

# Classifying Imbalanced Data: The Relevance of Accuracy and Feature Importance

Torben Widmann  
 Ulm University  
[torben.widmann@uni-ulm.de](mailto:torben.widmann@uni-ulm.de)

## Abstract

*The use of AI and ML algorithms can only contribute successfully to data-driven decision making if the underlying data is of sufficiently good quality. However, the effort of ensuring good data quality (DQ) must be proportionate to the potential impact of poor DQ. In this work, we therefore investigate the impact of DQ defects on the common and challenging task of classifying imbalanced data. We contribute to theory and practice by being the first to investigate the impact of DQ according to the particular DQ dimension accuracy and by examining the relevance of the importance of attributes with respect to the classification. Underpinning the significance of DQ, our experiments show that already few inaccuracies can lead to a considerably worse classification, that efficient data cleaning can be limited to a few attributes, and that distance-based algorithms are more affected by defects in less important attributes.*

**Keywords:** Data Quality, Accuracy, Feature Importance, Imbalanced Data Classification

## 1. Introduction

Data holds fundamental value and provides companies with a potential, sustainable competitive advantage (Hagiu, A., & Wright, J., 2020; Ngai et al., 2017). Indeed, companies strive to leverage their data in order to reduce costs, provide better products or services, improve customer relations (Raguseo, 2018), and to make better business decisions (Vassakis et al., 2018). In particular, artificial intelligence (AI) and machine learning (ML) algorithms are being used by organizations to harness the fundamental value of data (Sandeep et al., 2022). However, the results of quantitative analysis are ultimately highly dependent on the input data (Mans et al., 2015; Serhani et al., 2016), and insights from data analysis in real-world applications are often limited by insufficient veracity (Janssen et al., 2017). The reason for this is poor data quality (DQ), which can lead to a biased outcome of AI

and ML algorithms (Janssen et al., 2020) and thus imprecise conclusions, a lack of confidence in results, customer dissatisfaction, and increased costs (Hariri et al., 2019; Hazen et al., 2014; Redman, 1998).

This is particularly related to the challenge of classifying data, where accuracy and interpretability of classifiers are negatively affected (García et al., 2015; Zhu & Wu, 2004). Furthermore, in many real-world applications, such as finance, healthcare, or fault diagnosis, real-world data is commonly imbalanced, which further complicates classification (Fernández et al., 2018). A dataset is imbalanced if the number of instances in at least one class is much smaller than the number of instances in the other classes (Fernández et al., 2018; Stefanowski, 2016). In the case of two classes, the class with the larger proportion is called majority or negative class, while the smaller class is called minority or positive class. As a result of class imbalance, classifiers are biased towards the majority class and it is significantly more difficult for them to recognize patterns in the minority class, leading to the minority class being misclassified more often (Fernández et al., 2018; Stefanowski, 2016). Although imbalance can be related to DQ, e.g. if the real world is not correctly or completely represented by the data, it is generally an intrinsic property of the data and an independent complicating factor (He & Garcia, 2009).

Poor DQ can be caused by various errors, which translate into different DQ defects such as inaccurate, incomplete, conflicting, or outdated data (Chengalur-Smith et al., 1999; Ghasemaghahi & Calic, 2019). Thus, DQ is conceived as a construct of several dimensions (Lee et al., 2002; Redman, 1997) and requires a good and economically oriented management (Heinrich et al., 2018). However, since ensuring good DQ is time-consuming and costly (Kruse et al., 2015), it is crucial to put the effort and the associated costs in relation to the possible impact of poor DQ which in turn requires the assessment of possible consequences. Nevertheless, even data mining practitioners who spend a large part of their time on data processing and verifying its quality (Gupta & Gupta, 2019; Kruse et al., 2015) are often

unaware of the impact of DQ defects on subsequent analytics tasks (van Hulse & Khoshgoftaar, 2009). Thus, it is necessary to investigate the impact of DQ defects on data-driven decision making to assess potentially biased and inaccurate models (Feldman et al., 2018; Picado et al., 2020).

Since poor DQ and imbalanced data are both ubiquitous and serious challenges to data-driven decision making, it is of utmost importance to assess the extent to which different DQ defects affect classifiers in the context of imbalanced data. In this line, we extend the previous literature on the impact of attribute defects, i.e., defects within the independent variables, by assessing the impact of attribute defects with new properties to improve targeted data cleaning. In contrast to previous studies, we do not investigate the effects of random noise but specifically the effects of different inaccuracies. In addition, we examine the relevance of the importance of attributes with respect to the classification by corrupting attributes depending on the feature importance. For both investigations, we vary the amount of DQ defects and use ten imbalanced datasets with binary classes to evaluate and compare the performance of six classification algorithms.

With our study, we contribute to theory and practice in two major ways. We are the first to provide insights into the impact of attribute defects on classifying imbalanced data particularly by the DQ dimension accuracy. Furthermore, we are the first to shed light on the relevance of feature importance to the impact of attribute defects. Our experiments highlight the importance of DQ as a critical factor for the successful application of AI and ML algorithms. Even a few but large inaccuracies lead to a significantly worse classification, while tree-based algorithms classify the positive class best among the examined classifiers when attribute defects are present. In addition, practitioners can utilize these insights to clean their (imbalanced) data more efficiently by focusing on the most important attributes, since defects within the 75% least important attributes are almost equally insignificant.

The remainder of this work is organized as follows. First, we give an overview of prior works on the impact of DQ defects in the context of imbalanced data. Then, we describe our experimental design and present the results of the conducted experiments. Finally, we discuss these results and limitations of the experimental design and conclude with a final reflection.

## 2. Related Work

With the large body of existing work on the impact of poor DQ on the performance of AI & ML algorithms, there are two different perspectives on the topic of DQ and thus on the DQ defects that have been studied

(Heinrich & Helfert, 2003). Some of these works take the perspective of so-called quality of design, i.e. whether the data are fit for use (e.g., T. Jylhä & Suvanto, 2015; V. Jylhä et al., 2016). Others take the perspective of the so-called quality of conformance, i.e., to what degree the available data represent the associated real-world entities. These works consider DQ defects either by dimensions such as accuracy and completeness (e.g., Heinrich et al., 2021; Wang & Xu, 2019) or as general noise in the data (e.g., Gupta & Gupta, 2019; Johnson & Khoshgoftaar, 2022).

In the following, we discuss studies on the impact of DQ defects specifically on the classification of imbalanced data. Despite its common occurrence and high impact on classification with poor DQ (Weiss & Provost, 2003), DQ defects in imbalanced data have not been addressed in many studies. Most of these works (e.g., Kennedy et al., 2021; Seiffert et al., 2014; van Hulse & Khoshgoftaar, 2009) only consider defects in the class variable.

To the best of our knowledge, there are four papers on the impact of attribute errors on the classification of imbalanced data. Following the guidelines of Webster and Watson (2002), we searched the databases ScienceDirect, ACM Digital Library, IEEE Xplore, AIS Library, and Google Scholar. We used combinations of terms from {impact, consequences, (poor) data quality, defects, issues, noise, inaccuracies, imbalanced data, skewed data} and performed a forward and backward search starting from highly relevant papers, which resulted in four papers that are presented in the remainder of this section.

Both Folleco, Khoshgoftaar, van Hulse, and Bullard (2008) and Folleco, Khoshgoftaar, and Napolitano (2008) investigate DQ defects in the class as well as attribute variables when investigating the classification performance for imbalanced data. In their experiments, they insert different levels of attribute defects into the data by replacing a predefined number of attribute values with random values drawn from the respective attribute distribution but from the opposite class. In addition to this inserted noise, they assess the significance of each attribute by ranking all attributes according to the Kolmogorov-Smirnov (KS) statistic in order to insert the attribute defects into attributes with varying significance. However, only the average impact across multiple subsets of corrupted attributes defects is presented and the individual impact of different defects is not investigated. The two studies differ in the sense that Folleco, Khoshgoftaar, van Hulse, and Bullard (2008) distinguish among different classifiers as another control factor, concluding that the random forest is the most robust, whereas Folleco, Khoshgoftaar, and Napolitano (2008) consider only the decision tree

classifier but compare the effect of different sampling techniques.

In another study, Folleco, Khoshgoftaar, and Bullard (2008) take the same approach as the previous two papers, but they insert exclusively attribute defects and no class defects into the data. Besides consistent results, they further show that classification performance drops rapidly once the five most significant attributes are corrupted.

Puri and Gupta (2019) also focus exclusively on attribute defects and use only the decision tree classifier while comparing different sampling techniques. They additionally group the examined datasets based on their imbalance ratio, but do not specify the characteristics of the attribute defects or their insertion procedure.

Overall, the impact of attribute defects on the classification of imbalanced data is only examined using different levels of random noise within the attributes. Nevertheless, DQ defects can be more versatile than random noise (Firmani et al., 2016). Indeed, DQ is distinguished according to several dimensions since DQ defects can have diverse causes resulting in defects with different characteristics and extents (Sidi et al., 2012). For example, they may have different properties such as varying degrees of inaccuracy. Consequently, the impact of DQ defects should be investigated according to DQ dimensions such as accuracy.

In addition, previous studies have not explicitly considered the importance of attributes with respect to the classification, since the KS statistic assesses the similarity of two distributions, but not a mutual dependence (Lopes et al., 2007). Thus, the relevance of the location of defects among the attributes should be studied individually and depending on the respective feature importance.

As a result of this research gap, this study examines in particular the impact of attribute defects according to the DQ dimension accuracy and also considers the feature importance of corrupted attributes in addition to the defect level.

### 3. Design of the Experiments

In this section, we present our experimental design by first explaining the methodology and introducing the parameters according to which we insert different attribute defects. Then, we present the selection of datasets and classifiers used to evaluate the impact of the inserted attribute defects.

#### 3.1. Experimental Design and DQ Defects

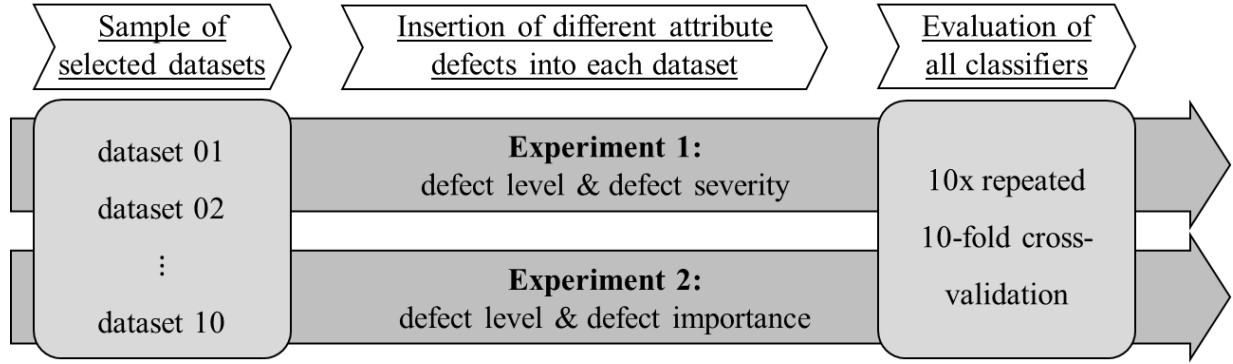
In this study, we conduct two experiments to evaluate the impact of different attribute defects on the classification of imbalanced data. In the first

experiment, we insert different degrees of inaccuracy to the attributes to investigate the impact of attribute defects according to the DQ dimension accuracy. We call this the defect severity of the attribute defects. In the second experiment, we examine the defect importance, i.e., the importance of each attribute for the classification measured by the feature importance, to evaluate how much this importance affects the impact of the defects. In addition, we vary the defect level, i.e., the amount of attribute defects for both experiments.

In order to obtain meaningful and significant results for both experiments, we use a sample of ten publicly available imbalanced datasets into which we insert the defects. We then use six classifiers to classify the imbalanced data for each of these datasets and evaluate the performance of each classifier using the average result across all datasets. To further increase the validity, we evaluate the classification of each dataset using a ten times repeated, stratified 10-fold cross-validation. In this process, each dataset is randomly divided into ten equal parts, so that both classes retain their initial distribution in each of these ten parts. Then, each of these parts is used once as a test set, while the other nine parts are used to train each model. After all ten possible combinations, this process is repeated ten times for different random partitions. As a result of this procedure, shown in Figure 1, we obtain the average classification performance over all datasets and the cross-validation of each classifier for the different inserted attribute defects.

The performance of all these classifications is measured by the F-measure (F1) of the positive class, which is the harmonic mean of recall and precision. The performance of this class is of particular importance when classifying imbalanced data. Due to the class imbalance, the positive class is much more difficult to predict while at the same time being usually of more interest (Sun et al., 2015; Yin et al., 2013). Therefore, the F1 constitutes a reasonable choice to compare the classification performance for imbalanced data.

For both experiments, we vary the defect level  $L$ , indicating the proportion of corrupted values. Thus, the number of inserted attribute defects for each dataset is determined by  $L \times |A| \times |N|$ , with the number of instances  $|N|$  and the number of attributes to be corrupted  $|A|$  in the respective dataset. In our experiments, the defect level  $L$  can take the values 5%, 10%, 15%, 20%, and 25% (cf. Puri & Gupta, 2019). In both experiments, we first consider these five levels of attribute defects, and once the defect level is set, both an instance and an attribute are randomly selected and corrupted until the specified number of attribute defects is reached. In this procedure, a certain value cannot be selected and corrupted multiple times.



**Figure 1. Illustration of the experimental design of our two experiments, where different attribute defects are inserted into imbalanced datasets, and then their impact on classifiers is evaluated with a cross-validation**

In the first experiment, we examine defects according to the DQ dimension accuracy for the previously determined number of defects. The corresponding parameter, defect severity  $S$ , controls the inaccuracy of the attribute values by varying the initial values to a certain degree rather than randomly changing them. The severity with which the values are corrupted depends on the estimated standard deviation of the respective attribute of the individual dataset, and the sign of the deviation is chosen randomly. For this purpose, the standard deviation of each attribute is estimated by  $\sigma = \frac{1}{5}(\max - \min)$  where  $\max$  and  $\min$  refer to the maximum and minimum values of the respective attribute (cf. Sáez et al., 2014). The defect severity  $S$  is varied in the steps  $0.25\sigma$ ,  $0.5\sigma$ ,  $1\sigma$ ,  $2\sigma$ , and  $3\sigma$ . Thus, as a result of defect level and defect severity, there are 25 combinations of attribute defects for each dataset in the first experiment.

In our second experiment, attribute defects are no longer inserted across all attributes, but only into specific attributes depending on their importance to the classification. Based on the feature importance of each attribute, we determine a ranking and assign the attributes to the respective quartile  $Q_i$  of the ranking. Thus, the parameter defect importance  $I$  results in four values  $Q_1$ ,  $Q_2$ ,  $Q_3$ , and  $Q_4$  where  $Q_1$  represents the most important 25% of attributes for each dataset and  $Q_4$  represents the least important 25% of attributes. Then, for a given defect importance, we insert random attribute defects into attributes within the corresponding quartile  $Q_i$  until the defect level is reached. Since the number of attributes to be corrupted  $|A|$  is equal to the number of attributes of the respective defect importance, the absolute number of attribute defects is reduced compared to the first experiment. As a result of defect level and defect importance, there are 20 combinations of attribute defects for each dataset in the second experiment.

To assess a reliable feature importance of each attribute, we use two common approaches, namely

decision tree feature importance as well as permutation feature importance, and two methods for each (Bolón-Canedo & Alonso-Betanzos, 2019). Tree-based algorithms provide information scores that indicate the reduction of the information criterion (e.g., Gini impurity) in all splits. Depending on this information score, a ranking can be determined which attributes are most important for the classification. For the feature importance approach based on permutation, the degradation of a classifier’s performance is measured after randomly rearranging the values of each single attribute, where a larger performance degradation means that the respective attribute is more important. For both approaches, we use the efficient and accurate algorithms random forest (RF) and extreme gradient boosting (XGB) (Dewi, Christine, and Rung-Ching Chen, 2019; Li et al., 2020) to rank the attributes of each dataset according to the respective feature importance. Finally, the attributes of each dataset are sort according to their average rank of all feature importance methods (Bolón-Canedo & Alonso-Betanzos, 2019).

In total, we apply the six selected classifiers to all ten imbalanced datasets and use a ten times repeated, stratified 10-fold cross-validation to receive reliable results in both experiments. In addition to each benchmark performance with no attribute defects inserted, there are 25 and 20 combinations of parameter values in the respective experiments resulting in 276,000 classifiers built in the course of this work.

### 3.2. Datasets and Classifiers

For our experiments, we use ten publicly available datasets from different domains and with different characteristics, as summarized in Table 1. Thus, they provide a broad sample for a valid evaluation. The number of instances in the datasets varies from 1,687 to 382,154 while the number of attributes ranges from 3 to 21. All datasets have a binary target class, i.e., each can be distinguished into a single majority (negative) class and a single minority (positive) class. The ratio of

negative instances to positive instances, the imbalance ratio (IR), ranges from 3.15 to 58.56.

There are three datasets from the financial industry. The *CreditApproval* dataset<sup>1</sup> contains demographic and economic information about individuals applying for credit. The label indicates whether the corresponding credit was approved. Similarly, the goal of classifying the *LoanRepay* dataset<sup>2</sup> is to predict whether a loan will be default, with the loan being repaid in most cases. The same is true for the *LoanDefault* dataset<sup>3</sup>, which is synthetically created based on data from a financial institution.

The next two datasets refer to customer relationship management (CRM), each containing information about customers and respective services. For the *MarketingCampaign* dataset<sup>4</sup>, the goal is to predict whether a telephone marketing campaign will be successful and whether the potential customer will use the service offered. At the other end of the customer journey is the *CustomerChurn* dataset<sup>5</sup>. Here, the goal is to predict whether a customer will cancel a contract.

Both medical datasets are highly imbalanced and designed to predict a disease based on personal and medical data. The *StrokePrediction* dataset<sup>6</sup> contains mainly personal data, while the *CancerPrediction* dataset<sup>7</sup> contains metrics from radiological scans.

For the *PredMaintenance* dataset<sup>8</sup>, synthetic data was derived from real-world predictive maintenance

data to predict whether machines should be maintained ahead of schedule. The *SoftwareMeasures* dataset<sup>9</sup> comes from the NASA KC1 software project and contains software quality metrics to predict software defects. In the classification of the *OnTimeGraduation* dataset<sup>10</sup>, the student's grade point average over four semesters is used to determine whether the student is graduating in standard time.

For each of these datasets, the impact of DQ defects is examined using six commonly used classifiers. These are the learning algorithms Naïve Bayes (NB), k-nearest neighbors (kNN), logistic regression (LR), decision tree (DT), random forest (RF), and support vector machine (SVM). Their respective, well-established parameters are adjusted to a certain degree in order to obtain a suitable baseline for the evaluation. During the experiments, these parameters remain unchanged.

For the NB algorithm, the Gaussian kernel was used and the prior probabilities were specified with the respective class ratio. For the kNN classifier, the three closest instances weighted by their distance are used for prediction. The LR is performed using the L1 norm as the penalty term (lasso regression). In addition, during the optimization, the instances are weighted in a balanced way based on their class, so that positive instances receive higher weights than negative instances. For both DT and RF, we use the Gini impurity as an information criterion to identify the best splits

**Table 1. Characteristics of the selected sample of imbalanced datasets to assess the impact of attribute defects on classification**

| Name of Dataset   | Domain      | Number of Instances | Number of Attributes | Number of Positive Instances | Imbalance Ratio |
|-------------------|-------------|---------------------|----------------------|------------------------------|-----------------|
| CreditApproval    | Banking     | 4,521               | 8                    | 521 (11.52%)                 | 7.68            |
| LoanRepay         | Banking     | 9,578               | 13                   | 1,533 (16.01%)               | 5.25            |
| LoanDefault       | Banking     | 10,000              | 3                    | 333 (3.33%)                  | 29.03           |
| MarketingCampaign | CRM         | 45,211              | 16                   | 5,289 (11.70%)               | 7.55            |
| CustomerChurn     | CRM         | 7,043               | 19                   | 1,869 (26.54%)               | 2.77            |
| StrokePrediction  | Medicine    | 43,400              | 10                   | 783 (1.80%)                  | 54.42           |
| CancerPrediction  | Medicine    | 11,183              | 6                    | 260 (2.32%)                  | 42.01           |
| PredMaintenance   | Engineering | 10,000              | 6                    | 339 (3.39%)                  | 28.50           |
| SoftwareMeasures  | Engineering | 2,096               | 21                   | 325 (15.51%)                 | 5.45            |
| OnTimeGraduation  | Education   | 1,687               | 4                    | 135 (8.00%)                  | 11.50           |

<sup>1</sup><https://www.kaggle.com/datasets/enesztrk/credit-approval>

<sup>2</sup><https://www.kaggle.com/datasets/swetashetye/lending-club-loan-data-imbalance-dataset>

<sup>3</sup><https://www.kaggle.com/datasets/kmlldas/loan-default-prediction>

<sup>4</sup><https://www.kaggle.com/datasets/hariharanpavan/bank-marketing-dataset-analysis-classification>

<sup>5</sup><https://www.kaggle.com/datasets/soheiltehranipour/it-customer-churn>

<sup>6</sup><https://www.kaggle.com/datasets/shashwatwork/cerebral-stroke-predictionimbalanced-dataset>

<sup>7</sup><https://www.kaggle.com/datasets/sudhanshu2198/microcalcification-classification>

<sup>8</sup><https://archive.ics.uci.edu/ml/datasets/AI4I+2020+Predictive+Maintenance+Dataset>

<sup>9</sup><https://www.openml.org/search?type=data&sort=runs&id=1067&status=active>

<sup>10</sup><https://www.kaggle.com/datasets/oddyvirgantara/on-time-graduation-classification>

within the trees. The SVM algorithm is instantiated using the radial basis function kernel to transform the data into an infinite dimensional feature space.

## 4. Impact of Attribute Defects in Imbalanced Data

This section contains the results of the two experiments presented in the previous section. First, we investigate the impact of attribute defects according to the dimension accuracy. Second, defects are selectively inserted into attributes of different importance with respect to the classification. In addition, the defect level is varied for both experiments.

### 4.1. Experimental Results – Defect Severity

The randomly selected and corrupted values with five specified defect levels and five different defect severities result in 25 F1 values for each classifier, averaged over all ten datasets. We examine the course of the F1 for each defect severity as the defect level increases, which is shown in Figure 2 for all six classifiers separately. Consistent with the literature (e.g., Seiffert et al., 2014), the performance of all six classifiers decreases with increasing defect level, as indicated by the negative slopes of the plotted lines. In particular, as the defect severity increases, the F1 for all classifiers deteriorates considerably, i.e., the less accurate the imbalanced data, the more the F1 decreases. Even for a defect level of only 5%, the classifiers can predict the positive instances significantly worse as the defect severity increases. Moreover, for all classifiers,

few but severe defects, i.e.,  $L = 5\%$  and  $S = 3\sigma$ , have a clearly larger negative impact on the prediction of the positive class than many but minor defects, i.e.,  $L = 25\%$  and  $S = 0.25\sigma$ . This implies for the prediction of the positive class that the defect severity, i.e., the inaccuracy of the data, has more serious consequences for our datasets than the defect level, i.e. the amount of inaccuracies. Overall, the results show that DQ is a fundamental requirement when classifying imbalanced data since classification performance is highly affected by inaccurate data. This is particularly true for large deviations from the original values, i.e., for a high defect severity. All selected classifiers are affected by the poor DQ and are clearly worse at detecting positive instances.

However, there are some differences between the classifiers in terms of the impact of inaccurate data. Both the absolute F1 obtained and the slopes of the curves differ among the six classifiers. Both tree-based classifiers, DT and RF, achieve the best F1 for all combinations of defect level and defect severity. On average, the DT is even able to outperform the RF. Furthermore, both tree-based algorithms are most robust to the progression of F1 as both defect parameters increase, which is consistent with the literature (Shwartz-Ziv & Armon, 2022). Table 2 shows the percentage change of F1 compared to the benchmark for all classifiers with  $S = 0.5\sigma$ . Here, the F1 for both tree-based algorithms is the least affected compared to the initial F1. Nevertheless, the F1 of the LR classifier deteriorates proportionally least for high defect levels, e.g.  $L = 25\%$ . Thus, for many attribute defects, the LR gradually catches up in classifying the positive class of imbalanced data (cf. Kirasich et al., 2018).

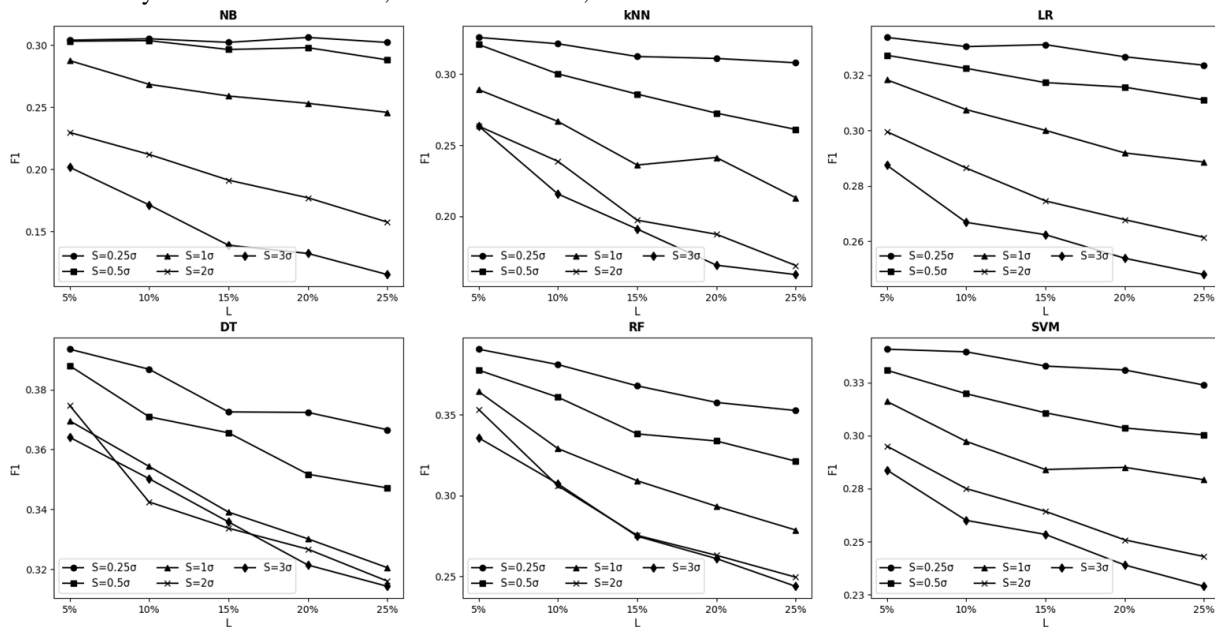


Figure 2. Variation of the F1 for each classifier as the defect level increases for different defect severities

**Table 2. Percentage change of F1 for  $S = 0.5\sigma$  compared to F1 without attribute defects**

| $S=0.5\sigma$ | $L=5\%$ | $L=10\%$ | $L=15\%$ | $L=20\%$ | $L=25\%$ |
|---------------|---------|----------|----------|----------|----------|
| <b>NB</b>     | -8.35   | -8.21    | -10.37   | -9.92    | -12.89   |
| <b>kNN</b>    | 0.07    | -6.35    | -10.37   | -14.98   | -18.55   |
| <b>LR</b>     | -1.33   | -2.75    | -4.30    | -4.80    | -6.19    |
| <b>DT</b>     | 2.66    | -1.86    | -3.27    | -6.94    | -8.15    |
| <b>RF</b>     | 3.41    | -1.17    | -7.40    | -8.60    | -12.01   |
| <b>SVM</b>    | -1.06   | -4.36    | -7.06    | -9.20    | -10.16   |

## 4.2. Experimental Results – Defect Importance

In the second experiment, there are 20 F1 values for each classifier, resulting from five defect levels and four defect importance values. The course of the F1 for each classifier is presented in Figure 3. For all classifiers, it is noticeable that corrupted attributes in each of  $Q_2$ ,  $Q_3$ , and  $Q_4$  have a distinct smaller impact on the classification performance than attributes in  $Q_1$ , i.e., the most important attributes. Thus, defects in the 25% most important attributes in terms of classification are crucial for predicting the positive class, while the 75% of attributes that are less important for classification have a relatively small impact on predicting the positive class. Even for a small defect level of  $L=5\%$ , the F1 drops significantly as the defect importance changes from  $I = Q_2$  to  $I = Q_1$ . This observation holds for all classifiers considered. Therefore, the F1 does not deteriorate in equal steps as the defect importance increases. Instead, the F1, and thus predicting the positive class, is considerably worse when attribute defects are present in the first quartile of importance compared to the other quartiles. This finding is especially true for higher defect levels, as the lines resulting from  $I = Q_1$  deviate even further from the other lines as the defect level increases. For the defect importances  $I = Q_2$  to  $I = Q_4$ , both the absolute F1 and its progression with increasing defect level behave quite similarly for each classifier, with only minor differences. However, there is also a similarity among all defect importances. For any given defect importance, the F1 decreases fairly linearly as the defect level increases for all classifiers.

Although this is true for all six classifiers, there is a striking difference between the classifiers. For the defect importances  $Q_2$ ,  $Q_3$ , and  $Q_4$ , the classification performance of the classifiers NB, LR, DT, and RT does almost not degrade with increasing defect level, but rather remains constant. The slopes of the lines resulting from  $Q_2$ ,  $Q_3$ , and  $Q_4$  are close to zero, indicating that a larger amount of attribute defects does not significantly further affect the ability to predict the positive class. In contrast, the F1 for the distance-based algorithms kNN and SVM are much more affected by attribute defects in  $Q_2$ ,  $Q_3$ , and  $Q_4$  as the defect level increases. The respective resulting curves show a negative slope for

both classifiers, meaning that the detection of positive instances is negatively affected as more defects occur in relatively unimportant attributes. However, for non-distance-based algorithms, defects in the second, third, and fourth quartiles of important attributes are relatively insignificant for the classification performance of the positive class.

## 5. Discussion and Conclusion

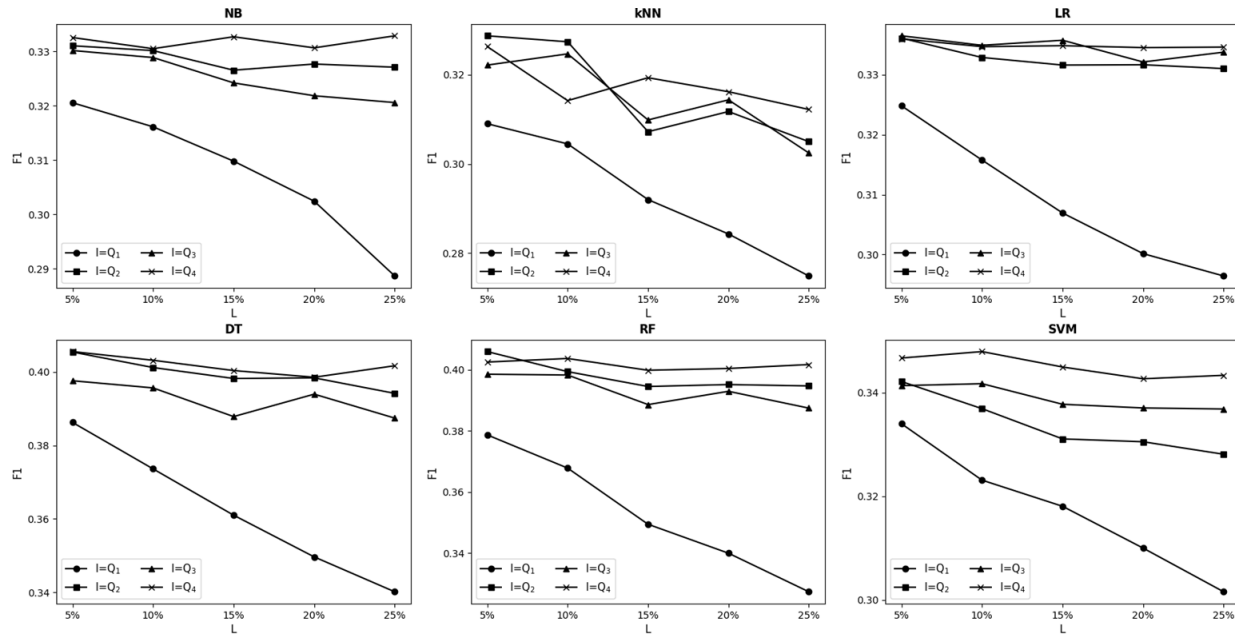
In this section, we discuss the previously presented results as well as limitations of our work, point out future research, and conclude with our contributions.

### 5.1. Implications for Theory and Practice

In this paper, we investigated the impact of attributes with poor quality on the classification performance of imbalanced data using six classification algorithms on ten different datasets. Building on the related literature, in addition to different defect levels, we specifically investigated DQ defects according to the dimension accuracy, i.e., different magnitudes of inaccuracies in the attribute values, and defects within attributes of different importance with respect to the classification. The experiments enrich the research on the impact of DQ defects and the findings provide important implications for both science and practice.

Overall, the results of our experiments highlight the importance of measuring and improving DQ in order to ensure reliable AI and ML algorithms, provide trustworthy data-driven decision support, and ultimately prevent biased decision making. The significant impact of inaccuracies on classifying imbalanced data for all classifiers studied emphasizes the relevance of data accuracy for AI and ML algorithms. Especially large deviations should be cleaned. Among the six classifiers, DT and RF are best at predicting positive instances, as attribute defects are present in imbalanced data. Thus, for tabular data, the tree-based algorithms are best suited to classify instances (cf. Shwartz-Ziv & Armon, 2022), even when DQ defects are present in imbalanced data. Therefore, for inaccurate and imbalanced data, we recommend using DT or RF when the minority class is the primary concern (cf. Folleco, Khoshgoftaar, van Hulse, & Bullard, 2008).

However, even though DT and RF perform best among the classifiers studied in terms of absolute F1 and across all parameter combinations, their performance is still strongly affected by the level and severity of the attribute defects. Hence, their results should also be carefully examined. Especially, when there are many attribute defects with large inaccuracies in imbalanced data, practitioners should carefully consider whether an



**Figure 3. Variation of the F1 for each classifier as the defect level increases for each defect importance**

algorithm such as LR is more appropriate, since the F1 will not deteriorate as much in percentage terms.

In the second experiment, it was shown that the feature importance of attributes affected by DQ defects is a major factor for classification performance. While affected attributes in the second, third, and fourth most important quartile are almost equally insignificant for classification performance, the defects within attributes of the most important quartile cause large losses in the F1. Thus, the DQ of the most important 25% of attributes is crucial for the classification of the minority class, and consequently, efficient data cleaning can be performed by focusing on the most important attributes.

Furthermore, distance-based algorithms such as kNN and SVM tend to be more affected by poor DQ in less important features than other classifiers. Even seemingly unimportant attributes affect the distances between all data points, leading to a biased prediction by distance-based methods. In contrast, the LR with the l1 norm, for example, sets coefficients of less important attributes close to 0, thus reducing their biasing effect. Therefore, distance-based algorithms should be avoided if high-quality data cannot be ensured for all attributes. We also recommend that, for efficient manners, in particular the most important attributes should be cleaned before performing any analysis – although ideally all attributes are cleaned.

## 5.2. Limitations and Future Research

Despite our contributions to academia and practice on the impact of DQ defects on the performance of AI and ML algorithms, our work has limitations that can

serve as a starting point for further research. First, we focus our evaluation on the F1 as this metric is well suited to assess the classification of the positive class, i.e., the class of interest. Nevertheless, further evaluation measures such as the area under the receiver operator curve could be determined and compared in continuing studies. Second, our work addresses the research gap by investigating the impact of DQ defects according to the particular DQ dimension accuracy. However, there are more DQ dimensions such as completeness, consistency, and currency, which exhibit other characteristics and thus may lead to other control parameters. Consequently, future work could extend the study of the impact of DQ defects on classifying imbalanced data to other DQ dimensions and even combinations thereof. Finally, in addition to a broader investigation of the impact of DQ defects on the classification of imbalanced data, the classifiers could be examined in combination with techniques such as sampling, boosting and bagging (e.g., Khoshgoftaar et al., 2011). These techniques can positively affect the classification performance for imbalanced data and should be studied in the context of DQ defects as well.

## 5.3. Conclusion

Real-world data of poor quality leads to poor results from AI and ML algorithms (Janssen et al., 2020; Mans et al., 2015). Especially in the already challenging case of classifying imbalanced data, it is therefore crucial to know the consequences of certain DQ defects. Thus, this work extends the current research by examining the impact of specific DQ defects. First, we contribute to

theory and practice by examining in particular the impact of inaccuracies with varying magnitude, i.e., defects according to the DQ dimension accuracy. Second, we extend the literature by investigating the relevance of feature importance with respect to the classification by studying the effects of corrupted attributes with different importance. Our findings extend the understanding of the impact of DQ defects on the classification of imbalanced data and clearly emphasize that the outcome of AI and ML algorithms must be assessed together with the quality of the underlying data. Underpinning the importance of DQ, we recommend comprehensively measuring and improving DQ according to its dimensions.

## 6. References

- Bolón-Canedo, V., & Alonso-Betanzos, A. (2019). Ensembles for Feature Selection: A Review and Future Trends. *Information Fusion*, 52, 1–12.
- Chengalur-Smith, I. N., Ballou, D. P., & Pazer, H. L. (1999). The Impact of Data Quality Information on Decision Making: An Exploratory Analysis. *IEEE Transactions on Knowledge and Data Engineering*, 11(6), 853–864.
- Dewi, Christine, and Rung-Ching Chen (2019). Random Forest and Support Vector Machine on Features Selection for Regression Analysis. *International Journal of Innovative Computing, Information and Control*, 15(6), 2027–2037.
- Feldman, M., Even, A., & Parmet, Y. (2018). A Methodology for Quantifying the Effect of Missing Data on Decision Quality in Classification Problems. *Communications in Statistics - Theory and Methods*, 47(11), 2643–2663.
- Fernández, A., García, S., Galar, M., Prati, R. C., Krawczyk, B., & Herrera, F. (2018). *Learning from Imbalanced Data Sets*. Springer International Publishing.
- Firmani, D., Mecella, M., Scannapieco, M., & Batini, C. (2016). On the Meaningfulness of “Big Data Quality”. *Data Science and Engineering*, 1(1), 6–20.
- Folleco, A., Khoshgoftaar, T., & Bullard, L. (2008). Analyzing the Impact of Attribute Noise on Software Quality Classification. In *International Conference on Software Engineering & Knowledge Engineering*, San Francisco, USA.
- Folleco, A., Khoshgoftaar, T. M., & Napolitano, A. (2008). Comparison of Four Performance Metrics for Evaluating Sampling Techniques for Low Quality Class-Imbalanced Data. In *International Conference on Machine Learning and Applications (ICMLA)*, San Diego, USA.
- Folleco, A., Khoshgoftaar, T. M., van Hulse, J., & Bullard, L. (2008). Identifying Learners Robust to Low Quality Data. In *International Conference on Information Reuse and Integration (IRI)*, Las Vegas, USA.
- García, S., Luengo, J., & Herrera, F. (2015). *Data Preprocessing in Data Mining*. Springer International Publishing.
- Ghasemaghaei, M., & Calic, G. (2019). Can Big Data Improve Firm Decision Quality? The Role of Data Quality and Data Diagnosticity. *Decision Support Systems*, 120, 38–49.
- Gupta, S., & Gupta, A. (2019). Dealing with Noise Problem in Machine Learning Data-sets: A Systematic Review. *Procedia Computer Science*, 161, 466–474.
- Hagiu, A., & Wright, J. (2020). When Data Creates Competitive Advantage. *Harvard Business Review*, 98, 94–101.
- Hariri, R. H., Fredericks, E. M., & Bowers, K. M. (2019). Uncertainty in Big Data Analytics: Survey, Opportunities, and Challenges. *Journal of Big Data*, 6(1), Article 44, 1–16.
- Hazen, B. T., Boone, C. A., Ezell, J. D., & Jones-Farmer, L. A. (2014). Data Quality for Data Science, Predictive Analytics, and Big Data in Supply Chain Management: An introduction to the Problem and Suggestions for Research and Applications. *International Journal of Production Economics*, 154, 72–80.
- He, H., & Garcia, E. A. (2009). Learning from Imbalanced Data. *IEEE Transactions on Knowledge and Data Engineering*, 21(9), 1263–1284.
- Heinrich, B., & Helfert, M. (2003). Analyzing Data Quality Investments in CRM - A Model-based Approach. In *International Conference on Information Quality (ICIQ)*, Cambridge, MA, USA.
- Heinrich, B., Hopf, M., Lohninger, D., Schiller, A., & Szubartowicz, M. (2021). Data Quality in Recommender Systems: The Impact of Completeness of Item Content Data on Prediction Accuracy of Recommender Systems. *Electron Markets*, 31(2), 389–409.
- Heinrich, B., Hristova, D., Klier, M., Schiller, A., & Szubartowicz, M. (2018). Requirements for Data Quality Metrics. *Journal of Data and Information Quality (JDIQ)*, 9(2), Article 12, 1–32.
- Janssen, M., Brous, P., Estevez, E., Barbosa, L. S., & Janowski, T. (2020). Data Governance: Organizing Data for Trustworthy Artificial Intelligence. *Government Information Quarterly*, 37(3), Article 101493.
- Janssen, M., van der Voort, H., & Wahyudi, A. (2017). Factors Influencing Big Data Decision-Making Quality. *Journal of Business Research*, 70, 338–345.
- Johnson, J. M., & Khoshgoftaar, T. M. (2022). A Survey on Classifying Big Data with Label Noise. *ACM Journal of Data and Information Quality*, 14(4), Article 23, 1–43.
- Jylhä, T., & Suvanto, M. E. (2015). Impacts of Poor Quality of Information in the Facility Management Field. *Facilities*, 33(5/6), 302–319.
- Jylhä, V., Bates, D. W., & Saranto, K. (2016). Adverse Events and Near Misses Relating to Information

- Management in a Hospital. *Journal of the Health Information Management*, 45(2), 55–63.
- Kennedy, R. K. L., Johnson, J. M., & Khoshgoftaar, T. M. (2021). The Effects of Class Label Noise on Highly-Imbalanced Big Data. In *International Conference on Tools for Artificial Intelligence (ICTAI)*, Washington, DC, USA.
- Khoshgoftaar, T. M., van Hulse, J., & Napolitano, A. (2011). Comparing Boosting and Bagging Techniques With Noisy and Imbalanced Data. *IEEE Transactions on Systems, Man, and Cybernetics - Part a: Systems and Humans*, 41(3), 552–568.
- Kirasich, K., Smith, T., & Sadler, B. (2018). Random Forest vs Logistic Regression: Binary Classification for Heterogeneous Datasets. *SMU Data Science Review*, 1(3), Article 9.
- Kruse, S., Papotti, P., & Naumann, F. (2015). Estimating Data Integration and Cleaning Effort. In *International Conference on Extending Database Technology, Proceedings*, Brussels, Belgium.
- Lee, Y. W., Strong, D. M., Kahn, B. K., & Wang, R. Y. (2002). AIMQ: A Methodology for Information Quality Assessment. *Information & Management*, 40(2), 133–146.
- Li, H., Cao, Y., Li, S., Zhao, J., & Sun, Y. (2020). XGBoost Model and Its Application to Personal Credit Evaluation. *IEEE Intelligent Systems*, 35(3), 52–61.
- Lopes, R. H., Reid, I. D., & Hobson, P. R. (2007). *The Two-Dimensional Kolmogorov-Smirnov Test*. Proceedings of Science.
- Mans, R. S., Vanwersch, R. J. B., & van der Aalst, W. (2015). *Process Mining in Healthcare: Evaluating and Exploiting Operational Healthcare Processes*. Springer International Publishing..
- Ngai, E. W. T., Gunasekaran, A., Wamba, S. F., Akter, S., & Dubey, R. (2017). Big Data Analytics in Electronic Markets. *Electron Markets*, 27, 243–245.
- Picado, J., Davis, J., Termehchy, A., & Lee, G. Y. (2020). Learning Over Dirty Data Without Cleaning. In *ACM SIGMOD International Conference on Management of Data*, Portland, USA.
- Puri, A., & Gupta, M. K. (2019). Comparative Analysis of Resampling Techniques under Noisy Imbalanced Datasets. In *International Conference on Issues and Challenges in Intelligent Computing Techniques (ICICT)*, Ghaziabad, India.
- Raguseo, E. (2018). Big Data Technologies: An Empirical Investigation on their Adoption, Benefits and Risks for Companies. *International Journal of Information Management*, 38(1), 187–195.
- Redman, T. C. (1997). *Data Quality for the Information Age*. Arctech House.
- Redman, T. C. (1998). The Impact of Poor Data Quality on the Typical Enterprise. *Communications of the ACM*, 41(2), 79–82.
- Sáez, J. A., Galar, M., Luengo, J., & Herrera, F. (2014). Analyzing the Presence of Noise in Multi-Class Problems: Alleviating its Influence with the One-vs-One Decomposition. *Knowledge and Information Systems*, 38, 179–206.
- Sandeep, S. R., Ahamad, S., Saxena, D., Srivastava, K., Jaiswal, S., & Bora, A. (2022). To Understand the Relationship Between Machine Learning and Artificial Intelligence in Large and Diversified Business Organisations. *Materials Today: Proceedings*, 56(4), 2082–2086.
- Seiffert, C., Khoshgoftaar, T. M., van Hulse, J., & Folleco, A. (2014). An Empirical Study of the Classification Performance of Learners on Imbalanced and Noisy Software Quality Data. *Information Sciences*, 259, 571–595.
- Serhani, M. A., El Kassabi, H. T., Taleb, I., & Nujum, A. (2016). An Hybrid Approach to Quality Evaluation across Big Data Value Chain. In *International Congress on Big Data*, San Francisco, USA.
- Shwartz-Ziv, R., & Armon, A. (2022). Tabular Data: Deep Learning is not All you Need. *Information Fusion*, 81, 84–90.
- Sidi, F., Shariat Panahy, P. H., Affendey, L. S., Jabar, M. A., Ibrahim, H., & Mustapha, A. (2012). Data Quality: A Survey of Data Quality Dimensions. In *International Conference on Information Retrieval & Knowledge Management*, Kuala Lumpur, Malaysia.
- Stefanowski, J. (2016). Dealing with Data Difficulty Factors While Learning from Imbalanced Data. In *Studies in Computational Intelligence. Challenges in Computational Statistics and Data Mining* (1<sup>st</sup> ed., Vol. 605, pp. 333–363). Springer.
- Sun, Z., Song, Q., Zhu, X., Sun, H., Xu, B., & Zhou, Y. (2015). A Novel Ensemble Method for Classifying Imbalanced Data. *Pattern Recognition*, 48(5), 1623–1637.
- van Hulse, J., & Khoshgoftaar, T. (2009). Knowledge Discovery from Imbalanced and Noisy Data. *Data & Knowledge Engineering*, 68(12), 1513–1542.
- Vassakis, K., Petrakis, E., & Kopanakis, I. (2018). Big Data Analytics: Applications, Prospects and Challenges. In *Lecture Notes on Data Engineering and Communications Technologies: Vol. 10. Mobile Big Data - A Roadmap from Models to Technologies* (Vol. 10, pp. 3–20). Springer.
- Wang, D., & Xu, Z. (2019). Impact of Inaccurate Data on Differential Privacy. *Computers & Security*, 82, 68–79.
- Webster, J., & Watson, R. T. (2002). Analyzing the Past to Prepare for the Future: Writing a Literature Review. *MIS Quarterly*, 26(2), XIII–XXIII.
- Weiss, G. M., & Provost, F. (2003). Learning When Training Data are Costly: The Effect of Class Distribution on Tree Induction. *Journal of Artificial Intelligence Research*, 19, 315–354.
- Yin, L., Ge, Y., Xiao, K., Wang, X., & Quan, X. (2013). Feature Selection for High-Dimensional Imbalanced Data. *Neurocomputing*, 105, 3–11.
- Zhu, X & Wu, X. (2004). Class Noise vs. Attribute Noise: A Quantitative Study. *Artificial Intelligence Review*, 22(3), 177–210.