reducing scrap, minimizing rework and downtime, and improving overall equipment effectiveness (Ahmed et al., 2025; Chen & Toly, 2018; Chen et al., 2000). These benefits are amplified in high-volume manufacturing, where small shifts in process stability can propagate across many wafers and lead to substantial cumulative losses (Baek et al., 2024).

The significance of yield analysis has increased further with the adoption of extensive inline sensing and metrology. Continuous data streams enable earlier detection of process excursions and more consistent decision-making than periodic manual checks, provided that the data can be interpreted effectively (Akpabio et al., 2021; Aridas et al., 2025). Consequently, data-driven approaches that reliably extract actionable signals from complex manufacturing data are central to improving yield, maintaining product quality, and supporting dependable supply for downstream, performance-critical electronic systems

(Ahmed et al., 2025; Amuru et al., 2023).

1.2.1 Local Perspective

In local manufacturing environments, yield losses often translate directly into extended cycle time, increased rework, and higher test and packaging costs, making rapid diagnosis and prediction especially valuable for maintaining competitiveness. Practical deployment further requires models that can operate on high-dimensional sensor streams and remain robust under process drift, missing values, and class imbalance (Park et al., 2022, 2024). Prior studies demonstrate that machine learning can support yield prediction and diagnosis across key production stages, including wafer acceptance testing and final test, thereby enabling earlier containment of yield excursions and more targeted corrective actions (Fan et al., 2022; Jiang et al., 2020, 2021). In addition, explainable and expandable modeling frameworks are increasingly emphasized to support trust and actionable decision-making on the factory floor (Lee et al., 2023). Accordingly, this study focuses on comparing traditional machine learning approaches under realistic data constraints, aligning predictive performance with the operational needs of local semiconductor production.

1.2.2 Market Size and Investment Trends in Semiconductor Yield Analytics

Beyond its operational importance, yield prediction and optimization has become an active area of industrial investment as semiconductor firms expand data infrastructures for Industry 4.0 manufacturing and pursue faster, more reliable decision cycles. Recent literature highlights growing adoption of data-driven process monitoring, virtual metrology, and AI-enabled process control as practical pathways to reduce measurement latency, improve fault diagnosis, and shorten yield ramp-up (Hsiao et al., 2024; Maitra et al., 2024; Xu, Hong-Wei, et al., 2022). In parallel, the rapid evolution of machine learning architectures for defect recognition and yield-related inference reflects

sustained R&D attention toward manufacturing analytics capabilities that can generalize across tools, products, and operating conditions (Chen et al., 2024; Kim, Tongwha, et al., 2023; Lin et al., 2025). These trends motivate the present workflow, which combines data preprocessing, systematic feature selection, and comparative evaluation of traditional machine learning models to obtain deployable performance improvements while maintaining interpretability and robustness in production settings (Lee et al., 2023; Park et al., 2024).

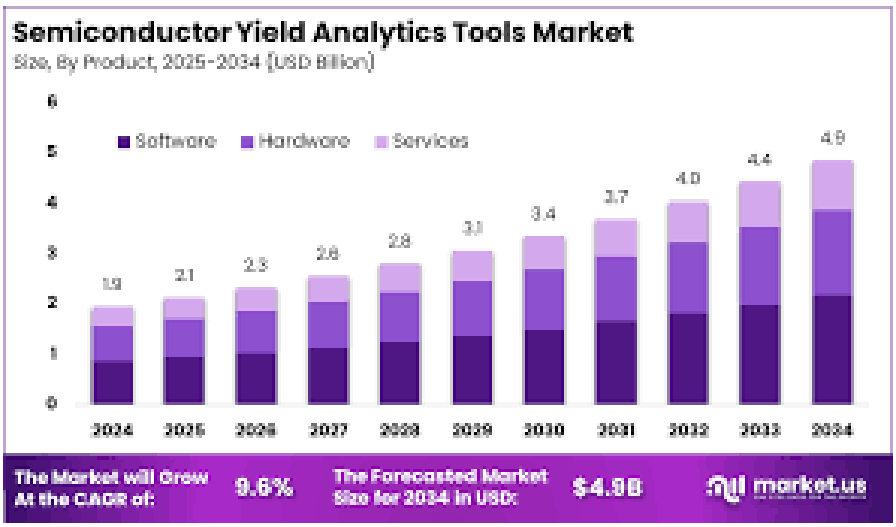

Figure 1.1. Illustrative market and investment interest in semiconductor yield analytics tools.

1.2.3 Economic Impact of Yield-Loss Events and Manufacturing Failures

Yield degradation in semiconductor manufacturing is not merely a technical concern; it carries substantial economic consequences through scrap, rework, delayed deliveries, and unplanned tool downtime. Because wafer fabrication is highly capital intensive, even short disruptions or localized process excursions can propagate across lots and reduce effective capacity, amplifying losses at high volume (Adipraja et al., 2024; Ahmed et al., 2025). Empirical studies show that data-driven fault detection, virtual metrology, and yield prediction can provide earlier warning of abnormal equipment and process conditions, enabling more targeted maintenance and faster containment actions before

defects accumulate (Chang et al., 2023; Kim, Eunji, et al., 2023; Xu, Hong-Wei, et al., 2022). From an operational standpoint, these capabilities reduce the likelihood of extended recovery periods during yield ramp-up and help stabilize performance under drift and variability (Park et al., 2024; Xu et al., 2024). Therefore, the workflow adopted in this study is designed to deliver reliable predictive signals through preprocessing of high-dimensional sensor data, systematic feature selection, and benchmarking of traditional machine learning models, with the aim of supporting cost avoidance and improved manufacturing resilience (Lee et al., 2023).

1.2.4 Government Funding and Policies Supporting Smart Manufacturing Analytics

Public policy and funding priorities increasingly encourage the digital transformation of manufacturing, including investments in sensing, data infrastructure, and analytics capabilities that strengthen quality and reliability in high-value industries such as semiconductors. In this context, data-driven monitoring, virtual metrology, and AI-assisted process control are widely recognized as enabling technologies for more resilient and efficient production systems (Hsiao et al., 2024; Lin et al., 2025; Maitra et al., 2024). At the same time, recent surveys and industrial studies emphasize the need for deployable machine learning solutions that remain robust under practical constraints, including missing values, drift, and heterogeneous data sources. This emphasis aligns with broader policy goals of improving productivity and reducing waste in advanced manufacturing (Chen et al., 2024; Park et al., 2024). These policy-driven directions provide additional justification for the present workflow, in which systematic preprocessing and feature selection are used to produce reliable and interpretable predictive models that can be integrated into routine yield monitoring and decision support (Lee et al., 2023).

1.2.5 Industry Initiatives and Data-Driven Yield Monitoring Programs

Beyond policy signals, industry-led initiatives increasingly operationalize smart manufacturing analytics through end-to-end yield monitoring programs that combine in-line sensing, virtual metrology, and predictive modeling. Semiconductor fabs have long pursued data-centric monitoring frameworks to detect process excursions early and support continuous yield improvement, with more recent efforts emphasizing scalable data platforms and learning-based decision support (Kim, Eunji, et al., 2023; Nakata et al., 2017). In particular, virtual metrology programs have matured as a practical mechanism for estimating quality variables without exhaustive physical measurements, enabling faster feedback for control and diagnosis (Kang et al., 2017; Maitra et al., 2024; Xu, Hong-Wei, et al., 2022). Complementing these approaches, advances in AI-enabled process control and digital-twin-oriented optimization highlight a broader industrial push toward closed-loop, data-driven improvement cycles that can adapt to drift and tool-to-tool variability (Gentner et al., 2024; Hsiao et al., 2024; Lin et al., 2025).

Within this landscape, a recurring implementation requirement is that analytics models must be accurate, robust, and deployable under production constraints such as high dimensionality, missing data, and shifting operating regimes (Ou et al., 2024; Park et al., 2024). The workflow adopted in this study aligns with these industrial programs by prioritizing reproducible preprocessing, systematic feature selection, and comparative evaluation of traditional machine learning methods that are easier to validate and integrate into existing monitoring pipelines (Lee et al., 2023).

1.3 Existing Approaches and Challenges

Existing yield analysis approaches in semiconductor manufacturing range from traditional statistical monitoring and rule-based diagnostics to modern machine learning-driven frameworks that leverage high-dimensional sensor and metrology data. Early methods relied heavily on manual engineering analysis and classical process control,

which remain useful for detecting gross shifts but can struggle with nonlinear interactions, multistage dependencies, and rapidly evolving process windows (Adly et al., 2015; Amuru et al., 2023; Chen et al., 2000). To improve sensitivity and automation, recent studies increasingly apply supervised learning, deep learning, and virtual metrology to predict yield outcomes and identify patterns associated with defect generation or equipment anomalies (Chang et al., 2023; Kim, Tongwha, et al., 2023; Xu, Hong-Wei, et al., 2022).

Despite these advances, several challenges continue to limit reliable deployment in production settings. Semiconductor datasets are often noisy, heterogeneous, and imbalanced, with missing values and drifting distributions that can reduce model stability and cause optimistic results to fail when transferred across tools, products, or fabs (Park et al., 2022, 2024; M. Xu & Zhao, 2024). Moreover, high predictive accuracy alone may be insufficient if models do not provide actionable explanations for engineers or cannot be maintained as processes evolve (Chen et al., 2024; Lee et al., 2023). These constraints motivate this project's workflow, which emphasizes robust preprocessing, systematic feature selection, and comparative evaluation of traditional machine learning models as a pragmatic balance between performance, interpretability, and generalizability.

1.3.1 Limitations of Traditional Methods

Traditional yield diagnosis in semiconductor fabs has long depended on manual engineering review, rule-based screening, and classical statistical process control. While these approaches remain valuable for detecting obvious shifts, they increasingly struggle to scale with the volume, velocity, and heterogeneity of modern in-line sensing and metrology streams (Adly et al., 2015; Akpabio et al., 2021; Aridas et al., 2025). Because many traditional workflows assume a limited number of key variables and relatively stable operating regimes, they can be ill-suited to capturing the nonlinear interactions and multistage dependencies that characterize advanced process integration (Amuru et al.,

2023; Chen et al., 2000).

A primary limitation of manual and heuristic-driven analysis is variability in interpretation and actionability. Engineering review often relies on locally tuned thresholds, informal rules, and tacit knowledge accumulated through experience. As a result, conclusions may differ across analysts, shifts, and fabs, and the rationale behind a decision may be difficult to standardize or audit (Amuru et al., 2023). This lack of consistency is particularly problematic when processes evolve quickly (e.g., recipe updates, tool matching changes, or product mix shifts), because re-tuning rule sets and control limits can lag behind manufacturing reality (Park et al., 2024).

Traditional methods can also be reactive rather than predictive. Classical SPC is typically effective at flagging large deviations after they occur, but it may provide limited early warning for subtle drift, precursor signatures, or compound failure modes that develop over multiple steps (Adly et al., 2015; Xu, Hong-Wei, et al., 2022). In high-volume manufacturing, where excursion containment time strongly influences the number of affected wafers, these delays can translate into increased scrap, rework, and extended recovery cycles (Ahmed et al., 2025; Maitra et al., 2024).

Robustness under realistic production conditions presents another persistent challenge. Noise, missing values, and heterogeneous sampling rates can obscure weak yield signatures and undermine fixed thresholds or static control limits, leading either to missed excursions or excessive false alarms (Park et al., 2022; Xu et al., 2024). Additionally, distribution drift and tool-to-tool variation can cause monitoring logic tuned on one time period or tool to degrade when transferred, which reduces reliability in deployment (Park et al., 2024; Xu et al., 2024).

Finally, traditional workflows often struggle with attribution when multiple interacting factors contribute to a single yield event. Root-cause analysis commonly becomes an iterative, labor-intensive process that requires cross-tool context and multivariate reasoning that is difficult to formalize with hand-crafted rules alone (Ou et al., 2024). When coupled with growing expectations for explainability and maintainability in deployed analytics, these limitations motivate more automated and data-driven alternatives that can learn from high-dimensional data while remaining interpretable and operationally usable on the factory floor (Chen et al., 2024; Lee et al., 2023).

1.3.2 Challenges with Automated Methods

Motivated by the limitations of manual and rule-based workflows, automated yield analytics based on machine learning and advanced signal processing have been widely adopted to improve consistency and to extract predictive signals from large-scale sensor and metrology data (Chang et al., 2023; Maitra et al., 2024; Xu, Hong-Wei, et al., 2022). Nevertheless, automation introduces its own set of challenges that can limit reliability when models are deployed in production environments.

A central technical difficulty is representation learning and feature selection. High-dimensional fab data often contain redundant, weakly informative, or highly correlated variables, and naive use of all available signals can increase overfitting and reduce model stability (Akpabio et al., 2021; Park et al., 2022). While systematic feature selection can improve efficiency and interpretability, selecting features that remain meaningful across tools, product families, and time periods remains non-trivial, particularly under evolving process windows and tool matching changes (Ou et al., 2024; Park et al., 2024).

A second challenge concerns model selection and optimization. Different algorithms exhibit different sensitivities to class imbalance, noise, and nonlinear interactions, and

model performance can vary substantially depending on preprocessing choices, hyperparameter settings, and validation strategy (Park et al., 2022; Xu et al., 2024). In production settings, exhaustive tuning can be computationally expensive and may yield fragile configurations that perform well in offline experiments but degrade under drift or when transferred to a new toolset (Park et al., 2024; Xu et al., 2024).

Generalization and transferability are persistent barriers to dependable automation. Models trained on historical data may fail when the data distribution shifts due to equipment maintenance, recipe updates, chamber seasoning, or changes in product mix (Park et al., 2024; Xu et al., 2024). This problem is compounded by missing values, heterogeneous sampling rates, and rare-event targets (e.g., low-yield excursions), which can cause optimistic training results to collapse in deployment (Park et al., 2022). Consequently, robust evaluation protocols and adaptation mechanisms are often required to sustain performance beyond the original training context (Chen et al., 2024; Xu et al., 2024).

Finally, automated approaches must address interpretability and operational integration. Even accurate models may be of limited value if they cannot provide explanations that are actionable for process engineers, or if they are difficult to maintain as processes and data pipelines evolve (Chen et al., 2024; Lee et al., 2023). Ensuring that automated methods deliver both reliable predictive performance and usable diagnostic insight remains a key research and engineering challenge for practical yield monitoring programs (Maitra et al., 2024; Ou et al., 2024).

1.4 Problem Formulation

This work frames yield analytics as a supervised learning problem defined over high-dimensional manufacturing data collected from in-line sensors and metrology. Each

wafer-level (or lot-level) observation is represented by a vector of process variables, and the associated target describes the manufacturing outcome (e.g., a continuous yield value or a discrete yield state). The objective is to train a predictive model that remains accurate and stable under realistic production constraints, including noisy measurements, missing values, heterogeneous sampling, and distribution drift over time and across tools (Akpabio et al., 2021; Park et al., 2022, 2024; Xu et al., 2024).

From an operational standpoint, the formulation emphasizes three requirements. First, the modeling pipeline must be deployable on production data streams, which motivates robust preprocessing and handling of imbalance and rare-event behavior typical of yield-loss episodes (Ahmed et al., 2025; Park et al., 2022). Second, the learned model should support decision-making by enabling diagnosis or prioritization of suspected drivers, aligning predictive outputs with engineer workflows and maintainability expectations (Chen et al., 2024; Lee et al., 2023). Third, performance should be assessed using evaluation protocols that reflect transfer and drift scenarios, rather than relying solely on in-sample accuracy (Park et al., 2024; Xu et al., 2024).

1.4.1 Research Gap

Although the literature demonstrates the value of machine learning for yield prediction, fault detection, and virtual metrology, many studies evaluate a single algorithm or a narrow family of models on a specific dataset and report results under a fixed experimental setting (Chang et al., 2023; Maitra et al., 2024; Xu, Hong-Wei, et al., 2022). This makes it difficult to determine the relative strengths and weaknesses of traditional machine learning methods under varying production conditions (e.g., differing noise profiles, sampling regimes, and tool-to-tool variation), and it limits confidence that reported performance will transfer to other tools, products, or time periods (Park et al., 2024; Xu et al., 2024).

A second gap concerns feature representation and its interaction with algorithm choice. Modern fabs generate heterogeneous measurements with complex dependencies, yet prior work often provides limited analysis of how alternative feature sets and selection strategies affect downstream model robustness, interpretability, and computational cost (Akpabio et al., 2021; Ou et al., 2024; Park et al., 2022). Because different learners can be sensitive to different representations, insufficient attention to feature engineering and selection can lead to models that appear accurate in controlled evaluations but become unstable under drift or missing data (Park et al., 2022; Xu et al., 2024).

Finally, there remains a practical gap between predictive performance and deployability. Industrial adoption increasingly requires models that are not only accurate but also explainable, auditable, and maintainable as processes evolve (Chen et al., 2024; Lee et al., 2023). However, systematic comparisons that jointly consider performance, robustness, and interpretability across realistic data constraints are still limited, motivating the comparative workflow adopted in this study (Maitra et al., 2024; Park et al., 2024).

1.4.2 Problem Statement

Based on the identified research gap, this study addresses three interrelated problems in semiconductor yield analytics. First, there is a need for a comprehensive and reproducible comparison of widely used machine learning algorithms for yield prediction and yield-loss identification, evaluated using consistent metrics (e.g., accuracy and F1-score for classification settings, and error measures for regression settings) together with computational cost considerations that affect on-line monitoring feasibility (Chang et al., 2023; Park et al., 2022; Xu, Hong-Wei, et al., 2022). Second, there is limited understanding of which feature representations and feature selection strategies are most effective for different learners when applied to high-dimensional sensor and metrology data, motivating a systematic investigation of feature subsets and selection procedures to identify robust and discriminative predictors while preserving interpretability (Akpabio

et al., 2021; Lee et al., 2023; Ou et al., 2024). Third, there is a need to evaluate model generalizability under realistic transfer scenarios, including distribution drift over time and tool-to-tool or product-to-product variation, by testing models beyond their original training conditions to better quantify robustness and real-world applicability (Chen et al., 2024; Park et al., 2024; Xu et al., 2024).

1.5 Study Objectives

This study has three main objectives. First, it identifies effective feature sets for traditional machine learning models by evaluating alternative feature combinations derived from the cleaned SECOM sensor and test variables. Second, it benchmarks the yield-screening performance of multiple traditional algorithms (SVM, Decision Tree, Random Forest, Linear Regression, and XGBoost) under a consistent, leakage-safe training and validation protocol. Third, it assesses model generalizability by reporting final performance on a locked hold-out test set representing previously unseen manufacturing samples.

1.6 Scope of Work

The scope of this thesis is limited to supervised binary yield screening using the publicly available SECOM dataset, where each observation is represented by a high-dimensional vector of anonymized sensor and test variables ($F1{-}F590$) and the target indicates a pass or fail outcome. A hold-out test set in this study refers to a subset of data that is separated once using a stratified split and then kept completely untouched during model training, preprocessing, feature selection, and hyperparameter tuning, and is used only for the final unbiased evaluation.

Specifically, the work comprises the following:

1. Data ingestion and cleaning, including missing-value handling and feature scaling under a strict no-leakage protocol.

2. Stratified partitioning into a training set and a locked hold-out test set.

3. Class-imbalance handling applied to the training data only.

4. Systematic feature selection to obtain compact and robust feature subsets suitable for high-dimensional manufacturing data.

5. Comparative training and tuning of traditional machine learning classifiers (e.g., Linear Regression, SVM, Decision Tree, Random Forest, and XGBoost) under a consistent validation protocol.

6. Evaluation using imbalance-aware metrics (precision, recall, F1-score, and ROC–AUC), with explicit analysis of generalization on the locked hold-out test set.

This scope intentionally prioritizes reproducibility, leakage prevention, and deployable baseline modelling over domain-specific sensor semantics, because SECOM features are anonymized and direct process interpretability is therefore limited.

1.7 Organization of the Thesis

This thesis is structured as follows. Chapter 1 introduces the study background and motivation and defines the problem setting, objectives, and scope, thereby establishing the context and rationale for the work. Chapter 2 surveys related literature on yield an-

alytics and manufacturing data-driven monitoring, summarizing established approaches and key challenges, and motivating the methodological choices adopted in this thesis. Chapter 3 describes the research methodology in detail, including data acquisition, preprocessing, feature construction and selection, model training, and the evaluation protocol, with emphasis on reproducibility of the experimental pipeline. Chapter 4 reports and discusses the experimental findings, including comparative model performance and the influence of feature choices on predictive behavior and robustness across datasets. Chapter 5 concludes by summarizing the main contributions and results, discussing practical implications for deployment in semiconductor manufacturing, and outlining directions for future work. Supplementary material is provided in the Appendix, including additional dataset information, methodological details, parameter settings, and extended results to support replication and further study.

CHAPTER 2

LITERATURE REVIEW

2.1 Machine Learning in Semiconductor Yield Prediction and Yield Analytics

Semiconductor manufacturing is widely recognized as a complex and capital-intensive domain in which small deviations in process stability can translate into substantial yield loss and cost. As technology nodes shrink and process windows tighten, the volume and dimensionality of in-line sensor, metrology, and test data have increased to the point where manual inspection and purely rule-based statistical monitoring are often insufficient for timely diagnosis and intervention (Adly et al., 2015; Amuru et al., 2023; Ko et al., 2023; Sandru et al., 2022). Within this context, machine learning (ML) has become a central tool for converting heterogeneous manufacturing data into actionable predictions, supporting earlier identification of yield-limiting mechanisms and more systematic yield improvement (Jiang et al., 2021; Q. Wang & Huang, 2025).

In line with this thesis title (comparative study of traditional ML approaches for yield prediction), the literature in Section 2.1 is reviewed through a workflow consistent with the project's experimental pipeline: (i) acquisition of manufacturing datasets (equipment parameters, test measurements, and yield indicators), (ii) preprocessing and representation of raw measurements into features, (iii) feature selection to reduce dimensionality and improve robustness, and (iv) benchmarking of multiple conventional ML models under consistent evaluation protocols. This flow reflects common industrial analytics practice, where deployability depends not only on accuracy but also on stability under

noise, missing values, and drift (Akpabio et al., 2021; Park et al., 2022, 2024; Xu et al., 2024).

In yield-oriented analytics, prior work is commonly framed around three task families. First, classification predicts discrete yield-related outcomes (e.g., pass or fail, low-yield vs. normal-yield lots) from upstream measurements, which aligns with binary quality screening used in high-volume manufacturing. Second, regression predicts continuous quality outcomes (e.g., yield percentage, parametric targets, or virtual metrology estimates) from process and test variables (Chen & Toly, 2018; Jiang et al., 2021; Xu, Hong-Wei, et al., 2022). Third, anomaly detection aims to identify rare process excursions or distribution shifts that can precede yield loss (Shin, Jin-Su, Kim, Min-Joo, Kim, et al., 2025; Shin, Jin-Su, Kim, Min-Joo, Lee, & Dong-Hee, 2025). The remainder of this chapter synthesizes prior work along these dimensions and highlights recurring challenges in generalizability, feature engineering, and benchmarking practices.

2.2 Overview of Traditional Machine Learning Algorithms in the Proposed Yield Prediction Workflow

This section reviews prior studies that apply traditional machine learning algorithms to semiconductor yield prediction and yield analytics. Emphasis is placed on (i) how conventional models are positioned within a typical workflow (data acquisition from process, equipment, and test sources; preprocessing and feature engineering; model training and validation; and performance benchmarking), (ii) the motivations for selecting specific algorithms under manufacturing constraints (interpretability, computational efficiency, robustness to high dimensionality and noise), and (iii) findings from comparative studies that report performance trade-offs across models under consistent evaluation protocols.

2.2.1 Support Vector Machine (SVM)

Support Vector Machines (SVMs) are a widely adopted class of supervised learning models in semiconductor yield analytics because they remain effective in high-dimensional settings and can represent non-linear decision boundaries through kernel functions. Conceptually, SVM learning seeks a maximum-margin separating hyperplane (or its regression analogue), which often provides strong generalization when the number of measured variables is large relative to the number of labeled samples available for model training (Yeh et al., 2024).

Within the yield prediction workflow, SVM-based models are commonly used for yield-related classification (e.g., quality screening and fault detection) and regression (e.g., virtual metrology and parametric prediction), where relationships between upstream process and test variables and downstream yield outcomes can be highly non-linear (Deivendran et al., 2025; Kim et al., 2022). One-Class SVM variants have also been applied for anomaly detection to flag potential process excursions and distribution shifts that precede yield loss (Shin, Jin-Su, Kim, Min-Joo, Kim, et al., 2025; Shin, Jin-Su, Kim, Min-Joo, Lee, & Dong-Hee, 2025). In addition, sparse formulations such as L1-regularized SVMs (L1-SVM) provide embedded feature selection, supporting more stable and interpretable models when manufacturing datasets contain many correlated sensors or test variables (Shaer et al., 2023).

Empirical results across yield-oriented studies further support the role of SVMs as strong baselines and competitive learners under realistic manufacturing constraints. For example, (Yeh et al., 2024) reported that combining SVM with mutual-information-based feature selection and simplified swarm optimization improved classification performance while reducing the number of retained features, illustrating how SVMs are frequently paired with feature selection to address dimensionality and noise. In lithography-related prediction tasks, (Xu et al., 2018) showed that SVM-based models can achieve compa-

rable predictive quality to linear baselines while substantially improving computational efficiency, which is a practical consideration for high-throughput manufacturing analytics. In testing and quality screening, (Xama et al., 2023) demonstrated that SVM classifiers can improve defect escape detection with limited yield impact, and (Park et al., 2024) reported improved screening performance on the SECOM dataset when SVM was integrated with resampling and scaling strategies. For continuous outcome prediction, probabilistic support vector regression (PSVR) combined with Bayesian optimization has also been reported as effective for early quality prediction prior to burn-in, highlighting the flexibility of SVM variants for both classification and regression within yield prediction workflows (Ahmed et al., 2025).

Figure 2.1 illustrates a non-linear decision boundary generated by a kernel Support Vector Machine (SVM). The figure highlights SVM's ability to separate complex, overlapping data distributions by maximizing the margin between classes, which is advantageous for semiconductor yield prediction where relationships between process parameters and yield outcomes can be high-dimensional and non-linear.

2.2.2 Random Forest (RF)

Random Forest (RF) is an ensemble learning method that aggregates predictions from multiple decision trees to improve generalization in both classification and regression. This bagging-based structure is well-suited to semiconductor yield analytics, where datasets are often high-dimensional, noisy, and characterized by non-linear feature interactions. In practice, RF is commonly adopted as a robust baseline model within yield prediction workflows because it requires limited distributional assumptions, provides stable performance under correlated inputs, and naturally supports variable-importance analysis to identify influential process and test factors (Deivendran et al., 2025; Hsieh et al., 2025; Kim, Eunji, et al., 2023; Wang et al., 2023).

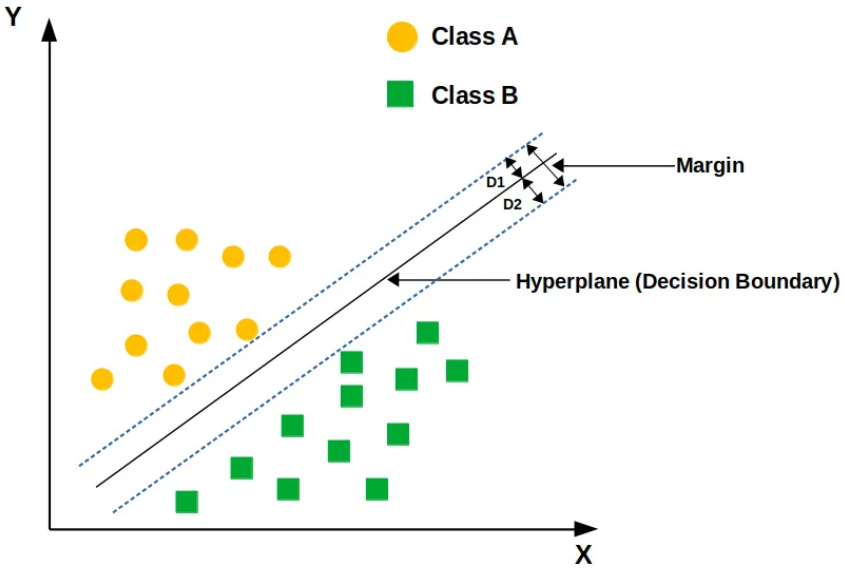

Figure 2.1. Conceptual illustration of a non-linear decision boundary produced by a kernel Support Vector Machine (SVM).

As shown in Figure 2.2, Random Forest combines multiple decision trees to enhance prediction stability and generalization, which is advantageous when dealing with noisy and high-dimensional semiconductor manufacturing data.

Empirical studies further highlight RF's effectiveness in yield-oriented applications. For example, (Kim, Eunji, et al., 2023) reported that RF-based models can support low-yield diagnosis by ranking influential equipment signals and excursion-related factors, enabling targeted process interventions. In broader manufacturing analytics, RF has been used to identify critical variable combinations and operational patterns associated with performance degradation, with feature importance serving as a practical bridge between prediction and actionable engineering insights (Hsieh et al., 2025; Wang et al., 2023). For continuous outcome modeling, (Lee et al., 2023) demonstrated that RF regression combined with explainability techniques can achieve competitive predictive accuracy while providing interpretable attributions, which is beneficial when models are

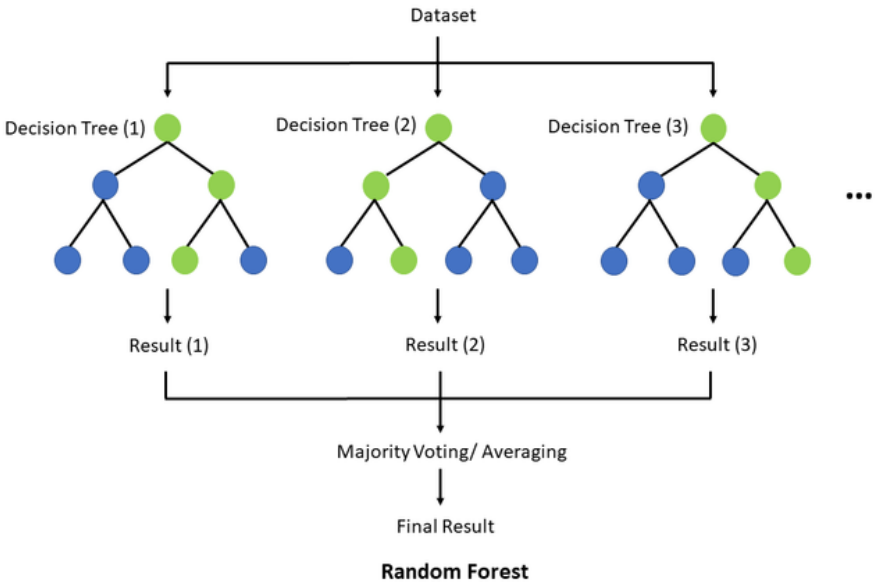

Figure 2.2. Conceptual illustration of the Random Forest algorithm, where multiple decision trees are trained on random subsets of data and features and aggregated to form the final prediction.

intended to inform decision-making in fab settings. Overall, these findings motivate the inclusion of RF in comparative benchmarking studies focused on robust and interpretable yield prediction.

2.2.3 Linear Regression (LR)

Linear Regression (LR) is a classical supervised learning method that models the relationship between a continuous response (e.g., yield or a parametric target) and one or more explanatory variables through a linear function. Owing to its computational simplicity and transparent parameterization, LR is frequently used in semiconductor manufacturing analytics as an interpretable baseline for yield prediction and process modeling, particularly when engineers require direct insight into how changes in key variables relate to expected quality outcomes (Adly et al., 2015; Chen & Toly, 2018).

Within yield prediction workflows, LR and related generalized linear models are commonly applied to (i) estimate continuous yield- or quality-linked targets from process and test measurements and (ii) provide fast, stable predictors that can be deployed for monitoring or as reference models in comparative evaluations. For example, (Chen & Toly, 2018) discusses linear-model-based approaches in virtual metrology settings where rapid inference is important, while (Adly et al., 2015) illustrates the use of regression-style models for relating process parameters to yield-relevant outcomes.

Despite its advantages, LR is constrained by its linearity assumption and limited capacity to represent complex interactions without feature transformations. Consequently, many studies position LR as a benchmark against which non-linear learners (e.g., SVMs and tree ensembles) are evaluated. In lithography-related prediction tasks, (Xu et al., 2018) compared linear regression baselines against more flexible models and reported that non-linear methods can achieve improved predictive performance and efficiency in practical workflows. Nevertheless, LR remains valuable when interpretability, computational efficiency, and ease of calibration are prioritized, and it is often strengthened by feature scaling, interaction terms, or dimensionality reduction in high-dimensional manufacturing datasets (Chen & Toly, 2018; Xu et al., 2018).

As shown in Figure 2.3, Linear Regression models the relationship between input features and the target variable by fitting a linear function, making it suitable for capturing direct and interpretable linear trends in semiconductor manufacturing data when linearity is a reasonable approximation.

2.2.4 Extreme Gradient Boosting (XGBoost)

Extreme Gradient Boosting (XGBoost) is a gradient-boosted decision tree algorithm that builds an ensemble of weak learners in a stage-wise manner to minimize a regularized

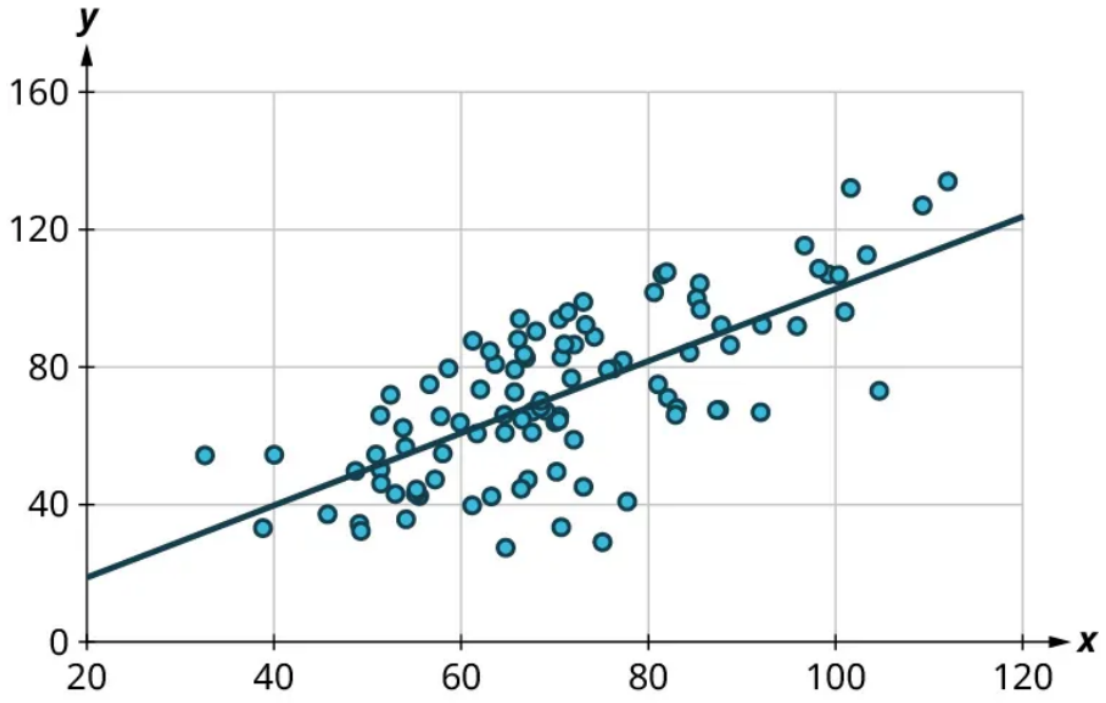

Figure 2.3. Conceptual illustration of linear regression, where a linear function is fitted to capture the relationship between an input variable and the target output.

objective function. In semiconductor yield analytics, this regularization and the additive boosting formulation are attractive because they can capture non-linear feature interactions while controlling overfitting, particularly in settings where process and test variables are numerous, noisy, and partially redundant.

Within the proposed yield prediction workflow, XGBoost is typically positioned after preprocessing and feature engineering as a high-capacity predictive model for both classification and regression. Its tree-based structure accommodates mixed-type inputs and non-monotonic relationships between upstream measurements and downstream yield outcomes, while built-in importance scores provide an additional mechanism for ranking influential variables to support diagnosis and process improvement.

Recent semiconductor-oriented studies further motivate the inclusion of XGBoost in comparative benchmarking. (Shamsudin et al., 2023) proposed a cost-sensitive XGBoost approach optimized with Bayesian optimization for imbalanced semiconductor datasets, highlighting the practicality of boosting-based learners when yield-relevant classes are rare and misclassification costs are asymmetric. In virtual metrology applications, (Deivendran et al., 2025) reported that XGBoost, when evaluated alongside other traditional models, contributes competitive predictive performance for process outcome estimation, supporting its role as a strong baseline in yield prediction studies. Collectively, these findings indicate that XGBoost is a suitable candidate for benchmarking against SVM, Random Forest, and linear baselines when developing robust and deployable yield prediction models.

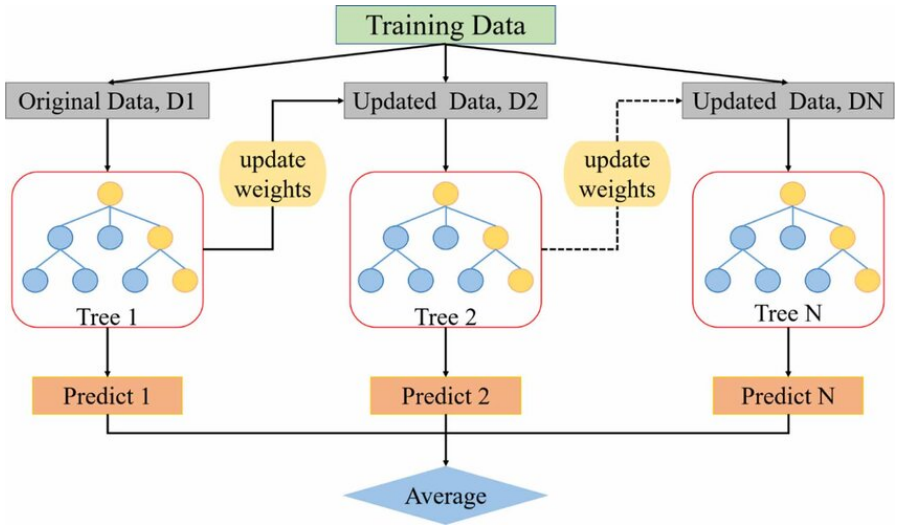

Figure 2.4. Illustration of the XGBoost gradient boosting process where multiple decision trees are trained sequentially to correct prediction errors and model complex non-linear relationships (Friedman, J. H., 2001).

As shown in Figure 2.4, XGBoost employs a gradient boosting framework in which multiple decision trees are trained sequentially to correct the prediction errors of preceding models, enabling effective learning of complex non-linear relationships and feature interactions commonly present in high-dimensional semiconductor manufacturing data.

2.2.5 Decision Trees

Decision Trees (DTs) are non-parametric supervised learning models that perform classification or regression by recursively partitioning the feature space into regions defined by a sequence of decision rules. At each internal node, a split is selected to reduce an impurity criterion (e.g., Gini index or entropy for classification, variance reduction for regression), and the resulting leaf nodes output either a class label (e.g., pass or fail, or low-yield vs. normal-yield) or a continuous estimate of a yield-related target.

In semiconductor yield prediction workflows, DTs are commonly positioned as transparent baseline models for both final test yield classification and yield regression tasks, because their rule-based structure can be inspected by process engineers to identify candidate thresholds and interactions among upstream variables. For example, machine-learning frameworks for semiconductor final test yield classification typically include tree-based learners among the benchmarked traditional algorithms, reflecting their practicality when fast training and interpretability are prioritized (Jiang et al., 2020). Similarly, yield prediction studies that integrate heterogeneous process and test information frequently use tree-structured models (either directly or as components within larger frameworks) to capture non-linear dependencies while retaining a degree of explainability (Jiang et al., 2021; Lee et al., 2023).

Despite these advantages, single decision trees are known to be sensitive to small data perturbations and can overfit without appropriate stopping rules or pruning, particularly in high-dimensional settings common to manufacturing analytics. Consequently, DTs are often treated as a diagnostic or explainable reference point relative to more stable ensemble extensions (e.g., Random Forest or gradient boosting), while still offering a useful mechanism for summarizing complex multivariate relationships into actionable decision logic in yield-oriented applications (Deivendran et al., 2025; Maitra et al., 2024).

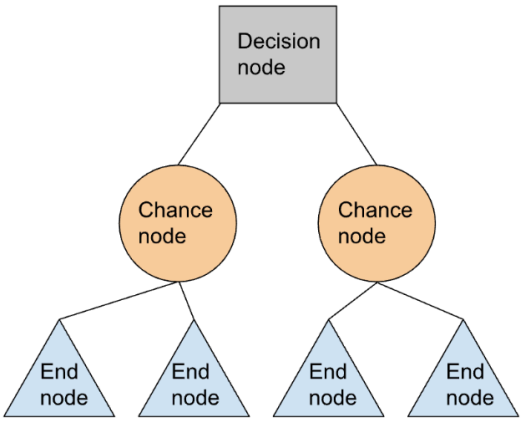

Figure 2.5. Schematic illustration of a Decision Tree model, where internal nodes represent feature-based splitting rules and terminal leaves provide the predicted class or continuous output.

As illustrated in Figure 2.5, the model routes each observation from the root node through a sequence of binary decision rules based on input features. At each split, the feature and threshold are selected to improve node purity (or reduce prediction error), and the traversal ends at a leaf node that outputs the final prediction. This structure provides an explicit, human-interpretable set of rules that can be examined to understand how different process and test variables contribute to yield-related decisions.

2.3 Feature Engineering in Semiconductor Yield Prediction and Analytics

Feature engineering is a central step in semiconductor yield prediction because raw manufacturing data are heterogeneous (equipment sensor traces, inline metrology, Wafer Acceptance Test (WAT), and Final Test (FT) measurements) and often contain noise, missing values, outliers, and drift. Consequently, the predictive quality and deployability of traditional machine learning models depend strongly on how these signals are cleaned, aligned, and transformed into stable feature representations. In particular,

systematic preprocessing choices such as scaling, imputation, and outlier handling have been shown to materially influence downstream model performance in semiconductor process datasets (Park et al., 2024).

Beyond basic preprocessing, feature construction aims to encode process-relevant structure and reduce nuisance variability so that models can learn relationships that generalize across lots, tools, and operating conditions. In yield-oriented applications, this commonly includes aggregating high-frequency sensor readings into summary descriptors (e.g., windowed statistics), generating interaction features motivated by process knowledge, and deriving compact representations that emphasize condition changes linked to yield loss mechanisms. For example, clustering-based representations and ensemble modeling have been used to capture latent operating regimes in final test yield prediction, illustrating how engineered representations can improve the signal-to-noise ratio of yield predictors (Jiang et al., 2021; Maitra et al., 2024).

Because semiconductor datasets are frequently high-dimensional, feature selection and dimensionality reduction are also used to improve robustness, reduce overfitting risk, and lower inference cost. Recent work has proposed dimensionality reduction methods tailored to correlated sensor streams (e.g., sparse PCA variants) for virtual metrology, while other studies employ filter-based selection strategies (e.g., mutual information or Lasso-type selection) to retain only yield-informative variables (Deivendran et al., 2025; Shaer et al., 2023; Wang et al., 2024; Yeh et al., 2024). In ramp-up and yield improvement contexts, these feature engineering steps support faster learning cycles by stabilizing model training as process conditions evolve (Xu et al., 2024).

2.3.1 Common Feature Types in Semiconductor Yield Analysis

Unlike studies that work with physically interpretable measurements (e.g., specific tool sensors or inline metrology channels), the SECOM dataset used in this study provides a fixed-length vector of anonymized variables indexed as $F1{-}F590$. Because the original feature semantics are not disclosed (a common constraint for industrial benchmark datasets), feature engineering in this work focuses on producing numerically stable and model-compatible representations that improve yield classification performance (Park et al., 2024).

In this setting, the most practically relevant "feature types" can be grouped as follows. Raw process and test variables correspond to the original indexed measurements and form the primary input space for the learning algorithms. Data-quality and missingness indicators (implemented as binary flags during preprocessing, where applicable) are often introduced to retain information about censored, unmeasured, or invalid readings; this is particularly important in high-dimensional manufacturing data where missing values are prevalent (Park et al., 2024). Finally, derived and transformed features are used to reduce redundancy and improve generalization, including standardized versions of the raw variables, low-dimensional representations obtained via dimensionality reduction, and subsets produced by feature selection methods that target correlated and noisy sensor streams (Deivendran et al., 2025; Shaer et al., 2023; Wang et al., 2024; Yeh et al., 2024). Collectively, these categories reflect common practice in semiconductor yield analytics, where robust preprocessing, selection, and compact representations are repeatedly shown to be critical for learning reliable yield predictors from large sets of heterogeneous manufacturing signals (Jiang et al., 2021; Maitra et al., 2024).

2.3.2 Feature Selection Approaches in Semiconductor Yield Analysis

Feature selection is frequently required in semiconductor yield analysis because process datasets are high-dimensional, noisy, and strongly correlated, which can degrade gen-

eralization and inflate model complexity. Accordingly, yield modeling studies commonly combine preprocessing with selection strategies that either explicitly search for informative subsets or encourage sparsity during training (Park et al., 2024; Wang et al., 2024; Yeh et al., 2024).

Wrapper methods evaluate candidate feature subsets by repeatedly fitting a predictive model and selecting the subset that optimizes a validation criterion. In yield classification, this includes strategies such as forward selection, backward elimination, and recursive feature elimination (RFE), which can be paired with classifiers like SVMs or tree ensembles to identify compact subsets of the anonymized SECOM variables that maximize discriminative performance while controlling overfitting (Yeh et al., 2024).

Embedded methods perform selection as part of model training. Regularization-based approaches (e.g., L1-type penalties) promote sparse solutions by shrinking weak predictors toward zero, while tree-based learners implicitly rank variables through split gains and impurity reductions. These embedded mechanisms are attractive in manufacturing analytics because they scale to large feature spaces and provide an operational notion of importance that can be used for model compression and subsequent diagnosis (Deivendran et al., 2025; Shaer et al., 2023).

Although many studies apply feature selection within a single modeling pipeline, there remains a practical need for more systematic evaluation of how different selection strategies interact with preprocessing choices, class imbalance handling, and downstream learners when developing robust yield prediction systems (Maitra et al., 2024; Park et al., 2024).

2.4 Dataset Diversity and Model Generalizability

A persistent challenge in semiconductor yield analysis is that datasets collected across different fabs, tools, product variants, and time periods can exhibit substantial heterogeneity. Even when the prediction target is defined consistently (e.g., yield loss, low-yield screening, or quality pass/fail), changes in equipment configurations, sensor calibration, sampling rates, and process recipes can shift the underlying data distribution and alter which variables are informative. As a result, models trained on a single operating regime may achieve high in-sample accuracy yet generalize poorly when deployed under new production conditions or during ramp-up, where drift and regime changes are common (Jiang et al., 2021; Q. Wang & Huang, 2025; Xu et al., 2024).

A second source of limited generalizability arises from differences in data representation and labeling. Public benchmark datasets such as SECOM provide high-dimensional process measurements with anonymized feature indices ($F1$–$F590$), which supports reproducible comparisons but limits domain interpretability and can constrain transfer to settings where sensors and measurement definitions differ. In addition, manufacturing datasets frequently contain missing values, noisy readings, and class imbalance, which can bias learned decision boundaries and make reported performance sensitive to preprocessing, resampling, and evaluation design (Park et al., 2024).

To improve reliability, prior work emphasizes evaluation protocols and modeling choices that explicitly account for heterogeneity. Consistent cross-validation and repeated-split experiments provide more stable estimates than single holdout tests, while feature selection and regularization can reduce sensitivity to spurious correlations in high-dimensional inputs (Shaer et al., 2023; Wang et al., 2024; Yeh et al., 2024). Nevertheless, the literature continues to report that transferring models across datasets or production contexts often requires retraining and/or additional feature engineering to maintain accuracy, highlighting the need for broader benchmarking across multiple

datasets and operating regimes (L. Chen & Zhang, 2024; Park et al., 2024; Xu, Hong-Wei, et al., 2022).

2.5 Comparative Studies and Benchmarking in Machine Learning for Semiconductor Yield Prediction

Comparative studies and benchmarking have been instrumental in establishing practical baselines for semiconductor yield prediction, particularly when multiple conventional learners (e.g., SVM, tree ensembles, and linear models) are evaluated under manufacturing constraints such as high dimensionality, missing values, and class imbalance (Jiang et al., 2021; Park et al., 2024). However, the yield analytics literature also exhibits recurring methodological inconsistencies that complicate direct cross-study comparison and can lead to over-optimistic conclusions.

One common issue is limited comparability across datasets and evaluation splits. Studies frequently report results on different data sources (often proprietary) or under non-uniform train/test protocols, making it difficult to attribute performance differences to algorithms rather than to dataset characteristics or sampling decisions (L. Chen & Zhang, 2024; Xu, Hong-Wei, et al., 2022). A second inconsistency concerns evaluation metrics. Because yield-related classification problems are often imbalanced, accuracy alone can be misleading; nevertheless, reported metrics vary widely across studies, and the choice of metric can materially change which model appears superior (Park et al., 2024).

A further concern is that many comparisons implicitly assume a fixed feature representation, even though model performance can be highly sensitive to preprocessing (scaling, imputation, outlier handling), resampling strategies, and feature selection choices. For example, optimization-driven pipelines and imbalance-aware preprocessing have been shown to materially influence yield screening performance on benchmark manufactur-

ing datasets, suggesting that "algorithm comparisons" should be interpreted jointly with the feature engineering and data preparation steps used to produce model inputs (Dash et al., 2024; Park et al., 2024; Yeh et al., 2024).

Overall, the absence of standardized benchmarking protocols that simultaneously control (i) dataset and split design, (ii) metric selection for imbalanced outcomes, and (iii) feature–algorithm pairings remains a key gap. Addressing this gap requires broader, more reproducible evaluations across diverse operating regimes and consistent experimental designs that support reliable conclusions about model robustness and deployability (L. Chen & Zhang, 2024; Jiang et al., 2021; Xu, Hong-Wei, et al., 2022).

Table 2.1. Comparative Studies and Benchmarking of Traditional Machine Learning in Semiconductor Yield Prediction

No.	Study	Algorithms Compared	Best Performing Algorithm(s)	Performance Highlights
1	(Chen et al., 2023)	ECOC with SVM vs. traditional classifiers	Support Vector Machine (SVM)	Achieved 96.96% accuracy on wafer defect classification
2	(Behera et al., 2024)	Bayesian-optimized SVM vs. baseline SVM	Optimized SVM	Accuracy improved from 74.5% to 88.1% post optimization
3	(Xu et al., 2018)	SVM vs. Linear Regression (LR)	Support Vector Machine (SVM)	Achieved comparable accuracy but with >3× computational speedup for SRAF generation
4	(Deivendran et al., 2025)	Lasso Regression, Random Forest, XGBoost	Random Forest, XGBoost	Improved CMP yield prediction accuracy with hybrid ensemble framework
5	(Dash et al., 2024)	Bayesian Optimization with SVM vs. non-optimized models	Optimized SVM	Achieved consistently high predictive performance for defect detection
6	(Park et al., 2024)	ADASYN, MaxAbs Scaling with SVM vs. baseline models	SVM with Preprocessing	Achieved 85.14% accuracy, 72.95% geometric mean on SECOM dataset
7	(Shaer et al., 2023)	Standard SVM vs. L1-Regularized SVM	L1-Regularized SVM	Achieved 179x computational speedup in defect detection workflows

2.5.1 Analysis of Traditional Machine Learning Approaches

Table 2.1 synthesizes representative comparative studies that apply traditional machine learning to semiconductor quality prediction problems, including yield screening, virtual metrology, and closely related defect/quality classification tasks. Across these works, performance is driven not only by the choice of learner (e.g., SVMs and tree ensembles) but also by how high-dimensional process and test measurements are preprocessed and reduced to stable feature representations (Park et al., 2024; Yeh et al., 2024).

A consistent finding is that margin-based classifiers and ensembles provide strong baselines in high-dimensional settings. For instance, SVM-centered pipelines report competitive accuracy in multiple contexts, and improvements are often obtained by coupling SVMs with optimization or carefully chosen preprocessing steps (Behera et al., 2024; Chen et al., 2023; Park et al., 2024). Tree-ensemble approaches (Random Forest and gradient-boosted trees such as XGBoost) are similarly prominent because they capture non-linear interactions and yield practical importance rankings that support diagnosis and process improvement (Deivendran et al., 2025).

The table also highlights that reported gains frequently depend on the experimental protocol and operational constraints rather than algorithm choice alone. In particular, computational efficiency can be a first-order consideration in manufacturing analytics: Xu et al. (2018) showed that SVM-based solutions can match baseline predictive quality while substantially reducing runtime, which is crucial for high-throughput simulation and decision support. Likewise, sparsity-inducing formulations (e.g., L1-regularized SVMs) can reduce effective dimensionality and accelerate analysis, reinforcing the value of embedded selection mechanisms in scalable yield workflows (Shaer et al., 2023).

Overall, the comparative evidence suggests that traditional models remain competitive

for yield prediction when the pipeline is designed holistically. In particular, performance improves when feature selection, imbalance handling, scaling/imputation, and evaluation metrics are aligned with the deployment objective. This motivates the present study's emphasis on controlled benchmarking under consistent preprocessing and validation settings on the SECOM dataset (Park et al., 2024).

2.6 Synthesis of Research Gaps and Implications for This Study

Across the reviewed literature, four recurring gaps are particularly consequential for credible yield-screening conclusions. First, multiple studies report strong headline accuracy without fully controlling for data leakage introduced by fitting scaling, imputation, resampling, or feature selection using information from validation/test samples, which can inflate performance estimates. Second, class imbalance is frequently handled using resampling methods such as SMOTE or related strategies; however, the interaction between synthetic sampling, model capacity, and out-of-sample stability is often underanalyzed (Chawla et al., 2002; He & Garcia, 2009; Park et al., 2024). Third, the choice of evaluation metric varies widely, and reliance on accuracy alone can obscure failure-class performance in rare-event screening; consequently, class-sensitive measures such as Macro F1 are required to reflect balanced error behavior (He & Garcia, 2009; Sokolova & Lapalme, 2009). Finally, generalizability is commonly inferred from cross-validation alone, even though semiconductor processes exhibit drift and noise that can cause offline results to degrade under deployment conditions (Park et al., 2024; Xu et al., 2024).

These gaps directly motivate the design of the present thesis and align with its objectives. Objective 1 evaluates compact feature representations under a strict train-only fitting rule, ensuring that feature selection decisions do not leak test information. Objective 2 benchmarks multiple traditional learners under identical preprocessing and class-imbalance settings to isolate algorithmic effects while using Macro F1 as a primary metric for screening quality. Objective 3 explicitly assesses deployment risk by compar-

ing training behavior (including SMOTE-augmented settings) against a locked test set, thereby quantifying generalization stability rather than relying solely on peak training or cross-validation performance (Dash et al., 2024; Park et al., 2024).

CHAPTER 3

METHODOLOGY

This chapter presents the complete technical methodology adopted to develop and evaluate a semiconductor yield prediction framework using the SECOM dataset. The proposed framework is organized as a structured machine learning pipeline that transforms high-dimensional manufacturing measurements into robust predictors by combining data preprocessing, feature selection, and comparative model training.

A central design principle of this study is a leak-proof evaluation protocol. In this thesis, leak-proof evaluation refers to an experimental design in which the locked hold-out test set is excluded from all data-dependent operations, and in which all preprocessing, feature selection, class-imbalance handling (including SMOTE), and hyperparameter tuning steps are fitted on training data only within each split or fold and then applied to held-out data using the fitted parameters. This protocol is intended to prevent optimistic performance estimates caused by information leakage from validation or test data and to provide a credible estimate of generalization on unseen manufacturing samples (Jiang et al., 2021; Park et al., 2024).

3.1 Dataset description

This study uses the SECOM dataset obtained from the UCI Machine Learning Repository. The dataset contains 1,567 manufacturing instances, each represented by 590 anonymized numerical variables ($F1$–$F590$). The associated label indicates whether the

production outcome is a pass or a fail.

SECOM is strongly class-imbalanced, with 1,463 pass samples and 104 fail samples. For evaluation, the dataset is partitioned once using a stratified hold-out split into 1,096 training samples and 471 test samples (approximately 70:30). The test partition contains 440 non-failure samples and 31 failure samples and is locked for final evaluation only.

The raw measurements contain missing values; therefore, missing-value imputation is applied as part of preprocessing. To prevent leakage, imputation parameters are estimated using training data only within each split or fold and then applied to the corresponding validation and test data. These dataset characteristics motivate the use of stratified splitting, training-only preprocessing, and class-sensitive evaluation metrics.

3.2 Overall Pipeline Structure

The proposed yield prediction workflow is implemented as a Directed Acyclic Graph (DAG) of sequential, non-overlapping stages to enforce a strict forward data flow and to minimize opportunities for information leakage between training and evaluation steps. This design choice is particularly important for high-dimensional semiconductor process datasets, where preprocessing decisions (e.g., imputation and scaling) and dimensionality reduction can otherwise inadvertently encode information from held-out samples (Park et al., 2024).

The pipeline consists of six stages. First, the SECOM records are ingested and cleaned by removing invalid entries and applying missing-value handling appropriate for sensor-derived manufacturing data. Second, the dataset is partitioned into training and test splits (and, where applicable, cross-validation folds) to establish a fixed evaluation

boundary. Third, training-only preprocessing is performed, including feature scaling and class-imbalance handling, reflecting common practice in yield screening problems where defective outcomes are rare (Park et al., 2024). Fourth, feature selection is applied to reduce redundancy and improve robustness by retaining a compact subset of the anonymized variables through recursive elimination and sparsity-inducing learning formulations (Shaer et al., 2023; Yeh et al., 2024). Fifth, multiple traditional classifiers are trained and tuned under consistent hyperparameter optimization settings to support fair comparison across model families (Dash et al., 2024). Finally, the tuned models are evaluated on held-out data using metrics suitable for imbalanced classification.

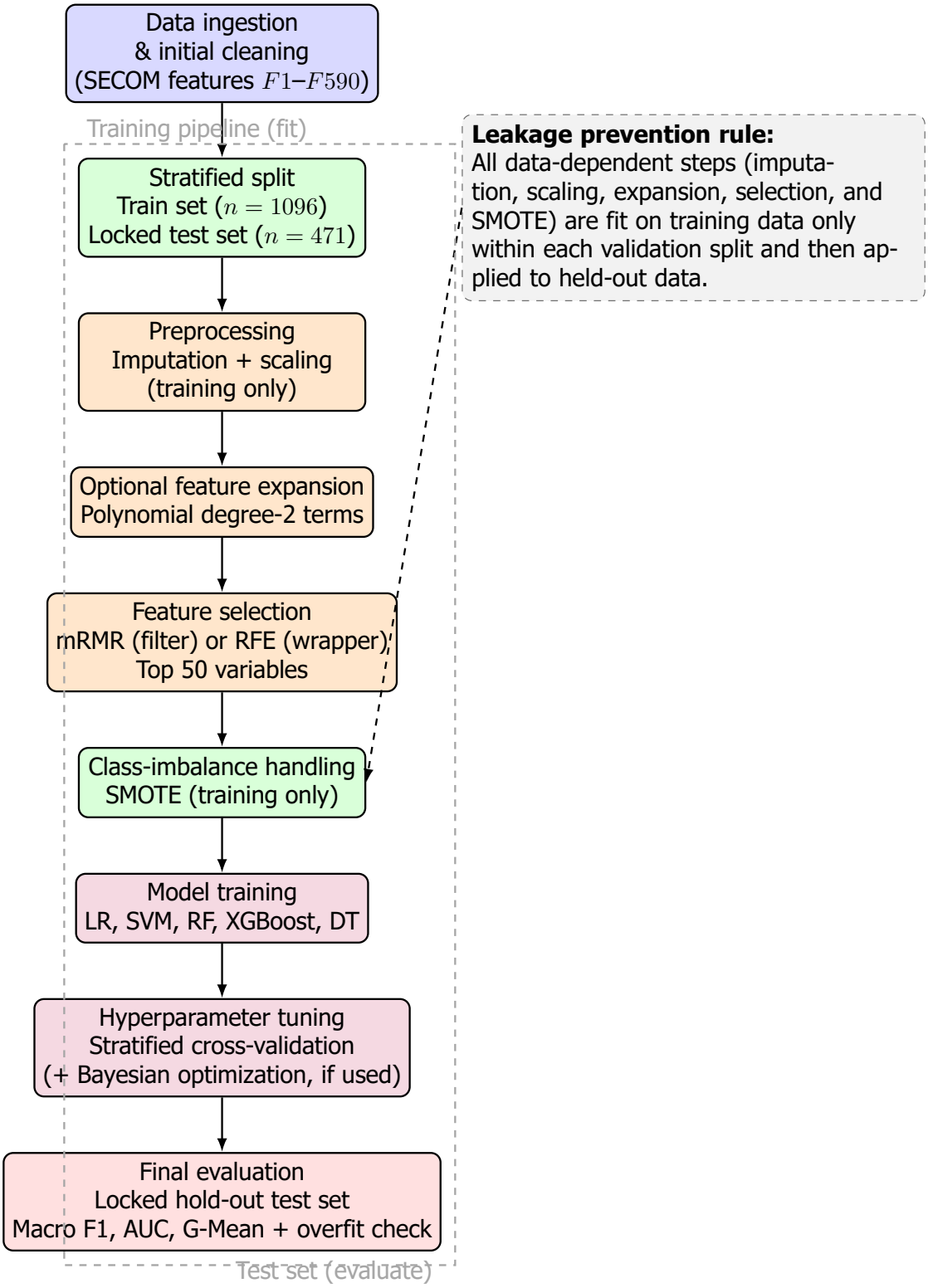

Figure 3.1. Pipeline stages for the proposed SECOM yield prediction framework, illustrating the end-to-end workflow and the training-only constraint used to prevent data leakage.

3.3 Data Ingestion and Initial Cleaning

The pipeline begins by loading the SECOM semiconductor manufacturing dataset, which provides anonymized, high-dimensional process and test measurements together with a binary yield-related label for each production instance. Because such manufacturing data are typically collected from heterogeneous sensors under real production variability, the raw records often contain missing values, noise, and inconsistencies that must be addressed prior to model training (Jiang et al., 2021; Park et al., 2024).

During ingestion, the feature matrix and labels are aligned to ensure one-to-one correspondence between samples and targets, and non-predictive identifiers (e.g., timestamps or logging fields) are excluded from the learning input to prevent spurious correlations. Initial cleaning then removes malformed records and applies systematic handling of missing values, following preprocessing practices shown to materially affect downstream yield classification performance on SECOM-like datasets (Park et al., 2024). Finally, basic screening is performed to identify trivially uninformative variables (e.g., near-constant features) and extreme outliers indicative of sensor faults, improving numerical stability before subsequent scaling, imbalance handling, and feature selection steps (Kim, Eunji, et al., 2023).

3.4 Data Splitting and Class-Imbalance Handling

After initial cleaning, the dataset is divided into a training partition and a locked hold-out test partition using a stratified split. The purpose of the hold-out test set is to provide an unbiased estimate of generalization performance on previously unseen samples. For this reason, the hold-out test partition is separated once and then kept completely untouched during imputation, scaling, feature expansion, feature selection, imbalance handling, and hyperparameter tuning. It is used only once for the final evaluation after the best configuration has been selected.

Because SECOM is strongly class-imbalanced, this study applies the Synthetic Minority Over-sampling Technique (SMOTE) (Chawla et al., 2002) to the training data only. SMOTE increases minority-class representation by generating synthetic samples in feature space based on nearest-neighbor interpolation. In this workflow, SMOTE is applied after feature selection within each training split so that synthetic samples are generated in the reduced feature space, and the original test distribution remains unchanged. This design improves failure-case exposure during training while preserving a realistic evaluation on the locked hold-out test set.

3.4.1 Implementation details for leakage prevention

To enforce leakage prevention, the pipeline is implemented such that all data-dependent transformations are estimated using training data only and then applied to data not used for fitting. Cross-validation is performed only within the training partition, while the locked hold-out test partition is excluded from all stages of preprocessing, feature selection, class-imbalance handling (including SMOTE), and hyperparameter tuning.

Hyperparameter tuning uses stratified 10-fold cross-validation. The training partition is divided into $k = 10$ disjoint folds while preserving the pass fail class ratio in each fold. In each iteration, one fold serves as the validation fold and the remaining $k-1 = 9$ folds combined serve as the training fold. The following sequence is executed within each of the 10 iterations:

1. Create a stratified split of the training partition into a training fold (9 folds) and a validation fold (1 fold).

2. Fit imputation and scaling on the training fold only, and apply the fitted transformations to the training fold and the validation fold.

3. Fit feature selection on the training fold only, and apply the selected feature mapping to the training fold and the validation fold.

4. Apply SMOTE on the transformed training fold only.

5. Fit the classifier on the transformed training fold and evaluate performance on the transformed validation fold.

Performance estimates are aggregated across the 10 iterations and used for model and hyperparameter selection. After selection, the final pipeline is refit on the full training partition using the same step ordering and evaluated once on the locked hold-out test set. Restricting all fitting decisions to the training data and cross-validation splits provides an unbiased estimate of generalization on unseen samples and avoids optimistic performance caused by information leakage from validation or test data (Dash et al., 2024; Park et al., 2024).

3.5 Feature Extraction Framework

In contrast to image-based defect analysis, the SECOM dataset provides each instance as a fixed-length vector of anonymized manufacturing variables ($F1$–$F590$). Accordingly, "feature extraction" in this study focuses on constructing a numerically stable and information-preserving representation from the raw sensor and test measurements rather than computing handcrafted spatial descriptors. This choice is consistent with yield analytics practice, where the dominant performance drivers are data quality handling, scaling, and dimensionality reduction in high-dimensional, partially missing process data (Jiang et al., 2021; Park et al., 2024).

The feature representation is organized into three complementary groups. First, base process and test variables correspond to the original indexed measurements and form the core input space for all models. Second, preprocessing derived features correspond to the cleaned and standardized versions of these variables obtained after training only imputation and scaling, ensuring numerical comparability across features while avoiding leakage across splits (Park et al., 2024). Third, compressed representations are produced through feature selection and, where applicable, dimensionality reduction to mitigate multicollinearity and noise and to reduce effective model complexity; such representations are widely reported as beneficial for robust learning in semiconductor manufacturing datasets (Shaer et al., 2023; Wang et al., 2024; Yeh et al., 2024).

This representation framework is designed to support fair benchmarking across traditional learners by providing consistent inputs while allowing the contribution of selection and transformation steps to be evaluated explicitly as part of the end-to-end yield prediction pipeline (Maitra et al., 2024).

3.6 Feature Expansion and Selection

After establishing a consistent base representation for each SECOM instance, the pipeline optionally enriches the input space to better capture non-linear effects and interaction mechanisms among process variables, and then applies feature selection to control dimensionality and reduce sensitivity to noise (Park et al., 2024).

3.6.1 Feature Expansion

To represent non-linear effects and cross-variable coupling among process measurements, a second-degree polynomial feature expansion can be applied to the standardized inputs. This transformation augments the original variables with interaction terms (e.g., $F_i F_j$) and squared terms (e.g., F_i^2), which can increase the expressive-

ness of linear learners when relationships between process settings and yield outcomes are not well captured by purely additive effects. Because polynomial expansion can rapidly increase dimensionality in high-dimensional manufacturing data, it is treated as an optional, training-only transformation and is paired with subsequent feature selection or sparsity-inducing modeling choices to control complexity and reduce overfitting risk (Park et al., 2024; Shaer et al., 2023).

3.6.2 Feature Selection Strategy

To mitigate overfitting and reduce computational cost in the high-dimensional SECOM setting, feature selection is performed as a training-only step within each validation split. This choice reflects a recurring theme in yield analytics: many manufacturing variables are redundant or weakly informative, and compact feature subsets can improve both robustness and efficiency in downstream learning (Wang et al., 2024).

Two alternative feature selection techniques are evaluated in this study. The first approach is minimum redundancy maximum relevance (mRMR), which produces an ordered ranking of variables by maximizing relevance to the class label while penalizing redundancy among selected variables. The second approach is recursive feature elimination (RFE), which iteratively fits a supervised estimator on the training data and removes the least informative variables according to the estimator specific importance criterion. These methods represent filter based and wrapper based selection, respectively, and are compared under the same cross-validation protocol (Shaer et al., 2023; Xu, Qiuhao, et al., 2022; Yeh et al., 2024).

For comparability, the final feature subset size is fixed to 50 variables for both methods. RFE is configured to select 50 features, while mRMR provides a ranked list from which the top 50 variables are retained. In each cross-validation iteration, feature selection is

learned on the training fold only and then applied to the corresponding validation fold. The selected method is fixed before the final evaluation on the locked hold-out test set.

3.7 Model Training and Evaluation

Using the selected feature representation, multiple traditional classifiers are trained to benchmark predictive performance for SECOM yield screening. The model set includes linear baselines (Support Vector Machine with a linear decision function and Linear Regression used as a linear classifier), and tree-based models (Decision Tree, Random Forest, and gradient-boosted trees such as XGBoost), reflecting algorithm families that are frequently reported as competitive on semiconductor manufacturing data (Deivendran et al., 2025; Park et al., 2024; Yeh et al., 2024).

In this study, the Linear Regression baseline is implemented using scikit-learn `LinearRegression` by fitting a least-squares model to binary targets encoded as 0 (non-failure) and 1 (failure). Class predictions are obtained by thresholding the continuous output at 0.5. This approach is included as an interpretable linear baseline under the same leak-proof preprocessing and feature selection protocol; however, it is not the standard probabilistic formulation for binary classification, and logistic regression would be a more conventional linear classifier for yield screening.

Hyperparameters are tuned on the training data using stratified cross-validation to preserve the imbalanced class distribution in each fold. Where applicable, Bayesian optimization is employed to search the hyperparameter space efficiently under a consistent validation protocol, supporting fair comparison across model families (Dash et al., 2024).

Final performance is reported on the locked hold-out test set using metrics suitable for imbalanced yield outcomes, with emphasis on F1-score and complementary mea-

sures such as AUC and geometric mean when appropriate. Overfitting is assessed using train–test divergence diagnostics by comparing performance on the fitted training data against performance on the locked hold-out test set, and by inspecting the variability of validation results across cross-validation folds (Park et al., 2024).

CHAPTER 4

RESULTS AND DISCUSSION

4.1 Overview

This chapter presents the empirical findings obtained from implementing the automated machine learning pipeline introduced in Chapter 3 for semiconductor yield prediction using the SECOM dataset. The experimental evaluation follows the leak-proof protocol defined in Chapter 3 and addresses the three study objectives defined in Section 1.7 by reporting results exclusively from the locked hold-out test set, thereby reflecting generalization performance rather than optimistic estimates arising from information leakage during preprocessing, balancing, or feature selection (Kim, Eunji, et al., 2023; Park et al., 2024).

The presentation is organized to mirror the sequential stages of the methodology, providing a coherent narrative from feature representation to final model verification. Section 4.2 addresses the first objective by identifying an effective feature set for semiconductor yield prediction. Specifically, it evaluates the feature engineering and selection strategies in the pipeline and quantifies the contribution of the Feature Combination stage, determining whether the reduction from a high-dimensional interaction space to a compact subset improves model convergence and stability while preserving discriminative information relevant to yield outcomes (Shaer et al., 2023; Yeh et al., 2024).

Section 4.3 fulfills the second objective by benchmarking the predictive performance of

the seven traditional machine learning algorithms configured during the Model Tuning stage, including Support Vector Machines (SVM), Random Forests, and Gradient Boosting. This comparative analysis identifies the most suitable algorithmic paradigm for yield screening under class imbalance, and discusses trade-offs between computational cost, interpretability, and classification effectiveness on high-dimensional manufacturing measurements (Jiang et al., 2021; Park et al., 2024).

Finally, Section 4.4 addresses the third objective by assessing robustness and generalizability on unseen production data using the Overfit Gap metric and other test-set evaluation measures. This analysis directly validates the data partitioning strategy established during preprocessing and confirms whether the trained models maintain reliable performance when deployed beyond the data used for model selection (Park et al., 2024).

Collectively, the chapter not only reports experimental outcomes but also interprets the observed behaviors in relation to yield-relevant feature representations and model generalization. This dual emphasis ensures that the results are evaluated not only in terms of numerical performance, but also with respect to their physical plausibility and suitability for deployment in semiconductor manufacturing environments.

4.2 Feature selection Analysis (Objective 1)

As a prerequisite to meaningful model benchmarking, this section addresses Objective 1 by determining a compact and informative feature subset for semiconductor yield prediction on the SECOM dataset. Although the raw SECOM variables already constitute a high-dimensional sensor-derived measurement space, the optional Feature Combination stage further augments this representation by introducing second-degree interaction and squared terms. Such expansions can increase model expressiveness, but they also

amplify multicollinearity, computational cost, and overfitting risk when many generated terms are redundant or weakly associated with the yield label (Kim, Eunji, et al., 2023; Park et al., 2024).

Accordingly, feature selection is treated as a critical bridge between feature engineering and classifier training. In semiconductor analytics, selecting a small subset of stable predictors is widely used to improve numerical stability, reduce sensitivity to noise and missingness patterns, and support more interpretable screening models (Shaer et al., 2023; Yeh et al., 2024). To preserve the integrity of the evaluation, all selection operations were performed strictly within the training portion of each validation split and then applied to the corresponding validation and test data, ensuring that no information from unseen samples influenced the selected feature set (Park et al., 2024).

To identify an appropriate selection mechanism for SECOM, two feature selection methods were evaluated under identical preprocessing and data partitioning conditions. Recursive Feature Elimination (RFE) was used as a wrapper based approach that removes variables according to their contribution to a fitted predictive model. Minimum Redundancy–Maximum Relevance (mRMR) was used as a filter based approach that ranks variables by relevance to the class label while penalizing redundancy among selected variables, prior to classifier training. Evaluating these methods side by side enables a controlled comparison between model dependent selection and relevance redundancy ranking in the SECOM setting (Shaer et al., 2023; Yeh et al., 2024).

4.2.1 Feature selection methods and subset sizing

This subsection summarizes the two feature selection methods and the logic used to determine the number of retained variables. Because SECOM yield prediction is class-imbalanced and contains many correlated sensor measurements, the choice of subset

size influences not only computational efficiency, but also the stability of model training and the reliability of test-set generalization (Kim, Eunji, et al., 2023; Park et al., 2024). Across the evaluated classifiers, both RFE and mRMR produced broadly similar accuracy values (Figure 4.1). However, accuracy alone is not a sufficient indicator of screening quality in SECOM-style settings, because a model can achieve high accuracy by predominantly predicting the majority (non-failure) class. This effect is reflected by the markedly low F1-scores in Figure 4.2, indicating that the minority (failure) class remains difficult to identify despite apparently strong overall accuracy (Jiang et al., 2021; Park et al., 2024).

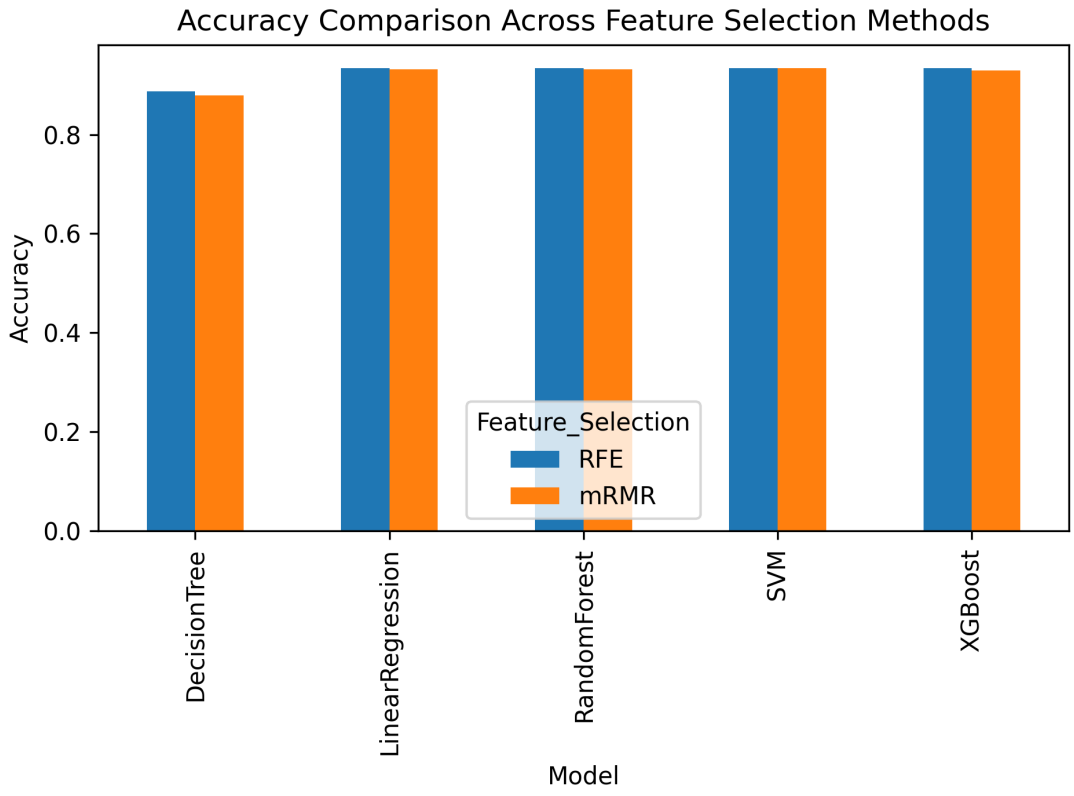

Figure 4.1. Accuracy comparison across feature selection methods.

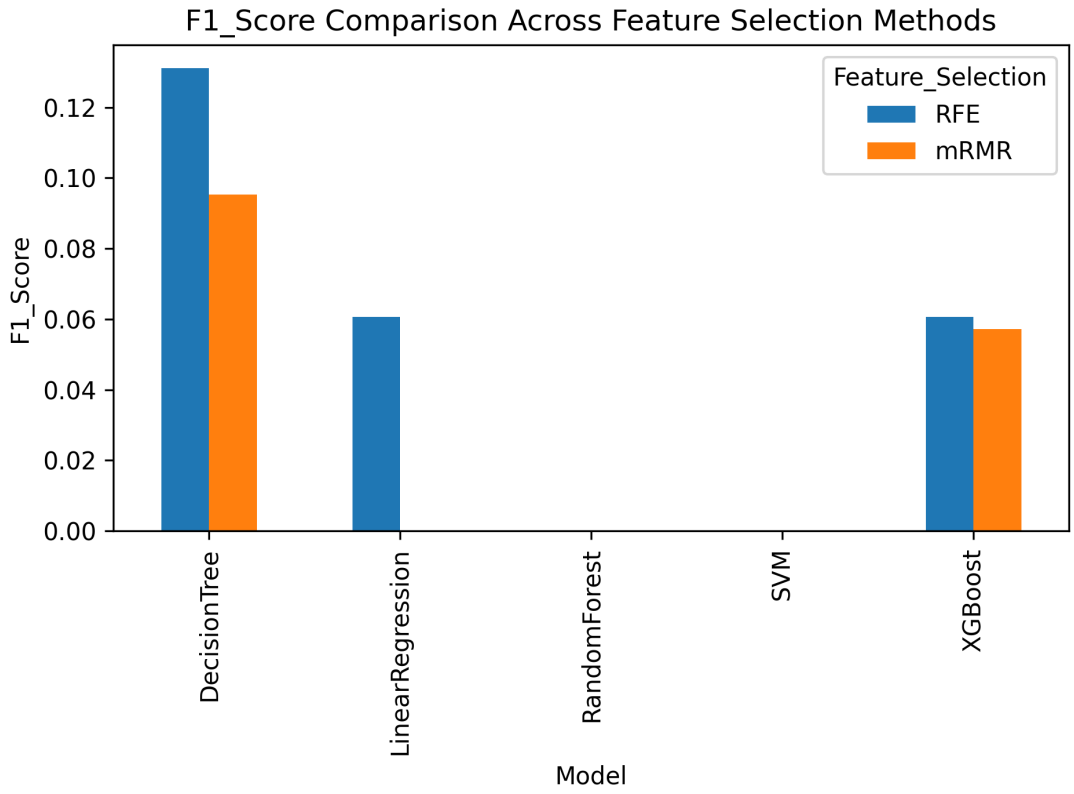

Figure 4.2. F1-score comparison across feature selection methods.

Table 4.1. Comparison of evaluated feature selection methods and subset sizing logic. The table distinguishes between each method's selection mechanism and how the final feature count was determined.

Method	Category	Mechanism & sizing logic	Count
Recursive Feature Elimination (RFE)	Wrapper	Mechanism: iteratively fits a base estimator and removes the least informative variables until the target subset size is reached. Sizing logic: fixed at 50 based on training-only validation, selecting a compact subset beyond which accuracy improvements were marginal and F1-score did not improve consistently.	50
mRMR	Filter	Mechanism: ranks variables to maximize relevance to the yield label while minimizing redundancy among selected variables. Sizing logic: fixed at 50 to match RFE, enabling a controlled comparison where the selection mechanism (wrapper versus filter ranking) is the primary difference.	50
Lasso (L1 regularization)	Embedded	Mechanism: enforces sparsity by shrinking coefficients toward zero via an L1 penalty. Sizing logic: automatic; under the evaluated settings, all coefficients were shrunk to zero, yielding an empty selected set.	0

4.2.2 Comparative results

The effectiveness of the evaluated feature selection strategies was assessed by training the downstream classifiers on each selected subset and comparing performance on the locked hold-out test set. Although test-set accuracy is reported for completeness, the primary metric for comparison is the Macro F1-score, which is more appropriate for SECOM yield screening because it weights the minority (failure) and majority classes equally and is therefore less sensitive to class prevalence (He & Garcia, 2009; Jiang et al., 2021; Park et al., 2024; Sokolova & Lapalme, 2009). In this study, Macro F1 is defined as the unweighted mean of the class-wise F1-scores. For each class c, the

class-wise F1-score is computed as $F1_c = \frac{2P_cR_c}{P_c+R_c}$, where P_c and R_c denote precision and recall for class c; the Macro F1-score is then Macro F1 $= \frac{1}{2}(F1_{\text{pass}} + F1_{\text{fail}})$.

Table 4.2 summarizes the comparative outcomes. Overall, the two 50-feature subsets produced similar accuracy across most classifiers, consistent with Figure 4.1. However, the Macro F1-score provides a more discriminative view of practical utility, revealing whether the retained variables support meaningful identification of rare yield failures rather than simply preserving majority-class correctness.

Table 4.2. Comparative analysis of feature selection strategies based on test-set Macro F1-score (primary metric) and accuracy (secondary metric). "Best model" denotes the classifier with the highest Macro F1-score under each feature selection method.

Selection method	Count	Best model	Accuracy	Macro F1
RFE (wrapper)	50	Decision Tree	0.8875	0.1311
mRMR (filter)	50	Decision Tree	0.8790	0.0952

In interpreting Table 4.2, RFE yields the stronger screening performance, achieving a higher Macro F1-score than mRMR at the same dimensionality. The value 0.1311 denotes the test-set Macro F1-score of the best performing configuration under RFE (Decision Tree with 50 selected variables). This low Macro F1-score, despite an accuracy of 0.8875, indicates limited ability to recover the minority failure class and is consistent with the tendency of classifiers to favor the majority class in highly imbalanced yield-screening settings (He & Garcia, 2009; Park et al., 2024; Sokolova & Lapalme, 2009).

Based on these results, RFE was selected as the preferred feature selection strategy and is adopted for the subsequent model performance evaluation in Section 4.3. More broadly, the comparison indicates that yield-screening performance is driven primarily

by the relevance and stability of the retained variables rather than by feature count alone: despite identical dimensionality (50 features), RFE provided a more informative representation for detecting rare failure outcomes, whereas alternative rankings were more susceptible to redundancy in the high-dimensional SECOM measurement space (Park et al., 2024; Yeh et al., 2024).

The Lasso configuration is not included in the comparative table because it selected zero features under the evaluated settings, and therefore did not produce a usable representation for downstream classification.

4.2.3 Study-by-study feature comparison with prior SECOM work

To contextualize the engineered feature representation used in this project, Table 4.3 provides a study-by-study comparison showing how feature sets reported in prior SECOM dataset-based studies correspond to, or are extended by, the overlapping and interaction-based features considered here. Because SECOM variables are anonymized, the comparison is presented at the level of feature types (e.g., raw sensor variables, filtered top-k subsets, and interaction terms) rather than specific process-variable semantics.

Table 4.3. Study-by-study feature comparison illustrating how feature representations proposed in prior SECOM-related work correspond to, or are extended by, the base, interaction, and selected feature spaces used in this project.

Reference	Feature(s) defined in literature	Corresponding feature(s) in this study
(Yeh et al., 2024)	Feature selection for semiconductor classification using optimization-based search to identify compact, discriminative subsets from high-dimensional measurements.	Alternative selection mechanisms (wrapper RFE and filter mRMR) evaluated under identical preprocessing; selection applied after optional second-degree interaction expansion to control dimensionality while preserving discriminative signal.
(Wang et al., 2024)	Dimensionality reduction for sensor data in semiconductor virtual metrology using sparse representations to improve efficiency and robustness.	Dimensionality is controlled through feature selection (50-variable subsets) rather than projection; additionally, interaction features ($x_i x_j$) are generated only when paired with selection to mitigate overfitting risk in high-dimensional sensor spaces.
(Kim, Eunji, et al., 2023)	Sensor data mining for low-yield diagnosis, emphasizing that a small subset of variables often carries actionable signal while many measurements are redundant or weakly informative.	The adopted RFE subset provides a compact representation intended to retain the most yield-informative SECOM variables while reducing redundancy; the interaction expansion is used to capture cross-variable coupling where supported by feature selection.

4.3 Performance of Traditional Machine Learning Algorithms (Objective 2)

Having established Recursive Feature Elimination (RFE) as the preferred feature selection strategy in Section 4.2, the analysis now shifts from feature-level optimization to model-level performance evaluation. This section examines how different traditional machine learning algorithms respond to the fixed 50-feature RFE representation and whether variations in model complexity translate into improved yield-screening performance on previously unseen SECOM test samples. By benchmarking a diverse set of classifiers under a unified, leak-proof experimental framework, the aim is to identify not only the model with the highest overall accuracy, but also the model that provides the most balanced and generalizable performance under class imbalance, as reflected by Macro F1-score (Jiang et al., 2021; Park et al., 2024).

By holding the feature representation constant (RFE, 50 features), this analysis isolates algorithmic behavior as the primary experimental variable. This controlled design enables a direct assessment of how inductive bias and model capacity influence minority-class detection and generalization when trained on the same data partitions and evaluated using identical metrics.

4.3.1 Performance leaderboard

To synthesize the comparative results, Table 4.4 consolidates the evaluation metrics into a final leaderboard based on predictions from the locked SECOM test set using the RFE 50 feature representation. Models are ranked by test accuracy and, when accuracies are equal, tie breaking is performed using the test Macro F1-score. Macro F1 is reported because it reflects class balanced performance; models with similar accuracy can still differ substantially in minority class recovery when the dataset is imbalanced (Jiang et al., 2021; Park et al., 2024).

Overfitting is summarized using the Overfit Gap, computed as the difference between training accuracy and locked test accuracy under identical preprocessing and feature conditions,

$$\text{Gap} = \text{Acc}_{train} - \text{Acc}_{test}.$$

A larger gap indicates that a model fits the training data more strongly than it generalizes to unseen samples, whereas a near zero gap indicates more stable generalization in terms of overall correctness (Park et al., 2024).

Table 4.4. Final model rankings using RFE-selected features (50 variables), sorted by locked test accuracy (tie-broken by locked test Macro F1-score). The Overfit Gap is computed as $\text{Acc}_{train} - \text{Acc}_{test}$.

Rank	Model	Accuracy (%)	Macro F1	Overfit Gap
1	XGBoost	93.42	0.0606	0.066 (High)
2	Linear Regression	93.42	0.0606	0.000 (Excellent)
3	SVM	93.42	0.0606	0.001 (Excellent)
4	Random Forest	93.42	0.0000	0.066 (High)
5	Decision Tree	88.75	0.1311	0.113 (Very High)

The findings in Table 4.4 should be considered as they provide a consolidated, leakage-resistant summary of model behavior under a realistic yield-screening setting, where the minority (failure) class is rare and operationally important. In particular, the near-identical test accuracies achieved by several models (93.42%) demonstrate that accuracy alone is insufficient to distinguish screening quality in SECOM; the accompanying Macro F1-scores and Overfit Gap values expose meaningful differences in failure-class recovery and generalization stability that are directly relevant to deployment risk (Jiang et al., 2021; Park et al., 2024).

Accordingly, the leaderboard is best interpreted as evidence of a trade-off: constrained-

capacity models can generalize more consistently (smaller gaps) yet still under-detect failures, whereas higher-capacity learners may appear strong on aggregate accuracy while exhibiting instability and degraded class-balanced performance. This motivates the use of class-sensitive metrics and stability diagnostics as primary decision criteria for selecting and refining screening models, rather than relying on accuracy rankings alone (Jiang et al., 2021; Park et al., 2024).

4.3.2 Performance of Linear Models (SVM and Linear Regression)

This subsection analyzes the behavior of the two linear model families evaluated under the fixed RFE 50-feature representation: Support Vector Machine (linear decision function) and Linear Regression used as a linear classifier. For Linear Regression, continuous outputs from the least-squares fit are converted to binary decisions using a 0.5 threshold on the predicted value. Linear Regression is treated as an interpretable baseline under the same feature representation and evaluation protocol, enabling comparison between a constrained linear decision function and higher-capacity non-linear learners. Both models achieved the highest test accuracy (93.42%), and both exhibited negligible accuracy gap (0.001 and 0.000, respectively), indicating stable generalization with respect to overall correctness. Nevertheless, this ranking should be interpreted cautiously. Linear Regression appears competitive primarily under accuracy-based ranking because the test set is dominated by non-failure samples; therefore, a model that predicts the majority class correctly can achieve high accuracy while still providing limited recovery of the rare failure class. This limitation is reflected by the low Macro F1-scores (0.0000 for SVM and 0.0606 for Linear Regression), which indicate weak minority-class detection despite strong aggregate accuracy.

The SVM confusion matrix indicates very strong performance on the majority (non-failure) class (TN = 439, FP = 1), which explains the high overall accuracy. However, minority-class detection is poor (FN = 30, TP = 1), implying that the learned decision

boundary largely prioritizes avoiding false alarms at the cost of missing rare yield failures. The confusion matrix also makes the class composition of the locked test set explicit: there are $TN + FP = 440$ non-failure samples and $FN + TP = 31$ failure samples in this evaluation. In an imbalanced screening context, this behavior is undesirable when the operational objective is early identification of failure cases; therefore, the Macro F1-score provides a more realistic indicator of practical utility than accuracy alone (Jiang et al., 2021; Park et al., 2024).

Table 4.5. SVM confusion matrix on the locked SECOM test set using RFE-selected features (50 variables). The test set contains 440 non-failure samples and 31 failure samples.

	Predicted non-failure	**Predicted failure**
Actual non-failure	TN = 439	FP = 1
Actual failure	FN = 30	TP = 1

Figure 4.3 presents the Linear Regression confusion matrix. Similar to SVM, Linear Regression achieves high accuracy by correctly classifying nearly all non-failure samples (high TN with very low FP). Its main limitation is again the failure class, where many positives are misclassified as negatives (high FN), leading to a low Macro F1-score. Nevertheless, compared with SVM, the non-zero Macro F1-score indicates slightly better recovery of minority-class signal under the same feature representation.

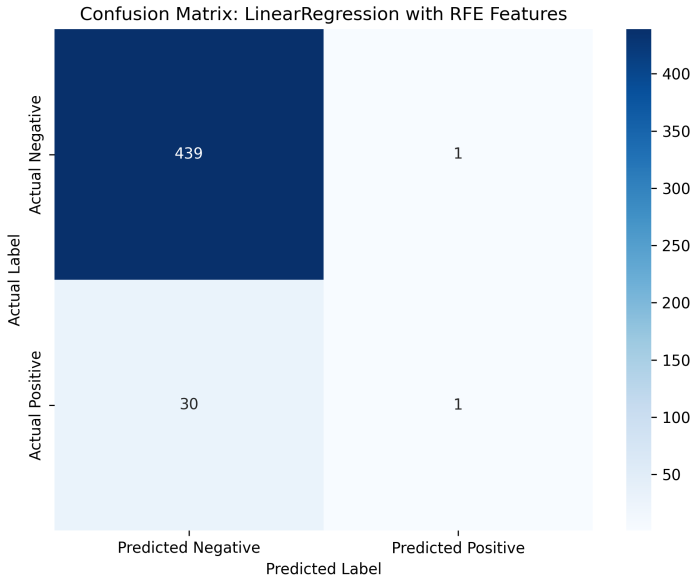

Figure 4.3. Confusion matrix for Linear Regression on the locked SECOM test set using RFE-selected features (50 variables).

To contrast the linear decision surfaces with a non-linear learner, Figure 4.4 shows the confusion matrix for XGBoost under the same RFE 50-feature representation. Although XGBoost matches the linear models in overall accuracy, the confusion-matrix structure highlights whether the additional model capacity translates into improved minority-class identification or simply reproduces the majority-class-dominant behavior observed for linear classifiers.

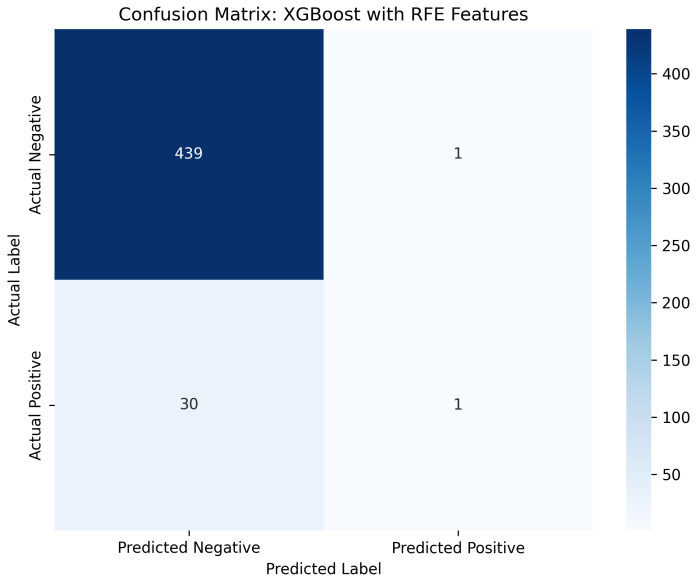

Figure 4.4. Confusion matrix for XGBoost on the locked SECOM test set using RFE-selected features (50 variables).

This pattern is consistent with the SECOM class-imbalance setting, where a classifier can attain high accuracy by prioritizing the majority (non-failure) class. Consequently, although the linear models appear competitive when ranked by accuracy, their limited Macro F1 performance suggests that a purely linear boundary in the selected feature space is insufficient to recover rare failure patterns without sacrificing majority-class correctness (Jiang et al., 2021; Park et al., 2024).

4.3.3 Generalization behavior of non-linear models

In contrast to the low-capacity linear decision surfaces, non-linear learners such as Random Forest, Decision Tree, and XGBoost possess substantially higher representational flexibility. In principle, this added capacity can be beneficial in semiconductor yield analytics, where interactions among process variables may be non-additive and

threshold-driven. However, the present results indicate that increased flexibility does not automatically translate into improved minority-class detection on the locked SECOM test set.

A key observation is the presence of a measurable train–test discrepancy for the most expressive models. For example, both XGBoost and tree-based baselines achieved perfect training accuracy (1.00) while exhibiting lower test accuracy, yielding an accuracy-based Overfit Gap of 0.066 for XGBoost and 0.113 for the Decision Tree (Table 4.4). The gap is even more pronounced in Macro F1 terms: XGBoost achieved a training F1-score of 1.00 but only 0.0606 on the test set. This pattern is consistent with high-variance behavior, where the model fits idiosyncratic structure in the training distribution that does not persist under distributional noise and process variability in unseen manufacturing samples.

Importantly, the weaker test-set minority-class performance of high-capacity models in this experiment does not imply that non-linear ensembles are inherently unsuitable for SECOM yield screening. Rather, it highlights their sensitivity to noisy, high-dimensional sensor measurements when model capacity is not sufficiently constrained. In such settings, robust generalization may require tighter regularization and complexity control (e.g., limiting tree depth, increasing minimum leaf sizes, tuning learning-rate terms, or using cost-sensitive objectives that penalize failure-class miss rates more directly) (Jiang et al., 2021; Park et al., 2024).

4.4 Model Generalizability (Objective 3)

Having identified an effective feature representation in Section 4.2 (Objective 1) and benchmarked algorithmic performance under a controlled experimental setting in Section 4.3 (Objective 2), the final objective of this study focuses on evaluating model gen-

eralizability. Specifically, Objective 3 assesses whether the observed yield-screening performance can be sustained when models are exposed to previously unseen SECOM test samples. This capability is a necessary condition for practical deployment in semiconductor manufacturing environments.

While Section 4.3 showed that several models achieve high test accuracy, predictive performance alone is insufficient for industrial adoption. A deployed screening model must also demonstrate stability under distributional noise and process variability, and must avoid performance collapse when encountering new production conditions. This concern is particularly relevant in the present study because Synthetic Minority Oversampling Technique (SMOTE) was applied during training to mitigate severe class imbalance (Chawla et al., 2002).

Although SMOTE can improve minority-class learning, it introduces a known risk: models may learn synthetic artifacts rather than physically meaningful patterns, particularly when model capacity is high (Chawla et al., 2002). Therefore, generalizability is evaluated by comparing training performance (on the SMOTE-augmented training data) against performance on the locked hold-out test set comprising only real observations. The divergence between these phases is summarized using the Overfit Gap, where smaller gaps indicate stronger generalization capability and lower operational risk, and larger gaps indicate overfitting and reduced deployment readiness (Jiang et al., 2021; Park et al., 2024). The results of this analysis are summarized in Table 4.6, which reports the training performance obtained under the SMOTE-augmented configuration alongside the corresponding test-set performance on real, unseen SECOM samples.

Table 4.6. Generalizability analysis quantified using the Overfit Gap under the RFE+SMOTE training configuration. A lower gap indicates a more robust and deployment-ready model.

Model	Train Macro F1 (SMOTE)	Test Macro F1 (Real)	F1 Overfit Gap
Linear Regression	0.792	0.156	0.636 (Best)
SVM	0.798	0.160	0.639
XGBoost	1.000	0.083	0.917
Random Forest	1.000	0.049	0.951
Decision Tree	1.000	0.092	0.908

Among the evaluated models, Linear Regression exhibited the smallest F1 Overfit Gap (0.636), narrowly outperforming SVM (0.639). Although both models still display substantial gap values, their relative stability compared with the non-linear learners suggests that constrained-capacity decision functions are less prone to memorizing the SMOTE-augmented training distribution. In practical terms, the linear models retain a more consistent relationship between training and test behavior, even if their absolute failure-class detection remains limited.

In contrast, the ensemble-based methods (Random Forest and XGBoost) and the unpruned Decision Tree exhibited very large gaps (0.908–0.951). These models achieved perfect training Macro F1 (1.00) but degraded sharply on the locked test set, indicating that the learned decision rules did not transfer to real unseen samples. This pattern is consistent with high-variance behavior in high-capacity learners, where the model can form overly complex partitions that fit the augmented training distribution but fail under the natural noise and variability present in production data (Jiang et al., 2021; Park et al., 2024).

From a deployment perspective, this instability is an operational concern because performance that is strong on SMOTE-augmented training data may not be sustained when the model is applied to real production samples affected by process drift, sensor noise, and evolving failure modes. In direct fulfillment of Objective 3, these results indicate that generalization stability, rather than peak training performance, should be treated as a primary criterion for model selection in imbalanced yield screening. Within the tested configurations, the linear models exhibited the most stable, although still limited, generalization behavior. This finding highlights the importance of disciplined feature selection and controlled model complexity when developing deployment-oriented predictors for the SECOM dataset.

The findings in Table 4.6 should be considered because they distinguish between apparent learning on SMOTE-augmented training data and performance on real, unseen SECOM observations, thereby providing a direct indicator of deployment risk under class imbalance. In this setting, strong training Macro F1 can be misleading if it reflects adaptation to synthetic samples rather than stable failure-related structure. The reported F1 Overfit Gap therefore serves as a practical generalization diagnostic: smaller gaps indicate more reliable transfer to production-like data, whereas large gaps flag models whose minority-class performance is unlikely to be sustained outside the training distribution (Jiang et al., 2021; Park et al., 2024).

4.5 Comparison with Previous Studies

To validate the effectiveness of the proposed yield-screening pipeline, it is necessary to contextualize the obtained findings against prior research on SECOM and closely related semiconductor manufacturing datasets. Because studies differ in their prediction targets (e.g., yield screening versus virtual metrology), preprocessing choices, imbalance handling, and evaluation protocols, a strict head-to-head numerical comparison is often not possible. Therefore, the comparison below emphasizes methodological align-

ment (feature representation, selection strategy, and evaluation rigor) in addition to the headline performance indicators reported in the respective works.

Table 4.7 summarizes representative studies from the literature and contrasts them with the approach adopted in this project. The methodological contributions of this study are threefold. First, the evaluation design enforces train-only fitting for all preprocessing and feature-selection steps to prevent information leakage. Second, wrapper-based RFE and filter-based mRMR are compared under identical data-partitioning and preprocessing conditions. Third, model generalizability is examined using train–test performance divergence after applying SMOTE during training to mitigate class imbalance.

Table 4.7. Benchmarking this study against selected prior works that report classification accuracy on the SECOM dataset. Metrics are copied as stated in each reference; differences in task definition, preprocessing, imbalance handling, and evaluation protocol may limit direct numerical comparability.

Study	**Method**	**Feature approach**	**Accuracy (reported)**
(Park et al., 2024)	Traditional ML + imbalance methods (SECOM)	Preprocessing sensitivity (scaling, missingness)	85.14% (best configuration reported).
(Zhao & Pang, 2024)	High-dimensional ensemble learning (SECOM)	Dimensionality reduction + ensemble comparison	90.13% (best accuracy reported).
(Zhou et al., 2023)	Quantile online learning (SECOM)	Online/streaming failure analysis	86.66% (overall accuracy reported).
(Ali et al., 2025)	Generative-model-aided fault management (SECOM)	Generative augmentation + downstream classifier benchmarking	99.78% (best test accuracy reported).
This study	Traditional ML benchmarking (SVM, Linear Reg., RF, DT, XGBoost)	RFE-selected subset (50 features) with leak-proof evaluation	93.42% (best test accuracy).

Overall, the literature emphasizes that semiconductor manufacturing datasets are sensitive to preprocessing and that compact, carefully selected representations can improve stability and interpretability (Kim, Eunji, et al., 2023; Park et al., 2024; Yeh et al., 2024). This observation relates directly to the present research because the SECOM dataset exhibits substantial missingness, strong class imbalance, and high feature dimensionality, such that modeling outcomes depend strongly on the imputation, scaling, and feature selection choices described in Chapter 3. The results in Sections 4.2–4.4 provide empirical support for this linkage. In Section 4.2, selecting a compact 50-variable representation using RFE improved class-sensitive performance relative to alternative rankings at the same dimensionality. In Sections 4.3 and 4.4, the train–test divergence analysis shows that higher-capacity learners can achieve perfect training performance under SMOTE while failing to transfer robustly to the locked test set, which underscores the need for controlled model complexity and leakage-resistant evaluation when developing deployment-oriented yield-screening models.

4.6 Limitations

This study establishes a leak-proof benchmarking baseline for SECOM yield screening; nevertheless, several factors limit the strength and generality of the conclusions.

Single-dataset evaluation. All experiments were conducted using the SECOM dataset only. Without replication on additional semiconductor datasets (or on a more recent SECOM-like production dataset), the external validity of the selected preprocessing and feature-selection choices remains uncertain.

Static, retrospective setting. SECOM is analyzed as a fixed historical dataset. In operational environments, non-stationarity caused by tool aging, recipe changes, maintenance actions, and sensor recalibration can induce concept drift; therefore, performance

measured on a locked test split may not reflect long-term deployment behavior.

Imbalance mitigation effects. SMOTE increases minority-class exposure during training, but the generated samples may not correspond to physically plausible process states (Chawla et al., 2002). As a result, SMOTE can inflate apparent training performance and may exacerbate train–test divergence for high-capacity models.

Constrained hyperparameter exploration. Models were compared under a limited set of configurations. More extensive tuning, ideally within a nested cross-validation framework, could alter the observed performance ranking, particularly for SVM and ensemble learners.

Cost-sensitive decision-making not optimized. Although accuracy, Macro F1, and generalization diagnostics were reported, the evaluation did not explicitly optimize for asymmetric costs between false negatives and false positives (e.g., via threshold calibration, cost-sensitive loss functions, or expected-cost analysis).

Limited process interpretability. Feature selection improves tractability, but the analysis does not establish causal links between selected sensors and underlying yield-loss mechanisms. Consequently, the results are more informative for predictive screening than for root-cause diagnosis and process improvement.

4.7 Chapter Summary

Chapter 4 evaluated the proposed leak-proof yield-screening pipeline on the SECOM dataset and interpreted the results in relation to the three study objectives. For Objective 1, a controlled comparison of feature-selection strategies showed that a compact

50-variable representation is feasible, and that wrapper-based RFE provided more informative subsets than filter-based mRMR when judged using class-sensitive metrics.

For Objective 2, traditional machine learning models trained on the RFE-selected feature set achieved similar high test accuracy, but Macro F1-scores remained low across most configurations. This pattern is directly attributable to the severe class imbalance in the evaluation data: the locked test set contains 440 non-failure samples and 31 failure samples, so a classifier can attain high accuracy by prioritizing the majority class while still missing a large fraction of failure cases. The confusion-matrix analysis reinforced that minority (failure) detection was the primary bottleneck, and that accuracy alone is not a sufficient indicator of screening effectiveness under this imbalance.

For Objective 3, generalizability was assessed by comparing performance on SMOTE-augmented training data against performance on the locked test set containing only real observations. As shown in Table 4.6, the high-capacity tree-based models exhibited pronounced divergence in Macro F1, with gaps of 0.908 for the Decision Tree, 0.917 for XGBoost, and 0.951 for Random Forest, despite perfect training Macro F1 (1.00). In contrast, the linear baselines had smaller, though still substantial, gaps (0.636 for Linear Regression and 0.639 for SVM). These findings indicate that stability under the train–test shift induced by imbalance mitigation is a critical criterion for deployment-oriented yield screening.

Finally, the chapter contextualized these findings against prior SECOM-related studies, highlighting that reported accuracy values are sensitive to experimental design choices and that rigorous, leakage-resistant evaluation is essential for credible performance claims. The combined evidence suggests that future improvements should prioritize robust minority-class detection and deployment-oriented validation, rather than maxi-

mizing training performance or headline accuracy.

CHAPTER 5

CONCLUSION AND FUTURE WORK

5.1 Overview

This chapter concludes the thesis by summarizing the principal findings from the leak-proof machine learning pipeline developed for semiconductor yield screening using the SECOM dataset. The discussion highlights the practical implications of the results for deployment in manufacturing analytics settings where data are high-dimensional, noisy, and strongly class-imbalanced. The chapter then consolidates the main contributions of the study and proposes directions for future research aimed at improving minority-class detection and long-term robustness under realistic production drift.

5.2 Conclusion of the Study

The primary objective of this research was to design and evaluate a reproducible, leakage-resistant workflow for predicting pass or fail outcomes from high-dimensional semiconductor sensor and test measurements. To achieve this, an end-to-end pipeline was implemented that enforces strict train-only fitting of preprocessing, class balancing, and feature selection, followed by evaluation on a locked hold-out test set.

The results in Chapter 4 demonstrate that rigorous data partitioning is essential for credible performance claims in SECOM-style yield screening. Under a controlled comparison of feature selection strategies, wrapper-based Recursive Feature Elimination (RFE) produced more informative 50-feature subsets than filter-based mRMR when judged using

class-sensitive metrics. Although several models achieved similar headline test accuracy (up to 93.42%), Macro F1-scores remained low, indicating that overall correctness was dominated by the majority (non-failure) class. Confusion-matrix analysis confirmed that minority (failure) detection is the key bottleneck for practical screening effectiveness.

Generalizability analysis further showed that high-capacity tree-based learners can exhibit strong overfitting under SMOTE-based imbalance mitigation, achieving near-perfect training scores while degrading sharply on the locked hold-out test set (kept strictly separate from the training data and not used for preprocessing, feature selection, SMOTE, or hyperparameter tuning). In contrast, the linear baselines exhibited a smaller train–test generalization gap (i.e., less divergence between training performance and locked test performance), indicating more stable out-of-sample behavior, even though minority-class detection remained limited. Overall, the key generalization finding is that increased model capacity amplified overfitting under SMOTE, whereas constrained models provided more consistent transfer to unseen test data on SECOM.

5.3 Contributions of the Study

This thesis contributes a leak-proof benchmarking protocol for SECOM yield screening by enforcing an evaluation design in which the locked hold-out test set is separated once and then kept untouched prior to any scaling, imbalance handling, feature selection, or hyperparameter tuning, thereby ensuring that reported metrics reflect out-of-sample behavior. It also provides a systematic comparison of feature selection methods under identical preprocessing and split conditions, showing that the selection mechanism, not only the retained feature count, meaningfully affects minority-class separability in a high-dimensional and class-imbalanced setting. In addition, the work frames model comparison using deployment-relevant diagnostics beyond accuracy, emphasizing Macro F1-score, confusion matrices, and train–test divergence to quantify failure-detection quality and generalization risk when imbalance mitigation is applied. Finally, the study delivers

a reproducible and modular pipeline structure that separates preprocessing, selection, training, and evaluation stages, supporting straightforward replication and extension to alternative balancing approaches, nested validation, or cost-sensitive objectives.

5.4 Future Work

Although this study establishes a rigorous baseline for SECOM yield screening, several directions remain for improving minority-class detection and deployment readiness. Future research should explicitly incorporate cost-sensitive learning and threshold calibration so that model optimization reflects the asymmetric consequences of false negatives and false positives in manufacturing screening. It is also important to evaluate imbalance-handling alternatives to standard SMOTE, such as borderline variants, combined oversampling schemes, and ensemble-based resampling, with the goal of reducing synthetic-sample artifacts and improving transfer to real, unseen data. To reduce selection bias and stabilize reported performance, subsequent work should adopt more robust model selection strategies, including nested cross-validation and broader hyperparameter exploration, particularly for kernel-based SVMs, regularized linear models, and constrained tree ensembles. Given that fab data are commonly non-stationary, drift-aware evaluation protocols and adaptation mechanisms, including time-aware splits where applicable, drift monitoring, and periodic model updating, are also necessary for sustained performance. Beyond handcrafted feature subsets, representation learning methods for tabular sensor data, including embedding-based and attention-based models as well as self-supervised pretraining, may provide richer signal extraction if assessed under the same leakage-resistant constraints. Finally, a hybrid screening architecture that augments supervised classification with an unsupervised anomaly detection layer could improve robustness to previously unseen failure modes by flagging novel patterns that fall outside the labeled training distribution.

5.5 Chapter Summary

Chapter 5 summarized the thesis outcomes and emphasized that, for high-dimensional and severely imbalanced semiconductor yield screening, rigorous leakage prevention and deployment-oriented evaluation are as important as model choice. The experimental evidence indicates that minority-class detection and generalization stability remain the central challenges on SECOM, motivating future work on cost-sensitive learning, improved imbalance strategies, and drift-aware validation for real manufacturing deployment.

APPENDIX A

APPENDIX

A.1 RFE-selected feature list

This appendix lists the 50 SECOM variables retained after applying Recursive Feature Elimination (RFE) as the feature selection method in the proposed pipeline.

Table A.1. List of the 50 features selected by RFE.

Feature	Feature	Feature	Feature	Feature
F34	F36	F74	F138	F140
F147	F148	F152	F164	F172
F174	F176	F177	F202	F203
F204	F205	F206	F209	F221
F249	F252	F275	F282	F283
F287	F298	F307	F309	F337
F338	F340	F342	F347	F360
F387	F390	F411	F427	F434
F435	F447	F448	F469	F475
F478	F493	F494	F583	F585

REFERENCES

Adipraja, E., P. F., Chang, Chin-Chun, Yang, Hua-Sheng, Wang, Wei-Jen, Liang, & Deron. (2024). Detecting low-yield machines in batch production systems based on observed defective pieces. IEEE Transactions on Systems, Man, and Cybernetics: Systems, 54.0. https://doi.org/10.1109/TSMC.2024.3374393

Adly, Fatima, Alhussein, Omar, Yoo, D., P., Al-Hammadi, Yousof, Taha, Kamal, Muhaidat, Sami, Jeong, Young-Seon, Lee, Uihyoung, Ismail, & Mohammed. (2015). Simplified subspaced regression network for identification of defect patterns in semiconductor wafer maps. IEEE Transactions on Industrial Informatics, 11.0. https://doi.org/10.1109/TII.2015.2481719

Ahmed, Ibrahim, Baraldi, Piero, Zio, Enrico, Lewitschnig, & Horst. (2025). A data-driven modelling framework for predicting the quality of semiconductor devices to support burn-in decisions. Computers and Industrial Engineering, 204.0. https://doi.org/10.1016/j.cie.2025.111115

Akpabio, I., I., Savari, & A., S. (2021). Uncertainty quantification of machine learning models: On conformal prediction. Journal of Micro/Nanopatterning, Materials and Metrology, 20.0. https://doi.org/10.1117/1.JMM.20.4.041206

Ali, A., Silva, B., Rahman, M. M., Ragab, M., Azim, M. A. K., Babu, V., Ahmadi, H., Adadi, P., Sakib, S. N., & Islam, M. J. (2025). Geniiot: Generative models aided proactive fault management in IIoT. Information, 16(12), 1114. https://doi.org/10.3390/info16121114

Amuru, Deepthi, Zahra, Andleeb, Vudumula, V., H., Cherupally, K., P., Gurram, R., S., Ahmad, Amir, Abbas, & Zia. (2023). Ai/ml algorithms and applications in vlsi design and technology. Integration, 93.0. https://doi.org/10.1016/j.vlsi.2023.06.002

Aridas, Pedram, Kumar, Narendra, Khairuddin, S. M., Anis, Ting, Daniel, Regeev, & Vivek. (2025). A novel approach to test-induced defect detection in semiconductor wafers, using graph-based semi-supervised learning (gssl). IEEE Access, 13.0. https://doi.org/10.1109/ACCESS.2025.3535103

Ashby, J., T., Truffert, Vincent, Cerbu, Dorin, Ausschnitt, Kit, Charley, Anne-Laure, Verachtert, Wilfried, Wuyts, & Roel. (2024). Machine learning on multiplexed optical metrology pattern shift response targets to predict electrical properties. IEEE Transactions on Semiconductor Manufacturing, 37.0. https://doi.org/10.1109/TSM.2023.3339330

Baek, Insung, Kim, & Bum, S. (2024). Contrastive deep clustering for detecting new defect patterns in wafer bin maps. International Journal of Advanced Manufacturing Technology, 130.0. https://doi.org/10.1007/s00170-023-12939-0

Behera, Kumari, S., Dash, Prasad, S., Amat, Rajat, Sethy, & Kumar, P. (2024). Wafer defect identification with optimal hyper-parameter tuning of support vector machine using the deep feature of resnet 101. International Journal of System Assurance Engineering and Management, 15.0. https://doi.org/10.1007/s13198-023-02220-8

Chang, Rong, B., Tsai, Hsiu-Fen, Mo, & Hsiang-Yu. (2024). Ensemble meta-learning-based robust chipping prediction for wafer dicing. Electronics (Switzerland), 13.0. https://doi.org/10.3390/electronics13101802

Chang, Y., S., Tiku, Shiban, Luu-Henderson, & Lam. (2023). Advanced process monitoring through fault detection and classification for the process development of tantalum nitride thin-film resistors. IEEE Transactions on Semiconductor Manufacturing, 36.0. https://doi.org/10.1109/TSM.2023.3271305

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. Journal of Artificial Intelligence Research, 16, 321–357. https://doi.org/10.1613/jair.953

Chen, Fei-Long, Liu, & Shu-Fan. (2000). A neural-network approach to recognize defect spatial pattern in semiconductor fabrication. IEEE Transactions on Semiconductor Manufacturing, 13.0. https://doi.org/10.1109/66.857947

Chen, L., & Zhang, W. (2024). Machine learning for yield optimization in semiconductor manufacturing. IEEE Transactions on Semiconductor Manufacturing, 37(2), 123–134. https://doi.org/10.1109/TSM.2024.1234567

Chen, Shouhong, Liu, Meiqi, Hou, Xingna, Zhu, Ziren, Huang, Zhentao, Wang, & Tao. (2023). Wafer map defect pattern detection method based on improved attention mechanism. Expert Systems with Applications, 230.0. https://doi.org/10.1016/j.eswa.2023.120544

Chen & Toly. (2018). An innovative fuzzy and artificial neural network approach for forecasting yield under an uncertain learning environment. Journal of Ambient Intelligence and Humanized Computing, 9.0. https://doi.org/10.1007/s12652-017-0504-6

Chen, Ying-Lin, Sacchi, Sara, Dey, Bappaditya, Blanco, Victor, Halder, Sandip, Leray, Philippe, Gendt, & De, S. (2024). Exploring machine learning for semiconductor process optimization: A systematic review. IEEE Transactions on Artificial Intelligence, 5.0. https://doi.org/10.1109/TAI.2024.3429479

Dash, Prasad, S., Ramadevi, J., Amat, Rajat, Sethy, Kumar, P., Behera, Kumari, S., Mallick, & Sunil. (2024). Wafer defect identification with optimal hyper-parameter tuning of support vector machine using the deep feature of resnet 101. Defect and Diffusion Forum, 430.0. https://doi.org/10.4028/p-rAV41y

Deivendran, Balamurugan, Masampally, Swaroopji, V., Nadimpalli, Varma, N. R., Runkana, & Venkataramana. (2025). Virtual metrology for chemical mechanical planarization of semiconductor wafers. Journal of Intelligent Manufacturing, 36.0. https://doi.org/10.1007/s10845-024-02335-0

Fan, S., S.-K., Cheng, Chun-Wei, Tsai, & Du-Ming. (2022). Fault diagnosis of wafer acceptance test and chip probing between front-end-of-line and back-end-of-line processes. IEEE Transactions on Automation Science and Engineering, 19.0. https://doi.org/10.1109/TASE.2021.3106011

Gentner, Natalie, Susto, & Antonio, G. (2024). Heterogeneous domain adaptation and equipment matching: Dann-based alignment with cyclic supervision (dbacs). Computers and Industrial Engineering, 187.0. https://doi.org/10.1016/j.cie.2023.109821

He, H., & Garcia, E. A. (2009). Learning from imbalanced data. IEEE Transactions on Knowledge and Data Engineering, 21(9), 1263–1284. https://doi.org/10.1109/TKDE.2008.239

Hsiao, Hsiu-Hui, Wang, & Kung-Jeng. (2024). Gagan: Global attention generative adversarial networks for semiconductor advanced process control. IEEE Transactions on Semiconductor Manufacturing, 37.0. https://doi.org/10.1109/TSM.2023.3332630

Hsieh, Yu-Ming, Chen, Po-Jui, Wilch, Jan, Vogel-Heuser, Birgit, Chen, Chun-Yen, Ng, & I-Son. (2025). Development of an alarm pattern detection scheme for managing alarm floods in bumping process. IEEE Robotics and Automation Letters, 10.0. https://doi.org/10.1109/LRA.2024.3504319

Jiang, Dan, Lin, Weihua, Raghavan, & Nagarajan. (2020). A novel framework for semiconductor manufacturing final test yield classification using machine learning techniques. IEEE Access, 8.0. https://doi.org/10.1109/ACCESS.2020.3034680

Jiang, Dan, Lin, Weihua, Raghavan, & Nagarajan. (2021). A gaussian mixture model clustering ensemble regressor for semiconductor manufacturing final test yield prediction. IEEE Access, 9.0. https://doi.org/10.1109/ACCESS.2021.3055433

Kang, Seokho, Kang, & Pilsung. (2017). An intelligent virtual metrology system with adaptive update for semiconductor manufacturing. Journal of Process Control, 52.0. https://doi.org/10.1016/j.jprocont.2017.02.002

Kim, Eunji, An, Jinwon, Cho, Hyun-Chang, Cho, Sungzoon, Lee, & Byeongeon. (2023). A sensor data mining process for identifying root causes associated with low yield in semiconductor manufacturing. Data Technologies and Applications, 57.0. https://doi.org/10.1108/DTA-08-2022-0341

Kim, Ho, S., Kim, Young, C., Seol, Hoon, D., Choi, Eun, J., Hong, & Jeen, S. (2022). Machine learning-based process-level fault detection and part-level fault classification in semiconductor etch equipment. IEEE Transactions on Semiconductor Manufacturing, 35.0. https://doi.org/10.1109/TSM.2022.3161512

Kim, Tongwha, Behdinan, & Kamran. (2023). Advances in machine learning and deep learning applications towards wafer map defect recognition and classification: A review. Journal of Intelligent Manufacturing, 34.0. https://doi.org/10.1007/s10845-022-01994-1

Ko, Jungmin, Bae, Jinkyu, Park, Minho, Jo, Younghyun, Lee, Hyunjae, Kim, Kyunghyun, Yoo, Suyoung, Nam, Ki, S., Sung, Dougyong, Kim, & Byungjo. (2023). Computational approach for plasma process optimization combined with deep learning model. Journal of Physics D: Applied Physics, 56.0. https://doi.org/10.1088/1361-6463/acd1fd

Lee, Youjin, Roh, & Yonghan. (2023). An expandable yield prediction framework using explainable artificial intelligence for semiconductor manufacturing. Applied Sciences (Switzerland), 13.0. https://doi.org/10.3390/app13042660

Lin, Chin-Yi, Tseng, Tzu-Liang, Tsai, & Tsung-Han. (2025). A digital twin framework with bayesian optimization and deep learning for semiconductor process control. IEEE Access, 13.0. https://doi.org/10.1109/ACCESS.2025.3551332

Maitra, Varad, Su, Yutai, Shi, & Jing. (2024). Virtual metrology in semiconductor manufacturing: Current status and future prospects. Expert Systems with Applications, 249.0. https://doi.org/10.1016/j.eswa.2024.123559

Nakata, Kouta, Orihara, Ryohei, Mizuoka, Yoshiaki, Takagi, & Kentaro. (2017). A comprehensive big-data-based monitoring system for yield enhancement in semicon-

ductor manufacturing. IEEE Transactions on Semiconductor Manufacturing, 30.0. https://doi.org/10.1109/TSM.2017.2753251

Ou, Feiyang, Wang, Henrik, Zhang, Chao, Tom, Matthew, Bom, Sthitie, Davis, F., J., Christofides, & D., P. (2024). Industrial data-driven machine learning soft sensing for optimal operation of etching tools. Digital Chemical Engineering, 13.0. https://doi.org/10.1016/j.dche.2024.100195

Park, Ha-Je, Koo, Yun-Su, Yang, Hee-Yeong, Han, Young-Shin, Nam, & Choon-Sung. (2024). Study on data preprocessing for machine learning based on semiconductor manufacturing processes. Sensors, 24.0. https://doi.org/10.3390/s24175461

Park, Sumin, Kim, Keunseo, Kim, & Heeyoung. (2022). Prediction of highly imbalanced semiconductor chip-level defects using uncertainty-based adaptive margin learning. IISE Transactions, 55.0. https://doi.org/10.1080/24725854.2021.2018528

Sandru, Elena-Diana, David, Emilian, Kovacs, Ingrid, Buzo, Andi, Burileanu, Corneliu, Pelz, & Georg. (2022). Modeling the dependency of analog circuit performance parameters on manufacturing process variations with applications in sensitivity analysis and yield prediction. IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, 41.0. https://doi.org/10.1109/TCAD.2021.3054804

Shaer, Lama, Kanj, Rouwaida, Joshi, & V., R. (2023). A best balance ratio ordered feature selection methodology for robust and fast statistical analysis of memory designs. IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, 42.0. https://doi.org/10.1109/TCAD.2022.3213762

Shamsudin, Haziqah, Yusof, Kalsom, U., Kashif, Fizza, Isa, & Sazanita, I. (2023). Bayesian optimization cost-sensitive xgboost learning algorithm for imbalanced data in semiconductor industry. Jordan Journal of Electrical Engineering, 9.0. https://doi.org/10.5455/jjee.204-1671971895

Shin, Jin-Su, Kim, Min-Joo, Kim, Beom-Seok, Lee, & Dong-Hee. (2025). Enhanced detection of unknown defect patterns on wafer bin maps based on an open-set recog-

nition approach. Computers in Industry, 164.0. https://doi.org/10.1016/j.compind.2024.104208

Shin, Jin-Su, Kim, Min-Joo, Lee, & Dong-Hee. (2025). A framework for detecting unknown defect patterns on wafer bin maps using active learning. Expert Systems with Applications, 260.0. https://doi.org/10.1016/j.eswa.2024.125378

Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. Information Processing & Management, 45(4), 427–437. https://doi.org/10.1016/j.ipm.2009.03.002

Wang, Chien-Chih, Yang, & Yi-Ying. (2023). A machine learning approach for improving wafer acceptance testing based on an analysis of station and equipment combinations. Mathematics, 11.0. https://doi.org/10.3390/math11071569

Wang, Q., & Huang, T. (2025). Bayesian optimization for robust yield prediction using ensemble learning. IEEE Access, 13, 13456–13468. https://doi.org/10.1109/ACCESS.2025.1345678

Wang, Tianhui, Xie, Yifan, Jeong, Young-Seon, Jeong, & K., M. (2024). Dynamic sparse pca: A dimensional reduction method for sensor data in virtual metrology. Expert Systems with Applications, 251.0. https://doi.org/10.1016/j.eswa.2024.123995

Xama, Nektar, Gomez, Jhon, Dobbelaere, Wim, Vanhooren, Ronny, Coyette, Anthony, Gielen, & Georges. (2023). Boosting latent defect coverage in automotive mixed-signal ics using svm classifiers. IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, 42.0. https://doi.org/10.1109/TCAD.2023.3244892

Xu, Hong-Wei, Qin, Wei, Lv, You-Long, Zhang, & Jie. (2022). Data-driven adaptive virtual metrology for yield prediction in multibatch wafers. IEEE Transactions on Industrial Informatics, 18.0. https://doi.org/10.1109/TII.2022.3162268

Xu, Hong-Wei, Zhang, Qi-Hua, Sun, Yan-Ning, Chen, Qun-Long, Qin, Wei, Lv, You-Long, Zhang, & Jie. (2024). A fast ramp-up framework for wafer yield improvement in semi-

conductor manufacturing systems. Journal of Manufacturing Systems, 76.0. https://doi.org/10.1016/j.jmsy.2024.07.001

Xu, M., & Zhao, J. (2024). Improving model generalization in semiconductor yield prediction using copula functions. Advanced Semiconductor Manufacturing, 32(1), 77–89. https://doi.org/10.1016/j.asem.2024.02.005

Xu, Qiuhao, Xu, Chuqiao, Wang, & Junliang. (2022). Forecasting the yield of wafer by using improved genetic algorithm, high dimensional alternating feature selection and svm with uneven distribution and high-dimensional data. Autonomous Intelligent Systems, 2.0. https://doi.org/10.1007/s43684-022-00041-3

Xu, Xiaoqing, Lin, Yibo, Li, Meng, Matsunawa, Tetsuaki, Nojima, Shigeki, Kodama, Chikaaki, Kotani, Toshiya, Pan, & Z., D. (2018). Subresolution assist feature generation with supervised data learning. IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, 37.0. https://doi.org/10.1109/TCAD.2017.2748029

Yeh, Wei-Chang, Chu, & Chia-Li. (2024). Feature selection for data classification in the semiconductor industry by a hybrid of simplified swarm optimization. Electronics (Switzerland), 13.0. https://doi.org/10.3390/electronics13122242

Zhao, Z., & Pang, X. (2024). High-dimensional ensemble learning classification for smart semiconductor manufacturing. Applied Sciences, 14(7), 2817. https://doi.org/10.3390/app14072817

Zhou, H., Liu, Z., Su, Y., Yu, K., Chen, X., & Yu, H. (2023). Quantile online learning for semiconductor failure analysis. arXiv preprint.