Anomaly Detection in Batch Manufacturing Processes using Localised Reconstruction Errors from 1-Dimensional Convolutional AutoEncoders

Mark Gorman, Dr Xuemei Ding, Professor Liam Maguire, and Professor Damien Coyle

Abstract

Multivariate batch time-series data sets within Semiconductor manufacturing processes present a difficult environment for effective Anomaly Detection (AD). The challenge is amplified by the limited availability of ground truth labelled data. In scenarios where AD is possible, black box modelling approaches constrain model interpretability. These challenges obstruct the widespread adoption of Deep Learning solutions. The objective of the study is to demonstrate an AD approach which employs 1-Dimensional Convolutional AutoEncoders (1d-CAE) and Localised Reconstruction Error (LRE) to improve AD performance and interpretability. Using LRE to identify sensors and data that result in the anomaly, the explainability of the Deep Learning solution is enhanced. The Tennessee Eastman Process (TEP) and LAM 9600 Metal Etcher datasets have been utilised to validate the proposed framework. The results show that the proposed LRE approach outperforms global reconstruction errors for similar model architectures achieving an AUC of 1.00. The proposed unsupervised learning approach with AE and LRE improves model explainability which is expected to be beneficial for deployment in semiconductor manufacturing where trustworthy results are critical and interpretable by process engineering teams.

Index Terms **- Deep Learning, Fault Detection and Classification, Semiconductor Manufacturing, Convolutional AutoEncoder, Reconstruction Error**

I. Introduction

Advanced Semiconductor manufacturing has pushed existing Machine Learning (ML) [1] approaches to the limits of their capability, leading many practitioners to seek improved performance through data-driven Deep Learning (DL) approaches. Fault Detection and Classification (FDC) [2] also referred to as Anomaly Detection (AD), defines the approach where anomalous operating equipment conditions can be detected through the real-time processing of time-series sensor data generated during wafer processing. Semiconductor processes operate in a multi-re-entrant, batch manufacturing environment accompanied by inherent variability within processes that utilise consumable parts [2]. Batch manufacturing refers to the mode in which equipment sets are operated with a predefined recipe of which there are multiple phases executed within a standard process duration. Traditional ML in this context is one where data is manually pre-processed, features engineered and selected by a human in the loop, and established workflows that heavily leverage empirical knowledge and subject matter expertise. Comparisons have been carried out [3] that implement this paradigm and demonstrate that the approach is viable. However, when a representative feature space is created through the aggregation of time-series signals, information loss is likely, therefore decreasing the capacity for an algorithm to detect anomalies [4]. Historical FDC approaches have involved Principle Components Analysis (PCA)[5], clustering methods such as k-Means or k-Nearest Neighbours (kNN) [6], Support Vectors [7], Decision Trees [8] or an ensemble of ML techniques [9]. However, in these instances, the modelled domain space is reduced through data aggregation, feature engineering and feature removal through a selection process. Deep Learning (DL) has been demonstrated to outperform traditional ML approaches and even human-level performance [10]. Although there are several fields of interest within DL, AutoEncoders (AE's) are effective in unsupervised AD [11] and dimensionality reduction [12]. AE's with Dense layers have been applied for AD [13] in batch manufacturing. However, Dense AE's are rigid in their network architecture potentially reducing the network's ability to capture time-based feature variation, resulting in less accurate models [14]. An alternative approach is a Convolutional AutoEncoder (CAE), created through the stacking of several non-linear layers that enable the network to extract hierarchical high-level features. Convolutional layers reduce the feature space through Pooling or increased kernel strides [14] and have been shown to perform well as feature extractors [12]. A 2D-Convolutional layer is preferred if data is related across both the x and y-axis [15]. Conversely, where inputs have only one axis of information dependence, a 1d-Convolutional layer is more suited to the input data shape [14]. 1d-CAEs are proficient extractors of high-level features along a single input axis [16]. For manufacturing use cases, the 1d-Convolutional kernel is applied along the time-dependent axis with the remaining axis containing input sensors [17]. Maggipinto et al [18] have implemented a hybrid approach where the 1d-CAE is used as a feature extractor for traditional ML AD algorithms. Zhang et al [19] implement a hybrid adoption of 1d-Convolutional layers for unsupervised feature extraction for other models. It is observed that 1d-Convolutional layers achieve high performance for feature extraction [14] in classification frameworks, however, a significant limitation is a requirement for the existence of labelled data. It is also noted that traditional

Manuscript submitted on 7th October 2021; revised 29th March 2022; revised 29th August 2022; accepted TBD. Date of Publication TBD.

M Gorman, X Ding, D Coyle, L Maguire are with the Intelligent Systems Research Centre (ISRC), School of Computing, Engineering and Intelligent Systems, Ulster University, Magee, Northern Ireland.

M Gorman is also with Seagate Technology in the Data Science Group (DSG), Springtown, Northern Ireland

Unsupervised ML AD approaches can benefit from 1d-CAE feature extraction [18]. However, a limitation of implementing the traditional ML techniques is the reduction of modelled feature space, creating further abstraction from the raw time-series trace data that decreases the interpretability of predictions. Here we propose a solution based on reconstruction error with 1d-CAE which preserves the entire trace for review by engineering teams.

II. Data Description

In this study, there are two datasets used to validate the proposed methodology. The Tennessee Eastman Process (TEP) and LAM 9600 Metal Etcher datasets are recognised benchmarks for AD in manufacturing experiments. The TEP dataset was initially introduced as an example for multivariate monitoring and control. For this experiment, fault classes 3, 9 and 15 have been removed as it has been previously reported that they lack observable deviation in mean or variance within the features [20]. Similarly, within the LAM 9600 Metal Etcher dataset, there are 19 input features along with varying process lengths. When comparing the 2 datasets, this dataset's contents are much closer to the process outputs observed in Semiconductor processing. Unlike the TEP dataset, the number of sample observations is limited. However, within the Metal Etcher dataset of 107 observations, there are 3 distinct experimental scenarios from which data is collected: Experiment 1 – 34 normal and 9 anomalies, Experiment 2 - 36 normal and 5 anomalies, Experiment 3 - 37 normal and 6 anomalies. Within the experiments, anomalies are determined based on an unsupervised thresholding method, however, retrospectively, predictions are assessed to determine the algorithm performance.

III. Methodology

A model training approach that facilitates the Subject Matter Expert (SME) to define a known good period of operation, also known as a ‘Golden Fingerprint’, in which the data sample accurately describes the intended mode of equipment operation. This approach is common practice within Semiconductor manufacturing and overcomes the need for detailed multi-class failure modes. The process of training the AE on known good process operation creates a novelty boundary i.e., an error threshold. The main assumptions within this framework are that the ‘Golden Fingerprint’ is an adequate representative sample that encompasses all ‘acceptable’ process variation and within future fault situations, error rates are sufficiently large for novel data and thus exceed the defined novelty boundary. Within the experiment, Nested Cross-Validation (nCV) is leveraged to optimise hyper-parameters and ensure generalisation performance is estimated thoroughly [21].

The 1d-CAE network is configured such that an encoder and decoder are symmetric in shape. The design forces the input shape to be condensed through 1d-Convolutional layers and reconstructed in the decoder using 1d-ConvolutionalTranspose Layers, where an identical shape to the input is produced. The shape dimensions are required to be identical so that loss can be measured across the network. Offline training generates an effective model that has learned the necessary novelty boundaries of known good observations and to determine appropriate threshold boundaries against which new observations are evaluated. Increased Convolutional kernel strides are preferred to pooling layers to reduce dimensionality and early stopping is implemented to monitor training loss. Train set normalisation weights, the trained model and the reconstruction error thresholds generated in the offline training process are preserved for online inference. At each Convolutional layer the Rectified Linear Activation Unit (ReLU) activation function is used. Adam is the chosen network optimiser which works to maintain parameter-specific learning rates that improve performance in sparse gradients whilst also taking per-parameter learning rates based on recent mean magnitudes of weight gradients resulting in a model that performs well in non-stationary scenarios.

Reconstruction error is the approach to calculating the difference between an input to a network and the reconstructed output. The lower the error in reconstruction the less likely it is that the given observation is an anomaly. Where the encoder g_ϕ and decoder f_θ are applied to input observations to generate a reconstruction from which relative errors can be derived. There are several methods to calculate reconstruction error, within this study Mean Absolute Error (MAE), Mean Squared Error (MSE) and Root Mean Square Error (RMSE) are evaluated. In the context of 1d-CAEs, each observation input column represents a different sensor, which has its specific error distribution. Condensing the reconstruction error into global representative statistics results in channel-specific error signals being diminished within the global error distribution. For this reason, a Localised Reconstruction Error (LRE) is preferred to Global Reconstruction Error (GRE) due to the independence given to each input channel or sensor when evaluating reconstruction

TABLE I
OUTER FOLD RESULTS FOR PROPOSED METHODOLOGY COMPARED TO SIMILAR STUDIES ON TEP DATASET

Ref	Network Type	Fault Labels Included	Metric	Result
Proposed	**1d-CAE with Localised Reconstruction Error**	**1,2,4,5,6,7,8,10,11,12,13,14,16,17,18,19,20**	**AUC**	***1.00***
[14]	1d-CAE	4,6,9,11,12,13	Accuracy	96.3
[19]	1d Convolutions & SDAE	4,5,9,11,14	Recognition Rate	93.29
[23]	Variational Recurrent AutoEncoder	1,2,4,6,7,8,10,11,12,13,14,15,16,17,18,19,20	Accuracy	96.3
[22]	2d Convolutions & Generative model	1,2,4,6,8,12	Accuracy	97.87

error thresholds. Each channel is evaluated separately, and subsequently, an aggregation of all channels occurs such that if any resulting channel returns an anomalous result the entire observation is registered as anomalous. Area Under the Curve (AUC) of the Receiver Operating Characteristics (ROC) is the preferred performance metric applied here.

IV. Experimental Results

Table I presents the experimental results of the proposed 1d-CAE with the LRE algorithm when compared to similar studies applying deep learning networks to the TEP dataset for FDC. Within the table, the general network style, chosen metric of evaluation and fault categories included in the experiment are compiled. While including more fault categories within the experiment the proposed method outperforms a 1d-CAE feature learning approach with a SoftMax classifier [14], semi-supervised deep generative model [22] and a stacked denoising autoencoder method [19] when comparing the rate of AD achieved. Furthermore, for a similar number of fault categories included in the experiment, the proposed 1d-CAE with LRE has achieved a higher AD rate than an approach based on variational recurrent autoencoders [23]. The result achieved here demonstrates that the proposed method generalises well across multiple fault scenarios when compared to alternative approaches. As the LAM 9600 experiment observation number is small when compared to the TEP dataset, this renders nCV impractical, therefore, the best hyperparameter settings that are obtained through the TEP experiment are used as the default settings in the LAM 9600 Etcher new training network. Table II details the hyperparameters which obtained the best results within the nCV experimental framework from the TEP dataset and were successfully used in the LAM etcher experiment. Overall, the results achieved across several parameter setups highlight that a smaller kernel size is preferred as well as the MAE reconstruction approach. Conversely, the Loss function has a negligible impact on the model performance, as shown in Table II. Comparing experiments applied to the LAM 9600 etcher dataset, the proposed 1d-CAE with LRE outperforms both a traditional ML framework [24] and a Bayesian approach [25] as shown in Table III. Solely using reconstruction errors facilitate the decomposition of actual inputs and predicted observations from the model to improve explainability. Figure 1 is a decomposition of the raw reconstruction errors for several samples which are both faults and normal observations from the TEP dataset. The decomposition demonstrates how the proposed architecture can be leveraged to improve prediction interpretability. In detail, fault 6 (plot label 6 in Figure 3) is a failure across several sensors and throughout the processing duration after the event of fault occurrence. However, in contrast, fault 4 is more isolated and occurs only on a subset of sensors. Furthermore, fault 5 is extremely subtle highlighting the approach's interpretability and sensitivity to different fault categories.

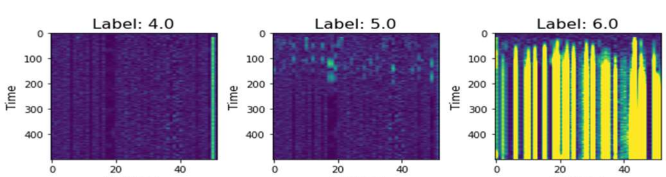

Figure 1 – Reconstruction Error Decomposition for a subset of normal and fault categories from the TEP dataset. In these examples, time and sensors are represented on the y and x axis respectively, with the colour representing the magnitude of reconstruction error

TABLE II
TEP INNER LOOP HYPER-PARAMETER MEAN AUC

Optimiser	Learning Rate	Kernel Size	Loss	Reconstruction Method					
				byChannel			byObservation		
				Reconstruction Error Equation					
				mae	mse	rmse	mae	mse	rmse
Adam	0.00001	10	mae	0.938	0.952	0.937	0.845	0.918	0.908
			mse	0.937	0.951	0.937	0.833	0.903	0.887
		100	mae	0.955	0.969	0.950	0.864	0.935	0.932
			mse	0.949	0.968	0.948	0.868	0.936	0.934
	0.0001	10	mae	***0.974***	0.968	0.972	0.902	0.940	0.937
			mse	***0.974***	0.970	***0.974***	0.904	0.941	0.939
		100	mae	0.965	0.963	0.962	0.842	0.921	0.912
			mse	0.959	0.963	0.957	0.866	0.928	0.924
	0.001	10	mae	0.972	0.967	0.971	0.910	0.941	0.939
			mse	0.973	0.969	0.973	0.914	0.942	0.940
		100	mae	0.933	0.946	0.935	0.758	0.846	0.812
			mse	0.923	0.949	0.923	0.714	0.827	0.768

TABLE III
MEAN AUC ACHIEVED FOR THE 3 EXPERIMENTAL SCENARIOS WITHIN THE LAM 9600 METAL ETCHER DATASET

Ref	Approach	Metric	Result
Proposed	Proposed Method	AUC	**1.00**
[24]	Traditional ML	AUC	0.94
[26]	GMM & Bayesian Method	Accuracy	0.90

V. Discussion

In this study, we demonstrate the 1d-CAE DL algorithm combined with local error reconstruction outperforms alternative approaches benchmarks in batch manufacturing. Considering algorithm performance, maximum accuracies can be achieved using both LRE and GRE, as shown in Tables II, however, improved robustness is demonstrated when byChannel LRE is used. Furthermore, the reconstruction error method preserves the point-to-point error rates providing the opportunity to review detailed predictions thus creating a transparent model. Interpretability is addressed in several of the comparative studies, however, only at the global scale through the clustering of latent network node weights [14], [19]. To encourage further solution adoption, predicted observation interpretability is required. The proposed approach leverages reconstruction errors, shown in figure 1 as a core component that identifies the regions of a higher error rate by observation. Similar studies that implement a classification approach are restricted in the number of anomalies that can be detected by the model. However, the proposed thresholding approach is suited to infinite fault scenarios as there is no prescribed fingerprint of a fault, only that the defined anomaly threshold has been breached. Also, the reconstruction error decomposition facilitates the requirement that future features of importance remain in the domain space. Model training on a 'golden fingerprint' allows the network to effectively generate the novelty boundary that facilitates infinite fault category generalisation for future unknown fault scenarios observed in the dataset.

VI. CONCLUSIONS AND FURTHER WORK

This study has demonstrated that the 1d-CAE algorithm applied with LRE for anomaly detection outperforms existing studies that use the TEP and LAM 9600 datasets. Using the experimental paradigm, the models trained using known 'Golden Fingerprint' observations can effectively learn a novelty boundary, outside of which anomalous operating scenarios can be identified. Furthermore, the study has demonstrated that AD model explainability can be improved greatly using the local reconstruction error methodology. The LRE approach provides independence to each feature when determining the observation result whilst maintaining the granularity required to decompose features along the time-dependent axis so that consumers of the algorithm results can better understand the mechanistic reason for the fault from a process perspective. The improved explainability accompanied by the unsupervised algorithm configuration framework results in a solution for model development with a potential for reduced cost of ownership to engineering teams in a production factory deployment environment. Further experimentation is required to explore the impact of various contamination percentages of anomalies in the training data. Similarly, the base assumption of reconstruction error-based methods is the anomalous behaviour far enough deviates from the training data to break the error threshold. A potential issue for Semiconductor AD applications is the seasonality, non-stationary and multi-modal data scenarios introduced through natural process variation and consumable equipment parts ageing. Some of this variation can be accounted for within the model, however, large deviations from the original training data will increase the likelihood of an increased false positive rate.

VII. REFERENCES

[1] Z. Ge, Z. Song, S. X. Ding, and B. Huang, "Data Mining and Analytics in the Process Industry: The Role of Machine Learning," *IEEE Access*, vol. 5, pp. 20590–20616, Sep. 2017, doi: 10.1109/ACCESS.2017.2756872.

[2] J. Moyne and J. Iskandar, "Big Data Analytics for Smart Manufacturing: Case Studies in Semiconductor Manufacturing," *Processes*, vol. 5, no. 3, p. 39, 2017, doi: 10.3390/pr5030039.

[3] T. Lee and C. O. Kim, "Statistical comparison of fault detection models for semiconductor manufacturing processes," *IEEE Trans. Semicond. Manuf.*, vol. 28, no. 1, pp. 80–91, 2015, doi: 10.1109/TSM.2014.2378796.

[4] J. Wang, Y. Ma, L. Zhang, R. X. Gao, and D. Wu, "Deep learning for smart manufacturing: Methods and applications," *J. Manuf. Syst.*, vol. 48, pp. 144–156, Jul. 2018, doi: 10.1016/j.jmsy.2018.01.003.

[5] G. A. Susto, M. Terzi, and A. Beghi, "Anomaly Detection Approaches for Semiconductor Manufacturing," *Procedia Manuf.*, vol. 11, no. June 2017, pp. 2018–2024, 2017, doi: 10.1016/j.promfg.2017.07.353.

[6] G. Verdier and A. Ferreira, "Adaptive mahalanobis distance and k-nearest neighbor rule for fault detection in semiconductor manufacturing," *IEEE Trans. Semicond. Manuf.*, vol. 24, no. 1, pp. 59–68, 2011.

[7] J. Wang, J. Feng, and Z. Han, "Discriminative Feature Selection Based on Imbalance SVDD for Fault Detection of Semiconductor Manufacturing Processes," *J. Circuits, Syst. Comput.*, vol. 25, no. 11, pp. 1–21, 2016, doi: 10.1142/S0218126616501437.

[8] K. Kerdprasop and N. Kerdprasop, "A data mining approach to automate fault detection model development in the semiconductor manufacturing process," *Int. J. Mech.*, vol. 5, no. 4, pp. 336–344, 2011, doi: 10.4028/www.scientific.net/AMM.52-54.1171.

[9] H. Rostami, J. Blue, and C. Yugma, "Automatic equipment fault fingerprint extraction for the fault diagnostic on the batch process data," *Appl. Soft Comput. J.*, vol. 68, pp. 972–989, 2018, doi: 10.1016/j.asoc.2017.10.029.

[10] K. Grace, J. Salvatier, A. Dafoe, B. Zhang, and O. Evans, "Viewpoint: When will ai exceed human performance? Evidence from ai experts," *Journal of Artificial Intelligence Research*, vol. 62. pp. 729–754, 2018, doi: 10.1613/jair.1.11222.

[11] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science (80-.).*, vol. 313, no. 5786, pp. 504–507, Jul. 2006, doi: 10.1126/science.1127647.

[12] J. Y. Xifeng Guo, Xinwang Liu, En Zhu, "Deep Clustering with Convolutional Autoencoders," *Conf. Pap.*, no. October, pp. 118–125, 2017, doi: https://doi.org/10.1007/978-3-319-70096-0_39.

[13] H. Lee, Y. Kim, and C. O. Kim, "A deep learning model for robust wafer fault monitoring with sensor measurement noise," *IEEE Trans. Semicond. Manuf.*, vol. 30, no. 1, pp. 23–31, Feb. 2017, doi: 10.1109/TSM.2016.2628865.

[14] S. Chen, J. Yu, and S. Wang, "One-dimensional convolutional auto-encoder-based feature learning for fault diagnosis of multivariate processes," *J. Process Control*, vol. 87, pp. 54–67, 2020, doi: 10.1016/j.jprocont.2020.01.004.

[15] M. Macas and C. Wu, "An unsupervised framework for anomaly detection in a water treatment system," *Proc. - 18th IEEE Int. Conf. Mach. Learn. Appl. ICMLA 2019*, no. 15027268, pp. 1298–1305, 2019, doi: 10.1109/ICMLA.2019.00212.

[16] S. Kiranyaz, T. Ince, and M. Gabbouj, "Real-Time Patient-Specific ECG Classification by 1-D Convolutional Neural Networks," *IEEE Trans. Biomed. Eng.*, vol. 63, no. 3, pp. 664–675, 2016, doi: 10.1109/TBME.2015.2468589.

[17] E. Kim, S. Cho, B. Lee, and M. Cho, "Fault Detection and Diagnosis Using Self-Attentive Convolutional Neural Networks for Variable-Length Sensor Data in Semiconductor Manufacturing," *IEEE Trans. Semicond. Manuf.*, vol. 32, no. 3, pp. 302–309, Aug. 2019, doi: 10.1109/TSM.2019.2917521.

[18] M. Maggipinto, A. Beghi, and G. A. Susto, "A deep learning-based approach to anomaly detection with 2-dimensional data in manufacturing," *IEEE Int. Conf. Ind. Informatics*, vol. 2019-July, pp. 187–192, 2019, doi: 10.1109/INDIN41052.2019.8972027.

[19] C. Zhang, J. Yu, and S. Wang, "Fault detection and recognition of multivariate process based on feature learning of one-dimensional convolutional neural network and stacked denoised autoencoder," *Int. J. Prod. Res.*, 2020, doi: 10.1080/00207543.2020.1733701.

[20] S. Heo and J. H. Lee, "Fault detection and classification using artificial neural networks," *IFAC-PapersOnLine*, vol. 51, no. 18, pp. 470–475, Jan. 2018, doi: 10.1016/j.ifacol.2018.09.380.

[21] D. Krstajic, L. J. Buturovic, D. E. Leahy, and S. Thomas, "Cross-validation pitfalls when selecting and assessing regression and classification models," *J. Cheminform.*, vol. 6, no. 1, p. 10, Mar. 2014, doi: 10.1186/1758-2946-6-10.

[22] T. Ko and H. Kim, "Fault Classification in High-Dimensional Complex Processes Using Semi-Supervised Deep Convolutional Generative Models," *IEEE Trans. Ind. Informatics*, vol. 16, no. 4, pp. 2868–2877, 2020, doi: 10.1109/TII.2019.2941486.

[23] F. Cheng, Q. P. He, and J. Zhao, "A novel process monitoring approach based on variational recurrent autoencoder," *Comput. Chem. Eng.*, vol. 129, Oct. 2019, doi: 10.1016/j.compchemeng.2019.106515.

[24] T. Lee, K. B. Lee, and C. O. Kim, "Performance of Machine Learning Algorithms for Class-Imbalanced Process Fault Detection Problems," *IEEE Trans. Semicond. Manuf.*, vol. 29, no. 4, pp. 436–445, 2016, doi: 10.1109/TSM.2016.2602226.

[25] J. M. J. M. Nawaz, M. Z. M. Z. Arshad, and S. J. S. J. Hong, "Fault diagnosis in semiconductor etch equipment using Bayesian networks," *J. Semicond. Technol. Sci.*, vol. 14, no. 2, pp. 252–261, 2014, doi: 10.5573/JSTS.2014.14.2.252.

[26] J. Yu, "Semiconductor manufacturing process monitoring using gaussian mixture model and Bayesian method with local and nonlocal information," *IEEE Trans. Semicond. Manuf.*, vol. 25, no. 3, pp. 480–493, 2012, doi: 10.1109/TSM.2012.2192945.