MACHINE LEARNING BASED LOG-ANALYSIS FOR AUTOMATED ANOMALY DETECTION

Narinder Gupta
Guru Kashi University, Talwandi Sabo

ABSTRACT

Many sensors are used in a single production process, making it difficult to pinpoint the exact source of a problem. More than one process cycle is required to make a semiconductor wafer. There are many cycles in this process, and it is difficult to spot abnormalities in time, thus the process continues until it is complete. The cost of producing these wafers is high, and a process failure can have a significant impact on both time and money. As a result, anomaly detection in semiconductor production can benefit greatly from machine learning. A manufacturing facility may interrupt the operation and fix the problematic equipment if irregularities in the production process could be discovered or predicted sooner. As a result, semiconductor producers would see an improvement in process yield and a reduction in expenses.

Keywords: machine learning, based log-analysis, automated, anomaly detection

I. Introduction

Machine learning and artificial intelligence can be applied to manufacturing, particularly in the field of semiconductor manufacturing. Many sensors monitor the manufacturing process and the semiconductors produced in semiconductor manufacturing facilities. A company's production process may be further optimised by using the data provided by sensors on various equipment. This data is currently solely utilised for debugging when an issue occurs.

II. Aims

Analog Devices' semiconductor production data is the major subject of this investigation (ADI). Although ADI is now implementing SPC and limit monitoring in their manufacturing process, they have had minimal success. Out-of-control processes and temporal irregularities can't be detected using these techniques very often (Chen, *et al.* 2021). For this reason, it's

impossible to define separate limitations for each data channel since the manufacturing process includes so many recipes and parameters. Many sensors are used in a single semi-conductor manufacturing process, making it impossible to keep track of all the information. A reactive rather than proactive strategy to anomalous occurrences has been adopted by ADI, relying on the data to investigate issues after they have happened rather than flagging and studying abnormalities as they occur. Anomaly parameters are notoriously difficult to pin down by hand, even if they are found.

III. Objectives

At ADI, the existing anomaly detection protocol gives a great opportunity for the implementation of machine learning-based anomaly detection techniques. We are interested in seeing if a pipeline can be built to automatically recognise unusual occurrences. Models that can be applied to a wide range of machines and recipes require a modest bit of domain information in the form of tagged data. Analog Devices' anomaly detection procedures are detailed in detail in Chapter 2 of this thesis, which is a consolidation and development of three earlier theses. We want to expand on these earlier research to create a real-time anomaly detection system. As it is now, a lot of flaws in a manufacturing process may go unnoticed until the entire process is complete; this may take anywhere from a few hours, to days, or even weeks, depending on the process. Consequently, real-time analysis has the potential to save both time and money by allowing early termination of the process.

IV. Research Questions

1. What constitutes "normal" conduct, and how wide does it extend?
2. Is there a way to tell the difference between normal conduct and aberrant behaviour?
3. What is the threshold for deeming a person an anomaly?

V. Literature Review

For the purpose of better comprehending the anomaly detection problem before providing literature on cluster analysis and time series forecasting, this chapter conducts a literature study. In addition, a number of prediction models from the literature are discussed.

VI. Anomaly Detection

Time series data anomalies are described as points or sequences of points that depart from the expected behaviour of the data. Manufacturing, economics, transportation, and health care are just few of the areas where anomaly identification is a challenge. No one answer exists for anomaly identification since various domains may have different definitions of

what constitutes an abnormality. Time-series anomaly detection using machine learning has been attempted in both an unsupervised and semi-supervised manner. According to (Bhanage, *et al.* 2021), the quantity of information about the data provided, the level of monitoring is calculated. As an example, supervised learning may be used to a dataset that has already been labelled with information. Unsupervised or semi-supervised learning approaches are needed to categorise data that contains minimal or no labelling. When it comes to detecting novel anomalous patterns, supervised learning models generally fall short, even while they can recognise previously known anomalies. Furthermore, because anomalies in data occur (ideally) infrequently, the distribution of anomalous against normal samples is imbalanced.

Clustering approaches may be utilised for unsupervised anomaly detection, whereas neural network models can be employed for supervised learning. Anomalies can show themselves in a variety of ways. Extreme numbers or outliers that are outside of the process's normal range are the simplest anomalies to spot. Each sensor channel may be set to automatically identify these abnormalities if the defined threshold is breached (Bao, *et al.* 2018). For instance, it's difficult to deal with anomalies that occur within a regular operating range but don't follow the usual pattern of time. They occur often in production contexts and can be difficult to spot using SPC approaches, which limit monitoring.

VII. Methodology

Following a literature study, popular anomaly detection methods are discussed in this thesis. Analog Devices' earlier research on time-series averaging, cluster analysis, and time-series forecasting is incorporated into this section as well. Finally, we look at the datasets for anomaly detection and assess their attributes. An automated model for anomaly detection is then created and tested to see if it can appropriately identify unusual occurrences. Training our algorithm to recognise regular wafer cycles and predicting one step ahead of those cycles is the first stage in this process. Accuracy in recognising abnormalities is next assessed using the model. Experiments are conducted on a variety of process recipes, and the model's prediction abilities are evaluated. Over time rather than one time-step, this model may forecast abnormal trends in data. Anomaly detection is the subject of this chapter, which focuses on the precise methodologies we apply (Fredriksson Franzén, & Tyrén, 2021). To begin, the time-series averaging methods used to build the reference cycle are discussed. The clustering algorithms and their distinctions are next reviewed. MLP and LSTM

implementations are then described in detail. Lastly, a high-level look at our whole system for automatically detecting anomalies is provided.

VIII. Structural Time Distortion

When two time series are compared using a distance measure, Dynamic Time Warping (DTW) is used to calculate and compare the dissimilarity between them. Both series are more comparable with lower DTW distances. In order to minimise the DTW distance between two time series and map one onto the other, it repeatedly warps two time series (Du, *et al.* 2021). It first produces an n-by-m distance matrix for two time series, A = (a1, a2, an) and B = (b1, b2, bm), with lengths of n and m, respectively (Shin, *et al.* 2021).

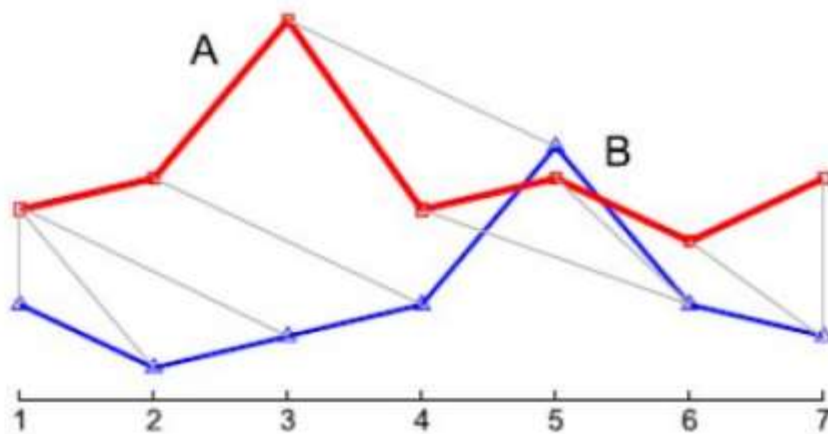

Figure 1: Mapping between Time-Series A and B

Summation of distances between A and B points, as well as three minimum distances around an element of interest (I, J), is Yi,j in this instance. P is the |ai bj |-norms's dimension. So that the Euclidean distance between two locations is employed, p is typically set to 2. Yi,j comes up with the ultimate answer to the question of how far apart the two series are. Alignment is depicted in Figure 4 where a set of two sets of data, one set of two sets of data, and one set of two sets of data are aligned. As the most often used measure of sequence dissimilarity, DTW can discover the most optimum global alignment between series (Shao, *et al.* 2020). When employing DTW, the two time series need not be equivalent lengths in order to benefit.

IX. Data characteristics

The plasma etching data is broken down into wafer cycles, each of which represents a single run of the plasma etcher. Either 600 or 300 timesteps make to a cycle (Recipe 920). (Recipe

945). For a particular parameter, Figure 8 demonstrates the difference between the two recipes. The biggest difference is the duration. For our studies, we eliminated six of the 31 total parameters from the dataset since they showed no fluctuation over the whole dataset (Yang, *et al.* 2021). Due to this, our models will not be as accurate as they might be if these characteristics were included. Anomalies and drift in the data are qualities we must keep in mind while developing our models.

XI. Gantt Chart

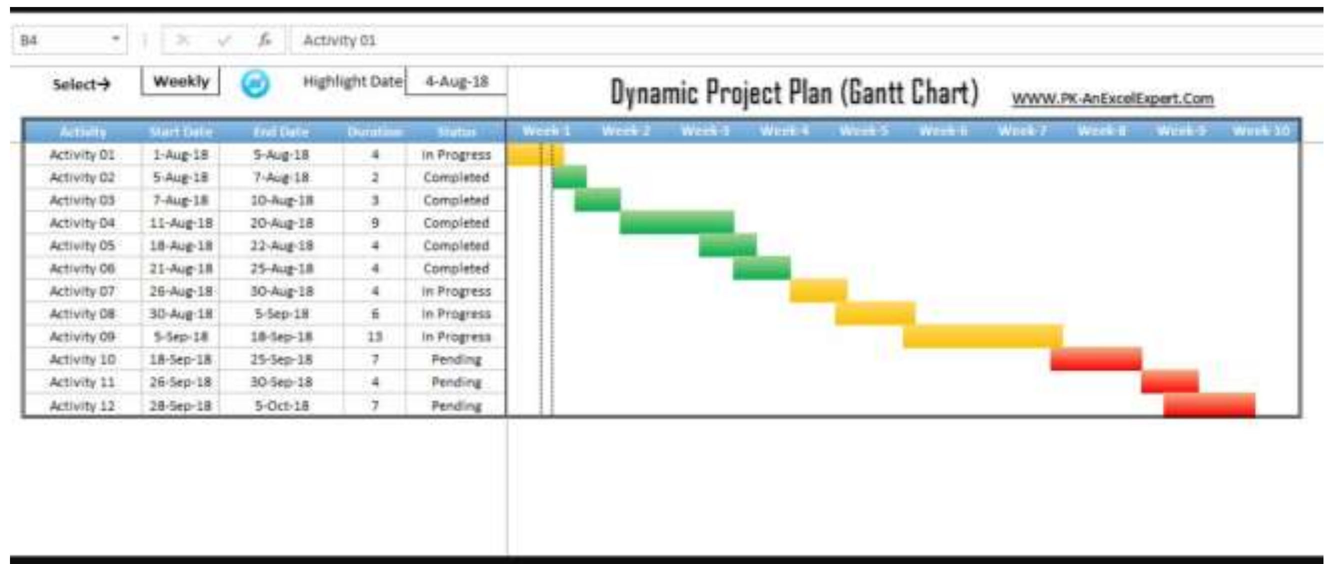

Figure 2: Gantt chart

(Source: Self-created)

XII. Conclusion

An automated pipeline for semiconductor manufacturing anomaly detection has been presented in this thesis. To begin, we create a reference cycle that represents an average time series of good and non-anomalous wafer cycles; we then perform cluster analysis on unlabeled wafer cycle data to identify normal and abnormal cycles; and finally we train neural networks to detect anomalous time-steps in a single cycle. We've demonstrated in our trials that even with only a limited quantity of labelled data, our pipeline is capable of detecting aberrant timesteps. Since our model relies on very little domain information, we may use an unsupervised technique to train it. It is our aim that our approach may be used to a wide range of industries, not simply those involved in the production of semiconductors. There are several possibilities for future study based on this work. To find out how much time is left till a future occurrence, statisticians and machine learning experts turn to a topic

known as survival analysis. Survival analysis can benefit from models other than neural networks, such as random forests and Bayesian approaches. The semiconductor manufacturing process might benefit from models like this. However, in order to fully train models for survival analysis, we would need a significantly bigger dataset of failure occurrences.

XIII. References

Astekin, M., Özcan, S., & Sözer, H. (2019, December). Incremental analysis of large-scale system logs for anomaly detection. In *2019 IEEE International Conference on Big Data (Big Data)* (pp. 2119-2127). IEEE. https://ieeexplore.ieee.org/abstract/document/9006593/?casa_token=hxpw6mTJT3wAAAAA:ECu2SCHL2tEKSVVouhjmG9wvqOGc1RwecRT5iJlieiLXzN8SxEQRIE__93o76-feOKZvbvPXwJk- [Accessed on 30-11-2021]

Bao, L., Li, Q., Lu, P., Lu, J., Ruan, T., & Zhang, K. (2018). Execution anomaly detection in large-scale systems through console log analysis. *Journal of Systems and Software*, *143*, 172-186. https://www.sciencedirect.com/science/article/pii/S0164121218301031?casa_token=b3VD5kVjYFAAAAAA:Cua2Ru72vJh13GDLrxedhJxWEaqfxjgYKI4Q3Z3YjXQTIhMlMcTZ5SZIw7zJvLvffChmE3WwarCt [Accessed on 30-11-2021]

Bhanage, D. A., Pawar, A. V., & Kotecha, K. (2021). IT Infrastructure Anomaly Detection and Failure Handling: A Systematic Literature Review Focusing on Datasets, Log Preprocessing, Machine & Deep Learning Approaches and Automated Tool. *IEEE Access*. https://ieeexplore.ieee.org/abstract/document/9615039/ [Accessed on 30-11-2021]

Cao, Q., Qiao, Y., & Lyu, Z. (2017, December). Machine learning to detect anomalies in web log analysis. In *2017 3rd IEEE International Conference on Computer and Communications (ICCC)* (pp. 519-523). IEEE. https://ieeexplore.ieee.org/abstract/document/8322600/?casa_token=8B_etk_kK6gAAAAA:mqJylYHUudSJcUUqux8-wnKTkRf_YaJl7Db4iIrJR9yXhYcUO1YXEQftBTnx5xKvpbvSYhVVKn8F [Accessed on 30-11-2021]

Chen, Z., Liu, J., Gu, W., Su, Y., & Lyu, M. R. (2021). Experience Report: Deep Learning-based System Log Analysis for Anomaly Detection. *arXiv preprint arXiv:2107.05908*. https://arxiv.org/abs/2107.05908 [Accessed on 30-11-2021]

Debnath, B., Solaimani, M., Gulzar, M. A. G., Arora, N., Lumezanu, C., Xu, J., ... & Khan, L. (2018, July). LogLens: A real-time log analysis system. In *2018 IEEE 38th international conference on distributed computing systems (ICDCS)* (pp. 1052-1062). IEEE. https://ieeexplore.ieee.org/abstract/document/8416368/?casa_token=IGIQ0ao1T4cAAAAA:bhijQZWWOino9I5C_DOI6ZU0qPejpVKxVeu-qkyd_hEk6ApUZgnZX1R-Eet54dl189pdZ4zAPEl9 [Accessed on 30-11-2021]

Du, Q., Zhao, L., Xu, J., Han, Y., & Zhang, S. (2021, September). Log-Based Anomaly Detection with Multi-Head Scaled Dot-Product Attention Mechanism. In *International Conference on Database and Expert Systems Applications* (pp. 335-347). Springer, Cham. https://link.springer.com/chapter/10.1007/978-3-030-86472-9_31 [Accessed on 30-11-2021]

Fredriksson Franzén, M., & Tyrén, N. (2021). Anomaly detection for automated security log analysis: Comparison of existing techniques and tools. https://www.diva-portal.org/smash/record.jsf?pid=diva2:1576656 [Accessed on 30-11-2021]

Hirakawa, R., Uchida, H., Nakano, A., Tominaga, K., & Nakatoh, Y. (2021). Anomaly detection on software log based on Temporal Memory. *Computers & Electrical Engineering*, *95*, 107433. https://www.sciencedirect.com/science/article/pii/S0045790621003943?casa_token=cfjgtm_Ev2MAAAAA:o63T4sPJTD8IhCbcLa3fQmtuMuRMAndlcJeyrENIZJ3Q7R-eqSR-DbznKs2n-wUVq8g5aloZUdre [Accessed on 30-11-2021]

Hirakawa, R., Uchida, H., Nakano, A., Tominaga, K., & Nakatoh, Y. (2021). Large Scale Log Anomaly Detection via Spatial Pooling1. *Cognitive Robotics*. https://www.sciencedirect.com/science/article/pii/S2667241321000173 [Accessed on 30-11-2021]

Hong, J., Park, S., Yoo, J. H., & Hong, J. W. K. (2020, November). Machine Learning based SLA-Aware VNF Anomaly Detection for Virtual Network Management. In *2020 16th International Conference on Network and Service Management (CNSM)* (pp. 1-

7). IEEE. https://ieeexplore.ieee.org/abstract/document/9269100/?casa_token=wytAnMa6IpsAAAAA:iW4nmieLE8C1VaVkrZ9fom3Qr6YBgnDDrC5nhr2RGdxr2hvtEYIBGeY_nMCvP9uHqvgJi9P4ObhC [Accessed on 30-11-2021]

Schmidt, T., Hauer, F., & Pretschner, A. (2020, September). Automated Anomaly Detection in CPS Log Files. In *International Conference on Computer Safety, Reliability, and Security* (pp. 179-194). Springer, Cham. https://link.springer.com/chapter/10.1007/978-3-030-54549-9_12 [Accessed on 30-11-2021]

Shao, W., Wang, Z., Wang, X., Qiu, K., Jia, C., & Jiang, C. (2020). LSC: Online auto-update smart contracts for fortifying blockchain-based log systems. *Information Sciences*, *512*, 506-517. https://www.sciencedirect.com/science/article/pii/S0020025519309260?casa_token=Zu98Df62yIkAAAAA:ZDpMsndd1Z8g0e3t1S0vB9ZJFYLIcwYESOq2GzjoZdVCtF6iyH5bRhZKdgcUIGglw68ZdjG9Jj8r [Accessed on 30-11-2021]

Shin, D., Khan, Z. A., Bianculli, D., & Briand, L. (2021, October). A Theoretical Framework for Understanding the Relationship Between Log Parsing and Anomaly Detection. In *International Conference on Runtime Verification* (pp. 277-287). Springer, Cham. https://link.springer.com/chapter/10.1007/978-3-030-88494-9_16 [Accessed on 30-11-2021]

Skopik, F., Wurzenberger, M., & Landauer, M. (2021). Smart Log Data Analytics: Techniques for Advanced Security Analysis. Skopik, F., Wurzenberger, M., & Landauer, M. (2021). Smart Log Data Analytics: Techniques for Advanced Security Analysis. [Accessed on 30-11-2021]

Tallón-Ballesteros, A. J., & Chen, C. (2020). Explainable AI: Using shapley value to explain complex anomaly detection ML-based systems. *Machine Learning and Artificial Intelligence: Proceedings of MLIS 2020*, *332*, 152. https://books.google.com/books?hl=en&lr=&id=hq4SEAAAQBAJ&oi=fnd&pg=PA152&dq=MACHINE+LEARNING+BASED+LOG-ANALYSIS+FOR+AUTOMATED+ANOMALY+DETECTION&ots=2bZzHPW7k-&sig=4IodPdO_0eiKyt_DEFC7Zm8diYM [Accessed on 30-11-2021]

Wang, Z., Tian, J., Fang, H., Chen, L., & Qin, J. (2021). LightLog: A lightweight temporal convolutional network for log anomaly detection on the edge. *Computer Networks*, 108616. https://www.sciencedirect.com/science/article/pii/S1389128621005119?casa_token=heZpk7kfOZsAAAAA:-7ix1Wgc3Ww8JrXNhj6wGccuDeldgKXZ-Yu8u1tH7oFfMUTjRGnRmfNtRaU-h3udOhCC5i8Zxtg3 [Accessed on 30-11-2021]

Yadav, R. B., Kumar, P. S., & Dhavale, S. V. (2020, June). A Survey on Log Anomaly Detection using Deep Learning. In *2020 8th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions)(ICRITO)* (pp. 1215-1220). IEEE. https://ieeexplore.ieee.org/abstract/document/9197818/ [Accessed on 30-11-2021]

Yang, L., Chen, J., Wang, Z., Wang, W., Jiang, J., Dong, X., & Zhang, W. (2021, May). Semi-supervised log-based anomaly detection via probabilistic label estimation. In *2021 IEEE/ACM 43rd International Conference on Software Engineering (ICSE)* (pp. 1448-1460). IEEE. https://ieeexplore.ieee.org/abstract/document/9401970/?casa_token=2OuRv9Pn0oYAAAAA:cAaQy0F9zpyyakMO8EmgRE7DtuQOKkqWOrYlorc_sgMhPYBl-cNapgXFCSAaaCG_vJapp4rQJCci [Accessed on 30-11-2021]