Available online at www.sciencedirect.com

ScienceDirect

IFAC PapersOnline CONFERENCE PAPER ARCHIVE

IFAC PapersOnLine 58-4 (2024) 174–179

Sustainability in Semiconductor Production via Interpretable and Reliable Predictions ⋆

Kiavash Fathi[I, 1,2] **Maria Stramaglia**[II, 3] **Marko Ristin**[III, 1]
Marcin Sadurski[IV, 1] **Tobias Kleinert**[V, 2]
Robert Schönfelder[VI, 3] **Hans Wernher van de Venn**[VII, 1]

[I]*kiavash.fathi@rwth-aachen.de*, [II]*maria.stramaglia@hitachienergy.com*
[III]*rist@zhaw.ch*, [IV]*sadu@zhaw.ch*, [V]*kleinert@plt.rwth-aachen.de*,
[VI]*robert.schoenfelder@hitachienergy.com*, [VII]*vhns@zhaw.ch*

Abstract: Sustainability in production stands out as one of the foremost achievements facilitated by Industry 4.0. With the continuous monitoring of production lines, it is now possible to detect quality deviations at their earliest occurrence and reduce the scrap production rate. This paper studies the sustainability in context of a partially monitored multistage semiconductor production line with multiple test units. Our goal is to provide a foundation for interpretable and reliable Product Quality Prediction (PQP) in industrial use cases with noisy quality check results and missing process information. To that end, we examine how the accumulated process information from different steps of the production line impacts the accuracy of the PQP models used later as base learners. Furthermore, we highlight the drawbacks of conventional model stacking on the accuracy due to base learner's prediction quantization. We propose instead an interpretable stacked model (SM) from base learners' predicted probability values, and demonstrate how it increases the accuracy and reliability of the predictions. To improve the results even further, we analyze different calibration methods both for base learners and the SM. Our final model leads to a 19.49% reduction in the binary estimated calibration error compared to the conventional models, and thus allows for increased PQP reliability.

Keywords: Model reliability, Interpretable AI, Quality prediction, Industry 4.0.

1. INTRODUCTION

One of the most significant impacts of Industry 4.0, *i.e.* the digitalization of the industrial processes, is the constant improvement in data collection from different processes and production assets in industrial manufacturing. The gathered data paves the way for sustainability and technological transformations in numerous areas such as *Product Quality Prediction* (PQP) (Lieber et al., 2013), and preventive and predictive maintenance (Bai et al., 2020).

In the domain of PQP, many industries opt for continuous quality monitoring to detect early quality deviations during production to reduce the scrap rate. This, in turn, leads to more sustainable production due to less material waste and energy consumption (Schulze Struchtrup et al., 2020). In particular, one of the industries heavily investing in PQP is the semiconductor manufacturing. Given the complexity of the products, the industrial settings in this area have progressed from single-stage manufacturing to multistage manufacturing systems involving multiple workstations (Arif et al., 2013). The current PQP methods in semiconductor industry face limitations due to delayed availability of the data from different production stages. Regardless of the source of information (process data, alarm data, *etc.*), they rely on the complete information from the production line (Melhem et al., 2017; Jebril et al., 2022). However, the complete process information is seldomly available. These methods are thus impractical for manufacturing settings where performance, sustainability and detection of errors at earliest occurrence are of utmost importance (Wang et al., 2022b). As a viable alternative, we start by training separate PQP models on the data accumulated during the initial production stage (see Fig. 1). Once the production starts, *i.e.*, as soon as data from any (new) stages of production is available, the test results from all of the test units (TUs) in the manufacturing process are predicted using these initially trained PQP models. We then continuously accumulate information from the production line to acquire more

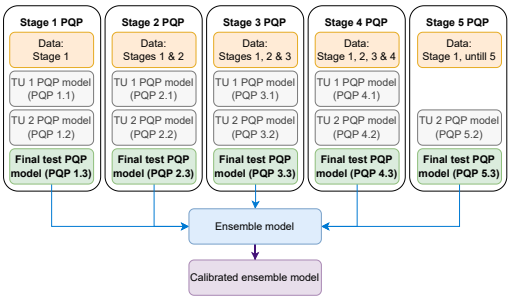


Fig. 1. Overview of our PQP approach

⋆ This research was supported by Innosuisse - Swiss Innovation Agency, Innosuisse Grant no: **51514.1 IP-ENG**, entitled **PreTune, Predictive Process Tuning by multi-perspective data analysis using a Digital Twin** in collaboration with Hitachi Energy Switzerland.

[1] Zurich University of Applied Sciences, 8400 Winterthur, Switzerland.
[2] Chair of Information and Automation Systems for Process and Material Technology, RWTH Aachen University, 52064 Aachen, Germany.
[3] Hitachi Energy Semiconductors, 5600 Lenzburg, Switzerland.

10.1016/j.ifacol.2024.07.213

accurate estimation of the product quality measured at the last TU. As our PQP models are trained for different stages of the production, we do not have to wait for the processes to finish to make a reliable and accurate PQP.

Additionally, there are sources of uncertainties in the semiconductor production lines which need to be taken into account for achieving reliable estimations. Calibrating probability estimations of the trained classifiers is commonly used to mitigate this issue in safety-critical systems (Vaicenavicius et al., 2019). To the best of our knowledge, calibrated probability estimations for PQP in the context of the industrial manufacturing has not been studied. We refer to them as the *Production Success Probability* (PSP) in this particular setting for clarity.

To demonstrate the utility of PSP, we study an operating production line as a use case. The production line contains multiple TUs with only the final TU having a fixed fail-pass criteria. Consequently, the results of the TUs preceding the final TU give us only blurred information, and therefore represent a source of uncertainty. The raw materials coming from different vendors also impact the quality, and are another source of uncertainty. This all makes it hard for the process engineer to make decisions about the final product quality.

To help reduce the uncertainties and improve the decision making, we propose to stack multiple PQP models and use their PSPs to predict the output of the final TU. We use model stacking (Dong et al., 2020) via logistic regression (LR) at the heart of our method, which gives us interpretability and also a boost in reliability without sacrificing the accuracy of the final PQP model. Using the PSP values for the stacked model (SM) also makes our approach demonstratively more data-efficient and evades much of the unwanted bias compared to the conventional calibration methods, which need separate datasets for calibration to reach a comparable level of reliability and accuracy.

Our main contributions are thus three-fold. We:

(1) Propose a PQP model which operates on partial information from the production line and at its earliest availability,
(2) Provide analysis on the performance boost given the information available from production line,
(3) Formulate an interpretable and data-efficient SM from the PQP predictions for better accuracy and increased reliability for industry.

The high accuracy of our approach as well as its ability to operate on incomplete data and handle uncertainties make it possible to use cost-effective PQP models in the industry. While the method is demonstrated on a semiconductor production line, it is easily generalizable to other fields of manufacturing as well.

2. PRODUCTION LINE UNDER STUDY

We demonstrate our approach on an operational production line from a working factory. It has important characteristics that guided the design of our method which we examine here. However, our method is not restricted to this line, and can be applied in a much wider range of settings.

The fail-pass criteria for intermediate quality tests are not fixed and are improved as more products are produced. However, the final TU is a fixed criteria which ensures the required specifications of the semiconductor are met when offered to the customers. Moreover, some processes which impact the final product quality are not monitored, see Fig. 2. Given the high uncertainty induced by the unmonitored processes, a probability value showing the chances of producing a successful product is far more practical than a simple binary output of the PQP (Vaicenavicius et al., 2019). By focusing on the probability values, we can infer the confidence of the classifier for predicting the final product quality given the currently available process information. Additionally, when trained and calibrated correctly, these well-adjusted probability values will also increase the accuracy of the predictions. The above-mentioned issues and characteristics make the method developed in this paper fundamentally different from single TU, observable and fixed quality criteria evaluation in semiconductor quality prediction research papers Jebril et al. (2022); Wang et al. (2022b); Heo et al. (2021).

3. RELATED WORK

There are mainly two families of approaches to PQP of multistage production lines. The first family leverages knowledge-based systems and the second family encompasses data-driven approaches. Data-driven approaches need only enough data and dispense of the domain knowledge, thus tend to be faster to develop and less error-prone. Md *et al.* provide a thorough analysis of both families (Md et al., 2022). Our work follows the data-driven approach for its benefits.

Given the limited quality and quantity of the available process and TU data from production lines, data-driven PQP can be challenging. In particular, the following issues have to be addressed according to the literature from both the data-driven PQP and process monitoring communities.

Unbalanced data: The available data from a production line is usually highly unbalanced, with either failed or successful products under-represented. At the beginning, there are many failed products, but their rate diminishes strongly with more experience and the quality monitoring of the production (Wang et al., 2022b). Consequently, PQP system must inherently deal with unbalanced data, either with under/oversampling, sample weighting, or other problem-specific solutions (Kovács, 2019). Moreover, accuracy measures which realistically reflect model performance are critical. Though these issues are well addressed in the machine learning community, they are still understudied in the domain of PQP (Melhem et al., 2017; Wang et al., 2019; Gu et al., 2023). We use different

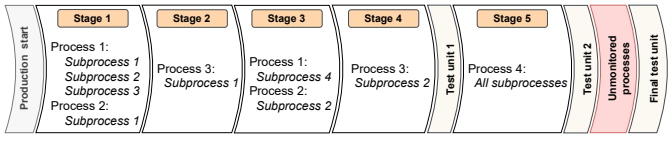


Fig. 2. Semiconductor production processes and TUs

sample weighting, among other parameters, in grid search to acquire the highest possible balanced accuracy.

Generalization and model interpretability: In the current Industry 4.0 settings, individualized batch productions are increasingly favored. This trend makes quality data scarce while simultaneously increasing its structural complexity. Data-driven models need to thus handle high-dimensional and small samples of quality information from the production line (Yu et al., 2021). One of the most effective ways to increase the generalization capability and stability of the trained models in this scenario is ensemble learning (Li et al., 2021). This method ensembles multiple models, called *base learners*, to boost performance. Moreover, interpretable ensemble models give insights on the importance of individual base learners (Dong et al., 2020). We embrace ensemble learning in our work to achieve both high accuracy and good interpretability.

Model reliability requirements: Serving as a consultant to the process engineer, the trained PQP model should both be accurate and indicate the confidence of the prediction (Silva Filho et al., 2023). Many learning algorithms extensively used in PQP, such as support vector machines (SVM) and boosted trees, push the predicted probability values away from 0 and 1. The prediction probabilities cluster closer to 0.5 which deteriorates their quality (Niculescu-Mizil and Caruana, 2005). With this shortcoming of the used boosted models in mind, we calibrate the PSP values coming from base learners and prevent the accuracy loss due to PSP binary quantization by stacking the PSP values.

Deployability of the model: In the similar vein, Yu *et al.* propose to stack different methods such as SVM, boosted tree, *etc.* to boost performance (Yu et al., 2021). Despite promising results in a fixed and stationary setting, such methods neglect the complexity of model maintenance in an ever-changing industrial setting (Huyen, 2022; Sculley et al., 2015). Instead of using different algorithms for training base learners, which is operationally difficult, we exploit the the fact that base learners trained for different stages use different subsets of the sensor readings from the production line. This allows us to train base learners in separation and greatly simplifies the maintenance in terms of machine learning operations (MLOps).

4. PRELIMINARIES

Model calibration: According to Dimitriadis et al. (2020) a binary probabilistic classifier $f : X \to [0, 1]$ with the instant space X and binary target space $Y = \{-, +\}$ is calibrated if

$$\forall s \in [0, 1] : P(Y = +|f(X) = s) = s \tag{1}$$

In numerous applications, especially safety-critical systems which employ machine learning, it is of great importance that the trained model expresses its true predictive uncertainty which raises the need for well calibrated models (Vaicenavicius et al., 2019). The reliability diagram serves as the main diagnostic tool to assess calibration. This diagram depicts the correlation between the observed event frequencies and the predictive probabilities. In the case of a well calibrated model, the aforementioned quantities must match and thus the plotted points in a reliability diagram are expected to lie close to the diagonal (Bröcker, 2008). In what follows, the best performing method used for model calibration in this paper is presented. For more details regarding other methods used in this paper please refer to Silva Filho et al. (2023).

Platt calibration: This method tries to calibrate a model by passing its posterior probabilities through a sigmoid function by minimizing the following cost function via gradient descent for a given set $(f(x_i), y_i)$

$$\underset{A,B}{argmin} \{-\sum_i y_i log(p_i) + (1 - y_i) log(1 - p_i)\} \tag{2}$$

where

$$p_i = \frac{1}{1 + exp(Af(x_i) + B)} \tag{3}$$

Binary estimated calibration error (Binary-ECE): In this paper, we have used the Binary-ECE as the performance measure for model calibration:

$$Binary\ ECE = \sum_{m=1}^{M} \frac{|B_m|}{N} |\bar{y}(B_m) - \bar{s}(B_m)| \tag{4}$$

where M denotes the number of bins in the reliability diagram, N is the number of data points, B_m represents the number of data points in each bin, and lastly $\bar{y}(B_m)$ and $\bar{s}(B_m)$ represent the proportion of positives and the average probability respectively.

5. PROPOSED METHOD

5.1 Data preparation

The process information along with the test results from different TUs in the production line, are stored separately in different tables of a database. As it can be seen in Fig. 2, many processes are revisited so that different subprocesses can be carried out on the currently produced batch. Thus, for generating input/output data for model training, the following steps have to be completed.

(1) **Separating subprocess data:** Based on factors such as part name, recipe name, *etc.*, the data from a process table is divided into different subprocess datasets.
(2) **Cleaning and annotating the subprocess data:** Given any arbitrary product, it is possible to find all the subprocess and, if available, test results using the product's ID. However, due to data storage, test and sensor reading errors, in some cases there will be missing information which has to be handle when the product quality model is deployed.
(3) **Subprocess data preprocessing:** Features which are categorical, *e.g.,* program name, recipe name, nuzzle ID, *etc.*, need to be encoded so that they can be used in input vector. In addition, the encoders must be stored so that they can be used in real-time as the production is in progress. Moreover, time stamps and delta time also need to be converted to integer or float values.

5.2 Model training

In this industrial project, for each stage of processes, separate models for predicting the test results for all

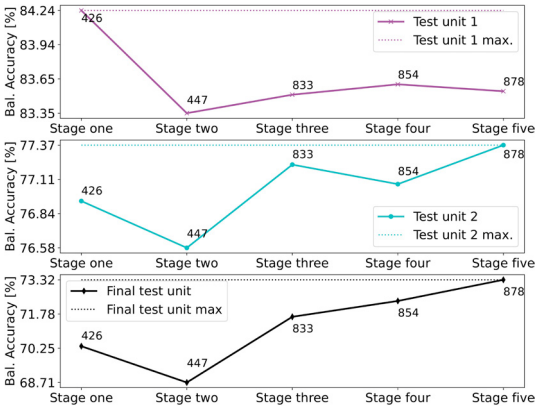


Fig. 3. Balanced accuracy of the trained base learners

three TUs are trained. Furthermore, given the additional challenges of unmonitored processes and the significant reduction of the number of available samples for training, from 2267 samples to only 756 samples, the development of the PQP models used for predicting the results from the last TU, depicted as green blocks in Fig. 1, are discussed extensively in subsections 5.4 and 5.5 of this paper. In short, in the current configuration, the PQP models for predicting the output of the first and second TU are merely gradient boosted trees trained separately for each stage. These models are also referred to as the base learners. For the final TU, in addition to the gradient boosted trees trained for each stage, these models are stacked and calibrated for enhancing the performance of the trained models and also to increase their reliability.

5.3 Base learners

In the conducted study, the base learners are gradient boosted trees, which are implemented using the XGBoost library (Chen and Guestrin, 2016). The potential changes in the production, *e.g.,* recipe changes, and also the need for a model suitably trained for a given manufacturing setting, raises the need for a generalizable solution which is perfectly tailored to the production line settings. Therefore, the hyperparameters of the base learners are acquired using grid search. In the current implementation, the hyperparameters *maximum tree depth*, *learning rate* and *number of estimators* are determined using grid search. Furthermore, to address the issue with imbalanced data in the production line and also promote the generalization capability of the base learners, the grid search also contains the hyperparameter *class weight*.

Moreover, to keep the ratio of the positive and negative classes almost constant during training and testing and also to avoid optimistic accuracy estimates, while splitting the available data from the production line, the method used for splitting the data is stratified. In addition, given the fact that the number of labeled samples in a production line is normally limited, in the proposed method, k-fold cross validation with is used to further increase the generalization power of the trained base learner. The balanced accuracy of the trained base learners is depicted in Fig. 3. The numbers next to each plotted point denote the number of features used in the predictor. As it can be seen in Fig. 3, the information attained during the second stage of assembly does not increase the PQP model prediction accuracy for any of the TUs. Furthermore, the accumulated information from stages three and four does not significantly increase the accuracy of the first PQP model. The second PQP model on the other hand shows a slight improvement in accuracy as more information is available from the production line compared to its accuracy when only trained with the data from the first stage. In addition, the results depicted in Fig. 3 clearly demonstrate the importance of accumulated stage data on the accuracy of predicting the final TU's output. Starting from the third stage, as more features are exposed to the base learners, the accuracy of the predictions is increased. This step-wise model training provides insight to the improved accuracy given the cost (number of features) while accumulating stage information which facilitates cost-effective PQP model development for multistage semiconductor production. Despite the high accuracy of the base learners (outperforming the similar work with the balanced accuracy of 71.15% and 64.56% in Melhem et al. (2017) and Tin et al. (2022) respectively), no assumption or prior knowledge from the production line is imposed on the training algorithm. In addition, the input vector to base learners in each step is also different which further promotes decorrelated bagging (James et al. (2013)).

5.4 Ensemble model

We tested several data fusion and model stacking strategies. The domain knowledge from the production line suggested that by fusing the first and second TU results in the input vector of the PQP model for the final TU, the accuracy should potentially increase; however, the results of the training rejected this hypothesis. For inspecting this issue, the Pearson correlation (Cohen et al. (2009)) between the TUs' results are calculated and a maximum correlation value of 0.091 between the first TU and Final TU is acquired. This denotes the fact that, given the information from an arbitrary TU, it is not possible to infer anything substantial for the other TUs in the production line. This issue can be blamed on the fact that these TUs inspect different key performance aspects of the produced semiconductors and thus their values are not correlated. Similarly, concatenating the predictions from the base learners for the different TUs to the input vector did not increase the accuracy of the ensemble model either.

As the last attempt, unlike conventional stacking approaches found in the related work, in this paper we use the predicted PSP of the last three base learners for predicting the results of the final TU, and map the generated vector from these PQP model to the final TU result. By employing the predicted PSP values from the PQP models, it is possible to prevent the potential information loss due to the quantization of the output to two possible outputs. This idea originates from knowledge distillation, where in the vanilla knowledge distillation case, the logits of a large deep model as the teacher are used to train a smaller network (Wang et al., 2022a). However, instead of training a smaller model and trying to make the distribution of the logits as close as possible, we merely use this so-called knowledge from different base learners to make better predictions. The results of the aforementioned model stacking resulted in a 0.18% increase in the balanced accuracy and a 19.49% reduction in binary-ECE. In addition, stack-

Table 1. SM parameters

Stage 3	Stage 4	Stage 5	Intercept
0.000	2.526	12.685	-7.630

Table 2. Balanced accuracy and binary-ECE

Model name	Balanced accuracy	Binary-ECE
Base learner (baseline)	73.32%	0.159
SM	73.50%	0.128
Isotonic - base learner	71.94%	0.161
Sigmoid - SM	**74.23%**	**0.128**
Binning - base leaner	68.41%	0.112
Binning - SM	70.67%	0.143

ing the base learners with LR addresses the issue of the boosted trees with the predicted probability values which are pushed away from 0 and 1 (Fig. 4(a) and Fig. 4(b)) resulting in a more reliable model than the base learners. The reliability test diagrams are generated from the test data which is not exposed to any steps of data cleaning, data preprocessing and model training. Fig. 4(a) clearly indicates the fact that the majority of the predicted PSP values, 112 data points, from a base learner are between 40% and 60% which reflect the model uncertainty while predicting the final product quality.

The coefficient of different PSP values and the intercept of the SM is depicted in Table 1. LR as one of the most interpretable models, provides insight into the most important PSP values and also how their combination helps predict the output of the final TU. Additionally, from MLOps point of view, our choice of LR simplifies ensemble model maintenance as a LR model is both straightforward to (re-)train and interpretable (Huyen (2022)).

5.5 Model calibration

As the last step of PQP model development, we aim to increase the reliability of the trained models via model calibration. We have used the Scikit-learn and PyCalib libraries (Pedregosa et al., 2011; Silva Filho et al., 2023) for model calibration. The reliability diagram of the best performing model can be seen in Fig. 4(c).

6. RESULTS

In this section, the balanced accuracy and the binary-ECE as the reliability criteria are compared for the developed PQP models. As can be seen in Table. 2, the proposed model stacking significantly increases the reliability of the base learners as the binary-ECE is reduced by 19.49%. In addition compared to the best base learner as the baseline, there is a 0.91% increase in the balanced accuracy of the model. The model with the lowest binary-ECE is the binning calibrated base learner. However, given the fact that the training data used for training the PQP models are eventually also used for calibrating the model, there is the unavoidable introduced bias in the model training (Niculescu-Mizil and Caruana, 2005). In case that the training data points were not as limited as in the current study, 756 samples each containing 878 features, a subset of the training data points should be set aside for model calibration to avoid performance drop.

7. DISCUSSION

Interpretability and reliability of data-driven models play a significant role in their acceptance as a solution in decision-making-related tasks. As in the studied production line, many industrial use cases suffer from the inherent uncertainty induced by factors such as raw material vendors, and also the lack of knowledge about unmonitored processes in their production line. In these stochastic and complex settings, it is vital to train models which clearly pinpoint the most important predictors in their decisions and also indicate how confident they are about their predictions. The model stacking from PSP values and the post hoc calibration presented in this paper aimed to shed light on the importance of reliability tests for data-driven models used in critical and cost sensitive multistage production settings. During development of the proposed method, we focused on having all the steps from data preprocessing up to model training as modular and generalizable as possible given the potential changes in the production lines which include, but are not limited to, updates in the recipe and optimization of production procedure. In addition, we did not incorporate any domain knowledge from semiconductor industry in model training and merely aimed to account for high stochasticity of the entire production.

8. CONCLUSION

The results of this study demonstrate a significant reduction in the binary-ECE by 19.49% with the side benefit of 0.91% increase in the balanced accuracy of the calibrated SM. We also demonstrate the steps needed to develop impactful PQP models for settings with limited and unbalanced training samples such that they can be easily reproduced in practice to other settings. In the future, we want to focus more on the maintenance of the developed PQP models. In particular, we plan to inspect the impact of data distribution shifts on the confidence levels of the PQP models. Moreover, we aim to develop a module for triggering a retraining process for either the base learner or the SM, or to trigger a recalibration of the PQP models. This would finally close the MLOps loop for effective data-driven solutions.

REFERENCES

Arif, F., Suryana, N., and Hussin, B. (2013). A data mining approach for developing quality prediction model in multi-stage manufacturing. *International Journal of Computer Applications*, 69(22).

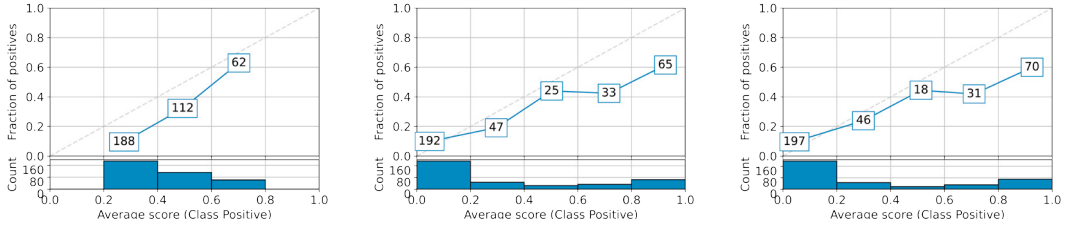
Bai, C., Dallasega, P., Orzes, G., and Sarkis, J. (2020). Industry 4.0 technologies assessment: A sustainability perspective. *International journal of production economics*, 229, 107776.

Bröcker, J. (2008). Some remarks on the reliability of categorical probability forecasts. *Monthly weather review*, 136(11), 4488–4502.

Chen, T. and Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, 785–794.

Cohen, I., Huang, Y., Chen, J., Benesty, J., Benesty, J., Chen, J., Huang, Y., and Cohen, I. (2009). Pearson correlation coefficient. *Noise reduction in speech processing*, 1–4.

Dimitriadis, T., Gneiting, T., and Jordan, A.I. (2020). Evaluating probabilistic classifiers: Reliability diagrams



(a) Base learner (b) SM (c) Sigmoid-calibrated SM

Fig. 4. Reliability diagrams of different models. The numbers in the boxes represent the number of samples in each bin.

and score decompositions revisited. *arXiv preprint arXiv:2008.03033*.

Dong, X., Yu, Z., Cao, W., Shi, Y., and Ma, Q. (2020). A survey on ensemble learning. front comput sci 14: 241–258.

Gu, J., Zhao, L., Yue, X., Arshad, N.I., and Mohamad, U.H. (2023). Multistage quality control in manufacturing process using blockchain with machine learning technique. *Information Processing & Management*, 60(4), 103341.

Heo, T., Kim, Y., and Kim, C.O. (2021). A modified lasso model for yield analysis considering the interaction effect in a multistage manufacturing line. *IEEE Transactions on Semiconductor Manufacturing*, 35(1), 32–39.

Huyen, C. (2022). *Designing machine learning systems.* " O'Reilly Media, Inc.".

James, G., Witten, D., Hastie, T., Tibshirani, R., et al. (2013). *An introduction to statistical learning*, volume 112. Springer.

Jebril, H.T., Pleschberger, M., and Susto, G.A. (2022). An autoencoder-based approach for fault detection in multistage manufacturing: a sputter deposition and rapid thermal processing case study. *IEEE Transactions on Semiconductor Manufacturing*, 35(2), 166–173.

Kovács, G. (2019). An empirical comparison and evaluation of minority oversampling techniques on a large number of imbalanced datasets. *Applied Soft Computing*, 83, 105662.

Li, Z., Chen, X., Wu, L., Ahmed, A.S., Wang, T., Zhang, Y., Li, H., Li, Z., Xu, Y., and Tong, Y. (2021). Error analysis of air-core coil current transformer based on stacking model fusion. *Energies*, 14(7), 1912.

Lieber, D., Stolpe, M., Konrad, B., Deuse, J., and Morik, K. (2013). Quality prediction in interlinked manufacturing processes based on supervised & unsupervised machine learning. *Procedia Cirp*, 7, 193–198.

Md, A.Q., Jha, K., Haneef, S., Sivaraman, A.K., and Tee, K.F. (2022). A review on data-driven quality prediction in the production process with machine learning for industry 4.0. *Processes*, 10(10), 1966.

Melhem, M., Ananou, B., Ouladsine, M., Combal, M., and Pinaton, J. (2017). Product quality prediction using alarm data: Application to the semiconductor manufacturing process. In *2017 25th Mediterranean Conference on Control and Automation (MED)*, 1332–1338. IEEE.

Niculescu-Mizil, A. and Caruana, R. (2005). Predicting good probabilities with supervised learning. In *Proceedings of the 22nd international conference on Machine learning*, 625–632.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830.

Schulze Struchtrup, A., Kvaktun, D., and Schiffers, R. (2020). A holistic approach to part quality prediction in injection molding based on machine learning. In *Advances in Polymer Processing 2020: Proceedings of the International Symposium on Plastics Technology*, 137–149. Springer.

Sculley, D., Holt, G., Golovin, D., Davydov, E., Phillips, T., Ebner, D., Chaudhary, V., Young, M., Crespo, J.F., and Dennison, D. (2015). Hidden technical debt in machine learning systems. *Advances in neural information processing systems*, 28.

Silva Filho, T., Song, H., Perello-Nieto, M., Santos-Rodriguez, R., Kull, M., and Flach, P. (2023). Classifier calibration: a survey on how to assess and improve predicted class probabilities. *Machine Learning*, 1–50.

Tin, T.C., Tan, S.C., and Lee, C.K. (2022). Virtual metrology in semiconductor fabrication foundry using deep learning neural networks. *IEEE Access*, 10, 81960–81973.

Vaicenavicius, J., Widmann, D., Andersson, C., Lindsten, F., Roll, J., and Schön, T. (2019). Evaluating model calibration in classification. In *The 22nd International Conference on Artificial Intelligence and Statistics*, 3459–3467. PMLR.

Wang, C., Zhang, S., Song, S., and Huang, G. (2022a). Learn from the past: Experience ensemble knowledge distillation. In *2022 26th International Conference on Pattern Recognition (ICPR)*, 4736–4743. IEEE.

Wang, G., Ledwoch, A., Hasani, R.M., Grosu, R., and Brintrup, A. (2019). A generative neural network model for the quality prediction of work in progress products. *Applied Soft Computing*, 85, 105683.

Wang, H.Y., Tsung, C.K., Hung, C.H., and Chen, C.H. (2022b). Designing the rule classification with oversampling approach with high accuracy for imbalanced data in semiconductor production lines. *Multimedia Tools and Applications*, 81(25), 36437–36452.

Yu, J., Pan, R., and Zhao, Y. (2021). High-dimensional, small-sample product quality prediction method based on mic-stacking ensemble learning. *Applied Sciences*, 12(1), 23.