Advances in Reliability, Safety and Security, Part 9 – Kołowrocki, Dąbrowska (eds.)
© 2024 Polish Safety and Reliability Association, Gdynia, ISBN 978-83-68136-21-0 (printed), ISBN 978-83-68136-08-1 (electronic)

Advances in Reliability, Safety and Security

ESREL 2024
Monograph Book Series

Multibranch Neural Network For Predicting Production Lot Quality In Semiconductor Industry

Joaquín Figueroa[a], Ibrahim Ahmed[a], Piero Baraldi[a], Enrico Zio[a,b]

[a]*Department of Energy, Politecnico di Milano, Milan, Italy*
[b]*MINES Paris-PSL, CRC, Sophia Antipolis, France*

Abstract

In semiconductor manufacturing, burn-in (BI) testing is expensive and time consuming, especially for new technologies. This work considers the possibility of using multimodal data collected during production for estimating the quality of the production process. The objective is to identify high quality lots for which the number of BI tests to be performed can be reduced, while keeping the early-life failure probability (ELFP) in line with user requirements. The developed method is based on a multibranch neural network, which receives in input signal measurements from production machines and wafer map images, and predicts the expected number of BI-relevant failures: if this number exceeds a preset threshold, all devices of the lot will be BI-tested ("Full Burn-in" – FB lot), otherwise the number of BI-tests can be reduced ("Reduced Burn-in" – RB lot) given the high quality of the lot. The proposed method has been validated considering a synthetic dataset that emulates real conditions in the semiconductor industry, where some FB lots cannot be identified using only one of the two sources of available data. Furthermore, we test the method considering different proportions between FB and RB data to mimic the condition of imbalanced data expected in real production processes. The results show the superiority of the proposed method compared to two other methods that use only a single source of data.

Keywords: semiconductor, production quality, burn-in testing, deep learning, imbalanced data

1. Introduction

In semiconductor manufacturing, burn-in (BI) tests are performed by exposing products to high stress conditions, such as high temperatures and voltages, to estimate the product early life failure probability (ELFP) and eliminate defective products. BI testing of semiconductor products is particularly important for their use in safety-critical industries, such as aerospace and automotive, which demand high reliability standards. However, BI tests are expensive and time consuming to perform (Suhir, 2019), particularly for new technologies, for which a 100% BI strategy is typically adopted (Kurz et al., 2021). Thus, efforts are being directed to safely reduce the number of tests to be performed on high-quality lots. In (Ahmed et al., 2023), a method based on probabilistic support vector regression (PSVR) is proposed to estimate the number of BI relevant failures using signals collected during the manufacturing process.

In this paper, we consider the possibility of exploiting multimodal data collected from multiple sources to estimate the quality of the production process. Indeed, failures can be originated by defects caused by abnormal conditions and malfunctions occurring in different stages of the production, during which different data are collected. For instance, some defects related to BI-relevant failures may originate during plasma etching and be identified by analyzing wafer maps, others may occur while performing plasma deposition and be they are not detected analyzing other signals.

In our work, we consider two sources of data collected during different stages of the production process: 1) process signals which contain information about the operation, and, possibly, malfunctioning of production machineries and 2) data from the results of tests performed on dies of wafers, which are collected in the form of images (typically referred to as wafer maps) and contain information about the quality of the products at a given stage of the production process.

The problem of developing data-driven methods to handle multimodal data has been addressed using a variety of approaches (Aria et al., 2020; Cao et al., 2021; Cho et al., 2023; Kim et al., 2021; Yang et al., 2021). In (Cao et al., 2021), satellite images and numerical data about climate and soil properties are used to build linear regression, random forest (RF) and long short-term memory network (LSTM) models for estimating rice crop yield. In (Aria et al., 2020), the authors use acoustic emission signals and microscopic images for estimating the damage size and remaining useful life (RUL) of degraded structures; the proposed method is based on LSTMs and Fully Convolutional DenseNets (FCDNs). In (Yang et al., 2021), images, numerical data and textual documents are used to develop a deep neural network for the prediction of the degradation level of a system of a nuclear power plant. In (Kim et al., 2021), process parameter values and time-series collected during the production of car windshield side molding are used to develop a multimodal neural network model for detecting faulty products. In (Cho et al., 2023), wafer maps and tabular data from wafer and package tests are used to build a multimodal neural network model for predicting the result of module tests at chip level. This latter work differs from our work in two ways: 1) the prediction of the quality of a single chip and not of a lot of chips; and 2) the prediction of the quality before module tests and not before burn-in tests.

In this work, we develop a multibranch neural network. One branch uses convolutional layers to process the images, whereas the other branch uses fully connected layers to process the signals. The output of the model is the number of BI-relevant failures in the lot, from which the early life failure probability (ELFP) is computed by applying the Clopper-Pearson (CP) estimator (Clopper and Pearson, 1934). If the predicted ELFP is smaller than a preset threshold, the lot will follow a "Reduced Burn-In" (RB) policy. Otherwise, it will follow a "Full Burn-In" (FB) policy.

Simulated data that mimic the behavior of the semiconductor production process are considered to verify the performance of the proposed method. Specifically, since early life failures can be caused at different production stages, we assume that some BI-relevant failures are not detectable using only one source of data. The capability of the method of dealing with unbalanced datasets has also been analyzed considering two different scenarios with different proportions of FB lots, ranging from a balanced to a highly imbalanced condition.

The remainder of this work is organized as follows. In Section 2, the problem formulation is given. In Section 3, the proposed method is presented. In Section 4, the case study is described and in Section 5 the obtained results are discussed. Finally, Section 6 discusses the conclusions and final remarks on the work done.

2. Problem formulation

A dataset $D = \{S(l), W(l), n(l), y(l)\}_{l=1,\dots,L}$ collected during semiconductor production is available, containing two modalities of data collected during production ($S(l)$ and $W(l)$), the number of BI-tested devices of the lot ($n(l)$) and the number of BI-relevant failures in the tested devices ($y(l)$). Specifically, $S(l) \in \mathbb{R}^m$ is a vector of electrical signals measured performing electrical tests on the l-th lots before BI and $W(l)$ is an image aggregating the wafer maps of the lot.

The objective of this work is to develop a method for predicting the number of BI-relevant failures ($\hat{y}^{test}$) for a lot for which the vector S^{test} of measured signals and the image W^{test} have been collected during production. Let n^{test} be the number of BI tests that will be performed. Once $\hat{y}^{test}$ has been predicted, the Clopper-Pearson (CP) estimator is used for the estimation of the ELFP of the lot, with confidence interval $(1-\alpha)$ (Clopper and Pearson, 1934). The CP estimator is the $(1-\alpha)$-quantile of a Beta distribution with parameters $a = \hat{y}(l) + 1$ and $b = n(l) - \hat{y}(l)$. The obtained ELFP is, then, compared against a predefined quality target, to determine whether the lot will need to undergo a full burn-in (FB) or a reduced burn-in (RB) policy. Figure 1 depicts the proposed formulation of the quality estimation problem.

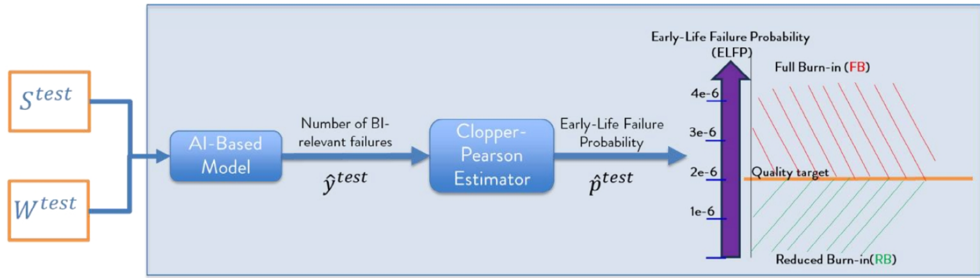

Fig. 1. Problem formulation for the production lot quality estimation.

3. Proposed method

To handle the available multimodal data, we develop a model based on a multibranch neural network. Specifically, the model receives in input the image of the wafer map W^{test} and the signal vector S^{test} and provides in output the expected number of BI-relevant failures $\hat{y}^{test}$ (Figure 2). For the branch of the model processing the wafer map images, convolution layers are used (Krizhevsky et al., 2012). They have been selected for their capability of sharing the model internal parameters and their sparse connectivity, which allow dealing with the large number of pixels in a single image, and their invariance to rotation and translation. In the last layer of this branch, the obtained feature maps are transformed into a feature vector h_W^{test} by using a flatten layer. With regard to the branch of the model processing the vector of signals S^{test}, a feedforward neural network (NN) is used to build the feature vector h_S^{test}. The features h_W^{test} and h_S^{test} extracted from both modalities, are, then, concatenated and given in input to fully connected layers for the prediction of the number of BI-relevant failures.

The model is trained using the error backpropagation algorithm (Rumelhart et al., 1986), which allows training the whole network simultaneously, i.e. all weights of the multibranch network are updated considering information from both sources.

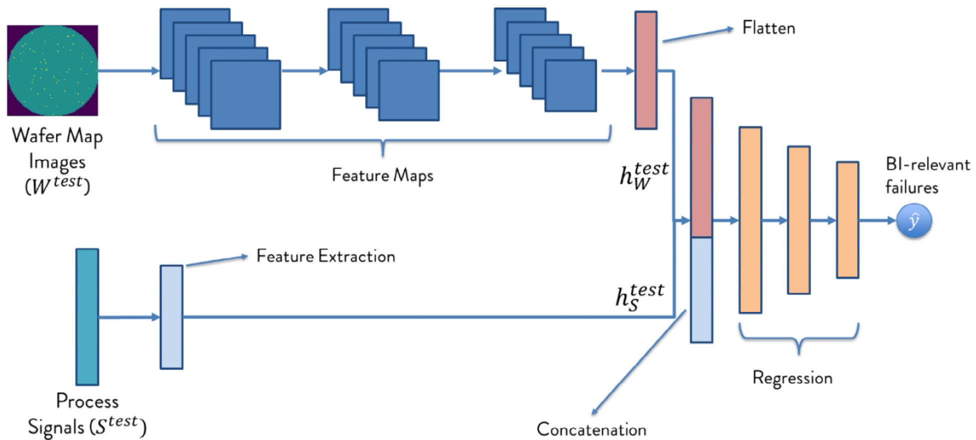

Fig. 2. Multibranch NN for the prediction of the number of BI-relevant failures.

3.1. Case study

The proposed method has been validated considering a synthetic dataset which emulates production signals and wafer map images in semiconductor production. Each pattern $(S(l), W(l), n(l), y(l)), l = 1, \ldots, L$, corresponds to a production lot, and each lot is formed by 25 wafers. The information available for each lot is an 8-dimensional vector $S(l)$ representing the production signals and a composite wafer map image $W(l)$. The vectors $S(l)$ are simulated using the procedure described in (Ahmed et al., 2023) and the images $W(l)$ are simulated considering the four classes of wafer maps shown in Figure 3. The class "None" is observed for wafers in normal conditions, where defects appear randomly due to the stochasticity of the process, whereas the other three classes ("Donut", "Edge-Ring" and "Edge-Loc") are observed for defective wafers due to the occurrence of abnormal conditions in different stages of the semiconductor production. The procedure followed for the simulation of the wafer map images is taken from (Maksim et al., 2019). In practice, the larger is the number of BI-relevant failures $y(l)$, the larger is the probability that a wafer of the lot belongs to the "Donut", "Edge-Ring" or "Edge-Loc" classes.

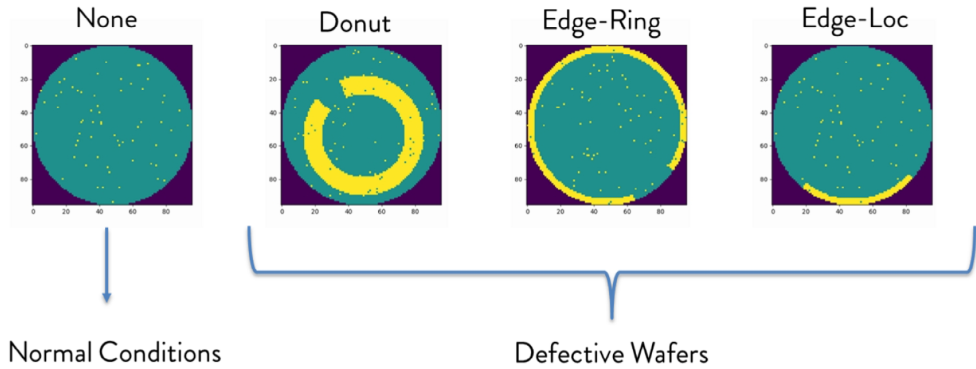

Fig. 3. Classes of wafer maps.

The dataset D is formed by 2000 patterns, 1800 of which are used for training the multibranch NN and 200 for testing its performance. To mimic that in semiconductor production some defects are not detectable in one source of data, no BI-relevant failures are simulated in the signals $S(l)$ of 20% of the lots of class FB and in the wafer maps of another 20% of the lots of class FB. As a result, 60% of the FB lots are detectable using anyone of the two data sources, whereas the remaining 40% of the FB lots are not detectable using only one data source. Furthermore, we consider two different proportions of FB/RB data to analyze the effect of using imbalanced data to develop the multibranch NN. This is done because the number of RB data in real production processes is expected to be larger than the number of FB data.

Figure 4 shows the two cases of balanced dataset (50% of patterns of class RB and 50% of class FB) and imbalanced dataset (95% of patterns of class RB and 5% of class FB).

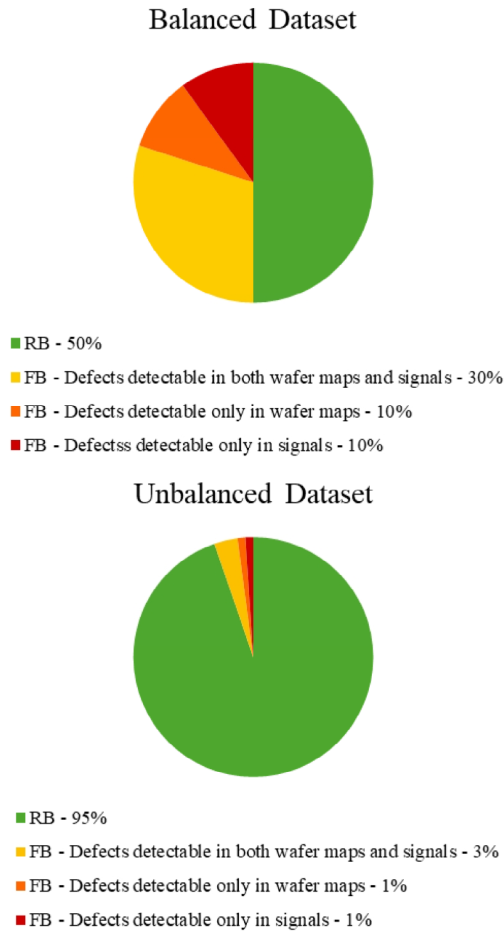

Fig. 4. Distribution of the lots in classes for the balanced (upper) and imbalanced (lower) cases.

4. Results and discussions

Table 1 reports the architecture of the developed multibranch NN. The performance of the proposed method is compared to that of models that receive in input a single source of data. Specifically, we consider a feedforward ANN processing only the signals and a CNN processing only the images.

Table 1. Proposed method architecture.

Branch	Parameter	Value
Signals	Number of layers	1
	Type of layers	Feedforward (F)
	Nodes per layer	64
	Activation function per layer	ReLU
Wafer Maps	Number of layers	7
	Type of layers	Convolutional (C)– Max Pooling (MP) – (C) – (MP) – (C) – (MP) – Flatten (FL)
	Nodes per layer	128 (C) – 0 (MP) – 64 (C) – 0 (MP) – 8 (C) – 0 (MP) – 0 (FL)
	Activation function per layer	ReLU(C) – None (MP) –ReLU (C) – None (MP) – ReLU (C) – 0 None (MP) – None (FL)
Fully connected layers for the regression task	Number of layers	4
	Type of layers	(F) - (F) - (F) - (F)
	Nodes per layer	64 – 32 – 8 – 1
	Activation function per layer	ReLU – ReLU – ReLU – Linear

Table 2 shows the classification accuracy on RB lots, i.e. the ratio between the number of RB lots correctly classified and the number of tested RB lots. When the dataset becomes imbalanced with a small fraction of FB patterns, the accuracy on RB increases. This is because the model's weights are tuned to minimize the loss of the training data, which contain more RB patterns than FB. In this case, the use of a multibranch NN allows obtaining a slight improvement of the performance when compared to the comparison models which use a single source of data. On the contrary, the accuracy on the FB data decreases when the models are trained on the imbalanced dataset (Table 3). Note that when the FB data are considered, the proposed method outperforms the comparison methods in the balanced case and reaches the same performance of the ANN model in the imbalanced case, which in turn achieves better performance than the CNN model.

Table 2. Classification accuracy on RB data of the multibranch deep neural network and of the two models that use a single source of data.

Method	Balanced Dataset	Imbalanced Dataset
Proposed Method	0.89 ± 0.01	0.94 ± 0.02
Comparison Method 1: ANN	0.43 ± 0.04	0.93 ± 0.02
Comparison Method 2: CNN	0.77 ± 0.03	0.86 ± 0.01

Table 3. Classification accuracy on FB data of the multibranch deep neural network and of the two models that use a single source of data.

Method	Balanced Dataset	Imbalanced Dataset
Proposed Method	0.99 ± 0.01	0.85 ± 0.01
Comparison Method 1: ANN	0.90 ± 0.01	0.85 ± 0.01
Comparison Method 2: CNN	0.98 ± 0.01	0.82 ± 0.06

Tables 4 and 5 report the performance of the models considering the data in which the defects are not detectable using only one source of data. In the case of the balanced dataset, the accuracy of the multibranch neural network is close to that of the model which uses the informative data. This result confirms that the multibranch NN is able to learn that the identification of some types of abnormal conditions requires to focus only on one source of data. On the contrary, in the case of the imbalanced dataset, the accuracy of the multibranch neural network remarkably decreases. This is due to the fact that too few examples of data in which the abnormal condition is detectable using only one source of data are available to train the model.

Table 4. Classification accuracy on the portion of FB data in which the defect is not detectable considering the signals.

Method	Balanced Dataset	Imbalanced Dataset
Proposed Method	0.98 ± 0.02	0.38 ± 0.05
Comparison Method 1: ANN	0.53 ± 0.02	0.35 ± 0.04
Comparison Method 2: CNN	1.00 ± 0.00	0.83 ± 0.02

Table 5. Classification accuracy on the portion of FB data in which the defect is not detectable considering the wafer maps.

Method	Balanced Dataset	Imbalanced Dataset
Proposed Method	1.00 ± 0.00	0.98 ± 0.02
Comparison Method 1: ANN	1.00 ± 0.00	1.00 ± 0.00
Comparison Method 2: CNN	0.93 ± 0.02	0.82 ± 0.09

5. Conclusions

A method for the prediction of the quality of semiconductor lots before BI testing is developed. The objective is to identify high quality lots for which it is possible to reduce burn-in testing whilst keeping high reliability standards. The developed method is based on a multibranch NN that receives in input multiple types of data and is trained to classify different types of abnormal conditions that may occur in the different phases of the production process.

The results obtained in a synthetic case study representative of a semiconductor production process show that the proposed method is successful in incorporating information from wafer maps and process signals. Specifically, when one modality is not informative about the abnormal conditions causing BI-relevant failures, the use of the other modality allows improving the accuracy of the quality estimation. The method is also shown more accurate than methods that use only one source of data, especially when the training set becomes imbalanced with a reduction of the proportion of lots that need full BI.

Future work will consider the development of models for further improving accuracy in case of imbalanced datasets, as this issue is likely to be present in real applications. This will involve the analysis of existing literature on imbalanced data in the context of machine learning and deep learning (Ren et al., 2023). Other future works are the application of the developed method to real production data collected within the iRel40 European co-funded innovation project (https://www.irel40.eu), and the development of methods to explain the multibranch NN functioning.

Acknowledgements

This work has been performed within the iRel40 European co-funded innovation project, 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. Also, the participation of Piero Baraldi is supported by the FAIR (Future Artificial Intelligence Research) project, funded by the NextGenerationEU program within the PNRR-PE-AI scheme (M4C2, investment 1.3, line on Artificial Intelligence).

Disclaimer

The document reflects only the author’s view, and the JU is not responsible for any use that may be made of the information it contains.

References

Ahmed, I., Baraldi, P., Zio, E., Lewitschnig, H. et al. 2023. Prediction of the Number of Defectives in a Production Batch of Semiconductor Devices. Proceedings of the 33rd European Safety and Reliability Conference (ESREL 2023), 2615–2620.

Aria, A., Lopez Droguett, E., Azarm, S., Modarres, M. 2020. Estimating damage size and remaining useful life in degraded structures using deep learning-based multi-source data fusion. Structural Health Monitoring 19(5), 1542-1559. https://doi.org/10.1177/1475921719890616

Cao, J., Zhang, Z., Tao, F., Zhang, L., Luo, Y., Zhang, J., Han, J., Xie, J. 2021. Integrating Multi-Source Data for Rice Yield Prediction across China using Machine Learning and Deep Learning Approaches. Agricultural and Forest Meteorology, 297. https://doi.org/10.1016/j.agrformet.2020.108275

Cho, H., Koo, W., Kim, H. 2023. Prediction of Highly Imbalanced Semiconductor Chip-Level Defects in Module Tests Using Multimodal Fusion and Logit Adjustment. IEEE Transactions on Semiconductor Manufacturing 36(3), 425–433. https://doi.org/10.1109/TSM.2023.3283101

Clopper, C. J., Pearson, E. S. 1934. The use of confidence or fiducial limits illustrated in the case of the binomial. Biometrika 26(4), 404–413.

Kim, G., Choi, J. G., Ku, M., Cho, H., Lim, S. 2021. A Multimodal Deep Learning-Based Fault Detection Model for a Plastic Injection Molding Process. IEEE Access, 9, 132455–132467. https://doi.org/10.1109/ACCESS.2021.3115665

Krizhevsky, A., Sutskever, I., Hinton, G. E. 2012. ImageNet Classification with Deep Convolutional Neural Networks. In Advances In Neural Information Processing Systems (1-9). https://doi.org/http://dx.doi.org/10.1016/j.protcy.2014.09.007

Kurz, D., Lewitschnig, H., Pilz, J. 2021. Flexible time reduction method for burn-in of high-quality products. Quality and Reliability Engineering International 37(6), 2900–2915. https://doi.org/10.1002/qre.2896

Maksim, K., Kirill, B., Eduard, Z., Nikita, G., Aleksandr, B., Arina, L., Vladislav, S., Daniil, M., Nikolay, K. 2019. Classification of Wafer Maps Defect Based on Deep Learning Methods With Small Amount of Data. 2019 International Conference on Engineering and Telecommunication (EnT), 1–5. https://doi.org/10.1109/EnT47717.2019.9030550

Ren, Z., Lin, T., Feng, K., Zhu, Y., Liu, Z., Yan, K. 2023. A Systematic Review on Imbalanced Learning Methods in Intelligent Fault Diagnosis. IEEE Transactions on Instrumentation and Measurement 72. https://doi.org/10.1109/TIM.2023.3246470

Rumelhart, D. E., Hinton, G. E., Williams, R. J. 1986. Learning representations by back-propagating errors. Nature 323(6088), 533–536.

Suhir, E. 2019. To burn-in, or not to burn-in: That’s the question. Aerospace, 6(3). https://doi.org/10.3390/aerospace6030029

Yang, Z., Baraldi, P., Zio, E. 2021. A multi-branch deep neural network model for failure prognostics based on multimodal data. Journal of Manufacturing Systems 59, 42-50. https://doi.org/10.1016/j.jmsy.2021.01.007