An Integrated Method for High-Dimensional Imbalanced Assembly Quality Prediction Supported by Edge Computing

YIXIONG FENG1, (Member, IEEE), TIANYUE WANG1, BINGTAO HU1, CHEN YANG1,2, AND JIANRONG TAN1

1State Key Laboratory of Fluid Power and Mechatronic Systems, Zhejiang University, Hangzhou 310027, China
2Institute of Wanxiang Group, Hangzhou 311215, China

Corresponding author: Bingtao Hu (hubingtao@zju.edu.cn)

This work was supported in part by the National Key Research and Development Program of China under Grant 2019YFB1705004, in part by the National Natural Science Foundation of China under Grant 51775489, and in part by the Zhejiang Provincial Natural Science Foundation of China under Grant LZ18E050001.

ABSTRACT With the rapid expansion of Industrial Internet of Things, cloud computing and artificial intelligence, many intelligent information services have been developed in smart factories. One of the most important applications is helping factory managers predict the quality of assembled products. Traditional prediction methods of assembly quality mainly focus on building classification or regression models with high accuracy. However, less attention is paid to high-dimensional and imbalanced data, which is a special but common scenario at real-life assembly quality prediction. In this paper, we first use random forest to reduce dimension and analyze critical-to-quality characteristics. Then, a SMOTE-Adaboost method with jointly optimized hyperparameters is proposed for imbalanced data classification in assembly quality prediction. In addition, edge computing is introduced to improve the efficiency and flexibility of quality prediction. Finally, the practicality and effectiveness of the proposed method are verified by a case study of wheel bearing assembly line, and the experimental results show that the proposed method is superior to other classification methods in assembly quality prediction.

INDEX TERMS Assembly quality prediction, high-dimensional, imbalanced data classification, SMOTE-Adaboost, edge computing.

I. INTRODUCTION

With the popularization and application of the industrial internet of things (IoT), artificial intelligence (AI) and cloud computing (CC), various devices in factory are connected to industrial IoT through wired and wireless sensor networks. Meanwhile, these new technologies promote the developments of industrial big data and generate many new tasks and requirements [1]. In response, various AI-based tasks are deployed to cloud computing center. Many industrial intelligent information services have been developed based on CC and AI. Cloud-based manufacturing service system provides an effective way to store and process large amounts of data, which can meet the needs of various AI tasks in smart factories. However, there are still some problems [2]. First, data transmission between terminal and cloud center may suffer unacceptable latency, which may cause errors for decision makers. Besides, many AI tasks need to be performed in some complex scenarios. If these tasks only rely on cloud computing resources, they will cause network congestion.

Edge computing is a new type of distributed computing method [3] and it has been widely used in smart city [4], smart healthcare [5] and other industrial areas [6], [7]. Compared with cloud computing, edge computing can guarantee shorter response time and higher reliability. Lin et al. [8] proposed a Multiclass Deep Q Network method to solve the workshop scheduling problem based on an edge computing framework. Lai et al. [9] used LSTM to recognize industrial electrical equipment with edge computing. Li et al. [10] proposed a hybrid computing solution and resource scheduling strategy for edge computing in smart manufacturing. Feng et al. [12] utilized a green scheduling of sustainable flexible workshop
FIGURE 1. The industrial internet of things for assembly products quality management.

with edge computing considering uncertain machine state. Sun et al. [13] proposed a double auction-based resource allocation method for mobile edge computing in industrial IoT. In summary, many intelligent information services based on edge computing have emerged and gradually become an import part to promote the digitalization of smart factory.

Among intelligent information services in smart factories, one key issue is to manage product quality [14]. From the perspective of managing product quality, many scholars have elaborated this problem from different aspects, such as reliability [15], useful life [16] and retrievability [17]. Besides, another hot topic is how to predict product quality. Take assembly product for example, the quality of product often requires special expensive and complicated testing equipment. In addition, the testing process needs long time. Therefore, predicting the assembly quality quickly and efficiently is helpful for both providing decision-making service for factory managers and reducing assembly time. Many scholars have focused on data-driven product quality prediction and developed lots of explorations. Diao et al. [18] proposed a quality prediction method for purifier carrier product based on improved principal component analysis and modified support vector machine. Wei et al. [19] proposed a kernel-based hybrid manifold learning and support vector machines algorithm for aero-engine product quality prediction. Yao and Ge [20] proposed an enhanced nonlinear Gaussian mixture regression for quality prediction of chemical product. Zhang and Ge [21] proposed a local parameter optimization of LSSVM method for residual CO\textsubscript{2} content. Meng et al. [22] proposed an improved radial basis function (RBF) networks for effluent quality prediction in wastewater treatment process. These scholars have made great contributions to quality prediction, but there still exist some problems. First, the number of qualified products is far greater than that of fault ones in a stable assembly line (i.e. the labels of assembly product quality are imbalanced). Therefore, quality prediction problem is transformed into imbalanced data classification problem. Second, there exist many quality characteristics in assembly process, and critical-to-quality characteristics (CTQs) are some of the key characteristics of products related to final quality. CTQs selection is helpful to reduce the dimension of the predictive model. Besides, analyzing the causality of quality problems plays a significant role in improving product quality for factory managers. Third, some quality prediction methods based on CC may cause latency and overfull bandwidth. To overcome these shortages, we first introduce edge computing into assembly quality prediction to improve the flexibility of quality prediction. Then, we establish a random forest model to select and analyze CTQs. At last, we combine Synthetic Minority Over-sampling Technique (SMOTE) [23] with adaptive boosting (Adaboost) [24] to solve imbalanced data classification problem, where the hyperparameters of SMOTE and Adaboost are jointly optimized to enhance the classification ability.

In conclusion, the main contributions of this paper are summarized as follows:

1. We formulate a framework for assembly quality prediction in industrial IoT environment supported by edge computing, which provides guidance for flexible processing industrial data.
2. A SMOTE-Adaboost method with jointly optimized hyperparameters is proposed for imbalanced data classification problem.
classification in assembly quality prediction, and experimental results verify the superiority of our proposed method.

(3) Based on our proposed quality prediction method, we select and analyze CTQs in wheel bearing assembly line by Random Forest, which provides guidance and reference for wheel bearing product assembly in practice.

The rest of paper is organized as follows: we formulate a framework for assembly quality prediction based on edge computing in section 2. The details of Random Forest-based CTQs selection and analysis methods are discussed in Section 3. A SMOTE-Adaboost method with jointly optimized hyperparameters is proposed in Section 4. We take the wheel bearing assembly as a case study and conduct the comparative experiments to verify the effectiveness of the method in Section 5. Conclusions and some future works are shown in Section 6.

II. A FRAMEWORK FOR PRODUCT ASSEMBLY QUALITY PREDICTION SUPPORTED BY EDGE COMPUTING

This section builds a four-layer architecture of industrial IoT and describes how they work. Furthermore, we briefly introduce our assembly quality prediction method.

A. THE INDUSTRIAL INTERNET OF THINGS FOR ASSEMBLY PRODUCTS QUALITY MANAGEMENT

To manage quality of assembly products and achieve real-time prediction of product assembly quality, an industrial IoT framework is formulated as shown in Fig. 1, which is the basis for implementing industrial intelligent information services.

1) PERCEPTION LAYER

Perception layer connects various physical devices (such as dimensional inspection equipment, vibration detection equipment, visual detection equipment, etc.) in assembly line through data acquisition terminals (such as sensors, industrial cameras and RFID readers). It guarantees the real-time acquisition of massive and heterogeneous data in assembly workshop.

2) EDGE DEVICE LAYER

Edge layer is located between the perception layer and the cloud center layer. Edge devices can not only transfer data, but also use limited hardware (such as PLC, industrial PC, Raspberry Pi, etc.) resources to convert data format, preprocess data, and analyze real-time data, which can support edge computing and provide edge intelligent services.

3) CLOUD CENTER LAYER

Various types of heterogeneous data collected from the edge device layer are managed by cloud center layer resource. With these data, some machine learning algorithms can be implemented to help factory managers make decisions.

4) SERVICE LAYER

Service layer is the entry of industrial IoT for factory managers. It provides various intelligent information services such as product quality management, workshop equipment management and workshop resource scheduling.
B. ASSEMBLY QUALITY PREDICTION METHOD SUPPORTED BY EDGE COMPUTINGS

Fig. 2 shows an assembly quality prediction illustration supported by edge computing. The data from industrial IoT includes historical data and real-time data, so we describe the specific application services for historical and real-time data, specifically. The details of this method are described as follows.

1) FOR HISTORICAL DATA

**Step1:** Historical data are collected by perception layer and stored in cloud center layer. Firstly, CTQs are selected according to the selection rules under the cloud computing environment. Secondly, the quality predictive model based on machine learning method is trained, where hyperparameters optimization technique is used to improve classification performance. Finally, the quality predictive model is established.

**Step2:** Deploying the CTQs selection rules and predictive models mentioned in Step 1 to the edge device layer.

2) FOR REAL-TIME DATA

**Step1:** Real-time data collected by perception layer are transmitted to the edge device layer. With the help of the edge device layer, the data pre-processing process is implemented. Data selection is performed according to the CTQs selection rules transmitted from the cloud center layer. Finally, the quality of the product is judged in real time based on the existing predictive models at the edge device layer.

**Step2:** The data pre-processed in the edge device layer is labeled by vibration detection equipment and temporarily stored in the edge device layer. The predictive model in edge device is regarded as an aided prediction tool for factory managers in this time. These labeled-data are transmitted to the cloud center layer when the network is free. Thus, the bandwidth pressure is reduced. The amount of data becomes larger when the sensor data are transmitted to the cloud center, which causes the data distribution change over time. Therefore, the existing quality predictive models in the cloud center layer will be updated with these time-insensitive data by some incremental learning methods. In this way, it helps to solve the problem caused by dynamic data [41], [42].

**Step3:** The updated predictive model is periodically transmitted to the edge layer when the network is free.

C. BRIEF DESCRIPTION OF THE PROPOSED ASSEMBLY QUALITY PREDICTION METHOD

A SMOTE-Adaboost with jointly optimized hyperparameters method is proposed for assembly quality prediction based on historical data. The flowchart of the proposed method is presented in Fig. 3. The proposed method includes two parts. The first part is to select and analyze the critical-to-quality characteristics (CTQs) in assembly process. To be more specific, we preprocess the initial data, and calculate quality characteristic importance based on random forest. With the help of sequential forward selection (SFS) strategy, we select the CTQs and get the data after selection. The second part is to predict the final assembly quality, where the training set and testing set are divided based on stratified sampling. After that, SMOTE algorithm is utilized to obtain balanced data set, and the Adaboost algorithm is used for quality classification. In addition, we utilize grid search method to optimize the hyperparameters of SMOTE and Adaboost jointly. Finally, we obtain the best quality predictive model and use it to predict the product quality. The details of the method are described in section 3 and section 4.

III. SELECT THE CTQs IN ASSEMBLY PROCESS BASED ON RANDOM FOREST

To improve the accuracy of quality prediction and select the CTQs related to final quality, this section focuses on the selection and analysis of CTQs in the assembly process, including two steps: data preprocessing and CTQs selection based on Random Forest.

A. DATA PREPROCESSING

The label of the dataset is text variable, so we convert non-numeric labels to numeric labels. Specifically, we convert the quality labels (OK, NG) to (0, 1), respectively.

In a stable assembly line, qualified products account for the majority, and only few products have quality problems. From the perspective of data mining, this phenomenon will cause imbalance in label category. Besides, dozens of quality characteristics are simultaneously measured in assembly line, which means the data have high-dimensional features. From [25] we found that SMOTE has little effect on most classifiers trained on high-dimensional data, so dimension should be reduced at first.

Due to the serious imbalance of the data, if the dimension reduction is performed directly, the result will be invalid [26]. In order not to add new data and reduce the amount of calculations, we generate a small balanced data set with random under-sampling strategy. After that, the balanced dataset is used for dimension reduction.

B. CTQs SELECTION BASED ON RANDOM FOREST

The whole assembly process includes manufacturing and assembly of parts. The assembly process itself contains many features. Moreover, the more parts an assembly product contains, the more quality characteristics the manufacturing process has. Therefore, there are more quality characteristics in whole assembly process compared with other machining process (such as casting, milling, etc.). Take wheel bearing for example, the whole assembly process includes 24 quality characteristics in manufacturing process and 14 quality characteristics in assembly process. Therefore, the whole assembly process usually has dozens of quality characteristics, which makes dataset contains more features. Thus, it is necessary to select CTQs in assembly line. Data-driven CTQs selection is usually regarded as feature selection process. Random forest (RF) is an ensemble learning algorithm based on bagging strategy. The core idea of this algorithm is to use
bootstrap strategy to obtain sample set and establish several
decision trees, where the final classification is determined
by majority voting according to the results of decision tree
classification. In the process of training the RF model, the
importance of each feature is also calculated, so RF can be
used as an embedded feature selection method [27]. The
sequential forward selection (SFS) strategy [28] is also intro-
duced to decide the numbers of CTQs.

The details of the method are as follows:

Step1: Calculate the importance of each quality characteristic.
(a) Establish a random forest model consisting of $D$ CART
decision trees. At each node of the CART decision tree, we
divide the nodes according to the Gini index:

$$i(t) = 1 - P_{1}^{2} - P_{-1}^{2}$$ (1)

where $P_{1}$ means the probability that the selected sample
belongs to the qualified product, and $P_{-1}$ means the proba-
bility that the selected sample belongs to the default product.
(b) Calculate the change in the Gini index of the feature
at node $t$, $l$ and $r$ are the two child nodes after the branch of
node $t$:

$$\Delta i(t) = i(t) - P_{l}i(t_{l}) - P_{r}i(t_{r})$$ (2)

(c) Sum up the impurity changes for all nodes $t$ that used
$X$, and then averaged over all decision trees $D$ in the Ran-
don forest. The importance of each feature is calculated as follows:

$$FI(X) = \sum_{D} \sum_{l} \Delta i_{l}(t, D)$$ (3)

Step2: Sort the features according to their importance, and
we get the feature set $Q$, where $X_{n}$ means nth CTQ:

$$Q = \{X_{a}, X_{b}, \ldots, X_{n}\}$$

Step3: Use RF as the base learner, and add a feature from
the set $Q$ each time. Each iteration uses cross-validation to
calculate the Area Under Curve (AUC) score, and use the
maximum AUC score as the optimization target to obtain the
best feature subset $Q'$. Besides, we calculate the importance of each quality charac-
teristics according to formula (3), and use step 2 to sort the
quality characteristics in the assembly process. This tech-
nique not only provides an analysis and decision-making
method for factory managers in quality problems, but also
helps factory managers to focus on monitoring and optimiz-
ing CTQs.

IV. A QUALITY PREDICTION METHOD BASED ON
SMOTE-ADABOOST WITH JOINTLY OPTIMIZED
HYPERPARAMETERS
In order to predict the assembly quality under the condition of
data imbalance, this section builds a predictive model for the
historical data in the assembly line, including two steps: the
establishment of the predictive model and the optimization of
hyperparameters.

A. PREDICTION OF ASSEMBLY QUALITY BASED ON
SMOTE AND ADABOOST
For the imbalanced data set, we use stratified sampling to
divide the original data set into training and testing sets to
ensure that the proportion of the training and testing set
categories is similar to the original data set.

SMOTE is a common oversampling method. The main idea of
the algorithm is to synthesize new samples from existing
minority samples to deal with imbalanced data [23], which
is widely used for data imbalance classification in concept
drift detection [29], fault detection [30] and other industrial
scenarios [31], [32]. Adaboost is an ensemble learning algo-
rithm based on the boosting framework [24]. Compared with
other machine learning algorithms such as Random Forest
and Support Vector Machine (SVM), this algorithm pays
more attention to classification error samples, and iteratively
improves the prediction performance of models. This method
is widely used in various industrial studies (such as network
intrusion detection [33], smart healthcare diagnosis [34], etc.)
due to its state-of-the-art performance. Therefore, we com-
bine the advantages of SMOTE and Adaboost algorithms
to form a new imbalanced classification algorithm called
SMOTE-Adaboost.

The details of the method are as follows:

Step1: According to the original imbalanced data set, calcu-
late the number of samples that need to be synthesized for
the minority class sample $N_{min}$ and the majority class sample
$N_{max}$:

$$N = \frac{N_{max}}{N_{min}} - 1$$ (4)

Step2: For the minority samples $x_{j}$, find $k$ neighboring
similar sample points based on Euclidean distance. After that,
randomly select several sample points and synthesize new
sample points according to the following formula:

$$X_{new} = X_{j} + rand(0, 1) \times (X_{j} - X_{w})$$ (5)

where $X_{w}$ represents the wth minority sample point.
Step3: Repeat step 2 to synthesize $N'N_{min}$ new samples,
combine with the original data set and get a balanced data
set.

Step4: Initialize the weight distribution of the training data
according to the balanced data set in step 3.

$$D_{1} = (w_{11}, w_{12}, \ldots, w_{1N}), w_{1i} = \frac{1}{N}, i = 1, 2, \ldots, N$$ (6)

Step5: For $m = 1, 2, \ldots, M$
(a) Use the training set with weight distribution $D_{m}$ to train
CART to get the base classifier.

$$G_{m}(x) : \chi \rightarrow \{1, -1\}$$ (7)

(b) Calculate the error of $G_{m}(x)$ in the training set

$$e_{m} = \sum_{i=1}^{N} P(G_{m}(x_{i}) \neq y_{i})$$ (8)
(c) Calculate the coefficient of $G_m(x)$

$$\alpha_m = \frac{1}{2} \log \frac{1 - e_m}{e_m}$$  \hspace{1cm} (9)

(d) Update the weights of the training set

$$D_{m+1} = (w_{m+1,1}, w_{m+1,2}, \ldots, w_{m+1,N})$$ \hspace{1cm} (10)

$$w_{m+1, i} = \frac{w_{mi}}{Z_m} \exp(-\alpha_m y_i G_m(x_i)), \hspace{0.5cm} i = 1, 2, \ldots, N$$ \hspace{1cm} (11)

$$Z_m = \sum_{i=1}^{N} w_{mi} \exp(-\alpha_m y_i G_m(x_i))$$ \hspace{1cm} (12)

where $Z_m$ is the normalization factor.

The learner $f(x)$ iterates according to the following formula:

$$f_m(x) = f_{m-1}(x) + \alpha_m G_m(x)$$ \hspace{1cm} (13)

In order to prevent overfitting, we added the regularization coefficient $v$, so the iteration formula is transformed into:

$$f_m(x) = f_{m-1}(x) + v\alpha_m G_m(x)$$ \hspace{1cm} (14)

Step6: Construct a linear combination of the base classifier, and finally get the classifier.

$$G(x) = \text{sign} \left( \sum_{m=1}^{M} v_m \alpha_m G_m(x) \right)$$ \hspace{1cm} (15)

### B. SMOTE-ADABOOST WITH JOINTLY OPTIMIZED HYPERPARAMETERS

We notice in step2, 5 and 6 that SMOTE algorithm and Adaboost algorithm all contain hyperparameters, which need to be set before algorithms are trained. The setting of hyperparameters affects the performance of the predictive model. Previous studies mainly focus on the hyperparameters in the classification or regression model. Moreover, the hyperparameter of SMOTE also affects the classification results. Therefore, we consider SMOTE and Adaboost as a whole and propose a SMOTE-Adaboost with jointly optimized hyperparameters method to improve the performance of the quality predictive model. Specifically, we focus on $k$ in SMOTE (the number of nearest neighbors of the selected few samples), $m$ (the number of decision trees) and $v$ (regularization coefficient) in Adaboost. The maximum AUC score is chosen as the optimization goal to obtain the best hyperparameters.

The main parameters of hyperparameter optimization methods [35] are grid search method, random search method, heuristic algorithm and so on. In this paper, the process of training model is deployed on cloud-center, so we concern more about the recognition accuracy of the predictive model than the time it takes to train the model. Therefore, we optimize the above three hyperparameters based on the grid search method to obtain the best quality predictive model.

### C. PERFORMANCE METRICS

For classification problems, confusion matrix is often used to measure the performance of the algorithms. As shown in Table 1, in this case, confusion matrix contains the following four cases:

1. True positive (TP): correctly classified as a qualified product.
2. False positive (FP): incorrectly classified as a qualified product.
3. True negative (TN): correctly classified as a fault product.
4. False negative (FN): incorrectly classified as a fault product.

In this paper, five classification performance indicators are chosen as performance metrics. The specific calculation formula is as follows.

The accuracy (ACC) represents the ability of the model to classify correctly:

$$\text{accuracy} = \frac{TP + TN}{TP + FN + FP + TN}$$ \hspace{1cm} (16)

The recall (also called sensitivity) indicates how many qualified products in the sample are predicted correctly.

$$\text{recall} = \frac{TP}{TP + FN}$$ \hspace{1cm} (17)

Specificity indicates how many fault products in the sample are predicted correctly. The physical significance of specificity in our assembly quality prediction is the probability of detecting the fault products. In other words, the higher the specificity score, the stronger the ability of the model to detect fault products.

$$\text{specificity} = \frac{TN}{TN + FP}$$ \hspace{1cm} (18)

G-mean is a comprehensive indicator for evaluating the performance of imbalanced data models. It is defined as the geometric mean of recall rates for all categories, and it takes into account both sensitivity and specificity. The higher the G-mean, the better the classification performance.

$$G – \text{mean} = \sqrt{\frac{TP}{TP + FN} \times \frac{TN}{TN + FP}}$$ \hspace{1cm} (19)

AUC is defined as the area enclosed by the coordinate axis under the ROC curve. It is a comprehensive performance classification indicator, which is commonly used to measure classification performance [36]. The higher the AUC, the better the algorithm performance. Therefore, AUC and G-mean are two comprehensive metrics that consider both the qualified products and fault products.

| Actual | Predicted  | Qualified product | Fault product |
|--------|------------|------------------|--------------|
|        |            | True Positive (TP) | False Negative (FN) |
|        |            | False Positive (FP) | True Negative (TN) |

### TABLE 1. Confusion matrix for performance evaluation.
In this section, the industrial IoT of wheel bearings assembly line is implemented to verify the practicality and effectiveness of proposed method. The experiments include two parts: the process of CTQs selection and the classification results of proposed method. We finally summarize the experimental results.

A. EXPERIMENTAL BACKGROUND

Wheel bearings are important parts of cars, and their quality affects the life of the entire car. The whole assembly process includes the manufacturing and assembly of flanges, outer rings and other parts in a wheel bearing assembly plant. As shown in Fig. 4, wheel bearings contain 11 parts, and there are many quality characteristics tests in the entire assembly line. As a complex assembly product, wheel bearings have high requirements for their assembly accuracy. Besides, they also have to meet qualified standards for performance.

The final assembly quality of the wheel bearing is tested by special vibration detection equipment, where the result is qualified or fault. The vibration detection equipment will apply a load to the product and make it rotate at a speed. This process needs a lot of time. Furthermore, vibration detection equipment requires customized special machines, which are expensive. Therefore, predicting the assembly quality of wheel bearing products with data-driven methods have the potentials to replace special equipment, which can save equipment cost and testing time.

Fig. 5 presents the framework of a wheel bearing assembly line in industrial IoT supported by edge computing. The whole assembly line includes manufacturing machines and assembly machines, and each machine is equipped with sensors to monitor the quality characteristic of the wheel bearing product. All the machines are connected to the industrial computer (though OPC UA, WIFI, etc.), which can be regarded as an edge device and provides data storage and computing capabilities. The industrial computer can upload or download the data and the model with private cloud center though HTTP/MQTT protocol.

To verify the effectiveness of our proposed assembly quality method, a historical dataset was collected from private cloud center as data sources for overall assembly quality prediction analysis. The dataset includes 38 quality characteristics and 1 final quality label (qualified or fault products). All quality characteristics are continuous random variables and there exist two types of distributions includes Gaussian distributions and non-Gaussian distributions, as shown in Fig. 6. Table 2 shows the specific quality characteristics of the partial samples. There are 2888 samples in the dataset, including 2786 samples of qualified products and 102 samples of fault products. The imbalance ratio of the dataset is about 27.3:1.

B. EXPERIMENTS AND RESULTS

1) PROCESS OF CTQ SELECTION

To reduce the dimension of predictive model and analysis the CTQs in assembly line, random forest is implemented to calculate the relationship between final quality and each quality characteristic. We calculate the importance of each quality characteristic, and sort them in a descending order, as shown in Fig. 7. Through the sequential forward selection (SFS) strategy, we finally select two quality characteristics in the assembly process and six quality characteristics in the
part of manufacturing process as the CTQs, and the details of CTQs are in Table 3. This method provides a basis for optimizing product quality.

2) CLASSIFICATION RESULTS WITH OTHER MACHINE LEARNING METHODS

In order to verify the effectiveness of the proposed predictive model, we take five performance indicators (ACC, AUC, recall, specificity, and G-mean) commonly used in classification algorithms as the evaluation indicators of quality predictive models. All the experiments in our study are deployed in python 3.6 environment and are run on a desktop computer equipped with a dual 2.3 GHz Intel i5 processor and 8 GB RAM.

After CTQs selection, the stratified sampling strategy is used to divide the training set and testing set to ensure that the category proportion of each data set is same. 5 cross validation is utilized in all experiments to avoid overfitting, and the description of training set and testing set in each iteration is shown in Table 4. It is worth noting that SMOTE is applied during cross-validation to avoid overoptimistic, as said in [43], [44]. We use the training set to establish the SMOTE-Adaboost predictive model, and jointly optimize the hyperparameters of the model by grid search method. The description of dataset after SMOTE is shown in Table 5. Finally, the value of each hyperparameter of SMOTE-Adaboost is determined to be $k = 6$, $m = 110$, and $v = 0.9$. We name the optimized quality predictive model as SMOTE-Adaboost_m.

To verify the classification effect between different classification methods under the same standard, we use the same

### TABLE 2. The detailed description of quality characteristic.

| Process               | Quality characteristics | sample1 | sample2 | sample3 | sample4 | sample5 |
|-----------------------|-------------------------|---------|---------|---------|---------|---------|
| Manufacturing process | Outer diameter (mm)     | 84.950  | 84.959  | 84.944  | 84.949  | 84.951  |
|                       | Outer flange channel waviness (mm) | 0.133   | 0.429   | 0.133   | 0.239   | 0.314   |
|                       | Inner circle channel roundness (mm) | 1.349   | 1.142   | 0.800   | 0.216   | 0.425   |
|                       | Ball bore deviation (mm)  | -0.177  | -2.020  | -1.330  | -1.374  | -0.958  |
|                       | Magnetic encoder parallel difference (mm) | 0.182   | 0.009   | 0.047   | 0.176   | 0.017   |
| Assembly process      | Flange outside diameter (mm) | 56.994  | 56.995  | 56.970  | 56.975  | 56.996  |
|                       | Position 1 diameter jump | 0.014   | 0.011   | 0.015   | 0.010   | 0.008   |
|                       | Position 2 diameter jump | 0.010   | 0.006   | 0.012   | 0.007   | 0.005   |
|                       | Bearing clearance(µm)   | -21.100 | -18.300 | -22.700 | -29.100 | -20.200 |
|                       | error of parts(µm)      | 121.500 | 113.700 | 75.100  | 105.700 | 114.900 |
|                       | Scaling ring pressing position(µm) | 32.738  | 32.708  | 32.705  | 32.709  | 32.745  |
|                       | Position 1 pressure(N)  | -560.300 | -566.800 | -638.900 | -517.700 | -671.700 |
|                       | Position 2 pressure(N)  | 39570.300 | 39072.300 | 39403.200 | 39275.400 | 38974.000 |
|                       | Magnetic encoder press position(µm) | -0.100  | -0.104  | -0.133  | -0.104  | -0.098  |
|                       | The depth of the installation(µm) | 32.678  | 32.695  | 32.688  | 32.681  | 32.699  |

| Final quality | vibration testing | Qualified | Qualified | Qualified | Fault | Qualified |
|---------------|-------------------|-----------|-----------|-----------|-------|-----------|

### TABLE 3. The detailed description of CTQs.

| Variable number | critical-to-quality characteristics (CTQs)       | Process              |
|-----------------|-------------------------------------------------|----------------------|
| V22             | Inner ring channel waviness                     | Manufacturing process|
| V27             | Bearing clearance                               | Assembly process     |
| V7              | Outer flange channel waviness                   | Manufacturing process|
| V1              | Inner circle channel roundness                  | Manufacturing process|
| V5              | Ball bore deviation                             | Manufacturing process|
| V29             | error of parts                                  | Assembly process     |
| V19             | Alignment between raceway                       | Manufacturing process|
| V13             | Ball waviness                                   | Manufacturing process|

### TABLE 4. Description of each dataset.

| Dataset          | Qualified product | Fault product | Total  |
|------------------|-------------------|---------------|--------|
| Training set     | 2228              | 82            | 2310   |
| Testing set      | 558               | 20            | 578    |
| Total            | 2786              | 102           | 2888   |

### TABLE 5. Description of each dataset after SMOTE.

| Dataset          | Qualified product | Fault product | Total  |
|------------------|-------------------|---------------|--------|
| Training set     | 2228              | 2228          | 4456   |
| Testing set      | 558               | 20            | 578    |
| Total            | 2786              | 2248          | 5034   |
Y. Feng et al.: Integrated Method for High-Dimensional Imbalanced Assembly Quality Prediction

3) CLASSIFICATION RESULTS WITH OTHER SAMPLING ALGORITHMS

To verify the effects of sampling algorithm, we select different sampling algorithms (SMOTE without optimization, Adaptive Synthetic Sampling, and without sampling) to conduct the experiments. The results are shown in Table 9, which indicate that the proposed method tend to have better generalization results. Although the proposed method is not as good as the predictive model without sampling strategy in ACC and recall score, the specificity of the predictive model without sampling strategy is very low, which means this predict model has a poor ability to detect fault products.

4) CLASSIFICATION RESULTS CONSIDER THE CTQS SELECTION

We conduct the experiment with CTQs selection and without CTQs selection. As the results shown in Table 10, the predictive model selected by CTQs not only has better prediction
effect in AUC, specificity, G-mean score, but also reduces the calculation amount of the training model.

C. DISCUSSION

In summary, some conclusions can be drawn from the above results.

From the perspective of high-dimensional data, Table 10 shows that the classification result of predictive model with CTQs selection is better, which means there are surely some quality characteristics (as shown in Table 3) in the assembly process that are redundant or have few correlations with final quality. Besides, Fig. 7 demonstrates the importance of each CTQ. Thus, selecting the CTQs is crucial for predicting assembly quality and analyzing quality problems.

In terms of imbalanced data, Table 6 and Table 9 show that the combination of SMOTE and Adaboost outperforms the integration of other sampling algorithms (i.e. ADASYN) and classification algorithms (i.e. SVM, Logistic Regression, Decision Tree, Random Forest). Moreover, Table 8 means that the SMOTE-Adaboost method with jointly optimized hyperparameters can improve the classification performances. This result also shows that our proposed method is helpful for imbalanced data classification problems.

Accuracy, recall, AUC, G-mean and specificity are five metrics widely used in various classification problem. However, the accuracy may be deceiving in imbalanced classification problem and are highly sensitive to changes in data [37]. From the perspective of real-life assembly line, the factory managers pay more attention to fault product, which is effectively reflected by specificity score. AUC and G-mean are two comprehensive metrics that consider both the qualified products and fault products. Thus, AUC, G-mean and specificity are more important metrics in our imbalanced assembly quality prediction scenario and widely used in kinds of imbalanced classification problems [31], [38]–[40]. Although some methods have better ACC and recall score, they are worse in specificity score, which denotes these models are poor to detect fault products. In conclusion, the proposed method is a useful tool for high-dimensional and imbalanced data classification in assembly quality prediction.

VI. CONCLUSION

Aiming at the problem of high-dimensional and imbalanced data in assembly process of assembly products, this paper proposes a prediction method of assembly quality based on edge intelligent service. Firstly, a framework for quality prediction of product assembly supported by edge computing is formulated. Secondly, we use Random Forest to select the CTQs and rank them in a descending order, which provides insights for factory managers to analyze quality issues. At last, we propose a SMOTE-Adaboost predictive model with jointly optimized hyperparameters, which solves the problem of assembly quality prediction under the imbalance of category. Experimental results demonstrate that our proposed method is superior to the integration of other sampling algorithms and classification algorithms, which conveys that the proposed method can perform better in assembly quality prediction and fault products detection.

Our study not only enhances the capability of product quality control, but also provides intelligent information services for factory managers. But some issues still exist and need further study. The proposed method in this article does not consider the uncertainty of CTQs selection, so we will focus on the mechanism of each quality characteristic and final assembly quality in the future. Besides, the data in the assembly are incremental, so another research direction is to promote the adaptation of predictive model with incremental data in edge computing scenarios.

REFERENCES

[1] F. Tao and Q. Qi, “New IT driven service-oriented smart manufacturing: Framework and characteristics,” IEEE Trans. Syst., Man, Cybern., Syst., vol. 49, no. 1, pp. 81–91, Jan. 2019.
[2] Q. Qi and F. Tao, “A smart manufacturing service system based on edge computing, fog computing, and cloud computing,” IEEE Access, vol. 7, pp. 86769–86777, 2019.
[3] W. Shi, J. Cao, Q. Zhang, Y. Li, and L. Xu, “Edge computing: Vision and challenges,” IEEE Internet Things J., vol. 3, no. 5, pp. 637–646, Oct. 2016.
[4] Y. Feng, B. Hu, H. Hao, Y. Gao, Z. Li, and J. Tan, “Design of distributed cyber–physical systems for connected and automated vehicles with implementing methodologies,” IEEE Trans. Ind. Inform., vol. 14, no. 9, pp. 4200–4211, Sep. 2018.
[5] P. Pace, G. Aloi, R. Gravina, G. Caliciuri, G. Fortino, and A. Liotta, “An edge-based architecture to support efficient applications for healthcare industry 4.0,” IEEE Trans. Ind. Inform., vol. 15, no. 1, pp. 481–489, Jan. 2019.
[6] Q. Fan and N. Ansari, “On cost aware cloudlet placement for mobile edge computing,” IEEE/CAA J. Automatica Sinica, vol. 6, no. 4, pp. 926–937, Jul. 2019, doi: 10.1109/JAS.2019.1911564.
[7] Z. Sheng, S. Pfersich, A. Eldridge, J. Zhou, D. Tian, and V. C. M. Leung, “Wireless acoustic sensor networks and edge computing for rapid acoustic monitoring,” IEEE/CAA J. Automatica Sinica, vol. 6, no. 1, pp. 64–74, Jan. 2019.
[8] C.-C. Lin, D.-J. Deng, Y.-L. Chih, and H.-T. Chiu, “Smart manufacturing scheduling with edge computing using multiclass deep Q network,” IEEE Trans. Ind. Inform., vol. 15, no. 7, pp. 4276–4284, Jul. 2019.
Y. Feng et al.: Integrated Method for High-Dimensional Imbalanced Assembly Quality Prediction

[9] C.-F. Lai, W.-C. Chien, L. T. Yang, and W. Qiang, “LSTM and edge computing for big data feature recognition of industrial electrical equipment,” IEEE Trans. Ind. Informat., vol. 15, no. 4, pp. 2469–2477, Apr. 2019.

[10] L. Li, K. Ota, and M. Dong, “Deep learning for smart industry: Efficient manufacture inspection system with fog computing,” IEEE Trans. Ind. Informat., vol. 14, no. 10, pp. 6465–6473, Oct. 2018.

[11] X. Li, J. Wan, H.-N. Dai, M. Imran, M. Xia, and A. Celesti, “A hybrid computing solution and resource scheduling strategy for edge computing in smart manufacturing,” IEEE Trans. Ind. Informat., vol. 15, no. 7, pp. 4225–4234, Jul. 2019.

[12] Y. Feng, Z. Hong, Z. Li, H. Zheng, and J. Tan, “Integrated intelligent green scheduling of sustainable flexible workshop with edge computing considering uncertain machine status,” J. Cleaner Prod., vol. 246, Feb. 2020, pp. 119070, doi: 10.1016/j.jclepro.2019.119070.

[13] W. Sun, J. Liu, Y. Yue, and H. Zhang, “Double auction-based resource allocation for mobile edge computing in industrial Internet of Things,” IEEE Trans. Ind. Informat., vol. 14, no. 10, pp. 4692–4701, Oct. 2018.

[14] A. Kusiak, “Data mining: Manufacturing and service applications,” Int. J. Prod. Res., vol. 44, no. 18–19, pp. 4175–4191, Sep. 2006.

[15] Y. Feng, Z. Hong, G. Tian, Z. Li, J. Tan, and H. Hu, “Environmentally friendly MCDM of reliability-based product optimisation combining DEMATEL-based ANP interval uncertainty and Vlse Kriterijumska Optimizacija Kompromisno Resenje (VIKOR),” Inf. Sci., vol. 442–443, pp. 128–144, May 2018.

[16] Y. Feng, Y. Gao, G. Tian, Z. Li, H. Hu, and H. Zheng, “Flexible process planning and End-of-Life decision-making for product recovery optimization based on hybrid disassembly,” IEEE Trans. Auton. Sci. Eng., vol. 16, no. 4, pp. 311–326, Oct. 2019.

[17] Y. Feng, M. Zhou, G. Tian, Z. Li, Z. Zhang, Q. Zhang, and J. Tan, “Target disassembly sequencing and scheme evaluation for CNC machine tools using improved multiobjective ant colony algorithm and fuzzy integral,” IEEE Trans. Syst., Man, Cybern., Syst., vol. 49, no. 12, pp. 2438–2451, Dec. 2019.

[18] G. Diao, L. Zhao, and Y. Yao, “A dynamic quality control approach by improving dominant factors based on improved principal component analysis,” Int. J. Prod. Res., vol. 53, no. 14, pp. 4287–4303, Jul. 2015.

[19] Z. Wei, Y. Feng, Z. Hong, R. Qu, and J. Tan, “Product quality improvement method in manufacturing process based on kernel optimisation algorithm,” Int. J. Prod. Res., vol. 55, no. 19, pp. 5597–5608, Oct. 2017.

[20] L. Yao and Z. Ge, “Nonlinear Gaussian mixture regression for multimode quality prediction with partially labeled data,” Expert Syst. Appl., vol. 59, pp. 226–234, Oct. 2016, doi: 10.1016/j.eswa.2016.04.027.

[21] T. S. Sethi and M. Kantardzic, “Handling adversarial concept drift in streaming data,” Expert Syst. Appl., vol. 97, pp. 18–40, May 2018, doi: 10.1016/j.eswa.2017.12.022.

[22] D. A. Ross, J. Lim, R.-S. Lin, and M.-H. Yang, “Incremental learning for robust visual tracking,” Int. J. Comput. Vis., vol. 77, nos. 1–3, pp. 125–141, May 2008, doi: 10.1007/s11263-007-0075-7.

[23] R. Blagus and L. Lusa, “Joint use of over- and undersampling techniques and cross-validation for the development and assessment of prediction models,” BMC Bioinf., vol. 16, no. 1, pp. 1–10, Dec. 2015.

[24] M. S. Santos, J. P. Soares, P. H. Abreu, H. Araujo, and J. Santos, “Cross-validation for imbalanced datasets: Avoiding overoptimistic and overfitting approaches [Research Frontier],” IEEE Comput. Intell. Mag., vol. 13, no. 4, pp. 59–76, Nov. 2018, doi: 10.1109/MCI.2018.2866730.

[25] W. Feng, W. Huang, and W. Bao, “Imbalanced hyperspectral image classification with an adaptive ensemble method based on SMOTE and rotation forest with differentiated sampling rates,” IEEE Geosci. Remote Sens. Lett., vol. 16, no. 12, pp. 1879–1883, Dec. 2019.

[26] B. M. Haddad, S. Yang, L. J. Karam, J. Ye, N. S. Patel, and M. W. Braun, “Multifeature, sparse-based approach for defects detection and classification in semiconductor units,” IEEE Trans. Autom. Sci. Eng., vol. 15, no. 1, pp. 145–159, Jan. 2018.

[27] W. Hu, J. Gao, Y. Wang, O. Wu, and S. Maybank, “Online AdaBoost-based parameterized methods for dynamic distributed network intrusion detection,” IEEE Trans. Cybern., vol. 44, no. 1, pp. 66–82, Jan. 2014.

[28] G. C. Gutierrez-Tobal, D. Alvarez, F. del Campo, and R. Hornero, “Utility of AdaBoost to detect sleep apnea-hypopnea syndrome from single-channel airflow,” IEEE Trans. Biomed. Eng., vol. 63, no. 3, pp. 636–646, Mar. 2016.

[29] J. Bergstra and Y. Bengio, “Random search for hyper-parameter optimization,” J. Mach. Learn. Res., vol. 13, pp. 281–305, Feb. 2012.

[30] S. Wang and X. Yao, “Diversity analysis on imbalanced data sets using ensemble models,” in Proc. IEEE Symp. Comput. Intell. Data Mining, Mar. 2009, pp. 324–331.

[31] H. He and E. A. Garcia, “Learning from imbalanced data,” IEEE Trans. Knowl. Data Eng., vol. 21, no. 9, pp. 1263–1284, Sep. 2009.

[32] S. Barua, M. M. Islam, X. Yao, and K. Murase, “MMWOTE-majority weighted minority oversampling technique for imbalanced data set learning,” IEEE Trans. Knowl. Data Eng., vol. 26, no. 2, pp. 405–425, Feb. 2014.

[33] S. Chiang Tan, J. Watada, Z. Ibrahim, and M. Khalid, “Evolutionary fuzzy ARTMAP neural networks for classification of semiconductor defects,” IEEE Trans. Neural Netw. Learn. Syst., vol. 26, no. 5, pp. 933–950, May 2015.

[34] H.-J. Kim, N.-O. Jo, and K.-S. Shin, “Optimization of cluster-based evolutionary undersampling for the artificial neural networks in corporate bankruptcy prediction,” Expert Syst. Appl., vol. 59, pp. 226–234, Oct. 2016, doi: 10.1016/j.eswa.2016.04.027.

[35] A. Province Outstanding Youth Fund and the National Outstanding Youth Fund.

[36] above research areas. He has been the Principal Investigator of four projects published over 200 journal articles and conference proceedings papers in the above research areas. He has been the Principal Investigator of four projects supported by the Natural National Science Foundation of China and one project supported by the National High-Tech Research and Development Program of China (863 Program). He is also the Winner of the Zhejiang Province Outstanding Youth Fund and the National Outstanding Youth Fund.

YIXIONG FENG (Member, IEEE) received the B.S. and M.S. degrees in mechanical engineering from Yanshan University, Qinhuangdao, China, in 1997 and 2000, respectively, and the Ph.D. degrees in mechanical engineering from Zhejiang University, Hangzhou, China, in 2004. He is currently a Professor with the Department of Mechanical Engineering, Zhejiang University. He is also a member with the State Key Laboratory of Fluid Power and Mechatronic Systems, Zhejiang University. His research interests include mechanical product design theory, intelligent automation, and advance manufacture technology. He has published over 200 journal articles and conference proceedings papers in the above research areas. He has been the Principal Investigator of four projects supported by the National Natural Science Foundation of China and one project supported by the National High-Tech Research and Development Program of China (863 Program). He is also the Winner of the Zhejiang Province Outstanding Youth Fund and the National Outstanding Youth Fund.
TIANYUE WANG received the B.S. degree in mechanical engineering from Shandong University, China, in 2018. He is currently pursuing the M.Phil. degree with the State Key Laboratory of Fluid Power and Mechatronic Systems, Zhejiang University. His research interests include industrial big data, quality prediction, and data driven information service.

BINGTAO HU received the B.S. degree in mechanical engineering from Zhejiang University, Hangzhou, China, in 2014, where he is currently pursuing the Ph.D. degree. He is also a member with the State Key Laboratory of Fluid Power and Mechatronic Systems, Zhejiang University. His research interests include intelligent transportation, mechanical product design theory, and advance manufacture technology.

CHEN YANG is currently pursuing the Ph.D. degree in electronics and information with Zhejiang University. He engaged in industrial Internet research with the China Wanxiang Group. His research interests include the industrial Internet, intelligent manufacturing, and data driven information service.

JIANRONG TAN received the M.S. degree in engineering and the Ph.D. degree in science.

He is currently a specially appointed Professor from Zhejiang University, a Ph.D. Supervisor, an Academician with the Chinese Academy of Science, the Dean of Mechanical Engineering, the Associate Supervisor of CAD&CG State Key Laboratory, the Head of the Institute of Engineering and Computer Graphics, Zhejiang University, and the Chief Supervisor of Engineering Graphics State Fundamental Courses. He holds a concurrent post of the Deputy Board Chairman in China Engineering Graphic Society, the Board Chairman in Chinese Mechanical Engineering Society, the Associate Supervisor of Mechanical Discipline of the National High School Tutoring Research Seminar, and a Supervisor of Engineering Graphics Tutoring Board in the Minister of Education. He mainly engaged in mechanical designing and theory, and the research in digital designing and manufacture. He gathered his 15 years’ working experience in manufacture and his theory in science, and proposed the technique in combination of batch and customization which is used for multitudinous customization, the status in engineering transition, the fuzzy status, the technology simulation of random status modeling and digitalized prototype integration, the combination of figures and geometry in multicomponent correlation in intricate equipment, the analysis of multilevel disposing and multiparameter matching. He has got the National Pride fourth, which includes the Second Prize in National Technology Progress twice, the First Prize in National Excellent Education Achievement. He has also got the First Prize in provincial-level technology progress for six times. He put his technology into the software. He set and got 12 copyrights of computer software, which achieved a lot in manufacturing enterprises. He has published eight pieces of monograph or compile. He has 142 pieces of article which has been searched by SCI/EI, which has been cited for 1600 times, among his typical 48 pieces of article has been cited for 746 times.

* * *