applied sciences

Article

Designing a Hybrid Equipment-Failure Diagnosis Mechanism under Mixed-Type Data with Limited Failure Samples

Cheng-Hui Chen [1,*], Chen-Kun Tsung [2,*] and Shyr-Shen Yu [1]

[1] Department of Computer Science and Engineering, National Chung Hsing University, Taichung 407224, Taiwan
[2] Department of Computer Science and Information Engineering, National Chin-Yi University of Technology, Taichung 411030, Taiwan
[*] Correspondence: star90154@gmail.com (C.-H.C.); ckt@ncut.edu.tw (C.-K.T.)

Abstract: The rarity of equipment failures results in a high level of imbalance between failure data and normal operation data, which makes the effective classification and prediction of such data difficult. Furthermore, many failure data are dominated by mixed data, which makes the model unable to adapt to this type of failure problem. Second, the replacement cycle of production equipment increases the difficulty of collecting failure data. In this paper, an equipment failure diagnosis method is proposed to solve the problem of poor prediction accuracy due to limited data. In this method, the synthetic minority oversampling technique is combined with a conditional tabular generative adversarial network. The proposed method can be used to predict limited data with a mixture of numerical and categorical data. Experimental results indicate that the proposed method can improve 6.45% compared to other similar methods when equipment failure data account for less than 1% of the total data.

Keywords: Mixed-Type Data; fault diagnosis; SmoteNC ctGAN; limited failure

check for updates

Citation: Chen, C.-H.; Tsung, C.-K.; Yu, S.-S. Designing a Hybrid Equipment-Failure Diagnosis Mechanism under Mixed-Type Data with Limited Failure Samples. *Appl. Sci.* **2022**, *12*, 9286. https://doi.org/10.3390/app12189286

Academic Editor: Amelia Zafra

Received: 16 August 2022
Accepted: 13 September 2022
Published: 16 September 2022

1. Introduction

Coronavirus disease 2019 has severely affected manufacturing and service industries worldwide, which has prompted corporations to focus on ensuring the stable delivery of orders. Consequently, equipment stability has become a key problem. In general, data on equipment failure are sparse. Given the high level of imbalance between failure data and regular-operation data, failure data cannot be effectively classified and predicted. Moreover, machine equipment is limited by its replacement cycles, which further increases the difficulty in collecting failure data [1,2]. Such data, which are time-limited and rare, are referred to as "limited data" [3]. Due to the properties of limited data sets, learning models often categorize the big data according to the normal operation conditions and do not diagnose fault type, which is crucial in the manufacturing industry.

Applications of the categorization and prediction of limited data include cancer diagnosis, scam trading identification, and equipment fault diagnosis. The failure comes from mechanical issues or abnormal data [4–6], and we discuss the data from mechanical issues in this study. Among these three applications, equipment fault diagnosis is the most difficult. Since equipment must be replaced periodically, with the new machine often having different operating procedures from the previous machine, the fault diagnosis model of the previous machine becomes outdated. Therefore, a new diagnosis model must be developed, which requires new failure data to be accumulated. Limited data has become a hot topic recently, and such data are analyzed using two methods. The first method is few-shot learning, which is a learning method aimed at overcoming the difficulties involved in classifying minority class data. In few-shot learning, small data are used to identify minority classes, and a feature extractor is used to perform small-sample tasks, thereby effectively extracting valuable information from small samples. In [7–11], few-shot learning was adopted to detect the manufacturing data of the minority class. By ensuring that no new data are generated, this method effectively prevents overfitting. The second method used to analyze limited data involves generating additional minority data or reducing the quantity of majority data to balance the data, increase the focus of the classifier on minority data, and enhance the accuracy of minority prediction. The most representative algorithm of this method is the synthetic minority oversampling technique (SMOTE) [12], in which simulated minority data are used to achieve data balance. Liu et al. employed a generative adversarial network (GAN) to simulate equipment failure data and adopted a long-short term memory network for fault prediction [13,14]. In [15–17], a hybrid approach that involves combining a GAN with the SMOTE was adopted for processing limited data. This approach solved the overfitting problem of the SMOTE and fulfilled the requirement of a GAN-based training model for considerable data. However, advancements in sensor technology and the Internet of Things (IoT) have increased the complexity of the status of environmental data obtained in machine operation processes, which has resulted in the emergence of data consisting of hybrid features (i.e., categorical and numerical features). The aforementioned methods are inapplicable to such data [18].

Machines have complicated structures and are vulnerable to various types of faults. The nonlinear relationship between performance parameters and faults increases the difficulty in overcoming the imbalanced nature of a data set containing limited data with hybrid features. In [12], a SMOTE-based technique (i.e., Synthetic Minority Over-Sampling Technique for Nominal and Continuous (SmoteNC)) was proposed to balance and process data with continuous and categorical features. In [19], an approach named Conditional Tabular Generative Adversarial Network (ctGAN) was used to establish an adversarial network with hybrid features. However, the SMOTE is prone to overfitting, and a ctGAN requires considerable data for training (Table 1).

Table 1. Analysis of previous literature.

	A Small Amount of Minority Class to Synthesize New Fault Data	Mixed-Type Data	Synthetic Data Representation	Solution
SmoteNC [12]	YES	YES	NO	The ctGAN was used to overcome the drawback of SmoteNC, namely the lack of sample representativeness.
GAN [14]	NO	NO	YES	CtGAN can overcome the inability to apply to Mixed-Type Data.
ctGAN [19]	NO	YES	YES	The oversampling method can increase the minority class data, which can provide enough data for ctGAN training model.
SmoteNC–ctGAN	YES	YES	YES	

In response to these problems, an equipment-fault diagnosis method that involves combining SmoteNC and ctGAN is proposed in this paper. This method comprises three stages. First, SmoteNC is used to simulate hybrid features to balance the data. Second, the simulated and real data are inputted into a ctGAN to generate new fault characteristic data. Third, real data are used to verify the reliability of the data produced by the ctGAN. In this paper, a novel fault diagnosis system, namely SmoteNC–ctGAN, is proposed for handling hybrid limited data. The proposed model can simultaneously handle imbalanced data with continuous and categorical features and fulfill the demand of a ctGAN for considerable training data; thus, the proposed model provides a solution for equipment fault diagnosis by using limited failure data.

The rest of this paper is organized as follows: We explain the proposed algorithm SmoteNC–ctGAN in Section 2; we compare this to other similar methods in Section 3; the case study and numerical results are reported in Section 4; finally, the conclusions and future works are presented in Section 5.

2. Materials and Methods

In this paper, a novel fault diagnosis system, namely SmoteNC–ctGAN, is proposed for handling hybrid limited data. SmoteNC and ctGAN have been proven to be effective models in the literature and are used in various fields. However, we found two shortcomings when these models were applied to equipment failure prediction. First, SmoteNC lacks a verification mechanism in the simulation process, which leads to insufficient authenticity of the simulated fault data. Second, ctGAN requires a lot of training data when simulated fault data. Therefore, in order to solve the above problems, SmoteNC can be used to generate a large amount of simulated fault data from a small amount of real fault data. Then, the simulated and real fault data are added to the ctGAN training process to solve the scarcity of fault data. The proposed method comprises three steps, namely data collection and preprocessing, limited data generation, and model learning and application. The framework of the proposed system is depicted in Figure 1.

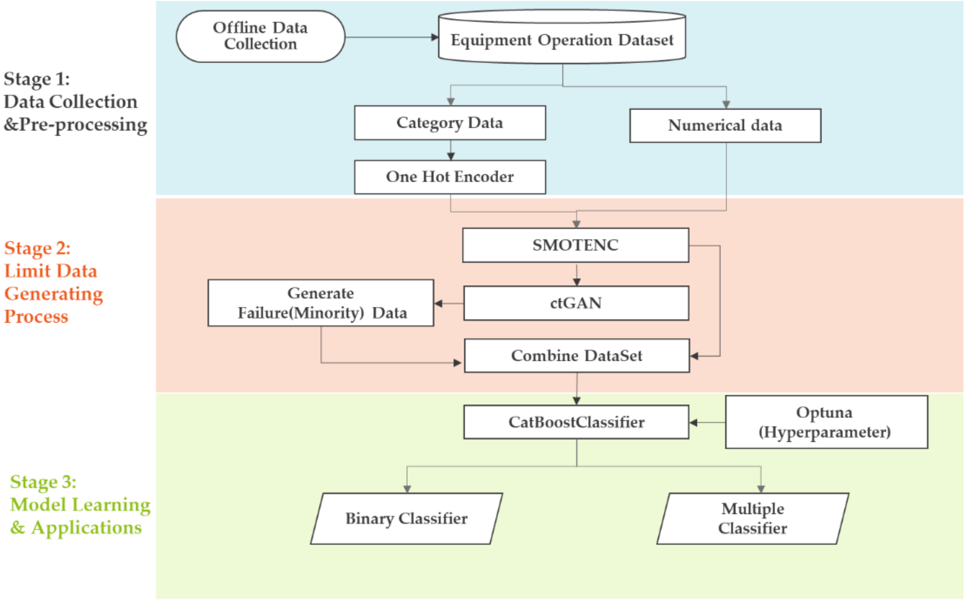


Figure 1. Framework of the proposed SmoteNC—conditional tabular generative adversarial network (ctGAN) system.

2.1. *Data Collection and Pre-Processing*

The machine operation processes involved in the production of different products are complicated and require different operating settings. IoT sensors receive different information when different products are produced. The differences in the sensor data received for each product is not considered an abnormality.

During data collection and preprocessing, data complexity (e.g., English text, punctuation, and Chinese text) might result in errors in a subsequent analysis. Therefore, product names are transformed through one-hot encoding to facilitate the next step.

2.2. *Limit-Data Generating Process*

The limited data generation process comprises three stages. First, SmoteNC is used to simulate hybrid features to balance the data. Second, the simulated failure data and real failure data are input into a ctGAN to generate additional simulated failure data. Finally, the real failure data are used to verify the reliability of the generated simulated failure data. The limited data generation process is illustrated in Figure 2.

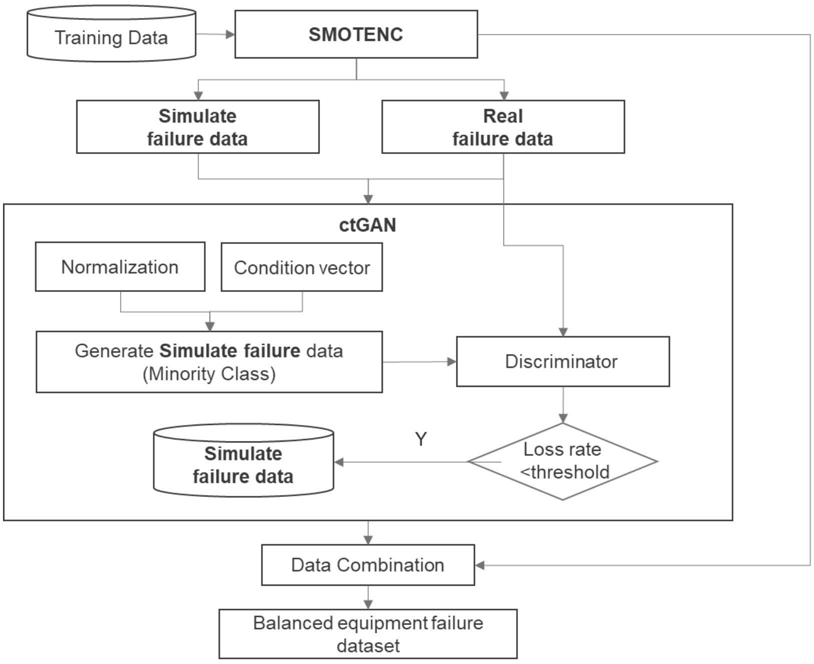


Figure 2. Limit Data Generating Process.

2.2.1. Synthetic Minority Oversampling Technique-Nominal Continuous (SmoteNC)

SmoteNC, which is based on the k-nearest neighbor algorithm, generates new samples of the minority class by using the k-nearest neighbors (Algorithm 1). It randomly generates eigenvectors using the minority features of the k-nearest neighbor. However, because the nearest neighbor of a categorical feature cannot be computed based on distance, this study replaced "distance" in the computation with "frequency." Of the k-neighbors computed using numerical features, the minority categorical sample with the highest frequency was replicated as a new sample [20]. By increasing the quantity of minority data, the data was balanced. In addition, the numerical feature generation indicators are presented in Equation (1):

$$S_{new} = S_i + rand(0,1) * \left(S_i - S_j\right) \tag{1}$$

where S_{new} is the synthetic new data, S_i is the minority class data, S_j is one of the K nearest neighbors of S_i, and T is the training data set. S_i, $S_j \in T$, $rand(0, 1)$ is randomly generated random numbers from 0 to 1.

Algorithm 1: SmoteNC(Pseudocode)

Input: Training data set *T*
Which contains failure dataset (minority class) *S*
Output: Synthesized failure dataset *Snew*
User defined parameter for k-nearest neighbors (Default k = 5)

```
for i in len(S)
 $KNN(s_i, k, T)$ // $x \in X$
   do
      if (numerical feature):
         Generate new failure feature datausing Equation (1)
      else:
         Find the highest frequency data in the k nearest
      j++
   while (j < k)
   Snew←New fail data
Return Snew
```

Studies have used SmoteNC on data sets with only nominal features [12,21]. In [14], a random forest was used in advance to classify data with nominal features and eliminate data in the error category, thereby increasing data representativeness. These methods for handling imbalanced data were input into a ctGAN as training data to fulfill the demand of neural network training, namely, the requirement for considerable training data. Subsequently, the ctGAN was used to overcome the drawback of SmoteNC, namely, the lack of sample representativeness.

2.2.2. Conditional Tabular Generative Adversarial Network (ctGAN)

A ctGAN is a modified conditional GAN applicable to solving problems involving hybrid characteristics, which cannot be solved using a GAN (Algorithm 2). The operation of a ctGAN involves three stages. First, normalization is conducted to process complicated data combinations. Second, condition vectors and sample training methods are used to learn the original data distribution. Finally, a discriminator is used to determine the loss rate threshold and verify whether the simulated minority data generated in the second stage are close to the real minority sample. ctGAN loss function and network structure adopt the [22,23] frameworks, respectively, and the generation network structure is G(m, cond) [19,24].

$$\begin{cases} c_0 = cond \oplus m \\ c_1 = c_0 \oplus ReLU(BN(FC_{|cond|+|m|\to 256}(c_0))) \\ c_2 = c_1 \oplus ReLU(BN(FC_{|cond|+|m|+256\to 256}(c_1))) \\ \hat{\alpha}_i = tanh(FC_{|cond|+|m|+512\to 1}(c_2)) \\ \hat{\beta}_i = gumbel_{0.2}(FC_{|cond|+|m|+512\to n_i}(c_2)) \\ \hat{d}_i = gumbel_{02}(FC_{|cond|+|z|+512\to |D_i|}(h_2)) \end{cases} \tag{2}$$

Algorithm 2: ctGAN(Pseudocode)

```
Input: the training set of the fault data F, which includes
The original training set S and the synthetic Snew by SmoteNC
Output: Synthesized failure dataset Gnew by ctGAN
//user setting Generate number
for i in len(F)
    normalization and condition vector
    R←Generate new failure feature data using Equation (2) //input F
    D←Discriminator (R, S) //Calculate the loss rate of real fault
    Data S and synthetic data. (loss function [9], network structure [10])
    Gnew←D
return Gnew
```

The proposed method has two advantages. First, the minority class features exhibited by the generated minority data samples depend on the condition vector and sample training method. These data are different from those obtained solely by simulating the distance between minority class data. Second, in the proposed method, a discriminator is used to verify whether the simulated minority class samples are representative of the real minority class samples.

2.2.3. Data Combination

In this study, the simulated and actual fault data are integrated, with the aim of reaching a balance between fault and normal operation data. There are three sources of integration. The first is the real operation data, which contains the fault and normal operation data. Second, the fault data generated by SmoteNC simulation can greatly increase the data for a very small number of fault samples, but the authenticity may not be enough. Finally, the equipment failure data simulated by ctGAN are closer to the real data after being screened by the discriminator. However, the disadvantage is that this requires a variety of training data. Therefore, we merged the real data and the two simulated data, so that the fault and normal operation data are balanced (Algorithm 3).

Algorithm 3: SmoteNC–ctGAN (Pseudocode)

```
Input: Training set $T$,
which contains failure dataset (minority class) $S$
Output: Synthesized failure dataset $T_{new}$
1.  $S_{new} \leftarrow SmoteNC(T)$
2.  $G_{new} \leftarrow ctGAN(S, S_{new})$
3.  $T_{new} \leftarrow T \cup S_{new} \cup G_{new}$
4.  $Return\ T_{new}$
```

Finally, we present the synthetic training dataset (T_{new}):

$$T_{new} = T \cup S_{new} \cup G_{new} \tag{3}$$

2.3. Model Learning and Applications

In the model learning and application stage, the effects of simulated and real failure data in equipment failure diagnosis are verified. Two classification tasks are performed in this stage. The first task involves classifying the failure diagnosis results as failure or nonfailure data. The second task involves multicategory classification, where the fault diagnosis results are further classified according to the type of failure, such as tool wear failure (TWF), heat dissipation failure (HDF), power failure (PWF), overstrain failure (OSF), and random failure (RNF).

Since obtaining equipment operation data is difficult, insufficient data are collected for deep learning in most studies [4,25]. Therefore, in this study, the CatBoost classifier [26], which is a popular classifier, was employed to classification tasks. The following section details this classifier and the Optuna hyperparameter learning method [27].

2.3.1. CatBoost Classifier and Optuna

CatBoost, which is an ensemble-learning algorithm based on a gradient boosting decision tree, employs an ordered boosting algorithm and a greedy algorithm to solve problems related to iterative gradient descent [15] and to reduce the risk of overfitting. To increase model accuracy and enhance model performance, the Optuna hyperparameter learning method [16] was adopted in the present study. The training parameters in this method are detailed in Section 3.

2.3.2. Model Evaluation

In equipment fault diagnosis, the rate of true positives is considerably higher than that of true negatives. Furthermore, a high false positive rate results in the constant triggering of a failure alarm, which decreases user confidence in a fault diagnosis model. Therefore, in this study, recall rate, accuracy [28], and balanced accuracy [29] were selected to evaluate the proposed model. In addition, balanced accuracy is an overall indicator for a small number of fault data. A confusion matrix [29–31] for fault diagnosis evaluation metrics is presented in Table 2. The equations for calculating these indicators are presented in Equations (4)–(6) below:

$$\text{Recall rate} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{4}$$

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{5}$$

$$\text{Balanced accuracy} = \left(\frac{\text{TP}}{(\text{TP} + \text{FN})} + \frac{\text{TN}}{(\text{TN} + \text{FP})}\right)/2 \tag{6}$$

Table 2. Confusion matrix for fault diagnosis dataset.

		Actual Condition	
		Failure	Normal
Prediction Condition	Failure	TP (True Positive)	FP (False Positive)
	Normal	FN (False Negative)	TN (True Negative)

3. Results

The results section details the fault data set; the design of relevant model parameters; and the experiment results, including those for recall, accuracy, and balanced accuracy.

3.1. Dataset Description

In this study, the UCI AI4I 2020 Predictive Maintenance Dataset [32] was used to verify the performance of the proposed model [33]. This data set contains 10,000 data points, of which, 3.4% of points represent fault data. Each data point has 12 features, specifically, six equipment operation features and six equipment fault features. These features are detailed in the following text.

The six equipment operation features of the data points are as follows:

1. Product ID: Product ID, which represents categorical data, is a key feature used to distinguish the type of product processed and consists of a letter Low (50%), medium (30%), High (20%) as product quality variants.
2. Air temperature: Air temperature, which represents numerical data, refers to the temperature of the environment (between 2 K and 300 K after normalization).
3. Process temperature (K): Process temperature, which represents numerical data, refers to the temperature of the production process.
4. Rotational speed (rpm): Rotational speed, which represents numerical data, refers to the rotational speed of the main shaft.
5. Torque (Nm): Torque represents a type of numerical data and is generally equal to 40 Nm where ε = 10 and no negative values.
6. Tool wear (min): Tool wear, which represents numerical data, refers to the tool operation time.

The six equipment fault features of the data points are as follows:

7. Tool wear failure (TWF): Tool wear failure causes a process failure.
8. Heat dissipation failure (HDF): Heat dissipation causes a process failure.
9. Power failure (PWF): Power failure causes a process failure.
10. Overstrain failure (OSF): OSF refers to the failure caused by overstrain in the production process.
11. Random failures (RNF): RNFs are failures whose cause cannot be determined. Their occurrence probability in the production process is 0.1%.
12. Machine failure: The original two-category label (0 represents normal, and 1 represents failure) was changed into a multicategory label (0 represents normal, 1 represents TWF, 2 represents HDF, 3 represents PWF, 4 represents OSF, and 5 represents RNF) to verify the multicategory prediction accuracy of the proposed model.

3.2. Experiment Setting

The experimental method adopts a three-fold cross-validation, and explains the average number of training and testing data sets after each round of segmentation. Please refer to the following, Table 3.

Table 3. The number of training and testing data sets after each round of segmentation.

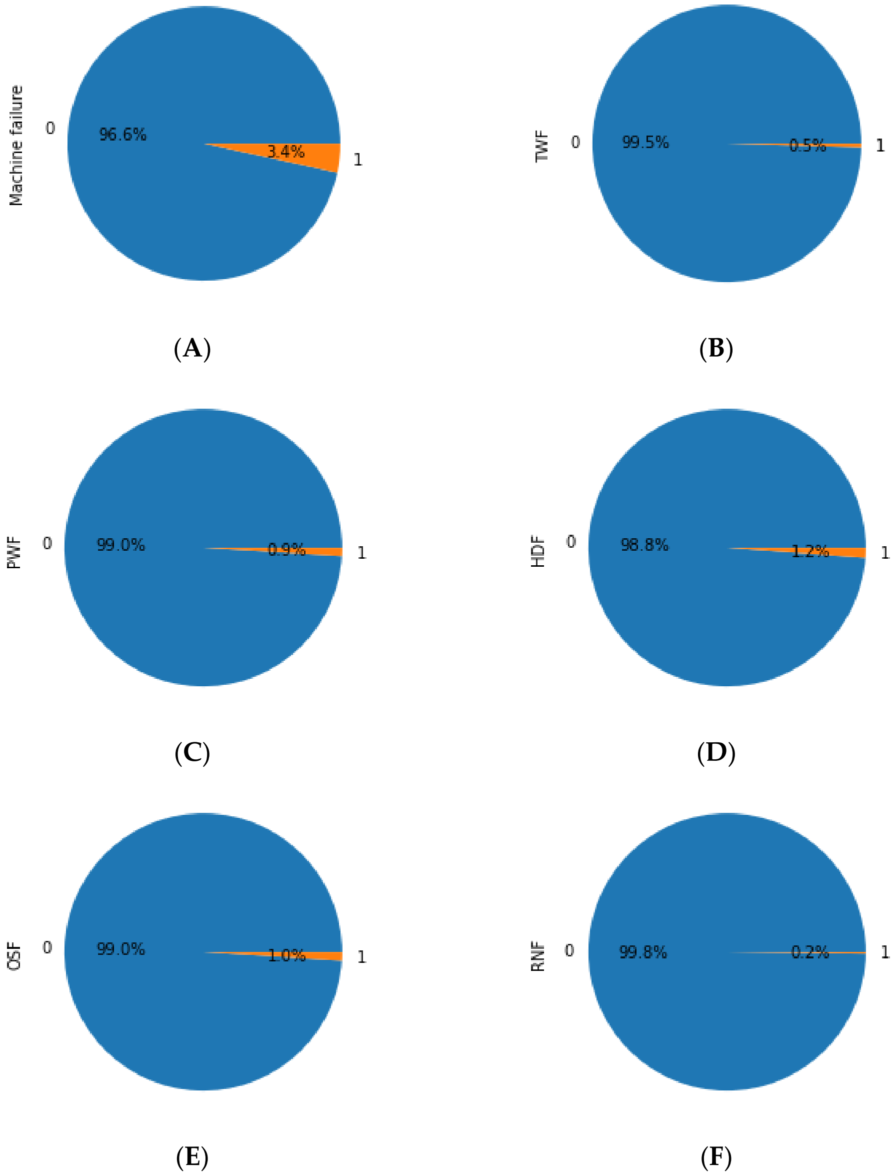
Round (Cross Validation)	**Total**	**Training Set**	**Test Set**
Each Round	10,000 (100%)	6700	3300

3.3. Parameter Setting

All parameters that are applied to ctGAN and CatBoost are listed in Tables 2 and 3, respectively.

3.4. Experiment Results

Equipment fault data exhibit two types of features: categorical features (related to the product information) and numerical features. Considerable imbalance was observed in the collected data, with the fault data accounting for a small proportion of the collected data (Figure 3, in which 1 and 0 represent fault and normal data, respectively).



(**A**) (**B**)

(**C**) (**D**)

(**E**) (**F**)

Figure 3. (**A**) Machine Failure Minority class 3.4%. (**B**) Tool wear failure (TWF) Minority class 0.5%. (**C**) Heat dissipation failure (HDF) Minority class 0.9%. (**D**) Power failure (PWF) Minority class 1.2%. (**E**) Overstrain failure (OSF) Minority class 1.0%. (**F**) Random failures (RNF) Minority class 0.2%.

The severe imbalance in the collected data set increased the difficulty of obtaining accurate model predictions. The prediction results obtained for the TWF, HDF, PWF, OSF, and RNF diagnoses with the confusion matrix and other methods for processing limited data were compared.

According to the results presented in Tables 4–13, the experimental results of five failures are presented. The proposed method exhibited a higher recall rate and balanced accuracy than did the other methods for the diagnoses of all types of failure. Perfect recall was achieved with the proposed method in PWF diagnosis and OSF diagnosis, and balanced accuracy achieved the best result at 98.77% and 98.05%, respectively. Although some false positive results were obtained using the proposed method in the aforementioned fault diagnoses, its performance was acceptable. For the TWF diagnosis, a balanced accuracy of 91.02%, which was 17.08% higher than the second-performing diagnosis, was observed, and the accuracy was only behind by 2.45%. For the HDF diagnosis, a balanced accuracy of 97.61% was observed, and other recall and accuracy data were higher than the second-performing ctGAN + CatBoost. For the RNF diagnosis, a balanced accuracy, recall rate, and accuracy of 63.84, 85.71%, and 42.06%, respectively, were achieved using the proposed method. This kind of random failure has no exact failure type, which makes it difficult to grasp. In addition, SmoteNC + CatBoost exhibited excellent performance in the HDF and OSF diagnoses, but performed unsatisfactorily in the PWF and RNF diagnoses. ctGAN + CatBoost exhibited favorable performance in the HDF and OSF diagnoses, but performed unsatisfactorily in the PWF and RNF diagnoses.

Table 4. The number of samples generated by SMOTE-NC and ctGAN.

Failure Mode	Total Traning Set	Traning Set (Original Training Data)	SMOTE-NC	ctGAN
TWF	19,658	6700 (Contains 27 Failure)	6673 (Failure)	6700 (Failure)
HDF	19,658	6700 (Contains 80 Failure)	6620 (Failure)	6700 (Failure)
OSF	19,658	6700 (Contains 62 Failure)	6638 (Failure)	6700 (Failure)
PWF	19,658	6700 (Contains 60 Failure)	6640 (Failure)	6700 (Failure)
RNF	19,658	6700 (Contains 12 Failure)	6688 (Failure)	6700 (Failure)
Machine Failure	19,658	6700 (Contains 221 Failure)	6479 (Failure)	6700 (Failure)

Table 5. Parameter settings for the ctGAN.

Parameter	Value
echo	10
Size of the output samples	Generator: (256,256) Discriminator: (256,256)
Optimizer	Adam
Learning Rate	0.0002
Loss Function	lower-bound (ELBO) loss
Activation	ReLU
Number of generated failure data	6700 (Same as the number of failures in the training set)

Table 6. Parameter settings for CatBoost (results of parameter optimization based on the Optuna method).

Parameter	**Value**
Iterations	50
Depth	6
Learning rate	0.18176
Early stopping rounds	10
Bagging temperature	0.8278
Iterations	50
Depth	6

Table 7. Prediction results for the TWF diagnosis obtained using the confusion matrix.

		Actual	
		Failure	Normal
Prediction	Failure	17	244
	Normal	2	3037

Table 8. Prediction results obtained for the TWF diagnosis using other methods used for processing limited data.

Method	**Recall Rate**	**Accuracy**	**Balanced Accuracy**
CatBoost (non-oversampling)	0.0000	**0.9942**	0.5000
SmoteNC + CatBoost	0.3684	0.9718	0.6719
ctGAN + CatBoost	0.5263	0.9500	0.7394
SmoteNC + ctGAN + CatBoost (The proposed method)	**0.8947**	0.9255	**0.9102**

Table 9. Prediction results obtained for the HDF diagnosis using the confusion matrix.

		Actual	
		Failure	Normal
Prediction	Failure	34	63
	Normal	1	3202

Table 10. Prediction results obtained for the HDF diagnosis using other methods used for processing limited data.

Method	**Recall Rate**	**Accuracy**	**Balanced Accuracy**
CatBoost (non-oversampling)	0.5143	**0.9948**	0.7571
SmoteNC + CatBoost	0.9429	0.9888	0.9661
ctGAN + CatBoost	**0.9714**	0.9785	0.9750
SmoteNC + ctGAN + CatBoost (The proposed method)	**0.9714**	0.9806	**0.9761**

Table 11. Prediction results obtained for PWF diagnosis with the confusion matrix.

		Actual	
		Failure	Normal
Prediction	Failure	35	80
	Normal	0	3185

Table 12. Prediction results obtained for the PWF diagnosis using other methods used for processing limited data.

Method	Recall Rate	Accuracy	Balanced Accuracy
CatBoost (non-oversampling)	0.4857	**0.9942**	0.7427
SmoteNC + CatBoost	1.0000	0.9579	0.9787
ctGAN + CatBoost	1.0000	0.9715	0.9856
SmoteNC + ctGAN + CatBoost (The proposed method)	**1.0000**	0.9758	**0.9877**

Table 13. Prediction results obtained for the OSF diagnosis using the confusion matrix.

		Actual	
		Failure	Normal
Prediction	Failure	36	127
	Normal	0	3137

To demonstrate the effectiveness of the proposed method for a multicategory fault diagnosis, the five types of failures were mixed and labeled (the normal condition, TWF, HDF, PWF, OSF, and RNF were labeled 0–5, respectively). The results obtained in the multicategory fault diagnosis are presented in Tables 14–17.

Table 14. Prediction results obtained for the OSF diagnosis using other methods used for processing limited data.

Method	Recall Rate	Accuracy	Balanced Accuracy
CatBoost (non-oversampling)	0.5833	**0.9952**	0.7915
SmoteNC + CatBoost	0.9722	0.9870	0.9797
ctGAN + CatBoost	0.9722	0.9742	0.9732
SmoteNC + ctGAN + CatBoost (The proposed method)	**1.0000**	0.9615	**0.9805**

Table 15. Prediction results obtained for the RNF diagnosis using the confusion matrix.

		Actual	
		Failure	Normal
Prediction	Failure	6	1911
	Normal	1	1382

Table 16. Prediction results obtained for the RNF diagnosis using other methods used for processing limited data.

Method	Recall Rate	Accuracy	Balanced Accuracy
CatBoost (non-oversampling)	0.0000	**0.9979**	0.5000
SmoteNC + CatBoost	0.2857	0.8615	0.5742
ctGAN + CatBoost	0.0000	0.9882	0.4951
SmoteNC + ctGAN + CatBoost (The proposed method)	**0.8571**	0.4206	**0.6384**

Table 17. Results obtained using different methods in multicategory fault diagnosis.

	Recall Rate	Accuracy	Balanced Accuracy
CatBoost (non-oversampling)	0.2868	**0.9687**	0.6423
SmoteNC + CatBoost	0.7881	0.9670	0.8809
ctGAN + CatBoost	0.8305	0.9082	0.8708
SmoteNC + ctGAN + CatBoost (The proposed method)	**0.9068**	0.8712	**0.8883**

The results presented in Tables 14–17 indicate that CatBoost classified most data into the normal category and, thus, exhibited a recall rate of 28.68%. SmoteNC + CatBoost exhibited high accuracy in the multicategory prediction but had a relatively low recall rate. ctGAN + CatBoost exhibited a recall rate of 83.05%. Finally, SmoteNC + ctGAN + CatBoost exhibited a recall rate of 90.68% and a balanced accuracy of 88.83%.

4. Discussion

Given that the failure data accounted for only 3.4% of the total data and were divided into five failure modes (i.e., 0.5%, 0.9%, 1.2%, 1.0%, and 0.2% of the data were classified into the TWF, HDF, PWF, OSF, and RNF minority classes, respectively), the data were highly imbalanced, which increased the difficulty in failure prediction. The experiment's results indicated that the proposed method was superior to the other methods in the five-category failure classification. Although the proposed method exhibited a high false positive rate, from the equipment diagnosis perspective, the cost of sudden machine downtime considerably exceeds that of system misdiagnosis.

4.1. The Focus of Prediction Is to Detect Equipment Failure, Not Normal Operation

From the perspective of equipment failure prediction, the cost of false negatives considerably exceeds that of false positives. Therefore, this study aimed to increase the possibility of false positives and, thus, reduce the possibility of false negatives. The experiment results revealed that prior to oversampling, the recall rate of the proposed method was approximately 0. This result suggests that the equipment exhibits normal operation; thus, the possibility of equipment failure can be ignored.

The experiment results for RNF prediction revealed that most methods did not exhibit favorable prediction results. Despite achieving a recall rate of 85.71%, the proposed method generated an excessive number of false alarms, possibly due to the insufficient scope of data collection. To overcome this problem, an IoT sensor can be attached to the equipment, or the collected data can be expanded. In the future, the authors of this study will analyze the reason for the aforementioned problem and collect operation data according to the identified reason. Finally, the multicategory prediction results revealed that some categories of equipment failure correlated. Due to the overlap of the categories, they could not be effectively separated. This problem is commonly encountered in equipment failure prediction. To solve this problem, a new failure category can be established to relabel correlating failure categories, thereby increasing the failure prediction accuracy of the proposed method.

4.2. Necessity of Processing Data with Hybrid Features in Limited Data Sets

The categorical data features of a product include product type. This information is crucial to the prediction of equipment failure. Different product types require drastically different allocations of production resources and sensor values. Therefore, processing data with hybrid features in limited data sets is essential.

4.3. Interpretability of the Equipment Failure Prediction Results

In this study, a tree-based model was selected for failure prediction. This model has a certain degree of interpretability and can be used for problem analysis, correctly predicting failures, and analyzing the causes of failures. The results of equipment failure prediction can be interpreted using tree algorithms to determine the reason for failure and to implement preventive measures accordingly [33]. Moreover, these results can be interpreted using GAN algorithms [34] to analyze the reason for each minority data class. The aforementioned analysis overcomes the overfitting problem of GAN algorithms.

5. Conclusions

In this paper, a method is proposed for predicting limited failure data with high accuracy to overcome the limitations associated with the processing of limited data with hybrid features. The experimental results indicate that the proposed method can improve 6.45% compared to similar methods when equipment failure data account for less than 1% of the total data. The proposed model can simultaneously handle imbalanced data with continuous and categorical features and can fulfill the demand of a ctGAN for considerable training data; thus, the proposed model provides a solution for equipment fault diagnosis by using limited failure data. Moreover, given the interpretability of the tree-based structure used in the proposed method, its results can be easily interpreted, and the reason for failure can be easily determined, thereby improving the maintenance efficiency.

Given that equipment deterioration is a time-series problem, failure data might be sequential in nature. Future studies can employ time-series production and prediction models in the failure data generation process to increase the accuracy of the generated data.

Author Contributions: Conceptualization, C.-H.C.; methodology, C.-H.C.; software, C.-H.C.; validation, C.-H.C.; formal analysis, C.-H.C. and C.-K.T.; investigation, C.-H.C.; resources, C.-H.C. and C.-K.T.; data curation, C.-H.C.; writing—original draft preparation, C.-H.C.; writing—review and editing, C.-K.T., S.-S.Y. and C.-H.C.; visualization, C.-H.C.; supervision, S.-S.Y.; project administration, C.-H.C.; funding acquisition, C.-H.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Ministry of Science and Technology of the Republic of China grant number MOST 109-2221-E-167-030-MY3.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Chang, Y.I.; Shei-Dei Yang, S.; Chuang, Y.C.; Shen, J.H.; Lin, C.C.; Li, C.E. Automatic Classification of Uroflow Patterns via the Grading-based Approach. *J. Inf. Sci. Eng.* **2022**, *38*, 463–477.
2. Shen, J.H.; Chen, M.Y.; Lu, C.T.; Wang, R.H. Monitoring spatial keyword queries based on resident domains of mobile objects in IoT environments. *Mob. Netw. Appl.* **2020**, *27*, 208–218. [CrossRef]
3. Hu, Y.; Liu, R.; Li, X.; Chen, D.; Hu, Q. Task-Sequencing Meta Learning for Intelligent Few-Shot Fault Diagnosis with Limited Data. *IEEE Trans. Industr. Inform.* **2022**, *18*, 3894–3904. [CrossRef]
4. Jeong, S.; Shen, J.H.; Ahn, B. A study on smart healthcare monitoring using IoT based on blockchain. *Wirel. Commun. Mob. Comput.* **2021**, *2021*, 1–9. [CrossRef]
5. Liu, J.C.; Yang, C.T.; Chan, Y.W.; Kristiani, E.; Jiang, W.J. Cyberattack detection model using deep learning in a network log system with data visualization. *J. Supercomput.* **2021**, *77*, 10984–11003. [CrossRef]
6. Yang, C.T.; Liu, J.C.; Kristiani, E.; Liu, M.L.; You, I.; Pau, G. Netflow monitoring and cyberattack detection using deep learning with ceph. *IEEE Access* **2020**, *8*, 7842–7850. [CrossRef]
7. Wang, H.; Li, Z.; Wang, H. Few-shot steel surface defect detection. *IEEE Trans. Instrum. Meas.* **2022**, *71*, 1–12. [CrossRef]
8. Zhang, A.; Li, S.; Cui, Y.; Yang, W.; Dong, R.; Hu, J. Limited data rolling bearing fault diagnosis with few-shot learning. *IEEE Access Pract. Innov. Open Solut.* **2019**, *7*, 110895–110904. [CrossRef]
9. Zhang, J.; Wang, Y.; Zhu, K.; Zhang, Y.; Li, Y. Diagnosis of interturn short-circuit faults in permanent magnet synchronous motors based on few-shot learning under a federated learning framework. *IEEE Trans. Ind. Inform.* **2021**, *17*, 8495–8504. [CrossRef]
10. Zhang, T.; Chen, J.; He, S.; Zhou, Z. Prior knowledge-augmented self-supervised feature learning for few-shot intelligent fault diagnosis of machines. *IEEE Trans. Ind. Electron.* **2022**, *69*, 10573–10584. [CrossRef]
11. Zhou, X.; Liang, W.; Shimizu, S.; Ma, J.; Jin, Q. Siamese Neural Network Based Few-Shot Learning for Anomaly Detection in Industrial Cyber-Physical Systems. *IEEE Trans. Ind. Inform.* **2021**, *17*, 5790–5798. [CrossRef]
12. Chawla, N.V.; Bowyer, K.W.; Hall, L.O.; Kegelmeyer, W.P. SMOTE: Synthetic Minority over-Sampling Technique. *J. Artif. Intell. Res.* **2002**, *16*, 321–357. [CrossRef]
13. Liu, H.; Zhao, H.; Wang, J.; Yuan, S.; Feng, W. LSTM-GAN-AE: A promising approach for fault diagnosis in machine health monitoring. *IEEE Trans. Instrum. Meas.* **2022**, *71*, 1–13. [CrossRef]

14. Moon, J.; Jung, S.; Park, S.; Hwang, E. Conditional tabular GAN-based two-stage data generation scheme for short-term load forecasting. *IEEE Access Pract. Innov. Open Solut.* **2020**, *8*, 205327–205339. [CrossRef]
15. Dablain, D.; Krawczyk, B.; Chawla, N.V. DeepSMOTE: Fusing Deep Learning and SMOTE for Imbalanced Data. *IEEE Trans. Neural Netw. Learn. Syst.* **2022**, 1–15. [CrossRef]
16. Sharma, A.; Singh, P.K.; Chandra, R. SMOTified-GAN for Class Imbalanced Pattern Classification Problems. *IEEE Access* **2022**, *10*, 30655–30665. [CrossRef]
17. Kim, J.; Park, H. OA-GAN: Overfitting Avoidance Method of GAN Oversampling Based on XAI. In Proceedings of the 2021 Twelfth International Conference on Ubiquitous and Future Networks (ICUFN), Jeju Island, Korea, 17–20 August 2021; pp. 394–398.
18. Mukherjee, M.; Khushi, M. SMOTE-ENC: A Novel SMOTE-Based Method to Generate Synthetic Data for Nominal and Continuous Features. *Appl. Syst. Innov.* **2021**, *4*, 18. [CrossRef]
19. Xu, L.; Skoularidou, M.; Cuesta-Infante, A.; Veeramachaneni, K. Modeling Tabular Data Using Conditional GAN. *arXiv* **2019**, arXiv:1907.00503.
20. Pradipta, G.A.; Wardoyo, R.; Musdholifah, A.; Sanjaya, I.N.H.; Ismail, M. SMOTE for Handling Imbalanced Data Problem: A Review. In Proceedings of the 2021 Sixth International Conference on Informatics and Computing (ICIC), Jakarta, Indonesia, 3–4 November 2021; pp. 1–8.
21. Xu, Z.; Shen, D.; Nie, T.; Kou, Y. A Hybrid Sampling Algorithm Combining M-SMOTE and ENN Based on Random Forest for Medical Imbalanced Data. *J. Biomed. Inform.* **2020**, *107*, 103465. [CrossRef]
22. Martin, A.; Chintala, S.; Bottou, L. Wasserstein GAN. *arXiv* **2017**. [CrossRef]
23. Zinan, L.; Khetan, A.; Fanti, G.; Oh, S. PacGAN: The Power of Two Samples in Generative Adversarial Networks. *arXiv* **2017**. [CrossRef]
24. Chunsheng, A.; Sun, J.; Wang, Y.; Wei, Q. A K-Means Improved CTGAN Oversampling Method for Data Imbalance Problem. In Proceedings of the 2021 IEEE 21st International Conference on Software Quality, Reliability and Security (QRS), Hainan, China, 6–10 December 2021; pp. 883–887.
25. Nascita, A.; Montieri, A.; Aceto, G.; Ciuonzo, D.; Persico, V.; Pescape, A. XAI meets mobile traffic classification: Understanding and improving multimodal deep learning architectures. *IEEE Trans. Netw. Serv. Manag.* **2021**, *18*, 4225–4246. [CrossRef]
26. Dorogush, A.V.; Ershov, V.; Gulin, A. CatBoost: Gradient Boosting with Categorical Features Support. *arXiv* **2018**, arXiv:1810.11363.
27. Akiba, T.; Sano, S.; Yanase, T.; Ohta, T.; Koyama, M. Optuna: A next-Generation Hyperparameter Optimization Framework. In Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery Data Mining, Anchorage, AK, USA, 4–8 August 2019; ACM: New York, NY, USA, 2019.
28. Matthews, B.W. Comparison of the Predicted and Observed Secondary Structure of T4 Phage Lysozyme. *Biochim. Biophys. Acta-(BBA)-Protein Struct.* **1975**, *405*, 442–451. [CrossRef]
29. Henning, B.K.; Ong, C.S.; Stephan, K.E.; Buhmann, M.J. The Balanced Accuracy and Its Posterior Distribution. In Proceedings of the 2010 20th International Conference on Pattern Recognition, Istanbul, Turkey, 23–26 August 2010; pp. 3121–3124.
30. Ioannis, M.; Rallis, I.; Georgoulas, I.; Kopsiaftis, G.; Doulamis, A.; Doulamis, N. Multiclass Confusion Matrix Reduction Method and Its Application on Net Promoter Score Classification Problem. *Technologies* **2021**, *9*, 81. [CrossRef]
31. Stephen, S.V. Selecting and Interpreting Measures of Thematic Classification Accuracy. *Remote Sens. Environ.* **1997**, *62*, 77–89. [CrossRef]
32. Dua, D.; Graff, C. *UCI Machine Learning Repository*; University of California Irvine: Irvine, CA, USA, 2017.
33. Matzka, S. Explainable Artificial Intelligence for Predictive Maintenance Applications. In Proceedings of the 2020 Third International Conference on Artificial Intelligence for Industries (AI4I), Irvine, CA, USA, 21–23 September 2020; pp. 69–74.
34. Mendel, J.M.; Bonissone, P.P. Critical thinking about explainable AI (XAI) for rule-based fuzzy systems. *IEEE Trans. Fuzzy Syst.* **2021**, *29*, 3579–3593. [CrossRef]