A Deep Variational Autoencoder for Spatial Patterns Clustering in Wafemap Measurement Data

Anonymous Author(s)
Affiliation
Address
email

Abstract

There exists a large number of process deviations in semiconductor manufacturing processes. Automated root cause analysis and decision-making help to significantly improve the effectiveness of manufacturing processes. Manufacturing defects reveal typical patterns in wafer measurement data. Spatial patterns recognition in wafermap data improves the efficiency of finding production issues during different process steps, as early as possible. In this paper, we introduce a deep learning approach for recognition and clustering of spatial patterns in wafermap test data in an unsupervised fashion. First, measurement values are pre-processed, then, a deep variational autoencoder is used to extract a low-dimensional representation of the wafermaps. Finally, various structures in the latent space are detected and wafers assign to the extracted clusters. Extensive simulations show that the proposed approach outperforms the best existing methods over a real-world dataset[1].

1 Introduction

Recently, "Industry 4.0" [1] outlined new challenges for semiconductor industry on producing smaller lot sizes, yet customer-specific products. Many companies have started to add new paradigms to their manufacturing processes, in order to stay in increasingly global markets. Automated root cause analysis and decision making with reduced human intervention has potential to efficiently improve effectiveness of semiconductor manufacturing. To this end, proposing an algorithm to detect defects and cluster them from given sensory data is an inevitable task [2]. Manufacturing defects reveal typical shapes (patterns) in measurement wafer test data (*e.g.* rings, spots, repetitive patterns). Spatial patterns recognition in wafer test data is an essential step for finding root causes of production issues. Several methods based on traditional image processing approaches have been proposed [3, 4]. There exist several recent machine learning techniques recognizing more complex patterns in wafer measurement test data. They have been proposed based on supervised training of mixture models [5], singular value decomposition [6], neural networks [7], and support vector machines [8]. Although such techniques are powerful, their supervised setting still requires a human expert to manually label a training dataset. To remove the subjective factors from wafermap pattern recognition task, one can take the advantage of an unsupervised approach, which in turn reduces costs and classification errors. In addition, a method needs to automatically detect the hidden dependencies between different types of wafer defects, that helps identification of unknown (or overlooked) patterns. Self-organizing neural networks [9], self-organizing maps [10], and dimensionality reduction based methods (*e.g.* [11, 12]) are among this category of methods. In this paper, we develop a deep variational autoencoder approach for extracting useful information from the wafer test data, then use the extracted low-dimensional latent features for recognizing and clustering spatial patterns in the wafermap measurements. In Section 3, we evaluate the performance of the proposed method compared to several well-known

[1] Implementation codes and dataset will be online for camera-ready version.

Submitted to 32nd Conference on Neural Information Processing Systems (NIPS 2018). Do not distribute.

methods. We show that our approach yields better separated wafermaps clusters with more structural behavior for different number of clusters and size of latent space.

2 Proposed method

2.1 Data pre-processing

Our available real dataset, provided by Infineon Technologies (`http://infineon.com`), consists of 6 wafer lots, each has 50 wafers containing 17509 chips. Each chip is measured with 20 different tests (features) and its position within a wafer is stored as a tuple. We consider each test measurement of a wafer as a bitmap. Overall, we have 6000 wafermaps, where each one represented as a bitmap of size 193x115 pixel. Data pre-processing is a fundamental step to clean the data before designing a machine learning model [13]. We apply several consecutive pre-processing steps to raw wafermaps, which are depicted in Figure 1: (1) We utilize a median absolute deviation (MAD)-based outlier detection method by modifying the common Z-score mechanism [14]; (2) Wafermaps are binarized by replacing the present values with 1 and the missing values (holes) with 0. Mathematical morphology mechanism [15] is then used to close small holes in the wafer area and find contours of the wafer; (3) Missing values in wafer area are inpainted with values reconstructed from neighborhood information around each missing region, using Chui-Mhaskar inpainting algorithm [16] via solving the biharmonic equations; (4) After feature normalization, wafers are smoothened using the median filtering procedure within a sliding window. A sample pre-processed wafermap is depicted in Figure 2. The cleansed wafermaps can then be used for further feature extraction and clustering tasks.

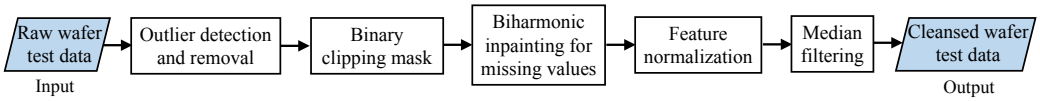


Figure 1: Wafermap pre-processing procedure

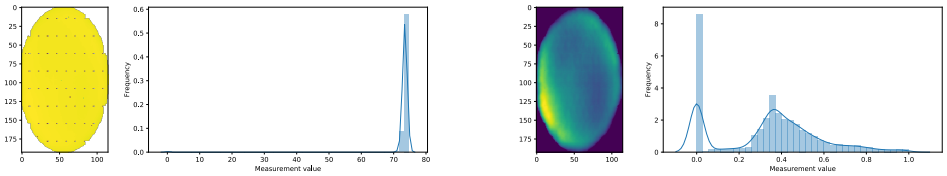


Figure 2: A raw wafermap test data (left) and the output of our pre-processing procedure with clearly visible crescent moon pattern (right).

2.2 A Deep Variational Autoencoder

The pre-processed wafermaps can be seen as N individual dataset containing iid samples of a discrete random variable X. Now, we want to extract a low-dimensional representation (*i.e.* latent variable Z) of the data to overcome the *curse of dimensionality* [17]. To this end, our approach is based on autoencoding variational Bayes [18].

We assume latent variable Z has a prior distribution $p(z) = \mathcal{N}(0, I)$ and X is conditioned on Z with likelihood $p_\theta(x|z;\theta)$, which gives us the latent variable model $p_\theta(x,z) = p_\theta(x|z)p(z)$. Since Bayesian inference of the latent features directly form posterior $p_\theta(z|x)$ is intractable for complicated distributions, a new conditional distribution $q_\phi(z|x)$ is introduced as an approximation. We assume the posterior $q_\phi(z|x)$ is a multivariate Gaussian distribution with a diagonal covariance $\mathcal{N}(\mu, \sigma^2 I)$ and is referred to *encoder* (or *recognition model*). Conditional distribution $p_\theta(x|z)$ is referred to *decoder*. Then, the parameters of this model ϕ, θ are calculated by minimizing the Kullback-Leibler (KL)-divergence between $q_\phi(z|x)$ and $p_\theta(z|x)$, *i.e.* maximizing the following loss function:

$$\mathcal{L}(\theta, \phi; x) = -D_{KL}(q_\phi(z|x)||p_\theta(z|x)) = \mathbb{E}_{q_\phi(z|x)}\Big[\log \frac{p_\theta(z|x)}{q_\phi(z|x)}\Big] \tag{1}$$

Using conditional distribution $p(z|x) = \frac{p(x,z)}{p(x)}$, one can get Evidence Lower Bound (ELBO) [19, 20]:

$$\mathcal{L}(\theta, \phi; x) = \overbrace{\mathbb{E}_{q_\phi(z|x)}[\log p_\theta(x|z)]}^{\text{Reconstruction Error}} - \overbrace{D_{KL}(q_\phi(z|x)||p(z))}^{\text{Regularization}} \tag{2}$$

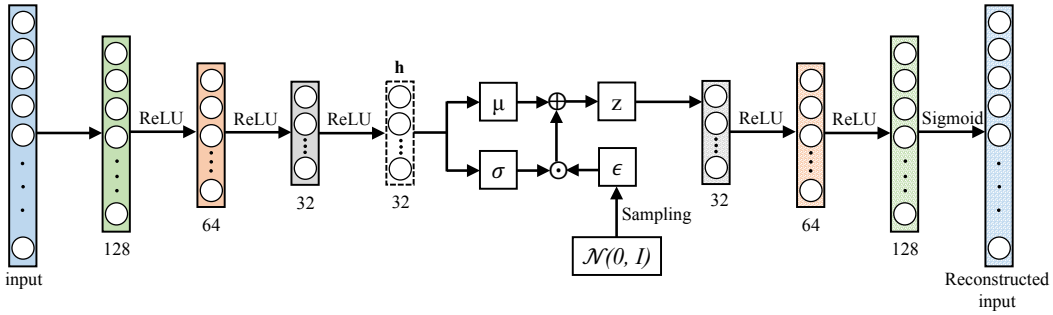


Figure 3: Architecture of our deep variational autoencoder with its reparametrization trick.

We use a deep Variational Auto-Encoder (VAE) neural network [21] to implement both encoder and decoder. A neuron i in the neural network takes an input vector $\mathbf{x}_i \in \mathbb{R}^d$ and maps it to an output vector $\mathbf{y}_i \in \mathbb{R}^{d'}$ with deterministic mapping (forward propagation) $\mathbf{y}_i = \varphi(\mathbf{W}_i\mathbf{x}_i + \mathbf{b}_i)$, where $\varphi(.)$ is differentiable non-linear activation function, $\mathbf{W}_i \in \mathbb{R}^{d\times d'}$ is a weight matrix and $\mathbf{b}_i \in \mathbb{R}^{d'}$ is a bias. We used Rectified Linear Unit (ReLU) and Sigmoid as activation functions of neurons in hidden layers and output layer, respectively. Model parameters $\theta = \{\mathbf{W}, \mathbf{b}\}$ are trained with Stochastic Gradient Descent (SGD) algorithm with respect to loss function $\mathcal{L}$ by iteratively updating $\theta \leftarrow \theta - \eta\frac{\partial\mathcal{L}}{\partial\theta}$ (backpropagation), where η is learning rate. We used RMSprop [22], an adaptive learning rate in which the gradient is divided by a running average of its recent magnitude. Optimizing the loss function in Equation (2) with SGD is problematic, because the latent vector z used in $p_\theta(x|z)$ must be sampled from distribution $q_\phi(z|x)$. However, the random sampling is a non-differentiable operation, hence it cannot be used for calculating gradients during backpropagation. We use a reparametrizarion trick to shift the randomness of the sampling operation into a random noise vector ϵ which allows us to express individual samples of reparametrized random variable $\tilde{z} \sim q_\phi(z|x)$ deterministically as $\tilde{z} = \mu + \sigma \odot \epsilon$, where $\epsilon \sim \mathcal{N}(0, I)$ and operator $\odot$ is an element-wise product. The neural network architecture with its reparametrization trick, used in our experiment, is depicted in Figure 3.

As mentioned before, the distributions $p(z) = \mathcal{N}(0, I)$ and $q_\theta(z|x) = \mathcal{N}(\mu, \sigma^2 I)$ are normally distributed, so the regularization term in Equation (2) can be rewritten as:

$$D_{KL}(q_\phi(z|x)||p(z)) = \frac{1}{2}\sum_{j=1}^{J}\Big(1 + \log(\sigma_j^2) - \mu_j^2 - \sigma_j^2\Big) \tag{3}$$

where J is the dimension of the latent vector z and the distribution parameters are mapped from a hidden layer $\mathbf{h}$ of the encoder as $\mu_j = \mathbf{W}_\mu\mathbf{h} + \mathbf{b}_\mu$ and $\log(\sigma_j^2) = \mathbf{W}_\sigma\mathbf{h} + \mathbf{b}_\sigma$. We assume that the decoder is Bernoulli distributed and the reconstruction error in Equation (2) can be approximated by binary cross-entropy as:

$$\mathbb{E}_{q_\phi(z|x)}[\log p_\theta(x|z)] \approx \frac{1}{L}\sum_{i=1}^{L}\Big(\log y_i + (1 - x_i)\log(1 - y_i)\Big) \tag{4}$$

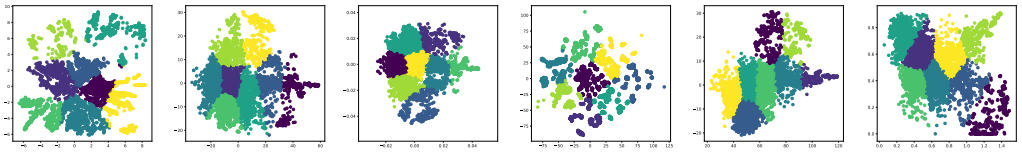
where x_i is the input of the encoder, y_i is the output of the decoder and L is the number of latent samples $z_1, \ldots, z_L$ used to approximate the expected value of the log-likelihood. The wafermap dataset X consisting of N samples can be potentially very large. Only a random subset (minibatch) of M wafers is sampled from the whole dataset. Hence, the loss function for each minibatch is $\mathcal{L}(\theta, \phi; X) \simeq \frac{N}{M}\sum_{i=1}^{M}\mathcal{L}(\theta, \phi; x_i)$. The number of sampled latent variables L in each iteration can be set to 1 for sufficiently large number of minibatches M [18] (in our implementation $L = 1, M = 100$).

2.3 Wafermaps Patterns Clustering

We have described a mechanism for non-linear mapping of high-dimensional wafer measurement data into a low-dimensional representation. Now, we specify how to group the extracted latent features into clusters based on a distance measure. Wafermaps with similar patterns should be considered in the same cluster and dissimilar wafermaps should be clustered in different groups. There exist two types of clustering methods (*i.e.* hierarchical and partitioning) that can be applied for clustering of the wafermaps. We used one algorithm from each category, namely *Hierarchical agglomerative* [23] and *k-means clustering*.

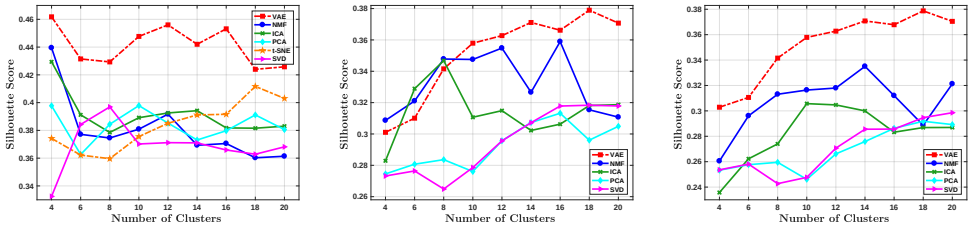
3 Evaluation

In this section, we evaluate the performance of the proposed variational autoencoder-based method compared to the other commonly used decomposition methods for spatial wafermaps patterns clustering. Figure 4 shows the eight detected clusters in two dimensional latent space for different feature extraction methods. The same pre-processing and clustering algorithm used for the competing methods. To evaluate how similar a latent vector is to the other vectors in its own cluster compared to the other clusters, the Silhouette metric [24] was used. The higher this score is, the better clustering accuracy the method will have. Figure 5 shows the Silhouette score of the competing methods over various low-dimensional latent spaces for two clustering methods.

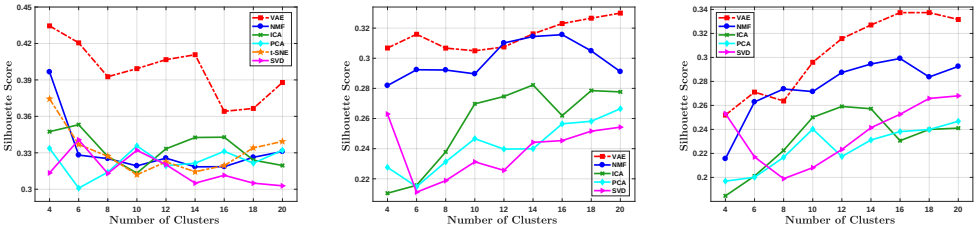


(a) VAE (b) PCA (c) ICA (d) t-SNE (e) T-SVD (f) NMF

Figure 4: Wafermaps projected into two dimensional latent feature space and clustered into 8 clusters. Feature extraction methods: (a) Our Variational Autoencoder; (b) Principal Component Analysis [25]; (c) Independent Component Analysis [26]; (d) t-Distributed Stochastic Neighbor Embedding [27]; (e) Truncated Singular-value Decomposition [6]; (f) Non-Negative Matrix Factorization [28].



(a) k-means, LATE_DIM = 2 (b) k-means, LATE_DIM = 4 (c) k-means, LATE_DIM = 5



(d) Agglom., LATE_DIM = 2 (e) Agglom., LATE_DIM = 4 (f) Agglom., LATE_DIM = 5

Figure 5: Evaluation of different methods with two clustering methods: k-means and agglomerative clustering. Our approach based on variational autoencoder (VAE) yields better separated clusters in terms of Silhouette score in comparison with the other methods in majority of cases.

4 Conclusion

Systematic defects in manufacturing industry are caused by a malfunction in a process equipment or human errors. Automated detection of such production issues and automated root cause analysis will improve the efficiency of semiconductor production. Manufacturing defects often exhibit patterns in measured test data. Recognizing spatial patterns and their categorization are essential tasks in root cause identification of the production issues. In this paper, after pre-processing procedures, we extracted the most characteristic features of the data using a deep variational autoencoder neural network, in order to recognize the wafermap patterns. We then utilized the extracted low-dimensional features in clustering task to group the wafers into meaningful clusters based on their spatial patterns. Finally, we experimentally showed the performance superiority of the proposed approach over a real dataset, in comparison with several well-known methods.

References

[1] H. Kagermann, J. Helbig, A. Hellinger, and W. Wahlster. *Recommendations for Implementing the Strategic Initiative INDUSTRIE 4.0: Securing the Future of German Manufacturing Industry ; Final Report of the Industrie 4.0 Working Group*. Forschungsunion, 2013.

[2] Chia-Yu Hsu. Clustering ensemble for identifying defective wafer bin map in semiconductor manufacturing. *Mathematical Problems in Engineering*, 2015.

[3] L. Breaux and B. Singh. Automatic defect classification system for patterned semiconductor wafers. In *Proceedings of International Symposium on Semiconductor Manufacturing*, pages 68–73, Sep 1995.

[4] Frederic Duvivier. Automatic detection of spatial signature on wafermaps in a high volume production. In *Proceedings of the 14th International Symposium on Defect and Fault-Tolerance in VLSI Systems*, DFT '99, pages 61–, Washington, DC, USA, 1999. IEEE Computer Society.

[5] Jinho Kim, Youngmin Lee, and Heeyoung Kim. Detection and clustering of mixed-type defect patterns in wafer bin maps. *IISE Transactions*, 50(2):99–111, 2018.

[6] Kamal Taha, Khaled Salah, and Paul D Yoo. Clustering the dominant defective patterns in semiconductor wafer maps. *IEEE Transactions on Semiconductor Manufacturing*, 31(1):156–165, 2018.

[7] F.L. Chen, Sheng-Che Lin, K. Yih-Yuh Doong, and K.L. Young. Logic product yield analysis by wafer bin map pattern recognition supervised neural network. *2003 5th International Conference on ASIC. Proceedings (IEEE Cat. No.03TH8690)*.

[8] Li-Chang Chao and Lee-Ing Tong. Wafer defect pattern recognition by multi-class support vector machines by using a novel defect cluster index. *Expert Systems with Applications*, 36(6):10158 – 10167, 2009.

[9] Chuan-Yu Chang, ChunHsi Li, Jia-Wei Chang, and MuDer Jeng. An unsupervised neural network approach for automatic semiconductor wafer defect inspection. *Expert Systems with Applications*, 36(1), 2009.

[10] F. Di Palma, G. De Nicolao, G. Miraglia, E. Pasquinetti, and F. Piccinini. Unsupervised spatial pattern classification of electrical-wafer-sorting maps in semiconductor manufacturing. *Pattern Recognition Letters*, 26(12):1857 – 1865, 2005.

[11] Jianbo Yu and Xiaolei Lu. Wafer map defect detection and recognition using joint local and nonlocal linear discriminant analysis. *IEEE Transactions on Semiconductor Manufacturing*, 29(1):33–43, 2016.

[12] Gal Mishne and Israel Cohen. Multi-channel wafer defect detection using diffusion maps. In *Electrical & Electronics Engineers in Israel (IEEEI), 2014 IEEE 28th Convention of*, pages 1–5. IEEE, 2014.

[13] Davide Chicco. Ten quick tips for machine learning in computational biology. *BioData Mining*, 2017.

[14] Christophe Leys, Christophe Ley, Olivier Klein, Philippe Bernard, and Laurent Licata. Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median. *Journal of Experimental Social Psychology*, 49(4):764–766, 2013.

[15] Chris Solomon and Toby Breckon. *Fundamentals of Digital Image Processing: A practical approach with examples in Matlab*. John Wiley & Sons, 2011.

[16] Charles K. Chui and H.N. Mhaskar. Mra contextual-recovery extension of smooth functions on manifolds. *Applied and Computational Harmonic Analysis*, 28(1):104–113, Jan 2010.

[17] R. Bellman and Rand Corporation. *Dynamic Programming*. Rand Corporation research study. Princeton University Press, 1957.

[18] Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. *CoRR*, abs/1312.6114, 2013.

[19] Matthew D. Hoffman, David M. Blei, Chong Wang, and John Paisley. Stochastic variational inference. *J. Mach. Learn. Res.*, 14(1):1303–1347, May 2013.

[20] David M. Blei, Alp Kucukelbir, and Jon D. McAuliffe. Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112(518):859–877, 2017.

[21] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016.

[22] T. Tieleman and G. Hinton. Lecture 6.5—RmsProp: Divide the gradient by a running average of its recent magnitude. COURSERA: Neural Networks for Machine Learning, 2012.

[23] Gabor J Szekely and Maria L Rizzo. Hierarchical clustering via joint between-within distances: Extending ward's minimum variance method. *Journal of classification*, 22(2):151–183, 2005.

[24] Peter J. Rousseeuw. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20(Supplement C):53 – 65, 1987.

[25] Tiago J Rato, Jakey Blue, Jacques Pinaton, and Marco S Reis. Translation-invariant multiscale energy-based pca for monitoring batch processes in semiconductor manufacturing. *IEEE Transactions on Automation Science and Engineering*, 14(2):894–904, 2017.

[26] Jong-Min Lee, S Joe Qin, and In-Beum Lee. Fault detection and diagnosis based on modified independent component analysis. *AIChE journal*, 52(10):3501–3514, 2006.

[27] George C Linderman, Manas Rachh, Jeremy G Hoskins, Stefan Steinerberger, and Yuval Kluger. Efficient algorithms for t-distributed stochastic neighborhood embedding. *arXiv preprint arXiv:1712.09005*, 2017.

[28] Reinhard Schachtner. *Extensions of non-negative matrix factorization and their application to the analysis of wafer test data*. PhD thesis, 2010.