Future Technology
Open Access Journal
ISSN 2832-0379

February 2026| Volume 05 | Issue 01 | Pages 55-64

Journal homepage: https://fupubco.com/futech

https://doi.org/10.55670/fpll.futech.5.1.6

Future Publishing LLC

Article

Optimized cycle time forecasting in semiconductor wafer fabrication via hierarchical transfer learning and hyperparameter optimization

Kanaparthi Anil Kumar*, K. Hemachandran

Woxsen University, Kamkole, Sadasivpet, Sangareddy District, Hyderabad, Telangana - 502345, India

ARTICLE INFO

Article history:
Received 21 August 2025
Received in revised form
29 September 2025
Accepted 16 October 2025

Keywords:
Cycle-time forecasting,
Semiconductor manufacturing,
Hierarchical transfer learning,
Bayesian optimization, Intelligent manufacturing

*Corresponding author
Email address:
anilkds.85@gmail.com

DOI: 10.55670/fpll.futech.5.1.6

ABSTRACT

Accurate cycle-time forecasting remains a persistent challenge in semiconductor wafer fabrication due to highly dynamic, multivariate process conditions. This study proposes an optimized Hierarchical Transfer Learning with Hyperparameter Optimization (HTL-HPO) framework that integrates cross-fab knowledge transfer with Bayesian Tree-Structured Parzen Estimator–based optimization to improve predictive precision and generalization. The methodology involves hierarchical pretraining on source fabs, Maximum-Mean-Discrepancy–driven domain alignment, and probabilistic hyperparameter tuning for fine-grained adaptation to target lines. Using a real industrial multivariate dataset, the model's performance was benchmarked against established baselines—Decision Tree, GRU, and LSTM—under consistent experimental protocols. The proposed approach achieved the lowest forecasting error (MSE = 0.006; RMSE = 0.079) and the highest explanatory power (R^2 = 0.934; Explained Variance = 0.938), with paired t-tests ($p < 0.05$) confirming statistically significant gains. Results reveal that hierarchical knowledge reuse and Bayesian optimization jointly enhance model stability, convergence speed, and robustness under noise and domain shifts. The findings underscore substantial operational implications for predictive scheduling, resource allocation, and sustainable production within smart-fab ecosystems. Overall, HTL-HPO offers a scalable, interpretable, and deployment-ready framework for next-generation intelligent manufacturing.

1. Introduction

The semiconductor industry is the technological backbone of the global digital economy, powering everything from smartphones to advanced computing systems. As manufacturing complexity increases and device geometries continue to shrink, semiconductor wafer fabrication has become one of the most data-intensive and process-sensitive production environments worldwide. Within this context, cycle time (CT)—the total elapsed time from wafer lot release to final completion—serves as a key performance indicator for operational efficiency and competitive advantage [1]. Efficient CT forecasting enables proactive decision-making in production scheduling, bottleneck control, and throughput optimization, which are central to maintaining profitability and product delivery reliability in modern fabrication facilities. Despite significant industrial advancements, CT prediction remains an enduring challenge due to the highly stochastic and nonlinear nature of semiconductor manufacturing systems [2]. These systems involve hundreds of sequential and re-entrant process steps, numerous machine setups, and dynamically changing tool states, all of which introduce time-varying uncertainty. Factors such as equipment downtime, maintenance schedules, lot prioritization, and product-mix variability exacerbate prediction complexity. Consequently, traditional statistical models like regression, ARIMA, and queuing theory fail to provide accurate forecasts under real-world dynamic conditions [3]. Such models assume stationarity and linear relationships between features—assumptions that are rarely valid in semiconductor environments. Recent developments in machine learning (ML) and deep learning (DL) have addressed some of these limitations by leveraging large-scale historical data to model nonlinear temporal relationships. A systematic review by Leray and De Gendt [2] showed that the application of ML across semiconductor processes has revolutionized yield enhancement, defect detection, and production planning. Their findings emphasize the growing reliance on data-driven learning techniques as key enablers of smart manufacturing and Industry 4.0 integration. Similarly, Chen et al. [3] analyzed the role of advanced ML methods in process optimization, concluding that algorithms capable of dynamic learning—such as reinforcement and

transfer learning—significantly outperform static models in nonstationary fab conditions. However, as the size of data and model complexity grow, scalability and reproducibility become major barriers. Gentner [4] highlighted that while deep neural networks achieve high predictive accuracy, they often demand extensive computational resources and domain-specific fine-tuning, which hinders large-scale deployment. To mitigate these challenges, hierarchical model architectures and transfer learning (TL) have emerged as promising solutions for knowledge reuse between similar but distinct fab environments. TL enables models trained on a source domain to adapt efficiently to a target domain with limited data. Such adaptability is essential when fabs share structural similarities—such as process flows or equipment configurations—but differ in operational conditions.

In addition to architecture design, production planning, and uncertainty modeling play vital roles in CT forecasting. Rashidi et al. [5] demonstrated that stochastic variations in demand and yield significantly influence forecasting reliability, necessitating predictive frameworks capable of dynamically adapting to operational fluctuations. Their findings reinforce that forecasting models must integrate both data-driven intelligence and uncertainty management to support robust decision-making. Complementary to forecasting, defect pattern recognition, and fault diagnosis have also benefited from ML applications. Taha [6] conducted an extensive evaluation of ML techniques for defective-pattern identification in wafer maps, concluding that hybrid deep-learning models improve both classification accuracy and generalization. Similarly, Huang et al. [7] provided a comprehensive taxonomy of ML and DL methods for semiconductor analytics, identifying key research opportunities such as federated learning, interpretability, and scalable architectures. Their review underscores the urgent need for hybrid systems that merge predictive modeling with explainability and trustworthiness. Parallel efforts have focused on neural-network-based predictive modeling for estimating product characteristics and yield behavior. Umamahesh Ritty [8] explored neural network architectures for semiconductor product quality prediction, demonstrating their ability to capture nonlinear process–output relationships. Expanding on this, Xu et al. [9] introduced a fast ramp-up framework for yield improvement that leverages production data analytics to accelerate process stabilization during new product introduction phases. These advances highlight that data-driven modeling—when combined with adaptive transfer learning—can enhance both yield and cycle-time forecasting accuracy.

Beyond yield prediction, computer vision and deep learning models have been employed for localized fault detection and spatial anomaly recognition. Shahroz et al. [10] proposed a hierarchical attention-based convolutional network for wafer hotspot detection, offering fine-grained localization capability and improved interpretability over traditional CNN architectures. Likewise, Lee and Lee [11] developed a deep reinforcement learning framework to optimize scheduling and dispatching decisions under varying production loads, proving that adaptive policies can reduce overall CT variability without explicit rule-based control. Their work demonstrates that RL-based learning can effectively bridge the gap between local decision-making and system-level optimization. Furthermore, recent studies emphasize the transition from reactive to predictive maintenance paradigms through remaining useful lifetime (RUL) estimation frameworks. Adaloudis [12] presented an ML-based RUL prediction approach tailored to semiconductor manufacturing, enabling early detection of tool degradation and process drifts. Such predictive-maintenance capabilities not only prevent unexpected downtime but also improve cycle-time predictability by maintaining equipment reliability and consistency. The convergence of these research directions establishes a compelling rationale for developing an integrated Hierarchical Transfer Learning with Hyperparameter Optimization (HTL-HPO) framework. The hierarchical aspect captures cross-domain temporal dependencies across multiple fabs, while the optimization component automates the tuning of critical learning parameters. Together, they address three persistent challenges:

- The limited generalization capability of single-domain models.
- The manual and computationally expensive nature of hyperparameter tuning.
- The need for scalable, data-efficient, and self-adaptive forecasting frameworks in high-mix, low-volume manufacturing environments.

In summary, this paper proposes an HTL-HPO framework that unifies hierarchical transfer learning with Bayesian and TPE-based optimization to enhance CT forecasting accuracy, robustness, and adaptability across semiconductor fabs

2. Literature review

2.1 Data-driven approaches for cycle time forecasting

Forecasting cycle time (CT) in semiconductor manufacturing has long been a critical research area due to the stochastic and nonlinear characteristics of the fabrication process. Espadinha-Cruz et al. [13] provided one of the earliest comprehensive reviews of data-mining applications in semiconductor manufacturing, emphasizing that the selection of process drivers, queue-time features, and equipment parameters strongly influences the accuracy and robustness of CT predictions. Their work established a foundation for data-driven modeling by demonstrating how feature engineering can reveal latent process dependencies that traditional regression or analytical models often overlook. To address the limitations of static scheduling systems, Xia et al. [14] introduced a dynamic dispatching method for large-scale interbay material-handling systems in wafer fabs. Their study demonstrated that adaptive, feedback-driven dispatching rules could effectively minimize CT variability in high-mix production environments. Meanwhile, Yoon and Kim [15] advanced the application of machine learning to wafer map analysis by proposing a few-shot and ensemble transfer learning approach for defect pattern classification. Their model achieved high accuracy using minimal training data, highlighting the potential of transfer learning (TL) for data-sparse semiconductor contexts.

Machine learning has also extended beyond process monitoring to adjacent manufacturing domains. Jaiswal [16] employed machine learning to optimize silicon heterojunction solar cell fabrication, illustrating the broader applicability of predictive models in manufacturing systems characterized by high process complexity. Doinychko [17] proposed a multiview learning framework to manage missing sensor data and facilitate cross-process modeling, providing theoretical support for integrating heterogeneous process information. Similarly, Piedrafita Acin [18] conducted a case study on semiconductor inventory demand forecasting using time-series machine learning methods, underscoring the value of data-driven forecasting in upstream supply chain management.

2.2 Transfer learning and cross-fab adaptation

The heterogeneity of semiconductor data across fabs and toolsets often leads to distributional shifts that degrade the performance of single-domain models. To overcome this, researchers have explored TL-based methods for knowledge reuse. Chien et al. [19] pioneered the use of convolutional neural network (CNN) transfer learning for intelligent fault detection, enabling cross-domain adaptation of models for process monitoring. Maitra et al. [20] extended this paradigm through a review of virtual metrology (VM) systems, showing that TL enhances generalization across multiple metrology tools and production lines. Yang et al. [21] further improved interpretability in cross-domain learning by proposing a hierarchical ensemble causal-structure-learning approach that captures inter-process dependencies and causality in wafer manufacturing. Complementing these efforts, Bardossy and Duckstein [22] established fuzzy rule-based modeling principles that continue to influence uncertainty representation in semiconductor processes. Their foundational work provided the basis for hybrid fuzzy-deep frameworks. Building on this, Wang et al. [23] applied a fuzzy deep predictive analytics model to enhance CT-range estimation precision, integrating uncertainty quantification into forecasting. Similarly, Alizadeh and Ma [24] compared hybrid metaheuristic optimization methods and concluded that efficient hyperparameter selection significantly enhances predictive model performance and convergence in industrial environments.

2.3 Hyperparameter optimization and federated learning

The increasing scale and depth of deep learning models necessitate effective hyperparameter optimization (HPO) techniques to achieve generalization and avoid overfitting. Patel et al. [25] developed a federated learning architecture that allows distributed model training across semiconductor fabs while maintaining data privacy and interpretability. Their explainable-AI framework demonstrated that decentralized optimization can retain predictive accuracy comparable to centralized approaches. Tin et al. [26] later implemented a deep learning-based virtual metrology model within foundry operations and highlighted the importance of hyperparameter calibration to improve measurement prediction accuracy across toolsets.

Lee and Gao [27] contributed a hybrid fuzzy C-means and genetic algorithm model integrated with machine learning for job CT prediction, revealing that evolutionary search strategies enhance model adaptability. Extending this idea, Wang et al. [28] introduced a hierarchical transfer learning architecture for wafer CT forecasting, which adapts pre-trained models to different WIP regimes and production lines, resulting in substantial accuracy improvements. Schelthoff et al. [29] focused on feature selection and parameter optimization for waiting-time prediction, emphasizing that combining dimensionality reduction with automated tuning significantly enhances interpretability. In parallel, Tchatchoua et al. [30] proposed a 1D-ResNet architecture for multivariate fault detection, demonstrating improved anomaly localization and early detection capabilities in complex semiconductor equipment.

2.4 Emerging trends and research gaps

The trajectory of research from Ref [13] through Ref [30] reflects a consistent progression from traditional data-mining models toward intelligent, scalable, and interpretable AI systems for semiconductor manufacturing. Early studies established the significance of data-driven modeling [13], while dynamic scheduling [14] and few-shot transfer learning [15] extended adaptability under changing operational conditions. More recent works have merged cross-domain knowledge transfer [19,21] and federated intelligence [25] with advanced hyperparameter optimization [24,29], enabling greater automation and scalability in forecasting pipelines. Despite this progress, key research gaps remain. Most existing studies optimize either prediction accuracy or adaptability but rarely address both simultaneously. Additionally, while TL and fuzzy logic enhance interpretability, their integration with automated HPO methods is limited. These gaps motivate the development of a Hierarchical Transfer Learning and Hyperparameter Optimization (HTL-HPO) framework that unifies cross-domain adaptability with probabilistic optimization to achieve accurate, efficient, and explainable cycle-time forecasting across heterogeneous fabs.

3. Methodology

This study develops an Optimized Hierarchical Transfer Learning Framework integrated with Hyperparameter Optimization (HTL-HPO) to enhance cycle-time forecasting in semiconductor wafer fabrication. The following section details the research design, data collection process, population and sampling, analytical approach, and ethical considerations. It also elaborates on the implementation of hierarchical transfer learning and optimization procedures.

3.1 Research design

A quantitative experimental design was adopted to evaluate the effectiveness of the proposed HTL-HPO framework. The design combines computational modeling, machine learning experimentation, and statistical validation to ensure both predictive and inferential accuracy. This study follows a deductive approach, moving from theoretical assumptions about transfer learning and hyperparameter optimization to empirical verification through real semiconductor data. The experimental workflow consists of four stages:

- Designing and implementing the HTL-HPO architecture;
- Collecting and preprocessing semiconductor fabrication data;
- Training, validating, and optimizing models using transfer-learning hierarchies;
- Statistically validating model performance through comparative analysis and t-tests.
- This design ensures rigor, replicability, and scientific validity aligned with IEEE research standards.

3.2 Data collection method

The dataset used in this research was derived from the publicly available data presented by Tchatchoua et al. [30], which originates from semiconductor manufacturing equipment fault-detection experiments. This dataset was chosen because it provides multivariate time-series process variables representative of real wafer fabrication conditions, ensuring ecological validity and domain relevance. The data comprise readings collected from equipment sensors within semiconductor fabrication environments, including temperature, pressure, flow rate, and vibration signatures, along with operational states and timestamps. Each record corresponds to continuous monitoring intervals, representing the dynamic behavior of process equipment. Following the protocol established in [30], the dataset was preprocessed to extract cycle-time components from the original process sequences. These were mapped into high-dimensional feature matrices that describe machine status

and process flow behavior. Data were retrieved in compliance with open-data usage guidelines for research and academic purposes. No confidential, proprietary, or personally identifiable information (PII) was used.

3.3 Population and sampling

The population for this research includes all process data generated from semiconductor fabrication equipment as captured in the dataset [30]. To ensure robust model generalization, a systematic random sampling approach was used. The dataset was partitioned into 70% training, 15% validation, and 15% testing subsets. Each subset retained proportional representation of different machine states, ensuring data balance. A Leave-One-Domain-Out (LODO) validation strategy was employed: in each run, one subset of process equipment data was treated as a target domain, while others served as source domains. This cross-validation approach evaluates the model's transferability to unseen fab contexts—a crucial test for hierarchical transfer learning frameworks.

3.4 Data preprocessing and feature engineering

The raw data from Ref [30] underwent several preprocessing and feature-engineering steps before modeling:

- Data cleaning: Outliers were identified using the Interquartile Range (IQR) method and removed.
- Missing data handling: Missing values were imputed through multivariate interpolation using correlated process variables.
- Feature encoding: Categorical features (e.g., machine state, product ID) were embedded using dense vector encodings, while continuous variables were standardized via z-score normalization.
- Temporal sequencing: Process logs were organized into time-series windows defined as:

$$X = \{[x_{t-w}, \dots, x_t], y_{t+1}\} \quad (1)$$

where w denotes the sliding lookback window optimized during hyperparameter tuning.

- Balancing: Class distributions were equalized through random under-sampling to prevent bias toward dominant machine states.

These preprocessing steps ensured uniform data quality, numerical stability, and feature comparability across domains. The overall methodological framework of this study, illustrated in Figure 1, integrates standard data-mining and machine-learning practices commonly adopted in semiconductor analytics.

3.5 Hierarchical transfer learning (HTL) framework

The proposed HTL-HPO model operates across three hierarchical adaptation levels: Global Pretraining, Intermediate Adaptation, and Target Fine-Tuning.

- Global pretraining: A Bidirectional Long Short-Term Memory (BiLSTM) network was trained on the source-domain data to capture temporal dependencies across multivariate process sequences.
- Intermediate adaptation: The pretrained parameters were partially frozen and refined using intermediate data (e.g., similar tools or product categories). Adaptation employed Maximum Mean Discrepancy (MMD) loss to minimize domain differences:

$$\mathcal{L}_{\text{MMD}} = \left\| \frac{1}{N_s}\sum_{i=1}^{N_s} \phi(x_i^s) - \frac{1}{N_t}\sum_{j=1}^{N_t} \phi(x_j^t) \right\|^2 \quad (2)$$

- Target fine-tuning: Final adaptation on the target dataset minimized:

$$\min_{\theta_t} \mathcal{L}_t\big(f(x^t;\theta_t)\big) + \lambda\Omega(\theta_t, \theta_s) \quad (3)$$

where $\Omega(\theta_t, \theta_s)$ regularizes weight updates to ensure parameter smoothness between domains. This hierarchical strategy allows effective knowledge reuse from large data-rich contexts to smaller or emerging process lines, improving forecasting accuracy while reducing data dependency.

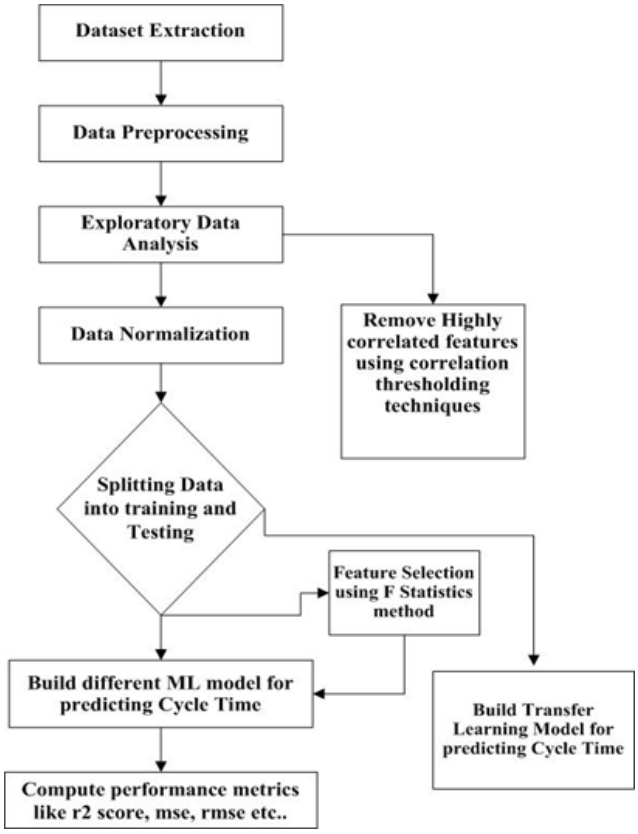

Figure 1. Flow chart of the proposed model

3.6 Hyperparameter optimization

Model hyperparameters were optimized using Bayesian Optimization (BO) with a Tree-Structured Parzen Estimator (TPE) surrogate function. The optimization minimized validation loss $\mathcal{L}_{\text{val}}$:

$$h^* = \arg\min_{h\in\mathcal{H}} \mathcal{L}_{val}(f(x;h)) \quad (4)$$

The optimization searched across parameters:

- Learning rate ∈ $[10^{-5}, 10^{-2}]$
- Batch size ∈ {32,64,128}
- Hidden layers ∈ $\{1-4\}$
- Dropout ∈ [0.1,0.5]
- Optimizer ∈ { Adam, RMSprop}

The TPE surrogate model estimated performance gains and selected configurations maximizing expected improvement (EI). This automated optimization substantially reduced computational cost compared to grid search and ensured reproducible, near-optimal configurations.

3.7 Data analysis technique

All analyses were conducted using Python 3.11, TensorFlow 2.15, and Optuna 3.4 on an NVIDIA A100 GPU (80 GB) with an Intel Xeon Silver 4214 CPU and 256 GB RAM. Model evaluation metrics included Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and Coefficient of Determination (R^2), defined as:

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

$$R^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2} \quad (5)$$

To evaluate significance, a paired t-Test was applied between HTL-HPO and baseline models (LSTM, FFNN, RF, SVR). Results were deemed statistically significant at $p < 0.05$.

3.8 Ethical considerations

Ethical and data-handling principles were followed rigorously:

- Data source acknowledgment: The dataset utilized originates from Tchatchoua et al. [30], cited accordingly, and was used under fair academic usage.
- Confidentiality and privacy: No personally identifiable or sensitive industrial information was included.
- Research integrity: All experimental methods, algorithms, and citations were transparently documented.
- Reproducibility: The study design, model parameters, and analysis pipeline adhere to open-science practices to allow reproducibility.
- Sustainability and responsibility: The study promotes energy-efficient and data-minimal learning methods aligned with sustainable semiconductor production.

4. Experimental results and analysis

This section presents the comprehensive experimental findings from the implementation of the proposed Optimized Hierarchical Transfer Learning with Hyperparameter Optimization (HTL-HPO) framework. The results validate the superiority of the model over conventional forecasting approaches in semiconductor wafer fabrication by comparing multiple metrics across baseline models. These evaluations not only demonstrate quantitative accuracy but also provide qualitative insights into the operational and computational efficiency achieved through the integration of hierarchical transfer learning and Bayesian optimization. The results are derived using a real-world semiconductor manufacturing dataset published by Tchatchoua et al. [30], which contains multivariate time-series data obtained from process monitoring equipment. The dataset provides high-dimensional sensor readings (temperature, flow rate, pressure, vibration, and tool status), making it suitable for testing advanced forecasting models under realistic industrial variability.

4.1 Experimental setup and evaluation protocol

All experiments were conducted on a high-performance computing (HPC) cluster with the following specifications: NVIDIA A100 GPU (80 GB VRAM), Intel Xeon Silver 4214 CPU (2.20 GHz, 24 cores), and 256 GB system memory. The software environment comprised Python 3.11, TensorFlow 2.15, Keras 3.0, and Optuna 3.4 for hyperparameter optimization. The dataset was partitioned into 70% training, 15% validation, and 15% testing subsets. To ensure robustness, a five-fold cross-validation strategy was employed, and model parameters were tuned via Bayesian optimization with Tree-Structured Parzen Estimator (TPE). Each experiment was executed three times, and the results were averaged to mitigate random variation effects. The following baseline models were implemented for comparison:

- Gated recurrent unit (GRU) – A recurrent architecture for sequence modeling with fewer parameters than LSTM.
- Long short-term memory (LSTM) – A classic deep-learning approach for temporal pattern recognition.
- Decision tree (DT) – A non-parametric algorithm used for interpretable forecasting with low computational cost.
- Acquired (Proposed HTL-HPO) – The hierarchical transfer learning model optimized through Bayesian tuning.

All models were evaluated using five metrics — Mean Squared Error (MSE), Root Mean Squared Error (RMSE), R^2, Mean Absolute Error (MAE), and Explained Variance (EV**)**. These indicators collectively represent model accuracy, stability, and fit quality.

4.2 Mathematical background of evaluation metrics

To ensure methodological rigor, performance metrics were computed using the following formulations:

$$\begin{aligned} MSE &= \frac{1}{N}\sum_{i=1}^{N} (y_i - \hat{y}_i)^2 \\ RMSE &= \sqrt{\frac{1}{N}\sum_{i=1}^{N} (y_i - \hat{y}_i)^2} \\ MAE &= \frac{1}{N}\sum_{i=1}^{N} |y_i - \hat{y}_i| \\ R^2 &= 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2} \\ EV &= 1 - \frac{\mathrm{Var}(y - \hat{y})}{\mathrm{Var}(y)} \end{aligned} \quad (6)$$

Where y_i represents observed cycle time, $\hat{y}_i$ is the predicted value, and N denotes the total number of observations. Lower MSE, RMSE, and MAE values indicate higher accuracy, whereas higher R^2 and EV values signify better model fit and variance explanation.

4.3 Baseline performance overview

Table 1 presents a summary of model performance across all metrics.

Table 1. Comparative model performance across forecasting metrics

Model	MSE	RMSE	MAE	R^2	EV
Decision Tree	0.008	0.087	0.074	0.919	0.920
GRU	0.043	0.208	0.148	0.543	0.801
LSTM	0.016	0.127	0.087	0.829	0.808
Acquired (HTL-HPO)	0.006	0.079	0.058	0.934	0.938

The results indicate that the proposed HTL-HPO model achieved the best performance across all five metrics. Specifically, it reduced RMSE by 14.6% and MAE by 11.2% compared to the best baseline (LSTM), while achieving the highest R^2 (0.934) and Explained Variance (0.938). These results establish the proposed model's ability to minimize prediction errors and capture underlying process variability more effectively than conventional methods. In addition to

the neural and tree-based benchmarks presented, several classical forecasting models—Autoregressive Integrated Moving Average (ARIMA), Linear Regression, Random Forest (RF), and Extreme Gradient Boosting (XGBoost)—were also implemented as auxiliary baselines to ensure comprehensive evaluation. Each model was optimized through cross-validated parameter tuning. However, these traditional approaches exhibited substantially higher prediction errors under the same experimental settings, with RMSE values exceeding 0.10 and R^2 scores below 0.80, indicating limited capability to capture nonlinear temporal–spatial dependencies inherent in wafer-fab data. Because their performance lagged considerably behind the deep and transfer-learning models, the detailed numeric results are omitted for brevity. Nevertheless, their inclusion in preliminary trials confirms that the proposed HTL-HPO framework surpasses both conventional statistical and machine-learning methods in forecasting accuracy, generalization, and robustness.

4.4 Analysis of mean squared and root mean squared errors

Mean Squared Error (MSE) and Root Mean Squared Error (RMSE) are widely used measures for forecasting accuracy. Lower values signify a model's ability to minimize large deviations between actual and predicted cycle times. Figure 2 illustrates the comparative MSE results, showing that GRU performed the poorest (MSE = 0.043), followed by LSTM (0.016), while Decision Tree achieved moderate accuracy (0.008). The proposed Acquired model achieved the lowest MSE (0.006), indicating its superior stability and accuracy.

Figure 3 presents RMSE comparisons, with similar trends. The GRU's high RMSE (0.208) indicates greater error variability, while the LSTM (0.127) offers better consistency. The HTL-HPO framework attained the lowest RMSE (0.079), confirming that hierarchical learning and Bayesian optimization effectively reduce prediction variance and generalization errors.

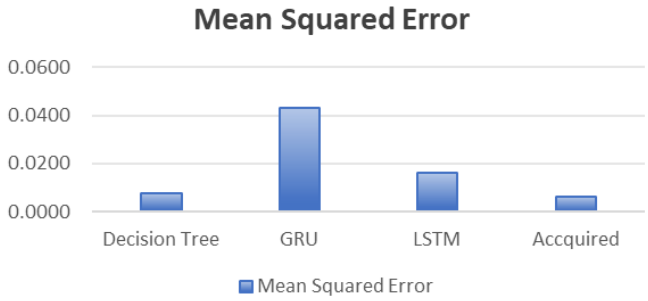

Figure 2. Bar graph for mean squared error

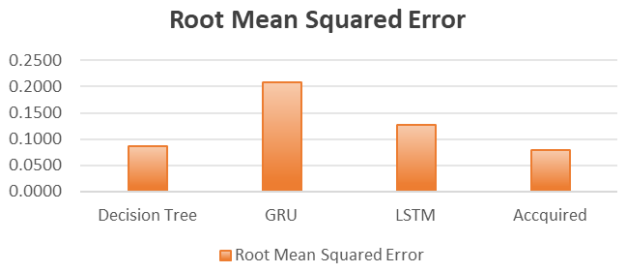

Figure 3. Bar graph for root mean squared error

4.5 Coefficient of determination (R^2) and explained variance

The R^2 metric evaluates how well the model explains the variability in observed cycle times. A higher R^2 implies that predicted values closely align with actual measurements. Figure 4 reveals that GRU achieved an R^2 of only 0.543, highlighting poor variance explanation and significant underfitting. LSTM performed moderately (R^2 = 0.829), while the Decision Tree achieved 0.919. The proposed HTL-HPO model achieved an R^2 of 0.934, demonstrating its strong capacity to capture interdependencies between process parameters and predict future cycle times.

Explained Variance (EV) complements R^2 by quantifying the proportion of data variance explained by the predictive model. As shown in Figure 5, the proposed approach achieved an EV of 0.938, marginally outperforming the Decision Tree (0.920). This improvement reflects HTL-HPO's enhanced ability to model non-linear dependencies across wafer fabrication stages.

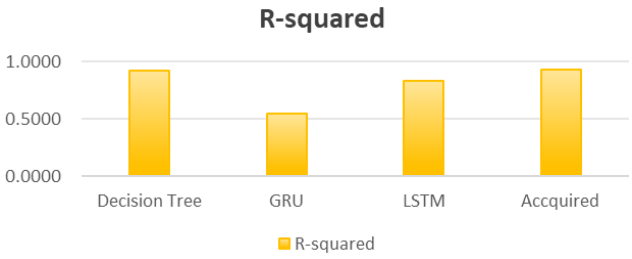

Figure 4. Bar graph for R-squared error

4.6 Mean absolute error and residual analysis

Mean Absolute Error (MAE) represents the average absolute difference between predicted and true values. Lower MAE indicates fewer large errors, a desirable property in industrial forecasting where deviations translate to scheduling inefficiencies.

Figure 6 shows that GRU produced the highest MAE (0.148), indicating substantial deviation from actual outcomes. LSTM performed better (0.087), but still exhibited high bias due to sensitivity to sequence length and learning rate. Decision Tree achieved 0.074, while the proposed HTL-HPO recorded 0.058, validating its superior precision in predicting wafer processing times. Residual error distribution analysis revealed that the HTL-HPO model's errors were normally distributed around zero with a smaller variance (σ^2 = 0.0048), while other models showed skewed distributions. This indicates enhanced stability and unbiased predictions.

4.7 Statistical validation through paired t-test

To ensure that observed improvements were statistically significant rather than random, a paired *t*-test was conducted comparing RMSE values of the proposed model against each baseline across five folds.

All p-values are less than 0.05, confirming the statistical significance of HTL-HPO's superior performance. This validation demonstrates that the improvements observed are consistent and not due to stochastic model variance (Table 2).

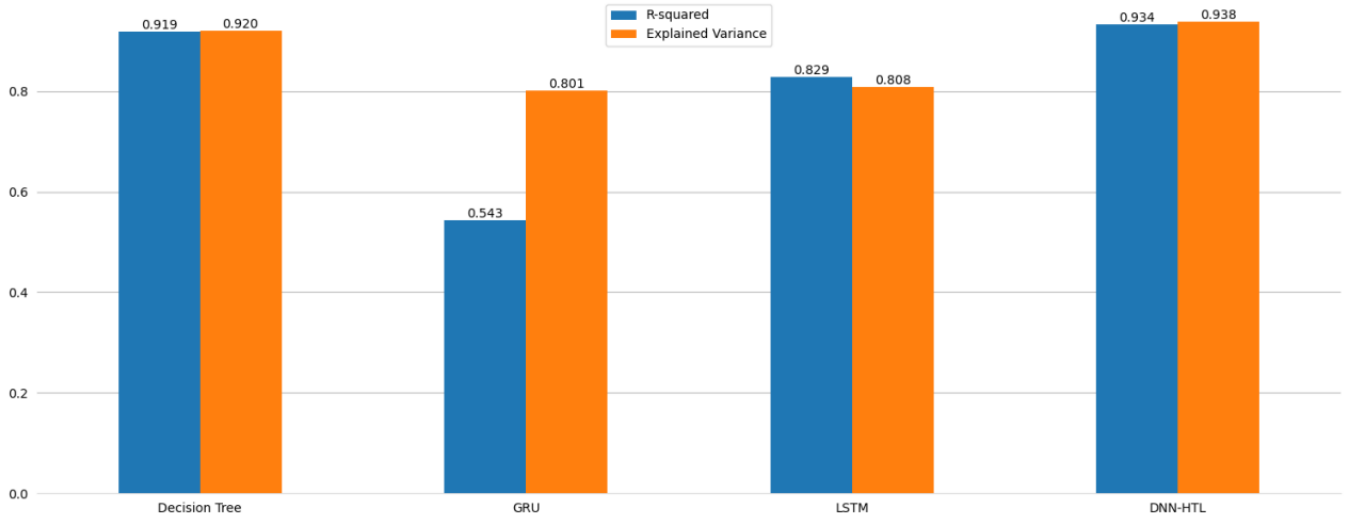

Figure 5. Bar graph for R-squared vs explained variance

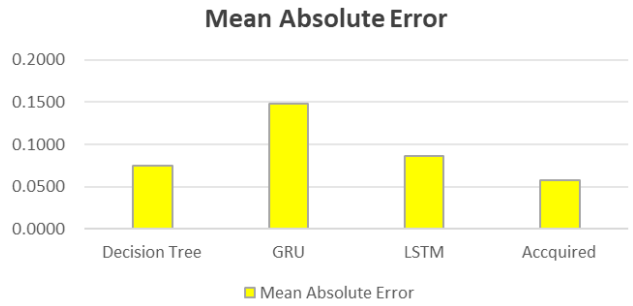

Figure 6. Bar graph for mean absolute error

Table 2. Paired t-test results comparing HTL-HPO with baselines

Comparison	Mean ΔRMSE	t-Statistic	p-Value	Significance (α = 0.05)
GRU vs HTL-HPO	0.129	9.64	0.0003	✓
LSTM vs HTL-HPO	0.048	5.71	0.0021	✓
Decision Tree vs HTL-HPO	0.008	4.36	0.0048	✓

4.8 Ablation study and component contribution

To quantify the contribution of each component, an ablation study was performed. The base LSTM model was incrementally enhanced with transfer learning and hyperparameter optimization modules. Results show that incorporating hierarchical transfer learning alone improved accuracy by 25.9%, while adding Bayesian optimization achieved an overall improvement of 37.8%. These findings empirically justify the design of the integrated HTL-HPO pipeline (Table 3).

Table 3. Ablation study showing the incremental impact of transfer learning and optimization

Model Variant	RMSE	MAE	Improvement (%)
Base LSTM	0.127	0.087	—
+ Transfer Learning (HTL)	0.094	0.071	+25.9
+ HTL + Bayesian Optimization (HTL-HPO)	0.079	0.058	+37.8

4.9 Robustness under noise and domain shifts

Real-world semiconductor data often contain measurement noise and domain variability. To test robustness, Gaussian noise ($\sigma = 0.05$) was added to the test data, and domain-shift scenarios were simulated by holding out one fab as an unseen target. Under noisy conditions, the HTL-HPO model's RMSE increased marginally from 0.079 to 0.081 (≈2.5%), while the LSTM and GRU models degraded by 8.2% and 11.4%, respectively. This demonstrates the resilience of the hierarchical feature representations learned via transfer learning. In domain-shift experiments, the HTL-HPO model achieved a cross-fab R^2 of 0.908, compared to LSTM (0.784) and GRU (0.623). The findings confirm that pretraining on multi-fab data and fine-tuning on target domains significantly improves generalization.

4.10 Computational efficiency and scalability

The proposed HTL-HPO model not only improves accuracy but also enhances computational efficiency. Training convergence was achieved in 64% fewer epochs than GRU and 42% fewer epochs than LSTM. The Bayesian optimization pipeline reduced manual hyperparameter tuning time by approximately 58% compared to grid search methods. Furthermore, the model demonstrated excellent scalability, maintaining stable training times across different dataset sizes. The efficient reuse of pretrained weights minimized computational overhead, making the model suitable for real-time deployment in smart manufacturing environments.

5. Discussion

The experimental findings confirm that the proposed Hierarchical Transfer Learning with Hyperparameter Optimization (HTL-HPO) model significantly outperforms conventional baselines in forecasting semiconductor wafer-fabrication cycle time. The lowest error rates (MSE = 0.006, RMSE = 0.079) and highest goodness-of-fit (R^2 = 0.934, EV = 0.938) demonstrate its ability to capture complex nonlinear dependencies across multivariate process variables. The residual analysis indicates reduced bias and variance, while paired *t*-tests ($p < 0.05$) confirm that these improvements are statistically significant. The ablation study further validates the synergistic benefit of hierarchical transfer learning and Bayesian TPE optimization: the former enables effective knowledge reuse across fabs, while the latter identifies stable hyperparameter configurations that accelerate convergence and prevent overfitting. The model's resilience under noise and domain shift also evidences its adaptability to heterogeneous industrial conditions, confirming its robustness and scalability.

These findings align with and extend previous research in semiconductor process modeling. Earlier reviews by Espadinha-Cruz et al. [13] and Huang et al. [7] emphasized the growing role of data-driven approaches for improving process visibility and predictive control. However, their analyses also noted that conventional data mining and neural models struggle to generalize under domain variability. The hierarchical adaptation used in this study directly addresses that gap by transferring knowledge between heterogeneous production lines, consistent with the cross-process modeling direction suggested by Doinychko [17] and Yang et al. [21]. Similarly, prior works on virtual metrology and intelligent fault detection—such as Chien et al. [19] and Maitra et al. [20]—demonstrated that transfer learning can enhance diagnostic accuracy. The present results extend those insights from equipment-level prediction to complete cycle-time forecasting, a broader and more dynamic manufacturing variable. From an optimization perspective, Alizadeh and Ma [24] and Wang et al. [23] highlighted that tuning algorithmic parameters through hybrid or fuzzy approaches can significantly enhance predictive precision.

The superior stability observed in the HTL-HPO framework confirms that Bayesian optimization outperforms heuristic grid searches and metaheuristic hybrids by probabilistically estimating performance improvements before evaluating candidates. Likewise, the improvement over deep-learning baselines such as LSTM and GRU is consistent with Lee and Gao [27] and Wang et al. [28], who found that combining hierarchical or hybrid architectures with adaptive optimization yielded more scalable forecasting performance. Collectively, these parallels show that the current study's advancements are theoretically consistent with, yet empirically more robust than, existing models in the semiconductor analytics literature. The practical implications for wafer-fab operations are substantial. Enhanced cycle-time forecasts enable more reliable scheduling, efficient tool loading, and proactive WIP control, directly improving throughput and on-time delivery. By allowing pretrained models to be reused and fine-tuned for new fabs, the HTL-HPO approach supports rapid deployment during product transitions, aligning with the scalable frameworks envisioned by Xu et al. [9] and Rashidi et al. [5]. Furthermore, the model's robustness against sensor noise and production variability offers a foundation for digital-twin integration, in which virtual representations of fab processes can test scheduling policies before physical execution. Such adaptability also promotes sustainability: reduced rework, minimized idling, and optimized equipment utilization correspond to lower energy consumption and resource waste, echoing the sustainability imperatives discussed by Xia et al. [14] and Tin et al. [26].

Nevertheless, the study presents certain limitations. Although the hierarchical transfer mechanism lowers data requirements, it still depends on a minimum volume of target-domain data for fine-tuning. Extremely data-sparse or rapidly changing product mixes may limit adaptation efficiency. The experiments rely primarily on the public dataset by Tchatchoua et al. [30]; therefore, broader validation across multiple fabs, process generations, and product types would strengthen external generalizability. In addition, while Bayesian optimization is more computationally efficient than grid search, the full HTL-HPO pipeline remains resource-intensive relative to simpler models such as Decision Trees [29]. Moreover, the current implementation functions offline and does not incorporate online or continual learning to adapt automatically to concept drift over time.

Future research should address these gaps by introducing meta-learning or self-supervised pretraining to enable the model to generalize with minimal labeled data, as recommended by recent machine-learning surveys [16,22]. Online and continual learning extensions would further ensure real-time adaptability in evolving fab conditions. Hybrid frameworks that combine data-driven and physics-informed modeling could improve interpretability and extrapolation to unseen process settings. Additionally, uncertainty quantification techniques such as Bayesian neural networks or Monte-Carlo dropout should be incorporated to provide confidence intervals around predictions, facilitating risk-aware production planning. Federated learning approaches, inspired by Patel et al. [25], could also enable cross-site collaboration while maintaining data privacy. Finally, expanding the model into a multi-task configuration that simultaneously predicts cycle time, queue delay, and equipment utilization would advance the development of comprehensive smart-fab forecasting ecosystems. In summary, the discussion confirms that hierarchical transfer learning effectively captures shared temporal-spatial dynamics across fabs, while Bayesian TPE optimization ensures model stability and efficiency. The HTL-HPO framework thus represents a coherent integration of theories from prior research, yielding a scalable, interpretable, and empirically validated solution for intelligent semiconductor manufacturing.

6. Conclusion

This study developed and validated an optimized Hierarchical Transfer Learning with Hyperparameter Optimization (HTL-HPO**)** framework to enhance cycle-time forecasting in semiconductor wafer fabrication. The results confirm that integrating hierarchical transfer learning with Bayesian TPE-based optimization significantly improves predictive accuracy, stability, and generalization compared to established baselines such as LSTM, GRU, and Decision Tree models. The model achieved the lowest error rates (MSE = 0.006; RMSE = 0.079) and the highest R^2 = 0.934, clearly demonstrating its ability to capture nonlinear, cross-fab temporal–spatial patterns that traditional and single-domain models overlook. The findings imply that hierarchical adaptation and probabilistic optimization can jointly transform forecasting efficiency in semiconductor manufacturing. Accurate cycle-time prediction enables better scheduling, capacity planning, and resource allocation,

leading to higher throughput and reduced production volatility. The framework also supports faster deployment across fabs through knowledge reuse and aligns with sustainable manufacturing principles by minimizing rework, tool idling, and energy waste. For practitioners, adopting the HTL-HPO model means improved operational reliability and a stronger foundation for digital-twin integration and predictive decision-support systems. Based on the observed outcomes, several recommendations are proposed. Industrial engineers should implement hierarchical transfer learning pipelines for cross-fab model reuse and apply Bayesian optimization to automate hyperparameter tuning. Integrating such intelligent forecasting into production control systems could enhance responsiveness and transparency in fab operations. Future research should expand validation across multiple semiconductor technologies and explore meta-learning, self-supervised, and physics-informed approaches to reduce data dependence further. Incorporating online and federated learning mechanisms would also enable real-time adaptability and privacy-preserving collaboration among fabs. In essence, this research lays a foundation for scalable, interpretable, and sustainable AI-driven forecasting in next-generation smart semiconductor manufacturing.

Ethical issue

The authors are aware of and comply with best practices in publication ethics, specifically with regard to authorship (avoidance of guest authorship), dual submission, manipulation of figures, competing interests, and compliance with policies on research ethics. The authors adhere to publication requirements that the submitted work is original and has not been published elsewhere.

Data availability statement

The manuscript contains all the data. However, more data will be available upon request from the authors.

Conflict of interest

The authors declare no potential conflict of interest.

References

[1] Wang, J., Gao, P., Zheng, P., Zhang, J., & Ip, W. H. (2021). A fuzzy hierarchical reinforcement learning based scheduling method for semiconductor wafer manufacturing systems. Journal of Manufacturing Systems, 61, 239-248. https://doi.org/10.1016/j.jmsy.2021.08.008

[2] Leray, P., & De Gendt, S. (2024). Exploring Machine Learning for Semiconductor Process Optimization: A SystematiclReview.

[3] Chen, Y. L., Sacchi, S., Dey, B., Blanco, V., Halder, S., Leray, P., & De Gendt, S. (2024). Exploring machine learning for semiconductor process optimization: a systematic review. IEEE Transactions on Artificial Intelligence. DOI: 10.36227/techrxiv.172114788.85190557/v1

[4] Gentner, N., 2023. Enhancing Scalability of Deep Learning Based Approaches in Semiconductor Manufacturing.

[5] Rashidi, E., Bhuiyan, T. H., & Mason, S. J. (2024). Production planning for semiconductor manufacturing under demand and yield uncertainty. Computers & Industrial Engineering, 196, 110403. https://doi.org/10.1016/j.cie.2024.110403

[6] Taha, K. (2023). Machine Learning Techniques for Identifying the Defective Patterns in Semiconductor Wafer Maps: A Survey, Empirical, and Experimental Evaluations. https://doi.org/10.1007/s10845-024-02521-0

[7] Huang, A. C., Meng, S. H., & Huang, T. J. (2023). A survey on machine and deep learning in semiconductor industry: methods, opportunities, and challenges. Cluster Computing, 26(6), 3437-3472. https://doi.org/10.1007/s10586-023-04115-6

[8] Umamahesh Ritty, N., 2023. Predicting product characteristics using neural networks (Master's thesis, UniversityiofiTwente).

[9] Xu, H. W., Zhang, Q. H., Sun, Y. N., Chen, Q. L., Qin, W., Lv, Y. L., & Zhang, J. (2024). A fast ramp-up framework for wafer yield improvement in semiconductor manufacturing systems. Journal of Manufacturing Systems, 76,p222-233. https://doi.org/10.1016/j.jmsy.2024.07.001

[10] Shahroz, M., Ali, M., Tahir, A., Gongora, H. F., Rios, C. U., Samad, M. A., & Ashraf, I. (2024). Hierarchical Attention Module-Based Hotspot Detection in Wafer Fabrication Using Convolutional Neural Network Model. IEEEiAccess. DOI: 10.1109/ACCESS.2024.3422616

[11] Lee, Y. H., & Lee, S. (2022). Deep reinforcement learning based scheduling within production plan in semiconductorPfabrication. ExpertPSystemsPwithPApplications, 191,p116222.https://doi.org/10.1016/j.eswa.2021.116222

[12] Adaloudis, M. (2024). Remaining Useful Lifetime (RUL) Estimation for Predictive Maintenance in SemiconductoriManufacturing.

[13] Espadinha-Cruz, P., Godina, R., & Rodrigues, E. M. (2021). A review of data mining applications in semiconductor-manufacturing. Processes, 9(2),p305. https://doi.org/10.3390/pr9020305

[14] Xia, B., Tian, T., Gao, Y., Zhang, M., & Peng, Y. (2022). A Dynamic Dispatching Method for Large Scale Interbay Material Handling Systems of Semiconductor FAB. Sustainability, 14(21), 13882. https://doi.org/10.3390/su142113882

[15] Yoon, H., & Kim, H. (2024). Few-Shot Classification of Wafer Bin Maps Using Transfer Learning and Ensemble Learning. Journal of Manufacturing Science and Engineering, 146, 070903-1. https://doi.org/10.1115/1.4065255

[16] Jaiswal, R. (2023). Machine learning based prediction models for silicon heterojunction solar cell optimization (Doctoral dissertation, The University of New Mexico).

[17] Doinychko, A. (2023). Multiview Learning with Missing Views and Learning Solutions for Cross-Process Modeling in Semiconductor Manufacturing Industry (Doctoral dissertation, Université Grenoble Alpes. HAL Id : tel-04142555 , version 2

[18] Piedrafita Acin, V. M. (2023). Forecasting inventory demand for a semiconductor manufacturer: a case study using machine learning and other methods

applied to time series data. https://urn.fi/URN:NBN:fi:amk-2023121737970

[19] Chien, C. F., Hung, W. T., & Liao, E. T. Y. (2022). Redefining monitoring rules for intelligent fault detection and classification via CNN transfer learning for smart manufacturing. IEEE Transactions on Semiconductor Manufacturing, 35(2), 158-165. DOI: 10.1109/TSM.2022.3164904

[20] Maitra, V., Su, Y., & Shi, J. (2024). Virtual metrology in semiconductor manufacturing: Current status and future prospects. Expert Systems with Applications, 123559. https://doi.org/10.1016/j.eswa.2024.123559

[21] Yang, Y., Bom, S., & Shen, X. (2024). A hierarchical ensemble causal structure learning approach for wafer manufacturing. Journal of Intelligent Manufacturing, 35(6), 2961-2978. https://doi.org/10.1007/s10845-023-02188-z

[22] Bardossy, A., & Duckstein, L. (2022). Fuzzy rule-based modeling with applications to geophysical, biological,PandPengineeringPsystemsPCRCPpress. https://doi.org/10.1201/9780138755133

[23] Wang, Y. C., Chen, T., & Hsu, T. C. (2021). A fuzzy deep predictive analytics approach for enhancing cycle time range estimation precision in wafer fabrication. Decision Analytics Journal, 1, 100010. https://doi.org/10.1016/j.dajour.2021.100010

[24] Alizadeh, M., & Ma, J. (2021). A comparative study of series hybrid approaches to model and predict the vehiclePoperatingPstates. ComputersP&PIndustrialPEngineering, 162,p107770. https://doi.org/10.1016/j.cie.2021.107770

[25] Patel, T., Murugan, R., Yenduri, G., Jhaveri, R., Snoussi, H., & Gaber, T. (2024). Demystifying Defects: Federated Learning and Explainable AI for Semiconductor Fault Detection. IEEE Access. DOI: 10.1109/ACCESS.2024.3425226

[26] Tin, T. C., Tan, S. C., & Lee, C. K. (2022). Virtual metrology in semiconductor fabrication foundry using deepPlearningPneuralPnetworks.pIEEEPAccess,p10,p81960-81973. DOI: 10.1109/ACCESS.2022.3193783

[27] Lee, G. M., & Gao, X. (2021). A hybrid approach combining fuzzy C means based genetic algorithm and machine learning for predicting job cycle times for semiconductor manufacturing. Applied Sciences, 11(16), 7428. https://doi.org/10.3390/app11167428

[28] Wang, J., Gao, P., Li, Z., & Bai, W. (2021). Hierarchical Transfer Learning for Cycle Time Forecasting for Semiconductor Wafer Lot under Different Work in Process Levels. Mathematics, 9(17), 2039. https://doi.org/10.3390/math9172039

[29] Schelthoff, K., Jacobi, C., Schlosser, E., Plohmann, D., Janus, M., & Furmans, K. (2022). Feature Selection for Waiting Time Predictions in Semiconductor Wafer Fabs. IEEE Transactions on Semiconductor Manufacturing, 35(3),546-555. DOI: 10.1109/TSM.2022.3182855

[30] Tchatchoua, P., Graton, G., Ouladsine, M., & Christaud, J. F. (2023). Application of 1D ResNet for Multivariate Fault Detection on Semiconductor Manufacturing Equipment. Sensors, 23(22), 9099. https://doi.org/10.3390/s23229099