Data Complexity Measures of Imbalanced Data Classification Problems

***Abstract*—We propose a series of data complexity metrics, which can measure the impact of various data factors on imbalanced classification before classifier training and testing. Our metrics can be divided into two categories according to our understanding of this problem. They are fundamental and non-fundamental factor metrics. The former is designed for fundamental factors that can independently impact the performance of imbalanced data classification, including overlap degree (OD) and noise degree (ND). The latter is designed for non-fundamental factors that can influence the performance only when fundamental factors exist, which include a small disjunct degree based on overlapped instances (SDO) and an imbalance ratio based on overlapped instances (IRO). We run experiments on real-world imbalanced datasets to verify the effectiveness of the proposed metrics. The correlation analysis of experimental results has shown that our metrics are valid and can be used to guide the selection of suitable solving approaches for imbalanced datasets.**

I. Introduction

The classification of imbalanced data is problematic in many application fields, including medical diagnosis [1], software detection [2], computer vision [3], bio-informatics [4] and others [5]. The main characteristic of this problem is the skewed data distribution of the dataset, which means that most instances belong to one class (the majority class), and the rest belong to the other (the minority class). This characteristic leads to the classification result bias toward the majority class.

Researchers have designed many techniques to handle this problem that mainly include two kinds: data-level and algorithm-level strategies. The former [6] aims to solve this problem by changing the data distribution of the imbalanced dataset to get a balanced dataset. The latter [7] addresses this problem by increasing the importance of the minority class in the model learning or decision process. In summary, these solving methods aim to decrease the impact of class imbalance.

However, several studies [8], [9] have found that the class imbalance is not the only factor that makes the classifier not work well in the imbalanced dataset. Other data factors such as noise, overlap, and small disjunct are more severe than class imbalance, which can make imbalanced data classification more challenging, as shown in Fig. 1. Hence, imbalance-solving methods that only focus on decreasing the impact of class imbalance are not effective when the imbalanced dataset contains other data factors.

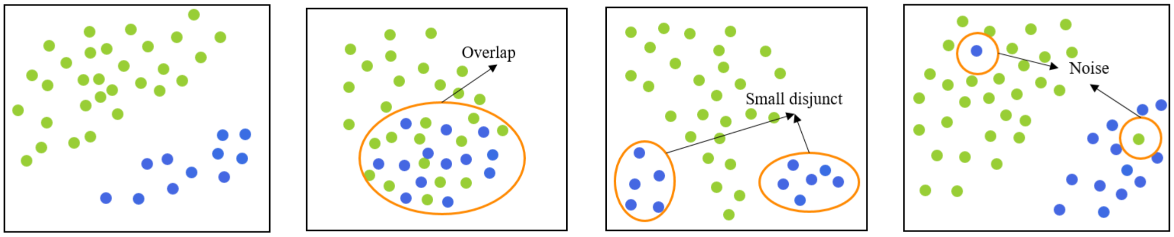
TABLE I
Correlation analysis results of the influence of a single data factor (SD is small disjunct.)

	Overlap	Noise	Imbalance	SD
SVM	-0.863	-0.970	0.260	-0.408
kNN	-0.918	-0.978	0.361	-0.412
RF	-0.877	-0.976	0.196	-0.401
DT	-0.914	-0.952	0.012	-0.474
AdaBoost	-0.882	-0.971	0.201	-0.401

Since the performance of the imbalanced data classification is strongly data-dependent, given an imbalanced dataset with low performance on the classification, one has no idea which data factor is the leading cause of the performance loss. Therefore, we aim to use data complexity metrics to analyze the relationship between the degradation of classification results and data factors of the training dataset, specifically noise, overlap, and small disjunct. Although a few studies have proposed data complexity metrics, their limitations are apparent. Most of them reflect the overall data complexity, which can only evaluate the difficulty of addressing this problem. Thus, they cannot be used to assess the impact of specific data factors and find the leading cause that makes the performance loss. To remedy the above limitations, we propose metrics for imbalanced datasets, which differ from prior works in two aspects. First, we emphasize measuring the difficulty of specific data factors for imbalanced data classification rather than the overall difficulty. Secondly, we provide a new type of classification for data complexity metrics based on our new understanding of the nature of this problem.

Our new understanding of this problem originates from analyzing the impact of various data factors. Based on the analysis results (Details are shown in Section IV.), we can divide data factors into two categories according to their characteristics. The former are fundamental factors that can independently influence the performance of imbalanced data classification, such as noise and overlap. The latter are non-fundamental factors that can only affect performance when fundamental factors exist, such as class imbalance and small disjunct. Therefore, an imbalanced data classification can be seen as a complicated classification mainly influenced by fundamental factors. Non-fundamental factors can enlarge the impact of fundamental factors. If an imbalanced dataset does not have fundamental factors, the effects of non-fundamental factors will be very limited or even ignored.

Then, we propose two types of data complexity metrics: fundamental factor metrics and non-fundamental factor metrics. They are metrics designed to evaluate fundamental factors and non-fundamental factors, respectively. The former includes overlap degree (OD) and noise degree (ND). The latter contains a small disjunct degree based on overlapped instances (SDO) and an imbalance ratio based on overlapped instances (IRO). We can use the proposed metrics to analyze the impact of



(a) The imbalanced dataset without other data factors.
(b) The imbalanced dataset with the overlap factor.
(c) The imbalanced dataset with the small disjunct factor.
(d) The imbalanced dataset with the noise factor.

Fig. 1. Imbalanced datasets with various data factors.

TABLE II
CORRELATION ANALYSIS RESULTS OF THE INFLUENCE OF CLASS IMBALANCE AND SMALL DISJUNCT UNDER FUNDAMENTAL FACTORS SCENARIOS

	Imbalance and noise	Small disjunct and noise	Imbalance and overlap	Small disjunct and overlap
SVM	-0.989	-0.869	-0.636	-0.588
kNN	-0.979	-0.720	-0.891	-0.511
RF	-0.983	-0.805	-0.585	-0.489
DT	-0.947	-0.875	-0.749	-0.502
AdaBoost	-0.997	-0.797	-0.689	-0.462

various data factors and determine which suitable approaches should be utilized before training on the dataset. We conduct experiments on real-world imbalanced datasets to assess the effectiveness of proposed metrics. The experimental results have shown that our metrics are effective and perform better than competing metrics.

Our main contributions lie in the following aspects: (1) We provide a new understanding of imbalanced data classification based on our analysis of the types of data factors. (2) We design two types of data complexity metrics, including fundamental factor metrics and non-fundamental factor metrics, which can be utilized to measure the difficulty of specific data factors for imbalanced datasets. (3) Our metrics can help researchers select suitable approaches to tackle imbalanced data classification before the classifier training.

We organized this paper as follows: In Section II, we provide a new understanding of the nature of the imbalanced data classification problem and point out weaknesses of related work. Section III presents two types of data complexity metrics for imbalanced datasets, which can be used to evaluate the difficulty of data factors. We conduct experiments and verify the effectiveness of our metrics in Section IV. In Section V, we use specific cases to illustrate that our metrics can guide imbalanced learning method selection. Section VI draws our conclusions.

II. THE NATURE OF THIS PROBLEM

Researchers have designed many methods to overcome imbalanced data classification. These methods mainly contain data-level and algorithm-level approaches. Data-level approaches aim to deal with this problem by changing the data distribution to obtain a balanced data distribution, such as SMOTE [6] and ADASYN [10]. Algorithm-level approaches [11] handle this challenge by increasing the importance of the minority class in the learning process. Most of these approaches assume that the leading cause of the deterioration of the classification performance is class imbalance. However, several studies [8], [9] have verified that class imbalance is not the only cause. Other data factors, including noise, overlap, and small disjunct, are more severe than class imbalance.

The noise appears because real-world data usually have many inconsistencies that negatively impact the data quality. The impact of noise is severe in the minority class for an imbalanced dataset [12]. If we directly oversampled the imbalanced dataset with noise, more noise will be generated blindly, which may degrade the classification performance. The overlap occurs when a region of the data space has a similar number of instances from each class. It can affect classification, particularly when the dataset is class-imbalanced. Prati et al. [13] have found that the loss of performance is influenced by the overlap and class imbalance. The small disjunct occurs when instances from the same class do not belong to a homogeneous region. Previous researchers have found that errors of classification are concentrated most heavily in small disjunct [14]. Japkowicz [15] has shown that the small disjunct is more responsible for the degradation in classification accuracy than class imbalance.

In this work, we aim to analyze further the impact of these data factors on imbalanced data classification. We first generate four groups of synthetic datasets that only change one data factor and datasets without the influence of other data factors. (Details of experiments can be seen in Section IV. Then, we train and test them on five commonly used classifiers. Finally, we make a correlation analysis to find the relationship between the impact of the data factor and the performance of classifiers. Table 1 shows that noise and overlap are strongly correlated with the performance of classifiers. We define these data factors as fundamental factors, as shown in Definition 1.

Definition 1: The fundamental factor for imbalanced data classification (FIDC). The fundamental factor can independently influence the performance of imbalanced data classification.

By contrast, small disjunct and class imbalance have weak correlations with the performance of classifiers. We further analyze the impact of small disjunct and class imbalance when imbalanced datasets contain fundamental factors. From Table 2, we can find that if an imbalanced dataset contains fundamental factors, small disjunct, and class imbalance are strongly correlated with the performance of classifiers. Then, we define small disjunct and class imbalance as non-fundamental factors as shown in Definition 2.

Definition 2: The non-fundamental factor for imbalanced data classification (NFIDC). The non-fundamental factor can only influence the performance of imbalanced data classification when fundamental factors exist.

Algorithm 1 Types of Data Factors

```
Input: the set of data factor $D_{factor} = \{factor_1, factor_2, ..., factor_n\}$, Correlation analysis results of the impact of the single data factor $correlation1$, Correlation analysis results of the impact of the data factor under fundamental factors scenarios $correlation2$.

for $i \leftarrow 1$ to $n$ do
    if $|correlation1| > 0.5$ then
        $factor_i$ is the fundamental factor;
    else if $|correlation2| > 0.5$ then
        $factor_i$ is the non-fundamental factor;
    end if
end for
Output: Types of data factors.
```

Specifically, we use the following Algorithm 1 to distinguish two types of data factors based on the correlation analysis results (Details are shown in Section IV). The concepts of FIDC and NFIDC have two advantages under the class imbalance scenario. First, it distinguishes data factors according to whether they can influence the performance of classifiers independently, which provides a deeper understanding of data factors. Thus, based on this concept, our proposed data complexity metrics can be more specific and targeted. By contrast, existing studies only indicate whether a data factor can affect imbalanced data classification, but they cannot determine the differences between the impacts of these data factors. Secondly, we can select corresponding strategies to tackle imbalanced data classification according to the categories of data factors that the imbalanced dataset contains. For example, if an imbalanced dataset has fundamental factors such as noise, we will first apply noise-solving methods to address the noise rather than use imbalance-solving methods directly.

We provide a new understanding of the imbalanced data classification problem based on these two definitions. This problem can be seen as a complicated classification mainly influenced by fundamental factors. Non-fundamental factors can enlarge the impact of fundamental factors. If an imbalanced dataset does not have fundamental factors, the effects of non-fundamental factors will be limited or even ignored.

Since the performance of the imbalanced data classification is strongly data-dependent, given an imbalanced dataset with low performance on the classification, one has no idea which data factor is the leading cause of the performance loss. Therefore, we aim to use data complexity metrics to analyze the relationship between the degradation of classification results and data factors of the training dataset.

The data complexity metrics for the classification problem originated in 2002. Ho and Basu [16] proposed a seminal work on the data complexity for classification, which studies 12 metrics that can evaluate the difficulty of the classification problem. These metrics can help to guide the selection of suitable classifiers for specific issues. However, studies [17], [18] have indicated that these data complexity metrics are ineffective when datasets are class imbalanced because the majority class denominates the metrics values. Thus, several researchers [19], [20] proposed complexity metrics for imbalanced data classification. Paper [21] introduced two data complexity metrics for the imbalanced data, which help explain the factors responsible for the deterioration in classifier performance. A framework complexity measurement (CM) [20] to study the relationship between the data complexity and imbalanced data problem was designed, which can be used to select suitable classifiers and approaches for dealing with imbalanced data with class overlap. Paper [22] provided a hardness estimate from the instance level, which can understand which instance is challenging to classify.

However, their limitations are apparent. Most of them reflect the overall data complexity, which can only evaluate the difficulty of addressing this problem. They cannot be used to assess the impact of specific data factors and find the leading cause of performance loss. Hence, we design a series of data complexity metrics to remedy their limitations.

III. Proposed Data Complexity Metrics

Based on understanding the nature of the imbalanced data classification in the previous section, we divide our data complexity metrics into fundamental data complexity metrics and non-fundamental data complexity metrics. They can measure the difficulty of fundamental factors and non-fundamental factors, respectively.

A. Fundamental Data Complexity Metrics

1) Noise Degree (ND): For each instance in the minority class of the imbalanced dataset, we calculate the Euclidean distance of its neighbors and find the k nearest neighbors based on the kNN method [23]. If the labels of all its k nearest neighbors differ from this instance, it is a noisy instance. The ND is defined as follows:

$$ND = \frac{N_{\text{noise}}}{N_{\min}} \tag{1}$$

Where N_{noise} is the quantity of noisy instances in the minority class and N_{min} is the quantity of minority class instances. The difference between ND and previous noise metrics lies in the former, focusing on the noise of the minority class because the performance of classifiers is mainly influenced by noise located in the minority class.

2) Overlap Degree (OD): Intuitively, an instance can be seen as overlapped if it has neighbors with different labels. Thus, we measure the overlap degree based on the kNN method. For each instance, we first calculate the Euclidean distance of its neighbors to find the k nearest neighbors. Then the instance overlap degree is defined as the percentage of the k neighbors for an instance that does not have the same class label. In addition, as we mentioned in noise degree, if the labels of all its k nearest neighbors differ from this instance, it is noise rather than overlap. We define instance overlap degree (IOD) as equation (2).

$$IOD\left(x_{i,j}\right) = \frac{kNN\left(x_{i,j}, D - D_j\right)}{k-1} \tag{2}$$

Then, based on IOD, we further calculate the overlap degree of the dataset by averaging the IOD of all minority class instances because the minority class instances mainly influence the performance of classifiers. We define overlap degree (OD) as equation (3).

$$OD = \frac{1}{N_{\text{min}}} \sum_{i=1}^{N_{\text{min}}} IOD\left(x_{i,j}\right) \tag{3}$$

Where N_{min} is the number of minority class instances and $x_{i,j}$ is the $i-th$ instances for class j. The details of the overlap degree are given in Algorithm 2.

Algorithm 2 Overlap Degree (OD)

Input: Dataset D, the quantity of the minority class samples N_{min}, the parameter of kNN method k.

```
for $i \leftarrow 1$ to $N_{min}$ do
    Calculate the number of its neighbors that have different labels with itself: $kNN\left(x_{i,j}, D - D_j\right)$;
    if $kNN\left(x_{i,j}, D - D_j\right) = k$ then
        $x_{i,j}$ is a noise and remove it from $D$;
    end if
end for
Calculate $IOD$ by equation (2);
Calculate $OD$ by equation (3);
```

Output: Overlap Degree (OD).

B. Non-fundamental Data Complexity Metrics

Non-fundamental factors can affect imbalanced data classification when the dataset contains fundamental factors. Therefore, our non-fundamental metrics measure the imbalanced dataset with fundamental factors.

1) Imbalance Ratio Based on Overlapped Instances (IRO): We design a metric named IRO to evaluate the imbalance ratio of the overlapped instances. The definition of IRO is shown as follows:

$$IRO = \frac{N_{omaj}}{N_{omin}} \tag{4}$$

Where N_{omin} and N_{omaj} represent the quantity of overlapped instances in the minority and majority classes, respectively.

2) Small Disjunct Degree based on Overlapped Instances (SDO): Related work on measuring the degree of small disjunct is very limited. Paper [24] proposed a quantitative metric for small disjunct, an error concentration curve. However, this metric is based on the results of classification, which cannot be used before the classifier training and testing. In this work, we design a metric named SDO to measure the small disjunct degree of the overlapped instances. We first use the DBCSAN method [25], [26] to find disjunct from overlapped instances. DBCSAN is a density-based non-parametric clustering method that can group closely packed points. Then, we find that we cannot determine whether a disjunct is a small disjunct since the definition of a small disjunct is influenced by the size of the dataset [24]. Thus, we measure the instance small disjunct degree (ISD) based on the largest disjunct in the dataset. For each instance, its disjunct degree is the number of instances in the largest disjunct divided by the number of instances in its disjunct, as shown in equation (5). In this way, the smaller disjunct, the larger the disjunct degree. Finally, we average ISD for all overlapped instances to get SDO by equation (6). The details of SDO are shown in Algorithm 3.

$$ISD\left(x_i\right) = \frac{disjunct_{max}}{disjunct(x_i)} \tag{5}$$

$$SDO = \frac{1}{N} \sum_{i=1}^{N} ISD\left(x_i\right) \tag{6}$$

C. Guidance of Usage

Table 3 summarizes our proposed data complexity metrics. We further introduce guidance on utilizing our metrics to handle imbalanced data classification. For researchers handling an imbalanced dataset, one can first use our metrics to measure the impact of various data factors. Then, we can select suitable approaches to solve according to the metrics values. For example, if the value of ND is very high (e.g., higher than 0.5), we should apply noise-solving approaches to address this dataset instead of directly using imbalanced solving approaches. By contrast, if the values of fundamental metrics are very small or nearly 0, we can directly apply imbalance-solving approaches to deal with the imbalanced dataset.

Algorithm 3 Small Disjunct Degree Based on Overlapped Instances (SDO)

Input: Dataset D, the quantity of instances in the dataset is N.

```
for $i \leftarrow 1$ to $N$ do
    if $IOD > 0$ then
        Use the DBSCAN method to find all clusters;
        Calculate the number of instances for each cluster; The number of instances for the disjunct that covers instance $x_i$ is $Disjunct(x_i)$;
        The cluster contains the largest number of instances is $Disjunct_{max}$;
    end if
end for
Calculate $ISD$ by equation (5);
Calculate $SDO$ by equation (6);
```

Output: Small Disjunct Degree Based on Overlapped Instances (SDO).

IV. Experiments

Our experiments contain two parts. We first use synthetic datasets experiments to analyze the effects of various data factors on imbalanced data classification. Then, we conduct real-world dataset experiments to assess the performance of our proposed data complexity metrics.

A. Experiments Setup

1) Baseline Classifiers: In experiments, we utilize five commonly used classifiers: support vector machine (SVM) [27], decision tree (DT) [28], kNN with k=5 [29], random forest (RF) [30], and AdaBoost [31]. All classifiers apply the default parameter in the scikit-learn library [32].

2) Evaluation Metrics: We use the area under the receiver operating characteristic curve (AUC) score to assess classification performance. AUC is a widely used measure for imbalanced data classification, which makes a trade-off between misclassified positive and correctly classified negative instances [33]. Correlation analysis can assess whether data factors impact imbalanced data classification and the performance of data complexity metrics. We use Spearman's rank correlation coefficient [34] to make a correlation analysis. This correlation coefficient can measure the rank between two variables with a monotonic function. The range of correlation is from -1 to 1, where -1 or 1 illustrates a perfect monotonously decreasing or increasing relationship, and 0 reflects no correlation between variables.

3) Competing Data Complexity Measures: We compare our overlap degree (OD) with three overlap measures: Fisher's Discriminant Ratio (F1), Volume of overlap region (F2) [16], and degOver [19]. The definition of F1 is shown as follows.

$$f = \frac{(\mu_1 - \mu_2)^2}{\sigma_1^2 + \sigma_2^2} \tag{7}$$

where $\mu_1, \mu_2, \sigma_1^2, \sigma_2^2$ are the means and variances of the two classes, respectively. We use the maximum f over all the feature dimensions for a multidimensional problem to describe the overlap level. F2 uses the maximum and minimum values of each attribute per class to calculate the size of the overlapping region. For instance, if an attribute whose values for class 0 range from 0.1 to 0.9, and the values for class 1 range from 0.6 to 1.2, then the size of the overlapping region for this attribute is 0.3. DegOver utilizes the 5-NN approach (k=5), which finds the 5-nearest neighbors for each sample and evaluates whether it is located in the overlapping or non-overlapping region.

We compare our noise metric (ND) with the outlier metric [9], which calculates the ratio of instances that belong to outliers. Our class imbalance metric (IRO) compares with the imbalance ratio, representing the ratio of instances in the majority and minority classes. Our small disjunct metric (SDO) compares with disjunct size (DS) [35]. DS first uses a slightly modified decision tree to form disjuncts. Then it calculates the number of instances in a disjunct divided by the number of cases covered by the largest disjunct.

B. The Impact of Single Data Factor

In this part, we conduct synthetic datasets experiments to analyze the impact of single data factors on imbalanced data classification.

1) Synthetic Datasets: We generate four groups of synthetic datasets to analyze the impact of four data factors. (1) Synthetic class imbalance datasets: The imbalance ratio is varied and datasets without other data factors. (2) Synthetic noise datasets: The noise level is varied under the class imbalance scenario. (3) Synthetic overlap datasets: The overlap level is varied under the class imbalance scenario. (4) Synthetic small disjunct datasets: The ratio of instances in small disjuncts varies under the class imbalance scenario. All synthetic datasets are generated by the scikit-learn library in Python and are binary classification datasets with four features that follow the same Gaussian distribution. Next, we provide more details about synthetic datasets.

For synthetic class imbalanced datasets, the minority class contains 100 instances, and the amount of instances in the majority class varies in the set {100, 1000, 2000, 3000, 4000, 5000, 6000, 7000, 8000, 9000, 10000}, where IRs are from 1 to 100. For the rest synthetic datasets, we provide them with the class imbalance scenario. The amount of instances in the minority class is 100, and the number of instances in the majority class varies in the set {500, 1000, 2000, 5000}, where IRs are 5, 10, 20, and 50, respectively. Then, we describe the characteristics of these synthetic datasets. We generate synthetic overlap datasets with different overlap levels by changing the distances between the centroids of two classes, varying from completely overlapped to completely separated. The distance in the set {0, 0.25, 0.5, 0.75, 1, 1.25, 1.5, 1.75, 2}. We generate synthetic noise datasets with different noise levels in the set {0, 0.05, 0.1, 0.15, 0.2, 0.25, 0.3}, where 0.05 means that 5% of the majority class instances are labeled as the minority class and the same amount of the minority class instances are marked as the majority class. The distances between the centroids of two classes for noise datasets are fixed at 2. We generate synthetic small disjunct datasets with the ratio of instances in small disjuncts varying in the set $\{0.1, 0.2, 0.3, 0.4, 0.5, 0.6\}$, where disjunct=0.1 means that 10% of the minority class instances in small disjuncts. We display some of the synthetic datasets in Fig. 2- 5.

2) Correlation analysis of the impact of the single data factor: Intuitively, we can find that noise and overlap can affect the performance of imbalanced data classification from Fig. 2. By contrast, class imbalance and small disjunct cannot. We further calculate the correlation analysis results. From Table 1, we can observe that noise and overlap are strongly correlated with the performance of imbalanced data classification. But class imbalance and small disjunct have low correlations. Hence, we conclude that noise and overlap are data factors that can independently influence the results of imbalanced data classification. Class imbalance and small disjunct cannot. Thus, noise and overlap are defined as fundamental factors according to Algorithm 1.

C. The Impact of the Data Factor under Fundamental Factors Scenarios

Since class imbalance and small disjunct cannot independently decrease the performance of imbalanced data classification, we aim to analyze further whether they can impact performance when the dataset contains fundamental factors. Therefore, we use four groups of synthetic datasets to analyze their impact. (1) Synthetic class imbalance datasets with overlap. (2) Synthetic class imbalance datasets with noise. (3) Synthetic small disjunct datasets with overlap. (4) Synthetic small disjunct datasets with noise. The details of the above four groups of datasets are the same as the previous synthetic datasets, but we set noise level = 0.1 and overlap level = 1, respectively.

From Table 2, we can find that class imbalance and small disjunct are highly correlated with the performance of imbalanced data classification when the dataset contains noise or overlap. Therefore, class imbalance and small disjunct are defined as non-fundamental factors according to Algorithm 1.

D. Real-world Datasets Experiments

1) Real-world Datasets: We use 26 real-world imbalanced datasets from imblearn [36] to conduct experiments. For each dataset, we calculate its ND, OD, IRO, and SDO. The details of the datasets are shown in Table 3, S(Samples), F(Features), IR(Imbalance ratio). From the values of data complexity metrics on real-world datasets, we can observe that most of them have high overlap degrees, which means that overlap is the major cause that makes imbalanced data classification challenging.

2) Results on Real-world Datasets: The results of correlation analysis on real-world imbalanced datasets are shown in Tables IV - VII. For noise metrics, we can find that ND shows high correlation values in all classifiers from Table IV. Fig. 6 provides an intuitive illustration of the correlation results on noise metrics. ND performs better than the competing noise metric because ND focuses on noisy instances in the minority

TABLE III
DESCRIPTION OF REAL-WORLD IMBALANCED DATASETS

Dataset	S	F	IR	OD	ND	IRO	SDO
ecoli	336	7	8.6	0.550	0.000	12.6	0.000
optical-digits	5620	64	9.1	0.089	0.000	10.4	0.000
satimage	6435	36	9.3	0.370	0.055	8.4	0.000
pen-digits	10992	16	9.4	0.003	0.004	9.9	0.000
abalone	4177	10	9.7	0.830	0.306	10.8	0.336
sick-euthyroid	3163	42	9.8	0.780	0.352	17.6	0.000
spectrometer	531	93	11	0.281	0.111	12.1	0.000
car-eval 34	1728	21	12	0.730	0.074	12.7	0.000
isolet	7797	617	12	0.328	0.008	12.9	0.000
us-crime	1994	100	12	0.706	0.378	15.7	0.147
yeast-ml8	2417	103	13	0.931	0.725	40.3	0.274
scene	2407	294	13	0.842	0.604	22.8	0.000
libras move	360	90	14	0.687	0.000	17	0.485
thyroid-sick	3772	52	15	0.856	0.285	17.4	0.000
coil-2000	9822	85	16	0.946	0.536	36.3	0.027
arrhythmia	452	278	17	1.000	0.750	87	0.000
solar-flare-m0	1389	32	19	0.714	0.461	37.8	0.478
oil	937	49	22	0.875	0.600	44.5	0.000
car-eval 4	1728	21	26	0.825	0.000	33.6	0.000
wine quality	4898	11	26	0.935	0.566	34	0.012
letter-img	20000	16	26	0.082	0.000	26.9	0.026
yeast-me2	1484	8	28	0.750	0.555	72	0.000
webpage	34780	300	33	0.461	0.129	37.2	0.037
ozone-level	2536	72	34	1.000	0.769	165	0.000
mammography	11183	6	42	0.430	0.188	50.7	0.266
abalone-19	4177	10	130	1.000	0.875	828	0.337

TABLE IV
THE CORRELATION RESULTS FOR NOISE METRICS ON REAL-WORLD DATASETS

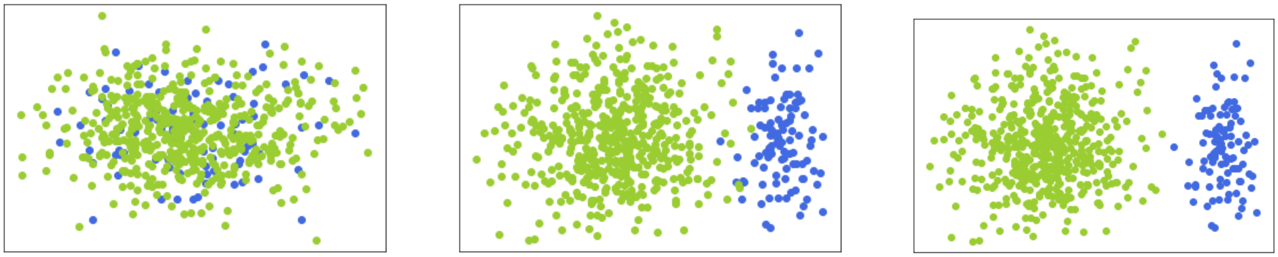
	outlier	ND
SVM	-0.549	$\mathbf{-0.804}$
kNN	-0.676	$\mathbf{-0.767}$
RF	-0.148	$\mathbf{-0.302}$
DT	-0.462	$\mathbf{-0.607}$
AdaBoost	-0.552	$\mathbf{-0.661}$

TABLE V
THE CORRELATION RESULTS FOR OVERLAP METRICS ON REAL-WORLD DATASETS

	F1	F2	degOver	OD
SVM	-0.076	0.109	-0.364	$\mathbf{-0.613}$
kNN	-0.074	-0.019	-0.634	$\mathbf{-0.965}$
RF	-0.339	-0.270	-0.399	$\mathbf{-0.521}$
DT	-0.248	0.156	-0.238	$\mathbf{-0.508}$
AdaBoost	-0.169	0.204	-0.326	$\mathbf{-0.584}$

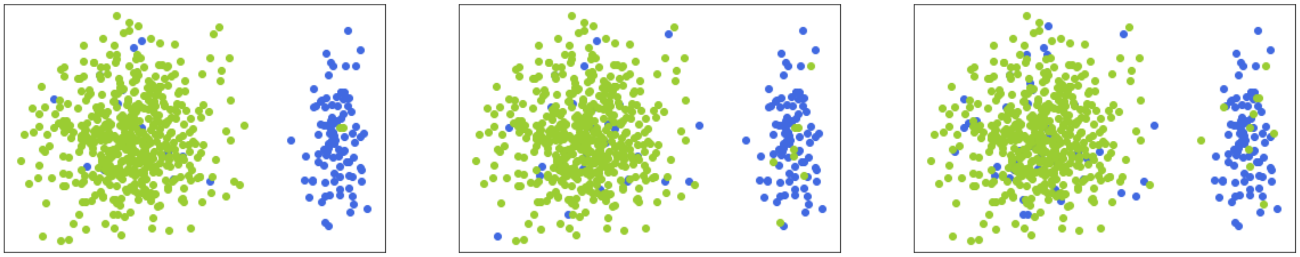
TABLE VI
THE CORRELATION RESULTS FOR SMALL DISJUNCT METRICS ON REAL-WORLD DATASETS

	DS	SDO
SVM	-0.209	$\mathbf{-0.280}$
kNN	-0.195	$\mathbf{-0.206}$
RF	$\mathbf{-0.114}$	-0.098
DT	-0.365	$\mathbf{-0.494}$
AdaBoost	-0.364	$\mathbf{-0.445}$



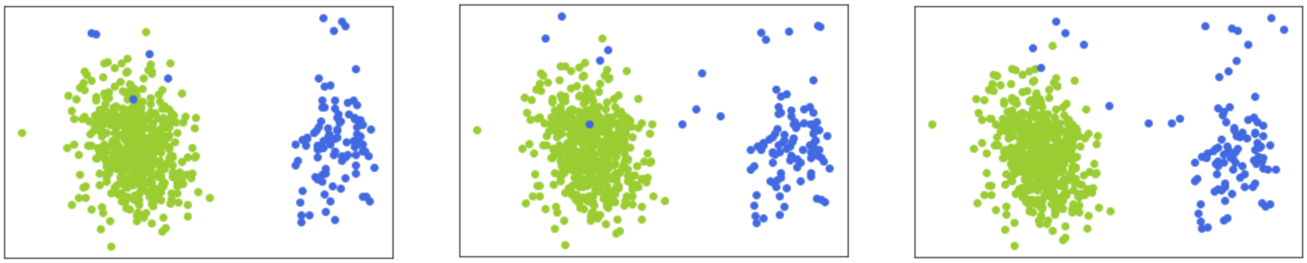
(a) distance=0, IR=5 (b) distance=1.5, IR=5 (c) distance=1.75, IR=5

Fig. 2. Synthetic overlap datasets.



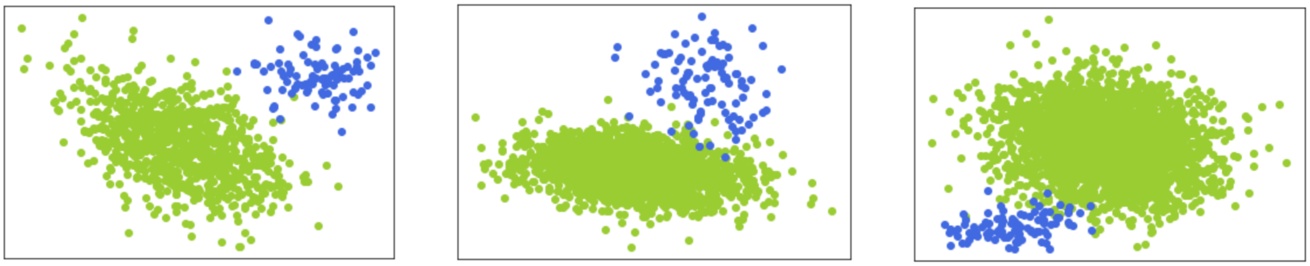
(a) noise=5%, IR=5 (b) noise=10%, IR=5 (c) noise =15%, IR=5

Fig. 3. Synthetic noise datasets.



(a) small disjunct ratio = 10%, IR=5 (b) small disjunct ratio = 20%, IR=5 (c) small disjunct ratio = 30%, IR=5

Fig. 4. Synthetic small disjuncts datasets.



(a) IR=10 (b) IR=20 (c) IR=30

Fig. 5. Synthetic class imbalance datasets.

TABLE VII
THE CORRELATION RESULTS FOR CLASS IMBALANCE METRICS ON REAL-WORLD DATASETS

	IR	IRO
SVM	-0.253	$\mathbf{-0.285}$
kNN	-0.245	$\mathbf{-0.280}$
RF	$\mathbf{-0.119}$	-0.112
DT	-0.294	$\mathbf{-0.334}$
AdaBoost	-0.298	$\mathbf{-0.337}$

class, which is the crucial part that influences the performance of imbalanced data classification.

OD achieves the highest correlation values in all overlap metrics, and most correlation values are larger than 0.5, as shown in Table V. Fig. 7 provides an intuitive illustration of the correlation results on overlap metrics. By contrast, competing overlap metrics perform much worse than OD. F1 uses the means and variances of two classes to calculate overlap levels, which are dominated by the majority class. The performance of F2 is terrible because it uses the maximum and minimum of two classes to find overlapped regions. However, noisy points and skewed data distribution can make the overlap region inaccurate. Moreover, degOver assigns the same overlap degree for all overlapped instances, which is not as accurate as OD. In summary, OD is a proper metric to describe the overlap degree for an imbalanced dataset. A dataset with a high OD value has a serious overlapped issue.

For class imbalance and small disjunct metrics, IRO and SDO have better performance than competing metrics, as shown in Table VI and Table VII. However, as we mentioned before, both class imbalance and small disjunct are non-fundamental factors; we may not directly use them to measure the dataset. IRO and SDO can be applied to assess the influence of class imbalance and small disjunct on imbalanced datasets with overlap issues.

V. DISCUSSION

This section explicitly studies two real-world imbalanced datasets from Table IV: letter-img and yeast-ml8. The letter-img dataset has IR = 26, which is high class imbalanced, but OD=0.0821 and ND=0, which means that the impact of overlap and noise is very small. Thus, we can handle this dataset by considering decreasing the effects of class imbalance rather than other data factors. This work utilizes four commonly used data-level class imbalance-solving approaches: SMOTE [6], ADASYN [10], Random over-sampling (ROS), and Random under-sampling (RUS). Table VIII shows the AUC scores of applying imbalance-solving approaches on various classifiers. We find that most classifiers perform well, even without using imbalance-solving strategies. Although the AUC scores in SVM and RF classifiers are low, they improve significantly by applying imbalance-solving techniques. This example illustrates that we can use our metrics to evaluate the impact of fundamental factors on imbalanced data classification. We can directly apply imbalance-solving methods without considering fundamental factors to address this dataset when values of fundamental factors metrics are low.

TABLE VIII
THE RESULTS ON LETTER-IMG DATASET

	None	SMOTE	ADASYN	ROS	RUS
SVM	0.8085	0.9499	0.9261	0.9420	0.9493
KNN	0.9892	0.9984	0.9983	0.9989	0.9582
RF	0.5000	0.9293	0.9526	0.9321	0.9482
DT	0.9643	0.9629	0.9566	0.9536	0.9660
AdaBoost	0.9630	0.9718	0.9794	0.9813	0.9733

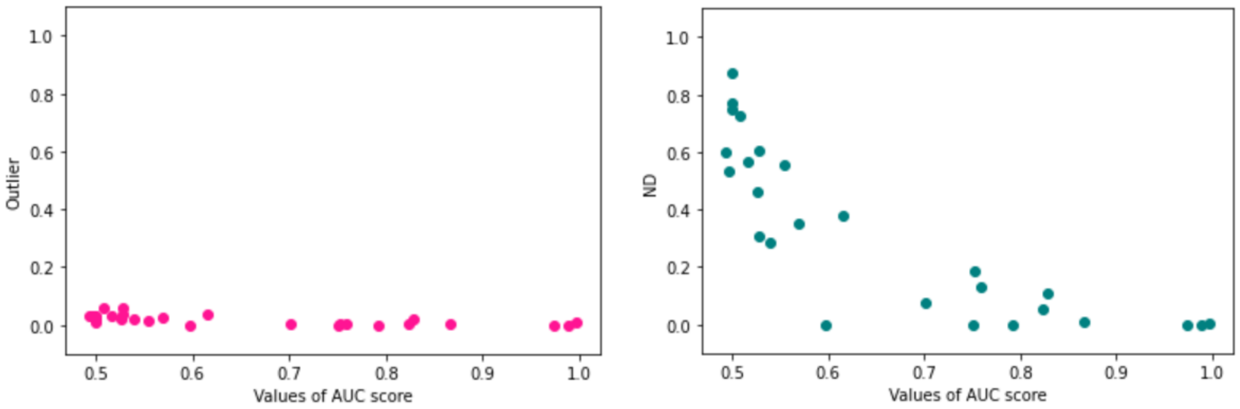
TABLE IX
THE RESULTS ON YEAST-ML8 DATASET

	None	SMOTE	ADASYN	ROS	RUS
SVM	0.5000	0.5194	0.5217	0.5162	0.5215
KNN	0.5091	0.5048	0.5162	0.5120	0.5680
RF	0.5000	0.5742	0.5492	0.5299	0.5783
DT	0.4549	0.5211	0.4801	0.5174	0.5443
AdaBoost	0.5114	0.5096	0.4824	0.4688	0.5576

In contrast, the yeast-ml8 dataset has IR=13, which is not high class imbalanced compared with the letter-img dataset, but OD=0.9318 and ND=0.7250, which means that the dataset is heavily influenced by overlap and noise. Table IX shows the AUC scores of using imbalance-solving approaches on various classifiers. We can observe that all classifiers perform poorly. The improvement from imbalanced solving techniques is also minimal. Some are near or equal to 0. This example shows that traditional imbalance-solving approaches are useless when imbalanced datasets contain fundamental data factors. This comparison again verifies that the class imbalance is not the leading cause of the degradation of imbalanced data classification since the letter-img dataset has a higher imbalance ratio than the yeast-ml8 dataset. Still, the former is much more challenging to solve than the latter.

VI. CONCLUSION

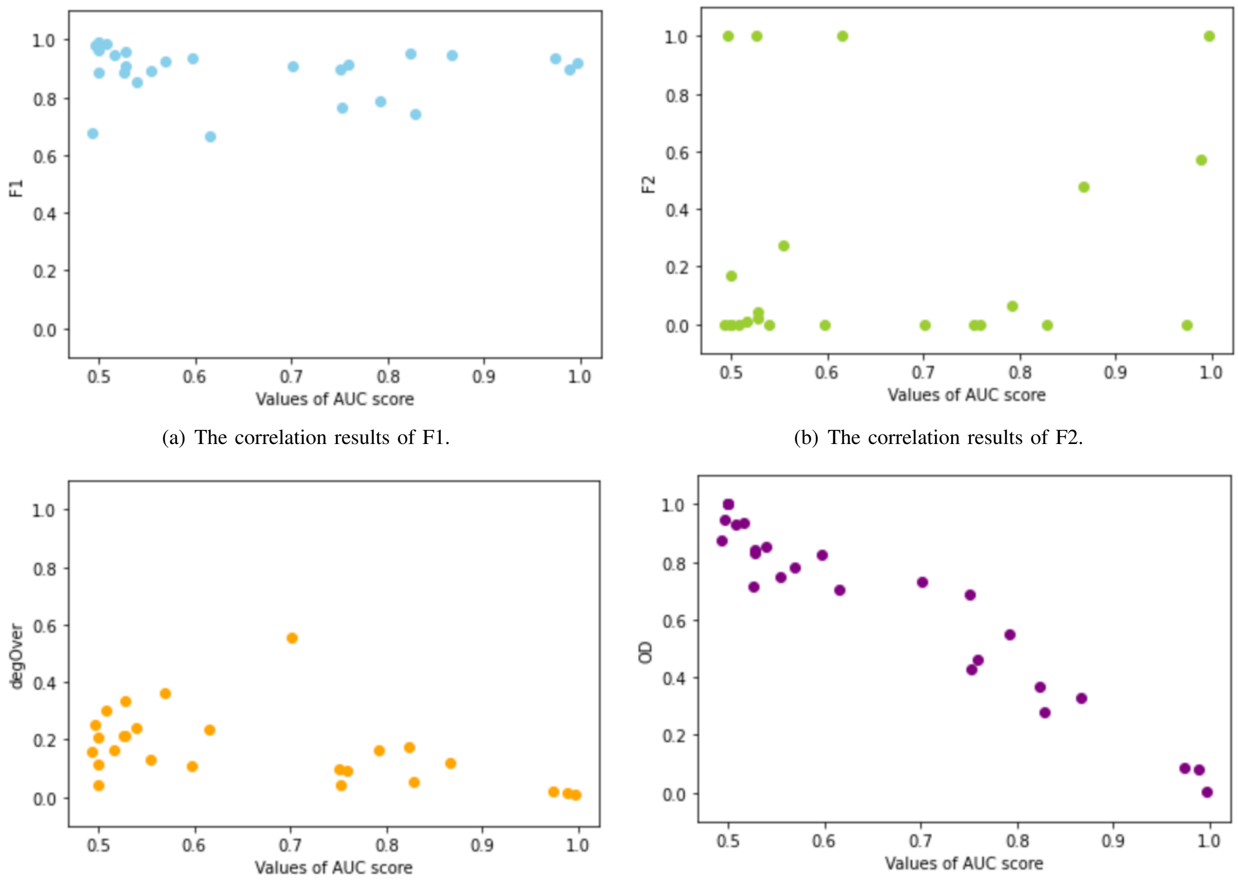
In this study, we propose a new understanding of imbalanced data classification based on the types of data factors. We divide data factors into two categories: fundamental and non-fundamental factors. The former can independently affect the performance of imbalanced data classification, and the latter cannot. The latter influences the performance when the imbalanced dataset contains fundamental factors. Therefore, imbalanced data classification can be seen as a complicated problem mainly influenced by fundamental factors, and non-fundamental factors can enlarge the impact of fundamental factors. Then we propose data complexity metrics overlap degree (OD) and noise degree (ND) for fundamental factors. For non-fundamental factors, we present a small disjunct degree based on overlapped instances (SDO) and an imbalance ratio based on overlapped instances (IRO). The experiments on real-world imbalanced datasets have shown that our metrics can be applied to analyze the intrinsic data factors for an imbalanced dataset before training and testing. The values of metrics can determine which addressing approaches are suitable to handle this imbalanced dataset.



(a) The correlation results of outlier metric.

(b) The correlation results of noise degree.

Fig. 6. The correlation results of Noise metrics.



(a) The correlation results of F1.

(b) The correlation results of F2.

(c) The correlation results of degOver.

(d) The correlation results of Overlap Degree.

Fig. 7. The correlation results of Overlap metrics.

References

[1] M. A. Mazurowski, P. A. Habas, J. M. Zurada, J. Y. Lo, J. A. Baker, and G. D. Tourassi, "Training neural network classifiers for medical decision making: The effects of imbalanced datasets on classification performance," *Neural networks*, vol. 21, no. 2-3, pp. 427–436, 2008.

[2] S. Wang and X. Yao, "Using class imbalance learning for software defect prediction," *IEEE Transactions on Reliability*, vol. 62, no. 2, pp. 434–443, 2013.

[3] O. Beijbom, P. J. Edmunds, D. I. Kline, B. G. Mitchell, and D. Kriegman, "Automated annotation of coral reef survey images," in *2012 IEEE conference on computer vision and pattern recognition*. IEEE, 2012, pp. 1170–1177.

[4] L. A. Bugnon, C. Yones, D. H. Milone, and G. Stegmayer, "Deep neural architectures for highly imbalanced data in bioinformatics," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 31, no. 8, pp. 2857–2867, 2019.

[5] G. Haixiang, L. Yijing, J. Shang, G. Mingyun, H. Yuanyue, and G. Bing, "Learning from class-imbalanced data: Review of methods and applications," *Expert systems with applications*, vol. 73, pp. 220–239, 2017.

[6] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: synthetic minority over-sampling technique," *Journal of artificial intelligence research*, vol. 16, pp. 321–357, 2002.

[7] Y. Cui, M. Jia, T.-Y. Lin, Y. Song, and S. Belongie, "Class-balanced loss based on effective number of samples," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 9268–9277.

[8] J. A. Sáez, J. Luengo, J. Stefanowski, and F. Herrera, "Smote–ipf: Addressing the noisy and borderline examples problem in imbalanced classification by a re-sampling method with filtering," *Information Sciences*, vol. 291, pp. 184–203, 2015.

[9] K. Napierala and J. Stefanowski, "Types of minority class examples and their influence on learning classifiers from imbalanced data," *Journal of Intelligent Information Systems*, vol. 46, no. 3, pp. 563–597, 2016.

[10] H. He, Y. Bai, E. A. Garcia, and S. Li, "Adasyn: Adaptive synthetic sampling approach for imbalanced learning," in *2008 IEEE international joint conference on neural networks (IEEE world congress on computational intelligence)*. IEEE, 2008, pp. 1322–1328.

[11] S. H. Khan, M. Hayat, M. Bennamoun, F. A. Sohel, and R. Togneri, "Cost-sensitive learning of deep feature representations from imbalanced data," *IEEE transactions on neural networks and learning systems*, vol. 29, no. 8, pp. 3573–3587, 2017.

[12] J. Van Hulse, T. M. Khoshgoftaar, and A. Napolitano, "Experimental perspectives on learning from imbalanced data," in *Proceedings of the 24th international conference on Machine learning*, 2007, pp. 935–942.

[13] R. C. Prati, G. E. Batista, and M. C. Monard, "Class imbalances versus class overlapping: an analysis of a learning system behavior," in *Mexican international conference on artificial intelligence*. Springer, 2004, pp. 312–321.

[14] D. R. Carvalho and A. A. Freitas, "A hybrid decision tree/genetic algorithm for coping with the problem of small disjuncts in data mining," in *Proceedings of the 2nd annual conference on genetic and evolutionary computation*, 2000, pp. 1061–1068.

[15] N. Japkowicz, "Class imbalances: are we focusing on the right issue," in *Workshop on learning from imbalanced data sets II*, vol. 1723, 2003, p. 63.

[16] T. K. Ho and M. Basu, "Complexity measures of supervised classification problems," *IEEE transactions on pattern analysis and machine intelligence*, vol. 24, no. 3, pp. 289–300, 2002.

[17] N. Anwar, G. Jones, and S. Ganesh, "Measurement of data complexity for classification problems with unbalanced data," *Statistical Analysis and Data Mining: The ASA Data Science Journal*, vol. 7, no. 3, pp. 194–211, 2014.

[18] A. L. Brun, A. S. Britto, L. S. Oliveira, F. Enembreck, and R. Sabourin, "Contribution of data complexity features on dynamic classifier selection," in *2016 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2016, pp. 4396–4403.

[19] M. M. P. M. d. Silva, "Addressing data complexity in imbalanced contexts," Ph.D. dissertation, Universidade de Coimbra, 2018.

[20] M. N. Anwar, "Complexity measurement for dealing with class imbalance problems in classification modelling: a thesis submitted in fulfilment of the requirements for the degree of doctor of philosophy, massey university, 2012," Ph.D. dissertation, Massey University, 2012.

[21] D. Singh, A. Gosain, and A. Saha, "Weighted k-nearest neighbor based data complexity metrics for imbalanced datasets," *Statistical Analysis and Data Mining: The ASA Data Science Journal*, vol. 13, no. 4, pp. 394–404, 2020.

[22] A. C. Lorena, L. P. Garcia, J. Lehmann, M. C. Souto, and T. K. Ho, "How complex is your classification problem? a survey on measuring classification complexity," *ACM Computing Surveys (CSUR)*, vol. 52, no. 5, pp. 1–34, 2019.

[23] S. Zhang, X. Li, M. Zong, X. Zhu, and R. Wang, "Efficient knn classification with different numbers of nearest neighbors," *IEEE transactions on neural networks and learning systems*, vol. 29, no. 5, pp. 1774–1785, 2017.

[24] G. M. Weiss and H. Hirsh, "A quantitative study of small disjuncts," *AAAI/IAAI*, vol. 2000, no. 665-670, p. 15, 2000.

[25] M. Ester, H.-P. Kriegel, J. Sander, X. Xu *et al.*, "A density-based algorithm for discovering clusters in large spatial databases with noise." in *kdd*, vol. 96, no. 34, 1996, pp. 226–231.

[26] E. Schubert, J. Sander, M. Ester, H. P. Kriegel, and X. Xu, "Dbscan revisited, revisited: why and how you should (still) use dbscan," *ACM Transactions on Database Systems (TODS)*, vol. 42, no. 3, pp. 1–21, 2017.

[27] D. Nasien, S. S. Yuhaniz, and H. Haron, "Statistical learning theory and support vector machines," in *2010 Second International Conference on Computer Research and Development*. IEEE, 2010, pp. 760–764.

[28] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, *Classification and regression trees*. Routledge, 2017.

[29] R. O. Duda, P. E. Hart *et al.*, *Pattern classification and scene analysis*. Wiley New York, 1973, vol. 3.

[30] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[31] Y. Freund and R. E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," *Journal of computer and system sciences*, vol. 55, no. 1, pp. 119–139, 1997.

[32] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg *et al.*, "Scikit-learn: Machine learning in python," *the Journal of machine Learning research*, vol. 12, pp. 2825–2830, 2011.

[33] J. M. Johnson and T. M. Khoshgoftaar, "Survey on deep learning with class imbalance," *Journal of Big Data*, vol. 6, no. 1, pp. 1–54, 2019.

[34] M. Kendall and J. D. Gibbons, "Rank correlation methods. charles griffin book series," 1990.

[35] M. R. Smith, T. Martinez, and C. Giraud-Carrier, "An instance level analysis of data complexity," *Machine learning*, vol. 95, no. 2, pp. 225–256, 2014.

[36] G. Lemaître, F. Nogueira, and C. K. Aridas, "Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning," *The Journal of Machine Learning Research*, vol. 18, no. 1, pp. 559–563, 2017.