Proceedings of the 2022 Winter Simulation Conference
B. Feng, G. Pedrielli, Y. Peng, S. Shashaani, E. Song, C.G. Corlu, L.H. Lee, E.P. Chew, T. Roeder, and P. Lendermann, eds.

EQUIPMENT PROGNOSIS WITH GHOST POINT RESAMPLING METHOD UNDER IMBALANCED DATA RESTRICTION

Rifa Arifati, Jakey Blue*

Institute of Industrial Engineering, National Taiwan University
1, Section 4, Roosevelt Road
Taipei, 106017, TAIWAN (R.O.C)

ABSTRACT

Fault detection and classification are crucial in equipment monitoring since it helps to prevent unexpected breakdowns. Nowadays, this task is implemented by massively applying various machine learning and deep learning techniques to detect faults as early as possible. In addition to the semiconductor manufacturing complexity, such as the sensor data being in the form of multivariate time series of variant lengths, one challenge is that the process data are very often (extremely) imbalanced. This research proposes a Ghost Point Resampling methodology which consists of kernelizing the FDC data, calculating distance measures, upper sampling the minority class, and finally classifying the faults in the kernel space. The effectiveness of doing data augmentation in the invisible kernel space will be demonstrated by case studies.

1 INTRODUCTION

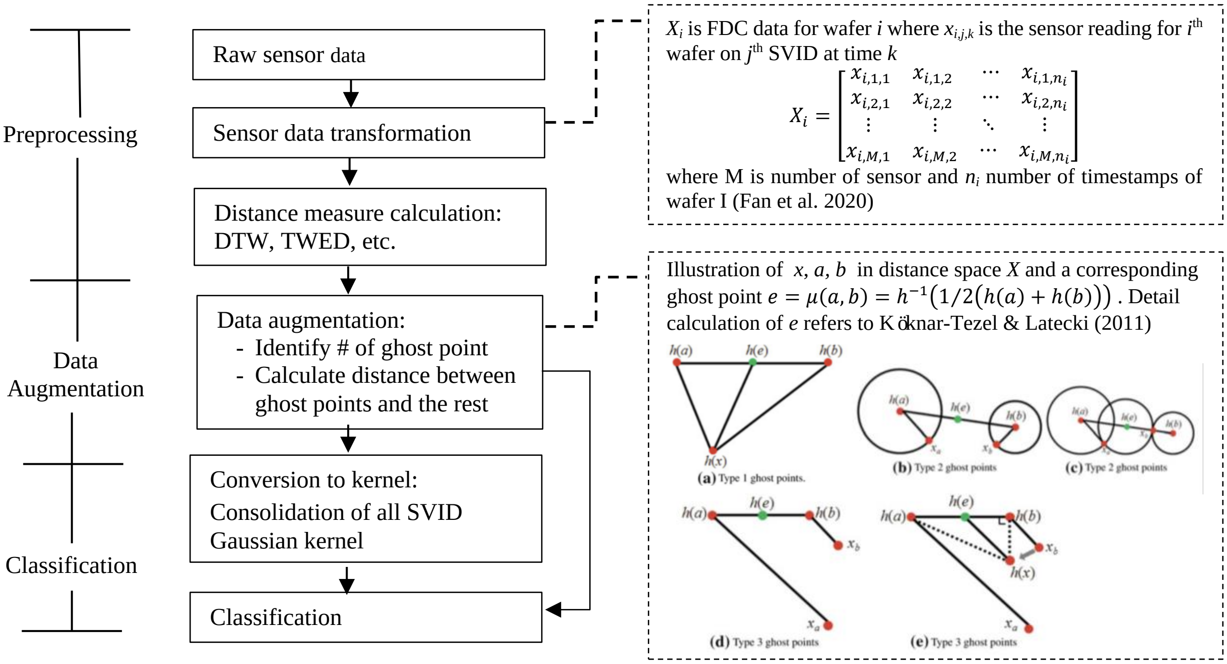
IC makers usually install comprehensive sensors to control the process to decrease the chances of fault occurrence, and consequently, massive production data are collected. To make sense of the big process data, engineers and researchers put enormous effort into the equipment condition modeling and try to set up rules/policies for optimal equipment control. However, an obvious drawback of the data-driven modeling techniques is the fault data are often of the minority class and result in the ineffectiveness of the imbalanced data analytics. Moreover, the data collected from multiple sensors exhibit sequence dependencies and often have variant lengths given the process nature. Not to mention that the sensor readings are also impacted by the recipe changes. It is, therefore, very challenging to classify the faults in the matured manufacturing production line.

Research on the classification of imbalanced data has emerged in the last two decades (see He and Garcia 2009; Guo et al. 2017), but unfortunately, the classification of imbalanced time series in semiconductor manufacturing is still rare to see. Thus, a framework for resolving the classification of highly imbalanced FDC data in semiconductor manufacturing is proposed in this research.

2 METHODOLOGY

As shown in Figure 1, the proposed framework consists of three major phases. In the preprocessing phase, raw time series data of wafer i are collected in tabular form and denoted as $\mathbf{X}_i$. The distance metric between two wafers is calculated by applying the Dynamic Time Warping (DTW) method. Given n wafers, an $n \times n$ distance matrix for each sensor will be obtained.

Following the Ghost Point Method proposed by Köknar-Tezel and Latecki (2011), we propose to consolidate the results of multiple time series into a unified distance (or dissimilarity) matrix. For the minority class, k synthetic points are created in the space of the distance matrix, which can be regarded as in the kernel space. Accordingly, a kernel-based classifier shall be applied. Because the data are sequence-dependent, Gaussian or negative kernels that preserve and reserve the order of the data should be used (Badiane and Cunningham 2022). Finally, the classifier is trained with the augmented data.



X_i is FDC data for wafer *i* where $x_{i,j,k}$ is the sensor reading for i^{th} wafer on j^{th} SVID at time *k*

$$X_i = \begin{bmatrix} x_{i,1,1} & x_{i,1,2} & \cdots & x_{i,1,n_i} \\ x_{i,2,1} & x_{i,2,2} & \cdots & x_{i,2,n_i} \\ \vdots & \vdots & \ddots & \vdots \\ x_{i,M,1} & x_{i,M,2} & \cdots & x_{i,M,n_i} \end{bmatrix}$$

where M is number of sensor and n_i number of timestamps of wafer I (Fan et al. 2020)

Illustration of x, a, b in distance space X and a corresponding ghost point $e = \mu(a,b) = h^{-1}\big(1/2\big(h(a) + h(b)\big)\big)$. Detail calculation of *e* refers to K ӧknar-Tezel & Latecki (2011)

Figure 1: The proposed framework of the ghost point resampling method for imbalanced time series.

3 RESULT AND ONGOING WORK

For the moment, we utilize PHM 2018 data to test the proposed framework. There are 17 sensors and 11 faults out of 3719 wafers in recipe 01. Different numbers of ghost points are created to train the model. The performance is shown in Figure 2. A quick conclusion can be learned that the ghost point method improves the SVM performance significantly.

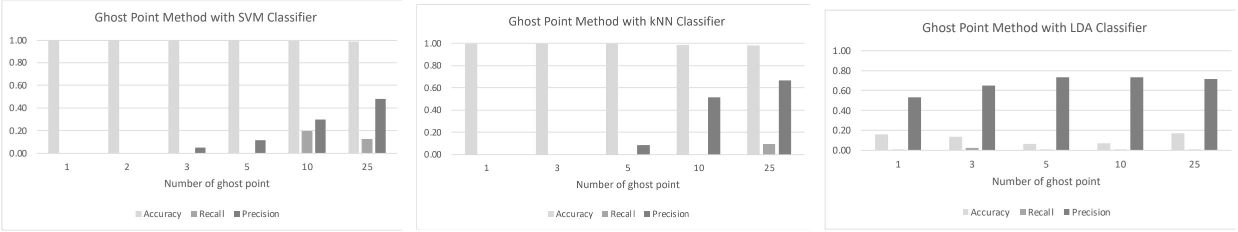


Figure 2: Testing results of the ghost point resampling applied to 2018 PHM challenge data.

Based on the preliminary results, we can see that the distance metrics, the consolidation of the multiple time series, and the kernel-based classifiers are all keys to the success of modeling the imbalanced faults. Methodological extensions are under development, and related experiments are being designed and tested.

REFERENCES

Badiane, M., and P. Cunningham. 2022. "An Empirical Evaluation of Kernels for Time Series". *Artificial Intelligence Review* 55(3): 1803-1820.

Fan, S. K. S., C. Y. Hsu, D. M. Tsai, F. He, and C. C. Cheng. 2020. "Data-driven Approach for Fault Detection and Diagnostic in Semiconductor Manufacturing". *IEEE Transactions on Automation Science and Engineering* 17(4): 1925-1936.

Guo, H., Y. Li, J. Shang, M. Gu, Y. Huang, and B. Gong. 2017. "Learning from Class-imbalanced Data: Review of Methods and Applications". *Expert Systems with Applications* 73: 220-239.

He, H., and E. A. Garcia. 2009. "Learning from Imbalanced Data". *IEEE Transactions on Knowledge and Data Engineering* 21(9): 1263-1284.

K ӧknar-Tezel, S., and L. J. Latecki. 2011. "Improving SVM Classification on Imbalanced Time Series Data Sets with Ghost Points". *Knowledge and Information Systems* 28(1): 1-23.