Journal of Applied Automation Technologies

Research Paper

Yield Drift Prediction in Wafer Lithography Using a TabTransformer-CatBoost Hybrid Model

Agnieszka Agata Filipek[1] and Patrycja Homa[1,*]

[1] Faculty of Electrical and Automatic Control, University of Warmia and Mazury in Olsztyn, Olsztyn, 10-719, Poland
*Corresponding author: patrycja.ho@uwm.edu.pl

Abstract. Yield drift in wafer lithography is often caused by weak interactions among tool state, product context, reticle history and metrology response, and fixed-limit monitoring is thus unable to detect early-stage degradation. Develop a TabTransformer-CatBoost hybrid model for lot-level drift prediction based on mixed manufacturing records in this paper. The pipeline has already aligned the collected 186,420 batches, which were gathered over a period of 14 months, encoding product, layer, scanner, track, formula, and maintenance status. It transmitted the learned representations, along with standardized continuous variables and explicit time indicators, to the CatBoost residual learner thru. Chronological rolling-origin validation avoids the leakage of future process states into training. In the held-out three-month interval, the mean absolute error of the hybrid was 0.842 percentage points and its area under the precision-recall curve for detecting a yield loss greater than 2.0 points was 0.913. It improves the error by 18.9% compared to CatBoost and by 24.1% over a standalone TabTransformer. At a fixed false-alarm rate of 5%, 91.6% of the drift episodes are identified and have a median lead time of 7.4 lots. After plasma correction, the calibration error is still 0.021. Ablation and transfer experiments have shown that contextual categorical embeddings and process-age variables are complementary additions, especially for maintenance and product-layer modifications. Based on the above results, attention-based categorical representation can be combined with boosted-tree decision surfaces to achieve accurate, calibrated, and operationally interpretable early warnings for lithography yield control.

Keywords: *TabTransformer, CatBoost, Wafer Lithography, Yield Drift, Categorical Embedding, Semiconductor Manufacturing*

Received on 08 July 2023, Accepted on 27 November 2023, Published on 05 December 2023

Introduction

Optical lithography selects the locations for forming functional structures on a wafer, and the yield impact of a small process deviation is rarely limited to a single measured parameter. Dose, focus, overlay, resist thickness, post-exposure bake, reticle condition and local equipment history all affect the product and layer requirements. Therefore, modern fabs collect extensive fault-detection, metrology, equipment and manufacturing-execution data [1]. Statistical Process Control is needed for traceability, but univariate analysis fails to identify correlated changes that are individually small [2]. Virtual Metrology has extended the observable process state by estimating unavailable measurements from equipment signals [3]. Data-driven yield learning further links upstream traces with downstream electrical results [4]. The problem is that lithography records contain many high-cardinality categorical fields, delayed labels, irregular sampling, and maintenance-induced regime changes [5]. The aforementioned characteristics are insufficient to ensure a stable average error: the engineering predictor should be able to identify directional declines early and indicate risks within thresholds suitable for the operator [6]. The hierarchy of lithography control has added to the data problem. A scanner-level shift is shown in the lot-specific sampling plan, and the yield impact is determined by design density, exposure field and the scope of downstream redundancy. Equipment trace may contain thousands of samples, but the production data mart often only stores quantiles, slopes, alarms and recipe identifiers. Therefore, a predictor must operate based on an imperfect but operationally feasible representation. Distinguish between normal wear and tear of

https://doi.org/10.64972/jaat.2023v1.320p29e:389-403

equipment and damage caused by abnormal operation, such as unauthorized modifications to the dosing module. Due to these differences, a mixed-type model is used to condition numerical evidence on categorical context rather than treating all lots as one homogeneous group.

Tree ensembles are suitable for the heterogeneous factory data and can handle missing values by learning non-linear splits [7]. Deep tabular models can also be used to represent categories as trainable embeddings instead of arbitrary integer codes [8]. Attention-based architectures can condition each categorical field on the full production context [9], and this is necessary because the same scanner reading has different connotations in a critical layer and a non-critical layer. Gradient Boosting still performs well because it models sparse interactions with strong regulation [10]. Recently, semiconductor research has also emphasized sequential validation rather than random validation, as random folding mixes different process epochs and overestimates generalization capability [11]. Drift detection work shows that alarms should be judged by persistence, calibration and lead time in addition to classification accuracy [12]. However, most existing studies either select an end-to-end neural representation or a tree learner and have not explored the complementary behavior of contextual categorical attention and boosted residual partitioning. An actual case will also be presented in the test. Engineers do not act on a single predicted percentage; they determine whether the risk remains, if the affected lots share a tool-layer route, and whether the proposed drivers align with recent equipment operation data. Even if a model's overall error is excellent, if the probability calibration is inaccurate or the alerts arrive only after the yield results are confirmed, it may still be unsuitable. Alternatively, a marginally less accurate model may be suitable if it is to provide an early warning. Therefore, this paper will consider regression, classification, calibration and episode timing as related views of the same forecasting problem.

Therefore, this paper introduces a TabTransformer-CatBoost hybrid model for lot-level lithography yield-drift prediction. First, the model contextually embeds product, layer, tool, recipe, reticle, and maintenance categories; then, it combines these representations with continuous process summaries and temporal drift descriptors in a CatBoost learner. Event-time alignment, rolling-origin validation, probability calibration, drift-episode scoring, and grouped attribution are used in this study to link predictive performance with factory use. The main achievement is not the combination of two powerful algorithms, but a leakage-controlled fusion that uses attention for categorical meaning and boosting for discontinuous risk surfaces. Assess the accuracy, early-warning capability, ablation effects, temporal stability and cross-tool transfer of the hybrid model, and provide a reproducible foundation for selecting between the components or the hybrid. The paper also separates representation learning from operational inference. The neural encoder is not required to learn all the numerical features or issue the final alarm independently. Its limited function is to learn how the categorical states change. The boosted stage can then keep the sharp divisions, missing-value paths and monotonic operating thresholds that are natural in process tables. Divide and conquer can simplify the audit of a hybrid by attributing the changes to context representation, temporal descriptors or the downstream decision surface through controlled ablations.

Related Work

Yield Learning and Drift Monitoring

Manufacturing yield prediction has moved from rule-based excursion handling to supervised learning of multistage process data. Virtual Metrology models infer expensive or delayed measurements from tool traces and thus shorten the feedback loop [13]. Ensemble methods have been employed to address the nonlinear characteristics of wafer-level defect and yield data without a complete physical model [14]. Deep networks have extended this ability to high-dimensional signals, but their performance is very sensitive to label alignment and production-route consistency [15]. Temporal convolution and recurrent architectures can exploit ordered sensor traces [16], and graph formulations represent dependencies among operations, chambers or wafer maps [17]. The above methods are suitable for lithography when the raw time series are available, but many production repositories only expose engineered summaries linked to categorical route descriptors [18].

A second research stream is concept drift in industrial prediction. Distribution changes can occur gradually due to component aging, suddenly after maintenance, or repeatedly with product mix. Window-based retraining can track slow changes but may forget rare operating states. A detector is overly sensitive to changes in the signal and thus reports a fault falsely. Therefore, deployment studies are increasingly dividing sample-level errors into

episode-level detection and report warning delay, persistence and false-alarm burden. The first two are especially suitable for lithography; a single-lot fluctuation may be negligible, but a continuous decline across several related lots needs to be investigated or held.

Label construction is a relatively small but required part of yield modelling. The final electrical yield will be known several days after lithography because of all the other steps. Directly assigning all downstream losses to lithography would result in noisy supervision. Therefore, the studies use route-conditioned baselines, excursion screens or confirmed engineering dispositions to isolate probable upstream degradation. Reduce the number of labels and improve specificity. The current work uses a persistent deviation label and evaluates it at the episode level; thus, although the model can learn weak precursors, it does not claim that every predicted deviation has a single lithography cause.

Contextual Tabular Modeling

Categorical variables are required in semiconductor manufacturing to specify the products, layers, tools, recipes, reticles and routes for interpretation of continuous measurements. Target encoding is concise but may leak label information if not strictly fitted within each training window [19]. Entity embeddings learn distributed representations, but independent embeddings do not explicitly map one field to the other category in the same record [20]. TabTransformer introduces a self-attention mechanism for categorical tokens, allowing tool tokens to be contextualized based on layer, recipe, and maintenance state [21]. It has a good representation of rare combinations, but the end-to-end multilayer perceptron may smooth out the threshold effect near the process limits.

Gradient Boosting offers an additional inductive bias. CatBoost addresses the issue of ordered target statistics and symmetric trees for categorical data in [22]. It performs well for medium-sized mixed tables, supports missing values, and shows feature contributions; however, very high-cardinality interactions can be split across multiple tree nodes. Therefore, hybrid tabular models have started to combine learned embeddings with classical learners [23]. The open engineering question is how to build this fusion without introducing noise in the time validation, duplicating category information, or generating an alarm score that is accurate but poorly calibrated. The method developed below treats the contextual encoder as a representation module trained within each chronological fold and provides the boosted learner with explicit raw continuous and time-state features along with the embeddings.

The two model families have different types of failure. Attention can distribute information over rare category combinations and produce smooth representations, but the learned geometry changes during retraining and is not directly interpretable by an engineer. Boosted trees divide the feature space relatively well and uncover additive contributions; however, some splits are inefficient because of category interactions. Passing fold-specific contextual vectors to boosting is therefore feasible if the encoder is trained without future data and attribution dimensions are regrouped into process concepts. This Design does not consider the embedding coordinate a physical quantity and keeps the raw process features for engineering analysis.

Drift-Aware Hybrid Methodology

Event-Time Data Construction and Drift Labels

The observation unit is a completed lithography lot at the moment when its last eligible inline metrology result becomes available. Records are joined from the manufacturing execution system, scanner and resist-track summaries, reticle management system, maintenance logs, and downstream yield database. A strict event-time constraint excludes measurements posted after the prediction timestamp. This constraint prevents delayed inspection results and future equipment states from entering the feature vector.

The continuous variables include exposure-dose statistics, focus deviations, overlay residuals, critical-dimension summaries, resist thickness, bake temperature, alignment quality, environmental measurements, elapsed time since maintenance, processed-wafer count, and rolling tool-load indicators. The categorical variables include product family, technology node, layer group, scanner, track, recipe, reticle family, process route, shift, and maintenance class. Missing measurements are retained through availability indicators instead of being replaced by globally estimated values.

For lot i, the yield deviation is measured relative to a route-conditioned historical baseline. Let y_i denote the observed good-die yield, and let B_i be the median yield of eligible lots belonging to the same product-layer family within the preceding stable window. The signed yield-drift magnitude is defined as

$$d_i = B_i - y_i. \quad \text{Eq.(1)}$$

A positive d_i represents yield deterioration. The binary drift label uses an engineering threshold $\tau = 2.0$ percentage points. To suppress isolated measurement noise, the threshold must be exceeded by at least two of the next three eligible lots:

$$z_i = \mathbb{I}(d_i \geq \tau)\mathbb{I}\left[\sum_{j=i}^{i+2} \mathbb{I}\big(d_j \geq \tau\big) \geq 2\right]. \quad \text{Eq.(2)}$$

All temporal descriptors are calculated exclusively from observations available before lot i. For a continuous process variable x and its trailing window $\mathcal{W}_i$, the robust standardized displacement is

$$r_i(x) = \frac{x_i - median(\mathcal{W}_i(x))}{1.4826MAD(\mathcal{W}_i(x)) + \varepsilon}. \quad \text{Eq.(3)}$$

Here, MAD($\cdot$) denotes the median absolute deviation, and ε is a small positive constant that prevents numerical instability. The robust displacement reduces sensitivity to occasional metrology spikes while retaining gradual movements in the process center.

Rules for data quality are applied before the calculation of temporal features. Duplicate equipment messages are avoided using the original equipment timestamp and ingestion sequence. Measurements outside the physically qualified range are marked as unavailable instead of being Winsorised to appear valid. The approved recipe modification will generate a transition indicator and reset the corresponding trailing window. Maintenance age is expressed as both elapsed hours and processed-wafer count because calendar time alone cannot account for wear caused by high-utilization periods of equipment.

A total of 186,420 qualified lots were collected over 14 months. The first eight months are for the initial training, the next three months are for rolling validation, and the final three months will be used for the independent test. Each validation origin is 2 weeks behind. Product-layer groups with fewer than 80 training observations are mapped to rare-context tokens but are still in the validation and test sets. This way can assess the system's performance in cold-weather or sparse-observation production without missing any hard-to-collect data.

Training weights compensate for the highly unequal production volume among product-layer groups. For group g containing n_g training lots, the sample weight is

$$w_i = min\left(4, \sqrt{\frac{\bar{n}}{n_{g(i)}}}\right) \quad \text{Eq.(4)}$$

where $\bar{n}$ is the median group size and $g(i)$ identifies the group of lot i. The upper bound prevents a small number of rare lots from dominating model optimization.

Contextual TabTransformer Encoder

Each categorical field is converted into a trainable token that includes a category embedding and a field-identity embedding. The field component is an identical integer index from a different category. The input token of lot i for categorical field c is

$$\boldsymbol{e}_{i,c} = \boldsymbol{E}_c\big[k_{i,c}\big] + \boldsymbol{q}_c \quad \text{Eq.(5)}$$

where $\mathbf{E}_c$ is the category-embedding matrix, $k_{i,c}$ is the category index, and $\mathbf{q}_c$ is the embedding representing the identity of field c.

The encoder has four Transformer blocks, eight attention heads, a 32-dimensional embedding for each categorical field, pre-layer normalization and a dropout rate of 0.1. Not a time-series model, the order of the categorical tokens is determined by field identity. Self-attention is used to model the conditional relationships between products, layers, scanners, tracks, recipes, masks, routes, and maintenance statuses.

For attention head h in Transformer layer l, contextual interaction is calculated as

$$\boldsymbol{A}_h^{(l)} = softmax\left(\frac{\boldsymbol{Q}_h^{(l)}\boldsymbol{K}_h^{(l)\top}}{\sqrt{d_h}}\right)\boldsymbol{V}_h^{(l)} \quad \text{Eq.(6)}$$

where d_h is the dimension of one attention head. Residual connections preserve the original categorical information, while gated feed-forward layers refine nonlinear interactions identified by the attention mechanism.

The final contextual representation combines mean pooling with trainable attention pooling:

$$\boldsymbol{h}_i = \left[\frac{1}{C}\sum_{c=1}^{C}\boldsymbol{H}_{i,c}^{(L)} \| \sum_{c=1}^{C}\alpha_{i,c}\boldsymbol{H}_{i,c}^{(L)}\right] \quad \text{Eq.(7)}$$

subject to

$$\alpha_{i,c} = \frac{exp\left(\boldsymbol{v}^\top \boldsymbol{H}_{i,c}^{(L)}\right)}{\sum_{c'=1}^{C} exp\left(\boldsymbol{v}^\top \boldsymbol{H}_{i,c'}^{(L)}\right)}. \quad \text{Eq.(8)}$$

Here, C denotes the number of categorical fields, L is the final Transformer layer, and $\|$ represents vector concatenation. Mean pooling preserves distributed contextual information, whereas attention pooling emphasizes categorical fields that are especially informative for the current lot.

The encoder is optimized using a Huber regression objective and a focal classification objective:

$$\mathcal{L}_{enc} = \frac{1}{N}\sum_{i=1}^{N} w_i \mathcal{H}_\delta\left(d_i - \hat{d}_i^{enc}\right) + \lambda_f \frac{1}{N}\sum_{i=1}^{N} w_i FL(z_i, p_i^{enc}). \quad \text{Eq.(9)}$$

The Huber component estimates the continuous yield-drift magnitude without allowing a few extreme events to dominate training. The focal component emphasizes difficult persistent-drift samples while reducing the influence of easily classified stable lots.

A masked-category reconstruction task is added as a representation regularizer. During training, 5% of the nonmissing category tokens are replaced with an unknown token. The encoder then predicts the original category from the remaining contextual fields:

$$\mathcal{L}_{total} = \mathcal{L}_{enc} + \lambda_m \sum_{c\in\mathcal{M}_i} CE\left(k_{i,c}, \hat{k}_{i,c}\right) \quad \text{Eq.(10)}$$

where $\mathcal{M}_i$ is the set of masked fields and $\mathrm{CE}(\cdot)$ denotes cross-entropy loss. The reconstruction weight is set to $\lambda_m = 0.10$. This auxiliary task discourages the encoder from compressing all categorical information into the yield label and improves robustness to new reticles, routes, and equipment combinations.

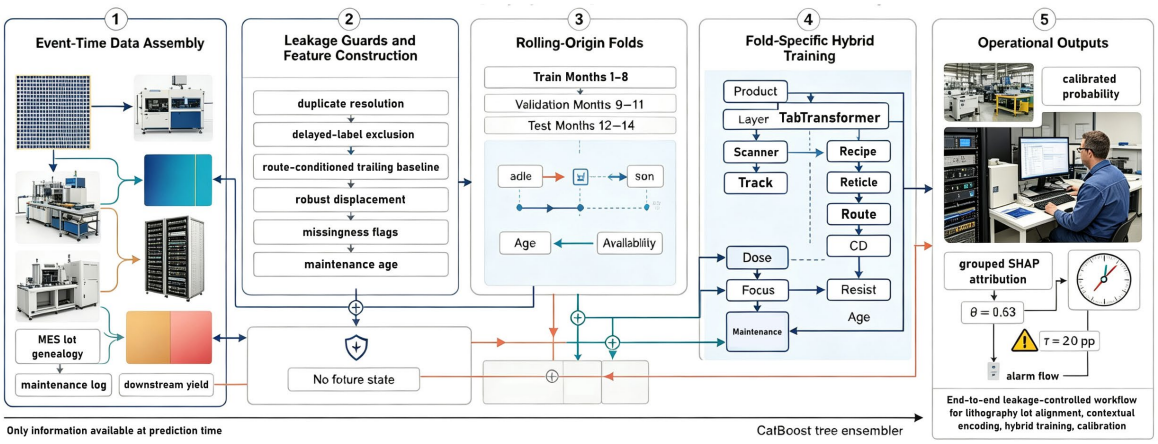

Figure 1. End-to-end leakage-controlled workflow for lithography lot alignment, contextual encoding, hybrid training, calibration, and drift-episode scoring.

Optimization uses AdamW with an initial learning rate of 3×10^{-4}, cosine learning-rate decay, weight decay of 10^{-5}, and chronological batches of 1,024 lots. Early stopping is determined only from the current rolling validation slice. After training, contextual vectors are generated for the corresponding training and validation

records using the fold-specific encoder. Embeddings trained with observations from a later origin are never reused for an earlier origin. Figure 1 presents the complete data-processing and representation-learning workflow, including event-time extraction, leakage prevention, contextual tokenization, fold-specific training, and chronological embedding export.

CatBoost Fusion, Calibration, and Interpretation

The CatBoost stage receives the contextual representation $\mathbf{h}_i$, standardized continuous variables, robust temporal displacements, measurement-availability indicators, and a restricted subset of stable raw categorical variables. Highcardinality fields already represented by the TabTransformer are excluded from the raw CatBoost inputs to avoid redundant categorical statistics. Let $\mathbf{u}_i$ denote the resulting fusion vector. The predicted yield-drift magnitude is generated by an additive ensemble of symmetric decision trees:

$$\hat{d}_i = f_0 + \eta \sum_{m=1}^{M} T_m(\boldsymbol{u}_i) \qquad \text{Eq.(11)}$$

where f_0 is the initial prediction, T_m denotes the m-th symmetric tree, M is the total number of trees, and η is the learning rate. The final configuration contains 1,200 depth- 8 trees, a learning rate of 0.035, and an L2 regularization coefficient of 5.

The boosted learner is optimized using a combination of Huber regression and pairwise ranking. The Huber component limits the influence of exceptionally large yield deviations, whereas the ranking term encourages deteriorating lots to receive higher scores than nearby stable lots from the same product-layer context:

$$\mathcal{L}_{CB} = \sum_{i=1}^{N} w_i \mathcal{H}_\delta\big(d_i - \hat{d}_i\big) + \lambda_r \sum_{(i,j)\in\mathcal{P}} ln\big[1 + exp\big(-\hat{d}_i + \hat{d}_j\big)\big]. \qquad \text{Eq.(12)}$$

The comparison set $\mathcal{P}$ contains a deteriorating lot i and a stable lot j observed within a nearby production interval. Pairs are constructed within the same product-layer family so that naturally lower-yield products are not incorrectly treated as drift cases.

The continuous prediction is converted into an initial risk score relative to the engineering threshold τ. Because this score is not guaranteed to represent an empirical probability, isotonic regression is subsequently fitted using the most recent chronological validation slice:

$$s_i = \frac{1}{1 + exp\big[-\big(\hat{d}_i - \tau\big)/T\big]}, \tilde{p}_i = g_{iso}(s_i) \qquad \text{Eq.(13)}$$

where T is a temperature parameter, $g_{\text{iso}}(\cdot)$ is the fitted isotonic mapping, and $\tilde{p}_i$ is the calibrated probability of persistent yield drift. Calibration parameters are estimated after model selection and remain fixed during independent test evaluation.

An operational warning is not issued from a single high-risk prediction. The calibrated probability must exceed threshold θ for two consecutive eligible lots in the same scanner-layer stream:

$$A_i = \mathbb{I}(\tilde{p}_i \geq \theta)\mathbb{I}(\tilde{p}_{i-1} \geq \theta). \qquad \text{Eq.(14)}$$

The threshold is set to have a 5% false-alarm rate in the validation data. Two consecutive lots need to be taken to smooth out the isolated fluctuations and retain the sensitivity of continuous deterioration. An episode is deemed to have occurred when the first alarm is triggered ten lots before and two lots after the confirmed start of the episode. Warning lead time is then defined as the number of eligible lots from the first alarm to the confirmed start time. A positive value is a prompt alarm; a negative value is a delayed alarm.

Model interpretation is based on SHAP contributions. Individual embedding dimensions are not assigned direct physical meanings because they encode distributed categorical relationships. Their absolute contributions are instead aggregated into predefined engineering groups:

$$\Phi_{i,g} = \sum_{j\in g} \big|\phi_{i,j}\big| \qquad \text{Eq.(15)}$$

https://doi.org/10.64972/jaat.2023v1.320p29e:389-403

where $\phi_{i,j}$ is the SHAP contribution of feature j for lot i, and g represents an engineering group. The groups include scanner optics, resist-track state, reticle history, route context, maintenance age, and environmental conditions. Their definitions are fixed before test evaluation to prevent post hoc regrouping.

The Three Prediction Quality Measures are shown below. Mean absolute error is the size of the continuous drift error. The area under the precision-recall curve shows discrimination in imbalanced data. Expected calibration error is calculated based on ten equal-frequency probability bins to determine the discrepancy between predicted and observed drift frequencies. It also provides event recall rate, accuracy rate, false positive rate, and median warning lead time to describe operational behavior.

Hyperparameter Selection follows a hierarchical temporal order. Tree depth, learning rate, L2 regularization, embedding dimension and ranking-loss weight are chosen from the first two rolling validation origins. The third origin is for probability calibration and alarm-threshold setting. The search criteria include normalized mean absolute error, precision-recall performance, and calibration error. Prediction delay is considered a deployment constraint and is not included in the optimization objective.

Figure 2 shows how contextual categorical representations, continuous process measurements, temporal drift descriptors, CatBoost fusion, probability calibration, persistent alarm logic and grouped engineering explanations are combined in a single prediction system.

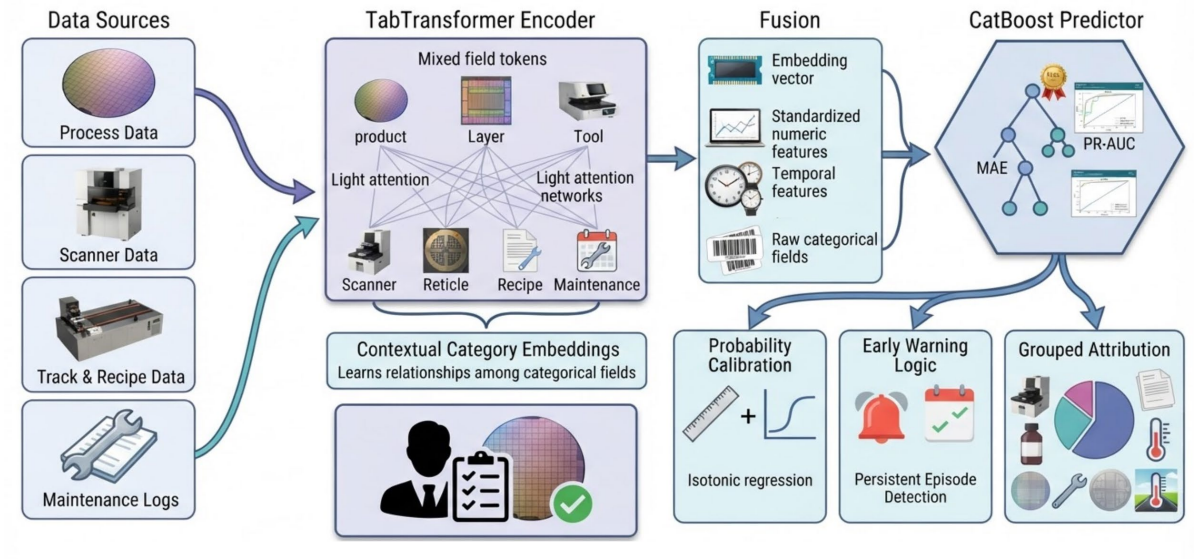

Figure 2. Architecture of the TabTransformer-CatBoost fusion model with calibrated early-warning and grouped attribution outputs.

Temporal Evaluation and Engineering Analysis

Experimental Protocol and Benchmark Configuration

Over 14 months, six scanners and five linked resist tracks provided 186,420 lithography lots for this study. The median lot has 23 eligible continuous variables, 10 categorical fields, 18 robust temporal descriptors and 23 availability indicators. Downstream yield is available for 96.8% of aligned lots; the remaining lots are not included in supervised loss but are kept for computing historical process windows. The last three-month test period had 38,614 lots and 1,284 cases of continuous drift. Label prevalence is 8.7% at the block level. Hyperparameters are selected at each rolling origin, and then fixed for the test period. CatBoost uses 1,200 depth-8 trees, learning rate 0.035, L2 regularization 5, and ordered boosting. The hybrid uses 256-dimensional pooled contextual features. All reported confidence intervals are obtained by 1,000 block-bootstrap samples over tool-day blocks.

Baselines include elastic net, random forest, XGBoost, CatBoost, a multilayer perceptron, FT-Transformer, and standalone TabTransformer. Preprocessing and chronological divisions are the same. Previous work has shown that random splitting can overstate the performance of evolving production streams [24], so a random-fold result is presented only as a diagnostic. Probability Calibration Follows Model Selection. Episode scoring has a

fixed tolerance window and, therefore, is consistent with the industrial drift evaluation in [25] that distinguishes early warnings from isolated sample classifications. One NVIDIA A100 GPU is used for neural encoding and a 32-core CPU for boosting in the computation of measurements; prediction latency includes feature assembly at a batch size of 256.

The target baseline is recalculated at each prediction time and not derived from the completed dataset. The median baseline yield of the product-layer groups is between 91.7% and 99.2%, so an absolute yield threshold is not suitable. Persistent episodes have a median length of 6 lots and a 90th-percentile of 21 lots. Maintenance occurs in 3.8% of the lot history over the past 48 hours. Missingness is structured: overlay summary availability is 94.1%, critical-dimension metrology 87.6%, and resist-thickness metrology 72.4%. Given the different rates of various products and routes, a single complete-case subset would have to exclude some older and lower-volume products. Availability flags can inform the model that the absence is due to the sampling plan and keep the measured-value channel when present.

Five random seeds are used to repeat the benchmark for neural networks. Variation among seeds is small for the hybrid: Test MAE is 0.833-0.854, and the range of precision-recall area is 0.907-0.918. Standalone TabTransformer is less stable, with an MAE of 1.072 to 1.146, and thus the boosted stage absorbs some representation variability. Hyperparameter selection consumes about 41 GPU-hours and 63 CPU-hours, and the final encoder and booster are used for 2.8 GPU-hours and 1.6 CPU-hours, respectively. The above costs are suitable for weekly or event-triggered retraining; they are not suitable for retraining after every batch.

To check whether the results are sensitive to the selection of the 2.0-point label, the whole evaluation will be repeated at thresholds of 1.5 and 2.5. At 1.5 points, the prevalence of the event is 13.2% and the hybrid has a precision-recall area of 0.901; at 2.5 points, the prevalence is 5.6% and the area is 0.924. MAE remains the same as the regression targets are continuous. The stable ranking of the model indicates that mixed advantages are not due to a single convenient binarization. Episode recall is also recalculated for confirmation windows of two-of-three and three-of-five lots. The latter has a longer event and increases apparent recall by 1.9 points, but it is not formally started. Therefore, the more responsive two-of-three definition will be retained for the main analysis.

Figure 3. Dataset chronology and drift structure: (a) monthly lot volume, persistent-drift prevalence, and maintenance events; (b) dose-focus displacement density with yield-drift contours and decision boundary; (c) product-layer coverage and rare-context share across chronological folds.

https://doi.org/10.64972/jaat.2023v1.320p29e:389-403

Figure 3 is cited multiple times through its subpanels: Figure 3(a) shows the monthly lot volume, drift prevalence and maintenance markers; Figure 3(b) shows the joint distribution of dose displacement and focus displacement with the 2.0-point drift boundary; Figure 3(c) compares product-layer coverage across chronological folds and highlights 17 rare contexts. The proportion has risen from 6.1% in month 3 to 11.8% in month 12, and the number of lots is still between 11,420 and 15,870 per month. The above separation shows that the test period is not merely a small sample but also a later and more deviated process period.

Predictive Accuracy, Calibration, and Early Warning

On the held-out interval, the hybrid achieves an MAE of 0.842 percentage points, root mean square error of 1.317 points, R-squared of 0.782, and precision-recall area of 0.913. CatBoost is the strongest single-stage baseline at 1.038 MAE and 0.865 precision-recall area, whereas standalone TabTransformer reaches 1.109 and 0.841. FT-Transformer records 1.074 and 0.852. The hybrid therefore reduces MAE by 18.9% relative to CatBoost and by 24.1% relative to standalone TabTransformer. Block-bootstrap 95% intervals for the MAE difference against CatBoost are 0.161-0.231 points, excluding zero. A recent comparison of tabular learners notes that boosted trees frequently dominate on heterogeneous medium-scale data [26]; the present result shows that contextual embeddings can improve that strong base when categories have process-dependent meaning.

Error stratification shows where the aggregate gain originates. On common product-layer contexts, the hybrid improves CatBoost MAE from 0.986 to 0.821. In the rare case above, the corresponding values were 1.412 and 1.167; although the absolute gain was larger, a performance difference still existed. Precision at the selected threshold for the critical layer is 81.3%, and for the non-critical layer it is 75.9%, as the tighter process window of the critical layer produces more coherent precursors. Worse performance after maintenance: MAE is 0.973 in the first 24 hours, and then it returns to 0.851 after 72 hours. Explicitly maintained-age accounts for about half of this recovery in the ablation analysis.

Figure 4 is referenced once, and each sub-panel contains a different comparison: Figure 4(a) shows MAE and precision-recall area for the eight models with bootstrap intervals, and Figure 4(b) plots error against prediction latency and model size. The hybrid needs 6.8 ms per 256-lot batch and is 47 MB; CatBoost is 3.1 ms and 31 MB; TabTransformer is 9.6 ms and 62 MB. The operating point is still below the 10ms batch-latency budget. The Pareto diagram also indicates that random forest is relatively computationally expensive for the level of accuracy it offers, and elastic net is fast but cannot capture interaction-driven drift.

Figure 4. Accuracy-efficiency benchmark: (a) MAE and precision-recall area with 95% block-bootstrap intervals; (b) prediction latency versus model size, colored by drift recall and annotated with the deployment budget.

Threshold Sensitivity is evaluated at 3% and 10% validation false-alarm targets. At 3%, episode recall is 86.2% and precision is 84.7%; at 10%, recall increases to 95.8% and precision drops to 66.5%. A 5-percentage-point increase is chosen, which will be about 7.1 reviewable alarms per scanner-week. Alarm clusters with the same scanner, layer group and 12-hour interval are grouped into a single engineering ticket, reducing the ticket count to 4.6 per scanner-week. The quantity of work is not directly related to the percentage of false alarms at the list level.

Calibration is needed because the alarm threshold has been set as a probability instead of an absolute value. The expected calibration error of the hybrid before calibration was 0.063, and it overpredicted risk in the highest decile. Isotonic correction lowers the value to 0.021 and reduces the Brier score from 0.079 to 0.066. CatBoost is still 0.038 after the same correction. Calibration practices for imbalanced industrial classifiers recommend

reporting reliability together with discrimination [27]. At the selected threshold theta=0.63, the hybrid has achieved 91.6%-episode recall, 78.4% precision and a 5.0% false-alarm rate. Median lead time is 7.4 lots; 75% of the detected episodes are flagged at least four lots before confirmation.

The magnitude estimates are evaluated for decision support as well as absolute error. For lots predicted to lose at least 3.0 points, observed mean loss is 3.26 points and the 10th-90th percentile interval is 2.18-4.71. CatBoost predicts 2.91 points for the same selected group, showing conservative compression of severe outcomes. The hybrid's Huber and ranking losses reduce this compression without producing unstable extremes: only 0.3% of predictions exceed the maximum drift observed in training, and those values are clipped for alarm presentation but retained in the error audit. Residuals show mild heteroscedasticity with increasing drift magnitude, so bootstrap intervals are calculated over tool-day blocks instead of relying on normal-error assumptions.

Figure 5 is cited once through four complementary diagnostic panels. Figure 5(a) gives reliability curves and sample counts per probability bin; Figure 5(b) shows precision-recall traces with the theta=0.63 operating point; Figure 5(c) reports lead-time distributions for CatBoost, TabTransformer, and the hybrid; Figure 5(d) displays recall and false-alarm rate as persistence changes from one to four consecutive lots. The hybrid's median lead time exceeds CatBoost by 2.1 lots, and requiring two lots removes 38% of isolated alarms while reducing episode recall by only 1.7 points. A three-lot rule lowers false alarms to 3.4% but sacrifices 6.9 recall points, so two-lot persistence is retained.

Figure 5. Calibrated warning behavior: (a) reliability curves with equal-frequency bin counts; (b) precision-recall curves and selected operating point; (c) episode lead-time distributions with quartiles; (d) recall and false-alarm trade-off under one- to four-lot persistence rules.

Ablation, Temporal Robustness, and Engineering Interpretation

Ablation removes a fusion component to observe the effect. Removing contextual embeddings increases MAE from 0.842 to 0.992 and reduces the precision-recall area from 0.913 to 0.881. After removing temporal displacement features, they are 0.931 and 0.892. Replace fold-specific encoder training with a single encoder trained on the entire pretest interval; although this configuration reduces MAE to 0.811, it leaks later validation context and is thus excluded. Therefore, representation learning must be performed within each time window for this reason. Research on evaluation leakage has shown that a relatively small preprocessing shortcut can

lead to considerable overoptimism [28]. Removing missingness flags has a small overall effect, but recall after maintenance drops from 89.7% to 83.2%; thus, the measurement availability itself contains state information.

Interactive analysis supports the design of labor division. If Contextual Embeddings are replaced by one-hot categories, CatBoost needs 18% more trees and still fails to improve the MAE by 0.119. When only the neural network processes continuous features, the MAE increases by 0.087; therefore, direct boosted splits are still effective near the process limit. Attention inspection shows that layer tokens give the highest average weight to recipe and product, and scanner tokens pay close attention to maintenance class and track. Attention weights are not shown as explanations; rather, their pattern indicates that the encoder has used cross-field context instead of independent embeddings.

Figure 6 is cited once with three views of robustness. Figure 6(a) is an ablation waterfall showing MAE increments and confidence intervals; Figure 6(b) traces monthly MAE, drift recall, and population-stability index across the test era; Figure 6(c) shows a heat map of transfer MAE from each training scanner to each target scanner. Monthly hybrid MAE ranges from 0.801 to 0.917 as population-stability index rises from 0.08 to 0.27. Scanner S6, introduced late in the period, is the hardest target at 1.104 MAE without adaptation. Fine-tuning the embeddings on 600 labeled S6 lots reduces this to 0.936, and refitting only CatBoost yields 0.958. Domain shift in semiconductor tools often concentrates in equipment-specific offsets while route relations remain partly shared [29], which is consistent with the advantage of limited embedding adaptation.

Figure 6. Robustness and transfer analysis: (a) ablation waterfall with MAE increments and confidence intervals; (b) monthly MAE, episode recall, and population-stability index; (c) cross-scanner transfer MAE heat map with limited-shot adaptation annotations.

Temporal Retraining Simulations Compare Fixed, Monthly and Drift-Triggered Policies. The fixed model trained in month 11 has a 0.917 MAE after 14 months. Monthly retraining maintains the monthly MAE at 0.879 but needs three full updates. A trigger at population-stability index 0.20 performs one update and stops at 0.866 MAE. Calibration decays faster than ranking: without recalibration, the expected calibration error is 0.049, and the precision-recall area remains above 0.90. A light-weight weekly isotonic refresh with recently confirmed labels restores the error to 0.026 without modifying the encoder or trees. Therefore, the frequency of probability maintenance may be higher than that of representation retraining.

Grouped attribution provides a process-level explanation rather than a list of embedding coordinates. Across all alarms, scanner optics accounts for 27.8% of absolute attribution, route context 21.4%, resist-track state 16.9%, maintenance age 13.7%, reticle history 11.2%, and environment 9.0%. This ranking changes by episode type. In

dose-drift episodes, scanner optics rises to 41.3%; after track maintenance, resist-track state and missingness together reach 35.6%. SHAP-based explanations should be interpreted as model attributions rather than causal effects [30]. To test stability, the top-two group overlap is computed across 20 bootstrap refits and reaches 0.84 for persistent alarms, compared with 0.61 for isolated high-risk lots. Stable grouped explanations are therefore shown to engineers only after the persistence rule is met.

Figure 7 is cited once and links the model output to three representative episodes. Figure 7(a) overlays observed yield, predicted drift, calibrated probability, the 0.63 alarm threshold, and maintenance events for a focus-control degradation; the first warning precedes confirmation by nine lots. Figure 7(b) is a grouped-attribution stream graph for a dose drift, where scanner-optics contribution rises from 22% to 48% before the yield boundary is crossed. Figure 7(c) presents counterfactual risk curves for focus correction, dose recentering, and reticle replacement; recentering dose by 0.7 robust standard deviations lowers predicted risk from 0.86 to 0.37, while reticle replacement changes it only to 0.79. Counterfactual analysis is constrained to observed process ranges because unconstrained feature perturbations can be operationally invalid [31].

Engineering review of 120 stratified alerts provides a qualitative check on model outputs. Reviewers, blinded to model identity, rate 76 hybrid alerts as directly actionable, 29 as informative but requiring additional metrology, and 15 as nonactionable. The corresponding actionable count for CatBoost is 61. Agreement between two reviewers is 0.78 by Cohen's kappa. The most useful explanation pattern is a paired rise in scanner-optics and maintenance-age attribution accompanied by a coherent focus displacement. Route-context dominance without a supporting continuous deviation is more often judged nonactionable. These findings do not replace a prospective trial, but they show why grouped contributions should be displayed with the underlying feature trajectories and availability state.

Figure 7. Engineering case analysis: (a) focus-control episode with observed yield, predicted drift, calibrated alarm, thresholds, and maintenance markers; (b) time-varying grouped attribution for a dose-drift episode; (c) constrained counterfactual risk curves for focus, dose, and reticle interventions.

The error audit shows the following two types of failures. First, 4.6% of the missed episodes are in the product-layer context with fewer than 80 historical lots, and the category tokens collapse to a rare-context representation. Second, a sudden reticle contamination can cause a yield loss in one lot and leaves no upstream sequence for early warning. Online adaptation can reduce the gradual distribution mismatch but cannot provide a lead time for an instantaneous unmeasured event [32]. False alarms are concentrated after the legitimate recipe qualification change; adding an approved-change indicator reduces this subset by 28% without changing the overall recall. In practice, the system will open an engineering review and not carry out automatic process

correction. Human-in-the-loop deployment is suitable because attribution and counterfactual outputs guide the focus of inspection but do not determine physical causality [33].

Conclusion

This work presented a drift-aware TabTransformer-CatBoost model for predicting lot-level yield deterioration in wafer lithography. Contextual attention represented the conditional meaning of high-cardinality production categories, and CatBoost fitted nonlinear thresholds across embeddings, continuous process summaries, temporal displacements, and availability states. Event-time alignment, fold-specific representation training, calibration, persistent episode logic and grouped attribution formed an integrated engineering workflow. The two model families were shown to have complementary inductive biases in this study, and together they supported both accurate yield estimation and practical early warning.

The current Design is limited to retrospective labels, plant-specific category taxonomies, and a lack of observations for newly qualified product-layer combinations. Group attribution describes the behavior of the model rather than the causal mechanism, and batch-level feature summaries cannot reflect sub-second scanner dynamics. Sudden contamination or hardware failure may also occur without a detectable lead-up, so no predictive model can guarantee an early warning time for all excursions. Transfer experiments show some portability among scanners, but more extensive data are needed at technology nodes and fabs.

Future work will connect the hybrid representation with self-supervised pre-training on unlabeled equipment traces, hierarchical rare-category embeddings, and continual calibration under a controlled update policy. Constraints guided by physics can enhance the credibility of counterfactual interventions, while federated evaluation can test generalization ability without exposing private production data. A prospective shadow deployment will measure the engineering review time, avoided wafer loss and alarm acceptance before any closed-loop operation is initiated.

Author Contributions

Agnieszka Agata Filipek contributes to conceptualization, methodology, software, validation, analysis, investigation, data collection, draft preparation, manuscript editing, visualization, supervision, project administration, and funding acquisition. Patrycja Homa contributes to validation, analysis, investigation, data collection, draft preparation. All authors have read and agreed with the manuscript before its submission and publication.

Funding

This research received no specific financial support from any funding agency.

Institutional Review Board Statement

Not applicable.

References

[1] Jiang, D., Lin, W., & Raghavan, N. (2020). A Novel Framework for Semiconductor Manufacturing Final Test Yield Classification Using Machine Learning Techniques. IEEE Access, 8, 197885-197895. https://doi.org/10.1109/access.2020.3034680

[2] Jiang, D., Lin, W., & Raghavan, N. (2021). A Gaussian Mixture Model Clustering Ensemble Regressor for Semiconductor Manufacturing Final Test Yield Prediction. IEEE Access, 9, 22253-22263. https://doi.org/10.1109/access.2021.3055433

[3] Kim, D., Kim, M., & Kim, W. (2020). Wafer Edge Yield Prediction Using a Combined Long Short-Term Memory and Feed- Forward Neural Network Model for Semiconductor Manufacturing. IEEE Access, 8, 215125-215132. https://doi.org/10.1109/access.2020.3040426

[4] Cho, M., Park, G., Song, M., Lee, J., Lee, B., & Kum, E. (2021). Discovery of Resource-Oriented Transition Systems for Yield Enhancement in Semiconductor Manufacturing. IEEE Transactions on Semiconductor Manufacturing, 34(1), 17-24. https://doi.org/10.1109/tsm.2020.3045686

[5] Choi, J. E., & Hong, S. J. (2021). Machine learning-based virtual metrology on film thickness in amorphous carbon layer deposition process. Measurement: Sensors, 16, 100046. https://doi.org/10.1016/j.measen.2021.100046

[6] Chen, C. H., Zhao, W. D., Pang, T., & Lin, Y. Z. (2020). Virtual metrology of semiconductor PVD process based on combination of tree-based ensemble model. ISA Transactions, 103, 192-202. https://doi.org/10.1016/j.isatra.2020.03.031

[7] Cai, H., Feng, J., Yang, Q., Li, W., Li, X., & Lee, J. (2020). A virtual metrology method with prediction uncertainty based on Gaussian process for chemical mechanical planarization. Computers in Industry, 119, 103228. https://doi.org/10.1016/j.compind.2020.103228

[8] Khakifirooz, M., Fathi, M., & Chien, C. F. (2021). Partially Observable Markov Decision Process for Monitoring Multilayer Wafer Fabrication. IEEE Transactions on Automation Science and Engineering, 18(4), 1742-1753. https://doi.org/10.1109/tase.2020.3017481

[9] Zhang, J., You, H., & Jia, R. (2020). Reliability hazard characterization of wafer-level spatial metrology parameters based on LOF-KNN method. Microelectronics Reliability, 107, 113599. https://doi.org/10.1016/j.microrel.2020.113599

[10] Chien, C. F., Hung, W. T., Pan, C. W., & Van Nguyen, T. H. (2022). Decision-based virtual metrology for advanced process control to empower smart production and an empirical study for semiconductor manufacturing. Computers & Industrial Engineering, 169, 108245. https://doi.org/10.1016/j.cie.2022.108245

[11] Wu, X., Chen, J., Xie, L., Chan, L. L. T., & Chen, C. I. (2020). Development of convolutional neural network based Gaussian process regression to construct a novel probabilistic virtual metrology in multi-stage semiconductor processes. Control Engineering Practice, 96, 104262. https://doi.org/10.1016/j.conengprac.2019.104262

[12] Susto, G. A., Maggipinto, M., Zocco, F., & Mcloone, S. (2020). Induced Start Dynamic Sampling for Wafer Metrology Optimization. IEEE Transactions on Automation Science and Engineering, 17(1), 418-432. https://doi.org/10.1109/tase.2019.2929193

[13] Shwartz-Ziv, R., & Armon, A. (2022). Tabular data: Deep learning is not all you need. Information Fusion, 81, 84-90. https://doi.org/10.1016/j.inffus.2021.11.011

[14] Farjon, G., & Bar-Hillel, A. (2022). Sparse Hierarchical Table Ensemble–A Deep Learning Alternative for Tabular Data. IEEE Access, 10, 75376-75384. https://doi.org/10.1109/access.2022.3190537

[15] Zhu, Y., Brettin, T., Xia, F., Partin, A., Shukla, M., Yoo, H., Evrard, Y. A., Doroshow, J. H., & Stevens, R. L. (2021). Converting tabular data into images for deep learning with convolutional neural networks. Scientific Reports, 11(1), 11325. https://doi.org/10.1038/s41598-021-90923-y

[16] Borisov, V., Broelemann, K., Kasneci, E., & Kasneci, G. (2022). DeepTLF: robust deep neural networks for heterogeneous tabular data. International Journal of Data Science and Analytics, 16(1), 85-100. https://doi.org/10.1007/s41060-022-00350-z

[17] Liu, Z., Zhou, Z., & Rekatsinas, T. (2021). Picket: guarding against corrupted data in tabular data during learning and inference. The VLDB Journal, 31(5), 927-955. https://doi.org/10.1007/s00778-021-00699-w

[18] Engelmann, J., & Lessmann, S. (2021). Conditional Wasserstein GAN-based oversampling of tabular data for imbalanced learning. Expert Systems with Applications, 174, 114582. https://doi.org/10.1016/j.eswa.2021.114582

[19] Shi, Q., Hu, M., Suganthan, P. N., & Katuwal, R. (2022). Weighting and pruning based ensemble deep random vector functional link network for tabular data classification. Pattern Recognition, 132, 108879. https://doi.org/10.1016/j.patcog.2022.108879

[20] Jabeur, S. B., Gharib, C., Mefteh-Wali, S., & Arfi, W. B. (2021). CatBoost model and artificial intelligence techniques for corporate failure prediction. Technological Forecasting and Social Change, 166, 120658. https://doi.org/10.1016/j.techfore.2021.120658

[21] Ahmed, I., Jeon, G., & Piccialli, F. (2022). From Artificial Intelligence to Explainable Artificial Intelligence in Industry 4.0: A Survey on What, How, and Where. IEEE Transactions on Industrial Informatics, 18(8), 5031-5042. https://doi.org/10.1109/tii.2022.3146552

[22] Minh, D., Wang, H. X., Li, Y. F., & Nguyen, T. N. (2021). Explainable artificial intelligence: a comprehensive review. Artificial Intelligence Review, 55(5), 3503-3568. https://doi.org/10.1007/s10462-021-10088-y

[23] Ferraro, A., Galli, A., Moscato, V., & Sperlì, G. (2022). Evaluating eXplainable artificial intelligence tools for hard disk drive predictive maintenance. Artificial Intelligence Review, 56(7), 7279-7314. https://doi.org/10.1007/s10462-022-10354-7

[24] Altendeitering, M., & Dübler, S. (2020). Scalable Detection of Concept Drift: A Learning Technique Based on Support Vector Machines. Procedia Manufacturing, 51, 400-407. https://doi.org/10.1016/j.promfg.2020.10.057

[25] Jayaratne, D., De Silva, D., Alahakoon, D., & Yu, X. (2021). Continuous detection of concept drift in industrial cyber-physical systems using closed loop incremental machine learning. Discover Artificial Intelligence, 1(1), 7. https://doi.org/10.1007/s44163-021-00007-z

[26] Halstead, B., Koh, Y. S., Riddle, P., Pears, R., Pechenizkiy, M., Bifet, A., Olivares, G., & Coulson, G. (2021). Analyzing and repairing concept drift adaptation in data stream classification. Machine Learning, 111(10), 3489-3523. https://doi.org/10.1007/s10994-021-05993-w

[27] Liu, A., Lu, J., & Zhang, G. (2021). Diverse Instance-Weighting Ensemble Based on Region Drift Disagreement for Concept Drift Adaptation. IEEE Transactions on Neural Networks and Learning Systems, 32(1), 293-307. https://doi.org/10.1109/tnnls.2020.2978523

[28] Kahraman, A., Kantardzic, M., & Kotan, M. (2022). Dynamic Modeling With Integrated Concept Drift Detection for Predicting Real-Time Energy Consumption of Industrial Machines. IEEE Access, 10, 104622-104635. https://doi.org/10.1109/access.2022.3210525

[29] Yang, C., Cheung, Y. M., Ding, J., & Tan, K. C. (2021). Concept Drift-Tolerant Transfer Learning in Dynamic Environments. IEEE Transactions on Neural Networks and Learning Systems, 33(8), 3857-3871. https://doi.org/10.1109/tnnls.2021.3054665

[30] Sun, Y., Sun, Y., & Dai, H. (2020). Two-Stage Cost-Sensitive Learning for Data Streams With Concept Drift and Class Imbalance. IEEE Access, 8, 191942-191955. https://doi.org/10.1109/access.2020.3031603

[31] Saqlain, M., Abbas, Q., & Lee, J. Y. (2020). A Deep Convolutional Neural Network for Wafer Defect Identification on an Imbalanced Dataset in Semiconductor Manufacturing Processes. IEEE Transactions on Semiconductor Manufacturing, 33(3), 436-444. https://doi.org/10.1109/tsm.2020.2994357

[32] Kahng, H., & Kim, S. B. (2020). Self-supervised representation learning for wafer bin map defect pattern classification. IEEE Transactions on Semiconductor Manufacturing, 34(1), 74-86. https://doi.org/10.1109/tsm.2020.3038165

[33] Kim, T., & Behdinan, K. (2023). Advances in machine learning and deep learning applications towards wafer map defect recognition and classification: a review. Journal of Intelligent Manufacturing, 34(8), 3215-3247. https://doi.org/10.1007/s10845-022-01994-1