mathematics

MDPI

Article

# Deep Smooth Random Sampling and Association Attention for Air Quality Anomaly Detection

Peng Wang [1], Minhang Li [2], Xiaoying Zhi [2], Xiliang Liu [2,*], Zhixiang He [2], Ziyue Di [2], Xiang Zhu [2], Yanchen Zhu [2], Wenqiong Cui [2], Wenyu Deng [2] and Wenhan Fan [2]

[1] Key Laboratory of Data Science and Smart Education Ministry of Education, Hainan Normal University, Haikou 570203, China; 050115@hainnu.edu.cn

[2] Faculty of Information Technology, Beijing University of Technology, Beijing 100124, China; minhangli@emails.bjut.edu.cn (M.L.); s202375050@emails.bjut.edu.cn (X.Z.); zxhe@bjut.edu.cn (Z.H.); dzy00@emails.bjut.edu.cn (Z.D.); zhuxiang@emails.bjut.edu.cn (X.Z.); zhuyanchen@emails.bjut.edu.cn (Y.Z.); cuiwq@emails.bjut.edu.cn (W.C.); wenyuuu@emails.bjut.edu.cn (W.D.); yefanwenhanzi@emails.bjut.edu.cn (W.F.)

* Correspondence: liuxl@bjut.edu.cn

**Abstract:** Real-time monitoring and timely warning of air quality are vital components of building livable cities and implementing the "Healthy China" strategy. Real-time, efficient, and accurate detection of air quality anomalies holds great significance. However, almost all existing methods for air quality anomaly detection often overlook the imbalanced distribution of data. In addition, many traditional methods cannot learn both pointwise representation and pairwise association, so they cannot solve complex features. This study proposes an anomaly detection method for air quality monitoring based on Deep Smooth Random Sampling and Association Attention in Transformer (DSRS-AAT). Firstly, based on the third geographical law, the more similar the geographical environment, the closer the geographical target features are. We cluster sites according to the surrounding geographic features to fully explore latent feature associations. Then, we employ Deep Smooth Random Sampling to rebalance the air quality datasets. Meanwhile, the Transformer with association attention considers both prior associations and series associations to distinguish anomaly patterns. Experiments are carried out with real data from 95 monitoring stations in Haikou City, China. Final results demonstrate that the proposed DSRS-AAT improves the effectiveness of anomaly detection and provides interpretability analysis for traceability, owing to a significant improvement with the baselines (OmniAnomaly, THOC, etc.). The proposed method effectively enhances the effectiveness of air quality anomaly detection and provides a reference value for real-time monitoring and early warning of urban air quality.

**Keywords:** air quality anomaly detection; imbalanced data processing; geographical third law; transformer; livable cities

**MSC:** 37M10; 68T07

check for updates

**Citation:** Wang, P.; Li, M.; Zhi, X.; Liu, X.; He, Z.; Di, Z.; Zhu, X.; Zhu, Y.; Cui, W.; Deng, W.; et al. Deep Smooth Random Sampling and Association Attention for Air Quality Anomaly Detection. *Mathematics* **2024**, *12*, 2048. https://doi.org/10.3390/math12132048

Academic Editor: Shih-Wei Lin

Received: 30 May 2024
Revised: 27 June 2024
Accepted: 28 June 2024
Published: 30 June 2024

## 1. Introduction

Air pollution, as a key factor affecting people's livelihood and health, has always been assigned great importance by the Chinese government and people. Therefore, in the outline of the 14th Five-Year Plan, urban air quality indicators were formally incorporated into the binding indicators of economic and social development, which indicates that air quality has become an important part of the process of national development. In order to continue to promote the blue-sky defense war, people's health rights must be ensured, and high-quality economic development must be promoted through continuous improvement of air quality. The State Council issued the Action Plan for Continuous Improvement of Air Quality at the end of 2023. The plan clearly puts forward the air quality improvement goals during the "14th Five-Year Plan" period and sets a timetable and roadmap for the continuous improvement of air quality. The implementation of this plan will not only improve the quality of the environment and enhance the quality of people's lives but also promote the optimization and upgrading of the economic structure, promote green and low-carbon development, and finally lay a solid foundation for achieving the Sustainable Development Goals.

In order to solve the problem of air pollution, air anomaly detection has become a key progress. It not only offers early alerts but swiftly responds to abrupt pollution incidents or unusual weather patterns. This significantly lessens the negative impact on public health and the environment. It also pinpoints anomalies, enabling further tracking of pollution sources. This provides crucial data for decision making by the relevant authorities. Moreover, anomaly detection helps oversee and maintain the data quality of the monitoring system. This ensures the precision and reliability of the monitoring data. It can also refine resource allocation, boosting efficiency and cost-effectiveness.

For example, traditional methods for detecting anomalies in air quality primarily rely on statistical models, data analysis, and simple machine learning models. They lack flexibility in their determination criteria and struggle to adequately address sudden human-induced situations [1]. These methods often overlook the spatiotemporal characteristics of real-world time series data, presenting certain limitations. Moreover, due to the unique nature of the datasets involved in anomaly detection tasks, the imbalanced feature distribution causes models to favor the majority class from training to fitting, leading to significant biases in evaluation metrics such as accuracy, making it difficult to achieve the desired performance, and significantly increasing the difficulty of improving results [2]. In recent years, with the upgrading of air monitoring stations, large-scale multimodal data have become available for research related to air quality. Models based on deep learning show significant potential in the field of air quality anomaly detection [3]. As data scale increases, deep learning outperforms traditional machine learning methods in performance and has more powerful data processing capabilities, more accurate detection capabilities, and better adaptability [4]. Among them, the attention mechanism brought by the Transformer model has attracted widespread attention due to its capability to handle sequential data, providing insights into time series anomaly detection.

In summary, this study focuses on anomaly detection in time series data, exploring solutions to the challenges faced by anomaly detection tasks on imbalanced datasets. An in-depth study is conducted to improve the effectiveness of air quality anomaly detection models. A data processing method named DSRS (Deep Smooth Random Sampling) is proposed to mitigate the impact of special data distributions and applied to the Transformer based on the association attention mechanism to accomplish the task of air quality anomaly detection. Here are the contributions of this study:

- Based on the principle of the geographical third law, combining POI data with the geographical location information of monitoring stations, exploring the impact of functional zoning on spatiotemporal data, and analyzing the potential environmental feature associations between stations to improve the effectiveness of anomaly detection;
- In response to the extreme imbalance of air quality datasets in reality, the fusion method of Deep Smooth Random Sampling (DSRS) is adopted for rebalancing processing to alleviate its impact on the training of anomaly detection models and improve accuracy;
- Introducing prior associations and sequence associations into the Transformer model to amplify the differences in performance features between anomaly and normal points in time series data to improve anomaly detection effectiveness;
- Matching and spatiotemporal analysis of anomaly detection results with complaint data contribute to addressing practical needs.

This study is structured as follows: First, in Section 2, some research status of unbalanced data processing and anomaly detection of time series data are briefly introduced. Section 3 explains our approach in detail. Section 4 then presents the results of our experiments, while Section 5 discusses the work we have carried out. Finally, Section 6 describes the conclusions of the research.

## 2. Related Works

### *2.1. Anomaly Detection*

Anomaly detection is a significant issue in computer science. With the development of big data and sensor technology, researchers utilize collected environmental monitoring data to establish various models for predicting and identifying anomalies [5]. Current research shows that traditional statistical methods, machine learning, and deep learning techniques are widely used in anomaly detection. Although these methods have made certain progress, they still face challenges such as detection accuracy and model generalization capability. This section will explore the current research status in the computer science field regarding anomaly detection and discuss existing technologies and methods.

#### 2.1.1. Traditional Methods

Nearest neighbor-based anomaly detection methods are divided into two categories: distance-based and density-based detection methods [6]. Distance-based methods detect anomalies by calculating the distance between abnormal data and normal data. Density-based techniques [7] compare the estimated density of each data point with its neighboring data points, considering data with lower estimated density as anomalies. One of the most classic algorithms is the K-Nearest Neighbors (KNN) algorithm and its variants, which determine whether a point is an outlier by calculating the distance between it and its K-Nearest Neighbors. The Local Outlier Factor (LOF) method [8] measures the relative isolation of given data points; however, its performance on scattered datasets is poor. Additionally, algorithm performance decreases when the density of abnormal data points is close to that of their neighborhood density and boundary data points. To further enhance the efficiency of this method, researchers have proposed improvements such as Connectivity-Based Outlier Factor (COF), Influence-Based Local Outlier Factor (INFLO), Label Driven Outlier Factor (LDOF), and Local Association-based Intensity (LOCI) [9]. Clustering-based anomaly detection techniques group data based on similarity or similar patterns, considering data not belonging to any cluster as anomalies. Common methods using clustering to detect outliers and noise points include Density-Based Spatial Clustering of Applications with Noise (DBSCAN) [10], which can handle noise-robust datasets and find all dense regions of samples as clusters; Ordering Points To Identify the Clustering Structure (OPTICS) [11], which can discover clusters of arbitrary shapes in large-scale data and has suitable robustness to noise points; and Cluster-Based Local Outlier Factor (CBLOF) [12], which detects outliers based on the LOF method using clustering. The core objective of clustering-based methods is to identify the structure of clusters; thus, these methods may overlook outlier data. The application of the mentioned algorithm in anomaly detection is manifested in many aspects. In the aspects of financial fraud detection, intrusion detection system, and industrial production, the abnormal behavior is identified by comparing. However, the performance of these algorithms in anomaly detection tasks is also limited by the setting of parameters and the quality of datasets.

#### 2.1.2. Shallow Learning Methods

Machine learning methods are widely used in anomaly detection, including many classic and convenient algorithms and techniques. For example, Support Vector Machines (SVM) [13] is a supervised learning algorithm capable of handling nonlinear problems. By mapping data to a high-dimensional feature space, it finds an optimal hyperplane to best separate normal points and outliers in the data. Decision Tree (DT) technology [14] is a simple and intuitive algorithm used for classification tasks. In anomaly detection, it splits data by selecting a specific feature or attribute, recursively dividing the dataset into smaller subsets, and guiding the data down different branches based on the different values of this feature. Random Forest (RF) is another ensemble learning method [15], which constructs multiple decision trees and combines them to improve the accuracy and robustness of anomaly detection models. Similar to traditional methods, shallow learning methods have a wide range of applications, but they are also limited by the quality of the data set and difficult to achieve the best performance

### 2.1.3. Deep Learning Methods

Deep learning models are effective technologies in the field of anomaly detection. Deep learning-based models have been applied in supervised, semi-supervised, and unsupervised modes. In supervised mode, models are trained using normal data and then used for anomaly detection. The key issue in this mode is obtaining precise labels for the inlier and outlier classes in each domain. In the semi-supervised mode used for anomaly detection, class labels are only available for inlier points. Unsupervised mode is widely applicable due to its ability to handle unlabeled datasets. In this mode, networks based on deep learning attempt to reconstruct the input at the output end, measuring reconstruction errors to rank outliers in the dataset. Many deep learning techniques, such as Adaptive Resonance Theory (ART), Generative Adversarial Networks (GAN), and Restricted Boltzmann Machines (RBM), have been applied in anomaly detection. Besides these networks, methods like Recurrent Neural Networks (RNN) are also popular.

Many researchers currently focus on multivariate time series anomaly detection. The Transformer model [16] has received widespread attention and application in time series prediction, anomaly detection, and classification due to its ability to handle sequence data. Ref. [17] proposed a new Multi-Scale Convolutional Recurrent Variational Autoencoder (MSCRVAE) model, which not only considers the temporal and spatial dependencies but also captures latent variable representations. Ref. [18] introduced a Dilated Convolutional Transformer-based GAN (DCT-GAN), which can improve the stability and robustness of anomaly detection in time series. Ref. [19] presented a lightweight, unsupervised multivariate time series anomaly detection algorithm, LUAD, which models multivariate time series to better capture representations between variables and the influence between time series. Ref. [20] proposed a dynamic network anomaly edge detection method that combines RegraphGAN with spatiotemporal encoding. Ref. [21] introduced a spatial–temporal knowledge graph network (STKGN) that uses continuous-time dynamic graphs to simulate the influence of events interacting within real nodes. Ref. [22] developed a real-time adaptive training algorithm, Spatially Adaptive Dynamic Convolutional Autoencoding ODER Anomaly Detection (STEAMCODER), which can effectively learn spatial features when the data volume is small and fully learn temporal features when the data volume is large. Compared with the first two types of methods, deep learning can cope well with large-scale data and improve computing efficiency and performance. With the maturity of technology, the application of such anomaly detection methods has gradually become the mainstream of various applications

In summary, current research on anomaly detection focuses on various application fields, continuously improving the performance of corresponding deep learning methods in data preprocessing, adjusting model parameters, and optimizing loss functions. Although the latest research results have demonstrated the superior performance of unsupervised deep learning in anomaly detection, its opacity and interpretability remain areas requiring continuous improvement in practical applications. Moreover, anomaly detection models also need corresponding improvements and optimizations based on their actual application scenarios, especially in learning information representation and finding distinguishable criteria between normal and abnormal data.

## *2.2. Imbalanced Data Processing*

In the context of anomaly detection in air quality, there is a significant disparity in the number of samples between different classes. The process of rebalancing aims to alter the distribution of imbalanced datasets through some mechanism to obtain a relatively balanced dataset. Existing rebalancing methods mainly include data-level and algorithm-level approaches.

Data-level rebalancing methods involve oversampling, undersampling, and mixed sampling techniques. Undersampling is a simple method of adjusting data distribution balance by removing the majority of class samples from the original dataset. Common undersampling methods include Ensemble Nearest Neighbors (ENN) [23], Self-Paced Ensemble (SPE) [24], Balanced Cascade [25], and Near Miss [26]. However, deleting sample data may lead to the loss of valuable information, potentially causing classifiers to miss important information related to the majority class, leading to overfitting. When the data imbalance problem is severe, this method is not feasible. Oversampling involves synthesizing minority samples according to certain rules to achieve data rebalancing. This includes methods such as Easy Ensemble [27], Synthetic Minority Oversampling Technique (SMOTE) [28], and Adaptive Synthetic Sampling Approach for Imbalanced Learning (ADASYN) [29]. However, this method is susceptible to noise and can prolong the training time of these models, and overfitting issues remain unresolved. Mixed sampling algorithms combine the advantages of undersampling and oversampling methods. Examples include the Synthetic Minority Oversampling Technique with TOMEK, an undersampling algorithm that aims to reduce the number of sample classes by removing those pairs that are closest to each other while retaining as much information as possible, and Quadratic Correction (SMOTETOMEK), Synthetic Minority Oversampling Technique with Ensemble of Nearest Neighbors (SMOTE-ENN) [30], and Integrated Forests for Synthetic Minority Oversampling Technique (IForest-SMOTE) [31], which have shown better performance in recent years. Many of the methods mentioned above are classical methods of data imbalance processing, which have a high degree of recognition and have been widely used, but the processing methods of time series data are still lacking in research.

At the algorithm level, there are techniques such as cost-sensitive learning [32], transfer learning [33], and self-supervised imbalanced learning for data rebalancing [34]. Cost-sensitive learning trains classifiers to set different penalties for different classification errors to reduce the model's preference for majority class samples. However, if the evaluation criteria of cost-sensitive learning overly emphasize a small portion of samples, it can be unfair to the majority of samples, and the algorithm may have weak generalization ability on datasets with different sample proportions. Transfer learning models the majority class samples and minority class samples separately and transfers the learning information of the majority class samples to the minority class samples. Self-supervised learning is an unsupervised learning method that mainly uses surrogate tasks to extract supervised information from large-scale unlabeled data to train models, thereby learning valuable representations for downstream tasks.

In summary, research on rebalancing imbalanced data is currently a hot topic [35]. Most researchers utilize various sampling methods to sample and rebalance data in specific domains, which, to some extent, alleviates the ratio of positive and negative samples in imbalanced data. However, these methods do not consider the distribution characteristics of time series data. Rebalancing for discrete points can instead disrupt the temporal characteristics of the corresponding task datasets, leading to a complete disconnection from reality.

## 3. DSRS-AAT

### *3.1. Overall Architecture*

First, we conduct statistical analysis on the air pollution data and meteorological data from each monitoring site and integrate these data for preliminary data preprocessing, including data cleaning and filling in missing values. Next, based on the theory of the third law of geography, we merge the air pollution and meteorological data of the target observation station and its strongly correlated observation stations to explore the complex spatiotemporal relationships. By spatial clustering of Points of Interest (POI) and spatial variability analysis, we select strongly correlated stations that co-vary with the target station, clustering sites with similar environments and extracting high-quality multidimensional time series data sequences, thus efficiently constructing the air quality dataset. After the initial construction of the dataset, we use the DSRS method to smooth and sample the data to alleviate the impact of imbalanced data, effectively improving the quality of the dataset. Subsequently, we input the processed dataset into a model based on the association mechanism and optimize the model parameters to achieve optimal performance. Finally, through comparative analysis with other baseline models and different ablation experiments, we demonstrate the superiority of our proposed method in performance and showcase the importance of each part of the design. The corresponding technical roadmap is shown in the Figure 1 below.

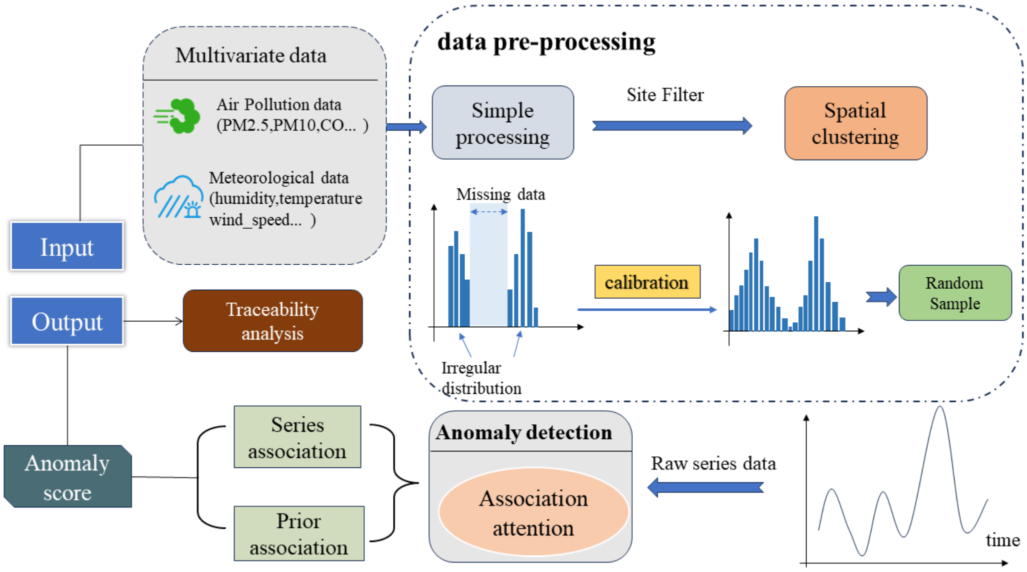

**Figure 1.** The technical flowchart.

### 3.2. Deep Smooth Random Sampling

In order to address the challenge of extreme imbalance in data for anomaly detection tasks, this study proposes a method based on Deep Smooth Random Sampling (DSRS) to mitigate the impact of imbalanced datasets and improve the performance of anomaly detection models. The advantage of this method lies in its integration of traditional sampling techniques while considering the contextual features of multidimensional time series data, thus preserving the original characteristics of the dataset while optimizing its quality to the fullest extent.

Inspired by the statistical properties of similar features in neighboring points with continuous target values [36], this study introduces the concept of deep smoothing, which essentially involves propagating statistical information of features within neighboring intervals. The main purpose of this process is to calibrate potentially biased estimates of feature distributions, especially for target values with few samples. Specifically, for a feature element $z$:

$$\mu_b = \frac{1}{N_b}\sum_{i=1}^{N_b} Z_i \tag{1}$$

$$\Sigma_b = \frac{1}{N_b - 1}\sum_{i=1}^{N_b}(z_i - \mu_b)(z_i - \mu_b)^T \tag{2}$$

where $N_b$ is the total number of samples that have been divided into small segments with intervals of $b$, a symmetric kernel $k(y_b, y_{b'})$ is applied to smooth the feature mean and covariance over the entire statistical heap $\beta$ based on the given feature data. The results are as follows:

$$\widetilde{\mu}_b = \sum_{b' \in \beta} k(y_b, y_{b'})\mu_{b'} \tag{3}$$

$$\check{\Sigma}_b = \sum_{b' \in \beta} k(y_b, y_{b'})\Sigma_{b'} \tag{4}$$

After obtaining both $\{\mu_b, \Sigma_b\}$ and $\{\widetilde{\mu}_b, \widetilde{\Sigma}_b\}$ simultaneously, the representation features of each input sample are calibrated according to the article [37].

$$\widetilde{z} = \widetilde{\Sigma}_b^{\frac{1}{2}} \Sigma_b^{-\frac{1}{2}}(z - \mu_b) + \widetilde{\mu}_b \tag{5}$$

After obtaining the calibrated representation features, a random spatial sampling technique is employed based on an effective fully random tree proposed in [38]. In this process, the data space is randomly partitioned into multiple subspaces until each data element in an independent subspace belongs to the same type. This method achieves time efficiency by randomly partitioning the space and selecting a split value in randomly chosen attributes, thus reducing the cost of computing information gain to find the splitting attribute and position. Below is the pseudocode for this method (Algorithm 1).

**Algorithm 1: Random Space Division Sampling.**

```
Input: Dataset S =$(s_1, s_2, s_3, \ldots s_n), s_i = (x, y), i = 1, 2, 3, \ldots n$, the number of trees
$NTree$ in a CRF
Output: Dataset $D$ after sampling
$NTree$= $log_2 d$ + 1;
if $NTree$ is an even then
    $NTree$ = NTree + 1;
end if
Construct $NTree$ complete random trees;
Find label noise points according to criterion
while the detected label noise is changed do
    Construct two more complete random tree, and $NTree$
    $= NTree$ + 2;
    Mark the label of each node with the majority label in
    the two new trees;
    for each sample P do
        Judge label(P) = label(parent(P));
    end for
end while
Remove all label noise points
Sample the class $C$ with the largest number of samples, denote its sampled result as
$C'$ and the rest as $R$
Return $D = R + C'$
```

### 3.3. *Association Attention in Transformer*

#### 3.3.1. Problem Definition

Considering a multivariate time series of length T:

$$X = (x_1, x_2, \ldots, x_T) \tag{6}$$

Each data point is obtained from a monitoring station at timestamp $t$, with data dimension d. The problem addressed in this study can be viewed as providing a time series sequence $X$ as input. For a test sequence $X_{test}$ of the same size as the training sequence, with length $T'$, the goal is to predict $y^{test} = (y_1, y_2, \ldots, y_{T'})$, where $y_t \in \{0, 1\}$. Here, 1 represents an anomalous data point, and 0 represents a normal data point. The core issue of unsupervised time series anomaly detection is determining whether $x_T$ is anomalous without labels.

As mentioned above, the emphasis in unsupervised time series anomaly detection is on learning informative feature representations and finding discriminative criteria for anomalies and non-anomalies. To some extent, the underlying basis of the model is similar to that of the anomaly Transformer [39], aiming to discover more information associations and address the problem by learning association differences. Association differences essentially amplify the impact of anomalies and non-anomalies in the evaluation, thereby enhancing the model's effectiveness in a more concise and efficient manner.

### 3.3.2. Core Mechanism of Association Attention

Taking into account the limitations and deficiencies of the Transformer model in handling anomaly detection tasks, this study draws inspiration from the more targeted anomaly Transformer model. It has been demonstrated to be effective on various types of time series datasets, optimizing the relevant techniques of association differences to better meet the demands of the task at hand (Figure 2).

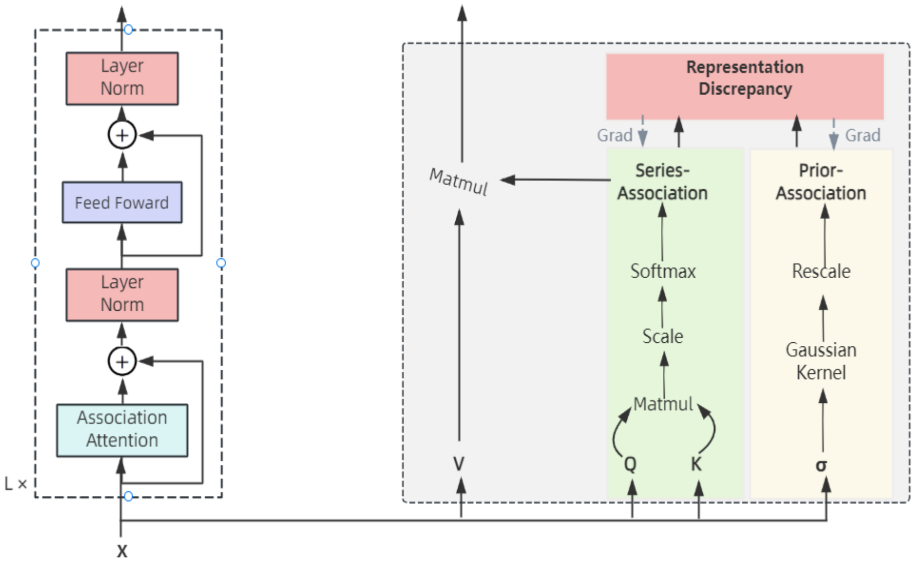

**Figure 2.** Core architecture based on association attention.

The input time series $X \in R^{N\times d}$ consists of elements $\mathcal{X}^l \in \mathbb{R}^{N\times d_{\text{model}}}, l \in \{1, \cdots, L\}$ representing the input to the $l$-th layer with $d_{\text{model}}$ channels, where the initial input $\mathcal{X}^0 = \text{Embedding}(\mathcal{X})$ represents the raw data processed through embedding. The association attention $(\cdot)$ function is used to compute the association attention, which is the core component for distinguishing between anomalous and non-anomalous points.

As mentioned in the Transformer model, the single-branch self-attention technique cannot simultaneously handle prior association and sequence association. To jointly process these two types of associations and effectively identify the differences between anomalous and non-anomalous points, a dual-branch architecture inspired by the anomaly Transformer is adopted, as illustrated in the above figure. This structure enables more effective derivation of association attention and subsequent determination based on it.

Firstly, the parameters are initialized:

$$Q, K, V, \sigma = X^{l-1}W_Q^l, X^{l-1}W_K^l, X^{l-1}W_V^l, X^{l-1}W_\sigma^l \tag{7}$$

Here, $Q, K, V, \sigma$ represent the query, key, and value for the self-attention mechanism.

To better understand the differences in associations and clarify the methods of the model presented in this paper, it is necessary to delve deeper into prior associations and sequential associations.

For the prior association branch, this paper employs a learnable Gaussian kernel to measure the priori of relative temporal distances. This design leverages the unimodal characteristic of the Gaussian kernel, thus structurally focusing more on the immediate time range. Moreover, a learnable scale parameter $\sigma$ is introduced for the Gaussian kernel, allowing the prior associations to adapt to different patterns of time series, such as variations in the length of anomaly segments.

Regarding the series association branch, this paper aims to learn the associations from the original sequences, enabling it to adaptively identify the most effective correlations. Notably, both methods maintain the time dependency of each time point, providing richer information than point-by-point representation. They reflect the prior associations based on adjacency sets and the learned actual associations, respectively, and the difference between these two should be capable of distinguishing between normal and abnormal states.

Below, we introduce the formulas for prior association $\mathcal{P}^l$ and series association $\mathcal{S}^l$.

$$\mathcal{P}^l = \text{Rescale}\left(\left[\frac{1}{\sqrt{2\pi}\sigma_i}\exp\left(-\frac{|j-i|^2}{2\sigma_i^2}\right)\right]_{i,j\in\{1,\cdots,N\}}\right) \tag{8}$$

$$\mathcal{S}^l = \text{Softmax}\left(\frac{\mathcal{Q}\mathcal{K}^{\mathrm{T}}}{\sqrt{d}}\right) \tag{9}$$

In the prior association, $\mathcal{P}^l \in \mathbb{R}^{N\times N}$ is generated based on the learned vector $\sigma \in \mathbb{R}^{N\times 1}$, where the $i$-th element $\sigma_i$ corresponds to the $i$-th time point. Specifically, for the $i$-th time point, its association weight with the $j$-th point is calculated using the Gaussian kernel function $G(|j-i|;\sigma_i) = \frac{1}{\sqrt{2\pi}\sigma_i}\exp(-\frac{|j-i|^2}{2\sigma_i^2})$. Additionally, the function Rescale $(\cdot)$ is used to convert the association weights into a discrete distribution $\mathcal{P}^l$, and $Softmax(\cdot)$ function is used to normalize the attention weights.

This study formalizes the association difference as the symmetric Kullback–Leibler (KL) divergence between the prior association and the sequence association, which represents the information gain between these two distributions. By averaging the association differences over multiple layers, the associations of multiple layers are combined into a more informative measure for a more balanced evaluation of the loss function, as follows:

$$Loss = \left[\frac{1}{L}\sum_{l=1}^{L}\left(\mathrm{KL}\left(\mathcal{P}^l_{i,:} \parallel \mathcal{S}^l_{i,:}\right) + \mathrm{KL}\left(\mathcal{S}^l_{i,:} \parallel \mathcal{P}^l_{i,:}\right)\right)\right]_{i=1,\cdots,N} \tag{10}$$

where $\mathrm{KL}(\cdot\,||\,\cdot)$ computes the degree of association between the discrete distributions $\mathcal{P}^l, \mathcal{S}^l$ in the same event space. The loss function measures the pointwise association difference of $\mathcal{X}^l$ relative to the multi-layer prior association $\mathcal{P}^l$ and sequence association $\mathcal{S}^l$, where the result of the $i$-th element is associated with the $i$-th time point in the time series. From the previous analysis, it can be inferred that the loss of anomalous points will be smaller than that of normal data points, which can serve as a potential differentiator. Unlike most reconstruction frameworks, which use a reconstruction component, this method does not utilize reconstruction. Although reconstruction can help detect anomalous behavior different from expected behavior, reconstructing the entire time series to obtain a reconstruction loss is not straightforward. Therefore, this approach can, to some extent, improve efficiency by reducing the impact of insufficiently considered latent information on representation information.

Incorporating normalized association differences into the anomaly detection criterion allows for balancing between representation information of the time series and association differences.

$$\text{Anomaly Score}(\mathcal{X}) = \text{Softmax}(Loss) \tag{11}$$

This represents an anomaly score for each data point in the time series, where data points corresponding to anomalies typically exhibit lower anomalous associations. Therefore, based on a hyperparameter $\delta$, one can determine whether a point is anomalous or not.

$$\mathcal{Y}_i = \begin{cases} 1 : \text{anomaly} & \text{Anomaly Score}(\mathcal{X}_i) \geq \delta \\ 0 : \text{normal} & \text{Anomaly Score}(\mathcal{X}_i) < \delta. \end{cases} \tag{12}$$

## 4. Experiments

### *4.1. Datasets*

The study area chosen for this research is Haikou City in Hainan Province, China. Hourly air pollution concentration data and corresponding meteorological data were collected from a total of 95 air monitoring stations distributed across four districts: Xiuying District, Longhua District, Qiongshan District, and Meilan District. These sites are shown in Figure 3 below. These data were used to construct the original dataset. The data span from2021-05-26 to 2023-03-11, encompassing features such as PM2.5, PM10, CO, $NO_2$, $SO_2$, $O_3$, air pressure, humidity, temperature, wind direction, and wind speed, totaling 11 variables. Table 1 presents the dataset used in the experiments.

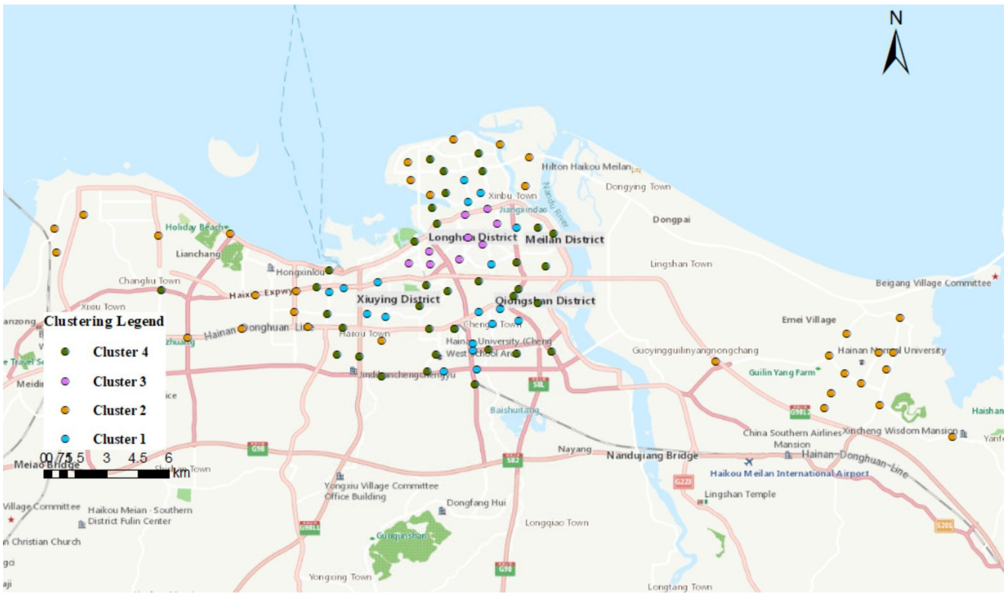

**Figure 3.** Site data clustering results.

**Table 1.** Description of experimental data sets.

| Type | Name | Units of Measurement |
|---|---|---|
| POI data | Longitude | - |
| | Latitude | - |
| Air quality data | PM2.5 | $\mu g/m^3$ |
| | PM10 | $\mu g/m^3$ |
| | CO | $\mu g/m^3$ |
| | $NO_2$ | $\mu g/m^3$ |
| | $SO_2$ | $\mu g/m^3$ |
| | $O_3$ | $\mu g/m^3$ |
| Meteorological data | Air pressure | hPa |
| | Humidity | % |
| | Temperature | °C |
| | Wind direction | - |
| | Wind speed | km/h |

Additionally, in order to facilitate the interpretability of the anomaly detection results later on, this study also incorporates a public complaint corpus dataset. The dataset encompasses six attribute columns: "event_startTime", "event_endTime", "event_title", "event_content", "reply_department", and "event_result", with a total of 3734 entries (Table 2).

**Table 2.** Complaint corpus dataset information.

| Field Name | Comments | Type |
|---|---|---|
| event_startTime | The time when the complainant initiated the complaint | Date |
| event_endTime | Completion time of processing by relevant departments | Date |
| event_title | Complaint event name | Natural language |
| event_content | Complaint event detail | Natural language |
| reply_department | Department for dealing with complaint event | Natural language |
| event_result | Result | Natural language |

### *4.2. Data Preprocessing*

Due to environmental factors, the data collection phase may encounter issues such as equipment malfunctions and transmission errors, leading to outliers and missing values. To maintain the continuity and integrity of the analysis, during the data preprocessing stage, this study employs linear interpolation to fill short-term missing values observed at some monitoring stations, while long-term missing values were addressed through multiple imputations.

Furthermore, considering the third law of geography, which suggests that similar geographic environments result in similar geographic features, analyzing the similarity and spatial association patterns among air quality data from monitoring stations can improve the quality of the air quality anomaly detection dataset and further optimize the effectiveness of air quality anomaly detection. In this study, POI (Points of Interest) data for the research area were obtained through Baidu's open API interface. POI data include various geographic entities such as commercial areas, cultural facilities, and transportation hubs. Subsequently, hierarchical clustering was employed to cluster all stations spatially. The hierarchical clustering method constructs a hierarchical tree structure to combine spatially adjacent and similar stations into clusters. This method does not require predefining the number of clusters, thus enabling the exploration of potential geographical patterns within the research area without prior knowledge. Analyzing the clustering results can help understand the geographic relationships among stations and thereby identify datasets with similar features for experimentation.

Using the hierarchical clustering method, all 95 monitoring stations were clustered, resulting in four different clusters. Combined with POI data, the visualization of the clusters is shown in Figure 3. Research reveals that stations in Cluster 1 and Cluster 3 are located in the city center with dense building facilities, including prominent urban structures such as commercial buildings, cultural institutions, and shopping centers. Stations in Cluster 2 are situated in peripheral areas, surrounded mainly by schools and parks. Stations in Cluster 4 are located in comprehensive areas, surrounded by various activity facilities. After clustering, four clusters of air quality monitoring datasets were obtained at the station level. Considering data quality and the direction of subsequent research, stations in Cluster 1 were selected as the experimental baseline data.

For the anomaly detection criterion, this study employs a statistical method known as the Interquartile Range (IQR), which measures the variability of data by dividing the dataset into quartiles. Specifically, 1.5 times the IQR was used as the threshold for identifying outliers in air quality data. Any data point exceeding 1.5 times the IQR was labeled as an anomaly. Additionally, based on actual complaints about air quality anomalies, manually labeled data points that conformed to the criterion were reviewed and incorporated to account for real-world scenarios. This method demonstrates advantages such as low complexity, fast computation speed, minimal tuning effort, and suitable interpretability in the experimental scenario of this study.

Following the methods described above, the processed original experimental dataset is presented in the Tables 3 and 4 below.

**Table 3.** Original data set information.

| Type | Amount | Anomaly/Normal Ratio |
|---|---|---|
| Features | 11 | — |
| Test data | 68,803 | 0.10 |
| Train data | 126,163 | 0.10 |

**Table 4.** Data set information after resampling.

| Method | Amount | Feature Dimension |
|---|---|---|
| ENN | 123,496 | 11 |
| ABSMOTE | 257,904 | 11 |
| RSDS | 43,800 | 11 |

### *4.3. Evaluation Metrics*

Precision is the most commonly used criterion for evaluating deep learning anomaly detection models. However, since anomaly detection tasks are aimed at learning from imbalanced data, using precision alone cannot effectively reflect the overall performance and results of the model in identifying anomalies in the entire dataset. To meet the practical needs of anomaly detection, precision (*p*-value), recall (R-value), and the harmonic mean of precision and recall (F1-value) are further adopted as evaluation metrics. Among these three metrics, anomaly detection focuses on the F1-value, where a higher value indicates better performance of the model. Table 5 presents the confusion matrix for air quality anomaly detection, where True Positive (TP) indicates the number of anomalous data points predicted as anomalous air quality, False Positive (FP) indicates the number of normal data points predicted as anomalous air quality, False Negative (FN) indicates the number of anomalous data points predicted as normal air quality, and True Negative (TN) indicates the number of normal data points predicted as normal air quality.

**Table 5.** Confusion matrix.

| | Actual Outlier | Actual Normal Point |
|---|---|---|
| Predicted outlier | True Positive (TP) | False Positive (FP) |
| Predicted normal | False Negative (FN) | True Negative (TN) |

Combining the definitions from the confusion matrix, the evaluation metrics used in this study, including *precision* (P), *recall* (R), and *F*1*-score* (F1), are represented by the following formulas:

$$Precision = \frac{TP}{TP + FP} \tag{13}$$

$$Recall = \frac{TP}{TP + FN} \tag{14}$$

$$F1 - score = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{15}$$

Furthermore, to better assess the model's performance, the ROC-AUC curve is included. It plots the True Positive Rate (*TPR*) against the False Positive Rate (*FPR*) as coordinates to depict the classifier's performance at different thresholds. The formulas for $TPR$ and $FPR$ are as follows:

$$TPR = \frac{TP}{TP + FN} \tag{16}$$

$$FPR = \frac{FP}{FP + TN} \tag{17}$$

AUC (Area Under the ROC Curve) is the area under the ROC curve, used to measure the performance of a classifier. The closer the AUC value is to 1, the better the classifier's performance; conversely, the closer the AUC value is to 0, the worse the classifier's performance.

In summary, the ROC curve and AUC value are two important indicators used to evaluate the performance of binary classification models. Through the ROC curve, we can intuitively understand the performance of the classifier at different thresholds, while through the AUC value, we can quantitatively evaluate the overall performance of the classifier.

### *4.4. Results*

#### 4.4.1. Baselines

In this study, we evaluate several representative baseline models, including OmniAnomaly [40], Deep-SVDD [41], and THOC [42], based on the DSRS-AAT model. Additionally, we compare two different sampling methods: undersampling, represented by ENN, and oversampling, represented by ABSMOTE [43]. All of the baseline models above take default parameters and can be implemented in the sklearn library or in the GitHub link given in the citation study. Through these comparative experiments, our aim was to demonstrate the superiority of the proposed method in anomaly detection tasks.

According to the results in Table 6, it can be observed that our proposed method achieved the best performance under the widely used F1-score metric. Without balancing processing, the F1-score of this model was, on average, 22% higher than that of the baseline models. After applying the imbalance processing method proposed in this study, the performance of each baseline model improved by approximately 22% on average. Overall, our method improves the performance by approximately 20% in the task of air quality anomaly detection.

**Table 6.** Main results.

| **Method** | | **None** | | | **ENN** | | | **ABSMOTE** | | | **RSDS** | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Metrics** | **P** | **R** | **F1** | **P** | **R** | **F1** | **P** | **R** | **F1** | **P** | **R** | **F1** |
| AAT | 81.12 | 40.91 | 54.39 | 80.96 | 42.09 | 55.38 | 79.39 | 38.78 | 51.57 | 85.64 | 53.35 | 65.69 |
| Deep-SVDD | 69.46 | 31.08 | 46.32 | 65.34 | 33.61 | 47.16 | 67.06 | 33.45 | 44.36 | 72.26 | 43.89 | 53.56 |
| Isolation Forest | 72.29 | 34.16 | 48.28 | 65.7 | 27.08 | 38.35 | 40.85 | 20.98 | 27.66 | 50.85 | 75.7 | 60.95 |
| BeatGAN | 62.18 | 33.04 | 46.07 | 63.86 | 31.14 | 45.89 | 60.46 | 30.29 | 42.16 | 75.61 | 37.16 | 51.06 |
| LOF | 67.41 | 30.6 | 43.86 | 66.37 | 29.06 | 42.15 | 66.78 | 31 | 42.56 | 68.15 | 30.15 | 44.16 |
| LSTM-VAE | 68.24 | 30.95 | 45.89 | 67.78 | 29.16 | 43.49 | 63.71 | 30.46 | 40.03 | 82.16 | 39.45 | 53.17 |
| OCSVM | 14.8 | 77.51 | 24.85 | 14.89 | 77.95 | 25 | 12.37 | 66.16 | 20.85 | 16.41 | 89.64 | 27.74 |
| OmniAnomaly | 78.01 | 34.01 | 49.78 | 72.56 | 31.05 | 49.02 | 72.59 | 37.23 | 50.54 | 79.16 | 48.26 | 61.26 |
| CL-MPPCA | 65.61 | 30.17 | 46.16 | 74.89 | 34.64 | 48.76 | 65.02 | 35.26 | 45.87 | 77.82 | 40.23 | 54.06 |
| THOC | 79.81 | 36.46 | 50.67 | 79.01 | 37.59 | 50.43 | 72.01 | 38.16 | 51.26 | 78.16 | 46.16 | 60.16 |

Furthermore, data analysis shows that traditional random processing methods such as ENN and ABSMOTE did not perform well on time series data tasks. These methods did not significantly improve the overall performance of the task and may even disrupt sequence features, leading to performance degradation in some cases. This further highlights the advantage of DSRS in addressing the imbalance problem in time series data. In the subsequent ablation experiment section, we will further support this conclusion with experimental results.

In the Figure 4, the height of the bars represents the F1-score values obtained from the experimental results for the respective methods after their integration with the model. A higher bar indicates a better performance, as demonstrated by the model.

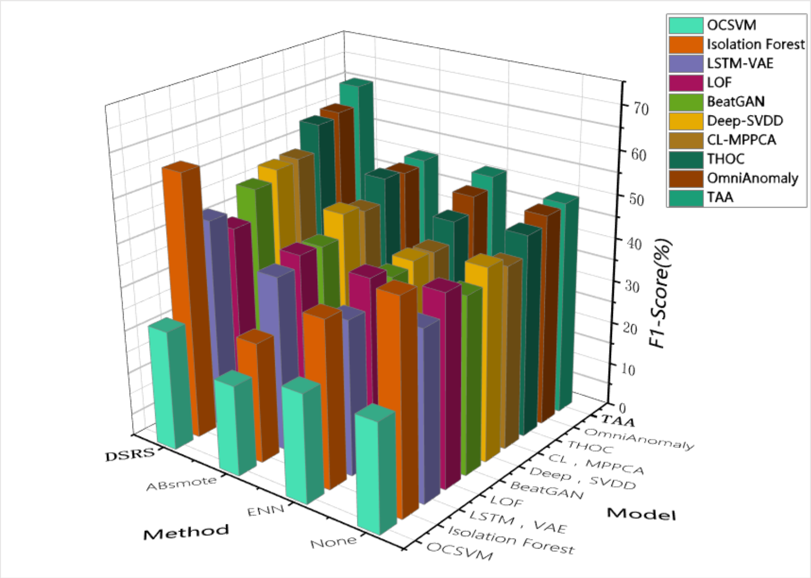

**Figure 4.** Comparison experiment between baseline model and fusion method.

The Receiver Operating Characteristic Curve (ROC curve) is a graphical tool that illustrates the performance of a classifier, showing the relationship between the True Positive Rate (TPR) and the False Positive Rate (FPR) of the classifier at different thresholds. The Area Under the Curve (AUC) value represents the area under the ROC curve, which is used to measure the performance of the classifier. A higher AUC value, closer to 1, indicates better classifier performance, while a lower AUC value, closer to 0, indicates poorer classifier performance. The gray dashed line in the graph represents an AUC value of 0.5, indicating a random classifier. It can be observed that under the DSRS method proposed in this study, the classifier's performance has significantly improved (Figure 5).

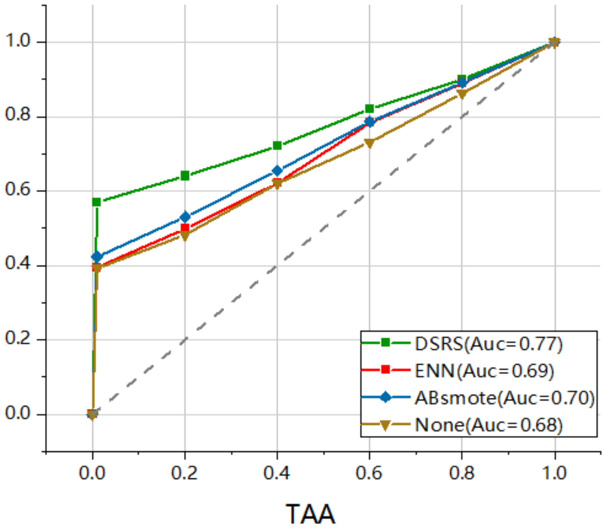

**Figure 5.** ROC curves (horizontal axis: False Positive Rate; vertical axis: True Positive Rate) for different methods.

### 4.4.2. Ablation Study

To gain deeper insights into the impact of the core components of our proposed method on the experimental results, we conduct a series of comparative experiments. As shown in the table below, by introducing the sequence association mechanism, the model's performance improved from 54.39 to 65.69, indicating that the sequence association mechanism played a crucial role in capturing the sequential relationships in the time series data. Building upon the association attention mechanism in Transformer (AAT), we apply the DSRS method to rebalance the experimental data, which significantly enhances the model's performance. Compared to other rebalancing methods, DSRS demonstrated superior performance in this experiment, demonstrating its effectiveness in addressing the imbalance issue in time series data. Regarding the selection of the kernel function for prior association, we compare the learnable power kernel function and the Gaussian kernel function. Although the power kernel function also achieves satisfactory results, the parameters of the Gaussian kernel function were easier to optimize and exhibited better performance in the experiment. For the statistical distance representing the loss function, we compare the L2 distance and the Jensen–Shannon divergence (JSD). The experimental results indicate that the L2 distance was not well suited for anomaly detection tasks, while JSD achieves better results, suggesting that JSD was more effective in measuring the distribution differences in time series data (Table 7).

**Table 7.** Ablation study result.

| Architecture | Imbalanced Method | Prior Association | Loss Distance | Avg F1-Score (%) |
|---|---|---|---|---|
| Transformer | DSRS | × | × | 50.62 |
| Ours | None | Gaussian | KL | 54.39 |
| | DSRS | Fix | KL | 50.16 |
| | | Power | KL | 62.01 |
| | | Gaussian | L2 | 38.26 |
| | | | JSD | 62.78 |
| | | | KL | 65.69 |

### 4.4.3. Anomaly Traceability Analysis

Existing research on air quality anomaly detection often struggles to balance the need for real-world contextual analysis with high precision on datasets, lacking interpretability in the results and requiring verification of the credibility of anomaly detection outcomes. This study further conducts a spatiotemporal analysis of pollution based on the actual air quality data of Meilan District in Haikou City, leveraging the experimental results of air quality anomaly detection. We match the pollution complaint corpus information spatiotemporally to achieve traceability in air quality anomaly detection, thereby expanding the continuity of anomaly detection tasks and enhancing their impact on real-world scenarios.

In this section, we analyze the complaint corpus dataset described above and introduce their spatiotemporal characteristics to match the results of anomaly detection. Spatial analysis of the latitude and longitude information in complaint corpus data can assist in verifying the effectiveness of the anomaly detection results in this study, further demonstrating the superiority of the proposed method. By matching the feedback of these two types of anomalies, we can attribute air pollution sources to both anthropogenic pollution and natural meteorological changes to some extent, aiding in optimizing the layout of air quality monitoring and improving the efficiency of environmental governance tasks by government departments.

We utilize reverse geocoding to restore the extracted geographical location information to latitude and longitude coordinates and visualize the geographical information in the pollution complaint data. Figure 6 illustrates the spatial distribution characteristics of the complaint corpus information using a heatmap. The results indicate that complaints are mainly concentrated in the comprehensive areas where residents are active, which is highly consistent with the selection of clustering sites in our experiments. This finding not only demonstrates the rationality and practical value of our selected site group in reflecting residents' concerns about air quality but also further validates the effectiveness of the site selection method.

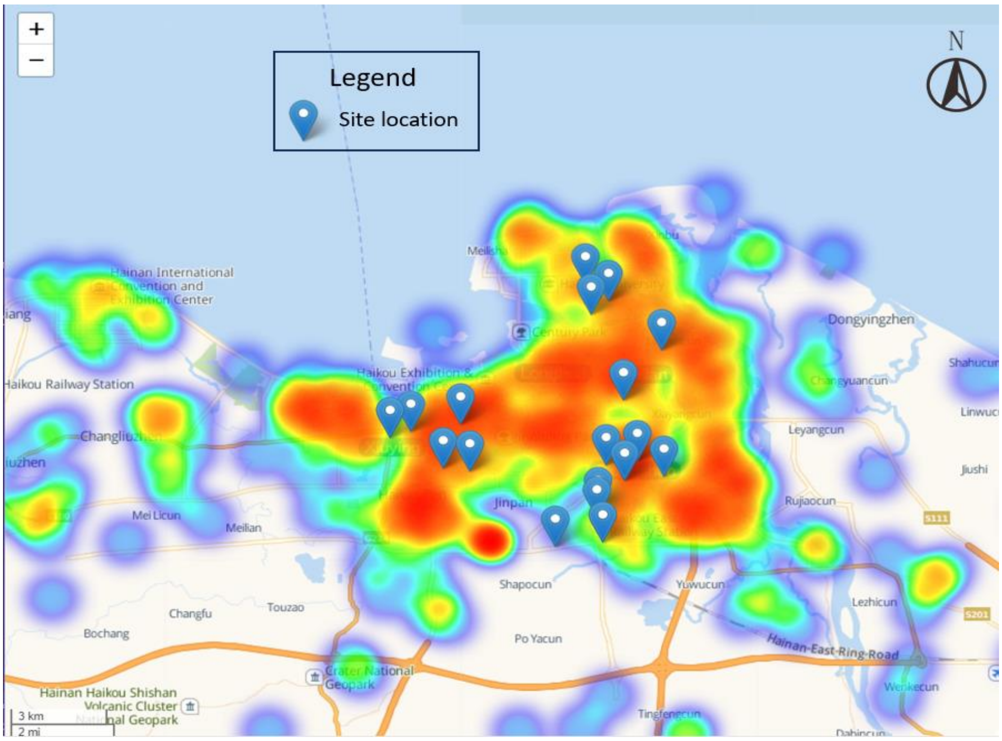

**Figure 6.** Spatial heat map of complaint corpus.

To analyze the performance and results of anomaly detection from a perspective different from traditional numerical analysis, this section compares the data from air monitoring stations in Haikou City with the addresses and pollutant information extracted in this study. We select the nearby area within one kilometer. Since the Earth is a sphere, calculations need to be performed using curved formulas. In this case, we use the Haversine formula to calculate the corresponding latitude changes within one kilometer, combined with the latitude and longitude coordinates of Haikou City (110.198, 20.044). According to this formula, the conversion formulas for latitude and longitude are as shown in Equations (12) and (13). The calculation of the longitude difference within one kilometer is as follows:

$$\delta lon = \frac{1}{(6371\ \text{km})/2\pi} \approx 0.008983^\circ \tag{18}$$

The calculation for latitude difference is as follows:

$$\delta lat = \frac{1}{((6371\ \text{km})/2\pi)\cos(20.044^\circ)} \approx 0.011406^\circ \tag{19}$$

Therefore, this study selects stations with latitude and longitude deviations between 0.008983 and 0.011406 for the analysis of anomalies. Here, the analysis is conducted on the data of station CEG41930016 on 29 December. This station is located at the intersection of Wenming East Road and Meiyuan Road, under the street lamp at Banqiao Seafood Plaza Wenming East Branch, with coordinates (110.371013, 20.037951). Specific geographic information is shown in the Figure 7.

By querying the map information, it is found that the distance between the two points is approximately 800 m, which meets the proximity principle of the station.

To trace the anomalies and demonstrate the association, the Figure 8 shows a segment of air quality time series data after 17:00 on that day.

From the Figure 8, it is evident that the air quality data sharply rose after 18:00 on that day, surpassing the anomaly baseline after 20:00, which was detected as an anomaly. However, the meteorological data shown in Figure 9 indicate that during this period, variables such as air pressure, wind speed, and temperature did not exhibit significant fluctuations. Therefore, it can be ruled out that the air quality anomaly was caused by meteorological changes.

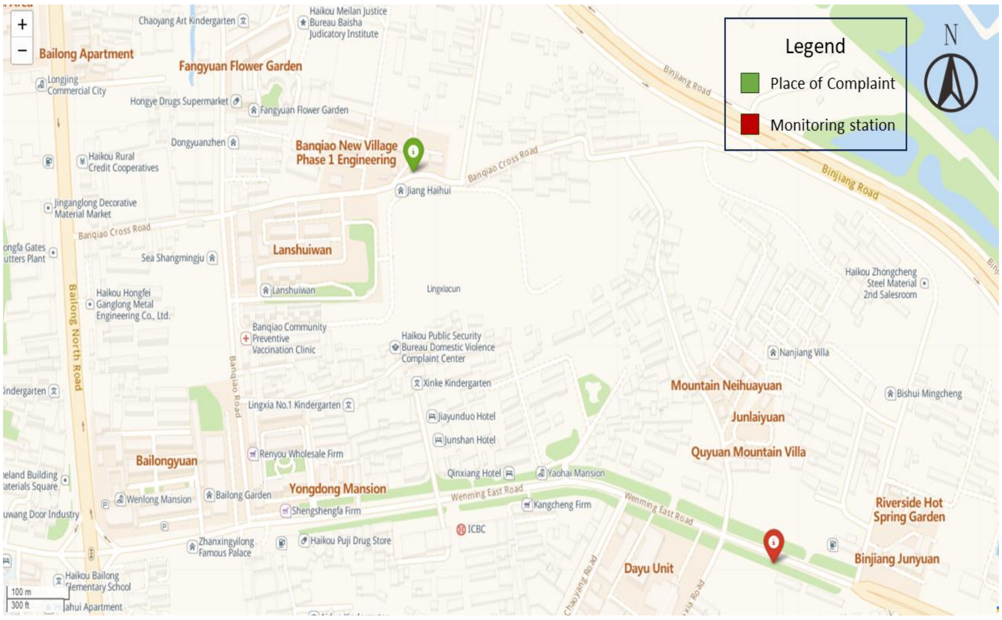

**Figure 7.** Specific geographic information.

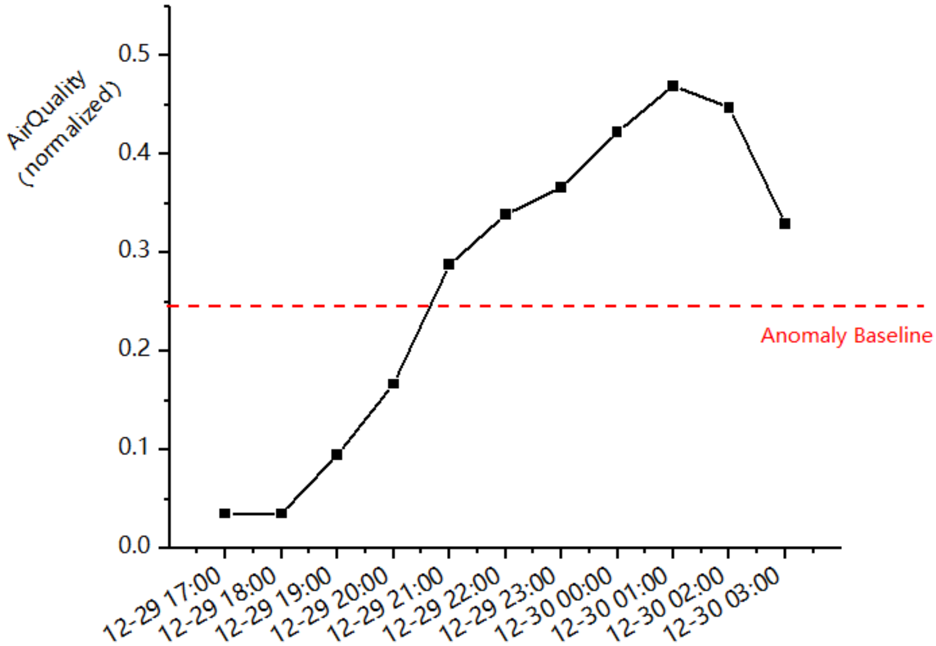

**Figure 8.** Complain about air quality data during the time of the incident.

Based on the provided information, "A resident complains of a gas leak at the Witte Cable Yard next to Phase II Blue Water Bay in Meilan", it was discovered that at 18:37 on the same day, a resident complained of a "gas leak" smell near the "Weite Cable Yard" next to "Lanshuibay Phase II" in Meilan District. Due to the similarity in timing between the incident and the detection, it was revealed by professionals during an inspection at 19:20 that it was the "gear oil odor" from the construction site rather than a gas leak. It is speculated that the volatile chemical substances in the gear oil evaporated and, combined with the air, caused a rapid increase in ozone in the air, thereby affecting the overall air quality. Furthermore, considering the subsequent decrease in data and no further complaints, it further confirms this sudden pollution caused by gear oil. This also demonstrates the comprehensiveness and accuracy of our air quality anomaly detection work. Not only can it accurately capture human-induced air pollution, but it also provides an additional channel for tracing complaints from residents, thereby adding value to improving air quality and meeting the needs of the people.

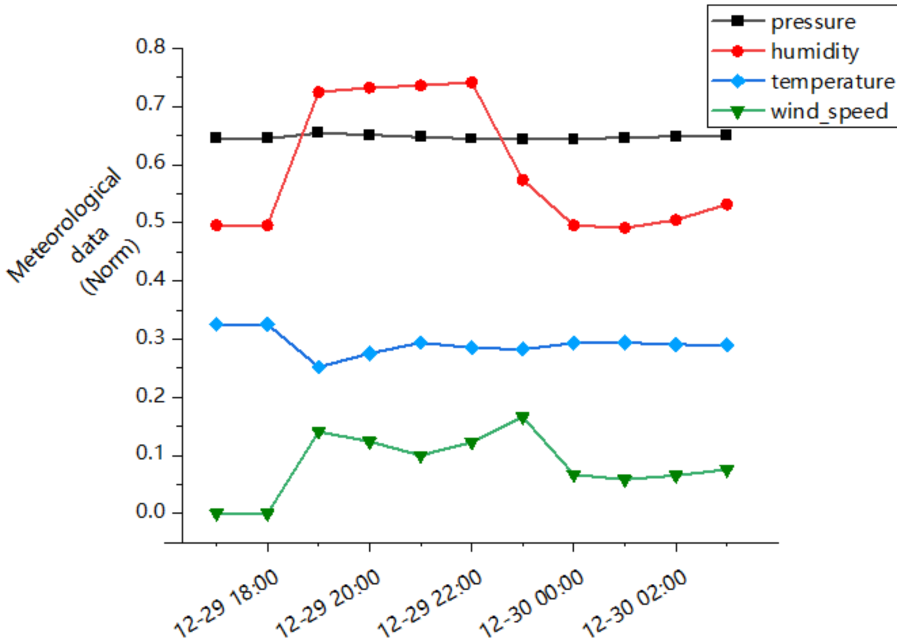

**Figure 9.** Complain about meteorological data during the time of the incident.

## 5. Discussion

### *5.1. The Advantages of Proposed Work*

The anomaly detection method proposed in this study has demonstrated its superiority across various experiments on real-world data. Our approach achieved the best results on three different methods and nine baseline model conditions. This approach, which addresses the issue of class imbalance in time series datasets for anomaly detection tasks, is a novel perspective to enhance detection performance. It offers a new entry point for research in anomaly detection, not limited to incremental improvements within different model architectures.

In addition, the association difference adopted in this paper can effectively amplify and identify the different performance patterns of abnormal points and normal points, thereby improving the detection effect. The new introduction of sequence associations enables architectures based on the Transformer model to be used in tasks dealing with time series data.

This integrated method can provide more inspiration for our subsequent studies. The method presented in this study also awaits validation on other types of time series datasets, with the hope of generalizing this approach to universal time series anomaly detection tasks.

### *5.2. Limitations in Experiment Settings*

Despite the promising results demonstrated by the experiments, this study still has certain limitations. A significant amount of effort was devoted to processing the dataset, striving to maximize the distinction between anomalous and normal patterns without compromising the temporal features. However, our dataset is not yet robust enough, which remains a challenging issue that urgently needs to be addressed in time series anomaly detection tasks. In view of the above problems, we will continue to collect more relevant data to make up for potential problems.

In comparison experiments with baseline models, the parameter setting process also requires further optimization. The lack of systematic verification could potentially lead to unexpected outcomes in experiments. Moreover, we also only conduct experiments on the air quality anomaly detection task data set in this paper, lacking the ablation comparison of multiple categories of data. In future work, we will try to use systematic tuning methods such as optimization algorithms to ensure the rationality of our parameter settings and avoid potential overfitting problems.

Subsequently, we also hope to enrich the evaluation index system to more effectively prove the advantages of our method. Because a single F1-score cannot be used as a complete performance index of the anomaly detection model, this point needs to be further explored and studied.

## 6. Conclusions

In this study, we utilize air quality data and meteorological data from 95 observation sites in Meilan District, Hainan Province, which constituted multidimensional time series data sequences, laying the foundation for dataset construction. Based on the third law of geography, we employ spatial clustering methods to associate data with similar environmental characteristics to enhance the quality of the dataset. To address the issue of imbalanced distribution of data in anomaly detection tasks, we propose the Deep Smooth Random Sampling (DSRS) method to resample existing datasets, which aims to alleviate the impact of data imbalance on the final results of anomaly detection tasks. Then, we integrate the Transformer model with the introduced association attention mechanism to magnify the difference between abnormal and normal patterns.

Under the above fusion method, the influence of data set imbalance on anomaly detection is successfully mitigated. In the comparison experiment with multiple baseline models and imbalance processing methods, the comprehensive performance of our method on real data sets is improved by about 20%, which is a relatively comprehensive improvement. In the future, we plan to explore a more robust evaluation index system to enrich the demonstration of the superiority of our method and strive to test it on air quality datasets from different regions under various circumstances.

**Author Contributions:** Conceptualization, X.L. and M.L.; methodology, M.L.; software, M.L.; formal analysis, P.W. and X.L.; investigation, X.Z. (Xiaoying Zhi), Z.D., X.Z. (Xiang Zhu), Y.Z., W.C., W.D., and W.F.; resources, P.W.; data curation, X.Z. (Xiaoying Zhi); writing—original draft preparation, M.L.; writing—review and editing, P.W., X.L., and Z.H.; visualization, M.L.; supervision, X.L.; funding acquisition, P.W. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research was funded by the Hainan Natural Science Foundation (120MS029).

**Data Availability Statement:** The datasets generated and analyzed during the current study are not publicly available but are available from the corresponding author upon reasonable request.

**Acknowledgments:** The authors would like to express sincere gratitude to the anonymous reviewers for their valuable feedback and constructive comments on our manuscript.

**Conflicts of Interest:** The authors declare no conflicts of interest.

## References

1. Chandola, V.; Banerjee, A.; Kumar, V. Anomaly detection: A survey. *ACM Comput. Surv. (CSUR)* **2009**, *41*, 1–58. [CrossRef]
2. Gao, B.; Kong, X.; Li, S.; Chen, Y.; Zhang, X.; Liu, Z.; Lv, W. Enhancing anomaly detection accuracy and interpretability in low-quality and class imbalanced data: A comprehensive approach. *Appl. Energy* **2024**, *353*, 122157. [CrossRef]
3. Pang, G.; Shen, C.; Cao, L.; Hengel, A.V.D. Deep learning for anomaly detection: A review. *ACM Comput. Comput. Comput. Comput. Comput. Surv. (CSUR)* **2021**, *54*, 1–38. [CrossRef]
4. Chalapathy, R.; Chawla, S.J. Deep learning for anomaly detection: A survey. *arXiv* **2019**, arXiv:1901.03407.
5. Smiti, A. A critical overview of outlier detection methods. *Comput. Sci. Rev.* **2020**, *38*, 100306. [CrossRef]
6. Mandhare, H.C.; Idate, S. A comparative study of cluster based outlier detection, distance based outlier detection and density based outlier detection techniques. In Proceedings of the 2017 International Conference on Intelligent Computing and Control Systems (ICICCS), Madurai, India, 15–16 June 2017; pp. 931–935.

7. Tang, B.; He, H. A local density-based approach for outlier detection. *Neurocomputing* **2017**, *241*, 171–180. [CrossRef]
8. Zou, D.; Xiang, Y.; Zhou, T.; Peng, Q.; Dai, W.; Hong, Z.; Shi, Y.; Wang, S.; Yin, J.; Quan, H. Outlier detection and data filling based on KNN and LOF for power transformer operation data classification. *Energy Rep.* **2023**, *9*, 698–711. [CrossRef]
9. Abhaya, A.; Patra, B.K. An efficient method for autoencoder based outlier detection. *Expert Syst. Appl.* **2023**, *213*, 118904. [CrossRef]
10. Çelik, M.; Dadaşer-Çelik, F.; Dokuz, A.Ş. Anomaly detection in temperature data using DBSCAN algorithm. In Proceedings of the 2011 International Symposium on Innovations in Intelligent Systems and Applications, Istanbul, Turkey, 15–18 June 2011; pp. 91–95.
11. Ankerst, M.; Breunig, M.M.; Kriegel, H.-P.; Sander, J. OPTICS: Ordering points to identify the clustering structure. *ACM Sigmod Rec.* **1999**, *28*, 49–60. [CrossRef]
12. Wang, X.; Duan, L.; Yu, Z.; He, C.; Bao, Z. Robust Multi-Kernel Nearest Neighborhood for Outlier Detection. In *IEEE Transactions on Knowledge and Data Engineering*; IEEE: Piscataway, NJ, USA, 2024.
13. Jordaan, E.M.; Smits, G.F.J.N.N. Robust outlier detection using SVM regression. In Proceedings of the 2004 IEEE International Joint Conference on Neural Networks, Budapest, Hungary, 25–29 July 2004; pp. 2017–2022.
14. Douiba, M.; Benkirane, S.; Guezzaz, A.; Azrour, M. An improved anomaly detection model for IoT security using decision tree and gradient boosting. *J. Supercomput.* **2023**, *79*, 3392–3411. [CrossRef]
15. Jha, R.S.; Ojha, K.; Mishra, A.; Mishra, R.; Kaushik, A. Cyber-Attacks and Anomaly detection on CICIDS-2017 dataset using ER-VEC. In Proceedings of the 2024 2nd International Conference on Disruptive Technologies (ICDT), Greater Noida, India, 15–16 March 2024; pp. 1453–1458.
16. Wen, Q.; Zhou, T.; Zhang, C.; Chen, W.; Ma, Z.; Yan, J.; Sun, L. Transformers in time series: A survey. *arXiv* **2022**, arXiv:2202.07125.
17. Xie, T.; Xu, Q.; Jiang, C. Anomaly detection for multivariate times series through the multi-scale convolutional recurrent variational autoencoder. *Expert Syst. Appl.* **2023**, *231*, 120725. [CrossRef]
18. Li, Y.; Peng, X.; Zhang, J.; Li, Z.; Wen, M. DCT-GAN: Dilated convolutional transformer-based GAN for time series anomaly detection. *IEEE Trans. Knowl. Data Eng.* **2021**, *35*, 3632–3644. [CrossRef]
19. Fan, J.; Liu, Z.; Wu, H.; Wu, J.; Si, Z.; Hao, P.; Luan, T.H. Luad: A lightweight unsupervised anomaly detection scheme for multivariate time series data. *Neurocomputing* **2023**, *557*, 126644. [CrossRef]
20. Guo, D.; Liu, Z.; Li, R. RegraphGAN: A graph generative adversarial network model for dynamic network anomaly detection. *Neural Netw.* **2023**, *166*, 273–285. [CrossRef] [PubMed]
21. Huai, Z.; Yang, G.; Tao, J. Spatial-temporal knowledge graph network for event prediction. *Neurocomputing* **2023**, *553*, 126557. [CrossRef]
22. Xu, P.; Gan, H.; Fu, H.; Zhang, Z. STEAMCODER: Spatial and Temporal Adaptive Dynamic Convolution Autoencoder for Anomaly Detection. *Knowl.-Based Syst.* **2023**, *279*, 110929. [CrossRef]
23. Wang, A.X.; Chukova, S.S.; Nguyen, B.P. Ensemble k-nearest neighbors based on centroid displacement. *Inf. Sci.* **2023**, *629*, 313–323. [CrossRef]
24. Liu, Z.; Cao, W.; Gao, Z.; Bian, J.; Chen, H.; Chang, Y.; Liu, T.-Y. Self-paced ensemble for highly imbalanced massive data classification. In Proceedings of the 2020 IEEE 36th International Conference on Data Engineering (ICDE), Dallas, TX, USA, 20–24 April 2020; pp. 841–852.
25. Chen, H.; Li, C.; Yang, W.; Liu, J.; An, X.; Zhao, Y. Deep balanced cascade forest: An novel fault diagnosis method for data imbalance. *ISA Trans.* **2022**, *126*, 428–439. [CrossRef]
26. Bao, L.; Juan, C.; Li, J.; Zhang, Y. Boosted near-miss under-sampling on SVM ensembles for concept detection in large-scale imbalanced datasets. *Neurocomputing* **2016**, *172*, 198–206. [CrossRef]
27. Liu, T.-Y. Easyensemble and feature selection for imbalance data sets. In Proceedings of the 2009 International Joint Conference on Bioinformatics, Systems Biology and Intelligent Computing, Shanghai, China, 3–8 August 2009; pp. 517–520.
28. 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]
29. He, H.; Bai, Y.; Garcia, E.A.; Li, S. ADASYN: Adaptive synthetic sampling approach for imbalanced learning. In Proceedings of the 2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence), Hong Kong, China, 1–8 June 2008; pp. 1322–1328.
30. Muntasir Nishat, M.; Faisal, F.; Jahan Ratul, I.; Al-Monsur, A.; Ar-Rafi, A.M.; Nasrullah, S.M.; Reza, M.T.; Khan, M.R.H. A Comprehensive Investigation of the Performances of Different Machine Learning Classifiers with SMOTE-ENN Oversampling Technique and Hyperparameter Optimization for Imbalanced Heart Failure Dataset. *Sci. Program.* **2022**, *2022*, 3649406. [CrossRef]
31. Zheng, Y.; Li, G.; Zhang, T. An Improved Over-sampling Algorithm based on iForest and SMOTE. In Proceedings of the 2019 8th International Conference on Software and Computer Applications, Penang, Malaysia, 19–21 February 2019; pp. 75–80.
32. Ling, C.X.; Sheng, V.S. Cost-sensitive learning and the class imbalance problem. *Encycl. Mach. Learn.* **2008**, *2011*, 231–235.
33. Weiss, K.; Khoshgoftaar, T.M.; Wang, D. A survey of transfer learning. *J. Big Data* **2016**, *3*, 9. [CrossRef]
34. Liu, H.; HaoChen, J.Z.; Gaidon, A.; Ma, T. Self-supervised learning is more robust to dataset imbalance. *arXiv* **2021**, arXiv:2110.05025.
35. He, H.; Garcia, E.A. Learning from imbalanced data. *IEEE Trans. Knowl. Data Eng.* **2009**, *21*, 1263–1284.

36. Yang, Y.; Zha, K.; Chen, Y.; Wang, H.; Katabi, D. Delving into deep imbalanced regression. In Proceedings of the International Conference on Machine Learning, Virtual, 18–24 July 2021; pp. 11842–11851.
37. Sun, B.; Feng, J.; Saenko, K. Return of frustratingly easy domain adaptation. In Proceedings of the AAAI Conference on Artificial Intelligence, Phoenix, AZ, USA, 12–17 February 2016.
38. Xia, S.; Wang, G.; Chen, Z.; Duan, Y. Complete random forest based class noise filtering learning for improving the generalizability of classifiers. *IEEE Trans. Knowl. Data Eng.* **2018**, *31*, 2063–2078. [CrossRef]
39. Xu, J.; Wu, H.; Wang, J.; Long, M. Anomaly transformer: Time series anomaly detection with association discrepancy. *arXiv* **2021**, arXiv:2110.02642.
40. Su, Y.; Zhao, Y.; Niu, C.; Liu, R.; Sun, W.; Pei, D. Robust anomaly detection for multivariate time series through stochastic recurrent neural network. In Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, Anchorage, AK, USA, 4–8 August 2019; pp. 2828–2837.
41. Ruff, L.; Vandermeulen, R.; Goernitz, N.; Deecke, L.; Siddiqui, S.A.; Binder, A.; Müller, E.; Kloft, M. Deep one-class classification. In Proceedings of the International Conference on Machine Learning, Stockholm Sweden, 10–15 July 2018; pp. 4393–4402.
42. Shen, L.; Li, Z.; Kwok, J. Timeseries anomaly detection using temporal hierarchical one-class network. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 13016–13026.
43. Majzoub, H.; Elgedawy, I. AB-SMOTE: An affinitive borderline SMOTE approach for imbalanced data binary classification. *Int. J. Mach. Learn. Comput.* **2020**, *10*, 31–37. [CrossRef]

**Disclaimer/Publisher's Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.