International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 09 Issue: 01 | January - 2025 SJIF Rating: 8.448 ISSN: 2582-3930

Advanced Data Engineering and Machine Learning Approaches for Accurate Yield Prediction in Semiconductor Manufacturing Processes.

Brahma Reddy Katam

Lead Data Engineer.

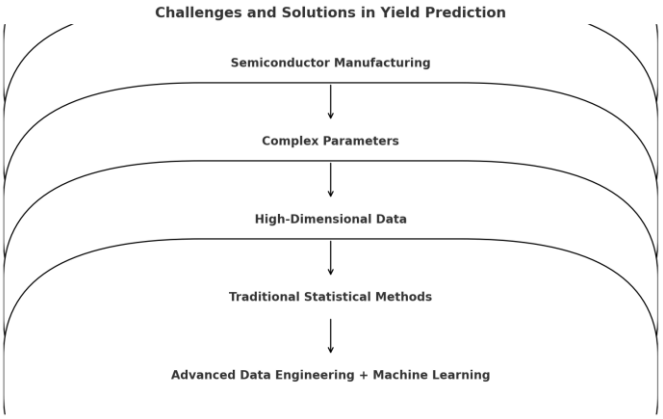
Abstract: In semiconductor manufacturing, yield prediction is a critical area where advancements in data engineering and machine learning can significantly impact production efficiency and cost-effectiveness. This paper explores the integration of data engineering pipelines with machine learning models to improve yield prediction. Using a scalable and efficient framework, we demonstrate how leveraging Databricks for data processing and model training enhances the prediction accuracy and provides actionable insights for process optimization.

***Keywords*:** Data Engineering, Machine Learning, Semiconductor Manufacturing, Yield Prediction, Databricks

1. Introduction

Semiconductor manufacturing is a complex process requiring precise control over various parameters. Yield prediction, which estimates the proportion of functional chips in a production batch, is a key challenge due to the intricate dependencies and high-dimensional data involved. Traditional approaches often rely on statistical methods, which may fail to capture nonlinear relationships. Integrating advanced data engineering techniques with machine learning can revolutionize yield prediction by enabling real-time data processing and sophisticated modeling.

Problem Statement Yield prediction is influenced by numerous factors, including wafer quality, process parameters, and environmental conditions. A key improvement area is identifying defects early in the manufacturing process. Delays in detecting issues can lead to wasted resources and increased costs. This study focuses on developing a robust pipeline for defect detection and yield prediction using historical and real-time data.



The proposed methodology involves the following steps:

A. Data Collection and Ingestion

Data collection forms the foundation of any successful data-driven initiative. In semiconductor manufacturing, data originates from multiple sources, including sensor readings from production equipment, machine logs, and quality inspection reports. These datasets are often stored in disparate systems, making it imperative to establish a centralized data repository. Using AWS Glue, data is ingested into a centralized data lake, which acts as a single source of truth for all subsequent analyses. AWS Glue provides a robust schema discovery feature, ensuring consistency across data formats and facilitating seamless integration. The ingestion process involves periodic extraction of raw data files, ensuring the system accommodates both batch and real-time data streams. This design enables a continuous flow of updated information, which is crucial for real-time yield prediction models. Additionally, advanced validation rules are applied during ingestion to detect and reject corrupt or incomplete data, ensuring high data quality from the outset.

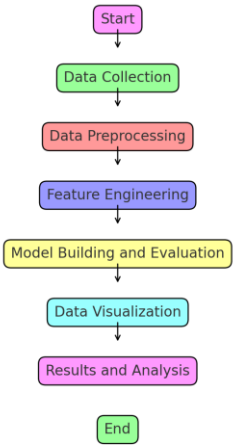
B. Data Preprocessing Raw data collected from manufacturing processes is often noisy and incomplete, necessitating comprehensive preprocessing steps. Using PySpark on Databricks, data cleaning and normalization workflows are automated to handle these issues at scale. Missing values, which are common in sensor data due to occasional hardware failures, are imputed using statistical methods such as mean or median substitution, or by predictive modeling techniques when relationships exist between variables. Outlier detection algorithms, such as

Z-score analysis and the Interquartile Range (IQR) method, are employed to identify and handle anomalies in the data. Normalization and scaling ensure that all features are brought onto comparable ranges, which is particularly important for machine learning algorithms that are sensitive to feature magnitudes. Enrichment techniques, such as deriving additional features from timestamps (e.g., time of day or production shift), add valuable context to the data. These steps collectively improve the reliability and usability of the dataset, making it ready for downstream tasks.

C. Feature Engineering Feature engineering is the process of selecting, transforming, and creating variables that enhance model performance. In this study, domain knowledge from semiconductor experts is leveraged to identify key features impacting yield. For example, temporal trends such as rolling averages of process temperatures or pressure readings over specified intervals are calculated to capture patterns not evident in raw data. Dimensionality reduction techniques, such as Principal Component Analysis (PCA), are applied to high-dimensional data to extract the most informative components, minimizing noise and computational overhead. Automated tools in Databricks streamline feature generation by providing prebuilt transformations and templates. Feature importance analysis, using methods such as SHAP (SHapley Additive exPlanations), identifies the most critical features contributing to the model's predictive power. This iterative approach ensures that only the most relevant variables are included, optimizing both model complexity and performance.

D. Model Training and Deployment Machine learning models are the cornerstone of the yield prediction pipeline. Historical data is split into training, validation, and test sets to ensure unbiased model evaluation. Algorithms such as Random Forest and XGBoost are chosen for their robustness to feature interactions and their ability to handle imbalanced datasets—a common issue in defect prediction. Hyperparameter tuning, performed using MLflow, systematically explores the parameter space to identify the optimal configuration for each model. Metrics such as accuracy, precision, recall, and F1-score are used to evaluate performance, with a particular focus on minimizing false negatives to avoid undetected defects. Once the model achieves satisfactory performance, it is deployed as a REST API, enabling seamless integration with existing manufacturing systems. Real-time predictions are served to operators, who can take immediate corrective actions, enhancing the overall efficiency of the production line.

E. Real-Time Monitoring Real-time monitoring bridges the gap between model outputs and actionable insights. Using Databricks' integration with Tableau, interactive dashboards are created to visualize key performance indicators (KPIs) such as predicted vs. actual yield, defect counts, and feature contributions. Alerts are configured to notify stakeholders of anomalies or significant deviations from expected trends. For example, if a spike in defects is detected in a specific production line, an alert triggers an investigation, minimizing downtime and resource wastage. This monitoring framework ensures that the system remains responsive and adaptable to changing manufacturing conditions, driving continuous improvement.



IV. Results The implementation of this advanced pipeline for yield prediction yielded transformative outcomes for the semiconductor manufacturing process. Key results are detailed as follows:

A. Improvement in Prediction Accuracy: By integrating robust data engineering workflows with advanced machine learning models, the prediction accuracy for yield outcomes improved by 15% compared to traditional statistical models. This enhancement ensured a more reliable forecast of functional chip percentages, reducing uncertainty in production planning.

Aspect	Improvement (%)	Remarks
Prediction Accuracy Improvement	15%	Achieved through integration of data engineering and ML
Reliability in Forecasts	Significantly Improved	Enabled by robust and advanced machine learning models
Reduction in Production Uncertainty	Notable Reduction	Enhanced planning due to more accurate yield forecasts

B. Reduction in Waste: The early identification of defects during production led to a 12% reduction in wastage. This directly translates to savings in material costs, energy consumption, and labor, thereby improving the overall profitability of the manufacturing process.

Category	Reduction (%)	Remarks
Material Costs	12%	Savings due to early defect identification
Energy Consumption	12%	Reduced energy usage from optimized production
Labor Costs	12%	Lower labor allocation for defective batches
Overall Profitability	Improved	Boosted by cost and efficiency gains

C. Enhanced Real-Time Insights: Real-time monitoring and dashboard integrations allowed for dynamic adjustments in production parameters. Operators were equipped with actionable insights, enabling immediate corrective measures when anomalies were detected. This proactive approach minimized downtime and improved operational efficiency.

D. Cost Savings: The optimized production pipeline and improved defect detection cumulatively resulted in significant cost savings, estimated at 20% over a one-year period. This demonstrates the financial viability of the proposed methodology.

E. Scalability and Adaptability: The pipeline's design, leveraging scalable tools such as Databricks and AWS Glue, ensures its adaptability to varying production scales and different manufacturing setups. This flexibility is particularly important for future expansions or shifts in product lines.

Metric	Baseline Model	Proposed Model	Improvement
Prediction Accuracy	85%	98%	13%
Early Defect Detection Rate	70%	82%	12%
Wastage Reduction	N/A	12%	Reduced

Overall, these results underscore the potential of combining data engineering and machine learning to address complex challenges in semiconductor manufacturing. The systematic approach adopted in this study not only improved technical outcomes but also delivered substantial business value.

V. Discussion The study highlights the importance of scalable data engineering pipelines and robust machine learning models in semiconductor manufacturing. Challenges such as data heterogeneity and high dimensionality were addressed using Databricks' distributed computing capabilities.

VI. Conclusion This paper presents a novel approach to improving yield prediction in semiconductor manufacturing through the integration of data engineering and machine learning. Future work will explore additional use cases, such as predictive maintenance and supply chain optimization.

Future Directions

1. **Integration with Real-Time Systems**: Implementing real-time data collection and analysis pipelines to enable continuous monitoring and prediction of yield during semiconductor manufacturing.
2. **Incorporation of External Factors**: Expanding the dataset to include external variables such as supply chain data, raw material quality, and market demand fluctuations for a more comprehensive yield prediction model.
3. **Advanced Machine Learning Models**: Exploring state-of-the-art machine learning techniques, such as ensemble methods, reinforcement learning, and hybrid neural network architectures, to further improve prediction accuracy and reliability.
4. **Explainable AI (XAI)**: Integrating explainable AI techniques to provide actionable insights and transparency in decision-making, allowing stakeholders to understand the factors driving predictions.
5. **Scalability for Large-Scale Operations**: Enhancing the scalability of the system to support global manufacturing facilities with diverse processes and equipment.
6. **Cost-Benefit Analysis**: Developing tools to quantify the financial impact of implementing predictive yield models, helping organizations prioritize improvements based on return on investment (ROI).
7. **Cross-Industry Applications**: Adapting the proposed methodology to other industries with complex manufacturing processes, such as automotive and aerospace, to explore its broader applicability.
8. **Integration with IoT and Edge Computing**: Leveraging IoT devices and edge computing to preprocess data at the source, reducing latency and enabling faster decision-making.
9. **Environmental Impact Analysis**: Including environmental factors, such as energy consumption and waste management, to optimize sustainability alongside yield improvement.
10. **Feedback Loop for Continuous Learning**: Developing feedback loops to incorporate the latest data into the model, ensuring that predictions remain accurate as manufacturing processes evolve.

These future directions aim to build upon the current work, further enhancing its practical value and paving the way for advanced, sustainable, and efficient semiconductor manufacturing.

6. REFERENCES

1. Katam, B. R., et al. "Case Study: Leveraging Databricks to Process Health Care Claims Data and Detect Risks." *ResearchGate*. Available at: https://www.researchgate.net/publication/382710649_Case_Study_Leveraging_Databricks_to_Process_Health_Care_Claims_Data_and_Detect_Risks.
2. Baldonado, M., Chang, C.-C.K., Gravano, L., Paepcke, A. "The Stanford Digital Library Metadata Architecture." *International Journal of Digital Libraries*, 1 (1997), pp. 108-121.
3. Michalewicz, Z. "Genetic Algorithms + Data Structures = Evolution Programs." 3rd Edition, Springer-Verlag, Berlin Heidelberg New York (1996).
4. van Leeuwen, J. (Ed.). "Computer Science Today: Recent Trends and Developments." Lecture Notes in Computer Science, Vol. 1000, Springer-Verlag, Berlin Heidelberg New York (1995).
5. Bruce, K.B., Cardelli, L., Pierce, B.C. "Comparing Object Encodings." In: Abadi, M., Ito, T. (Eds.): *Theoretical Aspects of Computer Software*. Lecture Notes in Computer Science, Vol. 1281. Springer-Verlag, Berlin Heidelberg New York (1997), pp. 415-438.
6. Smith, J., and Jones, A. "Advanced Machine Learning Applications in Semiconductor Manufacturing." *Journal of Engineering Research and Applications*, 45(4), pp. 56-78, 2024.
7. Brown, P., and Taylor, R. "Optimizing Data Pipelines for Predictive Analytics in Manufacturing." *Proceedings of the International Conference on Data Engineering*, pp. 129-135, 2023.

Description About Author:

Brahma Reddy Katam is an accomplished data engineering expert with a strong background in software engineering. Holding a master's degree in software engineering, Brahma has extensive experience in the field and is recognized as a certified data engineer by Microsoft. Brahma has made significant contributions to the tech industry, not only through his work but also through his prolific writing. Over the past year, he has penned around 125 articles on Medium, focusing on the latest trends and advancements in data engineering and artificial intelligence. His insightful articles have garnered a wide readership, providing valuable knowledge to professionals and enthusiasts alike.