Utilization of Machine Learning Techniques for Quality Monitoring and Prediction

Mohamed Ismail and Ahmed El-assal
Department of Mechanical Engineering
Benha Unversity
Benha 13518, Qalubia, Egypt
mohamed.allam@bhit.bu.edu.eg, ahmed.El-Assal@bhit.bu.edu.eg

Noha A. Mostafa
Mechanical Engineering Department, The British University in Egypt 11837, Cairo, Egypt
Industrial Engineering Department, Zagazig University 44519, Sharkia, Egypt
*Corresponding author: noha.mostafa@bue.edu.eg; namostafa@eng.zu.edu.eg
ORCID: 0000-0002-6410-3848

Abstract

Product quality is a key factor for manufacturing companies to evaluate their production capability and increase their market competitiveness. Today's Manufacturing processes have become more complicated and usually equipped with smart sensors that collect a massive amount of data along the manufacturing chain. This chain consists of a multistage of manufacturing processes to produce complex products to satisfy customer requirements. In multistage manufacturing systems, many factors may have interactive and cumulative effects on the final product quality. The purpose of this research is to introduce an intelligent real-time quality monitoring framework capable of predicting and identifying the quality deviations for multistage manufacturing systems as early as possible to reduce wastes of time and resources. We used different unsupervised and supervised machine learning techniques such as principal component analysis, support vector machine, neural network and random forest to consider the accumulative effect of different workstations and to construct the quality monitoring model. We used a complex semiconductor manufacturing dataset to evaluate the performance of the proposed framework. The results show the capability of the proposed framework to improve the performance of the quality monitoring process in the multistage manufacturing systems and to reduce both type I and type II errors.

Keywords:
Quality monitoring, Quality prediction, Multistage manufacturing, Industry 4.0, Machine Learning

1. Introduction

Product quality is a key factor for manufacturing companies to assess their production capability. According to (ISO 9000:2015), quality can be defined as "the degree to which a set of inherent characteristics fulfills requirements". Quality control seeks to monitor and predict the quality of products during the manufacturing processes. Traditional statistical process control methods (SPC) have been widely used in the industrial environment due to their simplicity and applicability. The successful application of traditional SPC is attributed to stable and low complexity processes, which thus limits the current application fields (Bai et al. 2019). Today's manufacturing processes have become more complicated with high dimension variables, uncertain and dynamic environments, and multistage of manufacturing processes (Cheng et al. 2018). Consequently, traditional SPC methods are not sufficient in dealing with current problems. Therefore, a more efficient method is needed to tackle quality monitoring problems.
Machine learning (ML), as a computational engine for data mining and pattern recognition, is capable of dealing with complex, high-dimensional multistage manufacturing processes (Ge et al. 2017). ML techniques can be used to transform the massive amount of data that had been collected by build-in or add-on smart sensors into valuable information that can explain the uncertainties and assist in making more informed decisions (Lee et al. 2013). The

output from the data mining models is a real-time quality index after each manufacturing workstation, which indicates the current performance so that the potential failures can be avoided before they occur (Kao et al. 2017).
Currently, most quality monitoring models focus on single-stage manufacturing or address the manufacturing chain as a single point (Arif et al. 2013). However, for the manufacturing chain, many factors (e.g., equipment, manufacturing variables, operators) may have interactive and cumulative effects on the final product quality (Kao et al. 2017). Also, most of the models need to wait until the end of the process to make the prediction, which might cause waste of time and high internal failure costs such as cost of scrap, rework, machine wear, etc.
The main objective of this paper is to introduce a real-time quality monitoring framework that uses data mining techniques to deal with complex and high-dimensional multistage manufacturing processes as well as the unbalanced nature of the manufacturing process.
The rest of this paper is organized as follows: in section 2, a brief literature review is given on the state-of-the-art works on quality monitoring and the integration between ML and quality monitoring. In section 3, the proposed quality monitoring framework is discussed. In section 4, a numerical example is used to verify the performance of the proposed framework and finally, in section 5, the conclusion and future work are given.

2. Literature Review

2.1 Statistical Process Control

Traditional SPC approaches have been widely used in the industrial environment since the 1920s due to their simplicity and applicability. The first SPC tool was pioneered by W.A. Shewhart (Shewhart 1931) to detect out-of-control points that indicate the existence of assignable causes. Later, Cumulative sum (CUSUM) and exponential weight moving average (EWMA) control charts were developed to detect small shift changes faster (Amini and Chang 2018). For the multivariable environment, the Hotelling T^2 control chart was developed to detect mean shifts for multivariable quality characteristics, usually less than ten variables (Li et al. 2006). Also, multivariate exponentially weighted moving averages (MEWMA) and multivariate cumulative sum (MCUSUM) were developed for multivariable environments; however, these methods suffer from unacceptable false alarms and can not detect out-of-control observations fast enough (Amini and Chang 2018). Since the needs and desires of the customers have become more sophisticated, the complexity in product structure has been increasing. Hence, multistage manufacturing systems (MMS) become more common to produce complex products. Most traditional SPC methods do not perform well in MMS as they are poorly capable of discriminating between changes at different stages and suffer from high false alarm rates (Shi and Zhou 2009).

2.2 Machine Learning

Over the past decades, data mining has played an essential role in knowledge discovery and decision support. ML serves as a computational engine for data mining, pattern recognition, information extraction and prediction. It can be categorized as supervised, unsupervised, semi-supervised and reinforcement learning (Ge et al. 2017).

2.2.1 Unsupervised Learning Methods

Unsupervised learning is used when the data set consists of unlabeled training examples $\{X_i\}_{i=1}^{N}$, where X is a feature vector without any corresponding target value Y. It is used for information extraction, dimensionality reduction, density estimation, data visualization, outlier detection, and process monitoring (Ge et al. 2017). Principal component analysis (PCA) and K-Means algorithms are used in this work to represent the accumulative effect for each workstation in the manufacturing chain and identify operational patterns for each workstation.

2.2.1.1 Principal Component Analysis (PCA)

PCA is a statistical technique that is used to transform the correlated variables from high dimensional space R^n to linearly uncorrelated variables, which is also called principal components (PCs) in a lower-dimensional space R^k where (n > k) while retaining the most information (Abdi and Williams 2010). The PCs can be calculated by computing the eigenvalue and eigenvectors of the covariance matrix of features. The resulting eigenvectors are used as PCs. The main applications in the industry are dimensionality reduction, process monitoring, abnormality detection, data visualization, and outlier detection (Ge et al. 2017).

2.2.1.2 K-Means algorithm

K-means algorithm is used to partition the data set into a specific number of K clusters. K-means algorithm randomly selects K points as initial centroids for each cluster, then assign data into the nearest cluster based on the distance between the cluster's centroid and data. There exist many methods to identify this distance; euclidean distance is one of the most used ones. Afterwards a new centroid is recalculated for each cluster. The process is repeated until the centroid of each cluster does not change or satisfy the tolerance value (Syakur et al. 2018). The main application of K-means clustering in the industry is to identify the different fault types and operating modes for manufacturing processes.

2.2.2 Supervised Learning Methods

Supervised learning is used when the data set consists of labeled training examples $\{X_i, y_i\}_{i=1}^{N}$ Where X is a feature vector with the corresponding target value y, the target value y can be discrete or continuous. If categorical output variables are set, the problem is formulated as a classification problem. Otherwise, the problem is formulated as a regression problem. Supervised learning can be used for fault diagnosis, process monitoring, quality prediction, and remaining useful life estimation (Ge et al. 2017). Supervised learning methods in the manufacturing processes include support vector machine, neural networks, decision trees, K-nearest, linear regression, and random forest.

2.2.2.1 Support Vector Machine (SVM)

The key idea of SVM is to create one or more hyperplane to separate different classes in a high dimensional space. SVM focuses on maximizing the margin, which is the distance between the closest training examples in the two classes. It is used when the data can be linearly separable; otherwise, the input vector needs to be mapped into higher dimensional space by using kernel tricks, then the hyperplane can be constructed (Tsai et al. 2009).

2.2.2.2 Artificial Neural Network (ANN)

The artificial neural network mimics how the human brain uses interconnected neurons to process information. ANN consists of an input, output and hidden layers. Each of them contains numbers of neurons that are connected to all neurons in the following layer. Each input neuron has an activation value; the neuron sends its value to the neurons in the following layer to which it is connected. For each receiving neurons, the activation values are calculated by an activation function. The weight of the connections are randomly assumed then the backpropagation algorithm is used during the training to find the optimal weights of the connections (Kotsiantis 2007).

2.2.2.3 k-Nearest Neighbor (k-NN)

A k-NN is a nonparametric leaning method based on the idea that instances that have similar properties will generally exist in proximity to each other. Therefore, a K number of nearest neighbors can be determined by calculating the Mahalanobis or Euclidean distance between the unclassed instance or input feature vector X and the training instances. The class of input vector can be assigned based on the majority vote of the K nearest instances in the training dataset (Cunningham and Delany 2007). The accuracy of the classification is affected by irrelevant features and noise.

2.2.2.4 Random Forest (RF)

RF is an ensemble learning method that uses multi decision trees to make the prediction (Breiman 2001). A number K of decision trees is built based on bootstrapped samples with n observations. Each decision tree is built using a randomly selected subset of k features. Each decision tree in the random forest will give a vote when a new sample needed to be classified. The final class for a new sample will be assigned based on the majority votes from all trees in the forest.

2.3 Integration Between Machine Learning And Quality Monitoring

ML is used in many industrial applications such as fault diagnosis, predictive maintenance, quality monitoring and improvement, and process optimization (Ge et al. 2017, Köksal et al. 2011, Weichert et al. 2019). Quality prediction models based on ML algorithms have been used in several manufacturing areas such as steel manufacturing (Li et al. 2019), semiconductor manufacturing (Kao et al. 2017, Arif et al. 2013, Al-Kharaz et al. 2019), 3D printing (Amini and Chang 2018), extrusion processes (García et al. 2019), battery manufacturing (Thiede et al. 2019), and automotive industry (Peres et al. 2019).
Arif et al. (2013) introduced an approach for quality monitoring in MMS using a cascaded quality monitoring system. They suggested using PCA to find intercorrelated variables relationships then applied the Iterative Dichotomiser algorithm (ID3) to identify the quality state of semiconductor wafers. The proposed method has an acceptable false alarm rate; however, the true positive (TP) rate was very low, which means it is not applicable for the quality prediction purpose. Kao et al. (2017) have proposed a combination of PCA with classification algorithms

such as SVM, decision trees and naïve base to predict the quality in the MMS. They used association rule mining for root cause analysis of defective products. Amini and Chang (2018) introduced an approach called a multi-layer classifier for process monitoring (MLCPM) to predict the likelihood of a successful printing process at critical layers. They suggest using K-means to assign each printing layer to a cluster then used the random forest algorithm to identify the most significant layer on the product quality. This method has a significant impact on reducing the dimensionality of the data and the computational time. However, the proposed framework did not take into account the cumulative and interactive effect from layer to layer during the printing process.
Most of the previous studies have applied the quality monitoring model in single-stage manufacturing or treated the manufacturing chain as a single point, which means that they built one prediction model for the whole process chain at the final stage. This single point model assumed that each workstation has an independent effect on the final product quality. It ignored the fact that each manufacturing workstation has an independent, cumulative and interactive effect on the final product quality. Also, most previous models had to wait until the end of the process to make the prediction, which might cause losses in resources.

3. Methodology

In this section, we introduce a real-time process monitoring framework that can monitor, predict and identify the quality deviations for complex, high-dimensional and multistage manufacturing systems as early as possible in order to reduce waste of resources and time and damage to the reputation of the company if defective products are dispatched to customers. The proposed framework is built and performed in two phases. Phase I is the model building phase, while phase II is the monitoring and updating phase.

3.1 Phase I: Data Collection And Model Building

This phase includes several steps related to data collection and preprocessing, dimensionality reduction, building the quality prediction model, and identifying the most significant stages in the manufacturing chain.

3.1.1 Data Collection

Data and measurements are collected from different sources along the manufacturing chain by using smart sensors. The collected dataset includes process parameters, state variables, environmental sensing variables, maintenance records, intermediate product characteristics and final product quality.

3.1.2 Data Preprocessing

The Manufacturing environment has many noises, outliers, missing values, anomaly events and inconsistencies. Raw data should be carefully preprocessed to remove inconsistent, redundant and misleading information. Data preprocessing helps to improve the quality of the collected raw data and the performance of the quality monitoring model (Xu et al. 2015). Data preprocessing includes data cleaning and data transformation.
The first step of data cleaning is to remove possible human errors and outliers as they spoil and mislead the training process, which in turn will reduce the model performance. Second, variables and instances with missing values above a defended threshold P1 and P2, respectively, should be removed. Third, the remaining missing values should be replaced by attribute mean or most probable value with Bayesian inference. Finally, columns with constant values should be removed as they do not carry any discriminative information (Xu et al. 2015).
After performing the data cleaning, the scale difference between manufacturing process variables should be adopted, so the dataset is normalized between zero and one using equation 1.

$$Y = \frac{(Y_{max} - Y_{min}) * (X - X_{min})}{(X_{max} - X_{min}) + Y_{min}} \tag{1}$$

Y_{max} and Y_{min} are users defended upper and lower bound (1 and 0, respectively), while X_{max} and X_{min} are the maximum and minimum values for each variable.

3.1.3 Feature Selection

Feature selection is the process of identifying a subset of features that is relevant to the quality prediction model. It helps to simplify the learning process, improve the model performance, reduce the computational time, and avoid overfitting. In this research, Fisher discriminant metric is used as a feature selection technique. It simply ranks the features based on the ratio between the inner class distance to inner class variance, as illustrated in equation 2 (Gu et al. 2011). Those features are selected based on the defined threshold P_3.

$$F(x^i) = \frac{\sum_{k=1}^{c} n_k\left(\mu_k^j - \mu^j\right)^2}{\sum_{k=1}^{c} n_k\left(\sigma_k^j\right)^2} \quad (2)$$

where μ_k^j , σ_k^j ,μ^j, σ^j and n_k are the mean and standard deviation for class k and the whole data set, respectively.

3.1.4 Tackling The Imbalance Problem

Another important issue for quality monitoring is insufficient instances for defective class samples. Since failure cases do not frequently happen, the number of positive (defective) class samples is relatively smaller than the negative (nondefective) class samples. When a small minority or the vast majority exists, the training model will be forced to favor the majority voice of the data in order to achieve higher accuracy. This problem is called the imbalance problem. In this research, Synthetic Minority Over-sampling Technique (SMOTE) technique is used to balance the training data set. This technique increases the number of defective (minority) class samples to be equal to the majority class samples (Chawla 2009). We can summarize SMOTE as for each instance x in the minor class, compute its Euclidean distance to all samples in the minor class and list its k nearest neighbor. Based on the percentage of the oversampling required, random samples (e.g., z_1, z_2, …, z_k) from its K nearest neighbors are selected. Then, for each selected z_i,a new sample S is synthesized using equation 3. the process is repeated until the required instance is achieved.

$$S = x + a(z_i - x) \quad (3)$$

where a is a random number between [0,1].

3.1.5 Cascaded Quality Model

In MMS, the quality of the output from a workstation is not only affected by the manufacturing operation variables in this workstation but also affected by the characteristics of the output from the previous workstation. Also, each manufacturing stage contributes to the final quality state (Arif et al. 2013). To consider the characteristic of MMS, partial and total quality concepts should be considered to describe the accumulation of quality variation along the manufacturing chain (Zhai et al. 2004). Cascade Quality Prediction Method (CQPM) [6,7] is implemented to mine the hidden relationship in MMS. There are three types of relationships in MMS: the relationship between manufacturing operation in each workstation, the relationship between workstations, and the relationship between manufacturing operation variables and final product quality R_1, R_2 and R_3, respectively. Figure 1 shows the concept of MMS, relationships in CQPM and partial/total quality concepts.

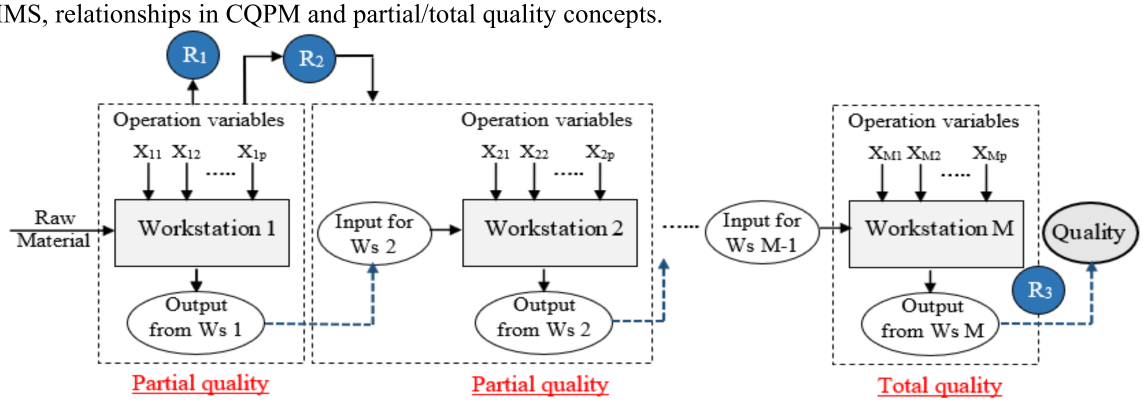

Figure 1. MMS concept, relationship in CQPM and partial/total quality concept

Based on the CQPM definition (Arif et al. 2013), we assume that the manufacturing line consists of M workstations. Each workstation has several manufacturing variables Xs and each intermediate product has several quality characteristics Ts. PCA can be used to transform the latent and correlated variables into a new set of dimensions while carrying the MMS characteristics. Equation 4 describes how the interaction between process variable ($X_{j,k}$) and the quality characteristics from previous work station $T_{(j-1),k}$ can affect the quality of the intermediate products (R_1 and R_2). Equation 5 illustrates that the quality state Q is a function of quality characteristics from the final workstation ($T_{M,s}$), which expresses R_3.

$$T_{j,s} = f\left(T_{(j-1),s}, X_{j,k}\right) = \sum_{k=1}^{p(j-1)} a_{s,k} T_{(j-1),s} + \sum_{k=1}^{p_j} a_{s,k} X_{j,k} \quad (4)$$

$$Q = f\left(T_{M,s}\right) \quad (5)$$

Where: $X_{j,k}$ is the k^{th} manufacturing operation variable in *a* j^{th} workstation (j = 1 to M and K = 1 to p), $T_{j,s}$ is the s^{th} quality characteristic of the intermediate product produced from the j^{th} workstation (s = 1 to r), $a_{k,j}$ is the amount of contribution of $X_{j,k}$ to $T_{j,s}$ or the eigenvector of the covariance matrix of the variables in PCA and Q represents the quality level.

3.1.6 Classification Algorithms

After identifying features that represent the MMS characteristics, several classification algorithms such as ANN, SVM, K-NN, random forest and logistic regression were used to predict the quality state of the final product.

3.1.7 Stage Selection

In MMS, not all workstations have the same effect on the final product quality. We suggest identifying the most significant workstations along the process chain. Then, at these significant workstations, we can build a quality checkpoint to predict the likelihood of the final product quality while intermediate products are still in the manufacturing process instead of checking the quality at the final stages like most of the previous works. In order to do this, we assume first that each workstation has a specific number of patterns, then we apply the K-means algorithm to assign each workstation to a specific cluster. The number of clusters for each workstation can be determined by the elbow method (Syakur et al. 2018). Finally, a feature selection technique can be applied to select the most significant workstations on the final product quality.

3.2 Phase II – Process Monitoring And Updating The Training Dataset

CQPM is applied to the manufacturing variables in each workstation and quality characteristics from the previous workstation. Then, when the first significant workstation is reached, the first quality prediction model will predict an initial result. This prediction provides the likelihood of the final product quality while the intermediate product still in the manufacturing process. If the likelihood of prediction is 0 (good product), the process continues until the next significant stage is reached. Otherwise, the process engineer or an automatic adaptation system can take action to save resources or to adopt the process parameters in the next workstation in a way that the intermediate product may have a chance to be good at the final workstation.

The training data should be updated periodically to cover any new patterns and reduce the chance of misclassification. We suggest using a critical threshold to trigger a warning when the model misclassification increases. Hence, the model can be updated by retraining the model with the initial training data plus the newly collected data.

4. Numerical Example

We used a complex semiconductor manufacturing dataset (called SECOM) (McCann and Johnston 2008) to verify the proposed framework. This dataset consists of 590 operation variables collected from different sensors along the semiconductor wafer manufacturing line and 1567 instances or samples and one quality variable, which is labeled as 1 for a non-conforming wafer and -1 for a conforming wafer. This data set is highly imbalanced with an approximate ratio of 14:1 between the two classes and it has 41,951 missing values distributed with different percentages along the manufacturing variables and 116 constant and irrelevant variables.

In order to deal with this complex dataset, we cleaned the dataset by removing features and instances which contain a percentage of missing values above thresholds P_1 = 50% and P2 = 26%. Then low variance and irrelevant features were removed from the dataset as they don't carry unique information; also, outliers were removed using the interquartile range (IQR) score as they decrease the model performance. The remained missing values were replaced by attribute means of each feature. All the data were normalized between zero and one using equation 1 and randomly split into 80% for training and 20% for testing. We ranked the manufacturing variables based on their importance using the fisher discriminant metric to remove variables that were not related to the quality monitoring process using threshold P_3 = 0.022. The previous steps helped to reduce the number of features to 63 features, which is a 936.5% reduction in dimension. We rebalanced the training dataset using the SMOTE algorithm so that the ratio between the two classes in the training set is approximately equal to 1.

In order to simulate the MMS scenario, we split the dataset into five groups; each one represents a manufacturing workstation as recommended by (Kao et al. 2017). We construct two quality monitoring models based on the proposed framework to assure the capability of the proposed system for quality monitoring in multistage manufacturing systems. The first model is based on a single-point approach and the second one is based on CQPM. The single point approach was build using the same previous steps, but instead of applying CQPM, we treat the whole data as a single point and build one quality model. The number of PCS that represent 95% of the variance in the data for the first model was 35 and for the second model is shown in table 1.

Table 1: Principal components at each manufacturing workstation

	Workstation 1	Workstation 2	Workstation 3	Workstation 4	Workstation 5
Principal Components	11	18	25	26	27

We used six ML algorithms, namely Naïve Bayes, KNN, SVM, random forest (RF), logistic regression (LR) and ANN, to predict the product quality level. We applied ten-fold cross-validation to avoid overfitting and grid search for hyperparameter tunning for each of the ML models. Accuracy, sensitivity and specificity metrics were used to evaluate the performance of the classification models and compare between the two quality monitoring approaches. The results shown in Figure 2 represent the performance of the classification models on the test set for the two quality monitoring approaches .

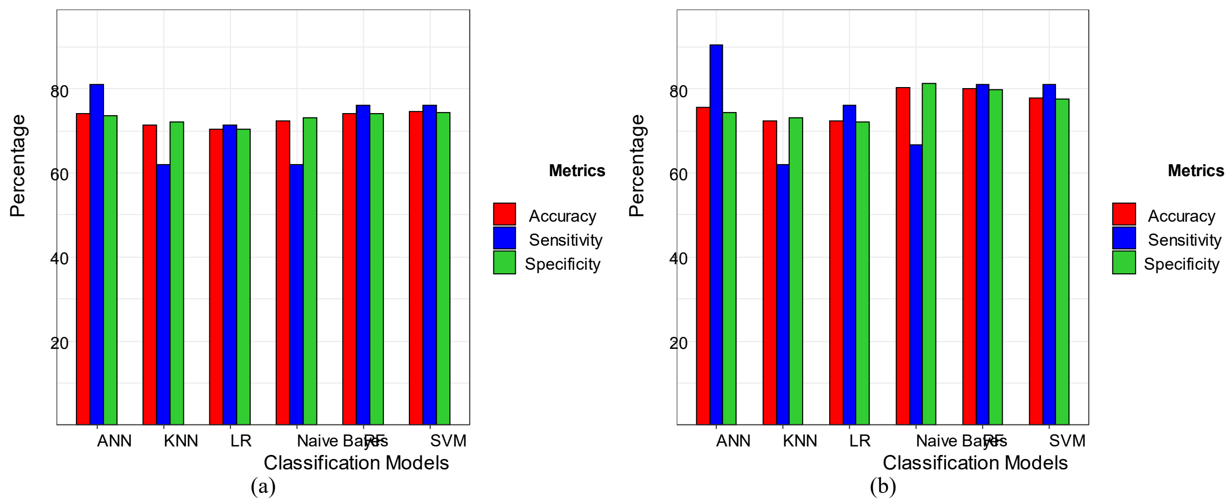

Figure 2: Model performance for each classification algorithms based on a) single-point approach and b) CQPM approach.

Detecting most of the non-conforming products with a minimum rate of false alarms is the most important aspect of the quality monitoring process. These two aspects can be described by sensitivity and specificity metrics. Results show that using the proposed framework based on CQPM is better for MMS than using the single point approach as the performance of the classification models in terms of sensitivity and specificity has improved. The average performance of all classification models increased in terms of sensitivity and specificity by 4.76% and 3.41%, respectively. Also, the average accuracy of all models has improved by 3.5%. With a sensitivity of 90.47% and specificity of 75.5%, the ANN model performs the best in detecting the non-conforming products with an acceptable value of false alarm rate compared to other models.

In order to identify the most critical and significant stages to final product quality, we applied the K-means algorithm to each manufacturing stage to identify operation patterns for every workstation. We used the elbow method to identify the optimum number of patterns for each workstation. Table 2 shows the number of patterns for each workstation. After identifying the pattern of each workstation, the random forest algorithm was applied in order to identify the most significant workstation to the product quality. Table 3 shows the importance of each workstation based on the random forest variable importance. We set a threshold of $P_4 = 0.2$ in order to identify the most significant workstation. We found that stages 1 and 3 are the most critical workstation, so we recommend building quality checkpoints at those stages to provide an initial prediction about the intermediate products' quality. If the likelihood of prediction is -1 (non-defective product), the process continues until the next significant stage is reached. Otherwise, the process engineer or an automatic adaptation system can take action to save resources or to adopt the process parameters in the next workstation in a way that the intermediate product may have a chance to be good at the final workstation. When the products reach the final workstation, the final quality state can be identified.

Table 2: number of patterns for each manufacturing workstation.

	Workstation 1	Workstation 2	Workstation 3	Workstation 4	Workstation 5
Number of clusters	5	5	4	3	4

Table 3: the importance of each workstation.

	Workstation 1	Workstation 2	Workstation 3	Workstation 4	Workstation 5
Importance	0.335	0.173	0.296	0.135	0.062

5. Conclusion And Future Work

In this research, we introduce an intelligent real-time quality monitoring framework capable of predicting and identifying the quality deviations for complex and high dimensional MMS as early as possible, especially at critical manufacturing stages. Our framework consists of two main phases: Phase I is the model building phase, which includes data collection and preprocessing, and quality prediction model building, while phase II includes monitoring the manufacturing process and updating the quality monitoring model to cover any new patterns and reduce the chance of misclassification.

The results from applying the proposed framework on a complex dataset that contains 1567 instances and 590 features show the capability of the proposed framework to improve the performance of the quality monitoring process in multistage manufacturing systems. The average performance of all classification models increased in terms of sensitivity and specificity by 4.76% and 3.41%, respectively. The ANN algorithm performed the best between all the classification models. We successfully identified the most critical stages which can be used as quality checkpoints.

For future work, we plan to improve the model performance by using more advanced feature selection, extraction and classification techniques. Moreover, we plan to use the critical stages as a quality checkpoint to provide the likelihood of producing non-conforming at the end of manufacturing chain products while the intermediate product passes through those stages.

References

Abdi, H, and Williams, LJ., Principal component analysis, *Wiley interdisciplinary reviews: computational statistics*, vol. 2, no. 4, pp.433–59, 2010.

Al-Kharaz, M, Ananou, B, Ouladsine, M, Combal, M, and Pinaton, J., Quality prediction in semiconductor manufacturing processes using multi-layer perceptron feedforward artificial neural network, *8th International Conference on Systems and Control, ICSC 2019* 423–8, IEEE. doi:10.1109/ICSC47195.2019.8950664.

Amini, M, and Chang, S., A review of machine learning approaches for high dimensional process monitoring, *IISE Annual Conference and Expo 2018*, no. May, pp.390–5, 2018.

Amini, M, and Chang, SI., MLCPM: A process monitoring framework for 3D metal printing in industrial scale, *Computers and Industrial Engineering*, vol. 124, no. July, pp.322–30, 2018. doi:10.1016/j.cie.2018.07.041.

Arif, F, Suryana, N, and Hussin, B., Cascade Quality Prediction Method Using Multiple PCA+ID3 for Multi-Stage Manufacturing System, *IERI Procedia*, vol. 4, no. May 2015, pp.201–7, 2013. doi:10.1016/j.ieri.2013.11.029.

Bai, Y, Sun, Z, Zeng, B, et al., A comparison of dimension reduction techniques for support vector machine modeling of multi-parameter manufacturing quality prediction, *Journal of Intelligent Manufacturing*, vol. 30, no. 5, pp.2245–56, 2019. doi:10.1007/s10845-017-1388-1.

Breiman, L., Random forests, *Machine Learning*, vol. 45, no. 1, pp.5–32, 2001. doi:10.1023/A:1010933404324.

Chawla, N V., Data mining for imbalanced datasets: An overview, In *Data mining and knowledge discovery handbook* 875–86, Springer.

Cheng, Y, Chen, K, Sun, H, Zhang, Y, and Tao, F., Data and knowledge mining with big data towards smart production, *Journal of Industrial Information Integration*, vol. 9, no. September, pp.1–13, 2018. doi:10.1016/j.jii.2017.08.001.

Cunningham, P, and Delany, SJ., k-Nearest neighbour classifiers, *Multiple Classifier Systems*, vol. 34, no. 8, pp.1–17, 2007.

García, V, Sánchez, JS, Rodríguez-Picón, LA, Méndez-González, LC, and Ochoa-Domínguez, H de J., Using regression models for predicting the product quality in a tubing extrusion process, *Journal of Intelligent Manufacturing*, vol. 30, no. 6, pp.2535–44, 2019. doi:10.1007/s10845-018-1418-7.

Ge, Z, Song, Z, Ding, SX, and Huang, B., Data Mining and Analytics in the Process Industry: The Role of Machine Learning, *IEEE Access*, vol. 5, pp.20590–616, 2017. doi:10.1109/ACCESS.2017.2756872.

Gu, Q, Li, Z, and Han, J., Generalized fisher score for feature selection, *Proceedings of the 27th Conference on Uncertainty in Artificial Intelligence, UAI 2011*, pp.266–73, 2011.

(ISO), IO for S., ISO 9000, 2015. Quality management systems–fundamentals and vocabulary, 2015.

Kao, HA, Hsieh, YS, Chen, CH, and Lee, J., Quality prediction modeling for multistage manufacturing based on classification and association rule mining, *MATEC Web of Conferences*, vol. 123, 2017. doi:10.1051/matecconf/201712300029.

Köksal, G, Batmaz, I, and Testik, MC., A review of data mining applications for quality improvement in manufacturing industry, *Expert Systems with Applications*, vol. 38, no. 10, pp.13448–67, 2011. doi:10.1016/j.eswa.2011.04.063.

Kotsiantis, SB., Supervised machine learning: A review of classification techniques, *Informatica (Ljubljana)*, vol. 31, no. 3, pp.249–68, 2007. doi:10.31449/inf.v31i3.148.

Lee, J, Lapira, E, Yang, S, and Kao, A., Predictive manufacturing system - Trends of next-generation production systems, *IFAC Proceedings Volumes (IFAC-PapersOnline)*, vol. 46, no. 7, pp.150–6, 2013. doi:10.3182/20130522-3-BR-4036.00107.

Li, F, Wu, J, Dong, F, et al., Ensemble Machine Learning Systems for the Estimation of Steel Quality Control, *Proceedings - 2018 IEEE International Conference on Big Data, Big Data 2018*,, pp.2245–52, 2019. doi:10.1109/BigData.2018.8622583.

Li, F, Runger, GC, and Tuv, E., Supervised learning for change-point detection, *International Journal of Production Research*, vol. 44, no. 14, pp.2853–68, 2006.

McCann, M, and Johnston, A., SECOM dataset UCI machine learning repository, 2008. https://archive.ics.uci.edu/ml/datasets/SECOM.

Peres, RS, Barata, J, Leitao, P, and Garcia, G., Multistage Quality Control Using Machine Learning in the Automotive Industry, *IEEE Access*, vol. 7, pp.79908–16, 2019. doi:10.1109/ACCESS.2019.2923405.

Shewhart, WA., "*Economic control of quality of manufactured product*." Macmillan And Co Ltd, London: Macmillan And Co Ltd, London.

Shi, J, and Zhou, S., Quality control and improvement for multistage systems: A survey, *IIE transactions*, vol. 41, no. 9, pp.744–53, 2009.

Syakur, MA, Khotimah, BK, Rochman, EMS, and Satoto, BD., Integration K-Means Clustering Method and Elbow Method for Identification of the Best Customer Profile Cluster, *IOP Conference Series: Materials Science and Engineering*, vol. 336, no. 1, 2018. doi:10.1088/1757-899X/336/1/012017.

Thiede, S, Turetskyy, A, Kwade, A, Kara, S, and Herrmann, C., Data mining in battery production chains towards multi-criterial quality prediction, *CIRP Annals*, vol. 68, no. 1, pp.463–6, 2019. doi:10.1016/j.cirp.2019.04.066.

Tsai, CF, Hsu, YF, Lin, CY, and Lin, WY., Intrusion detection by machine learning: A review, *Expert Systems with Applications*, vol. 36, no. 10, pp.11994–2000, 2009. doi:10.1016/j.eswa.2009.05.029.

Weichert, D, Link, P, Stoll, A, Rüping, S, Ihlenfeldt, S, and Wrobel, S., A review of machine learning for the optimization of production processes, *International Journal of Advanced Manufacturing Technology*, vol. 104, nos. 5–8, pp.1889–902, 2019. doi:10.1007/s00170-019-03988-5.

Xu, S, Lu, B, Baldea, M, et al., Data cleaning in the process industries, *Reviews in Chemical Engineering*, vol. 31, no. 5, pp.453–90, 2015. doi:10.1515/revce-2015-0022.

Zhai, J, Xu, X, Xie, C, and Luo, M., Fuzzy control for manufacturing quality based on variable precision rough set, *Proceedings of the World Congress on Intelligent Control and Automation (WCICA)*, vol. 3, pp.2347–51, 2004. doi:10.1109/wcica.2004.1342013.

Biographies

Mohamed Ismail is a teaching and research assistant in the Department of Mechanical Engineering, Benha University, Egypt. he graduated from the Department of Mechanical Engineering in 2018, Benha University, Egypt. he is currently working on a master thesis about quality monitoring in multistage manufacturing systems using machine learning techniques. His research interests are focused on industrial AI, Smart manufacturing, Quality monitoring, industry 4.0 and process improvement. He assists in teaching the following courses to UG students; quality control, hydraulic and pneumatic power systems, design of experiments, mechanical design and strength of materials.

Ahmed El-Assal is currently an emeritus professor in the Department of Mechanical engineering and Director of Quality Management in Benha faculty of engineering, Benha University, Egypt. He received his M.Sc. and Ph.D. degrees from Helwan University, Egypt and George Washington University, Washington, USA, in 1982 and 1988, respectively. He has about 40 years of academic experience in various positions at Helwan, George Washington, King Abdul Aziz and Benha universities. His research interests include Solid Mechanics, Fracture Mechanics and Materials Behavior, Mechanics of Composite Materials and its Applications, Reverse Engineering and industrial engineering. He teaches the following courses to UG and PG students; Mechanical Design, Materials Engineering, Material Handling, Computer-Aided Drafting for Design (CADD), Computer-Aided Design (CAD), Computer-Aided Manufacturing (CAM), Industrial Automation, Manufacturing, Computer Programming with Pascal and C languages with Applications. In addition, he is supervising P.h.D, M.Sc. students and graduation projects.

Noha A. Mostafa is an Assistant Professor of Industrial Engineering and Management, Zagazig University. She is currently a lecturer in The British University in Egypt. She got B.Sc. and M.Sc. in Industrial Engineering from Zagazig University. She was a visiting PhD student in Tokyo Institute of Technology, Japan in 2016. She finished her PhD degree in 2017 from Egypt-Japan University of Science and Technology (E-JUST), Egypt; the topic was the integration between different functions of the supply chain. Noha has broad research interests including supply chain management, logistics, sustainability, quality management, design thinking, value engineering, data analytics, and information systems. She is the founder and faculty advisor of IEOM student chapter in Zagazig University.