Proceedings of the 33rd European Safety and Reliability Conference (ESREL 2023)
Edited by Mário P. Brito, Terje Aven, Piero Baraldi, Marko Čepin and Enrico Zio
©2023 ESREL2023 Organizers. *Published by* Research Publishing, Singapore.
doi: 10.3850/978-981-18-8071-1_P620-cd

Process Data Analysis for Improved Burn-In Strategies Based on Complementary AI Models

Dr. Lars Langenberg

Pumacy Technologies AG, Germany. E-mail: lars.langenberg@pumacy.de

Rainer Pätzold

Pumacy Technologies AG, Germany. E-mail: rainer.paetzold@pumacy.de

Asad Khalid

Pumacy Technologies AG, Germany. E-mail: asad.khalid@pumacy.de

Burn-in (BI) tests are the industry standard to screen out the early life failures of semiconductors. Advanced sampling and test strategies allow to reduce BI times or sample size without affecting the defined quality targets. A new BI approach introduces a lot-specific health factor h that correlates to the probability of early failures. In our approach, the health factor of a specific wafer lot is derived from the Advanced Process Control (APC) system that logs all meta and logistics data of the production process, but not the raw sensor data. Complementary AI models were investigated to provide health indicators with a high correlation to early failures.
A practical study has been performed based on real APC data and several known BI defects. Due to the amount of data, big data strategies had to be applied and tested to reduce the computational demands. Our investigation shows that a combination of a binary classifier and an LSTM autoencoder model allow for a good assessment of the health factor. In addition, the autoencoder allows to identify and visualize potential issues of the production process via its loss function. That enables process engineers to assess and to investigate potential risks and issues of a specific lot.

Keywords: Semiconductor Devices, Burn-In Tests, Early Life Failure Rate, Advanced Process Control, Deep Neural Networks.

1. Introduction

Safety-critical applications of semiconductor devices, such as in automotive, aerospace or medical systems, require special measures in the production process to ensure that the demanding reliability targets are met. Typically, electronic devices exhibit an increased failure rate at the beginning of their lifetime (early life failures).

Burn-In (BI) is a widely used engineering method to identify faulty or weak items from a standard production (Block, Savits, 1997) and thereby to screen out early life failures. In BI tests, either 100% of the produced devices (100% BI) or random samples (BI study) are operated under accelerated stress conditions, such as increased temperature or voltage. Devices that do not meet the assured properties are considered as BI failures and removed from production (Kurz et al, 2018).

1.1. Advanced BI concepts

100% BI leads to long testing times and high costs. Therefore, advanced sampling and test concepts have been developed to reduce BI times and/or sample size without affecting the defined quality targets. A common approach is to evaluate the failure probability p in the early life of the devices in a BI study. Classically, this is done by computing the exact Clopper–Pearson upper bound for p (Kurz et al, 2018).

Advanced BI concepts take advantage of flexible sampling plans or previous BI studies for follower products (Kurz et al, 2021). For further burn-in improvements, recent approaches have investigated specific process data, such as the BI test measurements, to predict the number of defective units and their early failure probability (Baraldi et al, 2021).

1.2. Objective

While current BI concepts are mostly based on the overall production process and its observed failure rate, this paper proposes a new approach that analyzes the likelihood of early life failures of individual wafer lots within a given production process. A lot-specific health indicator *h* is introduced that correlates to the likelihood of early life failures depending on the individual production history of the wafer lot.

Thereby, the health indicator allows to reduce BI efforts: Wafer lots with an average or low health indicator need the full test regimen, as determined by the chosen BI concept. For wafer lots with an above average health indicator, however, BI sample size or BI time can be reduced without affecting the quality target.

Due to the low amounts of BI failures in the range of a few ppm, it is usually not possible to correlate BI failures to a single process step or a single root cause and, therefore, the health indicator must consider the whole production process.

2. Approach

For the efficient processing and reduction of the available process data into a single health indicator, a data chain was created that combines big data with machine learning algorithms:

(i) Acquisition of the APC data
(ii) APC data reduction
(iii) Machine learning using selected APC data sets (training data)
(iv) Combination and cross-validation of the AI models (by using test data)

2.1. *APC data acquisition*

Since many years, Advanced Process Control (APC) has become state of the art in the semiconductor industry for monitoring and controlling the very complex process flows and manufacturing parameters.

Usually built on top of the machine control systems, APC systems are deeply embedded in the semiconductor manufacturing process. APC comprises a set of control methods, such as Statistical Process Control (SPC), Fault Detection and Classification (FDC), Run-to-Run (R2R) control and Virtual Metrology (VM), see Moyne et al, 2016 and Schellenberger et al, 2010.

While originally intended for process control and automation with regard to systematic variations, current APC systems collect a vast amount of process data that document the production process of a wafer lot comprehensively.

The APC data does not include any sensor raw data or image data from visual inspections which are typically used for AI-based wafer analysis. If deviations from the standard process occur, however, measures are taken and thereby become visible in the APC data e.g., as error events, additional inspections, or extended handling times.

Therefore, the APC data can be seen as high-level logistics data or meta data that aggregates the underlying raw data.

Due to the amount of data, Big Data technologies are required to store and process the APC data in real-time, e.g., data lakes built with the Hadoop Distributed Filing Systems (Moyne et al, 2016).

2.2. *APC data reduction*

The amount of APC data may easily reach several gigabytes per wafer lot. As data sets from multiple wafers were required to train and to validate the AI models, it was crucial to further optimize the training data. Therefore, a set of data reduction techniques was combined into a data reduction pipeline, see Fig 1:

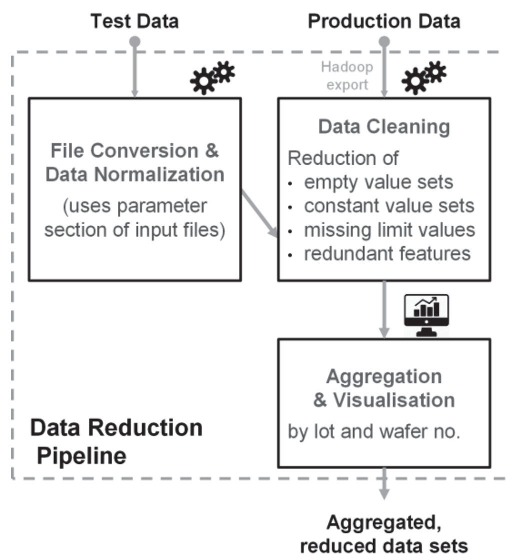

Fig 1. Schematic diagram of the data acquisition and data reduction steps.

- Basic data cleaning: remove empty or irrelevant data and features, such as process filenames.
- Label encoding: encode non-numeric features to numeric values.
- Feature reduction: remove features that have constant values or that correlate with other features.
- Sampling: use every n^{th} sample only.

2.3. *AI models*

In order to compare and to validate the AI models for APC data analysis, two different AI models were selected and trained.

- LSTM autoencoder and
- Binary classification.

2.3.1. *LSTM autoencoder model*

As semiconductor production has already reached a high-quality level, very low defect rates in the order of a few ppm are expected. Therefore, only few data sets from BI failures are available which leads to highly unbalanced data. Autoencoder models in general provide the advantage that they can be trained on the “good” data of passed lots, whereas the few data sets of failed lots can be reserved to validate the AI model.

A Long Short-Term Memory (LSTM) autoencoder is a type of neural network that can be used for anomaly detection in sequential data (Nguyen et al, 2021). It consists of two parts: the encoder and the decoder. The encoder maps the input sequence to a lower-dimensional space, while the decoder maps the lower-dimensional representation back to the original sequence. The autoencoder is trained to minimize the reconstruction error, which is the difference between the original and reconstructed sequence.

The loss function used for training the LSTM autoencoder is typically the mean squared error (MSE) between the input sequence and the reconstructed sequence.

The aim is to achieve a low reconstruction error/loss for the APC data of the wafer lots that passed BI and that serve as the training data for the model, and to determine a threshold for anomaly detection. Subsequently, the model is tested on the failed lots - the ones that exhibited faults during BI. If the APC data of the failed lots deviates from that of the passed lots, the model should generate a higher reconstruction error, generating anomalies.

Then, either the number of anomalies or a 90%-quantile of the loss function can be used as the lot-specific health indicator *h*.

2.3.2. *Binary classifier model*

The Binary Classifier is a type of machine learning model that can be used to classify data into two classes: positive and negative (Kiranyaz et al, 2021). In the context of the BI process, the binary classifier is used to differentiate between lots that fail the BI test and lots that pass the BI. The output of a binary classifier *p0* is a probability value, which represents the likelihood of the input belonging to one of the two classes.

In the context of APC data analysis, the probability *p0* that a given lot is classified as “passed” can now be used as its health indicator *h*.

The binary classifier can be evaluated using a confusion matrix, which summarizes the number of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). From the confusion matrix, various metrics can be calculated, such as the accuracy, precision, recall, and F1 Score, which provide insights into the performance of the classifier.

The F1 Score, the harmonic mean of precision and recall, is the commonly used metric for binary classification problems, as it considers both the precision and recall of the classifier. A higher F1 Score indicates better performance of the classifier:

$$F1_{score} = 2 \times \frac{precision \times recall}{precision + recall} \quad (1)$$

where,

$$precision = \frac{TP}{TP+FP} \quad (2)$$

$$recall = \frac{TP}{TP+FN} \quad (3)$$

In this paper, F1 scores are analyzed individually for passed and failed lots.

2.4. *Combined models*

By leveraging the strengths of each model discussed, a more nuanced understanding of the data can be achieved, and specific areas of interest come into focus.

The binary classifier provides a macroscopic view using the time analysis data, identifying whether a given lot is likely to pass or create

failures in the BI test. However, it does not provide detailed information about which specific aspects of the process are contributing to these outcomes. In addition, it requires balanced data sets of passed and failed lots which may be difficult to obtain in production processes with low failure rates.

This is where the LSTM autoencoder comes in - it provides a more microscopic view of the data by identifying specific patterns and anomalies within each time series.

3. Case Study

The case study is based on real APC data that was logged during a power semiconductor production process and retrieved from the manufacturer's data lake.

3.1. *APC data analysis*

In total, the investigated APC data covers a period of 22 months in which several BI failures occurred. Full data sets from 10 lots where defects were identified in BI tests (failed lots) and from 52 lots without known defects (passed lots) were collected to train and test the AI models.

The resulting 62 datasets consist of a time series of data samples that merge the APC data from all processing steps per selected wafer lot into one data file. The number of data samples per wafer lot is highly variable, with an average of about 15 million and a maximum of 26 million data sets (Fig 2). Each sample contains 52 features, such as the involved machines, carriers and actions, key process parameters, and the detected deviations.

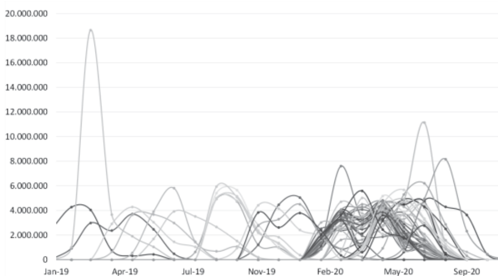

Fig 2. Time-series distribution of the APC datasets per lot and month.

After the data reduction steps described in 2.2, the data could be reduced to 20 independent features.

3.1. *Application of the LSTM autoencoder*

At first, an LSTM-based autoencoder model was selected and trained on the passed wafer lots. In addition, various strategies were tried to reduce the very large amount of training data and long training duration, such as data sampling and lot reduction. The accuracy of the autoencoder models was assessed by their loss functions, see the example shown in Fig 3.

This model configuration proved a slight difference between passed and failed wafer lots: For example, using 5 lots for the AI training and sampling every 5th data sample, the mean absolute error of the loss function for the failed lots increased by 26% compared to the passed lots (157 vs. 124).

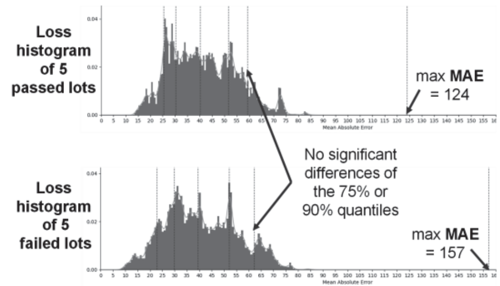

Fig 3. Mean average error distribution of the loss functions of 5 passed and 5 failed wafer lots, shown as histograms.

Still, the differences between passed and failed wafer lots remained small, and they were highly dependent on the sample size and the selected training datasets. In addition, it proved difficult to map the loss function to a single, meaningful health indicator h. Therefore, a different AI model was investigated.

3.2. *Application of the Binary Classifier*

To gain a deeper understanding of the APC data, we conducted an in-depth time analysis: While some features contained numerical values with a high variability, other features contained non-numeric values that stayed constant for several time steps. To address this, we calculated the duration for each unique value available for each feature by determining the difference between the first timestamp where the unique value was observed and the last timestamp (Fig 4). By performing this time analysis, we were able to

gain more insights into the duration of various events and patterns.

N unique values in feature x

duration	feature x value	number of samples
0.35	0	1241
112.07	1	776500
...	...	...

Fig 4. Data reduction based on feature duration.

In addition, a visualization tool helped to quickly identify those feature durations that differed significantly between passed and failed lots (Fig 5). Similar results could have been achieved by training and evaluating ML models on single features. But this approach would have required several weeks of processing time.

Fig 5. Visualization of features and their duration with regard to passed (“good”) and failed (“bad”) wafer lots.

The visual pre-selection allowed narrowing down the feature set to focus on the most significant features for AI training and evaluation. Distinct sub-sets of the APC data were created per feature that contained only the duration and the values.

For the binary classification model, a deep neural network with convolutional layers and a fully connected output layer was trained on each selected feature. As a result, a handful of features were separated, which provided significantly better results than others.

The resulting model confirms the results of the LSTM autoencoder, but with a much higher accuracy for passed lots, as shown with F1 scores of >90%, depending on the selected features, see Table 1. However, the F1 scores for failed lots were relatively low leading to further investigations how the time-analysis data could be restructured.

Table 1. Best F1 scores for AI models trained on a single feature.

Feature #	F1 score of failed lots	F1 score of passed lots
'Feat_1'	0,74	0,95
'Feat_2'	0,69	0,94
'Feat_3'	0,67	0,95
'Feat_4'	0,33	0,91
'Feat_5'	0,28	0,92
'Feat_6'	0,27	0,91
'Feat_7'	0,22	0,9

3.3. *Further data refinement and SHAP Analysis*

To further improve the prediction of failed lots, we investigated how combinations of multiple features would affect the accuracy of the AI. The data was restructured in such a way that each row would then denote the full sample for a single lot while the columns denote the count of unique values for the feature(s) Therefore, combining multiple features meant an increased number of columns.

M unique values in feature x

N LOTs

	x_ $128300	x_ $143910	x_ GB016	...
1E915233	1250050.0	403443490.0	2927910.0	...
1E029098	NaN	NaN	3091010.0	...
...	...	...	...	...

Fig 6. Restructured time-analysis data.

After re-training the binary classifier on this newly structured data, the accuracy could be improved for some, but not for all features, see Table 2. The results showed that the single feature ‘Feat_1’ provided perfect prediction accuracy - at least with the available test data - while combinations of different features did not improve the F1 scores compared with the single-feature approach.

Table 2. Best F1 scores for AI models trained on restructured time-analysis data.

Feature or combination	F1 score of failed lots	F1 score of passed lots
'Feat 1'	1	1
'Feat_1+Feat_2'	0,5	0,92
'Feat_4'	0	0,85
'Feat_1+Feat_2+Feat_2'	0	0,89
'Feat_2'	0	0,89

To further enhance the performance and to reduce model complexity, a SHAP analysis was applied on the model results.

Shapley additive explanations (SHAP) analysis is a method of interpreting the predictions of a machine learning model by assigning importance values to each input value of a feature to the model (Nohara et al, 2019). This method uses the concept of cooperative game theory to calculate how much each feature contributes to the prediction of the model.

In Shapley explanations, each feature value is considered as a player in a game where the prediction is the payout. The contribution of a feature value is calculated as the difference in the prediction when the feature is included compared to when it is excluded. This process is repeated for all possible combinations of features, and the contributions are averaged across all permutations to obtain the Shapley value of each feature. Shapley explanations can be used to identify important values, detect interactions, and assess the robustness of a model.

In the case of time analysis dataset, SHAP analysis was applied to the restructured time-analysis data, and it allowed to identify those feature values that contribute most to the model's predictive power.

As shown in Fig 7, the number of unique values, and consequently the amount of data, was reduced by a factor of 3x while maintaining a high level of model accuracy. Since SHAP analysis provides importance scores for each unique value, it also allows refining the APC data to retrain the LSTM autoencoder and to focus on a smaller subset of the input features.

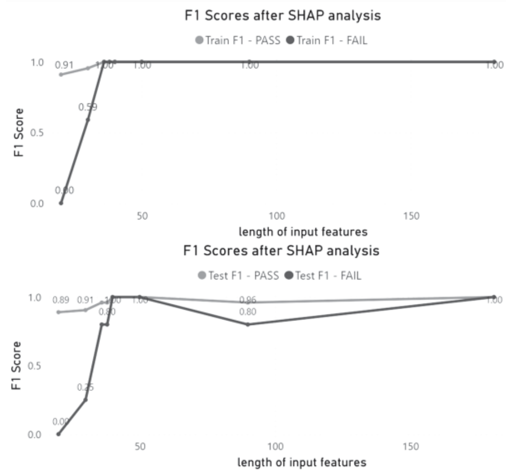

Fig 7. Comparison of F1 scores after SHAP analysis (top curve: training data / bottom curve: test data).

4. Results

With only basic data cleaning, the initial training runs of the LSTM autoencoder took up to 72 hours on an AI workstation. Therefore, the proposed data reduction pipeline was deployed, and training times could be reduced to a few hours for both autoencoder and binary classifier models. This allows to re-train the models on a regular basis if, for example, minor process changes or new types of failure occur.

With five selected data sets and using every fifth data sample, an LSTM autoencoder model was trained that detects differences between passed and failed lots, as seen in the maximum MAE of their loss functions. Still, the loss functions of both classes cannot be considered significantly different.

On the other hand, several AI models based on a binary classifier were developed that assess a wafer lot's health with a high accuracy, as shown by the corresponding high F1 scores. But the successful application of the binary classifier depends on the availability of training data from failed/passed wafer lots. Our test case provided data sets which contained enough BI failures to create these balanced data sets.

In addition, the binary classifier models directly compute the health indicator *h* per lot (as the probability *p0* that the lot belongs to the class "passed lots").

The very high F1 score of a single feature, however, could be an artifact resulting from the limited amount of "failed" data sets. Therefore, it might be better to use a combination of several features instead.

5. Conclusion and Outlook

Both investigated AI models, the LSTM autoencoder and the binary classifier, prove our initial assumption that issues or deviations in the production process are visible in its high-level APC data. Thus, the lot-specific APC data sets can be re-used to calculate the lot's health indicator *h*. For example, the BI can be reduced in terms of BI time and/or sample size if *h* is above average for a given production process.

In addition, the AI-based analysis of APC data has the advantage that it does not rely on physical process models or in-depth knowledge about the machines which is typically required to analyze low-level sensor data.

LSTM autoencoder models are the preferred tool if zero or very few APC data sets with known BI failures are available. If balanced data sets of passed/failed lots can be built, additional binary classifier models can be trained and optimized. These models can predict the health of a given wafer lot with a high accuracy. The output of the binary classifier *p0* can be used directly as the lot's health indicator.

Beyond the reduction of BI time, the combined AI approach generates further insights which cannot achieved by a single AI model.

Although the accuracy of the LSTM autoencoder is lower than the accuracy of the binary classifier, the LSTM autoencoder may help to visualize the results of the binary classifier: All deviations from "good" APC data sets, as represented by the training lots, appear as spikes (anomalies) in the loss function. An example is shown in the upper part of Fig 8.

In the time diagram, the timestamps of these anomalies can be assigned to a specific process step, machine, or activity (see Fig 8, lower part).

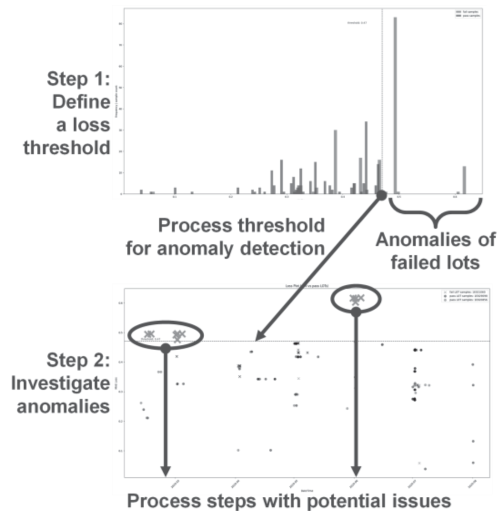

Fig 8. Histogram (top diagram) and time series analysis (bottom diagram) of the LSTM Autoencoder's loss function

Process engineers may use this visual tool to identify and to investigate the critical events that are the potential root causes of a low health indicator. Thereby, the autoencoder model is able to provide the missing explanation of a good or bad health assessment which the binary classifier model alone cannot produce.

Therefore, further efforts are ongoing to improve the LSTM autoencoder model and its anomaly detection based on our positive experience with the time-series data restructuring and SHAP analysis.

In addition, further research is required to enhance the existing BI reduction strategies with regard to the lot-specific health indicator.

Acknowledgement

iRel40 is a European co-funded innovation project that has been granted by the ECSEL Joint Undertaking (JU) under grant agreement No 876659. The funding of the project comes from the Horizon 2020 research programme and participating countries. National funding is provided by Germany, including the Free States of Saxony and Thuringia, Austria, Belgium, Finland, France, Italy, the Netherlands, Slovakia, Spain, Sweden, and Turkey.

References

Baraldi, P.,Medici S., Ahmed I., Zio, Lewitschnig, H. (2021). A Method based on Gaussian Process Regression for Modelling Burn-in of Semiconductor Devices. In: Leva, M., Patelli, E. Podofillini, L., Wilson, S. (eds), *Proceedings of the 31st European Safety and Reliability Conference (ESREL 2021)*, Singapore Research Publishing. https://doi.org/10.3850/978-981-18-2016-8_763-cd

Block, H., Savits, T. (1997) Burn-In. *Statistical Science Vol. 12, No. 1*, pp 1-13.

Kiranyaz, S., Avci, O., Abdeljaber, O., Ince, T., Gabbouj, M., & Inman, D. J. (2021). 1D convolutional neural networks and applications: A survey. Mechanical systems and signal processing, 151, 107398.

Kurz, D., Lewitschnig, H., Pilz, J. (2018). An Overview on Recent Advances in Statistical Burn-In Modeling for Semiconductor Devices. In: Pilz, J., Rasch, D., Melas, V., Moder, K. (eds) *Statistics and Simulation. IWS 2015. Springer Proceedings in Mathematics & Statistics, vol 231*, pp.371-380. Springer, Cham. https://doi.org/10.1007/978-3-319-76035-3_26

Kurz D., Lewitschnig H., Pilz J. (2021). Flexible time reduction method for burn-in of high-quality products. *Quality and Reliability Engineering International Vol.37, Issue6*, pp.2900-2915. https://doi.org/10.1002/qre.2896

Nguyen, H. D., Tran, K. P., Thomassey, S., & Hamad, M. (2021). Forecasting and Anomaly Detection approaches using LSTM and LSTM Autoencoder techniques with the applications in supply chain management. International Journal of Information Management, 57, 102282.

Nohara, Y., Matsumoto, K., Soejima, H., & Nakashima, N. (2019, September). Explanation of machine learning models using improved shapley additive explanation. In Proceedings of the 10th ACM international conference on bioinformatics, computational biology and health informatics (pp. 546-546).

Moyne, J., Samantaray, J., Armacost, M (2016) Big Data Capabilities Applied to Semiconductor Manufacturing Advanced Process Control. In: *IEEE Transactions on Semiconductor Manufacturing, vol. 29, no. 4*, pp. 283-291. doi: 10.1109/TSM.2016.2574130.

Schellenberger, M., Roeder, G., Öchsner, R., Schöpka, U., Kasko, I. (2010) Advanced process control – lessons learned from semiconductor manufacturing. In: *Photovoltaics International, 2010, Nr.9*, S.79-87. Retrieved from https://www.dr-production.de/content/dam/iisb2014/en/Documents/dr-production/2010_IISB_Schellenberger_PI.pdf