*Advances in Manufacturing Technology XXXV*
*M. Shafik and K. Case (Eds.)*


*doi:10.3233/ATDE220605*

# Feature Reduction and Selection for Use in Machine Learning for Manufacturing

Duaa ALRUFAIHI [a,b], Omogbai OLEGHE [c], Mohammed ALMANEI [a], Sandeep JAGTAP [a] and Konstantinos SALONITIS [a,1]
[a] *School of Aerospace, Transport and Manufacturing, Cranfield University, Cranfield MK43 0AL, UK.*
[b] *Saudi Industrial Development Fund, Riyadh, KSA*
[c] *Systems Engineering, Department of Engineering, University of Lagos, Nigeria*

**Abstract.** In a complex manufacturing system such as the multistage manufacturing system, maintaining the quality of the products becomes a challenging task. It is due to the interconnectivity and dependency of factors that can affect the final product. With the increasing availability of data, Machine Learning (ML) approaches are applied to assess and predict quality-related issues. In this paper, several ML algorithms, including feature reduction/selection methods, were applied to a publicly available multistage manufacturing dataset to predict the characteristic of the output measurements in (mm). A total of 24 prediction models were produced. The accuracy of the prediction models and the execution time were the evaluation metrics. The results show that uncontrolled variables are the most common features that have been selected by the selection/ reduction methods suggesting their strong relationship to the quality of the product. The performance of the prediction models was heavily dependent on the ML algorithm.

**Keywords.** Machine Learning, algorithms, Complex manufacturing systems.

## 1. Introduction

Nowadays, most of the manufacturing industries include multiple operational stages called “Multistations” [1]. Multistations are being used in semiconductor, automotive and aerospace industries. This type of manufacturing has strict quality requirements as the final product is affected by all earlier operations. Therefore, quality-related issues presented in the final product are a combination of complex factors or errors carried out between stages [2]. Besides the increased complexity of the manufacturing process, the availability of data generated from these manufacturing industries has also increased. Data has become widely available with the revolution of Industry 4.0, and it has been mainly used to solve problems through interpretation and processing [3]. Machine learning (ML) approaches have utilized the availability of data to solve problems related to manufacturing and multistage manufacturing. Recent ML approaches to quality-related issues in multistage manufacturing have involved testing the ability of classification ML algorithms for their power in predicting defects or faults [4]. Moreover, a couple of articles have investigated impact of feature reduction or feature selection techniques on the prediction accuracy of faults and defects [5], [6].

[1] Corresponding Author. k.salonitis@cranfield.ac.uk

This study investigates the effect of three feature reduction/selection methods, namely Principal Component Analysis (PCA), Correlation Based Feature Selection (CFS), and ReliefF, on predicting the characteristic of the output measurement of multistage manufacturing. The research was conducted using the publicly available multistage manufacturing dataset from Kaggle [7]. The dataset consists of information related to two-stage manufacturing. Each stage consists of some input variables and output measurements. Although most previous studies have focused on the accuracy of the prediction models in defect or fault classification, this study focuses on predicting the characteristic of the output measurement and the execution time. Moreover, this study utilizes the feature selection/ reduction techniques available to understand the multistage manufacturing system and help in the process of decision making.

## 2. Literature review

### *2.1. Machine learning in manufacturing problems*

ML is considered a multi-disciplinary field and can be applied to real-world problems with inherited complexity. A wide range of problems can be solved using ML that can be classified into these major classes: classification, anomaly detection, regression, clustering, and reinforcement [8]. Two main types of ML are identified, namely supervised and unsupervised methods.

In *supervised learning*, the ML algorithm learns from the historical data provided. The training model in supervised learning is provided with the correct and expected outputs. Therefore, the algorithm learns by comparing its output with the given input. Supervised learning can handle both regressions and classification tasks. In *unsupervised learning*, the ML algorithm recognizes and drives a pattern or a trend from uncategorized data. It is usually used with unlabelled data to discover the hidden pattern. Supervised ML approaches are more suitable for problems in the manufacturing environment since the data that originated from the manufacturing settings are usually labelled. Furthermore, it is used for statistical process control, which can also serve ML purposes [9]. For the present study, the following three ML algorithms are used: Random Forest (RF), Artificial Neural Networks (ANN) and Support Vector Regression (SVR).

While ML in manufacturing shows a lot of promising advantages, some challenges are still present:

- The availability of the relevant data, which is the most common challenge. Raw manufacturing data include irrelevant information and duplicated values that can directly affect the performance of the ML algorithm. Also, obtaining the data during the manufacturing process and the security-related aspect of the data poses a major challenge for the research area in manufacturing.
- The pre-processing of the data can significantly influence the results. It is common for the dataset to have missing values, and there are ML techniques that can be applied to replace these missing values. This can affect the original dataset, so it needs to be done carefully where bias and negative influence are reduced.
- The selection of the appropriate algorithms is challenging as there are different problems and requirements. Also, with advances in research, there are a lot of ML algorithms available in manufacturing

### *2.2. Feature Reduction/ Feature Selection*

Several studies have investigated the impact of feature selection or feature reduction on the accuracy of the model's prediction, specifically in fault detections. The fault detection accuracy in semiconductor manufacturing has been investigated by studying a classification with (pass and fail cases) problem using four classification algorithms with boosting techniques for the imbalanced data and different feature selection reduction techniques, including PCA [5]. In the same study, the evaluation metrics included the TP rate, FP rate, Precision, and F-measure. The prediction modelling was done using Weka software. Three feature selections (Boruta, Multivariate Adaptive Regression Spline (MARS) and Principal Component Analysis (PCA) were studied on a semiconductor dataset for a classification problem [6]. The impact of feature reduction and feature selection techniques, including PCA and CFS, on the defect prediction of software data, using the evaluation metric of the area under the receiver operating characteristic curve (AUC) and (IQR) Interquartile Range has also been studied [10]. Finally, 46 feature selection techniques (including CFS and Relief) using two classifier algorithms on more than twenty software defect datasets were studied [11].

## 3. Machine Learning techniques used in the present study

Three ML algorithms have been used in this study: They are Random Forest (RF), Artificial Neural Network (ANN), and Support Vector Regression (SVR).

Random forest (RF) is an ensemble ML algorithm that has been used for classification and regression [12]. It works by splitting the input data into data samples for regression tasks until specific requirements are met. For example, the splitting mechanism would stop if there were no observed improvements or the variation in the sample data becomes very small. Splitting input data into more samples is guided by using the best predictor of a random subset. These sample data are then used to predict the class or the independent variable - the final prediction results from combining all the predictions from the sample data.

Artificial Neural Networks (ANN) process information in a way such as the neurons in the human brain hence the name [13]. One of the most common NN types used for regression tasks is the Multilayer Perceptron (MLP). The MLP network has three main layers. The first layer is the input layer which consists of the input nodes. The second layer is the hidden layer, and the third layer is the output layer. The interconnections of the network have a weight that can be adjusted during the training process. Backpropagation is used to train the MLP, which consists of assigning random weights at the beginning of the training process and adjusting the weight to achieve an effectively hidden layer that can minimize the prediction error.

Support vector regression (SVR) is considered a type of the support vector machine algorithm (SVM) for linear and non-linear regression tasks. The construction of the loss function of the support vector regression is dependent on the error value between the predicted and the real value. If the error is greater than the threshold $\varepsilon$, the function will be built [14].

## 4. Feature reduction and selection methods

Feature reduction and selection steps are important steps used in reducing the number of features of a model, therefore, helping avoid the problem of multicollinearity or what is called "the curse of dimensionality" [10]. In fact, only a few number of features can determine the prediction performance. However, there is a major difference between feature reduction and feature selection methods. The number of features is reduced by selecting a subset of features from the original dataset in feature selection. However, in feature reduction, the number of features is reduced by creating new features that combine the original features in the dataset.

The advantages of using feature selection lie in reducing storage requirements, improving the performance model, and increasing the model's speed [15]. In feature selection, the algorithm uses the characteristic of the original data as a guide in selecting the appropriate feature, such as the distances between variables or their statistical dependencies. Feature selection is divided into two main categories. They are the filter method and the wrapper method. However, the wrapper method is computationally expensive; therefore, it has been deselected for this study.

In the present study, Correlation Based Feature Selection (CFS), ReliefF and Principal Component Analysis (PCA) methods are used. CFS is a feature selection method that ranks the features based on their correlation using an evaluation function. ReliefF feature selection is a multivariant and ranker algorithm [16]. That works by randomly selecting instances and evaluating that against two nearest neighbors. Finally, PCA is a very common feature reduction technique used to reduce the dimensionality of datasets [10]. It works by converting the original dataset into a reduced number of features.

## 5. Case study and dataset

This research project is based on studying and analyzing the dataset available from Kaggle called "Multi-stage continuous- flow manufacturing process" [7]. It is a time-series dataset that is perfectly structured with 116 features (labeled columns) and 14088 observations (rows) with a sample rate of 1 Hz. The data originated from an automotive multistage manufacturing process with two stages where information is collected from various sensors to a programmable logic controller (PLC) to a database.

The dataset is related to an extrusion process where the output from each stage (product's outer surface) of the process is measured in (mm). Machines operate in a combination of parallel and series. As a result, 15 output measurements of outer surface properties are recorded in each stage. Weka software is used to analyze the data (developed by the University of Waikato in New Zealand [17]).

## 6. Analysis

In any ML project that includes big data sets, the first step is to understand the data. Figs. 1 and 2 highlight some of the initial analyses undertaken using R for understanding the data set. The correlation test using Spearman's correlation coefficient was performed using R programming. Spearman's correlation coefficient has been employed in a similar study to understand the relationships between features and the target variable [4].

Moreover, this step includes the removal of attributes that contain a large amount of missing data. Although the cut-off value for removing missing values is not defined in the literature, in a similar domain such as this article [6], features with 55% missing values were removed. Another study has set the threshold to 85%, which has resulted in the removal of 10 features [4]. For this study, the missing value cut-off is set to 70%, resulting in removing of 4 features.

For developing the prediction models, two outputs were selected as target attributes. The selection was performed randomly: from stage 1, the output measurement 1 was selected, whereas from stage 2, the output measurement 7. Following this, feature selection or feature reduction was applied to the dataset. The search method for CFS was BestFirst. However, PCA and ReliefF use the ranker research methods. In the ranker search method, the user has to set the number of the ranked features. Since there is no rule for selecting the best set of variables [18], the accuracy of the prediction performance decreases when the set of attributes is less than 50%. Therefore, the top 40 attributes were selected. For PCA, the dataset was transformed into 40 attributes as well. Before applying PCA, the data was standardized to have a 0 mean and 1 standard deviation.

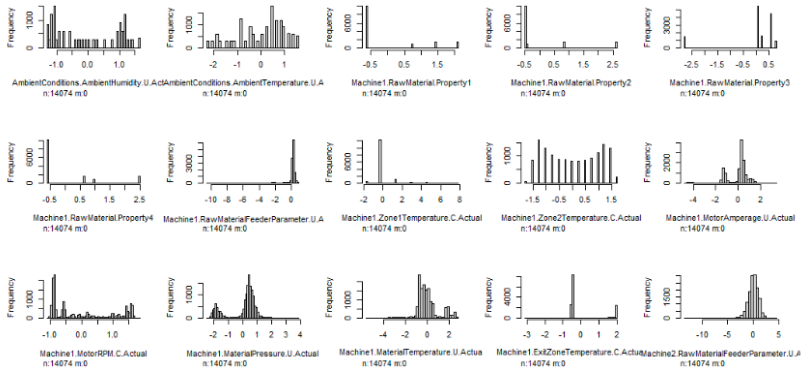


**Figure 1.** Distribution of some of the features using histogram after scaling the data.

The resulted features from the previous step were passed to the prediction models. When building the prediction models, data was split into 70% training and 30% testing ratios [19]. For the comparison, data without the application of feature selection or reduction is considered as the control group. Using the default configuration for the predictive models in Weka, 24 models were built for both stages.

Once the models have been trained, their performance is assessed using two criteria. The first criterion is the prediction accuracy using a regression task. Thus Root Mean Square Error (RMSE) is used to evaluate the accuracy. The second criterion is the total execution time required to build the model. Finally, a second experiment was performed to assess the performance and the impact of feature reduction/ selection on the case where the two stages are separated.

## 7. Results, analysis and discussion

### *7.1. Features selection*

Quality-related issues are of concern to manufacturers, particularly in the multistage industries. Therefore, this study has utilized three feature reduction and selection

methods (PCA, CFS, and ReliefF) to understand the features related to the characteristics of the output measurement and study their impact on the accuracy and execution time of the predictive models.

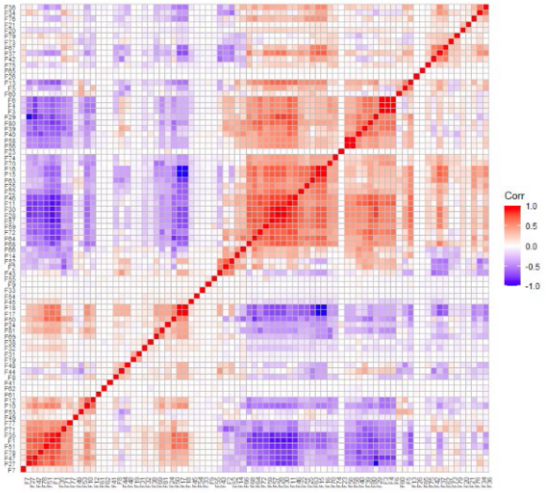


**Figure 2.** The correlation matrix using Spearman's correlation coefficient for the dataset.

The ReliefF method produced a list of 40 ranked features. The other filter method used: CFS has produced 6 features for the first stage and 9 for the second stage. The features produced from each method were classified into three categories: the stage at which they are measured, their control nature (controlled, uncontrolled or just raw material properties) and whether they are a stage output measurement.

A major percentage of the features filtered by the selection methods are uncontrollable features. In fact, most of the features that the three methods for both stages have picked up are also classified as uncontrollable features. It suggests their strong relationship with the output measurement. These uncontrollable factors, such as raw materials and ambient conditions, impose a challenge and can significantly impact on the quality of the products in multistage manufacturing [20].

### *7.2. Accuracy and execution time*

The prediction model results for the accuracy and the execution time are presented in Table 1. The worst and best case from each evaluation metric is colored in red and green respectively.

**Table 1.** Prediction model results for both stages

| | | Stage 1 | | Stage 2 | |
|---|---|---|---|---|---|
| | | Total time (s) | RMSE | Total time(s) | RMSE |
| Control | RF | 9.81 | 0.0772 | 11.57 | 0.0273 |
| | ANN | 312.2 | 0.1208 | 287.83 | 0.0402 |
| | SVR | 1813.7 | 0.2796 | 2293.17 | 0.0313 |
| PCA | RF | 18.13 | 0.1445 | 23.13 | 0.0291 |
| | ANN | 122.45 | 0.1433 | 128.05 | 0.0336 |
| | SVR | 471.96 | 0.2759 | 480.35 | 0.0319 |
| CFS | RF | 5.16 | 0.1048 | 6.98 | 0.0257 |
| | ANN | 4.31 | 0.2356 | 7.26 | 0.0304 |
| | SVR | 194.3 | 0.2847 | 238.21 | 0.0321 |
| ReliefF | RF | 161.34 | 0.0601 | 145.03 | 0.025 |
| | ANN | 235.94 | 0.1173 | 207.5 | 0.0302 |
| | SVR | 945.43 | 0.2735 | 880.4 | 0.032 |

The accuracy of prediction models has increased in some models but not in all. Some feature selection/ reduction has negatively impacted the accuracy of the models. Some major positive impacts on the accuracy happened at the cost of computational time. The accuracy and time execution are sometimes inversely proportional, except for some models where both criteria have improved. For example, stage 1 (ReliefF/ANN) has improved the accuracy by 3% and the time by 24% compared to the control. While (ReliefF/RF) has improved the accuracy by 22% while increasing the execution time by 16 times. This suggests that it is necessary for the user to define the requirements; if the accuracy is desired more than the execution time or if both criteria are important.

Within the three-feature selection/reduction methods, the performance of the prediction models varies depending on the regression algorithm. It indicates the dependability of the prediction performance on the selected ML algorithm [11]. Random forest showed its robustness in both aspects even without applying feature reduction/ selection.

Despite the wide range of applications of PCA, it did not perform well in terms of accuracy. The other filter methods generally have outperformed PCA. There are two possible reasons for that. The first reason concerns the dimensionality of the studied dataset. When the dimensionality of the data decreases, the performance of the PCA decreases as well. This is also demonstrated by the results of the second case, where each stage is evaluated separately, decreasing the dataset's dimensionality. This is also found in the work of [21]. The second reason is that filter-based methods such as CFS can outperform the performance of feature reduction when applied to supervised models [10].

The difference in the accuracy performance between the two stages was not as expected. A possible explanation is that the available data carries more information related to stage 2 rather than stage 1, and there is not enough data to explain the process of stage 1. The same was observed where the performance of prediction models decreased when both stages splitted, confirming that the available information was not enough to explain the whole system. A possible future work could be to test the method in more than 2 stages.

The application of PCA results on the data was a challenging task to interpret; because the result has been transformed into a linear combination of variables which makes the interpretation of the resulted features for the user more difficult than the features selection methods.

A possible limitation to this approach is the ranker-based search method used with PCA and ReliefF. Since there is no defined threshold to for the ranked attribute, someone might have to try a different proportion of the available features to find the best set without negatively affecting the accuracy.

## 8. Conclusion

This study has examined the impact of feature reduction and selection on multistage manufacturing datasets. The performance of the prediction models depends to a greater extent on the algorithm chosen. Some of the combinations of feature selection with RF show some promising results. Applying this approach to another data set with more than two stages could be an area to discover in future work. A great portion of the feature that has been selected is categorized as uncontrolled features suggesting their impact on the output of the manufacturing process. Monitoring these features could prevent the quality-

related issue from happening. In the case of raw materials, properties should be checked regularly to meet the customer's requirement and standards.

## References

[1] F. Tsung, et al. Statistical process control for multistage manufacturing and service operations: A review and some extensions. *International Journal of Services Operations and Informatics* 3 (2008), 191–204
[2] D. Djurdjanović, Y. Jiao, V. Majstorović. Multistage manufacturing process control robust to inaccurate knowledge about process noise. *CIRP Annals - Manufacturing Technology* 66(1) (2017) 437– 440.
[3] C. Shang, F. You. Data Analytics and Machine Learning for Smart Process Manufacturing: Recent Advances and Perspectives in the Big Data Era. *Engineering* 5 (2019), 1010–1016.
[4] R.S. Peres et al. Multistage Quality Control Using Machine Learning in the Automotive Industry. *IEEE Access* 7 (2019), 79908–79916.
[5] N. Kerdprasop. Feature selection and boosting techniques to improve fault detection accuracy in the semiconductor manufacturing process. *IMECS 2011 - International MultiConference of Engineers and Computer Scientists* 2011, 398–403
[6] D. Moldovan et al. Machine learning for sensor- based manufacturing processes. *IEEE 13th International Conference on Intelligent Computer Communication and Processing*, 2017, 147–154.
[7] Multi-stage continuous-flow manufacturing processI Kaggle. Available at: https://www.kaggle.com/supergus/multistage-continuousflow- manufacturing-process (Accessed: 18/08/2020)
[8] M. Almanei, et al. Machine Learning Algorithms Comparison for Manufacturing Applications. Advances in Manufacturing Technology XXXIV 2021, 377-382
[9] T. Wuest, et al. Machine learning in manufacturing: advantages, challenges, and applications. *Production & Manufacturing Research* 4(1) 2016, 23–45.
[10] M. Kondo, et al. The impact of feature reduction techniques on defect prediction models. *Empirical Software Engineering* 2019; 24(4): 1925–1963.
[11] A.O. Balogun et al. Impact of Feature Selection Methods on the Predictive Performance of Software Defect Prediction Models: An Extensive Empirical Study. *Symmetry* 2020; 12(7): 1147.
[12] K. Fawagreh, et al. Random forests: from early developments to recent advancements. *Systems Science & Control Engineering.* 2014; 2(1): 602–609
[13] C.F. Tsai, et al. Intrusion detection by machine learning: A review. *Expert Systems with Applications*. 2009. pp. 11994–12000
[14] J. Liu, et al. Comparative analysis of forecasting for air cargo volume: Statistical techniques vs. machine learning. *Journal of Data, Information and Management*. 2020; 1–13
[15] N. Sánchez-Maroño, et al. Filter methods for feature selection - A comparative study. *Lecture Notes in Computer Science* 2007. 178–187
[16] V. Bolón-Canedo, et al. On the scalability of feature selection methods on high-dimensional data. *Knowledge and Information Systems*. 2018; 56(2): 395–442
[17] M. Hall et al. The WEKA data mining software. *ACM SIGKDD Explorations Newsletter*. 2009; 11(1): 10–18.
[18] S. Vora, H. Yang. A comprehensive study of eleven feature selection algorithms and their impact on text classification. *Proceedings of Computing Conference 2017*. 440–449.
[19] E. Quatrini et al. Machine learning for anomaly detection and process phase classification to improve safety and maintenance activities. *Journal of Manufacturing Systems*. 2020; 56: 117–132.
[20] B. Lu B, X. Zhou. Quality and reliability oriented maintenance for multistage manufacturing systems subject to condition monitoring. *Journal of Manufacturing Systems*. 2019; 52: 76–85.
[21] G.T. Reddy et al. Analysis of Dimensionality Reduction Techniques on Big Data. *IEEE Access* 2020; 8: 54776–54788.