Proceedings of the ASME 2022
International Design Engineering Technical Conferences and
Computers and Information in Engineering Conference
IDETC-CIE2022
August 14-17, 2022, St. Louis, Missouri

DETC2022-89921

NORMALIZATION AND DIMENSION REDUCTION FOR MACHINE LEARNING IN ADVANCED MANUFACTURING

Jida Huang
Department of Mechanical and Industrial Engineering
University of Illinois at Chicago
Chicago, Illinois 60607
Email: jida@uic.edu

Tsz-Ho Kwok[*]
Mechanical, Industrial and Aerospace Engineering
Concordia University
Montreal, QC H3G 1M8, Canada
Email: tszho.kwok@concordia.ca

ABSTRACT

With the advances in sensing and communication techniques, data collection has become much easier in manufacturing processes. Machine learning (ML) is a vital tool for manufacturing data analytics to leverage the underlying informatics carried by data. However, the varieties of data formats, dimensionality, and manufacturing types hugely hinder the learning efficiency of ML methods. Data preparation is critical for exploiting the potential of ML in manufacturing problems. This paper investigates how data preparation affects the ML efficacy in manufacturing data. Specifically, we study the influences of data normalization and dimension reduction on the ML performance for various types of manufacturing problems. We conduct comparison studies of data with/without pre-processing on different manufacturing processes, such as casting, milling, and additive manufacturing. Experimental results reveal that different pre-processing methods have a distinct effect on learning efficiency. Normalization is helpful for both numerical and image data, while dimension reduction – this paper uses principal component analysis (PCA) – is not useful for low-dimensional numerical manufacturing data. Combining both normalization and PCA can significantly enhance the learning efficiency of high-dimensional data. After that, we summarize several practical guidelines for manufacturing data preparation for ML, which provide a valuable basis for future manufacturing data analysis with ML approaches.

1 Introduction

Smart manufacturing system is one of the critical elements for the fourth stage of the industrial revolution (Industry 4.0) [1]. One foundation for implementing the smart manufacturing system is data. Through data analysis, the manufacturing systems can understand the current operational status and predict the process condition to achieve the prognostics. In this paradigm, data has become a precious resource for systematic computational analysis of the manufacturing process, thus leading to more informed decisions and enhancing the efficacy of smart manufacturing [2, 3]. With the advances of sensing and communication technologies, data becomes much more accessible, and manufacturing data is experiencing explosive growth, which has reached over 1000 EB annually [4]. With the hidden intelligence of big data, manufacturing systems have become more "smart" to achieve the all-around monitoring, simulation, and optimization of production activities [5].

To leverage the value of information and knowledge embedded in data, machine learning (ML) has been a vital tool for exploring data capability. Currently, the primary use cases of ML in manufacturing are predictive quality and predictive maintenance. With machine learning approaches, predictive quality and yield automatically identify the root causes of process-driven production losses using continuous, multivariate analysis, powered by machine learning algorithms that are uniquely trained to understand each production process intimately [6]. ML methods help reduce common, painful process-driven losses (e.g., yield, waste, quality, and throughput) and optimize the production pro-

[*] Corresponding Author.

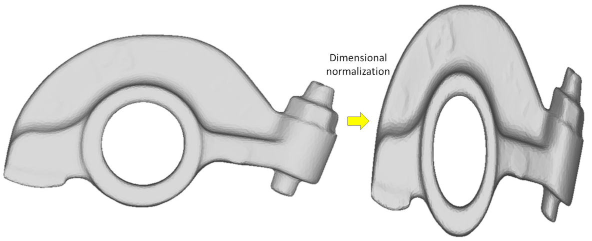

FIGURE 1: Applying normalization to a solid model by its *x*-, *y*-, and *z*-coordinates separately breaks the spatial relationship of the geometry.

cess, enabling the growth and expansion of product lines. Predictive maintenance ensures manufacturing systems continue to function without unnecessary interruptions by preempting a failure with ML algorithms. Predictive maintenance also leads to less maintenance activity, lower labor costs, and reduced inventory and materials wastage. To fully exploit the capability of ML methods for manufacturing systems, manufacturers need to know which data-driven solution is best suited for their own unique sets of challenges.

To find a suitable ML solution for a specific problem, the first thing is to cope with the data, known as data preparation. Since different manufacturing processes generate distinct data, it entails a great variety of formats, scale, and representations. Before selecting appropriate ML method, we need to handle the data variety and get the data prepared to applying ML for manufacturing problems. In data preparation, normalization and dimension reduction are two most fundamental methods for preprocessing the data. Studies [7,8] have shown that normalization is crucial to the performing ML. On one side, data normalization eliminates the biased weight because of the range of data and ensures the equity of features' numerical contribution from data. On the other side, it breaks the spatial relationship of the data. Fig. 1 illustrates that normalizing a solid model by its *x*-, *y*-, and *z*-coordinates separately makes a circle become an oval. As the features of the object may not equally important for the ML method, some features in the data can have a varying relevance while others are entirely irrelevant and redundant. The presence of unwanted features complicates the learning process and increases the feature space size. These features interfere with the useful features which confuse the learning algorithm and result in deterioration of the learning performance. It also increases the computational complexity of a machine learning algorithms [9]. Thus, the impact of normalization on the learning efficiency of manufacturing data needs investigation.

Dimensionality reduction also plays a vital role in machine learning, especially when working with data encompassing high-dimensional features (e.g., an image with a resolution of 64×64 contains 4096-pixel features). As manufacturing data may often contain noise or irrelevant information, which negatively affects the generalization capability of ML algorithms [10], dimension reduction can reduce the noise or irrelevant information of the data and also be beneficial for the computational efficiency of ML methods. However, reducing the dimension of the data loses some amount of information ingrained in data. Among various dimension reduction algorithms, Principal Components Analysis (PCA) is one of the most widely used methods [11]. PCA helps in data compression and hence reduces storage space and computation time. It also eliminates redundant features of data. However, PCA finds linear correlations between variables, which are sometimes undesirable. PCA can fail where mean and covariance are not enough to describe the datasets. Therefore, the impact of dimension reduction on the learning efficiency, especially for manufacturing data, needs further investigation.

The impact of data preparation on ML efficacy motivates the research question of this work. How do the data normalization and dimension reduction affect the performance of the ML methods for manufacturing data? To answer this question, we conduct normalization and dimension reduction (specifically PCA) on various types of data from typical manufacturing processes (e.g., CNC milling, 3D printing), and then compare the learning performance (in terms of the computational efficiency and prediction accuracy) of ML methods with and without data preparation. Through a comparison study, we can examine how normalization and dimension reduction affect the ML methods. Based on this, we derive the general manufacturing data preparation strategies, which could serve as guidelines for future manufacturing data preparation for ML.

We organized the rest of this paper as follows. Section 2 reviews the state-of-the-art data preparation for ML, and Section 3 introduces the study. Then, Section 4 presents the experiments and case studies, and Section 5 discusses the results and summarizes the design guideline for applying data preparation for manufacturing data. Finally, Section 6 concludes the paper and mentions the limitations and future directions of the work.

2 Related Works

There are abundant works that studied the integration of ML methods for manufacturing problems [12–15]. However, most of the current works focus on how to align the data-driven approaches with the manufacturing problem in specific applications. Other works [3,16] are focused on discussing the opportunities of embracing artificial intelligence with the manufacturing environment. For instance, Wang *et al.* [17] presented a review of deep learning methods and applications for smart manufacturing, and Wuest *et al.* [18] discussed the advantages, challenges, and applications of ML in manufacturing.

However, few works concentrate on the data preparation for ML in the manufacturing domain. Grzegorzewski *et al.* [19] discussed a systematic data pre-processing paradigm, including data quality and data preparation in industrial manufacturing. Some works [7, 8] investigated the effect of normalization on the ML performance, while others [20, 21] studied the dimension reduction methods for specific manufacturing problems. Rare works discussed the data preparation for learning efficiency. With the rapid development of artificial intelligence and the advances in manufacturing techniques, it urges the need of a systematic quantitative study of the influences of data preparation in ML methods for manufacturing data. This work does a comparative analysis of the data pre-processing methods on the effectiveness of ML methods.

3 Methodology

To investigate the impact of data preparation approaches on ML performance, we need to establish a fairly comparative framework. Figure 2 presents the overview of the proposed study paradigm. This method will first collect various types of data from typical manufacturing processes and then pre-process the collected raw data with and without the normalization and PCA. After that, it will apply the prepared data to ML methods and investigate the learning results comparatively.

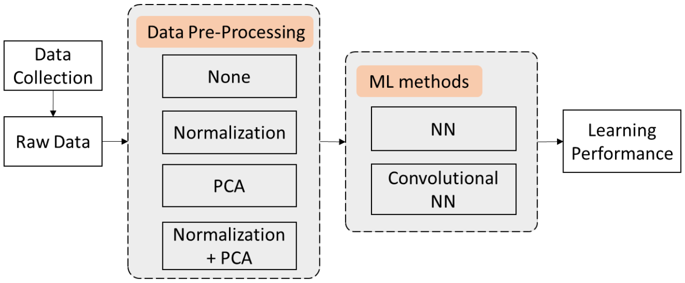

FIGURE 2: Overview of the proposed framework for machine learning of manufacturing data.

3.1 Data Collection

We mainly focus on analyzing data from the manufacturing domain in this work. Thus, we collect data from different manufacturing processes, such as milling, casting, and 3D printing. These data contain various features and represent typical manufacturing processes and products. Since each dataset will have its physical meanings and represented formats, this work will study two main categories: numerical parameters and image data. Numerical parameters usually describe the physical process parameters, while image data are directly the snapshots of the manufacturing processes or products.

3.2 Data Pre-Processing

This work will apply normalization and dimension reduction to the manufacturing data. This section presents the details of the two data pre-processing methods.

Normalization. The manufacturing datasets mainly include two types: numerical values and image data. The normalization for numerical data uses the standard score (z-score), which normalizes the data with center 0 and standard deviation 1. The main reason for selecting this method is that the parameters' value in the manufacturing process varies extensively, and there could be outliers. The z-score normalization transforms these different data scales into a similar numerical range without being affected severely by outliers exist. In terms of image data, since the intensity value of a pixel ranges from 0 to 255, there are no extremely large or small outliers. We apply a min-max normalization, transforming the minimum value of an image into 0 and the maximum value into 255. As image data contains the spatial relation of features, the min-max normalization helps preserve the relationships among the original input data.

Dimension reduction. This work mainly uses PCA for dimension reduction. PCA is a mathematical algorithm that reduces the dimensionality of the data while retaining most of the variation in the dataset [22]. Suppose the data has n observations with p numerical variables. The raw data can be represented as a $n \times p$ data matrix $\mathbf{X}$, whose j-th column is the vector x_j of observations on the j-th variable. We seek a linear combination of the columns of matrix $\mathbf{X}$ with maximum variance. Such linear combinations are given by $\sum_{j=1}^{p} \beta_j x_j = \mathbf{X}\beta$, where β is a vector of constants $\beta_1, \beta_2, \cdots, \beta_p$. The variance of any such linear combination is given by $var(\mathbf{X}\beta) = \beta^T S \beta$, where S is the sample covariance matrix associated with the dataset. Hence, identifying the linear combination with maximum variance is equivalent to getting a p-dimensional vector, which maximizes the quadratic form $\beta^T S \beta$. Viewed as an optimization problem, there are many approaches to find the optimal β values, such as singular value decomposition (SVD) [23].

3.3 Comparative Study Methods

To examine the impact of normalization and dimension reduction on learning efficiency, we employed ML methods to study various problems in manufacturing processes. We conduct the four pre-processing data options as shown in Fig. 2 on the input manufacturing data: 'None' means we take the raw data directly as input for ML methods, and we conduct normalization only, PCA only, and both on the raw data. We apply a fully connected neural network (NN) to both numerical and image data regarding the ML method. We also deploy the convolutional neural network (CNN) for the image data. With the different combinations of data pre-processing strategies and ML methods, we study and compare the learning efficiency in manufacturing data, thus providing insights on selecting appropriate data preparation for different manufacturing processes.

4 Experimental Results

We need to split the data for training, validation, and testing for the machine learning methods. In this work, we split the numerical manufacturing data (Section 4.1−4.3) into 0.7, 0.15, and 0.15, respectively; while for image data in Section 4.4−4.6, we separate them into 0.64, 0.16, and 0.2. We include the most significant components in dimensional reduction using PCA until the explained variance reaches 99%. All experiments are implemented in tensorflow[1] and run on a machine with 3.6 GHz 8-Core Intel Core i9 CPU and 16GB RAM.

4.1 Steel plates faults

Steel plates have many uses, ranging from household appliances to military. To manufacturing quality products from steel plates, it is important to inspect and correctly classify the type of surface defects in them. Here, we employ the Steel Plates Faults dataset [24], donated by Semeion, Research Center of Sciences of Communication, and it is licensed under a Creative Commons Attribution 4.0 International (CC BY 4.0) license. The dataset has 34 fields, including both inputs and outputs. The input vector comprises 27 indicators that describe the steel plate and the geometric shape of the defect, such as steel types, plate thickness, defect positions, areas, and perimeters. The output vector is a set of 7 one-hot encoded class labels (true or false) that classify the type of surface defects, e.g., stains, scratches, bumps.

TABLE 1: Accuracy of classifying the type of steel plate faults with various pre-processing options. (The larger the better)

Methods	NN size	Total	Train	Validate	Test
Raw	27-24-6	65.93%	63.63%	69.47%	73.16%
Norm	27-24-6	94.16%	95.27%	93.16%	90.00%
PCA	2-14-6	54.65%	55.07%	52.63%	54.74%
Both	16-18-6	94.32%	95.72%	92.63%	89.47%

We use cross-entropy minimization for training the classification model, and Table 1 summarizes the accuracy results for each pre-processing option, together with the size of NN. When the NN directly learns from the raw data, it finds the best result when the hidden layer has a size of 24, and the accuracy is about 65%. After normalization, the accuracy improves dramatically to about 94%, which reveals the importance of normalizing the data to eliminate the bias because of their ranges. When applying PCA to the raw data, it reduces the input dimension from 27 to 2 even when it requires holding 99% of the variance. This means that the bias is so significant, making all other parameters trivial. Because of this unfaithful loss of information, the accuracy reduces to about 55%, being the worst performance. After normalization, PCA reduces the input dimension to 16 instead of 2, which further confirms that the bias comes from the different ranges of data. Here, the accuracy is about 94%, similar to the one with only normalization. This shows that dimension reduction is not bad as long as the data is not biased, but it does not have much benefit for data that already has a small dimension.

4.2 Tool wear detection in CNC milling

In CNC milling, operators have to spend a lot of time checking if the tool is still good to use. There is a need for the identification of worn and unworn cutting tools based on the data from the machine's built-in sensors. We use the CNC milling dataset generated by the System-level Manufacturing and Automation Research Testbed (SMART) Lab at the University of Michigan. The dataset collects time-series data from 18 machining experiments with a sampling rate of 100 ms. The experiments carved an 'S' shape into the top surface of 2" × 2" × 1.5" wax blocks. Eight experiments used an unworn tool, while ten used a worn tool. The machining data contains the measurements from the 4 motors in the CNC (*x*-, *y*-, and *z*-axes and spindle), including tool condition, feed rate, clamping pressure, etc. We take every CNC measurement during the active machining operations as an independent observation and perform a supervised binary classification of the tool being unworn or worn. The input vector has 41 parameters, and there is only a Boolean output.

TABLE 2: Accuracy of tool wear detection with various pre-processing options. (The larger the better)

Methods	NN size	Total	Train	Validate	Test
Raw	41-42-1	61.52%	61.88%	59.66%	61.47%
Norm	41-36-1	92.10%	93.61%	88.77%	88.41%
PCA	2-46-7	55.58%	55.36%	56.16%	56.05%
Both	28-44-1	92.07%	94.19%	87.27%	86.91%

Table 2 summarizes the accuracy results for this test case. The results have the same trend as that of steel plates faults in Sec. 4.1. In short, the accuracy of learning from the raw data is about 62%, and normalization improves it to 92%. PCA on the raw data results in only 2 components remained and has a low accuracy of about 56%. Applying both normalization and dimension reduction gives a similar result (92%) to the one with only normalization. This further supports that normalization is important to the numeric manufacturing data, and dimension reduction is not meaningful because of the already small dimension.

[1] www.tensorflow.org

4.3 Surface roughness prediction in FFF 3D printing

Surface roughness is a major concern in 3D printing. It is useful to find out how the 3D printing parameters affect the surface roughness. We use a 3D printer dataset, which comes from the research by the Department of Mechanical Engineering at Selcuk University. The experiments used the Ultimaker S5 3D printer and filaments, and there are 50 observations. The dataset has both setting and measured parameters. The input vector takes 11 parameters, including later height, infill density, temperatures, print speed, etc.; and the output is the surface roughness.

TABLE 3: Mean squared error for predicting surface roughness from 3D printing parameters. (The smaller the better)

Methods	NN size	Total	Train	Validate	Test
Raw	11-11-1	7918	8093	7675	7418
Norm	11-19-1	311	59	809	883
PCA	4-10-1	8640	8424	9913	8288
Both	7-13-1	348	94	1183	590

For this regression problem, we use the scaled conjugate gradient to train the NN, and Table 3 reports the mean square errors in each case. Although the nature of this problem differs from the previous two (classification vs. regression), the learning performance has the same trend. With no pre-processing of the data, the error is as high as 7918. It decreases to only 311 with normalization. PCA on raw data reduces the input dimension from 11 to 4, and it has a large error of 8640; while PCA on normalized data has a slightly larger error than normalization only. This again confirms that normalization is important but not dimension reduction for numeric data, in both classification and regression problems.

4.4 Casting surface classification

Casting is one of the fundamental manufacturing processes, which usually pours a liquid material into a mold that contains a hollow cavity of the desired shape, and then allows the melt to solidify. A casting defect is undesired irregularities in the metal casting process. However, there are different defects arise in the casting process, like blowholes, pinholes, burr, shrinkage defects, mold material defects, pouring metal defects, metallurgical defects. These defects are unexpected anomalies and need to be detected timely in the casting process. Thus, an efficient quality inspection is critical to eliminate defective products. In this experiment, we use a dataset that contains images of parts from the casting process [25]. The dataset has 7348 gray-scaled images with the size of 100×100 pixels, and 57% of the images are defective parts. This experiment studies the effects of data with and without normalization and PCA on the neural network performance. Since the data format is in an image, we also applied the convolutional neural network (CNN) to study the learning efficiency. In CNN, the layers before the fully connected ones are also a dimension reduction process, since the filters and pooling operations reduce the original image size. Table 4 summarizes the learning accuracy and training time of different methods.

TABLE 4: Comparison of different approaches on defect detection accuracy of casting data. (The larger the better)

Methods	NN size	Train	Validate	Test	Train time (s/epoch)
Raw	10000-224-112-1	76.40%	77.87%	76.78%	2.27
Norm	10000-224-112-1	89.71%	79.66%	89.79%	2.28
PCA	1050-224-112-1	99.41%	96.43%	75.66%	0.53
Both	1050-224-112-1	99.35%	97.11%	97.20%	0.66
ConV	4608-224-112-1	98.87%	97.76%	98.88%	29.38

Table 4 shows that if we are using the raw data for the learning, each image contains 10,000 pixels, which is high-dimensional data, and the neural network achieves a 76.78% classification accuracy (part is good or defective) for the testing data. With normalization, the classification accuracy improves to 89.79%. This reveals that normalization is beneficial for eliminating the bias brought by the range within the original image. PCA reduces the dimension to 1050, and the training and validation accuracies improve significantly to 99.41% and 96.43%, but the testing one remains similar (75.66%). This reveals that the PCA can improve the training accuracy by reducing data dimensions. However, such dimension reduction is not beneficial to learn a faithful relation between the original data and the expected output, thus leading to the prediction ability on new data (testing data in this case) with lower fidelity. With normalization and PCA, we can see the prediction accuracy increases (97.2%). This confirms that the normalization is crucial for eliminating the bias included in original manufacturing data, no matter what type of formats (e.g., numerical values, image) the data is. After normalization, the PCA eliminates the redundant features carried by high-dimensional data, thus improving the testing accuracy from 89.79% to 97.2%. For image learning, CNN has been an effective method [26]. In this experiment, we apply it to the casting data. From Table 4, we can see the CNN (ConV[2] in the table) achieves a slightly better result than NN with both normalization and PCA; however, the training time of CNN is significantly

[2] The input dimension is the filters after convolution layers multiple the reduced image size, e.g., the image is reduced to 6×6 after pooling layers with 128 convolutional filters will result in $36 \times 128 = 4608$ features.

larger (29.38 $s/epoch$). This is mainly because there are a larger number of filters in convolution layers, even though the pooling operation has reduced the image size. The feeding dimension to the fully connected layer is 4608. The casting data studies show that both normalization and PCA are necessary for high-dimensional manufacturing data, and they are comparable to, but much faster than, CNN.

4.5 Steel surface defect classification

Defect inspection is a crucial step in guaranteeing the quality of industrial production. This experiment uses a steel plate defect inspection dataset [27]. This dataset collected six kinds of typical surface defects of the hot-rolled steel strip, i.e., rolled-in scale (RS), patches (Pa), crazing (Cr), pitted surface (PS), inclusion (In), and scratches (Sc). The original database includes 1,800 gray-scale images: 300 samples each of six different typical surface defects. Figure 3 illustrates some example images. To increase the variations and sufficiency for machine learning methods, it augmented the original data to six times the original data size, i.e., the augmented size of image data is 18,000. The image size in the dataset is 200 × 200 pixels, and thus the dimension of the raw data is 40,000. This experiment studies the effect of data with/without normalization and PCA on the neural network. Again, we also apply the CNN to study the learning efficiency.

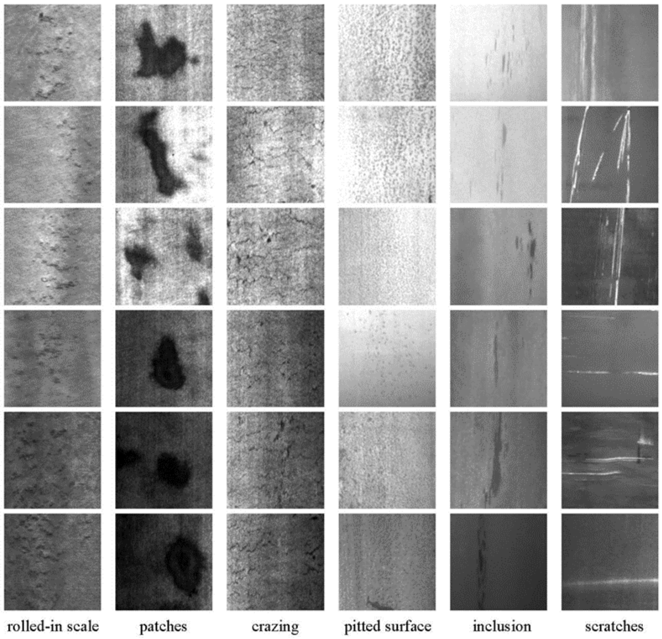

FIGURE 3: Example images of the steel plates defects [27].

Table 5 summarizes the learning accuracy and training time of different methods. This experiment shows that normalization can only slightly increase the defect detection accuracy in terms of training, validation, and testing. This reveals the data range bias within the high-dimensional data is not critical to the learning efficiency. PCA reduces the dimension largely from 40,000 to 721 and has a significant improvement in the testing accuracy from 14.72% to 66.39%. This reveals that the features highly correlate to each other within an image for defect detection. By eliminating the range bias among the data, the testing accuracy of PCA further improves to 81.76%. From these comparison studies, both PCA and normalization are critical to learning efficiency. It shows the effectiveness of eliminating the data range bias and finding the correlation encompassed within high-dimensional data. We also conducted the CNN on the steel surface classification. From the results, we can see that CNN achieves 95.28% testing accuracy. This is mainly because it can preserve the spatial correlation between features in convolution layers, while other methods transform the image into a vector, which loses the spatial relation. However, the computational cost of CNN is much higher than the NN with PCA and normalization. This implies that PCA and normalization with NN are still attractive if the computational time is a concern.

TABLE 5: Steel surface defect classification accuracy. (The larger the better)

Methods	NN size	Train	Validate	Test	Train time (s/epoch)
Raw	40000-224-112-1	17.15%	14.93%	14.72%	3.027
Norm	40000-224-112-1	28.12%	18.40%	17.01%	3.347
PCA	721-224-112-1	76.67%	68.89%	66.39%	0.052
Both	721-224-112-1	88.75%	82.29%	81.67%	0.050
ConV	9216-224-112-1	96.88%	95.83%	95.28%	21.53

4.6 Metal nut anomaly detection

The last dataset comprises manufacturing parts (specifically, metal nuts) anomaly detection, and it comes from Bergmann *et al.* [28, 29]. The dataset categorized the metal nuts into five classes: good (normal part with no defects), bent (the part is bent), color (the part has colors at a specific area), flip (the part is in an opposite orientation), and scratched (the part has scratches). Figure 4 shows some example images of these five categories. For the sake of computational efficiency, this experiment resizes the original images into 244 × 224. Thus, the raw data dimension is 50,716. We also conduct data augmentation on the dataset (increased from 313 images to 11,560). Table 6 summarizes differ-

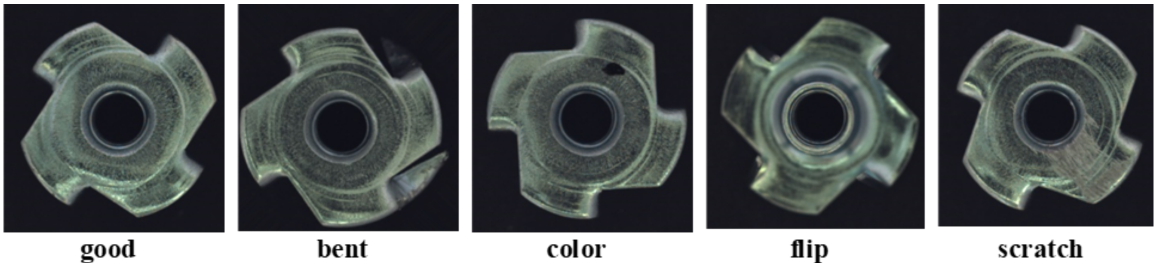

FIGURE 4: Example images of the metal nuts [28].

ent data pre-processing and their effects on learning efficiency.

TABLE 6: Metal nut defect detction accuracy. (The larger the better)

Methods	NN size	Train	Validate	Test	Training time (s/epoch)
Raw	50716-224-28-5	48.52	49.33	45.70	9.5459
Norm	50716-224-28-5	51.45	50.67	54.30	9.4434
PCA	4833-224-28-5	41.67	42.22	42.26	0.4151
Both	580-224-28-5	92.84	74.76	74.44	0.2529
ConV	25088-224-28-5	98.47	86.12	85.62	144.659

From the experimental results, we can observe that normalization slightly improves the learning accuracy, while only the PCA is not working. With both pre-processing of the high-dimensional data, the learning accuracy significantly improves to around 70-90%. This further confirms that elimination of the data range bias and correlation encompassed within data together is beneficial to the learning capability of NN. Similarly, CNN performs best by preserving the spatial relationship of features within the image while sacrificing the training time.

5 Discussion

In this work, we conduct a comparative study of the effect of data pre-processing on the learning efficiency of manufacturing data. Based on the analysis of the experimental results, we found several common patterns of the data pre-processing influences on manufacturing data across different applications. Based on the analysis of the experimental results, we found several common characteristics of the data preprocessing influences on manufacturing data across different manufacturing applications. Therefore, we summarize a few guidelines for data preparation for ML in manufacturing problems::

(1) **Normalization**: Normalization is essential for eliminating the bias brought by the scale of the data from different manufacturing processes either numerical values or images. It is especially important for numerical data, such as process parameters, to enhance the learning ability of ML methods.

(2) **PCA**: As a key dimension reduction method, PCA is beneficial to preserve the indispensable features of data. However, numerical manufacturing data usually contains low dimensional parameters, most of which have contributions to the learning objectives. Thus, PCA is undesired for numeric data pre-processing. While the image data is in a high-dimensional space, thus PCA is crucial for eliminating the redundancy information embedded in the data.

(3) **Image data**: We can get image data easily in manufacturing processes nowadays, and its structural nature brings the high-dimensionality of data. Thus, both normalization and PCA are inevitable pre-processing steps for image learning. A convolution neural network is attractive for manufacturing data analysis when computational resources are available.

Data pre-processing is crucial for the learning efficiency of ML methods in manufacturing data. This work summarizes the influences of data preparation for manufacturing data and provides several guidelines for the future developments of the ML paradigm in the intelligent manufacturing domain. We hope these guidelines serve as the groundwork for integrating the ML methods into manufacturing data analysis and solidify the basis for the establishment of the smart manufacturing system.

6 Conclusion

This paper studied the influences of data pre-processing on machine learning efficiency on various types of manufacturing data. We compared the data prepared with and without normalization and dimension reduction (PCA) for machine learning (ML). Based on the experimental results, we summarized several common design guidelines for data pre-processing within the manufacturing domain, thus providing a data preparation basis for future ML development in the manufacturing area. This work does not consider the size effect of image and other dimension reduction methods, such as manifold learning, which could be the future directions for data preparation in smart manufacturing.

Acknowledgement

This paper acknowledges the support of the Natural Sciences & Engineering Research Council of Canada (NSERC) grant #RGPIN-2017-06707.

REFERENCES

[1] P. Zheng, Z. Sang, R. Y. Zhong, Y. Liu, C. Liu, K. Mubarok, S. Yu, X. Xu, et al., Smart manufacturing systems for industry 4.0: Conceptual framework, scenarios, and future perspectives, Frontiers of Mechanical Engineering 13 (2) (2018) 137–150.

[2] G. Shao, S.-J. Shin, S. Jain, Data analytics using simulation for smart manufacturing, in: Proceedings of the Winter Simulation Conference 2014, IEEE, 2014, pp. 2192–2203.

[3] A. Kusiak, Smart manufacturing must embrace big data, Nature 544 (7648) (2017) 23–25.

[4] S. Yin, O. Kaynak, Big data for modern industry: challenges and trends [point of view], Proceedings of the IEEE 103 (2) (2015) 143–146.

[5] F. Tao, Q. Qi, A. Liu, A. Kusiak, Data-driven smart manufacturing, Journal of Manufacturing Systems 48 (2018) 157–169.

[6] J. Krauß, B. M. Pacheco, H. M. Zang, R. H. Schmitt, Automated machine learning for predictive quality in production, Procedia CIRP 93 (2020) 443–448.

[7] J. Sola, J. Sevilla, Importance of input data normalization for the application of neural networks to complex industrial problems, IEEE Transactions on nuclear science 44 (3) (1997) 1464–1468.

[8] D. Singh, B. Singh, Investigating the impact of data normalization on classification performance, Applied Soft Computing 97 (2020) 105524.

[9] G. Dougherty, Pattern recognition and classification: an introduction, Springer Science & Business Media, 2012.

[10] L. L. C. Kasun, Y. Yang, G.-B. Huang, Z. Zhang, Dimension reduction with extreme learning machine, IEEE transactions on Image Processing 25 (8) (2016) 3906–3918.

[11] I. K. Fodor, A survey of dimension reduction techniques, Tech. rep., Lawrence Livermore National Lab., CA (US) (2002).

[12] J. Huang, T.-H. Kwok, C. Zhou, W. Xu, Surfel convolutional neural network for support detection in additive manufacturing, The International Journal of Advanced Manufacturing Technology 105 (9) (2019) 3593–3604.

[13] J. Huang, L. J. Segura, T. Wang, G. Zhao, H. Sun, C. Zhou, Unsupervised learning for the droplet evolution prediction and process dynamics understanding in inkjet printing, Additive Manufacturing 35 (2020) 101197.

[14] J. A. Harding, M. Shahbaz, A. Kusiak, Data mining in manufacturing: a review (2006).

[15] G. Köksal, I. Batmaz, M. C. Testik, A review of data mining applications for quality improvement in manufacturing industry, Expert systems with Applications 38 (10) (2011) 13448–13467.

[16] A. Kusiak, Smart manufacturing, International Journal of Production Research 56 (1-2) (2018) 508–517.

[17] J. Wang, Y. Ma, L. Zhang, R. X. Gao, D. Wu, Deep learning for smart manufacturing: Methods and applications, Journal of manufacturing systems 48 (2018) 144–156.

[18] T. Wuest, D. Weimer, C. Irgens, K.-D. Thoben, Machine learning in manufacturing: advantages, challenges, and applications, Production & Manufacturing Research 4 (1) (2016) 23–45.

[19] P. Grzegorzewski, A. Kochanski, Data preprocessing in industrial manufacturing, in: Soft Modeling in Industrial Manufacturing, Springer, 2019, pp. 27–41.

[20] A. Chakraborti, H. P. Nagarajan, S. Panicker, H. Mokhtarian, E. Coatanéa, K. T. Koskinen, A dimension reduction method for efficient optimization of manufacturing performance, Procedia Manufacturing 38 (2019) 556–563.

[21] D. Zhao, Y. Bezgans, Y. Wang, W. Du, D. Lodkov, Performances of dimension reduction techniques for welding quality prediction based on the dynamic resistance signal, Journal of Manufacturing Processes 58 (2020) 335–343.

[22] M. Ringnér, What is principal component analysis?, Nature biotechnology 26 (3) (2008) 303–304.

[23] I. T. Jolliffe, J. Cadima, Principal component analysis: a review and recent developments, Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences 374 (2065) (2016) 20150202.

[24] M. Buscema, S. Terzi, W. Tastle, Steel Plates Faults, UCI Machine Learning Repository (2010).

[25] Kaggle, www.kaggle.com/ravirajsinh45.

[26] Y. LeCun, Y. Bengio, G. Hinton, Deep learning, nature 521 (7553) (2015) 436–444.

[27] Y. He, K. Song, Q. Meng, Y. Yan, An end-to-end steel surface defect detection approach via fusing multiple hierarchical features, IEEE Transactions on Instrumentation and Measurement 69 (4) (2019) 1493–1504.

[28] P. Bergmann, M. Fauser, D. Sattlegger, C. Steger, Mvtec ad–a comprehensive real-world dataset for unsupervised anomaly detection, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2019, pp. 9592–9600.

[29] P. Bergmann, K. Batzner, M. Fauser, D. Sattlegger, C. Steger, The mvtec anomaly detection dataset: a comprehensive real-world dataset for unsupervised anomaly detection, International Journal of Computer Vision 129 (4) (2021) 1038–1059.