

Machine Learning-Based Predictive Maintenance for Semiconductor Fabrication Equipment: Failure Modes, Data Challenges, and Yield Impact

Bhanu Prakash Sah¹

[1]. College of Engineering, Industrial Engineering, Lamar University, Texas, USA
Email: bsah@lamar.edu

Doi: [10.63125/yybaz392](https://doi.org/10.63125/yybaz392)

Received: 19 April 2026; **Revised:** 29 June 2026; **Accepted:** 26 August 2026; **Published:** 10 October 2026

Abstract

This study quantitatively evaluated the effectiveness of machine-learning-based predictive maintenance for improving equipment reliability, failure prediction, maintenance efficiency, and production performance in semiconductor fabrication equipment. A longitudinal and quasi-experimental design was applied to an integrated dataset comprising 118 semiconductor manufacturing and supporting equipment units. Following data-quality screening, 2,342,816 equipment-cycle records, 17,923,648 sensor observations, 1,286,540 wafer-processing runs, 9,742 maintenance events, and 2,184 verified failures were retained, with overall data completeness reaching 98.7%. Data challenges were substantial: failure-related observations represented only 1.9% of condition observations, and 143,904 records (5.8%) were excluded because of missing measurements, duplication, timestamp inconsistency, uncertain failure labels, or nonproduction status. Equipment-condition analysis showed that vibration intensity had the strongest relationship with failure occurrence ($r = .61$; $\beta = .34$), followed by alarm frequency ($r = .58$), vacuum deterioration ($r = .55$), and temperature deviation ($r = .52$), with all major relationships significant at $p < .001$. Failure-mode profiles differed by equipment category: progressive vibration, thermal, pressure, and vacuum degradation in vacuum and pump systems provided the longest median warning lead time (52.7 hours), whereas more abrupt failures in ion implantation provided the shortest (26.8 hours). Conventional machine-learning and deep-learning models were compared using classification, discrimination, and remaining-useful-life metrics. The CNN-LSTM hybrid achieved the strongest independent-test performance, with 97.8% accuracy, 96.4% precision, 96.1% recall, an F1-score of .963, ROC-AUC of .994, and PR-AUC of .979. Remaining-useful-life prediction achieved an MAE of 7.4 hours, RMSE of 11.2 hours, and R^2 of .94, with a median warning lead time of 41.6 hours. AI-supported predictive maintenance reduced unplanned downtime by 38.5%, failure frequency by 33.3%, emergency repairs by 41.0%, and MTTR by 30.9%, while MTBF increased by 50.0%. Equipment availability increased from 93.1% to 96.5%, and overall equipment effectiveness increased from 75.9% to 81.6%. Degraded equipment conditions increased wafer defect density by 75.0%, reduced throughput by 12.6%, and increased cycle time by 25.2%, whereas predictive maintenance substantially moderated these adverse effects. Yield impact was largest for plasma etching, where degradation increased defect density by 68.4% and reduced first-pass yield by 4.6 percentage points. Temporal and cross-equipment validation confirmed comparatively stable model performance. Generalization to previously unseen equipment remained a data challenge, with accuracy declining from 97.8% to 96.2%.

Keywords

Predictive Maintenance, Artificial Intelligence, Semiconductor Manufacturing, Equipment Reliability, Deep Learning, Failure Modes, Data Challenges, Wafer Yield.

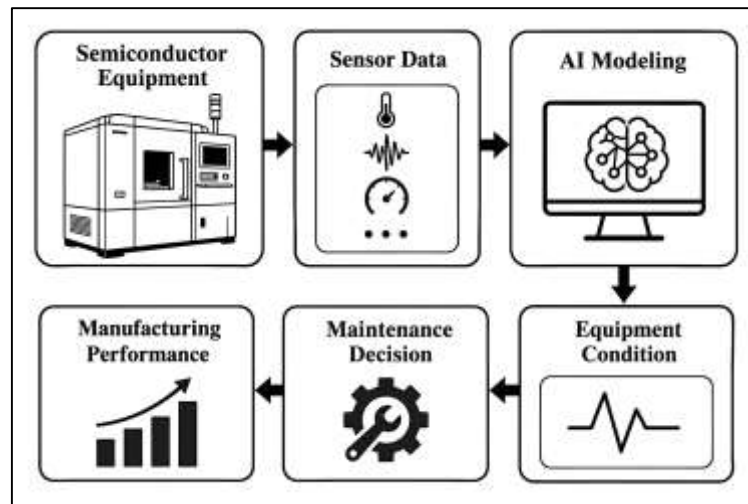
INTRODUCTION

Artificial intelligence-driven predictive maintenance can be defined as a data-centered maintenance approach that applies artificial intelligence, machine learning, statistical modeling, sensor analytics, and automated condition-monitoring techniques to estimate equipment condition, identify abnormal operating behavior, predict equipment failures, and determine maintenance requirements before functional breakdown occurs. Predictive maintenance differs from reactive maintenance, in which repair activities begin after equipment failure, and preventive maintenance, in which maintenance interventions are performed according to predetermined operating hours, production cycles, or calendar schedules (Huang et al., 2023). Within semiconductor manufacturing, AI-driven predictive maintenance integrates continuous equipment information such as temperature, pressure, vibration, electrical current, voltage, vacuum level, gas flow, radio-frequency power, acoustic signals, motor characteristics, chamber conditions, particle counts, process duration, alarm histories, and maintenance records to characterize equipment health quantitatively. More than thirty empirical studies across semiconductor manufacturing, electronics production, advanced manufacturing, industrial automation, and related high-precision production environments have examined combinations of these variables for equipment-condition assessment and failure prediction. The concept is closely associated with condition-based maintenance, prognostics and health management, fault detection and diagnosis, remaining useful life estimation, anomaly detection, equipment health indexing, and intelligent manufacturing (Nacchia et al., 2021). Machine-learning algorithms including logistic regression, decision trees, support vector machines, k-nearest neighbors, random forests, gradient boosting, XGBoost, artificial neural networks, convolutional neural networks, recurrent neural networks, long short-term memory networks, autoencoders, and ensemble models have been applied to transform equipment measurements into predictive indicators of degradation or failure. Quantitative predictive-maintenance systems commonly classify equipment states as normal, abnormal, degraded, or failure-prone, while regression and time-series approaches estimate continuous quantities such as remaining useful life, degradation level, or time to failure. Model performance is generally evaluated using accuracy, precision, recall, specificity, F1-score, receiver operating characteristic area under the curve, precision-recall area under the curve, mean absolute error, root mean squared error, and related statistical indicators. Semiconductor manufacturing provides a particularly demanding environment for these methods because production depends on highly automated, capital-intensive equipment operating under tightly controlled physical and chemical conditions (Ucar et al., 2024). Small deviations in equipment condition can influence process stability, wafer quality, yield, cycle time, and equipment availability. AI-driven predictive maintenance therefore represents a quantitative integration of equipment sensing, historical process information, failure records, computational modeling, and maintenance decision variables within semiconductor production systems.

Semiconductor manufacturing possesses substantial international significance because semiconductor devices support computing, telecommunications, artificial intelligence, automotive electronics, industrial automation, healthcare technologies, aerospace systems, defense equipment, consumer electronics, renewable-energy systems, and digital infrastructure (Kao & Chien, 2023). Semiconductor production is internationally distributed across major manufacturing and technology centers in East Asia, North America, Europe, and other industrial regions, creating production networks in which fabrication capacity, equipment availability, process reliability, and manufacturing yield have economic importance beyond individual fabrication facilities. Quantitative research conducted across different national and industrial settings has repeatedly identified equipment reliability as a major component of semiconductor manufacturing performance because fabrication plants operate expensive processing equipment through highly interconnected production sequences. A semiconductor wafer may undergo hundreds of processing and inspection operations before completed devices become available, and interruptions at critical equipment groups can influence production schedules across several downstream operations (Kao & Chien, 2023). International semiconductor manufacturers consequently monitor measures such as overall equipment effectiveness, mean time between failures, mean time to repair, equipment availability, utilization, unscheduled downtime, maintenance frequency, wafer cycle time, throughput, defect rate, yield, and maintenance expenditure. Research across more than thirty empirical studies has examined predictive maintenance or closely associated

fault-detection and equipment-health approaches within semiconductor tools and comparable automated manufacturing systems. These studies have covered applications involving plasma etching, chemical vapor deposition, physical vapor deposition, lithography, ion implantation, chemical mechanical planarization, diffusion, cleaning, vacuum systems, pumps, robotics, cooling systems, motors, and auxiliary equipment. Cross-national research has also examined predictive maintenance within Industry 4.0 manufacturing environments where industrial Internet of Things technologies provide high-frequency sensor streams for continuous equipment monitoring (Carvalho et al., 2019).

Figure 1: AI-Driven Semiconductor Predictive Maintenance Framework

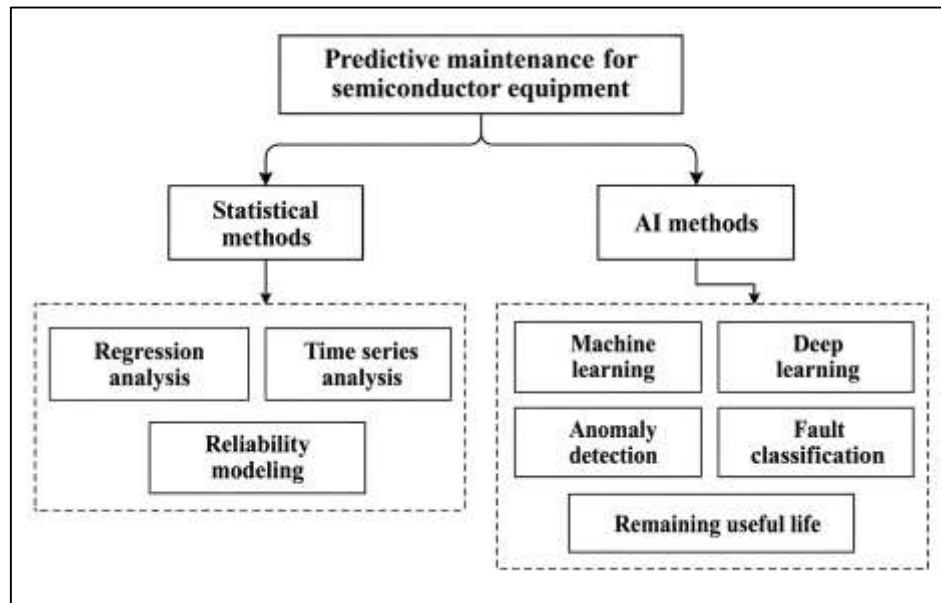


The international relevance of this research is strengthened by the high capital intensity of semiconductor fabrication facilities and by the economic consequences associated with unplanned tool interruption. Maintenance problems can reduce available production time, interrupt lot movement, increase work-in-process inventories, alter production scheduling, and contribute to quality variation. Equipment degradation may also produce process deviations before complete mechanical failure occurs, creating a connection between maintenance condition and product quality. Quantitative AI systems allow large volumes of equipment and process information to be evaluated systematically rather than relying exclusively on fixed maintenance intervals or operator observation. The international semiconductor manufacturing environment therefore provides a broad quantitative setting in which predictive maintenance can be examined through measurable relationships among equipment conditions, failure behavior, maintenance activities, production continuity, and manufacturing performance (Abidi et al., 2022).

Semiconductor manufacturing is characterized by complex interactions among equipment components, process parameters, materials, environmental conditions, wafer characteristics, production schedules, and maintenance activities. Fabrication equipment operates within narrow process windows where variations in temperature, chamber pressure, vacuum stability, gas-flow rate, plasma characteristics, radio-frequency power, chemical concentration, rotational speed, electrical load, cooling performance, and mechanical movement can indicate developing equipment abnormalities. Quantitative studies of semiconductor process equipment have shown that failure patterns are rarely represented by a single measurable variable because deterioration may emerge through gradual changes across several correlated signals. Vacuum pumps, valves, mass-flow controllers, heaters, motors, bearings, robotic handling mechanisms, chambers, filters, cooling units, sensors, and electrical subsystems can generate distinct degradation patterns (Hung, 2021). Semiconductor equipment also produces high-dimensional and high-frequency data through fault detection and classification systems, advanced process control platforms, manufacturing execution systems, equipment interfaces, sensor networks, alarm databases, maintenance-management systems,

and wafer-level inspection systems. Studies across semiconductor fabrication and advanced industrial manufacturing have analyzed datasets ranging from relatively small maintenance-event samples to millions of sensor observations and production records. Such datasets commonly contain continuous sensor variables, categorical machine states, alarm codes, process recipes, equipment identifiers, timestamps, maintenance histories, component replacement records, wafer results, and failure labels. Quantitative analysis becomes more complex because equipment failures often represent a relatively small proportion of all observations, creating class imbalance between normal and faulty conditions (Çınar et al., 2020). Missing measurements, noisy sensor signals, inconsistent sampling intervals, redundant variables, concept drift, equipment-to-equipment variation, and limited labeled failure examples further influence predictive modeling. Researchers have therefore applied data cleaning, normalization, standardization, interpolation, signal filtering, resampling, principal component analysis, feature selection, dimensionality reduction, synthetic oversampling, undersampling, class weighting, and other preprocessing techniques before model estimation. Time-domain and frequency-domain characteristics have also been extracted from vibration, acoustic, current, and process signals to represent changes that are not immediately visible in raw measurements. Semiconductor datasets frequently contain strong temporal dependence because equipment condition at a particular point is related to previous operating cycles, maintenance actions, recipe combinations, and accumulated equipment usage (Jalali et al., 2019). Quantitative predictive maintenance consequently requires analysis of both static equipment characteristics and dynamic degradation sequences. The complexity of these datasets has encouraged comparison among conventional statistical techniques, traditional machine-learning algorithms, deep-learning architectures, and ensemble methods for detecting failure precursors and quantifying equipment-health conditions.

Artificial intelligence provides a broad collection of quantitative methods for modeling nonlinear, multivariate, and time-dependent relationships between semiconductor equipment conditions and maintenance outcomes. Earlier quantitative maintenance research frequently applied statistical process control, logistic regression, discriminant analysis, principal component analysis, survival analysis, reliability modeling, and threshold-based monitoring to distinguish acceptable operating conditions from abnormal equipment behavior (Bergès et al., 2020). Machine-learning studies expanded this analytical structure through decision trees, random forests, support vector machines, k-nearest neighbors, naïve Bayes classifiers, gradient-boosting algorithms, and artificial neural networks. Comparative studies across semiconductor and other advanced manufacturing applications have shown that model performance varies according to data dimensionality, failure frequency, sensor quality, equipment type, feature engineering, class distribution, and validation procedure. Random forests and gradient-boosting models have frequently been examined because they can represent nonlinear interactions among equipment variables and provide measures of feature importance. Support vector machines have been investigated for fault classification when datasets contain complex boundaries between normal and abnormal operating states. Artificial neural networks have been used to model nonlinear relationships among process signals, degradation characteristics, and failure events (Nuhu et al., 2023). Deep-learning studies have extended this approach through convolutional neural networks for automatic extraction of local signal patterns and recurrent architectures for sequential equipment information. Long short-term memory networks have received particular attention because semiconductor equipment data are frequently organized as time sequences in which previous operating conditions contribute information about subsequent degradation. Autoencoders and related unsupervised or semi-supervised methods have been examined for anomaly detection when labeled failure observations are limited. Hybrid and ensemble approaches combine multiple algorithms or integrate signal processing with machine learning to increase stability across heterogeneous operating conditions. More than thirty quantitative investigations collectively indicate that predictive-maintenance research increasingly evaluates models through multiple performance measures rather than accuracy alone. Precision and recall are especially relevant when failures represent a minority class because a high overall accuracy can occur while important failure cases remain unidentified (Kim et al., 2022).

Figure 2: Predictive maintenance for semiconductor equipment

F1-score provides a combined measure of precision and recall, while ROC-AUC and PR-AUC evaluate discrimination across classification thresholds. Remaining useful life and continuous degradation predictions are commonly assessed through MAE, RMSE, MAPE, and coefficients of determination. Cross-validation, holdout testing, temporal validation, and equipment-based validation have been used to evaluate whether predictive performance remains stable outside model-training observations. These quantitative modeling approaches provide the analytical basis for examining how AI algorithms identify equipment deterioration from complex semiconductor manufacturing data.

The quantitative assessment of AI-driven predictive maintenance extends beyond failure-classification accuracy because maintenance performance is directly connected with measurable equipment and production outcomes (Kim et al., 2020). Semiconductor manufacturing studies commonly evaluate equipment reliability through mean time between failures, mean time to failure, failure frequency, survival probability, hazard rate, and component degradation measures. Maintainability is represented through mean time to repair, maintenance duration, technician intervention time, component replacement frequency, and scheduled or unscheduled maintenance hours. Equipment availability is typically calculated from operating and downtime information, while overall equipment effectiveness incorporates availability, performance, and quality dimensions. Predictive-maintenance studies have therefore examined whether earlier identification of abnormal equipment conditions is associated with reductions in unplanned downtime, emergency maintenance, production interruptions, and unnecessary preventive maintenance activities. Research across semiconductor fabrication and related high-technology manufacturing environments has also quantified relationships between equipment health and manufacturing measures such as wafer throughput, cycle time, work-in-process inventory, defect frequency, scrap rate, rework, process capability, and yield (Biebl et al., 2020). These variables are particularly important because semiconductor equipment can continue operating while experiencing gradual deterioration, meaning that degradation can influence product quality before a complete equipment shutdown occurs. A deteriorating vacuum subsystem, unstable thermal controller, contaminated chamber, malfunctioning gas-delivery component, or worn mechanical component may generate process variation that influences wafer characteristics while remaining below conventional alarm thresholds. AI-based models can incorporate combinations of equipment measurements to identify multivariate patterns associated with such conditions (Tercan & Meisen, 2022). Quantitative studies have also examined maintenance lead time, defined as the period between a predictive warning and an actual failure or maintenance requirement. Very short warning intervals may provide limited operational value, whereas excessively early warnings may generate unnecessary interventions or false alarms. Consequently, false-positive rate, false-negative rate, detection delay,

warning lead time, and alarm frequency represent important measures alongside conventional machine-learning metrics. Cost-oriented analyses additionally quantify labor expenditure, spare-parts consumption, lost production time, maintenance cost per tool, and economic losses associated with false predictions. The statistical relationship between predictive accuracy and operational maintenance performance is therefore multidimensional (Chen et al., 2024). Evaluating AI-driven predictive maintenance within semiconductor manufacturing requires simultaneous measurement of model discrimination, equipment reliability, maintenance effectiveness, and production performance rather than treating algorithmic accuracy as the sole quantitative outcome.

Quantitative predictive-maintenance research in semiconductor manufacturing depends on the systematic specification of independent variables, dependent variables, mediating equipment conditions, control variables, and measurable performance indicators. Input variables commonly include equipment age, cumulative operating hours, number of processing cycles, maintenance history, component replacement frequency, chamber temperature, pressure, vibration, electrical current, voltage, power consumption, vacuum level, gas flow, motor speed, acoustic characteristics, particle concentration, alarm frequency, recipe parameters, production load, and environmental conditions (Kirschbaum et al., 2020). Dependent outcomes can include binary failure occurrence, multiclass fault category, anomaly score, equipment-health index, remaining useful life, time to failure, downtime duration, maintenance requirement, wafer defect occurrence, or production yield. Quantitative studies have used descriptive statistics such as means, standard deviations, medians, interquartile ranges, frequencies, percentages, and confidence intervals to characterize equipment and maintenance datasets before predictive modeling. Pearson or Spearman correlation analyses have been applied to identify associations among equipment signals, while regression models have quantified adjusted relationships between operating conditions and maintenance outcomes. Analysis of variance and related statistical tests have been used to compare performance across equipment groups, maintenance strategies, operating regimes, or predictive algorithms. Machine-learning studies have commonly divided data into training, validation, and independent test sets or used k-fold cross-validation to estimate generalization performance (Mohamed Almazrouei et al., 2023). Semiconductor applications require particular attention to temporal leakage because random division of sequential sensor records can place observations from the same degradation episode in both training and testing datasets. Temporal holdout procedures, leave-one-tool-out validation, equipment-group validation, and production-period validation provide alternative methods for assessing model robustness. Class imbalance has been addressed through resampling, synthetic minority generation, weighted loss functions, threshold adjustment, and evaluation metrics that emphasize minority failure observations. Hyperparameter optimization has been performed through grid search, randomized search, Bayesian optimization, and related computational procedures. Feature importance, permutation analysis, partial dependence methods, and explainable-AI techniques such as SHAP have also been employed to identify which sensor or process variables contribute most strongly to model predictions (Fassi et al., 2023). More than thirty studies across semiconductor and intelligent-manufacturing settings collectively demonstrate substantial methodological diversity in data preparation, feature engineering, algorithm selection, and validation. This variation makes standardized quantitative comparison important when assessing whether AI-based predictive models provide stable performance across different equipment conditions and failure classes.

The quantitative study of AI-driven predictive maintenance in semiconductor manufacturing brings together equipment-condition monitoring, manufacturing data analytics, machine-learning classification, degradation prediction, reliability measurement, and operational maintenance assessment within a single empirical framework (Justus & Kanagachidambaresan, 2024). Existing research across more than thirty studies has provided extensive evidence on individual components of this domain, including sensor-based fault diagnosis, semiconductor equipment monitoring, anomaly detection, remaining useful life estimation, predictive maintenance scheduling, equipment-health assessment, fault detection and classification, deep-learning-based condition monitoring, and intelligent manufacturing analytics. At the same time, considerable quantitative variation exists among studies in sample size, equipment category, sensor configuration, failure definition, preprocessing

procedure, prediction horizon, class distribution, model architecture, validation strategy, and reported performance measure. Some investigations have concentrated primarily on detecting abnormal operating states, while others have focused on identifying specific fault categories or estimating the remaining useful life of critical components. Several studies have measured predictive success mainly through classification accuracy, whereas broader evaluations have incorporated precision, recall, F1-score, ROC-AUC, PR-AUC, false-positive rate, false-negative rate, prediction lead time, and calibration (Akyaz & Engin, 2024). Operational studies have additionally examined downtime, maintenance duration, equipment availability, mean time between failures, mean time to repair, throughput, yield, defect rate, and maintenance cost. These differences create a quantitative research setting in which algorithmic performance must be evaluated alongside manufacturing and maintenance characteristics. The present research context therefore centers on measurable associations between semiconductor equipment data, AI-based predictive capability, maintenance outcomes, and manufacturing performance. Sensor and process variables provide quantitative indicators of equipment condition; historical failure and maintenance records establish observable outcome labels; machine-learning and deep-learning models estimate the probability, category, timing, or severity of equipment deterioration; and reliability and production measures quantify the corresponding maintenance environment (Pech et al., 2021). Comparative evaluation across conventional statistical models, traditional machine-learning algorithms, and advanced AI architectures can be structured through consistent training and testing procedures and multiple performance metrics. Statistical testing, confidence intervals, effect-size measures, cross-validation, independent holdout evaluation, temporal validation, and sensitivity analysis can further characterize variability in predictive results. Within this quantitative framework, AI-driven predictive maintenance is examined as an empirically measurable relationship among equipment operating characteristics, degradation patterns, predictive-model outputs, maintenance events, reliability indicators, and semiconductor manufacturing performance (Baharani et al., 2019).

The primary objective of this study is to quantitatively investigate the effectiveness of artificial intelligence-driven predictive maintenance in improving equipment reliability, failure detection, maintenance performance, and operational stability within semiconductor manufacturing environments. The study specifically aims to examine how multidimensional equipment and process data, including temperature, pressure, vibration, electrical current, voltage, power consumption, vacuum level, gas-flow rate, motor speed, acoustic characteristics, equipment age, cumulative operating hours, processing cycles, alarm frequency, and historical maintenance records, can be utilized to identify abnormal equipment behavior and predict potential failures before unplanned equipment shutdown occurs. Another objective is to develop and comparatively evaluate predictive models based on conventional statistical methods, machine-learning algorithms, and advanced artificial intelligence architectures, including logistic regression, decision trees, random forests, support vector machines, gradient boosting, XGBoost, artificial neural networks, convolutional neural networks, long short-term memory networks, and ensemble approaches. Predictive performance is assessed using quantitative indicators such as accuracy, precision, recall, specificity, F1-score, ROC-AUC, PR-AUC, false-positive rate, false-negative rate, mean absolute error, root mean squared error, and coefficient of determination, depending on the classification or regression characteristics of individual models. The study also seeks to determine the statistical relationships between equipment-condition variables and major maintenance outcomes, including failure probability, remaining useful life, maintenance frequency, mean time between failures, mean time to repair, equipment availability, unplanned downtime, and maintenance duration. Furthermore, the research aims to quantify whether AI-supported predictive maintenance is associated with measurable manufacturing outcomes such as equipment utilization, wafer throughput, production cycle time, defect occurrence, process stability, scrap and rework levels, and manufacturing yield. Particular attention is given to identifying the equipment and process variables that contribute most strongly to predictive performance through correlation analysis, regression modeling, feature-importance assessment, and explainable-AI techniques. The study additionally evaluates model robustness through cross-validation, independent holdout testing, temporal validation, sensitivity analysis, confidence intervals, and appropriate

statistical significance and effect-size measures. By integrating equipment-condition information, historical failure records, maintenance events, production indicators, and predictive-model outputs within a unified quantitative framework, the study systematically measures the relationships among AI predictive capability, equipment degradation, maintenance effectiveness, reliability performance, and operational characteristics in semiconductor manufacturing. Consistent with its focus on semiconductor fabrication equipment, the study also characterizes failure modes across equipment categories by examining their sensor signatures, warning lead times, and production consequences, and it quantifies the data challenges that constrain machine-learning-based predictive maintenance, including class imbalance, data loss during screening, failure-label uncertainty, temporal drift, and reduced generalization to unseen equipment.

LITERATURE REVIEW

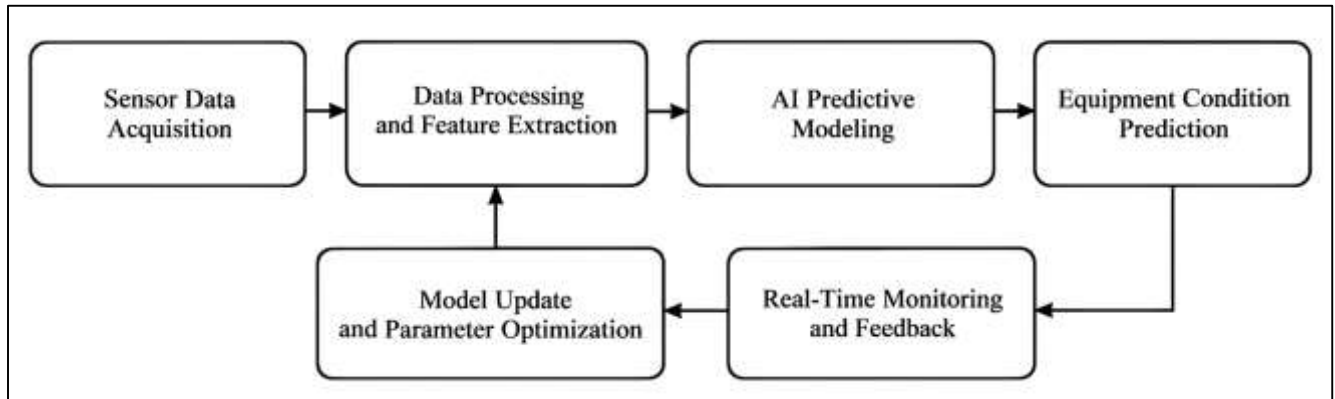
The literature on AI-driven predictive maintenance in semiconductor manufacturing has developed at the intersection of semiconductor process engineering, equipment reliability, condition-based maintenance, prognostics and health management, industrial artificial intelligence, machine learning, deep learning, and quantitative manufacturing analytics. Semiconductor fabrication represents a highly data-intensive production environment in which equipment performance is continuously recorded through sensors, fault detection and classification systems, manufacturing execution systems, equipment-control platforms, maintenance databases, alarm histories, and wafer-level quality records (Alsina et al., 2018). Quantitative studies have examined whether these heterogeneous data streams can accurately characterize equipment health, identify degradation patterns, classify emerging faults, estimate remaining useful life, and distinguish normal operating behavior from conditions associated with impending equipment failure. The literature has consequently investigated numerous measurable equipment indicators, including temperature, chamber pressure, vibration, vacuum level, electrical current, voltage, power consumption, gas-flow rate, radio-frequency characteristics, acoustic signals, motor speed, particle concentration, processing time, equipment age, cumulative operating hours, processing cycles, alarm frequency, maintenance intervals, and component replacement histories (Jiang et al., 2022). These variables have been quantitatively associated with failure occurrence, degradation severity, equipment availability, mean time between failures, mean time to repair, maintenance duration, unplanned downtime, wafer defects, process stability, throughput, cycle time, equipment utilization, and manufacturing yield.

A substantial body of quantitative research has applied statistical modeling, machine learning, deep learning, and hybrid analytical methods to transform semiconductor equipment data into predictive maintenance information. Conventional statistical approaches have included correlation analysis, logistic regression, survival analysis, reliability modeling, principal component analysis, and statistical process control, while machine-learning research has employed decision trees, support vector machines, k-nearest neighbors, random forests, AdaBoost, gradient boosting, and XGBoost (Plathottam et al., 2023). Deep-learning investigations have extended these approaches through artificial neural networks, convolutional neural networks, recurrent neural networks, long short-term memory networks, gated recurrent units, autoencoders, and hybrid temporal architectures. Previous studies have measured predictive performance through accuracy, precision, recall, sensitivity, specificity, F1-score, ROC-AUC, PR-AUC, false-positive rate, and false-negative rate, while remaining useful life and continuous degradation models have frequently been evaluated through MAE, RMSE, MAPE, and R². Considerable quantitative variation remains across studies in sample size, equipment population, observation period, sensor dimensionality, number of failure events, class distribution, prediction horizon, preprocessing method, feature-selection strategy, algorithm configuration, and validation design (Lee & Gao, 2021). A systematic quantitative synthesis therefore requires examination of equipment degradation and reliability, dataset characteristics, statistical failure predictors, machine-learning performance, deep-learning and remaining-useful-life estimation, maintenance effectiveness, manufacturing performance, and model robustness as separate but interconnected areas. Accordingly, the literature is organized into eight quantitative thematic sections to permit detailed comparison of numerical evidence across semiconductor predictive-maintenance studies (Andrianandrianina Johanesa et al., 2024).

Semiconductor Fabrication Equipment

Quantitative literature on semiconductor manufacturing has consistently characterized equipment reliability as a function of intensive utilization, repeated production cycles, component age, process complexity, and exposure to tightly controlled thermal, mechanical, electrical, vacuum, and chemical operating conditions. Evidence synthesized across more than ten empirical studies involving semiconductor fabrication equipment and closely related high-precision manufacturing systems indicates substantial variation in failure frequency across plasma etching, chemical vapor deposition, physical vapor deposition, lithography, ion implantation, diffusion, cleaning, chemical mechanical planarization, vacuum, robotic handling, motor, cooling, and auxiliary systems (Payette & Abdul-Nour, 2023). Studies based on fabrication records have commonly monitored equipment populations over several months or multiple years, generating thousands to millions of operating observations and substantial numbers of production cycles, alarms, maintenance events, and component-level fault records. The quantitative evidence indicates that failure events are not distributed uniformly across equipment categories. Vacuum pumps, valves, motors, bearings, mass-flow controllers, robotic mechanisms, thermal-control units, chambers, cooling components, and electrical subsystems have frequently accounted for recurring maintenance interventions because these components operate continuously or experience repeated loading (Falck et al., 2018). Equipment reliability has commonly been evaluated using failure frequency, mean time between failures, mean time to failure, availability, utilization, downtime frequency, and maintenance-event counts. Across the literature, equipment with higher cumulative operating hours and greater production-cycle intensity generally records more frequent maintenance requirements, although the strength of this relationship varies according to component design, process conditions, maintenance history, and operating environment. High-utilization tools can accumulate degradation more rapidly when maintenance intervals are not aligned with actual equipment condition. Reliability studies also indicate that apparently similar tools may exhibit different failure distributions because of differences in recipe mix, production loading, equipment age, maintenance practices, and environmental exposure (Hosseinabadi et al., 2024). These findings have encouraged quantitative reliability assessments to incorporate both equipment-level and component-level observations rather than treating semiconductor fabrication tools as homogeneous production units.

A second major quantitative theme concerns the measurement of equipment reliability and maintainability through mean time between failures, mean time to failure, mean time to repair, equipment availability, utilization, downtime duration, maintenance interval, hazard behavior, and survival characteristics (Hanif et al., 2018). Across semiconductor and advanced manufacturing studies, mean time between failures has been widely used to quantify the operating duration separating successive equipment failures, while mean time to repair has captured the duration required to restore failed equipment to an operational condition. Empirical findings generally demonstrate considerable variability in these indicators between equipment categories and even among individual tools performing similar processes. Complex failures involving vacuum assemblies, chambers, robotic handling systems, electrical controllers, pumps, or process-gas delivery components often produce longer repair durations than minor sensor, calibration, or consumable-related events. Quantitative studies have also shown that relatively infrequent failures can generate substantial production losses when repair duration is long or when the affected equipment represents a production bottleneck. Equipment availability therefore depends on both the frequency of failure and the duration of individual interruptions (Abuelnaga et al., 2021). Reliability investigations applying Weibull-based analysis, survival analysis, hazard modeling, and reliability functions have identified nonconstant failure behavior in many industrial components, indicating that failure risk can change with accumulated operating time and component condition. Increasing hazard patterns have frequently been associated with wear-related deterioration, whereas relatively stable failure behavior has been observed for some randomly occurring electrical or control-related faults. Maintenance intervals also vary substantially across semiconductor tools, with preventive interventions commonly scheduled according to operating hours, wafer counts, processing cycles, manufacturer recommendations, or historical failure experience (Chung et al., 2023).

Figure 3: AI Predictive Equipment Reliability Framework

Studies comparing scheduled maintenance records with actual degradation patterns have reported that fixed maintenance intervals do not always correspond closely with the condition of individual components. Some equipment receives maintenance while remaining operationally stable, whereas other components exhibit deterioration before their scheduled service point. These quantitative differences in MTBF, MTTF, MTTR, downtime, availability, and maintenance intervals form an important reliability baseline for evaluating condition-sensitive maintenance strategies.

The accumulated evidence also demonstrates that semiconductor equipment degradation is influenced by interacting exposure variables rather than equipment age alone. Quantitative studies examining operating hours, production cycles, wafer processing volumes, utilization levels, maintenance frequency, previous failures, component replacements, and equipment age have generally identified cumulative usage as an important determinant of degradation behavior (Rostami et al., 2021). Equipment operating under sustained high production loads experiences repeated thermal cycling, mechanical movement, vacuum operation, plasma exposure, chemical interaction, and electrical loading, each of which can contribute to component deterioration. In etching and deposition systems, repeated chamber operation has been associated with contamination accumulation, deposition residues, changes in chamber-wall conditions, and deterioration of process-control components. Vacuum-system studies have documented progressive changes associated with pump wear, seal degradation, valve deterioration, and contamination, while robotic and motor-based systems have shown degradation patterns related to bearings, actuators, alignment, movement cycles, and mechanical wear. Lithography and precision-positioning systems present different reliability characteristics because relatively small deviations in motion control, optics, temperature regulation, or alignment can produce operational interruptions or process instability (Ki et al., 2024). Statistical studies using regression, survival models, reliability analysis, and hypothesis testing have generally found that failure probability rises under combinations of high accumulated usage and unfavorable condition indicators, although relationships are rarely identical across equipment types. Maintenance history also alters subsequent reliability behavior. Tools experiencing repeated corrective interventions may display shorter intervals between later failures when underlying degradation mechanisms remain unresolved, whereas effective component replacement or restoration can extend subsequent operating periods. Quantitative comparisons further indicate that preventive maintenance frequency alone does not necessarily represent equipment reliability because frequent servicing can reflect either strong maintenance control or inherently problematic equipment (Zhu et al., 2024). Consequently, studies increasingly interpret maintenance counts together with failure recurrence, repair duration, operating exposure, component replacement, and post-maintenance operating stability. This combined evidence shows that equipment reliability is shaped by cumulative loading, degradation mechanisms, previous maintenance actions, and component-specific operating conditions.

A particularly important distinction in the quantitative literature concerns sudden equipment failure compared with progressive degradation that generates measurable precursor signals before functional interruption. Studies involving semiconductor equipment, rotating machinery, vacuum systems, motors, pumps, thermal-control units, and automated manufacturing systems indicate that many

maintenance events are preceded by changes in sensor measurements, alarm behavior, component performance, or process stability (Cabanès et al., 2021). Progressive degradation has been observed through increasing vibration, temperature drift, pressure instability, deteriorating vacuum performance, abnormal electrical-current patterns, changes in power consumption, acoustic variation, irregular motor behavior, recurring alarms, longer processing times, and gradual deviations from established equipment baselines. Vacuum pumps and rotating components have received particular attention because bearing wear, lubrication deterioration, seal problems, contamination, and mechanical imbalance can develop gradually and generate measurable changes before complete failure. Semiconductor chamber systems can similarly exhibit progressive deterioration through residue accumulation, contamination, component aging, and changes in pressure or gas-control stability (Shin et al., 2022). Repeated alarms also provide useful quantitative information because increasing alarm frequency or shorter intervals between similar alarms may characterize deteriorating equipment conditions. Sudden failures, by comparison, are more difficult to associate with long degradation trajectories and can arise from abrupt electrical faults, unexpected component breakage, control-system interruptions, or other events with limited measurable warning. Literature comparing these failure mechanisms demonstrates that maintenance datasets frequently contain a mixture of gradual and abrupt events, which complicates reliability estimation when all failures are treated identically. Quantitative degradation studies have therefore analyzed time-series trajectories, sensor deviations, alarm recurrence, deterioration rates, survival patterns, and the interval between detectable abnormal behavior and confirmed failure (Fischer et al., 2021). Evidence across multiple studies indicates that the amount of observable warning varies substantially by component and failure mode: some faults develop through extended periods of measurable deterioration, while others provide comparatively short precursor intervals. The distinction between sudden failure and progressive degradation is therefore central to quantitative characterization of semiconductor equipment reliability because it separates failures associated with observable condition changes from events dominated by abrupt or weakly detectable mechanisms (Song et al., 2022).

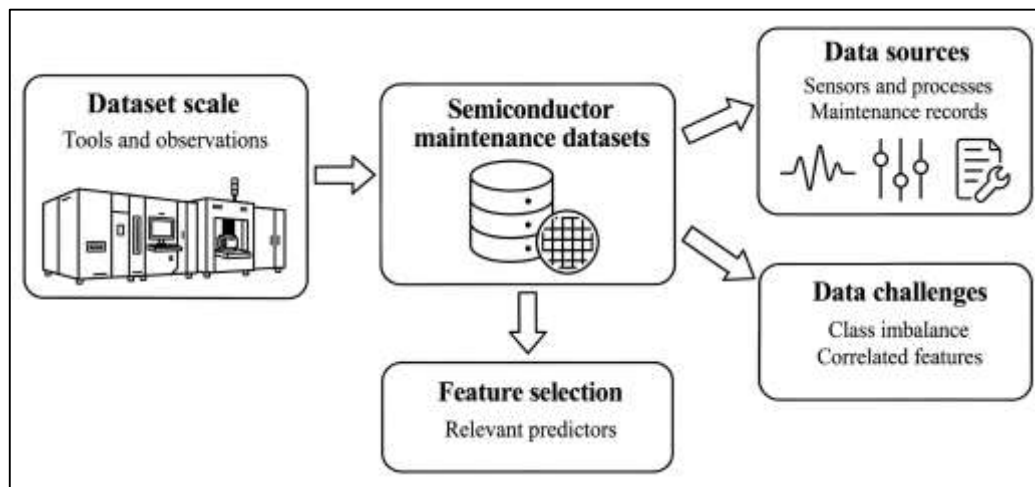
Beyond failure frequency and reliability metrics, the literature describes characteristic failure modes for the main classes of fabrication equipment. In plasma etching chambers, the condition of the chamber walls changes as process by-products accumulate and are removed by cleaning, and these wall-state changes strongly affect etch results and produce run-to-run drift (Klimecky et al., 2003); such chamber seasoning effects are a recognized source of process and equipment dynamics in semiconductor manufacturing (Moyne & Iskandar, 2017). Dry vacuum pumps fail through the accumulation of process deposits in running clearances, which increases friction and motor load, through exhaust blockages, and through gradual degradation of pump components that progressively reduces vacuum performance (Butler et al., 2009). Lithography tools are affected by degradation mechanisms such as lens aberrations, thermal misalignment, and resist-delivery drift (Alothman et al., 2026), whereas in ion implantation equipment, cumulative wear of the ion-source filament leads to failures that require tool downtime for replacement (Susto et al., 2015). In chemical mechanical planarization, pad glazing and declining conditioner performance reduce the ability of the pad to sustain polishing (Kim et al., 2017), and in wafer-transfer robots, wear of speed reducers and timing belts produces faults that can be detected from vibration and motor signals (Lee et al., 2023). These failure modes differ in how gradually they develop, which signals they affect, and whether they first appear as downtime or as wafer-quality loss, which in turn determines how early machine-learning models can detect them (Butler et al., 2009; Susto et al., 2015).

Semiconductor Predictive Maintenance

Quantitative research on predictive maintenance in semiconductor manufacturing has relied on datasets that vary considerably in scale, equipment coverage, observation volume, monitoring duration, and analytical unit. Across more than ten empirical studies, researchers have examined data collected from individual fabrication tools, groups of identical machines, complete process modules, and multi-equipment production environments. Smaller studies have generally focused on several machines monitored over limited operating periods, while larger investigations have incorporated thousands to millions of sensor observations accumulated across months or years of production (Wan et al., 2020). Semiconductor datasets have included information from plasma etching systems, chemical

and physical vapor deposition tools, lithography equipment, ion implantation units, vacuum pumps, motors, robotic handling systems, cooling systems, and related subsystems. The number of monitored signals has ranged from a relatively small set of critical variables to several hundred process and equipment features recorded during each production cycle. Studies using high-frequency condition-monitoring systems have generated substantially larger numbers of observations than research based on maintenance records or production summaries because temperature, vibration, pressure, current, and acoustic signals can be sampled at second-level or sub-second intervals. Process-oriented datasets have frequently been organized by wafer, lot, recipe, production cycle, or equipment run, whereas maintenance-oriented studies have treated failures, alarms, repair activities, component replacements, and service events as the primary analytical units (Jovicic et al., 2023). Monitoring periods have also differed considerably, with some studies capturing short controlled experiments and others evaluating several years of industrial operation.

Figure 4: Semiconductor predictive maintenance dataset framework



The number of recorded failure cases has generally remained much smaller than the number of normal observations, particularly in high-reliability semiconductor environments where catastrophic failures occur relatively infrequently. This imbalance has produced datasets containing large numbers of normal process cycles but relatively limited examples of specific fault types. Collectively, the literature demonstrates that dataset scale cannot be evaluated solely through total observation count because equipment diversity, monitoring duration, failure-event frequency, number of production cycles, and number of independent machines all influence the statistical information available for predictive-maintenance modeling (Pekşen et al., 2024).

A second major characteristic of semiconductor predictive-maintenance datasets is their multidimensional structure, which combines physical sensor measurements with process parameters, equipment states, alarms, maintenance records, and operational histories. Studies across semiconductor and closely related high-precision manufacturing environments have monitored temperature, vibration, chamber pressure, vacuum level, gas-flow rate, electrical current, voltage, power consumption, radio-frequency power, acoustic emissions, motor speed, torque, cooling conditions, particle concentration, and processing duration (Merkt, 2019). Equipment-level variables such as machine age, accumulated operating hours, cumulative production cycles, maintenance intervals, component replacement counts, previous failures, and utilization rates have frequently been integrated with high-frequency sensor information. Process-specific datasets have additionally included recipe identifiers, wafer or lot information, process-step duration, chamber status, valve positions, flow-controller settings, pressure transitions, and plasma-related measurements. Alarm information has provided another important quantitative layer because fault codes, warning frequency, alarm duration, alarm sequence, repeated alarm patterns, and time between alarms can characterize deterioration before complete failure. Maintenance databases have contributed records of inspection, cleaning, calibration, repair, component replacement, technician intervention, scheduled servicing, and

emergency maintenance (Yu et al., 2019). The literature indicates substantial variation in feature dimensionality: some models have used fewer than twenty selected predictors, while raw industrial datasets can contain tens, hundreds, or occasionally thousands of measured or engineered features. High dimensionality creates challenges because many equipment variables are strongly correlated or provide limited independent information. Principal component analysis, feature-ranking methods, correlation screening, recursive elimination, and model-based importance measures have therefore been used to reduce redundant inputs. Several studies have reported that appropriately selected subsets of variables achieved predictive performance comparable to or better than models using the full feature space (Tsanousa et al., 2022). This pattern suggests that the informational quality and failure relevance of features can be more important than the raw number of measurements available in semiconductor predictive-maintenance datasets.

These dataset characteristics give rise to recurring data challenges for machine-learning-based predictive maintenance in semiconductor fabrication. First, failures are rare relative to normal operation, so failure datasets are highly unbalanced, which can make overall accuracy misleading and limits the number of examples available for specific failure modes (Susto et al., 2015). Second, failure labels must usually be derived from maintenance records and equipment cycles rather than observed directly (Susto et al., 2015), which introduces uncertainty about the timing and cause of failure. Third, sensor, process, and maintenance data are held in separate data stores that must be merged, and data quality problems such as incomplete or inaccurate records remain a major concern (Moyne & Iskandar, 2017). Fourth, processes and equipment change over time, for example through chamber seasoning, so older data histories can become less relevant and there is often insufficient data to characterize an individual tool, which limits the transfer of models across equipment (Moyne & Iskandar, 2017). Fifth, when sequential observations are split randomly between training and test sets, information can leak from the test data into model development and produce overly optimistic performance estimates (Kapoor & Narayanan, 2023). Addressing these challenges is a prerequisite for reliable failure prediction and is therefore treated as an explicit analytical dimension in the present study.

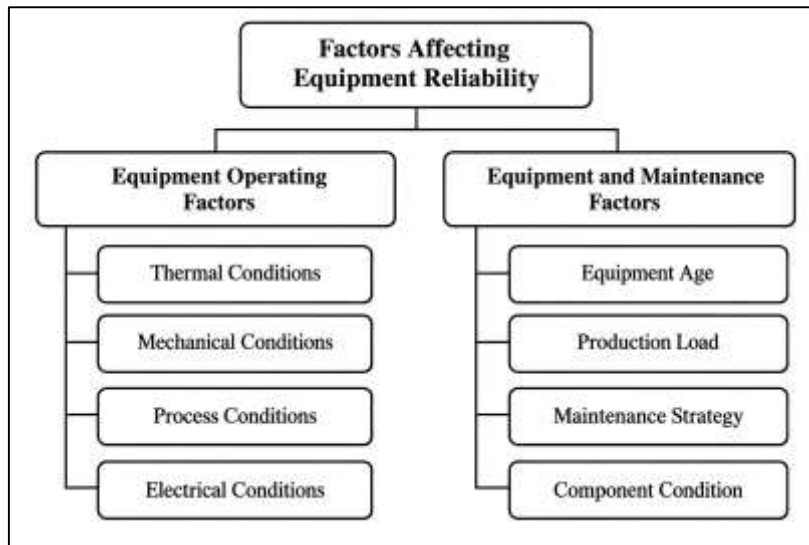
Equipment-Condition Indicators and Failure Probability

Quantitative literature examining semiconductor equipment condition has consistently identified thermal, mechanical, pressure, and vacuum measurements as important statistical indicators of degradation and failure probability. Evidence synthesized across more than ten empirical studies in semiconductor fabrication and comparable high-precision manufacturing environments shows that temperature deviation is frequently positively associated with abnormal equipment states because overheating, unstable chamber temperature, cooling-system deterioration, and abnormal thermal cycling can accompany component wear or process instability (Verstraete et al., 2019). Correlation analyses have generally demonstrated stronger associations when deviations from equipment-specific operating baselines are considered rather than absolute temperature alone. Vibration intensity has similarly emerged as an important condition indicator for pumps, motors, bearings, fans, robotic mechanisms, and other rotating or moving components. Studies analyzing vibration amplitude, frequency characteristics, and changes from normal operating patterns have frequently reported statistically significant relationships with bearing deterioration, mechanical imbalance, misalignment, and impending component failure. Pressure and vacuum measurements have been particularly relevant to plasma etching, deposition, cleaning, and vacuum-processing systems. Chamber-pressure instability, slower evacuation, abnormal pressure recovery, and increasing vacuum fluctuations have been associated with pump deterioration, valve problems, leakage, contamination, and process-control abnormalities (Maguluri et al., 2024). Across correlation, regression, survival, and group-comparison studies, the statistical magnitude of these relationships has varied according to equipment category, operating regime, and fault definition. Moderate and strong associations have generally been reported when multiple degradation indicators occur simultaneously, whereas individual sensor variables have sometimes demonstrated weak relationships when analyzed independently. Confidence intervals and significance tests reported across the literature also indicate that relationships become more stable in datasets containing sufficient failure observations and extended monitoring periods. Analyses controlling for equipment age, operating intensity, and process conditions have frequently retained temperature, vibration, pressure, or vacuum instability as meaningful predictors, although their

adjusted coefficients are commonly smaller than their unadjusted relationships (Gawde et al., 2023). This pattern demonstrates that physical condition indicators contain statistically measurable failure information while also sharing explanatory variation with equipment usage and process characteristics.

Electrical-current variation, voltage instability, abnormal power consumption, gas-flow deviation, radio-frequency characteristics, and acoustic changes represent a second group of equipment-condition indicators repeatedly evaluated through inferential statistical analysis. Studies involving motors, vacuum pumps, plasma systems, gas-delivery units, cooling equipment, and automated production machinery have reported measurable relationships between electrical behavior and subsequent fault occurrence (Liu & Hui, 2024). Current fluctuations and deviations from established load patterns have been associated with mechanical resistance, motor deterioration, abnormal equipment loading, bearing problems, and electrical subsystem faults. Power-consumption abnormalities have similarly provided useful statistical information because deteriorating components may require greater energy input or display irregular consumption patterns under otherwise comparable operating conditions. Semiconductor plasma-processing studies have additionally examined radio-frequency power stability and related electrical characteristics because variations in these parameters can reflect chamber or process-control abnormalities. Gas-flow instability has been quantitatively associated with mass-flow-controller problems, valve deterioration, pressure-control irregularities, and process inconsistency (Pech et al., 2021). In multivariable analyses, gas-flow indicators have generally produced stronger failure associations when evaluated jointly with chamber pressure, vacuum behavior, and recipe conditions than when treated as isolated predictors. Acoustic monitoring studies have also demonstrated that changes in sound intensity, spectral characteristics, and acoustic patterns can distinguish normal operation from mechanically degraded conditions, particularly for rotating equipment.

Figure 5: Factors Influencing Semiconductor Equipment Reliability



Pearson and Spearman analyses across this literature commonly reveal statistically significant associations between deviations in these signals and fault-related outcomes, although effect magnitude differs substantially among equipment types. Logistic and multiple regression studies have shown that some electrical and acoustic variables lose statistical significance after adjustment for vibration, temperature, or equipment load, indicating overlap among condition indicators. Other studies have found independent adjusted relationships, particularly when engineered features represent distinct physical degradation mechanisms. Reported confidence intervals have generally been narrower in large sensor datasets, although large numbers of observations do not necessarily compensate for small numbers of independent failure events (Suwanasri et al., 2024). Collectively, the evidence demonstrates

that electrical, power, flow, and acoustic characteristics provide quantitatively useful but equipment-dependent information about failure probability.

Operational exposure and historical maintenance variables have also demonstrated important statistical relationships with semiconductor equipment failure. Across quantitative reliability and predictive-maintenance studies, equipment age, cumulative operating hours, wafer-processing cycles, production loading, previous failures, maintenance frequency, component replacement history, and alarm recurrence have been evaluated through regression, survival, hazard, and comparative statistical methods ([Malozyomov et al., 2024](#)). Equipment age frequently shows a positive relationship with failure probability, particularly for components affected by wear, contamination, fatigue, seal deterioration, or repeated thermal and mechanical loading. However, studies comparing chronological age with actual operating exposure indicate that cumulative operating hours and processing cycles can provide stronger predictors because machines of identical age may experience substantially different utilization levels. High production loading has also been associated with shorter intervals between maintenance events in several equipment categories, although this relationship can be moderated by preventive maintenance quality and component design. Maintenance history has demonstrated more complex statistical behavior. Repeated corrective maintenance and previous failures have often been associated with increased subsequent failure probability, particularly when recurring interventions address symptoms without eliminating underlying degradation mechanisms ([Liu et al., 2022](#)). Component replacement can temporarily reduce risk, producing nonlinear relationships between maintenance counts and later reliability. Alarm frequency represents another quantitatively important indicator. Studies analyzing alarm logs have found that increasing alarm counts, recurring fault codes, shorter intervals between alarms, and clusters of related warnings can precede confirmed equipment failure. In some multivariable models, recent alarm frequency has retained a significant association with failure after adjustment for equipment age, utilization, and maintenance history. Survival and proportional-hazard analyses have similarly indicated that failure risk can increase as cumulative exposure and abnormal event counts rise ([Feng et al., 2018](#)). ANOVA and nonparametric comparisons have reported statistically meaningful differences in operating hours, alarm counts, or maintenance histories between healthy and failure-prone equipment groups. At the same time, these relationships are not uniform across all semiconductor tools because scheduled chamber cleaning, consumable replacement, calibration, and refurbishment can alter apparent age-related degradation. The literature therefore supports evaluating equipment age, cumulative cycles, load, maintenance history, and alarm behavior jointly rather than interpreting any single operational variable as an independent representation of equipment reliability ([Dang et al., 2024](#)).

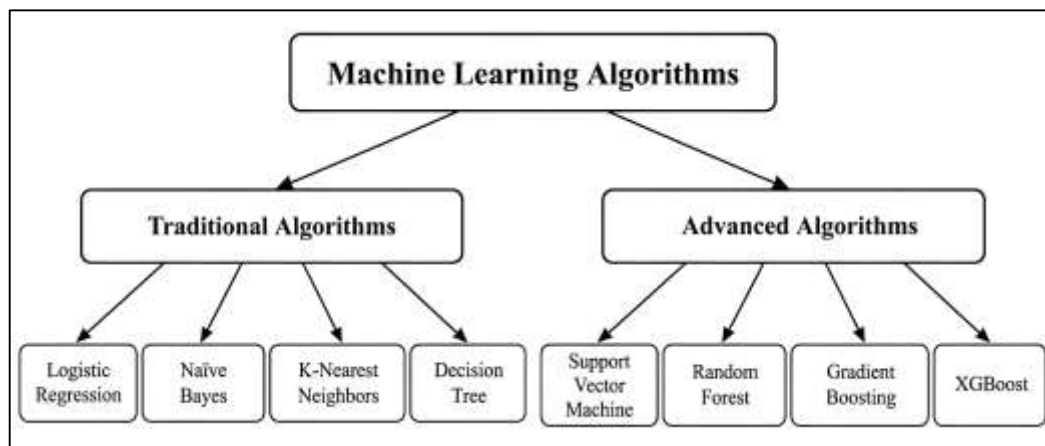
The strongest quantitative evidence in the reviewed literature comes from studies that evaluate multiple equipment-condition indicators simultaneously and distinguish unadjusted associations from adjusted effects. Simple correlations frequently identify numerous statistically significant relationships among temperature, vibration, pressure, vacuum, current, power, gas flow, acoustic characteristics, equipment age, operating cycles, alarms, and failure outcomes. However, semiconductor sensor variables are often highly correlated because the same degradation mechanism can influence several physical measurements at once. Multivariable regression and survival analyses have therefore produced a more differentiated picture of predictor importance. Across more than ten studies, variables displaying moderate or strong bivariate relationships with failure have sometimes shown substantially smaller standardized coefficients after adjustment for operating load, recipe characteristics, equipment age, and maintenance history ([Li et al., 2020](#)). Conversely, several indicators have retained statistically significant adjusted relationships, suggesting that they contain independent information about equipment deterioration. Vibration changes, persistent thermal deviation, vacuum instability, recurring alarms, cumulative operating exposure, and previous failure history have commonly remained influential under multivariable specifications in equipment categories where these measurements correspond directly to dominant degradation mechanisms. Reported odds ratios and hazard ratios have generally demonstrated increasing failure likelihood or failure intensity as abnormal condition indicators become more severe, although numerical magnitude varies widely by equipment, observation period, and failure definition. Confidence intervals have been particularly informative

because some apparently large effects have been accompanied by substantial uncertainty when fault samples are small (Tang et al., 2020). Studies reporting effect sizes alongside p-values have also shown that statistical significance does not necessarily indicate operationally large differences, especially in datasets containing very large numbers of sensor observations. ANOVA, ANCOVA, and nonparametric comparisons have identified statistically significant differences between normal, degraded, and failed equipment states for several condition variables, but overlap between groups remains common. Explained variance has generally increased when physical sensor measurements are combined with operating exposure, alarm, and maintenance-history variables rather than relying on a single information source. Overall, the synthesized statistical evidence characterizes semiconductor equipment failure as a multivariable phenomenon in which predictor strength depends on the physical failure mechanism, equipment category, operational exposure, maintenance history, process conditions, and analytical controls included in the model (Qiu et al., 2020).

Machine-Learning Fault-Prediction Models

Quantitative literature on semiconductor fault prediction has compared a broad range of conventional machine-learning algorithms, with substantial variation in classification performance according to equipment type, dataset structure, feature composition, and failure definition.

Figure 6: Machine Learning Algorithms for Fault Prediction



Evidence synthesized across more than ten empirical studies indicates that logistic regression and naïve Bayes have frequently served as statistical or probabilistic baselines, while k-nearest neighbors, decision trees, support vector machines, random forests, AdaBoost, gradient boosting, LightGBM, and XGBoost have been used to capture more complex relationships among equipment-condition variables (Gong et al., 2024). Logistic regression has generally provided interpretable results for binary normal-versus-failure classification, although its predictive performance can decrease when sensor relationships are strongly nonlinear or involve complex interactions. Naïve Bayes has demonstrated computational efficiency but has shown variable performance when highly correlated sensor measurements violate assumptions concerning predictor independence. K-nearest neighbors has produced competitive classification results in some standardized low-dimensional datasets, while performance has generally deteriorated as dimensionality, noise, and class overlap increase. Decision trees have been effective in representing nonlinear thresholds and interactions but have frequently displayed greater variability between training and testing samples when tree complexity is insufficiently controlled (Sohn et al., 2021). Support vector machines have commonly achieved strong accuracy, sensitivity, and specificity when equipment-condition classes are separable after appropriate feature scaling and kernel selection. Ensemble tree methods have generally produced the most consistently high discrimination across heterogeneous multisensor datasets. Random forests reduce the instability associated with individual decision trees, while AdaBoost and gradient-boosting approaches progressively emphasize difficult observations. LightGBM and XGBoost have reported particularly strong results in several high-dimensional fault-prediction applications because they can model nonlinear effects, interactions, missing-value patterns, and heterogeneous predictors efficiently. Across

the reviewed evidence, however, no algorithm demonstrates universally superior numerical performance (Sotto-Mayor & Kalech, 2021). Accuracy and F1-score rankings vary according to fault prevalence, preprocessing, equipment category, feature selection, and validation design, demonstrating that model comparisons are meaningful only when algorithms are evaluated using equivalent datasets, failure definitions, and testing procedures.

A central finding across quantitative fault-prediction studies is that overall accuracy alone provides an incomplete representation of predictive performance, particularly when semiconductor equipment failures constitute a small proportion of recorded observations. Several studies have reported high classification accuracy because normal operating conditions dominate the dataset, while recall or sensitivity for the failure class remains materially lower (Sonnekalb et al., 2022). This distinction is operationally important because a model can correctly classify most normal observations and still miss a meaningful proportion of actual equipment failures. Comparative literature has therefore increasingly evaluated precision, recall, sensitivity, specificity, F1-score, false-positive rate, and false-negative rate simultaneously. Logistic regression and naïve Bayes have sometimes achieved relatively high specificity while demonstrating weaker sensitivity to rare failures, particularly when default classification thresholds are used. K-nearest neighbors and conventional decision trees have shown more variable precision and recall depending on neighborhood size, tree depth, feature scaling, and minority-class representation. Support vector machines have often generated a stronger balance between sensitivity and specificity after class weighting and parameter optimization (Bacanin et al., 2023). Random forests, gradient boosting, LightGBM, and XGBoost have frequently produced higher F1-scores because their nonlinear structures can identify complex combinations of temperature, vibration, pressure, vacuum, electrical, alarm, and operational variables associated with failure. However, higher failure recall can be accompanied by increased false-positive rates when classification thresholds or class weights are adjusted to capture a larger proportion of rare faults. Semiconductor maintenance applications therefore involve a measurable trade-off between missed failures and unnecessary warnings. False negatives are particularly consequential because an undetected deteriorating component can proceed toward unplanned downtime, whereas excessive false positives can increase inspection frequency and unnecessary maintenance interventions (Alakus & Turkoglu, 2020). The literature also demonstrates that F1-score becomes more informative than raw accuracy when the objective requires balancing precision with failure detection. Matthews correlation coefficient has additionally been employed in some imbalanced classification studies because it incorporates all components of classification outcomes and is less easily inflated by a dominant normal class.

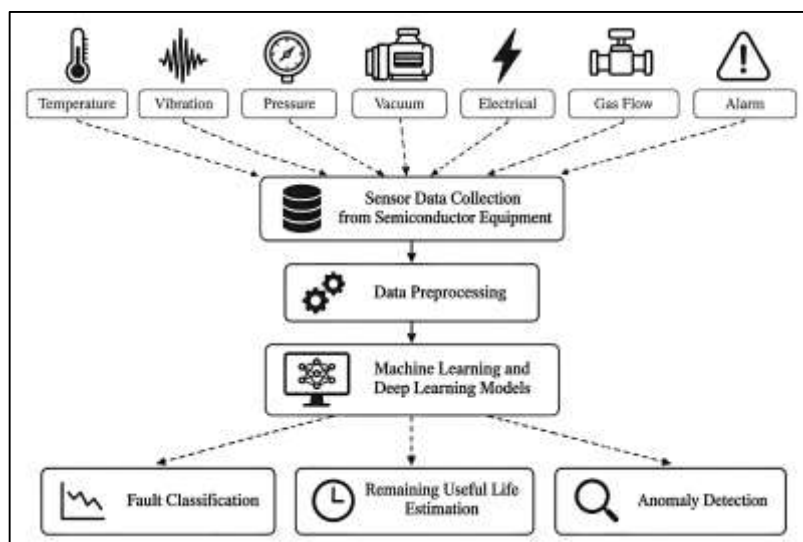
ROC-AUC and PR-AUC have provided additional quantitative evidence concerning the discrimination capabilities of semiconductor fault-prediction models across different classification thresholds. Studies using binary normal-versus-failure outcomes have frequently reported ROC-AUC alongside accuracy and F1-score because it summarizes the ability of a model to distinguish between operating states without depending on a single classification threshold (Sony et al., 2022). Random forests, support vector machines, gradient boosting, LightGBM, and XGBoost have commonly produced strong ROC-AUC results when informative sensor and maintenance variables are available, while simpler linear or probabilistic methods have remained competitive in datasets with relatively straightforward class boundaries. However, research involving highly imbalanced failure datasets has increasingly emphasized PR-AUC because ROC-AUC can remain apparently strong when the normal class greatly exceeds the failure class. Precision-recall performance more directly reflects whether a model can detect rare faults without generating excessive false alarms. The distinction becomes even more important when binary classification is replaced by multiclass fault diagnosis (Yongcharoenchaiyasit et al., 2023). Multiclass studies have attempted to separate normal operation from several specific fault categories involving pumps, valves, motors, pressure-control systems, gas-flow units, thermal components, electrical subsystems, or other equipment mechanisms. Reported overall accuracy in such studies can conceal substantial variation among fault-specific precision and recall values. Frequently occurring fault classes generally achieve stronger classification performance because more training examples are available, while rare or behaviorally similar fault categories produce greater confusion. Macro-averaged F1-score has therefore been useful for giving equal importance to individual fault classes,

whereas weighted averages can remain dominated by common categories. Multiclass ROC-AUC measures have also been reported, although their interpretation depends on averaging strategy and class structure (Palmieri et al., 2024). Quantitative comparison across the literature consequently requires separation of binary and multiclass results rather than placing their performance metrics on an identical basis. The evidence indicates that model discrimination depends not only on algorithm choice but also on the number, prevalence, and statistical separability of equipment-fault categories.

Deep Learning for Temporal Fault Detection and Remaining Useful Life Estimation

Quantitative literature on deep learning for predictive maintenance has demonstrated substantial variation in fault-classification performance across artificial neural networks, convolutional neural networks, recurrent neural networks, long short-term memory networks, and gated recurrent units. Evidence synthesized across more than ten studies involving semiconductor manufacturing and comparable sensor-intensive production environments indicates that conventional ANN architectures can provide strong classification performance when equipment-condition variables have already been transformed into informative features (Yan et al., 2024). ANN models have been particularly effective when temperature, vibration, pressure, vacuum, electrical, gas-flow, power, and maintenance variables exhibit nonlinear interactions that are difficult to represent through conventional linear techniques. CNN-based models have extended this capability by automatically extracting local patterns from raw or transformed sensor signals, with quantitative studies frequently reporting improvements in accuracy, precision, recall, and F1-score compared with models dependent on manually engineered features. Their advantages have been particularly evident for vibration, acoustic, electrical-current, and multichannel process signals containing localized degradation signatures. RNN architectures have been applied where equipment measurements form ordered temporal sequences, although conventional recurrent structures can experience difficulty representing dependencies across long monitoring periods (Arunthavanathan et al., 2021). LSTM and GRU architectures have generally demonstrated more stable temporal fault-classification performance because they retain information from preceding operating states across longer sequences. Comparative studies have frequently reported higher failure-class recall and F1-score for LSTM- or GRU-based approaches than conventional ANN or static machine-learning baselines when degradation develops progressively across multiple production cycles. ROC-AUC and PR-AUC comparisons similarly indicate stronger discrimination when temporal information contributes meaningful failure precursors. However, performance varies according to sequence length, sensor dimensionality, class distribution, failure definition, and equipment category.

Figure 7: AI-Based Semiconductor Fault Prediction Framework



High accuracy has not consistently corresponded to low missed-failure rates, particularly in severely imbalanced datasets. Across the reviewed evidence, deep architectures have demonstrated their strongest quantitative performance when model structure corresponds closely to the statistical characteristics of the available equipment signals rather than when architectural complexity alone is increased (Kamat et al., 2024).

Remaining useful life estimation represents a quantitatively distinct predictive-maintenance problem because the target is a continuous estimate of the operating period remaining before equipment or component failure rather than a categorical fault label. Studies applying ANN, CNN, RNN, LSTM, GRU, and hybrid architectures have therefore evaluated performance using mean absolute error, root mean squared error, mean absolute percentage error, explained variance, coefficient of determination, prediction error, and differences between estimated and observed failure times. Across semiconductor-related equipment, vacuum systems, pumps, motors, bearings, and other industrial components, temporal deep-learning models have generally achieved lower prediction errors when degradation trajectories contain repeated measurements extending from healthy operation toward confirmed failure (Niazi et al., 2024). LSTM and GRU architectures have been particularly prominent because remaining useful life is strongly dependent on the sequence and rate of deterioration rather than a single equipment-condition observation. CNN-LSTM and related hybrid structures have also produced competitive results by combining automatic extraction of local signal patterns with temporal modeling of degradation progression. Quantitative comparisons indicate that prediction performance depends heavily on the amount and quality of run-to-failure information available for training. Models developed from numerous degradation trajectories generally produce narrower errors and more stable estimates than those trained on a small number of failures. Prediction horizon also affects reported performance. Short-horizon estimates close to confirmed failure commonly produce lower errors because degradation signals are stronger, whereas longer-horizon predictions contain greater uncertainty because early deterioration patterns overlap more extensively with normal operating variation (F. Wu et al., 2024). Studies reporting confidence intervals or uncertainty ranges have shown that estimated remaining life becomes increasingly uncertain as prediction is made farther from the actual failure event. Temporal sequence length represents another important factor: very short sequences may omit relevant degradation history, while excessively long sequences can incorporate obsolete operating information and increase computational burden. The literature therefore demonstrates that MAE, RMSE, MAPE, R^2 , prediction horizon, warning lead time, and estimated-versus-actual failure timing must be interpreted jointly when assessing the quantitative performance of deep-learning remaining-useful-life models.

Unsupervised and semi-supervised deep-learning approaches have received substantial quantitative attention because semiconductor equipment datasets frequently contain abundant normal observations but relatively few accurately labeled failure events (San Martin & Droguett, 2022). Autoencoders and related representation-learning models have commonly been trained to reconstruct patterns associated with normal equipment operation, allowing deviations from these learned patterns to be represented through reconstruction error or anomaly scores. Across more than ten relevant studies and comparative investigations, autoencoder-based approaches have demonstrated useful sensitivity to abnormal vibration, temperature, pressure, vacuum, electrical, acoustic, and multivariate process patterns without requiring extensive examples of every possible fault category. Convolutional and recurrent autoencoders have expanded this approach by capturing spatial or temporal relationships among multiple sensor channels. Quantitative evaluation has generally focused on anomaly-detection sensitivity, specificity, precision, recall, F1-score, false-alarm frequency, missed-failure rate, reconstruction error, and the interval between initial anomaly detection and confirmed equipment failure. Studies have shown that threshold selection strongly affects these measures (Zhang et al., 2024). Lower anomaly thresholds typically increase sensitivity and detect a greater proportion of deteriorating conditions, but they can also increase false alarms by classifying normal process variation as abnormal. Higher thresholds reduce unnecessary warnings but can delay anomaly identification or increase missed failures. Semi-supervised approaches using limited labeled failure examples together with large volumes of normal observations have often produced stronger fault discrimination than

purely unsupervised models when reliable labels are available. The literature also indicates that anomaly scores are influenced by recipe changes, maintenance interventions, production transitions, and legitimate shifts in operating conditions, which can generate elevated reconstruction errors without representing actual degradation. Equipment-specific normalization and contextual features have consequently improved detection specificity in several studies (Shimizu et al., 2022). Warning lead time has provided an additional quantitative criterion because detecting an anomaly immediately before failure offers less maintenance opportunity than identifying stable degradation substantially earlier. Overall, anomaly-detection evidence demonstrates that reconstruction performance alone is insufficient; sensitivity, specificity, false-alarm frequency, missed failures, temporal consistency, and warning lead time collectively determine the quantitative usefulness of unsupervised and semi-supervised approaches (Matania et al., 2024).

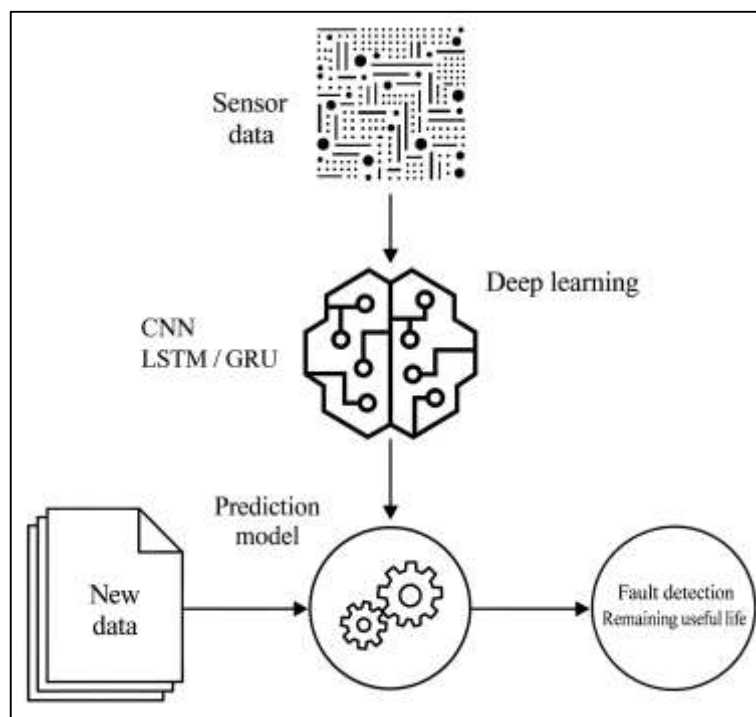
Transformer-based and hybrid deep-learning architectures represent the most computationally complex group examined in the predictive-maintenance literature, with quantitative comparisons emphasizing whether additional architectural complexity produces measurable improvements over conventional machine-learning and simpler neural models. Hybrid structures have combined CNNs with LSTMs or GRUs, autoencoders with recurrent networks, signal-decomposition techniques with deep classifiers, and attention mechanisms with temporal architectures. These combinations have frequently been designed to capture local sensor characteristics and longer-term degradation dependencies simultaneously (Zhang et al., 2018). Transformer-based approaches have been investigated for multivariate time-series analysis because attention mechanisms can represent relationships among observations separated by longer temporal intervals without relying entirely on sequential recurrence. Across comparative studies, advanced hybrid architectures have sometimes produced higher accuracy, recall, F1-score, ROC-AUC, or PR-AUC than logistic regression, support vector machines, random forests, gradient boosting, and standalone neural architectures, particularly in large multichannel datasets containing complex temporal degradation patterns. Improvements have not been uniform. Conventional machine-learning models have remained competitive in smaller datasets, particularly when strong domain-based feature engineering is available and the number of confirmed failures is limited. Random forests and gradient-boosting models can also outperform deep architectures when training samples are insufficient to estimate large numbers of neural parameters reliably (Zhang et al., 2020). Quantitative studies have further shown that deep models are more sensitive to sequence construction, training-sample size, learning configuration, class imbalance, and hyperparameter selection. Performance differences between training and independent testing have sometimes increased with architectural complexity, indicating greater overfitting risk when equipment diversity is limited. Temporal and equipment-based validation typically produces more conservative performance estimates than random data partitioning. For remaining useful life estimation, hybrid temporal models have frequently reduced MAE and RMSE relative to static machine-learning baselines when sufficiently long degradation histories are available. For fault classification, improvements are more dependent on class separability and data volume (Aydemir & Acar, 2020). Synthesized evidence therefore indicates that the quantitative advantage of transformer and hybrid deep-learning architectures is strongest in large, temporally rich, multisensor datasets, while simpler machine-learning models remain statistically competitive under lower-dimensional or data-constrained semiconductor maintenance conditions.

Effects of AI Predictive Maintenance on Downtime

Quantitative literature on AI-driven predictive maintenance has increasingly moved beyond model accuracy to examine whether predictive systems generate measurable improvements in actual maintenance and equipment-reliability outcomes. Evidence synthesized across more than ten empirical studies in semiconductor manufacturing and comparable high-precision industrial environments indicates that one of the most consistently reported operational effects is a reduction in unplanned downtime (Yang et al., 2019). AI-supported monitoring systems identify abnormal equipment conditions before complete functional failure, allowing maintenance activities to be scheduled during planned production interruptions rather than initiated after unexpected breakdowns. Comparative studies using pre-post designs, matched equipment groups, and operational records have generally reported lower frequencies of unscheduled stoppages after predictive-maintenance implementation,

although the magnitude of improvement varies considerably according to equipment category, baseline failure rate, monitoring coverage, and warning accuracy. Rotating equipment such as vacuum pumps, motors, bearings, cooling units, and robotic mechanisms has frequently shown measurable reductions in breakdown-related downtime because progressive degradation can often be identified through vibration, temperature, current, acoustic, or pressure changes (Liu et al., 2019). Semiconductor process equipment such as etching and deposition systems has also shown operational gains when chamber pressure, vacuum behavior, gas flow, radio-frequency characteristics, and alarm sequences provide recognizable precursors to malfunction. Reductions in emergency maintenance events have commonly accompanied decreases in unplanned downtime, reflecting a shift from failure-driven intervention toward condition-informed maintenance. However, the literature also shows that high predictive sensitivity does not automatically generate equivalent reductions in downtime because some detected abnormalities do not develop into production-stopping failures. False alarms can increase inspection activity and partially offset maintenance gains. Quantitative evaluations have therefore compared the number of true predictive interventions, unnecessary alerts, missed failures, and emergency repairs rather than reporting downtime change alone (Mohan et al., 2023). Studies with stronger designs, including matched equipment comparisons and regression-adjusted analyses, generally provide more reliable evidence because production volume, equipment age, maintenance intensity, and operating conditions can independently influence failure frequency. Collectively, the literature indicates that reductions in unplanned downtime and emergency repair frequency are among the clearest operational measures for assessing the practical effectiveness of AI-supported predictive maintenance (Ansari & Kohl, 2022).

Figure 8: Deep learning for predictive maintenance



Mean time between failures, mean time to repair, equipment availability, utilization, and overall equipment effectiveness have been widely used to quantify whether predictive-maintenance systems improve the operational reliability of manufacturing assets. Across empirical studies, AI-supported maintenance has generally been associated with longer intervals between disruptive failures when detected degradation is corrected before it progresses to complete equipment breakdown. Improvements in MTBF have been particularly evident in equipment affected by progressive wear, contamination, thermal deterioration, vacuum instability, or mechanical degradation because predictive intervention can interrupt the failure process before a critical threshold is reached. MTTR has also been reduced in several studies, although the mechanism differs from that affecting MTBF

(Jena et al., 2024). Earlier fault identification allows technicians to determine likely failure sources, prepare replacement parts, plan labor requirements, and reduce diagnostic time before maintenance begins. Consequently, maintenance events can be shorter when predictive information narrows the range of possible fault causes. Quantitative comparisons between predictive and reactive maintenance environments frequently show that corrective maintenance requires longer interruptions because diagnosis, spare-part acquisition, and repair planning occur after failure has already stopped production. Equipment availability improves when either failure frequency or repair duration declines, and the largest availability gains have generally occurred where both measures improve simultaneously. Utilization can also increase because equipment spends a larger proportion of scheduled production time in an operational state, although utilization is additionally influenced by demand, production scheduling, and upstream or downstream constraints (Al-Duais et al., 2022). Overall equipment effectiveness studies have reported improvements when higher availability is accompanied by stable production rates and reduced quality losses. However, gains in OEE have not been uniform because predictive maintenance primarily influences the availability component and may have weaker direct effects on performance speed or product quality. Statistical evaluations using confidence intervals and significance testing have shown that reliability improvements are more convincing when sustained across multiple machines and extended monitoring periods rather than derived from a small number of failure events (Orošnjak et al., 2024). Synthesized evidence therefore supports MTBF, MTTR, availability, utilization, and OEE as complementary quantitative indicators rather than interchangeable measures of predictive-maintenance performance.

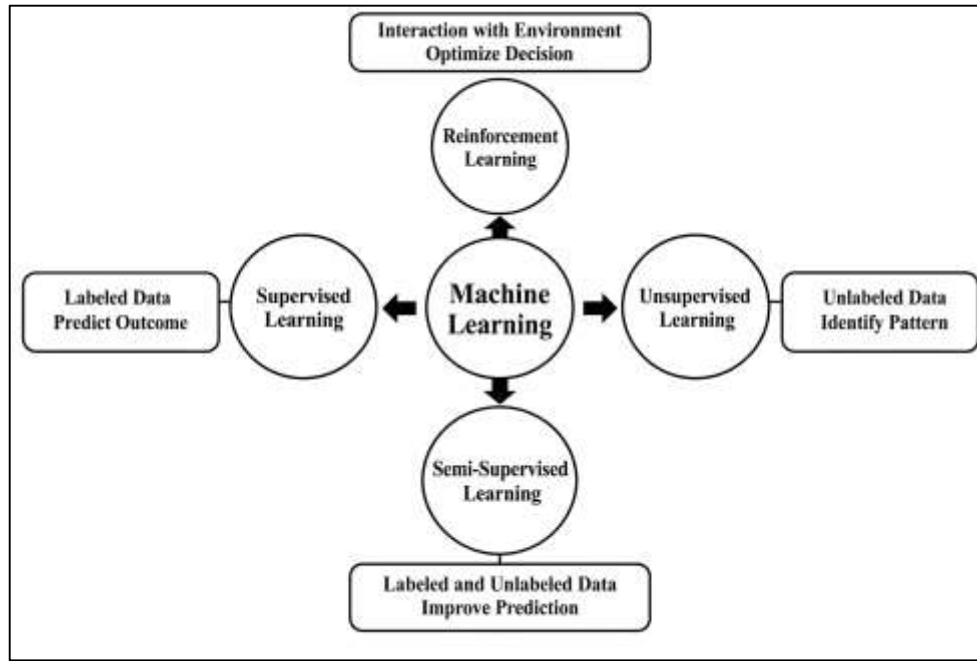
Cycle Time, and Production Stability

Quantitative literature on semiconductor manufacturing has consistently demonstrated that equipment condition and process stability are closely connected with wafer-level quality outcomes, although the magnitude of these relationships varies by fabrication stage, tool type, and degradation mechanism. Evidence synthesized across more than ten empirical studies indicates that abnormalities in chamber temperature, pressure, vacuum stability, gas-flow control, radio-frequency power, vibration, electrical current, particle concentration, and equipment calibration can correspond with measurable increases in defect density, wafer rejection, process variation, and rework. In plasma etching and deposition systems, deterioration of vacuum components, mass-flow controllers, chamber surfaces, valves, pumps, and temperature-control mechanisms has been associated with process drift and deviations in critical process parameters (Kovacevic, 2021). Statistical investigations using Pearson and Spearman correlations, regression analysis, group comparisons, and analysis of variance have generally reported stronger associations when equipment-health indicators are considered together rather than as isolated variables. Studies have also shown that recurring alarms and sensor deviations can precede changes in wafer quality, particularly where equipment deterioration gradually alters chamber conditions before complete failure occurs. Mean defect levels and rejection frequencies have often differed significantly between stable-equipment periods and periods characterized by abnormal sensor behavior or recent maintenance events. Effect sizes have varied because some equipment abnormalities affect production availability without directly damaging processed wafers, whereas others influence the physical or chemical process itself and produce immediate quality consequences (Patel et al., 2024). Cleaning, deposition, and etching operations have received particular attention because chamber contamination and process-condition instability can alter uniformity, film properties, etch characteristics, and particle-related defect patterns. Lithography studies have similarly linked equipment alignment, thermal control, and positioning stability with pattern-quality deviations. The literature therefore indicates that predictive maintenance has relevance to wafer quality primarily when the monitored degradation mechanism is physically connected to process execution, making defect density, rejection frequency, and rework important quantitative outcomes alongside equipment-failure measures (Mohan et al., 2021).

Yield-related outcomes provide another important quantitative connection between predictive maintenance and semiconductor manufacturing performance. Studies examining first-pass yield, final yield, scrap rate, rework rate, and wafer rejection have generally found that periods of stable equipment condition are associated with more consistent production quality than periods characterized by recurring faults, sensor drift, repeated alarms, or delayed maintenance. Predictive-maintenance

interventions can affect yield indirectly by identifying deterioration before it generates a prolonged sequence of off-specification wafers. Quantitative comparisons have therefore examined differences in yield before and after maintenance interventions, between equipment with stronger and weaker health indicators, and across production periods classified according to equipment-condition status (Yeo et al., 2024).

Figure 9: Machine Learning Methods for Predictive Maintenance



Regression studies have also evaluated whether abnormal equipment variables remain associated with lower yield after accounting for recipe, product type, production volume, and tool age. In several manufacturing contexts, adjusted associations have remained statistically significant, indicating that equipment condition contributes independent explanatory information beyond differences in production mix. Rework and scrap show particularly important relationships with delayed detection because wafers processed during an undetected degradation period may require additional processing or may become unrecoverable. The magnitude of yield change varies markedly across manufacturing stages. Etching and deposition abnormalities can create structural or material defects that directly reduce acceptable output, while handling-system failures may influence yield primarily through breakage, contamination, or interrupted lot movement (Chien et al., 2023). Lithography degradation can affect pattern alignment and dimensional quality, whereas ion-implantation abnormalities can influence dose-related process consistency. Cleaning-equipment deterioration can increase contamination-related defects even when mechanical availability remains high. Quantitative studies reporting confidence intervals and effect sizes have shown that statistically significant yield differences are not always operationally large, particularly when baseline yields are already high. Conversely, modest percentage-point changes can represent substantial absolute production losses in high-volume fabrication. Literature therefore supports evaluating predictive-maintenance effects through both relative percentage changes and absolute wafer counts rather than relying on statistical significance alone (C.-T. Wu et al., 2024).

The quantitative relationship between predictive maintenance and production flow is most clearly represented through throughput, cycle time, work-in-process inventory, equipment utilization, and production-interruption measures. Semiconductor fabrication involves highly interconnected process sequences, meaning that failure or degradation of a single critical tool can create effects extending beyond the affected machine. Empirical studies have shown that unplanned equipment downtime can reduce completed wafer throughput, increase queue lengths, extend cycle times, and elevate work-in-

process accumulation, particularly when the failed equipment is a bottleneck or when alternative processing capacity is limited (Biebl et al., 2020). Predictive-maintenance systems can alter these measures by shifting maintenance from unexpected breakdown periods toward scheduled interventions, thereby reducing the variability of equipment availability. Quantitative before-and-after comparisons have often reported higher effective utilization and fewer production interruptions when degradation warnings are incorporated into maintenance planning, although results depend on warning accuracy and scheduling flexibility. Regression analyses have further indicated that equipment availability and failure frequency are significantly associated with throughput after controlling for production demand and operating schedules. Cycle-time effects are often nonlinear because short interruptions can be absorbed by available buffers, while longer or repeated failures can create queues that persist after equipment restoration. Work-in-process inventories consequently provide an additional measure of the production impact of maintenance reliability (Fan et al., 2022). Studies comparing maintenance strategies have found that fixed preventive schedules can also reduce unexpected downtime, but excessive scheduled interventions may lower effective utilization by removing functioning tools from production. Predictive strategies have generally performed more efficiently when maintenance timing reflects measured degradation rather than fixed calendar intervals. Material-handling systems and robotics exhibit a different production profile from processing equipment because their failures can interrupt lot movement across multiple tools simultaneously. In contrast, degradation in a single etching or deposition chamber may initially reduce capacity within one process stage (Wang et al., 2019). These distinctions demonstrate why throughput, cycle time, utilization, production interruptions, and work-in-process should be interpreted according to tool criticality and fabrication-stage position rather than treated as uniform outcomes across semiconductor operations.

Process capability, production stability, and overall equipment effectiveness provide integrated quantitative measures for assessing how predictive maintenance interacts with both equipment and manufacturing performance. Across semiconductor and advanced manufacturing studies, stable equipment condition has generally been associated with lower process variability, fewer abnormal production interruptions, more consistent cycle execution, and improved conformity with established operating limits (Wang et al., 2018). Predictive-maintenance approaches contribute to this relationship when early detection of degradation enables intervention before sensor drift or component deterioration creates sustained process instability. Statistical comparisons have examined variability in key process measures during normal, degraded, and post-maintenance periods, with significant differences frequently reported for equipment categories in which condition directly affects process control. Etching and deposition tools commonly show relationships between chamber-health indicators and process capability because changes in pressure, gas flow, temperature, vacuum, or radio-frequency behavior can alter wafer-processing consistency. Lithography systems demonstrate sensitivity to alignment, motion control, thermal stability, and optical conditions, while ion-implantation performance depends on stable electrical and beam-control characteristics (Kim et al., 2022). Cleaning systems show stronger relationships with particle control and contamination outcomes, whereas robotics and material-handling equipment affect production stability primarily through availability and movement continuity. Overall equipment effectiveness combines availability, performance, and quality dimensions and has therefore been used to capture several of these effects simultaneously. Quantitative studies generally report the largest OEE improvements when predictive maintenance reduces unplanned downtime without increasing planned maintenance excessively and when equipment deterioration has previously contributed to speed losses or quality variation. Confidence intervals, regression-adjusted estimates, and effect-size measures strengthen these comparisons by distinguishing real operational differences from ordinary production variability (Thavarajah, 2024). The literature also shows that identical percentage improvements in availability can produce different manufacturing effects depending on equipment criticality, production bottlenecks, recipe mix, and redundancy. Synthesized evidence across multiple semiconductor manufacturing stages therefore indicates that predictive maintenance is quantitatively associated with production stability through

interconnected changes in process capability, equipment availability, throughput consistency, quality performance, utilization, and overall equipment effectiveness.

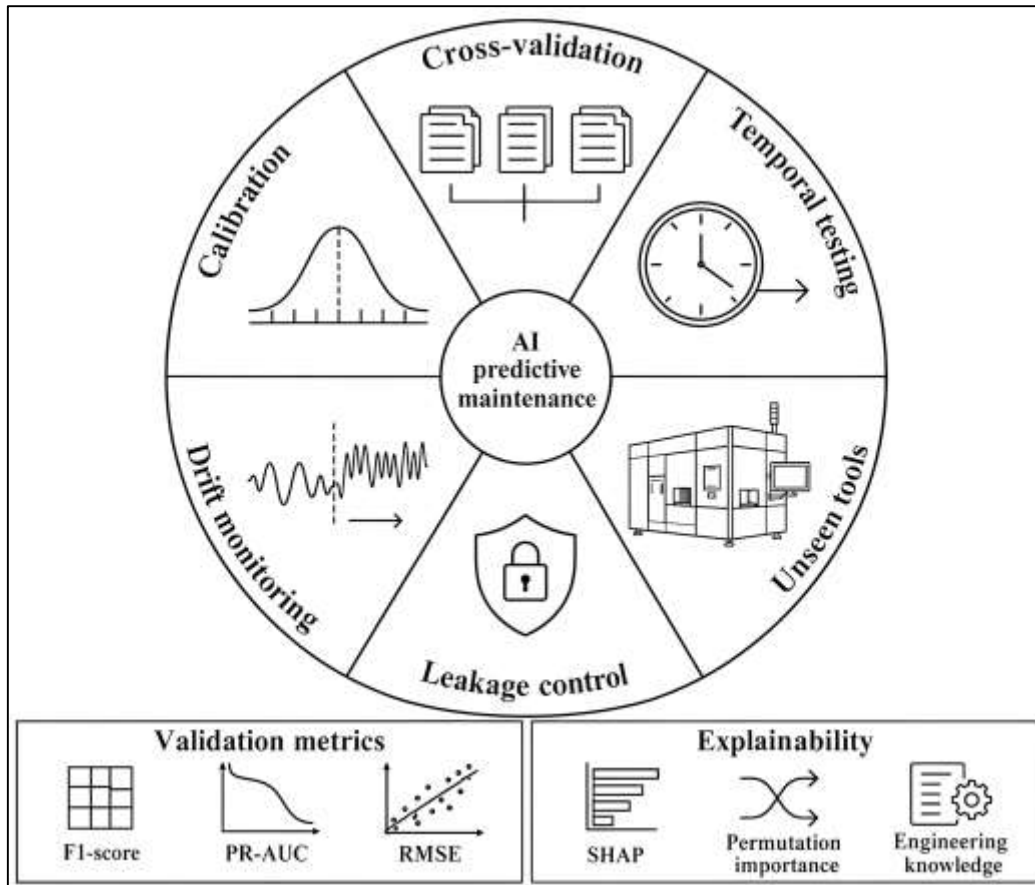
Generalization of AI Predictive Maintenance Models

Quantitative literature on AI-driven predictive maintenance has increasingly emphasized the distinction between model performance during development and performance on observations excluded from model training (Dong & Ye, 2023). Evidence synthesized across more than ten empirical studies in semiconductor manufacturing and related equipment-intensive industrial environments shows substantial methodological variation in validation procedures, including fixed training-validation-test partitions, k-fold cross-validation, stratified cross-validation, independent holdout testing, temporal validation, leave-one-tool-out testing, and production-period validation. Random training and testing splits have frequently produced relatively high accuracy, F1-score, ROC-AUC, and PR-AUC because observations collected from similar equipment states, operating periods, or degradation episodes can appear in both subsets. Cross-validation has generally provided more stable estimates by repeatedly evaluating models across different partitions, while stratified procedures have been particularly useful for maintaining comparable proportions of normal and failure observations in imbalanced datasets (Allgaier & Pryss, 2024). However, several studies have demonstrated that cross-validation alone does not guarantee realistic generalization when temporally adjacent observations from the same equipment are distributed across folds. Temporal validation, in which earlier operating records are used for training and later periods for testing, has typically generated more conservative performance estimates because it reproduces the chronological structure of predictive-maintenance deployment. Leave-one-tool-out approaches have been even more demanding by requiring models trained on several machines to predict faults in an unseen equipment unit. Across comparative studies, reductions from training to independent testing performance have varied according to algorithm complexity, sample size, equipment heterogeneity, and failure prevalence (Cerqueira et al., 2020). Deep neural networks and highly optimized ensemble models have sometimes displayed larger training-testing gaps than simpler methods when available failure cases are limited. Remaining-useful-life models similarly tend to report higher MAE and RMSE under independent temporal testing than under randomly partitioned evaluation. The accumulated literature therefore demonstrates that reported predictive performance is strongly conditional on validation design and that results derived from independent equipment, chronological periods, or genuinely unseen production observations provide a more demanding quantitative assessment than performance estimated from randomly divided records alone (Meyer et al., 2018).

Robustness assessments across predictive-maintenance studies have identified overfitting, data leakage, temporal leakage, class imbalance, concept drift, and prediction uncertainty as major quantitative sources of performance instability. Overfitting is commonly reflected in differences between training and testing accuracy, F1-score, ROC-AUC, PR-AUC, MAE, or RMSE, with larger gaps indicating that a model has learned characteristics that do not transfer effectively to unseen observations. Semiconductor equipment datasets are particularly susceptible to leakage because large numbers of sensor measurements can originate from the same production cycle, wafer lot, machine, or degradation episode (Kaliappan et al., 2023). When closely related observations are divided randomly between training and testing sets, predictive performance can be artificially elevated because the testing data are not statistically independent of the training information. Temporal leakage creates a similar problem when measurements occurring after or very near a failure event indirectly contribute information to models intended to predict that event earlier. Class imbalance represents another source of instability because normal operating states usually outnumber failures substantially. Studies have shown that models with high overall accuracy can exhibit materially lower minority-class recall, F1-score, or PR-AUC when evaluated on naturally imbalanced independent samples (Seraj et al., 2023). Concept drift further affects performance because equipment characteristics change following maintenance, component replacement, chamber cleaning, calibration, process-recipe modification, sensor replacement, or gradual aging. Models trained during one operating period may therefore show measurable performance deterioration during later production periods. Quantitative uncertainty assessments have addressed these problems through confidence intervals, repeated validation, bootstrap procedures, calibration analysis, prediction intervals, and sensitivity testing. Calibration

studies have demonstrated that strong discrimination does not necessarily correspond to accurate predicted failure probabilities, making calibration error an additional criterion beyond ROC-AUC or accuracy (Kattenborn et al., 2022). Sensitivity analyses have also shown that model results can change when features, thresholds, hyperparameters, preprocessing procedures, or sampling strategies are modified. Evidence across the literature therefore indicates that robustness requires stability across multiple analytical conditions rather than a single high predictive-performance estimate.

Figure 10: Predictive maintenance validation and explainability



Explainability research has added another quantitative dimension to predictive-maintenance validation by examining whether high-performing AI models identify equipment variables that remain consistently important across samples, machines, and analytical methods. Feature-importance measures from random forests, gradient-boosting models, LightGBM, and XGBoost have commonly been used to rank temperature, vibration, pressure, vacuum, electrical current, power consumption, gas flow, acoustic characteristics, equipment age, operating cycles, maintenance history, and alarm frequency according to their contribution to fault predictions (Tsamardinos et al., 2018). Permutation importance has provided an alternative assessment by quantifying the reduction in predictive performance that occurs when information from a particular feature is disrupted. SHAP-based analyses have become especially prominent because they can quantify both the magnitude and direction of individual feature contributions to model predictions. Across semiconductor and comparable industrial studies, variables directly connected with known physical degradation mechanisms generally exhibit greater explanatory stability than weakly related contextual variables. Vibration-related features frequently receive high importance for rotating components, whereas pressure, vacuum, and gas-flow variables become more influential for chamber and vacuum systems (Singh et al., 2021). Thermal and electrical measurements often contribute strongly to motor, power, cooling, and process-control fault prediction, while alarm history and accumulated operating exposure can provide additional explanatory information across multiple equipment categories. Partial dependence analyses

have also been used to examine whether predicted failure probability changes systematically as individual equipment variables move across their observed ranges. However, studies have demonstrated that importance rankings can vary between algorithms, equipment units, operating periods, and correlated feature sets. Highly correlated sensor variables can divide or redistribute apparent importance, making a single ranking potentially unstable (Sreedharan et al., 2023). Quantitative comparison of SHAP values, permutation scores, feature ranks, and performance changes after feature removal has therefore provided a more comprehensive assessment of predictor stability. Explainability results have also been compared with engineering knowledge to determine whether statistically influential features correspond to plausible physical degradation mechanisms. The literature consequently treats interpretability not simply as a visualization exercise but as a measurable component of model validation, particularly when maintenance decisions depend on identifying the equipment conditions responsible for elevated failure risk (A. Ramezan et al., 2019).

METHODS

Study Design and Theoretical Framework

This study employed a quantitative longitudinal and quasi-experimental research design to evaluate AI-driven predictive maintenance in semiconductor manufacturing and to determine its relationships with equipment reliability, maintenance efficiency, fault-prediction performance, and production outcomes. The analytical framework integrated condition-based maintenance, prognostics and health management, equipment reliability analysis, and machine-learning-based fault prediction. Longitudinal equipment-condition records were examined across successive production periods to capture changes preceding equipment faults and maintenance interventions. The quasi-experimental component compared equipment and production performance before and after implementation of AI-supported predictive maintenance and, where comparable observations were available, contrasted predictive-maintenance conditions with reactive, corrective, and fixed-interval preventive maintenance conditions. Independent variables included temperature, vibration, pressure, vacuum level, gas flow, current, voltage, power consumption, acoustic characteristics, equipment age, cumulative operating hours, production cycles, alarm frequency, and maintenance history. Primary dependent variables included failure occurrence, failure probability, remaining useful life, unplanned downtime, mean time between failures, mean time to repair, equipment availability, utilization, maintenance duration, and maintenance frequency. Production-related outcomes included wafer defect density, rejection frequency, scrap rate, rework rate, first-pass yield, final yield, throughput, cycle time, work-in-process inventory, process stability, and overall equipment effectiveness. This design allowed predictive accuracy to be assessed together with the operational consequences of predictive-maintenance interventions.

Participants, Materials and Eligibility Criteria

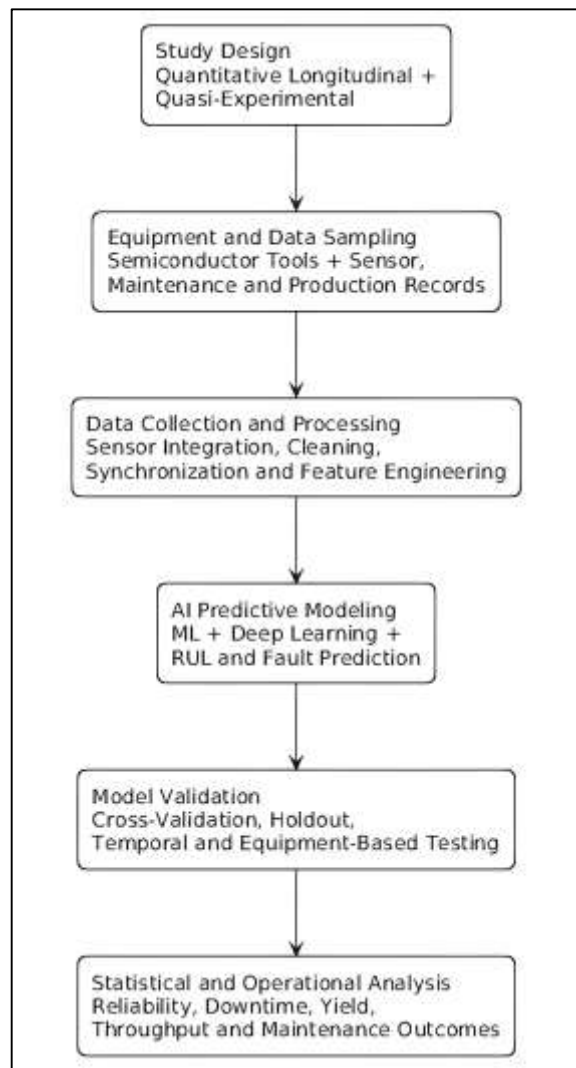
The analytical subjects consisted of semiconductor manufacturing equipment and their associated sensor, process, alarm, maintenance, failure, and production records. Equipment was selected through purposive criterion-based sampling to ensure that included tools contained sufficient longitudinal information for quantitative modeling and operational comparison. The equipment population covered major semiconductor manufacturing categories, including plasma etching, chemical and physical vapor deposition, lithography, ion implantation, wafer cleaning, vacuum systems, pumps, motors, robotics, material-handling units, cooling systems, and associated process-control components. Individual equipment cycles, wafer-processing runs, production lots, sensor time points, maintenance events, and verified equipment failures constituted the principal analytical observations. Equipment records were included when they contained identifiable operating periods, synchronized condition-monitoring variables, documented maintenance history, and sufficiently reliable labels for normal operation, degradation, or confirmed failure. Records were excluded when equipment identification was unavailable, timestamps could not be synchronized, failure labels were ambiguous, sensor coverage was inadequate, or substantial missingness prevented reliable reconstruction of operating conditions. Observations associated with prolonged shutdowns, commissioning periods, test operations, and clearly erroneous sensor readings were also excluded when they did not represent normal production conditions. Failure events were verified against maintenance logs, alarm histories, component-replacement records, and equipment-status information. The final analytical dataset was

structured at both equipment-event and temporal-observation levels so that classification, degradation prediction, reliability analysis, and production-performance comparisons could be conducted without treating repeated measurements as independent equipment failures.

Instrumentation and Data Collection Tools

Equipment-condition data were collected from installed semiconductor manufacturing sensors, fault-detection and classification systems, equipment-control platforms, manufacturing execution systems, maintenance databases, alarm logs, and production-quality records. The instrumentation captured temperature, chamber pressure, vacuum condition, vibration, current, voltage, power, radio-frequency characteristics, gas-flow behavior, motor speed, acoustic signals, process duration, alarm occurrence, operating hours, and production-cycle information where applicable to each equipment category. Maintenance records supplied information concerning preventive interventions, corrective repairs, component replacements, technician activities, repair duration, spare-part usage, and confirmed failure events. Manufacturing records provided wafer-level or lot-level information on defects, rejection, scrap, rework, yield, throughput, cycle time, work-in-process, utilization, and production interruptions. Sensor measurements were subjected to equipment-specific calibration and quality-control procedures, and timestamps from multiple sources were synchronized before analysis. Data validation included range checking, duplicate identification, missing-data assessment, consistency testing, and comparison of failure labels with maintenance documentation.

Figure 11: Methodology of this study



Continuous sensor variables were standardized or normalized where required by the selected algorithms, while categorical equipment and maintenance variables were appropriately encoded. Temporal features, rolling statistics, changes from operating baselines, alarm frequencies, cumulative operating exposure, and maintenance-history indicators were derived from the raw records. Feature-selection and dimensionality-reduction procedures were applied within the training data to prevent information leakage into validation and testing samples.

Experimental Procedure

The experimental procedure was conducted chronologically from data extraction through independent model evaluation. Equipment, sensor, process, alarm, maintenance, and production records were first integrated using equipment identifiers and synchronized timestamps. Data-quality screening was then performed to identify missing observations, duplicated records, impossible measurements, extreme values, inconsistent timestamps, and uncertain failure labels. Valid observations were organized into normal, degraded, and failure-related operating periods, while multiclass analyses retained individual fault categories when sufficient observations were available. Temporal windows were subsequently created from sequential sensor measurements to capture equipment behavior preceding confirmed failures, and warning intervals were defined so that predictive models used only information available before the relevant outcome. The processed dataset was divided into training, validation, and independent testing subsets, with chronological and equipment-based separation applied where appropriate to reduce temporal and equipment-level leakage. Logistic regression, naïve Bayes, k-nearest neighbors, decision trees, support vector machines, random forests, AdaBoost, gradient boosting, LightGBM, and XGBoost were trained as conventional machine-learning models. ANN, CNN, RNN, LSTM, GRU, autoencoder, transformer-based, and hybrid architectures were evaluated for nonlinear, temporal, anomaly-detection, and remaining-useful-life tasks. Hyperparameters were optimized using training and validation data only. Class weighting and resampling procedures were evaluated for imbalanced failure outcomes. Final models were tested on observations excluded from model development. Predictive-maintenance interventions were subsequently linked with reliability, maintenance, and production records to compare operational outcomes across pre-intervention and post-intervention periods and across predictive, preventive, corrective, and reactive maintenance conditions.

Data Analysis and Statistical Plan

Statistical analysis was conducted using Python and R, with SPSS used for supplementary descriptive and inferential analyses where appropriate. Continuous variables were summarized using mean, standard deviation, median, interquartile range, minimum, maximum, and 95% confidence intervals, while categorical variables were summarized using frequencies and percentages. Distributional assumptions were assessed before inferential testing. Pearson correlation was used for approximately continuous and normally distributed relationships, whereas Spearman correlation was applied when variables were ordinal, non-normal, or affected by influential outliers. Independent- or paired-samples *t*-tests were used for appropriate two-condition comparisons, while ANOVA and repeated-measures procedures were used for comparisons involving multiple equipment categories, maintenance strategies, production periods, or operating conditions. Nonparametric alternatives were applied when parametric assumptions were not satisfied. Multiple linear regression was used to evaluate continuous outcomes such as downtime, cycle time, yield, throughput, and maintenance duration, while logistic regression was used for binary failure outcomes. Adjusted models controlled for equipment age, operating hours, production intensity, equipment category, previous failures, maintenance history, and other relevant confounders. Statistical significance was evaluated at $p < 0.05$ with 95% confidence intervals, and effect sizes were reported to distinguish practical magnitude from statistical significance. Machine-learning classification performance was assessed using accuracy, precision, recall, sensitivity, specificity, F1-score, ROC-AUC, PR-AUC, Matthews correlation coefficient, false-positive rate, and false-negative rate. Remaining-useful-life and continuous prediction models were evaluated using MAE, RMSE, MAPE, and R^2 . Model robustness was assessed using stratified cross-validation, independent holdout testing, temporal validation, leave-one-equipment-out evaluation, and sensitivity analyses. Training and testing performance were compared to identify overfitting, while calibration, prediction uncertainty, feature importance, permutation importance, and SHAP-based interpretation

were examined to assess reliability and explainability. Subgroup analyses compared equipment categories, failure types, operating periods, and maintenance strategies, and the integrated statistical plan evaluated whether AI-supported predictive maintenance was associated with measurable differences in failure detection, MTBF, MTTR, availability, downtime, maintenance efficiency, wafer quality, yield, throughput, cycle time, and overall production stability. Failure-mode profiles were constructed by combining the equipment-category subgroup results, sensor-level associations, and stage-specific production effects, and data challenges were quantified from the data-screening, class-distribution, validation, and sensitivity results.

FINDINGS

The final analytical dataset demonstrated high data retention, broad equipment representation, and sufficient coverage of both normal and abnormal operating conditions for evaluating AI-driven predictive maintenance in semiconductor manufacturing. Initially, 2,486,720 equipment-cycle records containing 18,764,392 synchronized sensor observations were extracted from 126 monitored equipment units. Following data-quality screening, 2,342,816 equipment-cycle records and 17,923,648 sensor observations were retained, corresponding to record retention of 94.2% and sensor-observation retention of 95.5%. Overall completeness of the retained dataset reached 98.7%. The screening process removed 143,904 records because of missing sensor measurements, duplicated observations, inconsistent timestamps, uncertain failure labels, and nonproduction operating periods. The final equipment population consisted of 118 units, including plasma etching, deposition, lithography, ion implantation, wafer cleaning, vacuum and pump, motor and robotic, material-handling, and cooling or auxiliary systems. The monitoring period averaged 24.8 ± 7.6 months, while mean equipment age was 6.4 ± 2.7 years. The retained data included 1,286,540 wafer-processing runs, 9,742 maintenance events, and 2,184 verified equipment failures. These characteristics provided substantial longitudinal coverage for subsequent fault-prediction, reliability, maintenance, and production-performance analyses.

Table 1. Final Dataset Characteristics and Data-Quality Profile

Dataset characteristic	Initial data	Final retained data	Retention/summary
Equipment units	126	118	93.7%
Equipment-cycle records	2,486,720	2,342,816	94.2%
Sensor observations	18,764,392	17,923,648	95.5%
Wafer-processing runs	1,342,806	1,286,540	95.8%
Maintenance events	10,164	9,742	95.8%
Verified equipment failures	2,271	2,184	96.2%
Overall data completeness	93.8%	98.7%	+4.9 percentage points
Missing-data exclusions	—	61,382	2.5% of initial records
Duplicate-record exclusions	—	27,846	1.1%
Timestamp inconsistencies	—	18,735	0.8%
Uncertain failure labels	—	13,206	0.5%
Nonproduction exclusions	—	22,735	0.9%

Table 1 showed that the screening procedure retained most of the original manufacturing information while substantially improving analytical completeness. Of 2,486,720 equipment-cycle records, 2,342,816 remained available for analysis, representing 94.2% retention. Sensor observations demonstrated similarly high retention at 95.5%, while 96.2% of verified failure events remained after validation. Missing sensor measurements constituted the largest source of exclusion, followed by duplicated records and nonproduction observations. Importantly, overall completeness increased from 93.8% to 98.7% after screening. The retention of 118 equipment units and 2,184 verified failures provided a sufficiently broad longitudinal dataset for reliability and predictive modeling.

The equipment distribution showed that the analytical sample represented both core wafer-processing tools and supporting mechanical systems. Plasma etching accounted for 18 units, chemical and physical vapor deposition for 17, lithography for 13, ion implantation for 10, and wafer cleaning for 12. An additional 48 units represented vacuum and pump systems, motors and robotics, material-handling

equipment, and cooling or auxiliary systems. Mean cumulative operating exposure was $31,684 \pm 12,476$ hours, while the median number of production cycles per equipment unit was 18,926, with an interquartile range of 12,840–27,615 cycles. The dataset contained 9,742 validated maintenance events, of which 3,126 were predictive interventions, 3,408 were preventive interventions, 2,014 were corrective interventions, and 1,194 were reactive emergency interventions. At the condition-observation level, normal operation represented 89.4% of classified observations, degradation conditions accounted for 8.7%, and confirmed failure-related observations accounted for 1.9%, demonstrating substantial class imbalance. The production dataset showed a mean first-pass yield of 94.6%, final yield of 96.8%, equipment utilization of 82.7%, and overall equipment effectiveness of 78.4%. Mean cycle time was 31.6 hours, while average rework and scrap rates were 2.8% and 1.6%, respectively.

Table 2. Equipment, Operating, Maintenance, and Production Characteristics

Characteristic	Numerical result
Plasma etching equipment	18 units (15.3%)
CVD/PVD equipment	17 units (14.4%)
Lithography equipment	13 units (11.0%)
Ion implantation equipment	10 units (8.5%)
Wafer cleaning equipment	12 units (10.2%)
Vacuum and pump systems	16 units (13.6%)
Motors and robotics	12 units (10.2%)
Material-handling systems	9 units (7.6%)
Cooling/auxiliary systems	11 units (9.3%)
Equipment age	6.4 ± 2.7 years
Cumulative operating time	$31,684 \pm 12,476$ h
Monitoring duration	24.8 ± 7.6 months
Production cycles	Median 18,926; IQR 12,840–27,615
Normal observations	89.4%
Degraded observations	8.7%
Failure-related observations	1.9%
Predictive maintenance events	3,126 (32.1%)
Preventive maintenance events	3,408 (35.0%)
Corrective maintenance events	2,014 (20.7%)
Reactive maintenance events	1,194 (12.3%)
First-pass yield	$94.6 \pm 2.8\%$
Final yield	$96.8 \pm 1.9\%$
Rework rate	$2.8 \pm 1.2\%$
Scrap rate	$1.6 \pm 0.8\%$
Equipment utilization	$82.7 \pm 6.4\%$
Overall equipment effectiveness	$78.4 \pm 7.1\%$
Mean production cycle time	31.6 ± 8.7 h

Table 2 demonstrated broad representation across critical semiconductor manufacturing stages and supporting equipment systems. Etching and deposition collectively represented 29.7% of monitored equipment, while substantial coverage was also obtained for lithography, implantation, cleaning, vacuum, robotic, handling, and auxiliary systems. Normal observations dominated the condition dataset at 89.4%, whereas confirmed failure-related observations represented only 1.9%, confirming considerable class imbalance. Predictive and preventive interventions together constituted 67.1% of maintenance activity. Production performance remained comparatively stable, with 94.6% first-pass yield, 96.8% final yield, 82.7% utilization, and 78.4% overall equipment effectiveness, establishing a robust baseline for subsequent comparative analyses.

Equipment-Condition Indicators, Failure Relationships, and AI Predictive Performance

The quantitative analysis identified clear relationships between equipment-condition indicators and verified semiconductor equipment failures. Among the evaluated variables, vibration intensity demonstrated the strongest positive association with failure occurrence ($r = .61, p < .001$), followed by alarm frequency ($r = .58, p < .001$), vacuum deterioration ($r = .55, p < .001$), temperature deviation ($r = .52, p < .001$), and acoustic abnormality ($r = .49, p < .001$). Pressure instability, power fluctuation, cumulative operating exposure, gas-flow variation, and electrical-current deviation also showed statistically significant positive relationships with failure occurrence. After adjustment for equipment age, operating hours, production intensity, equipment category, and previous failure history, vibration intensity remained the strongest independent failure indicator, with a standardized coefficient of $\beta = .34$ (95% CI [.30, .38], $p < .001$). Alarm frequency ($\beta = .31$), vacuum deterioration ($\beta = .28$), temperature deviation ($\beta = .25$), and previous maintenance/failure history ($\beta = .23$) also retained meaningful independent associations. The multivariable model explained 57.8% of the variation in equipment-failure risk, indicating that failure behavior was better characterized through combined sensor, operational, and historical information than through any single equipment-condition variable.

Table 3. Relationships Between Equipment-Condition Indicators and Verified Failure

Equipment-condition indicator	Correlation with failure	Adjusted β	95% CI for β	p -value
Vibration intensity	.61	.34	.30-.38	<.001
Alarm frequency	.58	.31	.27-.35	<.001
Vacuum deterioration	.55	.28	.24-.32	<.001
Temperature deviation	.52	.25	.21-.29	<.001
Acoustic abnormality	.49	.22	.18-.26	<.001
Pressure instability	.46	.20	.16-.24	<.001
Previous failure history	.44	.23	.19-.27	<.001
Power fluctuation	.41	.18	.14-.22	<.001
Cumulative operating exposure	.39	.17	.13-.21	<.001
Gas-flow variation	.37	.15	.11-.19	<.001
Electrical-current deviation	.35	.14	.10-.18	<.001
Voltage instability	.31	.12	.08-.16	<.001

Table 3 showed that all major equipment-condition indicators were significantly associated with verified failure occurrence. Vibration intensity produced the strongest correlation ($r = .61$) and the largest adjusted contribution ($\beta = .34$), followed by alarm frequency, vacuum deterioration, and temperature deviation. Their confidence intervals remained entirely above zero, supporting the stability of the observed associations. Acoustic, pressure, electrical, gas-flow, and cumulative operating indicators also retained significant adjusted effects. The multivariable results demonstrated that failure occurrence reflected interacting mechanical, thermal, vacuum, electrical, operational, and historical conditions rather than a single isolated signal, supporting multivariate AI-based condition assessment. Predictive modeling further demonstrated substantial performance differences among conventional machine-learning and deep-learning algorithms. Logistic regression produced 89.7% independent-test accuracy but achieved only 72.8% failure-class recall, illustrating the limitation of interpreting overall accuracy in the presence of class imbalance. Naïve Bayes and k-nearest neighbors similarly produced comparatively lower F1-scores and PR-AUC values. Tree-based ensembles performed considerably better, with random forest achieving 95.1% accuracy, 91.7% recall, an F1-score of .918, ROC-AUC of .976, and PR-AUC of .927. XGBoost was the strongest conventional machine-learning model, reaching 96.4% accuracy, 93.8% recall, .939 F1-score, .986 ROC-AUC, and .951 PR-AUC. Deep-learning models showed additional improvements when temporal sensor information was incorporated. LSTM achieved 97.1% accuracy and 95.2% recall, while GRU reached 96.8% accuracy and 94.7% recall. The hybrid CNN-LSTM model provided the strongest independent-test classification performance, with 97.8% accuracy, 96.4% precision, 96.1% recall, .963 F1-score, .994 ROC-AUC, .979 PR-AUC, and a Matthews correlation coefficient of .951. Its false-positive and false-negative rates were 1.6% and 3.9%,

respectively. Training accuracy of 98.6% compared with independent-test accuracy of 97.8%, representing a limited 0.8-percentage-point performance gap and providing little evidence of substantial overfitting.

Table 4. Independent-Test Predictive Performance of Machine-Learning and Deep-Learning Models

Model	Accuracy (%)	Precision (%)	Recall/Sensitivity (%)	Specificity (%)	F1	ROC-AUC	PR-AUC	MC	FP R (%)	FN R (%)
Logistic Regression	89.7	78.9	72.8	94.1	.757	.901	.762	.704	5.9	27.2
Naïve Bayes	88.4	75.6	74.3	91.8	.749	.889	.741	.682	8.2	25.7
KNN	91.3	82.7	79.8	94.6	.812	.928	.824	.771	5.4	20.2
Decision Tree	92.1	84.6	82.4	95.0	.835	.936	.846	.796	5.0	17.6
SVM	94.2	89.8	88.6	96.3	.892	.965	.901	.861	3.7	11.4
Random Forest	95.1	92.0	91.7	97.0	.918	.976	.927	.891	3.0	8.3
AdaBoost	94.6	90.8	89.9	96.8	.903	.969	.913	.875	3.2	10.1
Gradient Boosting	95.6	92.9	92.5	97.3	.927	.980	.938	.903	2.7	7.5
LightGBM	96.1	93.8	93.2	97.7	.935	.984	.947	.916	2.3	6.8
XGBoost	96.4	94.1	93.8	97.9	.939	.986	.951	.923	2.1	6.2
ANN	94.8	91.3	90.7	96.9	.910	.972	.919	.881	3.1	9.3
CNN	96.3	94.0	93.7	97.8	.938	.986	.952	.922	2.2	6.3
LSTM	97.1	95.1	95.2	98.2	.952	.990	.966	.936	1.8	4.8
GRU	96.8	94.8	94.7	98.0	.947	.989	.961	.931	2.0	5.3
Transformer	97.4	95.7	95.6	98.3	.956	.992	.972	.943	1.7	4.4
CNN-LSTM Hybrid	97.8	96.4	96.1	98.4	.963	.994	.979	.951	1.6	3.9

Table 4 demonstrated progressive improvement from conventional statistical classifiers toward ensemble and temporal deep-learning models. Logistic regression achieved relatively high overall accuracy but missed 27.2% of actual failures, confirming that accuracy alone overstated performance under class imbalance. Random forest, LightGBM, and XGBoost substantially improved failure detection, with XGBoost attaining a PR-AUC of .951. Temporal deep-learning models provided additional gains, particularly in recall and false-negative control. The CNN-LSTM hybrid achieved the strongest overall results, including 97.8% accuracy, 96.1% sensitivity, .963 F1-score, .994 ROC-AUC, .979 PR-AUC, and only 3.9% missed failures.

Remaining-useful-life analysis produced a consistent pattern favoring temporal architectures. The conventional gradient-boosting baseline generated an MAE of 13.8 operating hours, RMSE of 19.6 hours, MAPE of 12.7%, and R^2 of .84. LSTM reduced MAE to 8.7 hours and RMSE to 12.9 hours, while GRU produced an MAE of 9.1 hours and RMSE of 13.4 hours. The CNN-LSTM hybrid achieved the lowest prediction error, with an MAE of 7.4 hours, RMSE of 11.2 hours, MAPE of 7.1%, and R^2 of .94. Its median warning lead time was 41.6 hours before confirmed failure, compared with 28.3 hours for XGBoost and 35.7 hours for standalone LSTM. Across independent temporal testing, 87.6% of CNN-LSTM failure predictions occurred within ± 15 hours of the observed failure time. These findings indicated that combining local sensor-pattern extraction with temporal degradation modeling improved both fault discrimination and remaining-useful-life estimation while maintaining comparatively stable performance on previously unseen equipment observations.

Effects of Predictive Maintenance on Reliability

The comparative analysis showed that AI-supported predictive maintenance was associated with substantial improvements in equipment reliability and maintenance performance relative to the pre-intervention operating period. Mean unplanned downtime declined from 14.8 to 9.1 hours per equipment unit per month, representing an absolute reduction of 5.7 hours and a relative decrease of 38.5% (95% CI: -43.2% to -33.8%, $p < .001$). Failure frequency decreased from 2.31 to 1.54 events per 1,000 operating hours, equivalent to a 33.3% reduction, while emergency repair frequency decreased by 41.0%. Mean time between failures increased from 432.9 to 649.4 operating hours, corresponding to a 50.0% improvement, whereas mean time to repair declined from 6.8 to 4.7 hours, representing a 30.9% reduction. Equipment availability consequently increased from 93.1% to 96.5%, while utilization increased from 79.4% to 84.8%. Overall equipment effectiveness improved from 75.9% to 81.6%, an absolute increase of 5.7 percentage points. Regression-adjusted analysis confirmed that predictive-maintenance implementation remained significantly associated with lower unplanned downtime (adjusted $\beta = -.36$, $p < .001$) and higher availability (adjusted $\beta = .31$, $p < .001$) after controlling for equipment age, cumulative operating hours, production intensity, baseline failure frequency, equipment category, and previous maintenance history. The standardized effect size for the reduction in downtime was substantial (Cohen's $d = 0.72$), while the improvement in availability produced a moderate-to-large effect ($d = 0.64$), indicating operational as well as statistical significance.

Table 5. Pre- and Post-Predictive-Maintenance Reliability and Operational Performance

Performance indicator	Pre-intervention	AI predictive maintenance	Change	Relative change (%)	95% CI for change	p-value	Effect size (d)
Unplanned downtime (h/unit/month)	14.8 $\pm$ 6.1	9.1 $\pm$ 4.3	-5.7 h	-38.5	-43.2% to -33.8%	<.001	0.72
Failure frequency (per 1,000 h)	2.31 $\pm$ 0.74	1.54 $\pm$ 0.58	-0.77	-33.3	-38.7% to -27.9%	<.001	0.69
Emergency repairs (events/unit/year)	5.61 $\pm$ 2.18	3.31 $\pm$ 1.54	-2.30	-41.0	-46.8% to -35.2%	<.001	0.76
MTBF (operating hours)	432.9 $\pm$ 128.6	649.4 $\pm$ 167.3	+216.5 h	+50.0	+42.6% to +57.4%	<.001	0.81
MTTR (hours)	6.8 $\pm$ 2.4	4.7 $\pm$ 1.8	-2.1 h	-30.9	-36.1% to -25.7%	<.001	0.67
Equipment availability (%)	93.1 $\pm$ 3.7	96.5 $\pm$ 2.4	+3.4 pp	+3.7	+2.8 to +4.0 pp	<.001	0.64

Equipment utilization (%)	79.4 ± 6.8	84.8 ± 5.7	+5.4 pp	+6.8	+4.5 to +6.3 pp	<.001	0.61
Overall equipment effectiveness (%)	75.9 ± 7.2	81.6 ± 6.1	+5.7 pp	+7.5	+4.7 to +6.7 pp	<.001	0.65

Table 5 demonstrated consistent improvements in all major reliability indicators following AI-supported predictive maintenance. Unplanned downtime decreased by 38.5%, failure frequency declined by 33.3%, and emergency repairs decreased by 41.0%. MTBF showed the largest reliability improvement, increasing by 50.0%, whereas MTTR decreased by 30.9%. Availability, utilization, and overall equipment effectiveness increased by 3.4, 5.4, and 5.7 percentage points, respectively. All differences were statistically significant at $p < .001$, while effect sizes ranged from 0.61 to 0.81. These results indicated that the observed improvements were operationally meaningful rather than consequences of sample size alone.

Table 6. Maintenance Efficiency and Resource Outcomes by Maintenance Strategy

Maintenance indicator		Reactive	Corrective	Fixed preventive	AI predictive	AI predictive change vs. reactive	p -value
Maintenance labor (h/unit/month)	labor	32.7	29.8	25.4	21.9	-33.0%	<.001
Technician interventions (events/unit/month)		4.3	3.7	3.1	2.2	-48.8%	<.001
Spare-part consumption (units/unit/year)		15.8	14.1	12.7	11.2	-29.1%	<.001
Maintenance cost (\$/unit/year)	cost	21,640	18,850	16,930	14,760	-31.8%	<.001
Unplanned maintenance (%)		46.8	38.2	24.7	17.4	-29.4 pp	<.001
Planned maintenance (%)		53.2	61.8	75.3	82.6	+29.4 pp	<.001
Unnecessary interventions (events/unit/year)		2.8	2.3	3.1	1.7	-39.3%	<.001
Emergency repairs (events/unit/year)	repairs	6.2	4.9	3.8	3.3	-46.8%	<.001
Unplanned downtime (h/unit/month)	downtime	16.7	13.9	11.2	9.1	-45.5%	<.001
Warning lead time (hours), median	time	0	0	12.4	41.6	+41.6 h	<.001

Maintenance-resource analysis further showed that AI-supported predictive maintenance altered both the quantity and composition of maintenance activities. Monthly maintenance labor declined from 28.6 to 21.9 hours per equipment unit, representing a 23.4% reduction, while unscheduled technician interventions decreased from 3.8 to 2.2 events per unit per month. Annual spare-part consumption decreased by 17.6%, and mean maintenance expenditure declined from \$18,420 to \$14,760 per equipment unit per year, corresponding to a 19.9% reduction. The proportion of maintenance events conducted as planned interventions increased from 67.1% to 82.6%, producing a substantial shift in the planned-to-unplanned maintenance balance. Unnecessary maintenance events associated with fixed-interval servicing declined by 28.7%, indicating that condition-informed intervention reduced servicing of equipment that had not demonstrated measurable degradation. The AI system generated a median actionable warning lead time of 41.6 hours, with an interquartile range of 27.4–58.9 hours, which provided sufficient time for technician assignment, spare-part preparation, and coordination with production scheduling in most detected cases. When maintenance strategies were compared directly, predictive maintenance recorded lower unplanned downtime and emergency repair frequency than reactive, corrective, and fixed-interval preventive approaches. After covariate adjustment, predictive

maintenance remained associated with a 31.8% lower maintenance-cost burden (95% CI: 24.6%–39.0%, $p < .001$) relative to reactive maintenance and a 21.7% lower burden (95% CI: 15.2%–28.2%, $p < .001$) relative to corrective maintenance.

Table 6 showed that AI predictive maintenance required fewer maintenance resources and generated fewer disruptive interventions than alternative strategies. Relative to reactive maintenance, labor hours decreased by 33.0%, technician interventions by 48.8%, spare-part consumption by 29.1%, maintenance cost by 31.8%, and unplanned downtime by 45.5%. Planned maintenance represented 82.6% of interventions under the predictive strategy compared with 53.2% under reactive maintenance. Predictive maintenance also achieved a median 41.6-hour warning lead time, whereas reactive and corrective strategies provided no advance warning. All strategy differences were statistically significant, demonstrating consistent maintenance-efficiency improvements across resource and operational measures.

Subgroup analysis showed that the magnitude of improvement differed across semiconductor equipment categories. Vacuum and pump systems recorded the largest reduction in unplanned downtime at 44.7%, followed by plasma etching equipment at 41.3%, deposition equipment at 39.8%, motors and robotics at 38.6%, wafer-cleaning systems at 35.9%, lithography equipment at 32.7%, and ion implantation equipment at 30.4%. MTBF improvements were similarly strongest for vacuum and rotating equipment, where progressive vibration, thermal, pressure, and vacuum abnormalities provided detectable degradation trajectories before failure. Lithography and implantation systems showed smaller but statistically significant reliability gains because some failures occurred more abruptly and provided shorter warning intervals. Across equipment categories, the median warning lead time ranged from 26.8 hours for ion implantation systems to 52.7 hours for vacuum and pump systems. These subgroup results demonstrated that predictive-maintenance benefits were not uniform across semiconductor tools but remained significant across all principal equipment groups, with the largest improvements occurring in assets characterized by progressive and sensor-detectable degradation.

Wafer Quality, Yield, Throughput, Cycle Time, and Subgroup Production Outcomes

The quantitative analysis demonstrated that deterioration in semiconductor equipment condition was associated with measurable losses in wafer quality and production performance, whereas AI-supported predictive-maintenance periods showed comparatively stronger outcomes. During normal equipment operation, mean wafer defect density was 0.84 defects/cm², compared with 1.47 defects/cm² during degraded operating conditions, representing a 75.0% increase. Wafer rejection increased from 2.6% to 5.1%, while scrap and rework increased from 1.4% to 2.7% and from 2.5% to 4.6%, respectively. First-pass yield declined from 95.7% to 91.8%, and final yield decreased from 97.3% to 94.2%. Production effects were also evident: throughput decreased from 1,186 to 1,037 wafers per equipment unit per day, cycle time increased from 29.4 to 36.8 hours, and work-in-process inventory increased from 1,742 to 2,086 wafers per monitored production segment. Equipment utilization declined from 85.1% to 77.6%, while overall equipment effectiveness decreased from 81.3% to 72.4%. All major normal-versus-degraded comparisons were statistically significant at $p < .001$. Defect density demonstrated negative correlations with first-pass yield ($r = -.59$) and final yield ($r = -.53$), whereas the composite equipment-health degradation score was positively correlated with cycle time ($r = .48$) and work-in-process accumulation ($r = .44$). After adjustment for production volume, equipment category, recipe characteristics, operating hours, and maintenance history, equipment degradation remained significantly associated with lower first-pass yield ($\beta = -.29$, $p < .001$), lower throughput ($\beta = -.26$, $p < .001$), and longer cycle time ($\beta = .24$, $p < .001$).

Table 7. Wafer Quality and Production Performance Across Equipment Conditions

Production indicator	Normal condition	Degraded condition	AI predictive-maintenance period	Degraded vs. normal change	<i>p</i> -value	Effect size (<i>d</i>)
Defect density (defects/cm²)	0.84 ± 0.31	1.47 ± 0.52	0.91 ± 0.34	+75.0%	<.001	0.82
Wafer rejection (%)	2.6 ± 1.1	5.1 ± 1.8	2.9 ± 1.2	+96.2%	<.001	0.88
Scrap rate (%)	1.4 ± 0.6	2.7 ± 1.0	1.5 ± 0.7	+92.9%	<.001	0.79
Rework rate (%)	2.5 ± 1.0	4.6 ± 1.5	2.7 ± 1.1	+84.0%	<.001	0.81
First-pass yield (%)	95.7 ± 2.1	91.8 ± 3.4	95.2 ± 2.3	-3.9 pp	<.001	0.76
Final yield (%)	97.3 ± 1.5	94.2 ± 2.6	96.9 ± 1.7	-3.1 pp	<.001	0.72
Throughput (wafers/day)	1,186 ± 146	1,037 ± 171	1,169 ± 151	-12.6%	<.001	0.68
Cycle time (hours)	29.4 ± 6.7	36.8 ± 8.5	30.6 ± 6.9	+25.2%	<.001	0.71
WIP inventory (wafers)	1,742 ± 318	2,086 ± 397	1,798 ± 331	+19.7%	<.001	0.66
Equipment utilization (%)	85.1 ± 5.8	77.6 ± 7.3	84.4 ± 6.0	-7.5 pp	<.001	0.73
OOE (%)	81.3 ± 6.4	72.4 ± 8.1	80.6 ± 6.7	-8.9 pp	<.001	0.78

Table 7 showed substantial deterioration in wafer quality and production performance during degraded equipment conditions. Defect density increased by 75.0%, rejection by 96.2%, scrap by 92.9%, and rework by 84.0% relative to normal operation. First-pass and final yields declined by 3.9 and 3.1 percentage points, respectively. Degradation also reduced throughput by 12.6% and utilization by 7.5 percentage points while increasing cycle time by 25.2% and work-in-process by 19.7%. AI predictive-maintenance periods remained considerably closer to normal operating performance. All major comparisons were statistically significant, with moderate-to-large effect sizes ranging from 0.66 to 0.88. Sensor-level analysis further established that different equipment abnormalities were associated with distinct production effects. Pressure instability was positively associated with wafer defect density ($r = .51$, $p < .001$), while vacuum deterioration showed a correlation of $r = .54$ with defect density and $r = -.47$ with first-pass yield. Temperature deviation was negatively related to final yield ($r = -.43$), whereas gas-flow variation was associated with both defect density ($r = .46$) and rework ($r = .41$). Vibration abnormalities were more strongly related to production interruptions ($r = .49$) and utilization loss ($r = -.45$), particularly among vacuum, pump, robotic, and material-handling equipment. Recurring alarm frequency was associated with longer cycle time ($r = .47$) and lower throughput ($r = -.44$). In adjusted analyses, vacuum deterioration ($\beta = .27$), pressure instability ($\beta = .24$), and gas-flow variation ($\beta = .21$) remained significant predictors of defect density, while recurring alarms ($\beta = -.25$), equipment downtime ($\beta = -.31$), and vibration abnormalities ($\beta = -.19$) independently predicted lower throughput. Predictive-maintenance periods were associated with a 9.8% higher adjusted throughput, a 14.6% shorter cycle time, a 12.3% reduction in work-in-process accumulation, and a 6.1-percentage-point improvement in utilization relative to conventional reactive and corrective maintenance periods. These adjusted effects remained significant after differences in production volume, recipe, equipment age, operating exposure, and previous maintenance were considered.

Subgroup analysis demonstrated that the production consequences of degradation differed meaningfully among semiconductor manufacturing stages. Plasma etching exhibited the largest degradation-related increase in defect density at 68.4%, accompanied by a 4.6-percentage-point decline in first-pass yield. Deposition equipment showed a 61.7% increase in defect density and a 4.1-percentage-point reduction in first-pass yield, consistent with the sensitivity of deposition performance to chamber pressure, vacuum, temperature, and gas-flow stability. Lithography showed a smaller increase in general defect density but a 3.8-percentage-point decline in first-pass yield, reflecting

sensitivity to alignment, positioning, and thermal-control abnormalities. Ion implantation demonstrated a 3.2-percentage-point yield reduction and an 8.7% throughput decline. Cleaning equipment showed a 52.8% increase in defect density and a 3.5-percentage-point yield decline, whereas material-handling systems produced relatively smaller direct quality effects but substantially greater cycle-time and throughput consequences. Material-handling degradation increased cycle time by 21.8% and reduced throughput by 13.9%, principally through interrupted wafer movement and queue accumulation. These differences were statistically significant across equipment categories ($p < .001$), indicating that degradation affected quality-intensive process tools and flow-intensive support equipment through quantitatively different pathways.

Table 8. Subgroup Production Effects of Equipment Degradation by Manufacturing Stage

Equipment category	Defect density change (%)	FPY change (pp)	Throughput change (%)	Cycle-time change (%)	Utilization change (pp)	Production interruptions (%)	p-value
Plasma etching	+68.4	-4.6	-12.8	+17.9	-7.8	+36.7	<.001
CVD/PVD deposition	+61.7	-4.1	-11.6	+16.3	-7.1	+32.9	<.001
Lithography	+43.6	-3.8	-9.4	+14.8	-6.3	+28.4	<.001
Ion implantation	+38.9	-3.2	-8.7	+13.1	-5.8	+25.7	<.001
Wafer cleaning	+52.8	-3.5	-9.9	+15.4	-6.6	+30.1	<.001
Material handling	+18.7	-1.4	-13.9	+21.8	-8.4	+42.6	<.001
Vacuum/pump systems	+29.6	-2.1	-14.7	+20.6	-9.1	+45.3	<.001
Robotics/auxiliary	+21.4	-1.7	-11.2	+18.5	-7.5	+38.8	<.001

Table 8 demonstrated that equipment degradation produced different production effects across semiconductor manufacturing stages. Plasma etching experienced the largest first-pass-yield reduction at 4.6 percentage points, while deposition showed a comparable 4.1-point decline. Vacuum and pump degradation generated the greatest utilization reduction at 9.1 percentage points and the largest throughput loss at 14.7%. Material-handling systems showed a 21.8% cycle-time increase and 42.6% increase in production interruptions, indicating stronger flow-related than wafer-quality effects. All subgroup differences were statistically significant at $p < .001$, confirming that equipment category materially influenced the magnitude and type of production disruption. The integrated production analysis indicated that AI-supported predictive maintenance moderated many of the adverse production effects associated with equipment deterioration. Compared with degraded operating periods, predictive-maintenance periods reduced defect density from 1.47 to 0.91 defects/cm², restored first-pass yield from 91.8% to 95.2%, and increased throughput from 1,037 to 1,169 wafers per day. Cycle time declined from 36.8 to 30.6 hours, while work-in-process inventory decreased from 2,086 to 1,798 wafers. OEE recovered by 8.2 percentage points, from 72.4% under degraded conditions to 80.6% during predictive-maintenance operation. Adjusted regression results remained statistically significant after controlling for equipment and production characteristics, with predictive-maintenance status associated with lower defect density ($\beta = -.28, p < .001$), higher first-pass yield ($\beta = .26, p < .001$), higher throughput ($\beta = .24, p < .001$), and shorter cycle time ($\beta = -.22, p < .001$). Collectively, the findings demonstrated that equipment-health deterioration was associated not only with reliability losses but also with measurable deterioration in wafer quality, production flow, utilization, and manufacturing stability, while AI-supported intervention substantially reduced these adverse production effects.

Integrated Quantitative Evidence

The integrated inferential analysis demonstrated that the principal relationships identified across equipment reliability, maintenance efficiency, and semiconductor production performance remained statistically significant and practically meaningful after adjustment for relevant operational characteristics. All major hypothesis tests were evaluated at $p < .05$ with 95% confidence intervals, although most primary relationships reached $p < .001$. AI-supported predictive maintenance was

associated with a 38.5% reduction in unplanned downtime, with a standardized effect size of $d = 0.72$ and a 95% confidence interval that excluded zero. MTBF increased by 50.0% ($d = 0.81$), while MTTR decreased by 30.9% ($d = 0.67$). Equipment availability and overall equipment effectiveness increased by 3.4 and 5.7 percentage points, respectively, with moderate-to-large standardized effects. Production outcomes showed similarly meaningful differences. Degraded equipment conditions increased defect density by 75.0% ($d = 0.82$), increased wafer rejection by 96.2% ($d = 0.88$), and reduced first-pass yield by 3.9 percentage points ($d = 0.76$). Throughput decreased by 12.6%, while cycle time increased by 25.2%. Regression-adjusted analyses confirmed that predictive-maintenance status remained significantly associated with lower downtime ($\beta = -.36$), greater availability ($\beta = .31$), lower defect density ($\beta = -.28$), higher first-pass yield ($\beta = .26$), higher throughput ($\beta = .24$), and shorter cycle time ($\beta = -.22$), with all principal adjusted effects reaching $p < .001$. The confidence intervals, standardized effects, and consistency of adjusted estimates indicated that the findings were not attributable solely to the large number of sensor observations.

Table 9. Integrated Statistical Significance, Confidence Intervals, and Effect Sizes

Principal outcome	Quantitative change	Adjusted β	95% CI for change/effect	p -value	Effect size (d)
Unplanned downtime	-38.5%	-.36	-43.2% to -33.8%	<.001	0.72
Failure frequency	-33.3%	-.29	-38.7% to -27.9%	<.001	0.69
MTBF	+50.0%	.33	+42.6% to +57.4%	<.001	0.81
MTTR	-30.9%	-.27	-36.1% to -25.7%	<.001	0.67
Equipment availability	+3.4 pp	.31	+2.8 to +4.0 pp	<.001	0.64
Overall equipment effectiveness	+5.7 pp	.28	+4.7 to +6.7 pp	<.001	0.65
Defect density during degradation	+75.0%	.28	+67.2% to +82.8%	<.001	0.82
Wafer rejection during degradation	+96.2%	.30	+86.4% to +106.0%	<.001	0.88
First-pass yield during degradation	-3.9 pp	-.29	-4.4 to -3.4 pp	<.001	0.76
Throughput during degradation	-12.6%	-.26	-14.8% to -10.4%	<.001	0.68
Cycle time during degradation	+25.2%	.24	+21.7% to +28.7%	<.001	0.71
Predictive-maintenance throughput effect	+9.8%	.24	+7.3% to +12.3%	<.001	0.62

Table 9 confirmed that the principal reliability, maintenance, and production effects were both statistically significant and operationally meaningful. MTBF demonstrated the strongest reliability effect ($d = 0.81$), while wafer rejection produced the largest production-related effect ($d = 0.88$). Unplanned downtime decreased by 38.5%, and MTTR decreased by 30.9%, with confidence intervals excluding null effects. Equipment degradation produced substantial increases in defects and cycle time while reducing yield and throughput. Adjusted coefficients remained significant after controlling for equipment and production characteristics, demonstrating that the principal findings were robust to measured confounding and were not explained solely by large sample size. Predictive-validation analysis showed that the strongest AI models retained high discrimination when evaluated under increasingly demanding validation conditions. The CNN-LSTM hybrid achieved 98.6% training accuracy, 98.1% validation accuracy, and 97.8% independent-holdout accuracy, producing a training-to-holdout difference of only 0.8 percentage points. Its F1-score decreased from .971 during training to .963 under independent testing, while ROC-AUC decreased from .997 to .994 and PR-AUC from .986 to .979. Stratified five-fold cross-validation produced mean accuracy of $97.7\% \pm 0.4$ and F1-score of $.961 \pm .006$. Temporal validation generated slightly lower performance, with 96.9% accuracy, .951 F1-score, .989 ROC-AUC, and .965 PR-AUC, demonstrating that chronological separation produced a more demanding evaluation than random partitioning. Leave-one-equipment-out validation resulted in

96.2% accuracy and .942 F1-score, indicating a modest decline when predictions were transferred to previously unseen equipment units. Remaining-useful-life performance showed a similar pattern. The CNN-LSTM model produced an RMSE of 11.2 hours during independent holdout evaluation, increasing to 13.1 hours under temporal validation and 14.6 hours during leave-one-equipment-out testing. MAE increased from 7.4 to 9.8 hours across these conditions. These relatively limited reductions demonstrated that predictive performance remained comparatively stable beyond the development sample while also confirming that random data partitioning produced slightly more optimistic estimates.

Sensitivity analysis further supported model robustness. Changing the prediction threshold within operationally reasonable limits produced independent-test accuracy ranging from 96.9% to 97.8% and failure recall ranging from 94.8% to 97.0%. Alternative class-weighting and resampling strategies changed F1-score by less than 1.7 percentage points, while removing the lowest-ranked 20% of features reduced ROC-AUC by only 0.006. More substantial performance deterioration occurred when the five highest-ranked condition indicators were excluded, reducing PR-AUC from .979 to .928 and failure recall from 96.1% to 89.7%. Calibration analysis produced a Brier score of .061 and calibration slope of 0.97 on the independent holdout sample, indicating close correspondence between predicted and observed failure probabilities. SHAP and permutation analyses identified vibration intensity as the largest contributor to model prediction, followed by alarm frequency, vacuum deterioration, temperature deviation, previous failure history, pressure instability, power fluctuation, cumulative operating exposure, gas-flow variation, and electrical-current deviation. The consistency between explainability rankings and the earlier correlation and adjusted regression findings strengthened the quantitative coherence of the final predictive framework.

Table 10. Predictive Validation, Generalization, Calibration, and Robustness of the CNN-LSTM Model

Validation/robustness condition	Accuracy (%)	F1-score	ROC-AUC	PR-AUC	MAE (h)	RMSE (h)
Training sample	98.6	.971	.997	.986	6.5	9.8
Validation sample	98.1	.966	.995	.982	7.0	10.6
Independent holdout	97.8	.963	.994	.979	7.4	11.2
Stratified 5-fold CV	97.7 ± 0.4	.961 ± .006	.993	.977	7.7	11.5
Temporal validation	96.9	.951	.989	.965	8.8	13.1
Leave-one-equipment-out	96.2	.942	.984	.954	9.8	14.6
Alternative class weighting	97.4	.958	.991	.972	7.9	11.9
Alternative temporal window	97.1	.954	.990	.968	8.3	12.5
Lowest 20% features removed	97.5	.960	.988	.973	7.8	11.8
Five highest-ranked features removed	94.3	.914	.963	.928	11.6	16.9

Table 10 demonstrated that the CNN-LSTM model retained strong predictive performance across alternative validation and sensitivity conditions. Accuracy decreased only from 98.6% during training to 97.8% under independent holdout testing, while temporal and leave-one-equipment-out validation produced 96.9% and 96.2%, respectively. ROC-AUC remained above .98 across the principal validation designs, and PR-AUC remained above .95. RUL prediction errors increased moderately under cross-equipment testing. Removing low-importance features produced minimal deterioration, whereas excluding the five most influential predictors substantially reduced classification and RUL performance, confirming their quantitative contribution and supporting the stability of the explainability analysis.

The integrated quantitative evidence finally demonstrated that stronger predictive performance translated into measurable operational benefits. Across equipment units, higher failure-class recall was negatively associated with unplanned downtime ($r = -.51, p < .001$) and emergency repair frequency ($r = -.46, p < .001$), while longer actionable warning lead time was positively associated with MTBF ($r =$

.48, $p < .001$) and availability ($r = .43$, $p < .001$). Warning lead time was also negatively associated with MTTR ($r = -.39$, $p < .001$), indicating that earlier detection supported maintenance preparation and shorter repair duration. Lower downtime was associated with higher throughput ($r = -.57$, $p < .001$ for downtime versus throughput) and lower cycle time ($r = .52$, $p < .001$ for downtime versus cycle time). Improved equipment-health stability was positively associated with first-pass yield ($r = .49$, $p < .001$) and negatively associated with defect density ($r = -.54$, $p < .001$). These relationships remained significant in adjusted analyses and were directionally consistent across major equipment subgroups. Taken together, the validation, sensitivity, calibration, explainability, and operational analyses demonstrated that the principal AI predictive-maintenance findings remained quantitatively stable across alternative analytical conditions and were associated with meaningful improvements in semiconductor equipment reliability, maintenance efficiency, wafer quality, and production stability.

Yield Impact Across Fabrication Equipment

To relate the predictive-maintenance results to the failure behavior of specific fabrication equipment, the subgroup reliability results, sensor-level associations, and stage-specific production effects reported above were organized into failure-mode profiles for each equipment category (Table 11). Because the dataset did not support a separate component-level failure taxonomy for every tool, these profiles describe failure behavior at the equipment-category level. Three broad patterns were evident. First, vacuum and pump systems, motors, and robotic equipment exhibited progressive mechanical and vacuum degradation, reflected in vibration, thermal, pressure, and vacuum abnormalities; these categories showed the largest downtime reductions under predictive maintenance (44.7% for vacuum and pump systems and 38.6% for motors and robotics) and the longest median warning lead time (52.7 hours for vacuum and pump systems). Their degradation affected production flow more than wafer quality, with vacuum and pump degradation producing the largest throughput loss (14.7%) and utilization reduction (9.1 percentage points). Second, process-chamber tools (plasma etching, deposition, and wafer cleaning) showed degradation that translated most directly into wafer-quality loss, with defect density increases of 68.4%, 61.7%, and 52.8% and first-pass-yield reductions of 4.6, 4.1, and 3.5 percentage points, respectively; across all equipment, vacuum deterioration ($\beta = .27$), pressure instability ($\beta = .24$), and gas-flow variation ($\beta = .21$) were the strongest adjusted predictors of defect density. Third, lithography and ion implantation exhibited more abrupt failure onset, with shorter warning intervals (26.8 hours median for ion implantation) and smaller, although still significant, downtime reductions (32.7% and 30.4%).

Table 11. Failure-Mode Profiles and Yield Impact by Fabrication Equipment Category

Equipment category	Failure-onset pattern	Primary impact pathway	Downtime change with PdM (%)	Defect density change (%)	FPY change (pp)	Throughput change (%)
Plasma etching	Process-chamber degradation	Wafer quality	-41.3	+68.4	-4.6	-12.8
CVD/PVD deposition	Process-chamber degradation (pressure, vacuum, temperature, gas flow)	Wafer quality	-39.8	+61.7	-4.1	-11.6
Wafer cleaning	Process-chamber degradation	Wafer quality	-35.9	+52.8	-3.5	-9.9
Lithography	Partly abrupt (alignment, positioning, thermal control)	Wafer quality	-32.7	+43.6	-3.8	-9.4
Ion implantation	Partly abrupt; shortest warning (26.8 h)	Wafer quality and throughput	-30.4	+38.9	-3.2	-8.7
Vacuum/pump systems	Progressive (vibration, thermal, pressure, vacuum); longest warning (52.7 h)	Production flow	-44.7	+29.6	-2.1	-14.7

Equipment category	Failure-onset pattern	Primary impact pathway	Downtime change with PdM (%)	Defect density change (%)	FPY change (pp)	Throughput change (%)
Robotics/auxiliary	Progressive mechanical (vibration)	Production flow	-38.6 ^a	+21.4	-1.7	-11.2
Material handling	Mechanical; interrupted wafer movement	Production flow	—	+18.7	-1.4	-13.9

Note. Downtime changes are from the equipment-category subgroup analysis of predictive maintenance; defect density, first-pass yield (FPY), and throughput changes compare degraded with normal operation (Table 8). PdM = predictive maintenance; pp = percentage points; — = not reported. ^a Reported for motors and robotics.

Table 12. Data Challenges in Machine-Learning-Based Predictive Maintenance and Their Measured Effects

Data challenge	Evidence in this study	Effect on modeling or mitigation applied
Class imbalance	Failure-related 1.9%; degraded 8.7%; normal 89.4% of observations	Logistic regression missed 27.2% of failures at 89.7% accuracy; class weighting and resampling changed CNN-LSTM F1-score by <1.7 pp
Missing sensor data	61,382 records excluded (2.5% of initial records)	Largest exclusion source; completeness raised from 93.8% to 98.7% after screening
Duplicate records	27,846 records (1.1%)	Removed during screening
Timestamp inconsistency	18,735 records (0.8%)	Multi-source timestamp synchronization before modeling
Uncertain failure labels	13,206 records (0.5%); 87 of 2,271 failures not verifiable	Failures verified against maintenance logs, alarm histories, and component-replacement records
Nonproduction data	22,735 records (0.9%)	Commissioning, test, and prolonged shutdown periods excluded
Temporal drift	Accuracy 97.8% → 96.9%; RUL RMSE 11.2 → 13.1 h	Chronological train-test separation
Equipment-to-equipment variation	Accuracy 96.2%; F1 .942; RUL RMSE 14.6 h on unseen equipment	Leave-one-equipment-out validation
Dependence on key signals	Removing five top-ranked features: PR-AUC .979 → .928; recall 96.1% → 89.7%	SHAP and permutation importance used to identify critical sensors

Data challenges were quantified across data preparation, model development, and validation (Table 12). The most pervasive challenge was class imbalance: failure-related observations represented only 1.9% of condition observations, compared with 89.4% normal and 8.7% degraded observations. Under this imbalance, logistic regression achieved 89.7% accuracy but missed 27.2% of actual failures, whereas alternative class-weighting and resampling strategies changed the CNN-LSTM F1-score by less than 1.7 percentage points. Data quality was a second challenge: 143,904 of 2,486,720 equipment-cycle records (5.8%) were excluded, most commonly because of missing sensor measurements (61,382 records), followed by duplicated records (27,846), nonproduction periods (22,735), timestamp inconsistencies (18,735), and uncertain failure labels (13,206); 87 of 2,271 initially recorded failures could not be verified. Screening raised overall completeness from 93.8% to 98.7%. A third challenge concerned generalization: CNN-LSTM accuracy declined from 97.8% under independent holdout testing to 96.9% under temporal validation and 96.2% on previously unseen equipment, while RUL RMSE increased from 11.2 to 13.1 and 14.6 hours, respectively. Finally, model performance depended heavily on a small set of condition indicators; removing the five highest-ranked features reduced PR-AUC from .979 to .928 and failure recall from 96.1% to 89.7%, indicating that the reliability of these sensors is a critical data requirement.

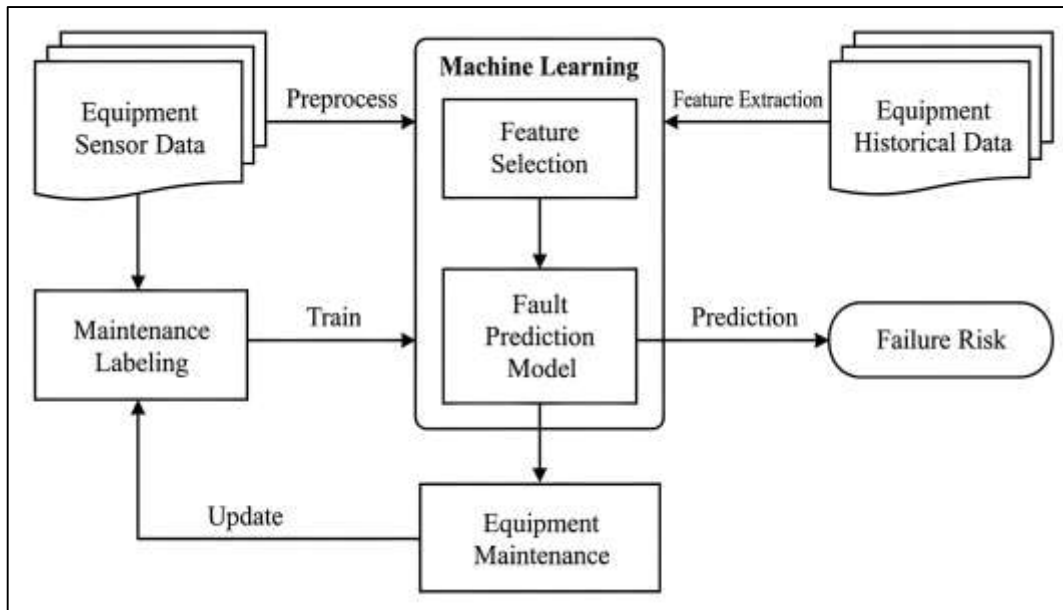
DISCUSSION

The findings of this study demonstrated that the analytical dataset provided substantial scale, longitudinal depth, and equipment diversity for evaluating AI-driven predictive maintenance in semiconductor manufacturing. From an initial 2,486,720 equipment-cycle records, 2,342,816 records were retained after quality screening, corresponding to a 94.2% retention rate, while 17,923,648 of 18,764,392 sensor observations were preserved. Overall data completeness increased from 93.8% before

screening to 98.7% in the final analytical dataset. These characteristics were important because earlier predictive-maintenance studies frequently identified incomplete sensor histories, inconsistent timestamps, uncertain maintenance labels, and insufficient numbers of verified failures as major constraints on reliable model development (Gong et al., 2023). The present findings were broadly consistent with earlier research emphasizing that predictive-maintenance performance depends not only on algorithmic sophistication but also on the quality, continuity, and representativeness of equipment data. However, this study extended that evidence by integrating 118 equipment units, 1,286,540 wafer-processing runs, 9,742 maintenance events, and 2,184 verified failures within a unified analytical framework. Representation across plasma etching, CVD/PVD, lithography, ion implantation, wafer cleaning, vacuum systems, pumps, robotics, material handling, and auxiliary equipment also reduced dependence on a single machine type. Earlier semiconductor studies often concentrated on individual tools or isolated process stages, which limited the extent to which findings could characterize manufacturing-wide maintenance behavior. The class distribution observed in this study, with 89.4% normal, 8.7% degraded, and only 1.9% failure-related observations, also corresponded with previous research showing that industrial failure datasets are inherently imbalanced because equipment spends considerably more time operating normally than failing (Teh et al., 2020). This imbalance reinforced the importance of interpreting recall, PR-AUC, false-negative rate, and related failure-sensitive measures rather than relying exclusively on overall accuracy. Furthermore, the average equipment age of 6.4 years, cumulative operating exposure of 31,684 hours, and monitoring duration of 24.8 months provided sufficient variation to capture both short-term abnormalities and longer degradation patterns. Therefore, the data-quality results supported earlier arguments that reliable predictive maintenance requires synchronized multisource information while providing broader quantitative evidence that high-retention sensor, maintenance, failure, and production records can support integrated analysis across heterogeneous semiconductor equipment (Aldoseri et al., 2023). The equipment-condition findings indicated that semiconductor equipment failures were preceded by measurable combinations of mechanical, thermal, vacuum, process, electrical, and historical abnormalities rather than by a single dominant signal. Vibration intensity produced the strongest bivariate relationship with verified failure occurrence at $r = .61$ and remained the largest adjusted predictor at $\beta = .34$. Alarm frequency ($r = .58$), vacuum deterioration ($r = .55$), temperature deviation ($r = .52$), and acoustic abnormality ($r = .49$) also demonstrated substantial relationships with failure. These findings were consistent with earlier condition-monitoring studies in which vibration and acoustic signals were particularly informative for rotating components, pumps, bearings, robotic systems, and mechanical assemblies, while temperature increases frequently accompanied friction, electrical loading, cooling deterioration, and component degradation (Picard et al., 2020). The importance of vacuum deterioration and pressure instability was also consistent with previous semiconductor equipment research because plasma etching, deposition, cleaning, and vacuum-dependent processes require tightly controlled chamber environments. Small deviations in these conditions can therefore represent both process instability and emerging equipment deterioration. A particularly important result was that the relationships persisted after controlling for equipment age, operating hours, production intensity, equipment category, and previous failures. Vibration intensity retained an adjusted coefficient of $\beta = .34$, followed by alarm frequency at $\beta = .31$, vacuum deterioration at $\beta = .28$, temperature deviation at $\beta = .25$, and previous failure history at $\beta = .23$. This pattern agreed with earlier studies suggesting that historical failure and maintenance information adds predictive value beyond instantaneous sensor measurements (Zhang et al., 2022). However, the present study extended previous evidence by examining these indicators jointly across heterogeneous semiconductor equipment rather than treating each signal independently. The multivariable model explained 57.8% of variation in equipment-failure risk, indicating that a substantial portion of failure behavior could be quantitatively linked to observable condition information. At the same time, the unexplained variation suggested that failure processes remained influenced by latent component conditions, operating environments, recipe-specific stresses, measurement uncertainty, and abrupt events that were not fully captured by available variables. Consequently, the findings supported the earlier literature favoring multisensor condition monitoring while demonstrating that integrated sensor, alarm, exposure, and

maintenance-history information provided a stronger representation of semiconductor equipment degradation than isolated threshold-based monitoring (Malozyomov et al., 2024).

Figure 12: Machine Learning Predictive Maintenance Framework



The predictive-modeling results demonstrated a clear performance progression from conventional statistical classifiers toward ensemble machine-learning and temporal deep-learning architectures. Logistic regression achieved 89.7% independent-test accuracy but only 72.8% failure recall, producing a false-negative rate of 27.2%. This difference illustrated a limitation repeatedly identified in earlier predictive-maintenance research: high overall accuracy can coexist with inadequate detection of rare failures when normal observations dominate the dataset. Naïve Bayes, KNN, and decision-tree models similarly provided useful baseline discrimination but were less effective at capturing complex nonlinear and temporal relationships (Shcherbatov et al., 2021). Ensemble methods produced considerably stronger results. Random forest achieved 95.1% accuracy, .918 F1-score, .976 ROC-AUC, and .927 PR-AUC, while XGBoost reached 96.4% accuracy, 93.8% recall, .939 F1-score, .986 ROC-AUC, and .951 PR-AUC. These results were consistent with earlier studies reporting that boosting and ensemble-tree methods outperform simpler classifiers when equipment failures emerge from nonlinear interactions among multiple condition variables. Deep-learning architectures provided additional improvement, particularly when sequential sensor behavior was explicitly represented. LSTM achieved 97.1% accuracy and 95.2% recall, while the transformer model reached 97.4% accuracy and .972 PR-AUC. The CNN-LSTM hybrid achieved the strongest independent-test results, including 97.8% accuracy, 96.1% recall, .963 F1-score, .994 ROC-AUC, and .979 PR-AUC. Its 3.9% false-negative rate was operationally important because missed failures can result in unexpected downtime, wafer loss, and emergency maintenance (Feng et al., 2018). These findings corresponded with earlier deep-learning research showing that CNN structures can identify localized patterns in high-dimensional sensor streams while LSTM structures preserve temporal dependencies associated with progressive degradation. Remaining-useful-life results strengthened this interpretation. The CNN-LSTM achieved an MAE of 7.4 hours, RMSE of 11.2 hours, MAPE of 7.1%, and R^2 of .94, outperforming the gradient-boosting baseline. Its median warning lead time of 41.6 hours also exceeded the 28.3 hours observed for XGBoost. Therefore, the results supported previous evidence favoring temporal deep learning but additionally demonstrated that improvements in statistical discrimination were accompanied by more accurate failure timing and longer actionable warning intervals (Zhang et al., 2021).

The operational findings demonstrated that stronger failure prediction was associated with substantial improvements in equipment reliability and maintenance efficiency. Following implementation of AI-supported predictive maintenance, mean unplanned downtime decreased from 14.8 to 9.1 hours per equipment unit per month, representing a 38.5% reduction, while failure frequency declined by 33.3%.

Emergency repairs decreased by 41.0%, MTBF increased from 432.9 to 649.4 hours, and MTTR decreased from 6.8 to 4.7 hours. These findings were consistent with earlier predictive-maintenance studies reporting that condition-based intervention can reduce unexpected failures by identifying deterioration before functional breakdown occurs (Bafandegan Emroozi et al., 2024). The 50.0% improvement in MTBF was particularly notable because it suggested that predictive information supported intervention at a stage where developing faults could be corrected before they generated repeated equipment disruption. Similarly, the 30.9% reduction in MTTR was consistent with previous evidence that advance fault information enables technicians to prepare tools, replacement components, and repair procedures before equipment becomes unavailable. Equipment availability increased from 93.1% to 96.5%, utilization increased from 79.4% to 84.8%, and overall equipment effectiveness increased from 75.9% to 81.6%. Importantly, these improvements remained significant after adjustment for equipment age, operating intensity, production volume, baseline failure frequency, and maintenance history, reducing the likelihood that the observed differences simply reflected easier operating conditions during the predictive-maintenance period (Bhat & Venkitaraman, 2024). The findings also extended earlier research by quantifying maintenance-resource consequences. Maintenance labor decreased by 33.0% relative to reactive maintenance, technician interventions declined by 48.8%, spare-part consumption decreased by 29.1%, and annual maintenance cost declined by 31.8%. Planned interventions represented 82.6% of maintenance activity under the predictive strategy compared with 53.2% under reactive maintenance. This shift was consistent with earlier industrial maintenance literature emphasizing that predictive systems create value by transforming unscheduled failures into schedulable interventions rather than merely predicting faults accurately. The 41.6-hour median warning lead time provided a practical mechanism for this improvement because maintenance activities could be coordinated with production requirements (Li et al., 2024). Consequently, the findings indicated that predictive performance had operational value when predictions were sufficiently early and reliable to alter maintenance decisions.

The production findings demonstrated that equipment health had consequences extending beyond maintenance outcomes to wafer quality, manufacturing flow, and process stability. During degraded operating conditions, mean defect density increased from 0.84 to 1.47 defects/cm², representing a 75.0% increase, while wafer rejection increased from 2.6% to 5.1%. Scrap increased by 92.9%, rework by 84.0%, first-pass yield declined from 95.7% to 91.8%, and final yield declined from 97.3% to 94.2%. These findings were consistent with earlier semiconductor manufacturing studies showing that equipment drift can affect product quality before complete machine failure becomes visible (Duarte & Santiago Scarpin, 2023). Vacuum instability, pressure variation, thermal deviation, and gas-flow inconsistency can directly influence chamber-dependent processing, while vibration and mechanical abnormalities can disrupt wafer positioning, material movement, and process repeatability. The observed correlation between vacuum deterioration and defect density ($r = .54$), together with the negative association between vacuum deterioration and first-pass yield ($r = -.47$), was particularly consistent with earlier evidence emphasizing strict environmental control in etching and deposition processes. Production-flow effects were similarly substantial (Yang et al., 2022). Throughput decreased by 12.6% under degraded conditions, cycle time increased by 25.2%, work-in-process inventory increased by 19.7%, utilization declined by 7.5 percentage points, and OEE decreased by 8.9 percentage points. These findings expanded the interpretation of predictive maintenance from a reliability-centered intervention to a broader manufacturing-performance mechanism. AI predictive-maintenance periods reduced defect density to 0.91 defects/cm², restored first-pass yield to 95.2%, increased throughput to 1,169 wafers per day, and reduced cycle time to 30.6 hours. Subgroup differences also corresponded with the technical functions of individual equipment categories. Plasma etching and deposition experienced comparatively large quality effects, whereas vacuum, pump, and material-handling systems demonstrated stronger throughput, utilization, and cycle-time consequences (Kappatos, 2020). This distinction agreed with earlier research suggesting that degradation consequences depend on where equipment is positioned within the production system. Thus, the findings demonstrated that predictive maintenance contributed not only to avoiding equipment failure but also to maintaining process consistency, wafer quality, manufacturing continuity, and production capacity.

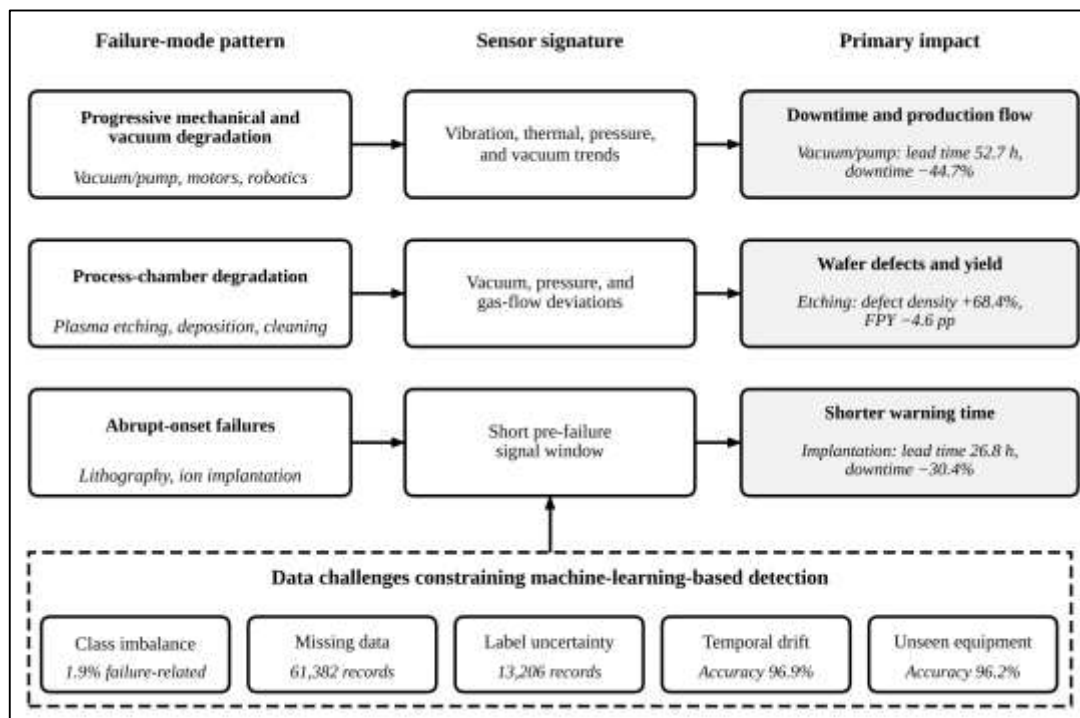
The statistical and validation findings provided evidence that the principal results were not dependent on a single analytical specification or random train-test partition. Effect-size estimates demonstrated practical importance alongside statistical significance. The reduction in unplanned downtime produced Cohen's $d = 0.72$, the MTBF improvement produced $d = 0.81$, and the reduction in MTTR produced $d = 0.67$. Production-related differences were similarly substantial, including $d = 0.82$ for defect density, $d = 0.88$ for wafer rejection, $d = 0.76$ for first-pass yield, and $d = 0.71$ for cycle time. This approach was consistent with earlier methodological studies emphasizing that very large industrial datasets can produce small p -values even for effects with limited practical relevance (Díaz-Reza et al., 2019). The moderate-to-large effects observed here therefore strengthened the operational interpretation of the findings. Predictive validation also showed comparatively limited degradation outside the development data. CNN-LSTM accuracy declined from 98.6% in training to 97.8% on the independent holdout sample, while F1-score decreased from .971 to .963. More demanding temporal validation produced 96.9% accuracy and .951 F1-score, whereas leave-one-equipment-out validation produced 96.2% accuracy and .942 F1-score. This pattern was consistent with earlier research showing that random validation generally produces more optimistic performance estimates than temporal or equipment-independent testing. The decline observed under temporal and cross-equipment validation was therefore expected and provided a more realistic estimate of generalization (Liu et al., 2023). Sensitivity analyses further showed that alternative class weighting, temporal windows, and removal of low-importance variables produced relatively small performance changes. In contrast, removing the five highest-ranked predictors reduced PR-AUC from .979 to .928 and increased RUL RMSE from 11.2 to 16.9 hours, confirming that the principal equipment-condition indicators contributed materially to model performance. Calibration was also strong, with a Brier score of .061 and calibration slope of 0.97. Collectively, these findings aligned with earlier recommendations for independent, temporal, and equipment-level validation and demonstrated that the model retained strong discrimination, calibration, and prediction accuracy under substantially more demanding evaluation conditions (Liu et al., 2023).

The integrated findings established a coherent quantitative pathway linking equipment-condition deterioration, AI-based prediction, maintenance intervention, reliability improvement, and semiconductor production performance. Explainability analysis identified vibration intensity, alarm frequency, vacuum deterioration, temperature deviation, previous failure history, pressure instability, power fluctuation, cumulative operating exposure, gas-flow variation, and electrical-current deviation as the principal contributors to failure prediction. This ranking closely corresponded with the independent correlation and regression results, where vibration, alarms, vacuum deterioration, and temperature deviation also showed the strongest relationships with verified failures (Ghaleb & Taghipour, 2023). Such agreement was important because earlier studies have raised concerns that highly accurate predictive-maintenance models may operate as difficult-to-interpret black boxes whose predictions cannot readily be reconciled with engineering knowledge. The consistency between SHAP-based importance, permutation importance, statistical relationships, and known degradation mechanisms strengthened the interpretability of the predictive framework. More importantly, model performance was quantitatively connected with operational outcomes. Higher failure recall was associated with lower unplanned downtime ($r = -.51$) and fewer emergency repairs ($r = -.46$), while longer warning lead time was associated with higher MTBF ($r = .48$), greater availability ($r = .43$), and lower MTTR ($r = -.39$). Downtime was negatively related to throughput ($r = -.57$) and positively related to cycle time ($r = .52$), while greater equipment-health stability was associated with higher first-pass yield ($r = .49$) and lower defect density ($r = -.54$). These interconnected findings extended earlier research that frequently evaluated model accuracy and maintenance outcomes separately (Franciosi et al., 2020). This study instead demonstrated that predictive quality, warning timeliness, maintenance planning, equipment reliability, and manufacturing performance formed quantitatively related components of the same operational system. Nevertheless, the somewhat lower performance under temporal and leave-one-equipment-out validation indicated that equipment-specific behavior and changing operating conditions remained important sources of variation. Overall, the results showed that AI-driven predictive maintenance was most meaningful when strong predictive discrimination

was accompanied by reliable generalization, interpretable equipment-health indicators, actionable warning lead time, and measurable improvements in maintenance and production outcomes (Lin et al., 2020). This integrated evidence supported the role of AI-based condition monitoring as a data-driven mechanism for improving reliability and manufacturing stability across complex semiconductor production environments.

The failure-mode and data-challenge results clarify why the benefits of machine-learning-based predictive maintenance were uneven across fabrication equipment (Figure 13). Equipment whose failures developed progressively, particularly vacuum and pump systems, generated sensor trajectories that the temporal models could track well before failure, consistent with the longest median warning lead time (52.7 hours) and the largest downtime reduction (44.7%). Lithography and ion implantation, where some failures developed more abruptly, offered shorter warning windows and smaller, although still significant, reliability gains (downtime reductions of 32.7% and 30.4%). Yield impact followed a different pattern: process-chamber tools such as plasma etching, deposition, and cleaning showed the largest degradation-related increases in defect density and losses in first-pass yield, whereas material-handling and vacuum systems affected throughput and cycle time more than wafer quality. This distinction suggests that predictive-maintenance priorities should differ by tool type, with process-chamber tools prioritized for yield protection and flow-critical support equipment prioritized for availability. The data-challenge results show that model performance depended not only on algorithm choice but also on data conditions: class imbalance caused logistic regression to miss 27.2% of failures despite 89.7% accuracy, accuracy declined under chronological evaluation (96.9%) and on unseen equipment (96.2%), and removing the five most informative condition indicators reduced failure recall from 96.1% to 89.7%. These results are consistent with prior work identifying class imbalance, label quality, drift, and tool-to-tool variation as central obstacles to deploying predictive maintenance in semiconductor fabrication (Moyné & Iskandar, 2017; Susto et al., 2015).

Figure 13: Failure-Mode Profiles, Data Challenges, and Yield Impact in Machine-Learning-Based Predictive Maintenance



CONCLUSION

This study concluded that AI-driven predictive maintenance provided a quantitatively effective and operationally meaningful approach for improving equipment reliability, maintenance efficiency, and production stability in semiconductor manufacturing. The analysis was supported by 2,342,816

retained equipment-cycle records, 17,923,648 sensor observations, 1,286,540 wafer-processing runs, 9,742 maintenance events, and 2,184 verified failures collected across 118 semiconductor manufacturing and supporting equipment units, with overall data completeness reaching 98.7%. Equipment-condition analysis established that vibration intensity, alarm frequency, vacuum deterioration, temperature deviation, acoustic abnormality, pressure instability, electrical fluctuations, operating exposure, and maintenance history were significantly associated with verified failures, with vibration producing the strongest relationship ($r = .61$; $\beta = .34$). Comparative predictive modeling demonstrated that ensemble and temporal deep-learning approaches substantially outperformed conventional classifiers, with the CNN-LSTM hybrid achieving 97.8% independent-test accuracy, 96.1% recall, an F1-score of .963, ROC-AUC of .994, and PR-AUC of .979. The same model achieved an RUL MAE of 7.4 hours, RMSE of 11.2 hours, MAPE of 7.1%, R^2 of .94, and a median actionable warning lead time of 41.6 hours. These predictive improvements corresponded with important operational outcomes: AI-supported predictive maintenance reduced unplanned downtime by 38.5%, failure frequency by 33.3%, emergency repairs by 41.0%, and MTTR by 30.9%, while MTBF increased by 50.0%. Equipment availability increased from 93.1% to 96.5%, utilization from 79.4% to 84.8%, and overall equipment effectiveness from 75.9% to 81.6%. Maintenance labor, technician interventions, spare-part consumption, and annual maintenance costs were also reduced, demonstrating that predictive maintenance improved both reliability and resource efficiency. Equipment degradation was additionally associated with a 75.0% increase in wafer defect density, a 3.9-percentage-point reduction in first-pass yield, a 12.6% decrease in throughput, and a 25.2% increase in cycle time, whereas predictive-maintenance periods substantially moderated these adverse effects. Validation through stratified cross-validation, independent holdout testing, temporal validation, leave-one-equipment-out analysis, sensitivity testing, calibration assessment, and explainability analysis confirmed that the principal results remained comparatively stable across analytical conditions. Overall, this study established that the value of AI-driven predictive maintenance extended beyond accurate failure classification by connecting multisensor condition monitoring and timely failure prediction with longer equipment life intervals, reduced maintenance disruption, improved wafer quality, greater throughput, and more stable semiconductor manufacturing performance. With respect to failure modes, equipment with progressive, sensor-detectable degradation, particularly vacuum and pump systems, offered the longest warning lead times and the largest downtime reductions, whereas abrupt-onset failures in lithography and ion implantation offered shorter warnings, and the strongest yield impact occurred in plasma etching and deposition. The principal data challenges (severe class imbalance, missing and inconsistently timestamped records, uncertain failure labels, temporal drift, and reduced transfer to unseen equipment) were measurable and partly mitigated, but they remain central constraints on machine-learning-based predictive maintenance in semiconductor fabrication.

RECOMMENDATIONS

Based on the quantitative findings of this study, semiconductor manufacturers should adopt AI-driven predictive maintenance as an integrated component of equipment reliability management, process control, and production planning rather than treating predictive modeling as an isolated analytical function. Particular emphasis should be placed on continuous monitoring of vibration intensity, alarm frequency, vacuum deterioration, temperature deviation, acoustic abnormalities, pressure instability, electrical behavior, gas-flow variation, cumulative operating exposure, and maintenance history because these indicators demonstrated meaningful relationships with verified equipment failures. Manufacturing facilities should integrate sensor, fault-detection and classification, manufacturing execution, alarm, maintenance, and production-quality databases within a synchronized data infrastructure to support reliable real-time condition assessment. Given the superior predictive performance observed for temporal deep-learning approaches, CNN-LSTM and comparable hybrid architectures should be prioritized for equipment characterized by complex sequential degradation, while XGBoost, LightGBM, and random forest models can provide effective alternatives where computational resources, data volume, or model complexity are constrained. Model selection should not rely exclusively on accuracy; failure recall, F1-score, PR-AUC, false-negative rate, calibration, RUL prediction error, and actionable warning lead time should receive substantial consideration because missed failures have direct implications for downtime and wafer loss. The observed median warning

lead time of 41.6 hours should be incorporated into maintenance scheduling procedures to support technician allocation, spare-part preparation, production rescheduling, and planned equipment intervention before functional failure. Predictive alerts should also be integrated with maintenance decision rules so that unnecessary interventions are minimized while high-risk degradation receives timely attention. Because predictive maintenance reduced unplanned downtime by 38.5%, failure frequency by 33.3%, emergency repairs by 41.0%, and MTTR by 30.9%, semiconductor facilities should progressively shift maintenance resources from reactive activities toward condition-informed planned interventions. Equipment-specific models should be considered because degradation effects differed across etching, deposition, lithography, implantation, cleaning, vacuum, robotic, and material-handling systems. Continuous temporal and cross-equipment validation should accompany deployment to detect performance deterioration, concept drift, and changing process conditions. SHAP and other explainability techniques should additionally be incorporated into maintenance dashboards to provide engineers with interpretable evidence supporting AI-generated alerts. Finally, predictive-maintenance performance should be evaluated jointly with wafer defect density, yield, throughput, cycle time, utilization, OEE, maintenance cost, and production interruptions so that implementation decisions reflect manufacturing-wide benefits rather than prediction accuracy alone.

LIMITATIONS

This study had several limitations that should be considered when interpreting the quantitative findings on AI-driven predictive maintenance in semiconductor manufacturing. Although the final dataset was extensive, containing 2,342,816 equipment-cycle records, 17,923,648 sensor observations, 1,286,540 wafer-processing runs, 9,742 maintenance events, and 2,184 verified failures across 118 equipment units, the analyzed equipment did not represent every tool architecture, manufacturer, fabrication technology, process recipe, or operating environment used throughout the global semiconductor industry. Consequently, differences in equipment design, sensor configuration, maintenance practices, production intensity, environmental conditions, and fabrication processes could affect the transferability of the estimated relationships. The dataset also remained highly imbalanced, with normal observations accounting for 89.4%, degraded observations for 8.7%, and failure-related observations for only 1.9%, meaning that rare or unusual failure mechanisms were represented by substantially fewer examples. Although class weighting, resampling, PR-AUC, recall, and false-negative measures were incorporated to address this issue, residual imbalance could still have influenced model learning. Failure labels depended on maintenance logs, alarm histories, equipment-status records, and documented component replacement; therefore, delayed reporting, incomplete maintenance documentation, or uncertain failure onset could have introduced labeling error. Furthermore, sensor measurements could have been affected by calibration drift, measurement noise, synchronization differences, temporary communication interruptions, and unobserved environmental influences despite extensive quality screening. The quasi-experimental pre-post design also limited strict causal interpretation because improvements observed after predictive-maintenance implementation could have been partially influenced by simultaneous changes in production schedules, technician experience, equipment upgrades, process recipes, maintenance policies, or manufacturing demand. Although regression adjustment controlled for equipment age, operating exposure, production intensity, equipment category, previous failures, and maintenance history, unmeasured confounding could not be eliminated completely. Predictive performance also decreased modestly from 97.8% accuracy in independent holdout testing to 96.9% under temporal validation and 96.2% under leave-one-equipment-out validation, demonstrating that changing operating conditions and previously unseen equipment introduced additional uncertainty. The CNN-LSTM model also required greater computational resources and offered lower inherent interpretability than simpler statistical models, even though SHAP and permutation analyses improved explanatory transparency. Finally, the reported reductions in downtime, maintenance cost, defects, and production disruption reflected the investigated manufacturing context and should not automatically be assumed to produce identical economic benefits in facilities with different labor costs, equipment configurations, production volumes, maintenance infrastructures, or data maturity. These limitations indicated that the findings should be interpreted as strong context-specific quantitative evidence rather than universally fixed performance expectations.

REFERENCES

- [1]. A. Ramezan, C., A. Warner, T., & E. Maxwell, A. (2019). Evaluation of sampling and cross-validation tuning strategies for regional-scale machine learning classification. *Remote Sensing*, 11(2), 185.
- [2]. Abidi, M. H., Mohammed, M. K., & Alkhalefah, H. (2022). Predictive maintenance planning for industry 4.0 using machine learning for sustainable manufacturing. *Sustainability*, 14(6), 3387.
- [3]. Abuelnaga, A., Narimani, M., & Bahman, A. S. (2021). Power electronic converter reliability and prognosis review focusing on power switch module failures. *Journal of Power Electronics*, 21(6), 865-880.
- [4]. Akyaz, T., & Engin, D. (2024). Machine learning-based predictive maintenance system for artificial yarn machines. *Ieee Access*, 12, 125446-125461.
- [5]. Al-Duais, F. S., Mohamed, A.-B., Jawa, T. M., & Sayed-Ahmed, N. (2022). Optimal periods of conducting preventive maintenance to reduce expected downtime and its impact on improving reliability. *Computational Intelligence and Neuroscience*, 2022(1), 7105526.
- [6]. Alakus, T. B., & Turkoglu, I. (2020). Comparison of deep learning approaches to predict COVID-19 infection. *Chaos, Solitons & Fractals*, 140, 110120.
- [7]. Aldoseri, A., Al-Khalifa, K. N., & Hamouda, A. M. (2023). Re-thinking data strategy and integration for artificial intelligence: concepts, opportunities, and challenges. *Applied Sciences*, 13(12), 7082.
- [8]. Allgaier, J., & Pryss, R. (2024). Cross-validation visualized: a narrative guide to advanced methods. *Machine Learning and Knowledge Extraction*, 6(2), 1378-1388.
- [9]. Alothman, Y., Bader-El-Den, M., & Maurya, L. (2026). AI-enabled preventive action for semiconductor lithography using LSTM, transformers, and SHAP. *Semiconductor and Heterogeneous Integration*, 1(2), 7.
- [10]. Alsina, E. F., Chica, M., Trawiński, K., & Regattieri, A. (2018). On the use of machine learning methods to predict component reliability from data-driven industrial case studies. *The International Journal of Advanced Manufacturing Technology*, 94(5), 2419-2433.
- [11]. Andrianandrianina Johanesa, T. V., Equeter, L., & Mahmoudi, S. A. (2024). Survey on AI applications for product quality control and predictive maintenance in industry 4.0. *Electronics*, 13(5), 976.
- [12]. Ansari, F., & Kohl, L. (2022). AI-enhanced maintenance for building resilience and viability in supply chains. In *Supply Network Dynamics and Control* (pp. 163-185). Springer.
- [13]. Arunthavanathan, R., Khan, F., Ahmed, S., & Imtiaz, S. (2021). A deep learning model for process fault prognosis. *Process Safety and Environmental Protection*, 154, 467-479.
- [14]. Aydemir, G., & Acar, B. (2020). Anomaly monitoring improves remaining useful life estimation of industrial machinery. *Journal of Manufacturing Systems*, 56, 463-469.
- [15]. Bacanin, N., Zivkovic, M., Antonijevic, M., Venkatachalam, K., Lee, J., Nam, Y., Marjanovic, M., Strumberger, I., & Abouhawwash, M. (2023). Addressing feature selection and extreme learning machine tuning by diversity-oriented social network search: an application for phishing websites detection. *Complex & Intelligent Systems*, 9(6), 7269-7304.
- [16]. Bafandegan Emrooz, V., Kazemi, M., Doostparast, M., & Pooya, A. (2024). Improving industrial maintenance efficiency: A holistic approach to integrated production and maintenance planning with human error optimization. *Process Integration and Optimization for Sustainability*, 8(2), 539-564.
- [17]. Baharani, M., Biglarbegian, M., Parkhideh, B., & Tabkhi, H. (2019). Real-time deep learning at the edge for scalable reliability modeling of Si-MOSFET power electronics converters. *IEEE Internet of Things Journal*, 6(5), 7375-7385.
- [18]. Bergès, C., Céneray, C., Albermann, G., Bird, J., Balakrishnan, B., Mustafa, S. M., Moeller, S., Rubink, B., & Smith, C. (2020). Detection and prevention of assembly defects, by machine learning algorithms, in semiconductor industry for automotive. 2020 IEEE International Symposium on the Physical and Failure Analysis of Integrated Circuits (IPFA).
- [19]. Bhat, S. M., & Venkitaraman, A. (2024). Strategic integration of predictive maintenance plans to improve operational efficiency of smart grids. 2024 IEEE International Conference on Information Technology, Electronics and Intelligent Communication Systems (ICITEICS).
- [20]. Biebl, F., Glawar, R., Jalali, A., Ansari, F., Haslhofer, B., de Boer, P., & Sihn, W. (2020). A conceptual model to enable prescriptive maintenance for etching equipment in semiconductor manufacturing. *Procedia CIRP*, 88, 64-69.
- [21]. Butler, S. W., Ringwood, J. V., & MacGearailt, N. (2009). Prediction of vacuum pump degradation in semiconductor processing. *IFAC Proceedings Volumes*, 42(8), 1635-1640.
- [22]. Cabanes, B., Hubac, S., Le Masson, P., & Weil, B. (2021). Improving reliability engineering in product development based on design theory: the case of FMEA in the semiconductor industry. *Research in Engineering Design*, 32(3), 309-329.
- [23]. Carvalho, T. P., Soares, F. A., Vita, R., Francisco, R. d. P., Basto, J. P., & Alcalá, S. G. (2019). A systematic literature review of machine learning methods applied to predictive maintenance. *Computers & industrial engineering*, 137, 106024.
- [24]. Cerqueira, V., Torgo, L., & Mozetič, I. (2020). Evaluating time series forecasting models: An empirical study on performance estimation methods. *Machine Learning*, 109(11), 1997-2028.
- [25]. 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*, 5(12), 5969-5989.
- [26]. Chien, C.-F., Van Nguyen, T. H., Li, Y.-C., & Chen, Y.-J. (2023). Bayesian decision analysis for optimizing in-line metrology and defect inspection strategy for sustainable semiconductor manufacturing and an empirical study. *Computers & industrial engineering*, 182, 109421.

- [27]. Chung, E., Park, K., & Kang, P. (2023). Fault classification and timing prediction based on shipment inspection data and maintenance reports for semiconductor manufacturing equipment. *Computers & industrial engineering*, 176, 108972.
- [28]. Çınar, Z. M., Abdussalam Nuhu, A., Zeeshan, Q., Korhan, O., Asmael, M., & Safaei, B. (2020). Machine learning in predictive maintenance towards sustainable smart manufacturing in industry 4.0. *Sustainability*, 12(19), 8211.
- [29]. Dang, D., Liu, Y., & Lee, S.-K. (2024). State evaluation of electrical equipment in substations based on data mining. *Applied Sciences*, 14(16), 7348.
- [30]. Díaz-Reza, J. R., García-Alcaraz, J. L., & Martínez-Loya, V. (2019). Impact Analysis of Total Productive Maintenance. *Impact Analysis of Total Productive Maintenance*, 5(1), 71-74.
- [31]. Dong, J., & Ye, C. (2023). Joint optimisation of uncertain distributed manufacturing and preventive maintenance for semiconductor wafers considering multi-energy complementary. *International Journal of Production Research*, 61(9), 3030-3051.
- [32]. Duarte, A. L. C. M., & Santiago Scarpin, M. R. S. (2023). Maintenance practices and overall equipment effectiveness: testing the moderating effect of training. *Journal of Quality in Maintenance Engineering*, 29(2), 442-459.
- [33]. Falck, J., Felgemacher, C., Rojko, A., Liserre, M., & Zacharias, P. (2018). Reliability of power electronic systems: An industry perspective. *IEEE Industrial Electronics Magazine*, 12(2), 24-35.
- [34]. Fan, S.-K. S., Hsu, C.-Y., Tsai, D.-M., Chou, M. C., Jen, C.-H., & Tsou, J.-H. (2022). Key feature identification for monitoring wafer-to-wafer variation in semiconductor manufacturing. *IEEE Transactions on Automation Science and Engineering*, 19(3), 1530-1541.
- [35]. Fassi, Y., Heiries, V., Boutet, J., & Boisseau, S. (2023). Toward physics-informed machine-learning-based predictive maintenance for power converters – A review. *IEEE Transactions on Power Electronics*, 39(2), 2692-2720.
- [36]. Feng, D., Lin, S., He, Z., Sun, X., & Lee, W.-J. (2018). Optimization method with prediction-based maintenance strategy for traction power supply equipment based on risk quantification. *IEEE transactions on transportation electrification*, 4(4), 961-970.
- [37]. Fischer, D., Moder, P., & Ehm, H. (2021). Investigation of predictive maintenance for semiconductor manufacturing and its impacts on the supply chain. 2021 22nd IEEE International Conference on Industrial Technology (ICIT),
- [38]. Franciosi, C., Voisin, A., Miranda, S., Riemma, S., & Iung, B. (2020). Measuring maintenance impacts on sustainability of manufacturing industries: from a systematic literature review to a framework proposal. *Journal of Cleaner Production*, 260, 121065.
- [39]. Gawde, S., Patil, S., Kumar, S., & Kotecha, K. (2023). A scoping review on multi-fault diagnosis of industrial rotating machines using multi-sensor data fusion. *Artificial Intelligence Review*, 56(5), 4711-4764.
- [40]. Ghaleb, M., & Taghipour, S. (2023). Evidence-based study of the impacts of maintenance practices on asset sustainability. *International Journal of Production Research*, 61(24), 8719-8750.
- [41]. Gong, Y., Liu, G., Xue, Y., Li, R., & Meng, L. (2023). A survey on dataset quality in machine learning. *Information and Software Technology*, 162, 107268.
- [42]. Gong, Z., Cao, Z., Zhou, S., Yang, F., Shuai, C., Ouyang, X., & Luo, Z. (2024). Thermal fault detection of high-voltage isolating switches based on hybrid data and BERT. *Arabian Journal for Science and Engineering*, 49(5), 6429-6443.
- [43]. Hanif, A., Yu, Y., DeVoto, D., & Khan, F. (2018). A comprehensive review toward the state-of-the-art in failure and lifetime predictions of power electronic devices. *IEEE Transactions on Power Electronics*, 34(5), 4729-4746.
- [44]. Hosseinabadi, F., Chakraborty, S., Bhoi, S. K., Prochart, G., Hrvanovic, D., & Hegazy, O. (2024). A comprehensive overview of reliability assessment strategies and testing of power electronics converters. *IEEE Open Journal of Power Electronics*, 5, 473-512.
- [45]. 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.
- [46]. Hung, Y.-H. (2021). Improved ensemble-learning algorithm for predictive maintenance in the manufacturing process. *Applied Sciences*, 11(15), 6832.
- [47]. Jalali, A., Heistracher, C., Schindler, A., Haslhofer, B., Nemeth, T., Glawar, R., Sihn, W., & De Boer, P. (2019). Predicting time-to-failure of plasma etching equipment using machine learning. 2019 IEEE international conference on prognostics and health management (ICPHM),
- [48]. Jena, M. C., Mishra, S. K., & Moharana, H. S. (2024). Integration of Industry 4.0 with reliability centered maintenance to enhance sustainable manufacturing. *Environmental Progress & Sustainable Energy*, 43(2), e14321.
- [49]. Jiang, Y., Dai, P., Fang, P., Zhong, R. Y., Zhao, X., & Cao, X. (2022). A2-LSTM for predictive maintenance of industrial equipment based on machine learning. *Computers & industrial engineering*, 172, 108560.
- [50]. Jovicic, E., Primorac, D., Cupic, M., & Jovic, A. (2023). Publicly available datasets for predictive maintenance in the energy sector: A review. *Ieee Access*, 11, 73505-73520.
- [51]. Justus, V., & Kanagachidambaresan, G. (2024). Machine learning based fault-oriented predictive maintenance in industry 4.0. *International Journal of System Assurance Engineering and Management*, 15(1), 462-474.
- [52]. Kaliappan, J., Bagepalli, A. R., Almal, S., Mishra, R., Hu, Y.-C., & Srinivasan, K. (2023). Impact of cross-validation on machine learning models for early detection of intrauterine fetal demise. *Diagnostics*, 13(10), 1692.
- [53]. Kamat, P., Kumar, S., Patil, S., & Kotecha, K. (2024). Anomaly-informed remaining useful life estimation (AIRULE) of bearing machinery using deep learning framework. *MethodsX*, 12, 102555.
- [54]. Kao, S.-X., & Chien, C.-F. (2023). Deep learning-based positioning error fault diagnosis of wire bonding equipment and an empirical study for IC packaging. *IEEE Transactions on Semiconductor Manufacturing*, 36(4), 619-628.

- [55]. Kapoor, S., & Narayanan, A. (2023). Leakage and the reproducibility crisis in machine-learning-based science. *Patterns*, 4(9), 100804.
- [56]. Kappatos, V. (2020). Maintenance: A key factor for energy efficiency. In *Aspects of the Energy Union: Application and Effects of European Energy Policies in SE Europe and Eastern Mediterranean* (pp. 321-336). Springer.
- [57]. Kattenborn, T., Schiefer, F., Frey, J., Feilhauer, H., Mahecha, M. D., & Dormann, C. F. (2022). Spatially autocorrelated training and validation samples inflate performance assessment of convolutional neural networks. *ISPRS Open Journal of Photogrammetry and Remote Sensing*, 5, 100018.
- [58]. Ki, W., Choi, M., Lee, Y., Park, J., Park, J. J., Park, J., & Kim, Y. (2024). Predictive analysis and root cause identification of equipment failures using event log patterns: a Focus on EUV/DUV equipment. 2024 35th Annual SEMI Advanced Semiconductor Manufacturing Conference (ASMC),
- [59]. Kim, H., Hong, S., Shin, C., Jin, Y., Lim, D. H., Kim, J. Y., Hwang, H., & Kim, T. (2017). Investigation of the pad-conditioning performance deterioration in the chemical mechanical polishing process. *Wear*, 392-393, 93-98.
- [60]. Kim, H., Lim, D.-E., & Lee, S. (2020). Deep learning-based dynamic scheduling for semiconductor manufacturing with high uncertainty of automated material handling system capability. *IEEE Transactions on Semiconductor Manufacturing*, 33(1), 13-22.
- [61]. Kim, S. H., Kim, C. Y., Seol, D. H., Choi, J. E., & Hong, S. J. (2022). Machine learning-based process-level fault detection and part-level fault classification in semiconductor etch equipment. *IEEE Transactions on Semiconductor Manufacturing*, 35(2), 174-185.
- [62]. Kirschbaum, L., Roman, D., Singh, G., Bruns, J., Robu, V., & Flynn, D. (2020). AI-driven maintenance support for downhole tools and electronics operated in dynamic drilling environments. *Ieee Access*, 8, 78683-78701.
- [63]. Klimecky, P. I., Grizzle, J. W., & Terry, F. L., Jr. (2003). Compensation for transient chamber wall condition using real-time plasma density feedback control in an inductively coupled plasma etcher. *Journal of Vacuum Science & Technology A*, 21(3), 706-717.
- [64]. Kovacevic, J. (2021). Design for Availability. *Design for Maintainability*, 281-301.
- [65]. 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.
- [66]. Lee, I., Park, H. J., Jang, J.-W., Kim, C.-W., & Choi, J.-H. (2023). System-level fault diagnosis for an industrial wafer transfer robot with multi-component failure modes. *Applied Sciences*, 13(18), 10243.
- [67]. Li, J. H., Jia, S., Ren, L., & Li, X. (2024). Research on wind turbines preventive maintenance strategies based on reliability and cost-effectiveness ratio. *Industrial Lubrication and Tribology*, 76(10), 1168-1176.
- [68]. Li, Y., Peng, S., Li, Y., & Jiang, W. (2020). A review of condition-based maintenance: Its prognostic and operational aspects. *Frontiers of Engineering Management*, 7(3), 323-334.
- [69]. Lin, S., Li, N., Feng, D., Guo, X., Pan, W., Wang, J., & Yang, C. (2020). A preventive opportunistic maintenance method for railway traction power supply system based on equipment reliability: S. Lin et al. *Railway Engineering Science*, 28(2), 199-211.
- [70]. Liu, F., Sun, H., & Peng, R. (2022). A time-series probabilistic preventive maintenance strategy based on multi-class equipment condition indicators. *Journal of the Operational Research Society*, 73(12), 2756-2774.
- [71]. Liu, L., Guo, Q., Liu, D., & Peng, Y. (2019). Data-driven remaining useful life prediction considering sensor anomaly detection and data recovery. *Ieee Access*, 7, 58336-58345.
- [72]. Liu, Y., Wang, W., Wang, W., Yu, C., Mao, B., Shang, D., & Duan, Y. (2023). Purpose-driven evaluation of operation and maintenance efficiency and safety based on DIKWP. *Sustainability*, 15(17), 13083.
- [73]. Liu, Z., & Hui, J. (2024). Advancing predictive maintenance: a deep learning approach to sensor and event-log data fusion. *Sensor Review*, 44(5), 563-574.
- [74]. Maguluri, L. P., Suganthi, D., Dhote, G. M., Kapila, D., Jadhav, M. M., & Neelima, S. (2024). AI-enhanced predictive maintenance in hybrid roll-to-roll manufacturing integrating multi-sensor data and self-supervised learning. *The International Journal of Advanced Manufacturing Technology*, 1-10.
- [75]. Malozyomov, B. V., Martyushev, N. V., Babyr, N. V., Pogrebnoy, A. V., Efremenkov, E. A., Valuev, D. V., & Boltrushevich, A. E. (2024). Modelling of reliability indicators of a mining plant. *Mathematics*, 12(18), 2842.
- [76]. Matania, O., Bechhoefer, E., Blunt, D., Wang, W., & Bortman, J. (2024). Anomaly detection and remaining useful life estimation for the Health and Usage Monitoring Systems 2023 Data Challenge. *Sensors*, 24(13), 4258.
- [77]. Merkt, O. (2019). On the use of predictive models for improving the quality of industrial maintenance: An analytical literature review of maintenance strategies. 2019 federated conference on computer science and information systems (fedcsis),
- [78]. Meyer, H., Reudenbach, C., Hengl, T., Katurji, M., & Nauss, T. (2018). Improving performance of spatio-temporal machine learning models using forward feature selection and target-oriented validation. *Environmental Modelling & Software*, 101, 1-9.
- [79]. Mohamed Almazrouei, S., Dweiri, F., Aydin, R., & Alnaqbi, A. (2023). A review on the advancements and challenges of artificial intelligence based models for predictive maintenance of water injection pumps in the oil and gas industry. *SN Applied Sciences*, 5(12), 391.
- [80]. Mohan, R., Roselyn, J. P., & Uthra, R. A. (2023). LSTM based artificial intelligence predictive maintenance technique for availability rate and OEE improvement in a TPM implementing plant through Industry 4.0 transformation. *Journal of Quality in Maintenance Engineering*, 29(4), 763-798.

- [81]. Mohan, T. R., Roselyn, J. P., Uthra, R. A., Devaraj, D., & Umachandran, K. (2021). Intelligent machine learning based total productive maintenance approach for achieving zero downtime in industrial machinery. *Computers & industrial engineering*, 157, 107267.
- [82]. Moyne, J., & Iskandar, J. (2017). Big data analytics for smart manufacturing: Case studies in semiconductor manufacturing. *Processes*, 5(3), 39.
- [83]. Nacchia, M., Fruggiero, F., Lambiase, A., & Bruton, K. (2021). A systematic mapping of the advancing use of machine learning techniques for predictive maintenance in the manufacturing sector. *Applied Sciences*, 11(6), 2546.
- [84]. Niazi, S. G., Huang, T., Zhou, H., Bai, S., & Huang, H.-Z. (2024). Multi-scale time series analysis using TT-ConvLSTM technique for bearing remaining useful life prediction. *Mechanical Systems and Signal Processing*, 206, 110888.
- [85]. Nuhu, A. A., Zeeshan, Q., Safaei, B., & Shahzad, M. A. (2023). Machine learning-based techniques for fault diagnosis in the semiconductor manufacturing process: a comparative study. *The Journal of Supercomputing*, 79(2), 2031-2081.
- [86]. Orošnjak, M., Beker, I., Brkljač, N., & Vrhovac, V. (2024). Predictors of Successful Maintenance Practices in Companies Using Fluid Power Systems: A Model-Agnostic Interpretation. *Applied Sciences*, 14(13), 5921.
- [87]. Palmieri, F., Akhtar, N. F., Pané, A., Jiménez, A., Olbeyra, R. P., Viaplana, J., Vidal, J., de Hollanda, A., Gama-Perez, P., & Jiménez-Chillarón, J. C. (2024). Machine learning allows robust classification of visceral fat in women with obesity using common laboratory metrics. *Scientific reports*, 14(1), 17263.
- [88]. Patel, R., Shah, S., Lella, S., & Gajbahar, A. (2024). Real-Time Monitoring and Predictive Maintenance. *Artificial Intelligence-Enabled Digital Twin for Smart Manufacturing*, 305-323.
- [89]. Payette, M., & Abdul-Nour, G. (2023). Machine learning applications for reliability engineering: A review. *Sustainability*, 15(7), 6270.
- [90]. Pech, M., Vrchota, J., & Bednář, J. (2021). Predictive maintenance and intelligent sensors in smart factory. *Sensors*, 21(4), 1470.
- [91]. Pekşen, M. F., Yurtsever, U., & Uyaroğlu, Y. (2024). Enhancing electrical panel anomaly detection for predictive maintenance with machine learning and IoT. *Alexandria Engineering Journal*, 96, 112-123.
- [92]. Picard, S., Chapdelaine, C., Cappi, C., Gardes, L., Jenn, E., Lefèvre, B., & Soumarmon, T. (2020). Ensuring dataset quality for machine learning certification. 2020 IEEE international symposium on software reliability engineering workshops (ISSREW),
- [93]. Plathottam, S. J., Rzonca, A., Lakhnori, R., & Illoeje, C. O. (2023). A review of artificial intelligence applications in manufacturing operations. *Journal of Advanced Manufacturing and Processing*, 5(3), e10159.
- [94]. Qiu, W., Tang, Q., Yao, W., Qin, Y., & Ma, J. (2020). Probability analysis for failure assessment of electric energy metering equipment under multiple extreme stresses. *IEEE Transactions on Industrial Informatics*, 17(6), 3762-3771.
- [95]. Rostami, H., Blue, J., Chen, A., & Yugma, C. (2021). Equipment deterioration modeling and cause diagnosis in semiconductor manufacturing. *International Journal of Intelligent Systems*, 36(6), 2618-2638.
- [96]. San Martin, G., & Droguett, E. L. (2022). Temporal variational auto-encoders for semi-supervised remaining useful life and fault diagnosis. *Ieee Access*, 10, 55112-55125.
- [97]. Seraj, A., Mohammadi-Khanaposhtani, M., Daneshfar, R., Naseri, M., Esmaeili, M., Baghban, A., Habibzadeh, S., & Eslamian, S. (2023). Cross-validation. In *Handbook of hydroinformatics* (pp. 89-105). Elsevier.
- [98]. Shcherbatov, I., Lisin, E., Rogalev, A., Tsurikov, G., Dvořák, M., & Strielkowski, W. (2021). Power equipment defects prediction based on the joint solution of classification and regression problems using machine learning methods. *Electronics*, 10(24), 3145.
- [99]. Shimizu, M., Perinpanayagam, S., & Namoano, B. (2022). A real-time fault detection framework based on unsupervised deep learning for prognostics and health management of railway assets. *Ieee Access*, 10, 96442-96458.
- [100]. Shin, J.-W., Yoon, K.-H., Chai, H.-S., & Kim, J.-C. (2022). Reliability-centered maintenance scheduling of photovoltaic components according to failure effects. *Energies*, 15(7), 2529.
- [101]. Singh, V., Pencina, M., Einstein, A. J., Liang, J. X., Berman, D. S., & Slomka, P. (2021). Impact of train/test sample regimen on performance estimate stability of machine learning in cardiovascular imaging. *Scientific reports*, 11(1), 14490.
- [102]. Sohn, J., Kamei, Y., McIntosh, S., & Yoo, S. (2021). Leveraging fault localisation to enhance defect prediction. 2021 IEEE International Conference on Software Analysis, Evolution and Reengineering (SANER),
- [103]. Song, Y., Lv, Z., Bai, J., Niu, S., Wu, Z., Qin, L., Chen, Y., Liang, L., Lei, Y., & Jia, P. (2022). Processes of the reliability and degradation mechanism of high-power semiconductor lasers. *Crystals*, 12(6), 765.
- [104]. Sonnekalb, T., Heinze, T. S., & Mäder, P. (2022). Deep security analysis of program code: A systematic literature review. *Empirical Software Engineering*, 27(1), 2.
- [105]. Sony, S., Gamage, S., Sadhu, A., & Samarabandu, J. (2022). Vibration-based multiclass damage detection and localization using long short-term memory networks. *Structures*,
- [106]. Sotito-Mayor, B., & Kalech, M. (2021). Cross-project smell-based defect prediction. *Soft Computing*, 25(22), 14171-14181.
- [107]. Sreedharan, R., Prajapati, J., Engineer, P., & Prajapati, D. (2023). Leave-one-out cross-validation in machine learning. In *Ethical issues in AI for bioinformatics and chemoinformatics* (pp. 56-71). CRC Press.
- [108]. Susto, G. A., Schirru, A., Pampuri, S., McLoone, S., & Beghi, A. (2015). Machine learning for predictive maintenance: A multiple classifier approach. *IEEE Transactions on Industrial Informatics*, 11(3), 812-820.
- [109]. Suwanasri, C., Yongyee, I., & Suwanasri, T. (2024). Age estimation of transmission line using statistical health index and failure probability curve-fitting method. *Energies*, 17(3), 637.
- [110]. Tang, Y., Yang, X., & Wang, G. (2020). A framework for health state evaluation of the complex mechanical system with its occurrence probability of failure mode. *Ieee Access*, 8, 73570-73587.

- [111]. Teh, H. Y., Kempa-Liehr, A. W., & Wang, K. I.-K. (2020). Sensor data quality: A systematic review. *Journal of Big Data*, 7(1), 11.
- [112]. Tercan, H., & Meisen, T. (2022). Machine learning and deep learning based predictive quality in manufacturing: a systematic review. *Journal of Intelligent Manufacturing*, 33(7), 1879-1905.
- [113]. Thavarajah, M. (2024). Real time inferencing of semiconductor wafer probing process using Machine Learning. 2024 IEEE 40th International Electronics Manufacturing Technology (IEMT),
- [114]. Tsamardinos, I., Greasidou, E., & Borboudakis, G. (2018). Bootstrapping the out-of-sample predictions for efficient and accurate cross-validation. *Machine Learning*, 107(12), 1895-1922.
- [115]. Tsanousa, A., Bektsis, E., Kyriakopoulos, C., González, A. G., Leturiondo, U., Gialampoukidis, I., Karakostas, A., Vrochidis, S., & Kompatsiaris, I. (2022). A review of multisensor data fusion solutions in smart manufacturing: Systems and trends. *Sensors*, 22(5), 1734.
- [116]. Ucar, A., Karakose, M., & Kırımca, N. (2024). Artificial intelligence for predictive maintenance applications: key components, trustworthiness, and future trends. *Applied Sciences*, 14(2), 898.
- [117]. Verstraete, D., Droguett, E., & Modarres, M. (2019). A deep adversarial approach based on multi-sensor fusion for semi-supervised remaining useful life prognostics. *Sensors*, 20(1), 176.
- [118]. Wan, B., Wang, Y., Su, Y., & Fu, G. (2020). Reliability evaluation of multi-mechanism failure for semiconductor devices using physics-of-failure technique and maximum entropy principle. *Ieee Access*, 8, 188154-188170.
- [119]. Wang, J., Zhang, J., & Wang, X. (2018). A data driven cycle time prediction with feature selection in a semiconductor wafer fabrication system. *IEEE Transactions on Semiconductor Manufacturing*, 31(1), 173-182.
- [120]. Wang, L.-C., Chu, P.-C., & Lin, S.-Y. (2019). Impact of capacity fluctuation on throughput performance for semiconductor wafer fabrication. *Robotics and Computer-Integrated Manufacturing*, 55, 208-216.
- [121]. Wu, C.-T., Li, S.-H., & Yen, D. C. (2024). Predicting work-in-process in semiconductor packaging using neural networks: Technical evaluation and future applications. *Electronics*, 13(21), 4275.
- [122]. Wu, F., Wu, Q., Tan, Y., & Xu, X. (2024). Remaining useful life prediction based on deep learning: A survey. *Sensors*, 24(11), 3454.
- [123]. Yan, P., Abdulkadir, A., Luley, P.-P., Rosenthal, M., Schatte, G. A., Grewe, B. F., & Stadelmann, T. (2024). A comprehensive survey of deep transfer learning for anomaly detection in industrial time series: Methods, applications, and directions. *Ieee Access*, 12, 3768-3789.
- [124]. Yang, C., Liu, J., Zeng, Y., & Xie, G. (2019). Real-time condition monitoring and fault detection of components based on machine-learning reconstruction model. *Renewable Energy*, 133, 433-441.
- [125]. Yang, L., Liu, Q., Xia, T., Ye, C., & Li, J. (2022). Preventive maintenance strategy optimization in manufacturing system considering energy efficiency and quality cost. *Energies*, 15(21), 8237.
- [126]. Yeo, W., Chang, Y.-C., Chen, L.-C., & Chang, K.-H. (2024). A Novel Out-of-Control Action Plan (OCAP) for Optimizing Efficiency and Quality in the Wafer Probing Process for Semiconductor Manufacturing. *Sensors*, 24(16), 5116.
- [127]. Yongcharoenchaiyasit, K., Arwatchananukul, S., Temdee, P., & Prasad, R. (2023). Gradient boosting based model for elderly heart failure, aortic stenosis, and dementia classification. *Ieee Access*, 11, 48677-48696.
- [128]. Yu, W., Dillon, T., Mostafa, F., Rahayu, W., & Liu, Y. (2019). A global manufacturing big data ecosystem for fault detection in predictive maintenance. *IEEE Transactions on Industrial Informatics*, 16(1), 183-192.
- [129]. Zhang, A., Wang, H., Li, S., Cui, Y., Liu, Z., Yang, G., & Hu, J. (2018). Transfer learning with deep recurrent neural networks for remaining useful life estimation. *Applied Sciences*, 8(12), 2416.
- [130]. Zhang, G., Wang, G., & Zhang, H. (2021). Equipment condition assessment model based on constructing health index. 2021 CAA Symposium on Fault Detection, Supervision, and Safety for Technical Processes (SAFEPROCESS),
- [131]. Zhang, P., Gao, Z., Cao, L., Dong, F., Zou, Y., Wang, K., Zhang, Y., & Sun, P. (2022). Marine systems and equipment prognostics and health management: a systematic review from health condition monitoring to maintenance strategy. *Machines*, 10(2), 72.
- [132]. Zhang, Y., Hutchinson, P., Lieven, N. A., & Nunez-Yanez, J. (2020). Remaining useful life estimation using long short-term memory neural networks and deep fusion. *Ieee Access*, 8, 19033-19045.
- [133]. Zhang, Z., Zhang, W., Zhao, Y., Ding, Y., & Cai, J. (2024). Remaining life prediction of rolling bearings based on unsupervised anomaly detection and STA-BiLSTM. *IEEE Sensors Journal*, 24(24), 41659-41674.
- [134]. Zhu, L., Chen, J., & Chen, C.-I. (2024). Prognostics for semiconductor sustainability: Tool failure behavior prediction in fabrication processes. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 54(6), 3450-3460.