Data Analytics and Visualization for Manufacturing Systems

Xinran Shi[1], Pengfei Chen[1], Yuchen Wen[1], Xiaowei Yue[1], and Yuqi Cao[2]
[1]School of Industrial and Systems Engineering
[2]School of Music

1. Introduction

1.1 Motivation

To achieve high-quality production monitoring, data visualization tools such as statistical control charts have been developed and applied to manufacturing systems (MFGSs) [1]. However, current data analytics and visualization tools in MFGSs only focused on low-dimensional small data. As the development of sensing technology, a large amount of in-line data can be generated from the sophisticated MFGSs [2]. Our team will try to bridge the gap between data analytics/visualization and advanced manufacturing. By integration of analytics/visualization and smart manufacturing, as shown in Fig. 1, we can make the MFGSs more informative and user-friendly.

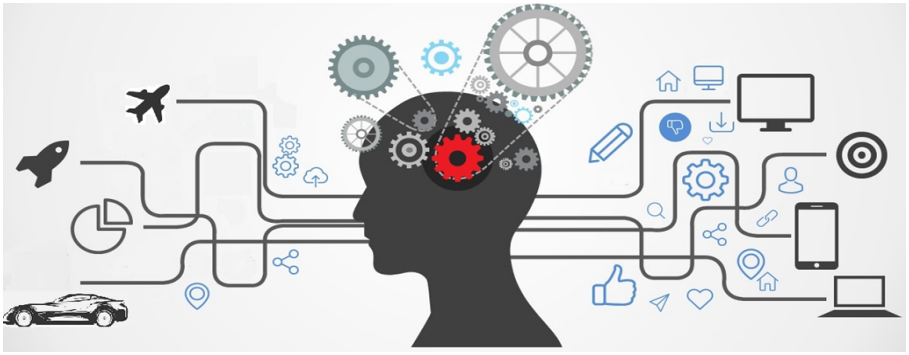

Fig. 1 Manufacturing systems ⇒Data analytics ⇒Data visualization

1.2 Problem definition

The objective of this research is to develop a Big Data analytics and visualization framework for complex manufacturing systems through systematic and deep integration of data analytics, visualization and manufacturing domain knowledge. The project consists of three modules, as shown in Fig. 2:

- Data collection, management, cleaning and preliminary analysis.
- Feature learning, correlation analysis, hierarchical clustering and predictive modeling.
- User-friendly interface allowing the user, with few data visualization/analytics skills, to interact with datasets and make intelligent decisions.

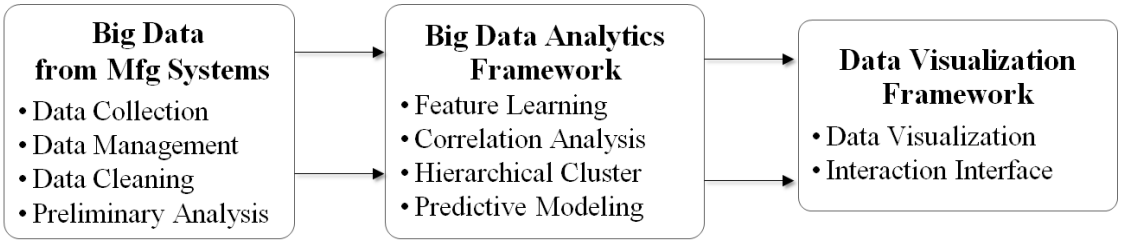

Fig. 2 Research modules

1.3 Data description

We used the Bosch datasets from Kaggle [3] (Table 1). Bosch records data at every step along its assembly lines, they have the potential to apply advanced big data analytics/visualization to improve these manufacturing processes. However, the intricacies of the data and complexities of the production line pose big challenges.

Table 1. Manufacturing datasets from Bosch production lines

Data file	Description	Size
train_date.csv	Training set date features	2.7 GB
train_numeric.csv	Training set numeric features	2.0 GB
train_categorical.csv	Training set categorical features	2.5 GB

From the dataset, there are more than one million parts samples, four production lines, 52 sections and more than two thousand features. For example, L3_S36_F3939 is a feature measured on line 3, station 36, and is feature number 3939. The physical meaning of production line, section and feature are not given due to confidential issue.

2. Survey

Data analytics and visualization is very important in manufacturing. For analytics, Wang and Mcgreavy proposed an automatic classification method for operation modes, but it is not suitable for failure prediction and correlation analysis [4]. Tseng proposed a data mining technique called rough set theory for quality assurance, but it is for machine process instead of multistage manufacturing like bosch [5]. Koonce and Tsai developed a data mining algorithm to find patterns in generic algorithms, which can be used to scheduling instead of multistage manufacturing [6]. Braha and Shmilovici used decision tree and neural network for advanced wafer cleaning, but the features were different to visualize [7]. Wang et al. proposed hierarchical clustering for defects classification, but it did not consider the correlation relationship among similar defects [8]. Researchers used a functional manifold model to regularize features and characterize cross-correlations [9], but this method was only suitable for functional data. Functional PCA [10], graphical LASSO [11] and sparse subspace clustering [12] methods were applied to describe conditional correlations among multiple variables. However, these methods cannot handle dynamic cross-correlations, which are very common in complex manufacturing systems. In our paper, we will explore the dynamic correlation maps during manufacturing processes. For more paper about data mining and analytics in manufacturing, please refer to [13]. Harding and Tiwari reviewed the progress of data mining in manufacturing [13], but they did not integrate data analytics with visualization. For visualization, there is a big gap between visualization experts and manufacturing engineers. Current methods such as statistical process control (SPC) can only monitor the small number of features as shown in Fig. 3 [14], which fails in Big data systems. Macgregor et al. proposed building multiple control charts to monitor manufacturing, while this kind of visualization lacks efficient feature learning and it is hard for engineers to do intelligent decision making [15]. Some visualization tools are provided for industrial analytics [16] and large scale manufacturing data learning [17], however, they did not have interactive

capability. Rohrer proposed visualization using 2D and 3D animation, but this visualization only works for manufacturing simulation instead of real manufacturing process [18]. Lindskog et al. [19] proposed visualization support for virtual design of manufacturing, however, it is an off-line visualization and not suitable for in-line monitoring.

In summary, existing approaches to analytics and visualization of smart manufacturing suffer from many drawbacks. Firstly, the feature learning does not consider the complex relationship among multi-stations; Secondly, few techniques consider dynamic correlation maps; Thirdly, the manufacturing visualization mainly shows the small data (limited key features). New data analytics integration with interactive visualization are highly required.

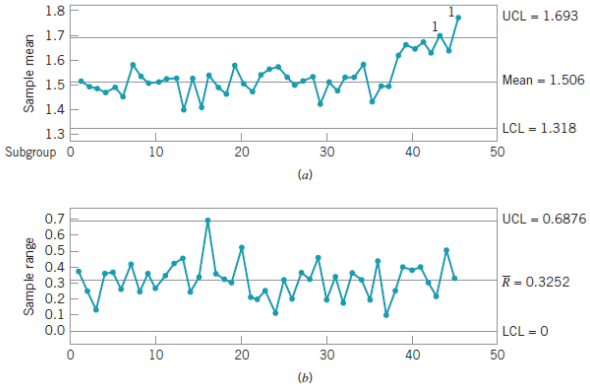

Fig. 3 Current SPC data visualization in MFGSs [14]

3. Proposed method

3.1 Intuition

For analytics, existing analysis mainly focuses on individual station [11, 12], but lack of systematic information fusion among multiple stations. Although some researchers did root-cause analysis and correlation analysis for industrial cases, none of them are lying on such a big size of data as we do. The big data in MFGSs brings more information about the system but increases the difficulty of understanding and illustration. Our team uses visualization techniques to remove the barrier between data analytics and engineer-wise practice.

Our visualization design has two contributions. Firstly, we bridge the gap between data analytics and data visualization in manufacturing applications and propose methodologies for deeply integrating these two types of techniques. Secondly, manufacturing data is not well-studied in the visualization field. For example, the spatial-temporal correlation is one of the most important characteristics of MFGSs, which is associated with not only datasets, but also physical structures. However, existing visualization tools did not provide this analysis. Our visualization tool shows the entire production line and our algorithm includes both numerical and time features.

3.2 Data cleaning for path identification

OpenRefine failed to clean the dataset due to the huge data size. We then turned to R and managed to extract paths and clean the data. Fig. 4 show the logic flow of this step.

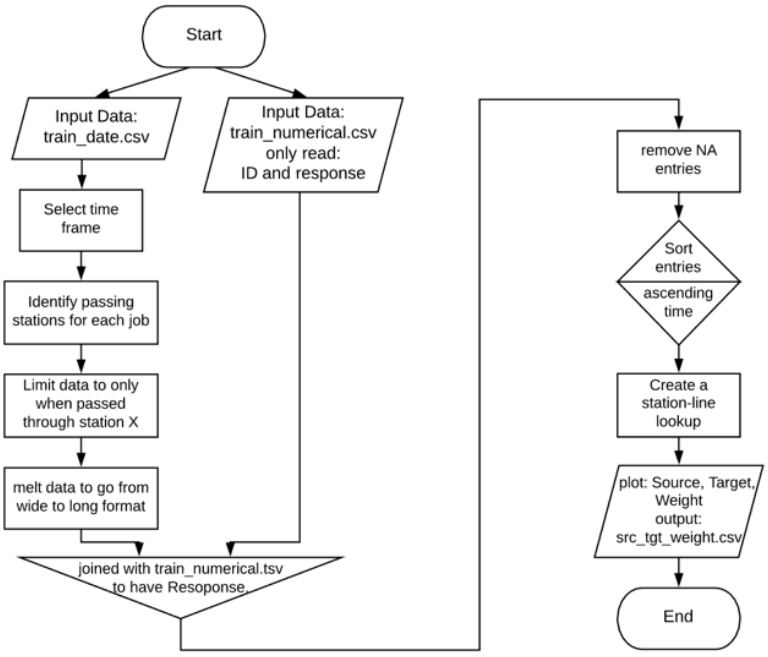

Fig. 4 Logic flow for path identification

3.3 Station-level data analysis and visualization

We use Microsoft Azure Storage container to store our data, and Microsoft Azure machine learning studio as the primary tool to analyze our data. For each station, we examine its features' correlation, and the correlation map is acquirable from the user interface. The results of correlation analysis are shown in Section 4.1. From the correlation analysis, we could see that each two neighbor features are strongly correlated. By comparing the similarity [13], we found that several features are similar. Therefore, we perform dimension reduction by generating new features by integrating two highly-correlated features. Since the response variable of the dataset is highly imbalanced, we use SMOTE [14] to increase the number of underrepresented cases, i.e. failure cases. After SMOTE, we compared different predictive models in Azure, as shown in Fig. 5. This station-level modeling lays a foundation for future multi-station analysis.

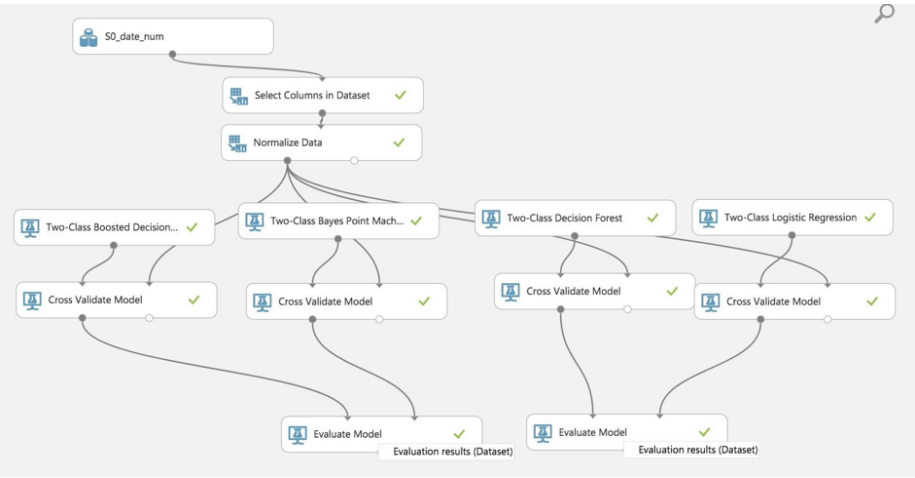

Fig. 5 Station-level modeling

3.4 Between-station data analysis

We further analyze the sensitive stations from station-level analytics. The sensitive stations are the stations that have significant influences on the response. The idea is to partition the overall variance in the response due to each station and the error. For the response y, we have a two-level factor variable x_i of each station, where i = 0, 1, 2, …, 51:

$$x_i = 0 \text{ Inactive}$$
$$x_i = 1 \text{ Active}$$

The number of responses belongs to "1" is significantly less than those belongs to the "0" class that happens 0.58% of the time, which suggests that the dataset is highly imbalanced. When one tries to develop a predictive model based on this dataset, the high imbalance issue will lead to a biased and inaccurate model. To address this challenge, we removed the duplicate rows to increase the rare case to 12.8%. We then applied Ward Hierarchical Clustering method to the failure dataset and success dataset. The results of hierarchical clustering can be shown in Section 4.2.

3.5 Predictive modeling

The dataset has over one million samples and over two-thousands features. There are numerical features, categorical features and time features. The meaning of features is not given due to confidential issue; hence we cannot apply engineering knowledge here. Since this is a large n and large p problem, and the structures of data are complicated, traditional statistical models do not work well. We tried the logistic regression, but it did not work well. Hence we adopt the ensemble learning method to handle the problem. Xgboost classifier is used here, which is a gradient boosting framework that has gained much popularity and attention recently. Due to the large number of features, as well as large number of samples, we first use 0.1 million data samples with all features to train a Xgboost model. After fitting model, we select the important features that have value greater than the threshold 0.005. This step will reduce the 2124 features to 68. Then we use these features to retrain the Xgboost model for the entire dataset. After cross validation, we can find the best model and the results will be shown in next sections.

3.6 Data visualization

We have investigated several visualization approaches then decide to use D3 BihiSankey diagram as our principal tool for the visualization module to address following consideration:

- Easy for deployment: D3, a versatile JavaScript library for visualizing data using web standards, is easy to deploy and maintain across multiple platforms. It is beneficial especially to applications that deal with complicated operating system and software environment in the manufacturing industry.
- Sophisticated cyclic network data flow: The bi-directional hierarchical Sankey (BihiSankey) graph [15] is a Sankey diagram variation first developed by Atkinson. It is suited for visualizing data flow in the network. Unlike conventional Sankey diagram, BihiSankey can work with cyclic networks.

- High-level-informative visualization and friendly user-data interaction: It also has flexible re-layout functionalities which enable friendly user-module interactions.

4. Experiments, Evaluations and Results

In this section, we would like to test the performance of our proposed methods. We list questions our experiments are designed to answer, including:

- How is the global (between-station) and local (within-station) correlations relationship among features?
- Which stations have significant probabilities of failures compared to others?
- Can we predict the failure rate for a sample that passes specific path?
- Can the D3 realize a dynamic failure rate ranking and visualization?
- Does the D3 have the excellent interactive capability for manufacturing engineers?

We will answer these questions in Section 4.1 to Section 4.4.

4.1 Global and local correlation maps

We can explore the global correlation map and local correlation map based on our proposed methods in Section 3.3 and 3.4. The global correlation map shown in Fig.6 (a) indicates the correlation relationship among all the features, while the local correlation map shown in Fig.6 (b) only illustrates the association relationship among single-station features. If we find one zone has stronger correlation pattern in the global correlation map, we will further explore the local data.

The correlation maps lay a solid foundation for causality assessment, which can be helpful for diagnosis of manufacturing systems.

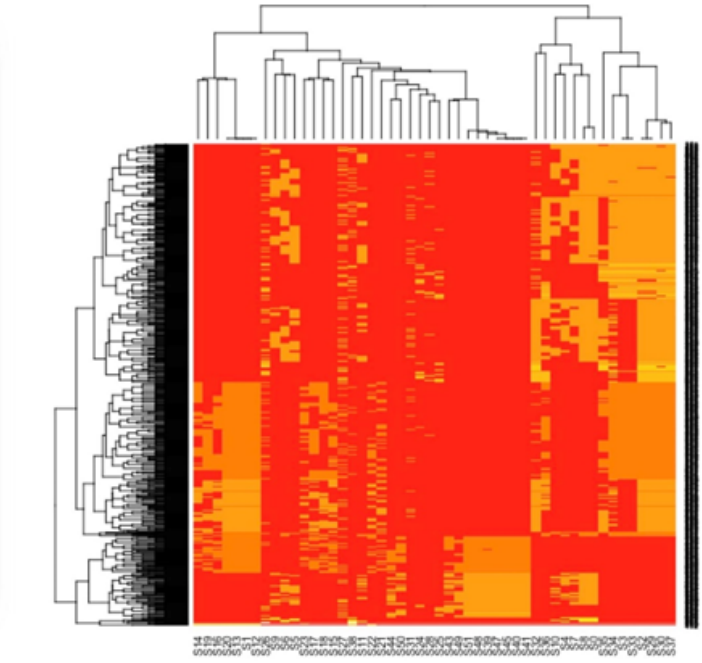

Fig. 6 (a) Global correlation map

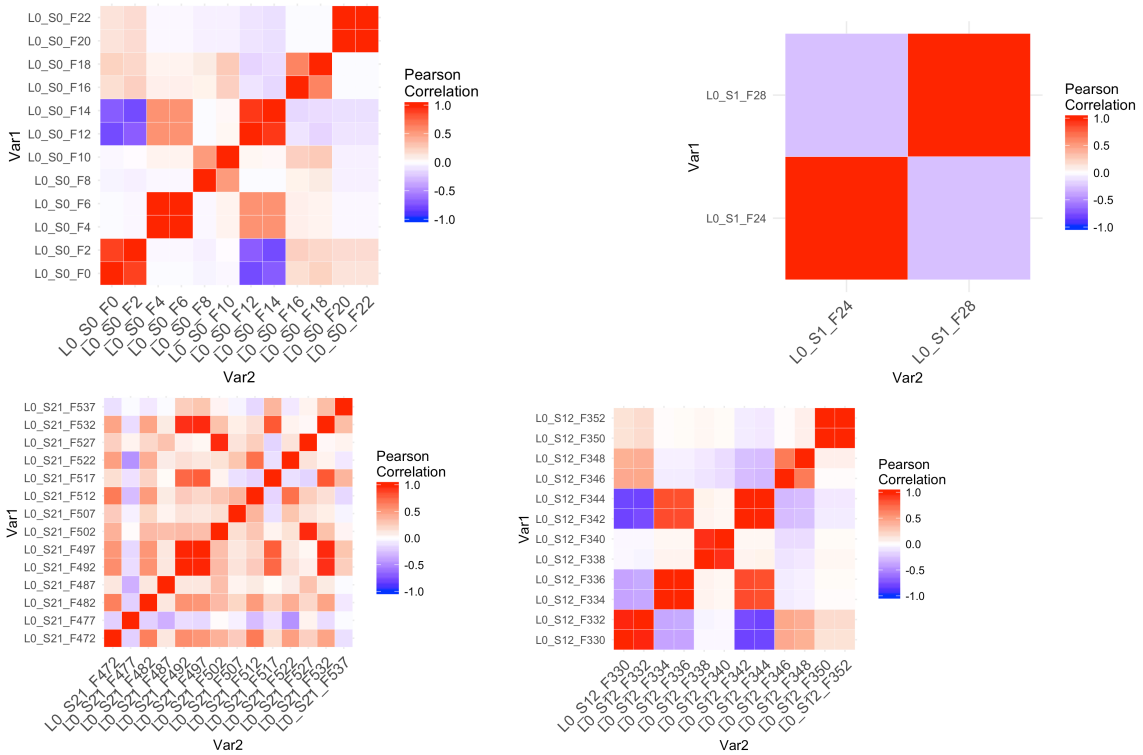

Fig. 6 (b) Feature correlation map of Station 0, 1, 21, 12

4.2 Ward hierarchical clustering

We used the Ward hierarchical clustering method to the failure dataset and success dataset. Their results are shown in Fig. 7 (a) and (b) respectively. Based on the hierarchical clustering, we can find the potential connections among different stations. Also, comparing clusters for failure and success datasets, one could conclude that Station 43, 44, 50, 49 are very sensitive to the responses, and they have a larger probability of resulting failures.

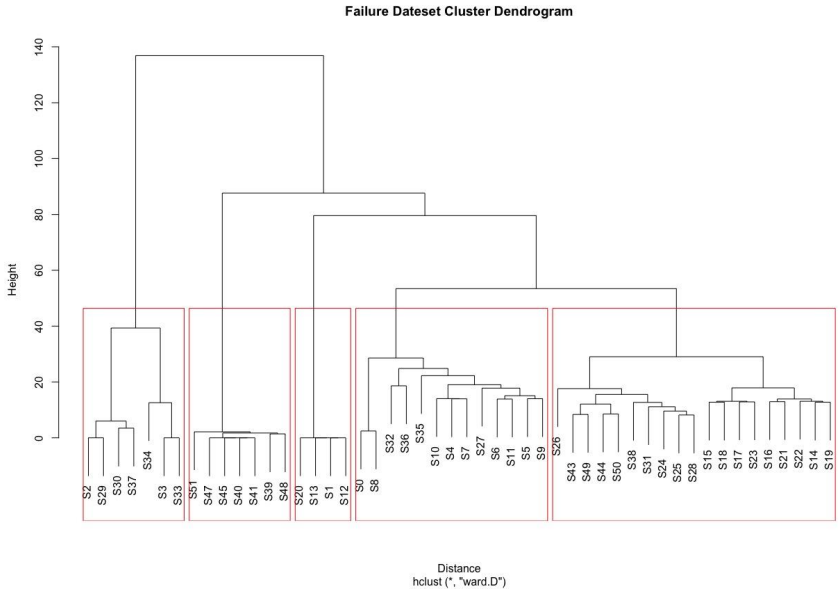

Fig. 7 (a) Failure ("Response == 1") Dateset Cluster Denfrogram

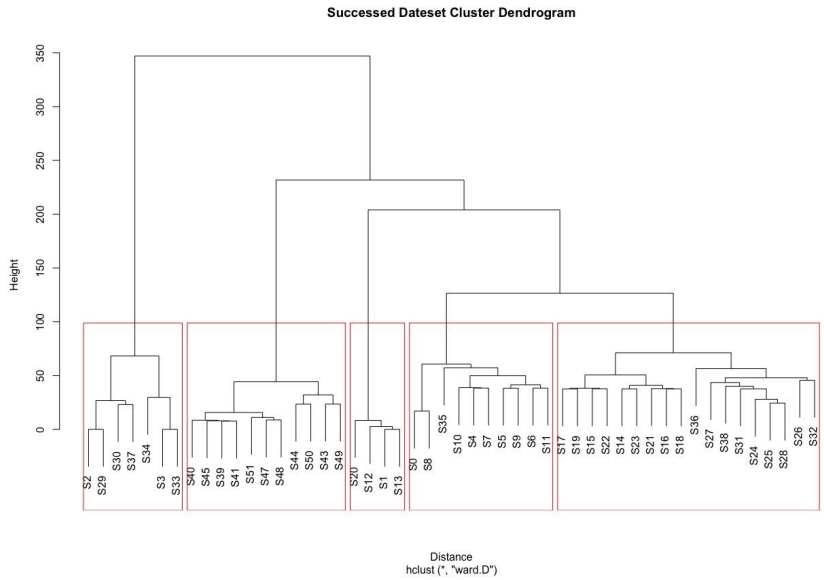

Fig. 7 (b) Success ("Response == 0") Dateset Cluster Denfrogram

4.3 Predictive modeling result

Fig.8 shows the feature selection results. As you can see from the graph, most of features are not important and after feature selection only 68 feature remains.

```
[  23  939 1018 1019 1029 1042 1160 1165 1168 1169 1173 1174 1178 1183
 1187 1189 1197 1213 1221 1222 1228 1231 1236 1238 1245 1252 1270 1278
 1294 1298 1311 1443 1458 1477 1490 1516 1549 1550 1585 1684 1852 1858
 1884 1887 1888 1893 1897 1911 1927 1936 1940 1959 1960 1975 1982 1983
 1985 1987 1988 1992 1994 1999 2006 2007 2010 2022 2063 2094]
```

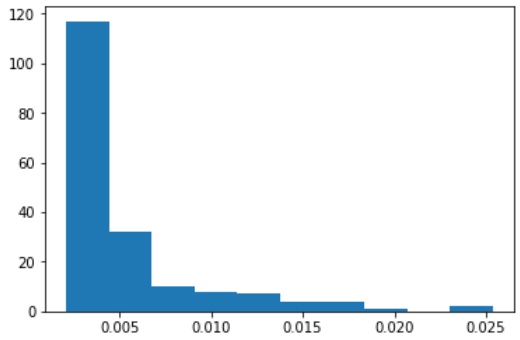

Fig. 8 Selected features and with feature importance value

Since this is a binary classification problem AUC under ROC value is used to measure the performance of the model. Since the data set is too large, we are not able to use cross validation to tune a large range of hyper parameters. We use cross validation to tune the two most important hyper parameter of the model, which is max_depth and base_score. From Fig.9, max_depth 10 and base_score 0.005 will give the best AUC under ROC at 0.735.

max_depth	3	5	10	3	5	10
base_score	0.005	0.005	0.005	0.05	0.05	0.05
AUC under ROC	0.684	0.708	0.735	0.651	0.678	0.692

Fig. 9 Tuning parameter and prediction result

4.4 Interactive interface

We developed our interactive interface based on methods in Section 3.6. Our data visualization interface consists of 3 modules:

(i) The top one is the BihiSankey diagram visualizing the full data flow graph of the production lines. It enables users to have a clear grasp of the production status on the system level. Each node represents a production line station, whose size is determined by the total product quantities at the station. The directional edges indicate the products' flowing through the production process, with the line-widths proportional to the product numbers (Fig. 10).

Due to the graph's complexity, we also create extra nodes to aggregate the corresponding stations (Fig. 11). All the nodes can be repositioned by users for optimum visual effect.

Besides of BihiSankey, we've also investigated other d3 techniques such as force-graph and conventional Sankey diagram. Neither of them worked out since force-graph cannot properly demonstrate the data flow, and normal Sankey failed to process cyclic graph.

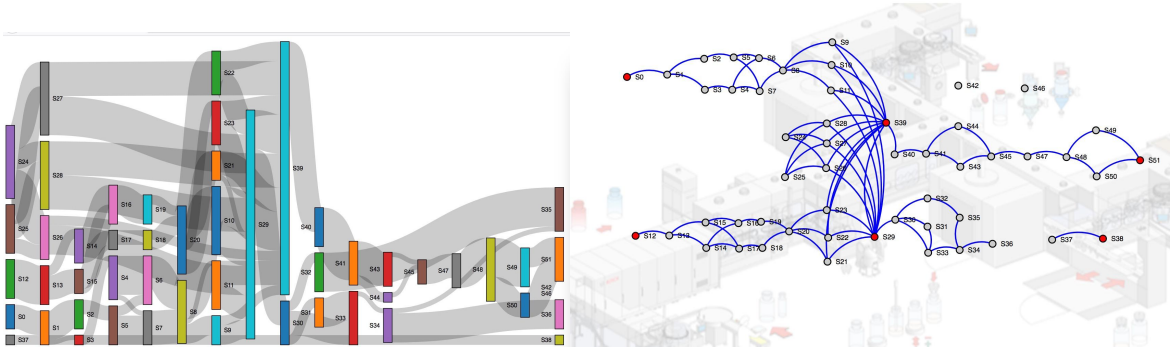

Fig. 10 (a) conventional Sankey diagram (b) Force-graph

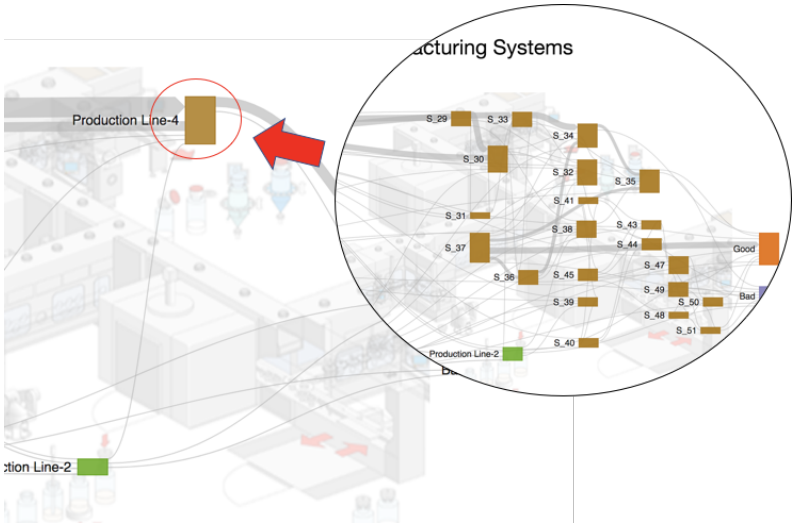

Fig. 11 Illustration of node aggregating / expanding

(ii) The bottom-left module displays the correlation maps of the sensors within an individual station. There are more than 1000 sensors over 52 stations in the graph. Each sensor corresponding to one feature in our predicting model. How to visualize the feature-wise information is quite challenging. After some experiments, we come to the solution with

utilizing the correlation maps as shown in Fig. 12. The patterns of those maps ideally exhibit the dynamics changes on the system level, which are difficult to observe on the single chart of an individual feature.

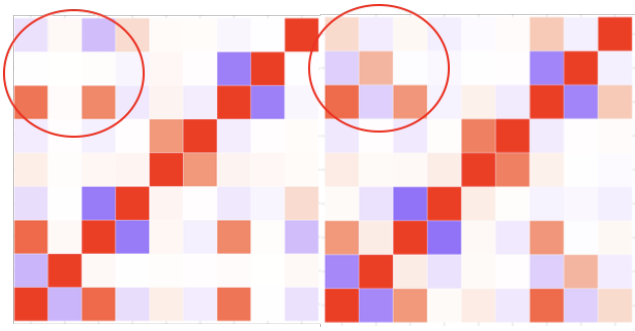

Fig. 12 Different pattern caused by feature dynamics change

(iii) The bottom-right module shows the ten stations with the highest product failure rate in real-time. Every 3 second, our system will re-calculate the failure rate of every station, rank them, and update the list, as shown in Fig. 13.

Identifying the failure product is important, however, it is more critical for manufacturing practitioner to find out why and how. That is why we decided to monitor the station failure rate in real time.

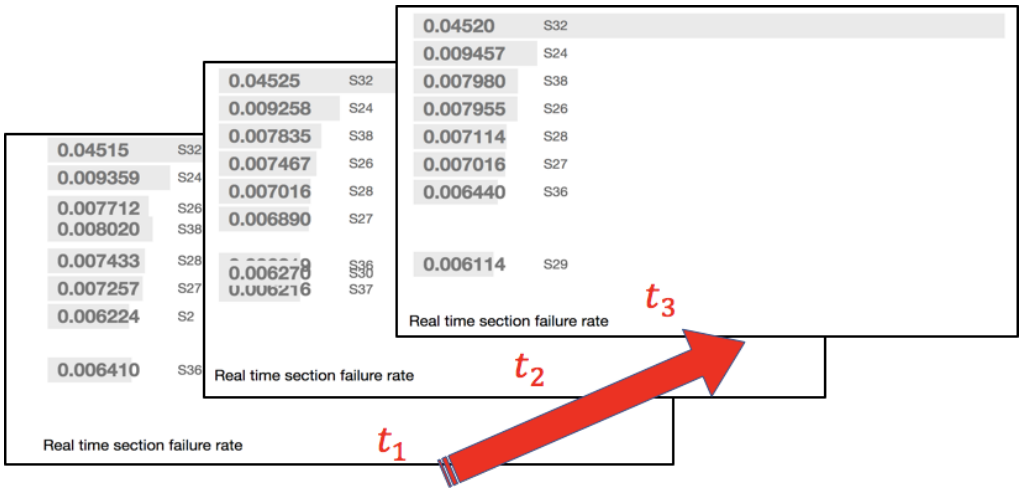

Fig. 13 Dynamic bar-chart with real-time station failure rate ranking

After integration of all three key modules, we can get our final data visualization interface, as shown in Fig. 14. We can explore the influence of each manufacturing station and the corresponding correlation map. Also, the failure rate from different stations are dynamically updated as the collection of data streaming. This interface is an example of data analytics and visualization for smart manufacturing.

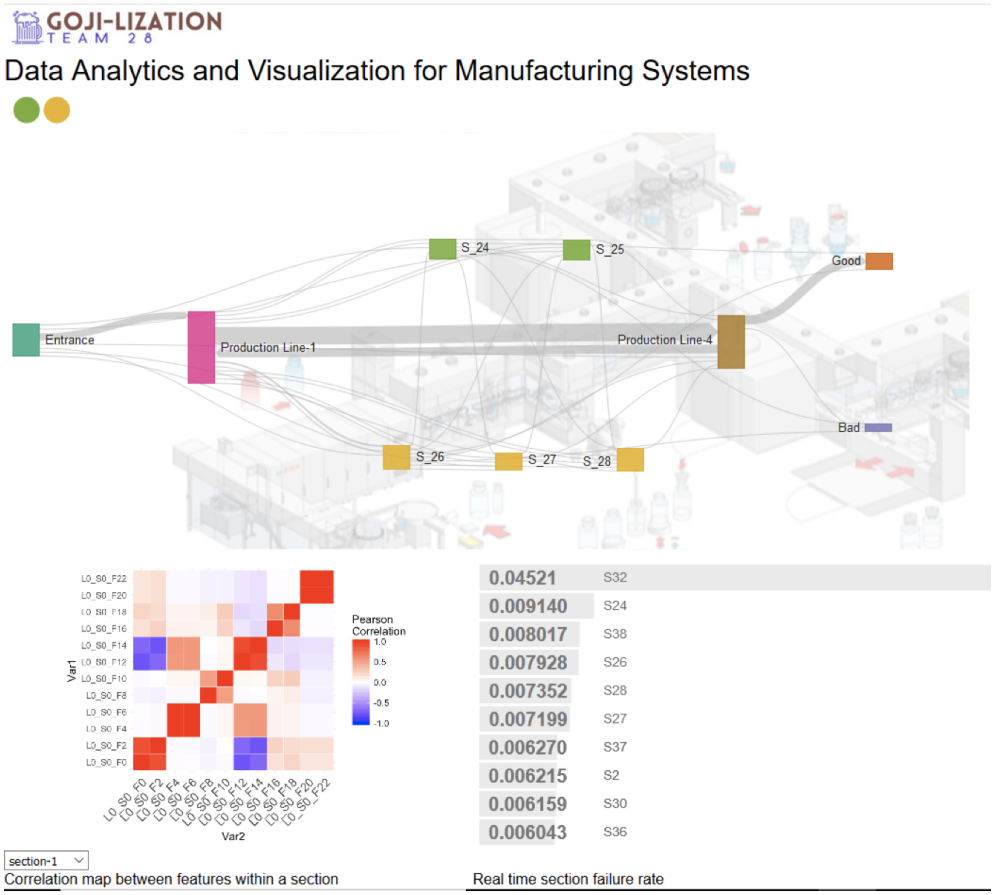

Fig. 14 Interface overview

5. Conclusions and Future Work

We highlighted our key experimental results in this section. The proposed Big Data analytics and visualization methodology consists of following critical experimental results:

- Identify the global and local correlation map
- Identify the key sections (40, 43, 49, 50) that result in large failure rate
- Predict the failure samples
- Develop an interactive interface for visualization

Our main contributions include:

- A systematic framework for big data analytics and visualization has been developed for smart MFGSs
- The global/local correlation map, hierarchical clustering, and predictive modeling are explored
- We provide an interactive D3 interface for manufacturing engineers

For the future work, we will try to promote the deep integration between data analytics, visualization and MFGSs. A cloud-based data analytics and visualization architecture will be utilized for smart manufacturing, as shown in Fig. 15.

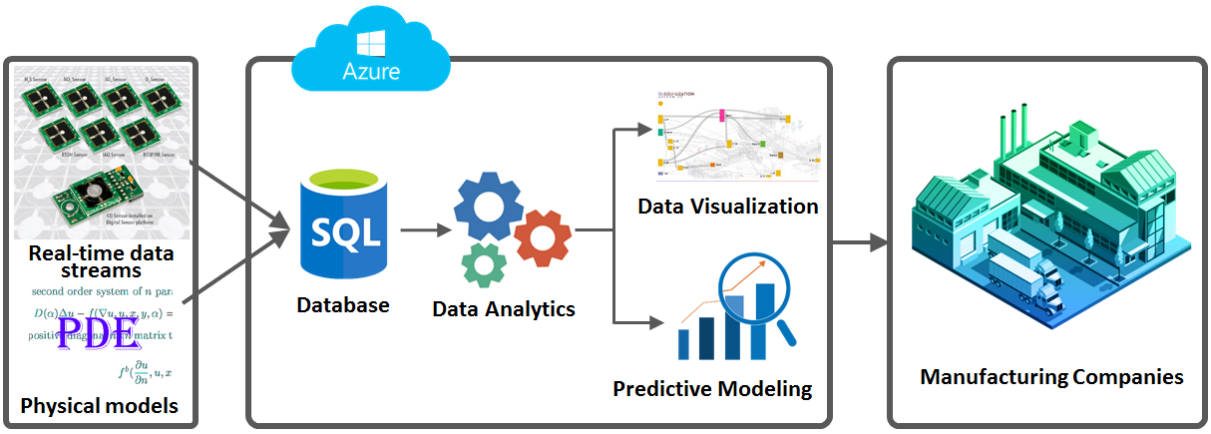

Fig. 15 Cloud-based smart manufacturing architecture

Distribution of team member effort

All team members have contributed similar amount of effort.

Reference

[1] Choudhary, A.K., Harding, J.A. and Tiwari, M.K., 2009. Data mining in manufacturing: a review based on the kind of knowledge. *Journal of Intelligent Manufacturing*, *20*(5), p.501.

[2] Thorsten Wuest, Daniel Weimer, Christopher Irgens & Klaus-Dieter Thoben (2016) Machine learning in manufacturing: advantages, challenges, and applications, Production & Manufacturing Research, 4:1, 23-45,DOI: 10.1080/21693277.2016.1192517

[3] Kaggle, "Bosch Production Line Performance," 2016. [Online]. Available: https://www.kaggle.com/c/bosch-production-line-performance.

[4] Wang, X.Z. and McGreavy, C., 1998. Automatic classification for mining process operational data. Industrial & Engineering Chemistry Research, 37(6), pp.2215-2222.

[5] Tseng, T.L.B., Leeper, T., Banda, C., Herren, S.M. and Ford, J., 2004, January. Quality assurance in machining process using data mining. In IIE Annual Conference. Proceedings (p. 1). Institute of Industrial and Systems Engineers (IISE).

[6]Koonce, D. A., & Tsai, S. C. (2000). Using data mining to find patterns in genetic algorithm solutions to a job shop schedule. Computers & Industrial Engineering, 38, 361–374. doi:10.1016/ S0360-8352(00)00050-4.

[7] Braha, D., & Shmilovici, A. (2002). Data mining for improving a cleaning process in the semiconductor Industry. IEEE Transactions on Semiconductor Manufacturing, 15(1), 91–101. doi:10.1109/66.983448.

[8] Wang, C. H., Kuo,W.,&Bensmail, H. (2006). Detection and classification of defects patterns on semiconductor wafers. IIE Transactions, 38, 1059–1068. doi:10.1080/07408170600733236.

[9] Chiou, J.M. and Müller, H.G., 2014. Linear manifold modelling of multivariate functional data. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 76(3), pp.605-626.

[10] Paynabar, K., Zou, C. and Qiu, P., 2016. A change-point approach for phase-i analysis in multivariate profile monitoring and diagnosis. Technometrics, 58(2), pp.191-204.

[11] Qiao, X., Guo, S. and James, G.M., 2017. Functional graphical models. Journal of the American Statistical Association, (just-accepted).

[12] Bahadori, M.T., Kale, D., Fan, Y. and Liu, Y., 2015, June. Functional subspace clustering with application to time series. In International Conference on Machine Learning (pp. 228-237).

[13] Choudhary, A.K., Harding, J.A. and Tiwari, M.K., 2009. Data mining in manufacturing: a review based on the kind of knowledge. Journal of Intelligent Manufacturing, 20(5), p.501.

[14] Montgomery, D.C., 2009. Introduction to statistical quality control. John Wiley & Sons (New York).

[15] MacGregor, J.F. and Kourti, T., 1995. Statistical process control of multivariate processes. Control Engineering Practice, 3(3), pp.403-414.

[16] Flath, C.M. and Stein, N., 2018. Towards a data science toolbox for industrial analytics applications. Computers in Industry, 94, pp.16-25.

[17] Nedelkoski, S. and Stojanovski, G., 2017, July. Machine learning for large scale manufacturing data with limited information. In Control & Automation (ICCA), 2017 13th IEEE International Conference on (pp. 70-75). IEEE.

[18] Rohrer, M.W., 2000. Seeing is believing: the importance of visualization in manufacturing simulation. In Simulation Conference, 2000. Proceedings. Winter (Vol. 2, pp. 1211-1216). IEEE.

[19] Lindskog, E., Berglund, J., Vallhagen, J. and Johansson, B., 2013. Visualization support for virtual redesign of manufacturing systems. Procedia CIRP, 7, pp.419-424

[20] Oksanen, J., Blanchet, F. G., Kindt, R., Legendre, P., Minchin, P. R., O'hara, R. B., ... & Wagner, H. (2013). Package 'vegan'. Community ecology package, version, 2(9).

[21] Lusa, L. (2013). SMOTE for high-dimensional class-imbalanced data. BMC bioinformatics, 14(1), 106.

[22] Maclean, M., 2018. D3 Tips and Tricks v3.x. Interactive Data Visualization in a Web Browser.

[23] Murray, S., 2017. Interactive Data Visualization for the Web: An Introduction to Designing with D3.

[24] Atkinson, N., https://github.com/Neilos/bihisankey

[25] Bostock, M., and Davies, J., https://github.com/d3/d3-sankey

[26] https://gist.github.com/emoruzzi/6f8140c4d903e64ae35dce0c971e488f#file-app-js-L404

[27] https://github.com/d3/d3/wiki