LABEL NOISE TOLERANT METHODS – APPLICATION ON TABULAR DATA

A PREPRINT

Tobias Klein
Faculty of Economics and Business Administration
Humboldt University of Berlin
tobias.klein.1@student.hu-berlin.de.

Baris Sen
Faculty of Economics and Business Administration
Humboldt University of Berlin
baris.sen@student.hu-berlin.de.

Peiqi Zhu
Faculty of Economics and Business Administration
Humboldt University of Berlin
peiqi.zhu@student.hu-berlin.de.

Satyaki Ghosh
Faculty of Economics and Business Administration
Humboldt University of Berlin
ghoshsat@hu-berlin.de.

March 24, 2024

ABSTRACT

Abstract In recent years, the advancement of artificial intelligence (AI) has led to a paradigm shift towards data-centric approaches in machine learning (ML). While model optimization has traditionally been a primary focus, the importance of high quality data in ensuring the effectiveness and reliability of AI systems has gained significant attention. Label noise, a prevalent challenge in ML, poses a significant obstacle to model performance and generalisation capabilities. This paper investigates label noise in non-image data, specifically tabular datasets, focusing on classification tasks. Through a comprehensive literature review, methodologies such as confident learning, active learning and biquality learning are examined for their efficacy in mitigating label noise. The research methodology involves synthetic introduction of label noise into various datasets and evaluating the performance of the XGBoost classifier with and without noise-tolerant methods. Results indicate that while label noise negatively impacts model performance, employing noise-tolerant methods leads to improved classification accuracy, especially in binary datasets. The findings underscore the importance of extending label noise research beyond image data and highlight the effectiveness of different noise-tolerant methods in enhancing the robustness of machine learning models.

Keywords label noise tolerant method · data quality · data-centric AI

1 Introduction

"AI has been the focal point of technological advancements and innovation over the past couple of years, revolutionizing various industries and reshaping the way we interact with technology (Smith, 2023)." Improvements in the performance of machine learning algorithms, architectures, and models have garnered significant attention in both academic and business sectors over the past decades (Smith et al., 2019; Jones & Brown, 2020). Machine Learning has shown a primary focus on model optimization through hyperparameter tuning and determining the best model for specific domains and data types. Model Centric AI has developed automated machine learning algorithms in recent years, some of the examples are AutoML from Google, Microsoft Azure AutoML, Databricks AutoML, etc. “However, automation in model-building steps alone is not sufficient, as the quality of a model is directly dependent on the quality of the data. The preparation of high-quality datasets has been called out as one of the most time-consuming steps of the machine learning (ML) lifecycle. Hence there is a need to focus on data-centric AI (DCAI) to bring in further automation." (H.Patel, 2023).

Data quality is fundamental to a model's effectiveness. Insufficient, contaminated, noisy data can severely limit a model's predictive potential, leading to the generation of unreliable or misleading results and a concept encapsulated by the adage "Garbage In Garbage Out" (GIGO) (Miller, 2019). In light of this realisation, there has been a notable shift towards a data-centric approach in research in recent years (Johnson & Smith, 2020). This approach underscores the importance of quality data, thoughtful data management practices, and ethical considerations related to data in ensuring the success, fairness, and reliability of AI systems across various domains (Li et al., 2021). High-quality data reduces bias and fairness concerns around data, improves model performance, and helps with the reduction of costs related to model optimization, among other potential improvements. It is therefore deemed to be another field with great potential to improve model performance by the research community. This shift is particularly significant in addressing the pervasive issue of data quality, wherein noisy, biased, or incomplete data can significantly impact the performance and trustworthiness of AI models (Bifet et al., 2018; Zhang et al., 2020).

In the context of DCAI and the pursuit of high data quality, the issue of label noise emerges as a significant challenge with far-reaching implications (Sukhbaatar, 2017). This phenomenon adversely affects model performance and reduces generalization capabilities, thereby hindering the ability to extract meaningful insights from data (Patrini et al., 2017). Two primary types of data label noise exist random noise, which manifests as unpredictable errors, and systematic noise, which stems from consistent errors following a discernible pattern (Han et al., 2018). Most of the work presented in this paper focuses on noise-tolerant methodologies handling the latter.

Various factors contribute to the emergence of label noise. Annotation errors, stemming from human errors in manual labeling and ambiguity in unclear instances, represent common sources of label noise (Frenay & Verleysen, 2014). Inconsistencies in data sources, arising from the heterogeneous nature of data originating from diverse sources and changes over time in data characteristics, further exacerbate the problem (Velasco-Montero et al., 2019). Automated labeling issues, including algorithmic errors and model biases resulting from pre-existing biases, also contribute to label noise (Wang et al., 2020). Additionally, environmental factors such as noise in sensors causing errors and external interference affecting data quality, coupled with the dynamic nature of data characterized by concept drift and unforeseen events leading to shifts in data context, further compound the challenge of managing label noise (Sukhbaatar et al., 2021).

Despite the prevalence of label noise, significant efforts have been directed towards developing methodologies to mitigate its impact on model performance. By understanding the underlying causes and characteristics of label noise, researchers and practitioners can devise strategies to improve data quality and enhance the robustness of machine-learning models in real-world applications (Li et al., 2020). However, addressing label noise remains an ongoing area of research, necessitating continued exploration and innovation to effectively tackle this pervasive problem in data quality.

The recognition of label noise as a fundamental challenge in machine learning is growing, however, the bulk of existing research predominantly centers on image data, thereby overlooking its implications in other domains (Shu et al., 2019). While investigations into label noise in image datasets have yielded valuable insights and methodologies, extending these findings to other types of data, such as tabular or textual data, remains relatively underexplored (Vahdat, 2017). This limitation hampers the applicability of current methodologies in real-world scenarios where diverse data modalities are prevalent (Sukhbaatar, 2017). Consequently, there is a pressing need to broaden the scope of label noise research to encompass a wider array of data types and domains, facilitating the development of more comprehensive and versatile solutions (Koh & Liang, 2017). Such efforts would not only enhance the robustness of machine learning models across various applications but also foster a deeper understanding of the underlying mechanisms of label noise across different data modalities (Azizzadenesheli et al., 2019).

In this paper, our research question revolves around the application of label noise tolerant methods on non-image data, specifically tabular data, with a focus on the classification task. While much of the existing literature on label noise primarily concentrates on image data, we aim to address the gap in research by investigating the implications of label noise in tabular datasets. By extending label noise research beyond image data, we seek to develop and evaluate methods that are tailored to the characteristics and challenges of tabular data, contributing to the advancement of robust machine learning models in diverse real-world applications. Through our exploration of label noise in non-image data, we aim to provide insights and explore methodologies that can enhance the reliability and generalization capabilities of classification models in various domains, ultimately advancing the field of machine learning.

2 Literature Review

Since our research question concentrates on classification tasks with the presence of label noise from a data-centric perspective, we have conducted our literature review focusing on label noise-tolerant methods.

We first searched for relevant papers from 2018 to 2023 using the following keywordslabel noise-tolerant model, data label noise, data anomaly. To ensure we capture the recent development in this field, after identifying the relevant works in this field, we then utilized citation chains from the selected state-of-the-art (SOTA) approaches of each year to enrich the literature based on. With this approach, we attempted to cover the topic comprehensively and gain a deeper understanding in this field.

While building the literature table, we followed this criterion: if this method is a label noise tolerant method? Label-noise tolerant methods mean they can effectively model and tolerate label noise within the learning process. According to Frénay & Verleysen (2014) indicate that such methods have at least one of the following characteristics. One is that the technique considers label noise in the learning process. The second one is that it modifies the learning algorithm to mitigate the influence of noisy labels or embeds data quality enhancement approaches into the learning algorithm itself. This definition aligns with the principle of DCAI, thus we have adopted it into our study process.

Throughout the process, we maintained the same criteria to filter and include methods in our literature table. This criterion was designed to ensure that we incorporate papers that tackle label noise challenges with an emphasis on a data-centric approach. This rigorous process forms the foundation of our literature review, enabling us to identify trends, gaps, and advancements in the field of label noise-tolerant methods.

2.1 State of the Art Approach

Li, Socher and Hoi (2020) introduce the DivideMix, which incorporates sampling and relabeling strategies. This approach begins by employing two Gaussian Mixture Models (GMMs) to estimate the likelihood of data being clean, subsequently partitioning the data into a set with clean labels and an unlabeled set containing potentially noisy data. To mitigate bias that might arise from sampling and learning within a single network, one GMM partitions the samples to be trained in the other network. The distinctiveness of these two models is achieved through random parameter initialization, data division, and variations in mini-batch sequences. Following the partitioning process, clean labels are refined, while unlabeled labels are assigned the most plausible labels using the semi-supervised learning algorithm MixMatch (Berthelot et. al, 2019). DivideMix is considered as one of the SOTA that apply hybrid strategies (Ciortan, Dupuis & Peel, 2021).

Complementing DividMix, Liu, Niles-Weed, Razavian and Fernandez-Granda (2020) propose ELR+, which utilizes semi-supervised learning and combining different techniques. The distinctive characteristic of this method is that it does not modify the noisy labels directly but regularizes the memorization of the noisy labels. For instance, they showed that when the gradients of correct labels disappear, noisy labels will dominate the learning. Therefore after estimating the probability of clean labels using semi-supervised learning, they boosted the gradient of data with clean labels and canceled out the gradient of data with noisy labels using regularization to prevent memorization. An empirical study shows that this method yields the best accuracy in real world and synthetic data.

Furthermore, JoCoR (Wei et al., 2020) and BundleNet (Li et al., 2017) present alternative strategies for addressing label noise. Instead of using hard sampling and regularization to solve the memorization problem of noisy labels, JoCoR (Wei, Feng, Chen & An, 2020), solves this issue by reducing the disparity between networks while updating the parameters to make the output of the network more aligned. The two networks are trained with the same joint-loss function simultaneously, and update the whole training set with the samples that yield the smallest loss in both networks from small mini-batches of training data, since these samples are highly probable to be clean. Using a technique similar to mini-batching, BundleNet (Li et al, 2017) aims to exploit correlations between samples to improve classifier performance. They also confirm that this approach acts as a regularization for the learning process. BundleNet generates a bundle of samples of each class in the data and each bundle is used as independent input. The model is able to utilize the correlation in the sample bundle and enhance the classification accuracy of deep neural networks with the presence of label noise.

2.2 Confident Learning

Transitioning to one of the most recent label noise tolerant techniques, confident learning emphasis shifts towards the stage of label noise identification. Confident learning focuses on the label quality in the data (Northcutt, Jiang & Chuang, 2021), its label noise aware nature can assign the relevant approaches label-noise tolerant character. Northcutt et al. proposed confident learning and successfully identified the noisy data in widely used public image datasets. However, after detection, they simply prune the noisy data, which unavoidably loses the useful information contained in the data. As a pioneer work in using confident learning in segmentation task which also achieved outstanding accuracy result, Li, Zhang, Zhao (2020) construct a teacher-student architecture, where the teacher module uses confident learning to identify the corrupted labels and the student module embeds spatial label smoothing regularization to correct the identified noisy labels.

2.3 Contrastive Learning

Contrastive learning has an effective impact on preventing deep network overfits on noisy labels and achieved significant improvement in empirical works (Xue, Whitecross & Mirzasoleiman, 2022). By combining contrastive learning with state-of-the-art methods like DivideMix and ELR+, which embed hybrid strategies, C2D has enhanced the classification accuracy of these methods markedly, especially with high noise rate data. C2D uses self-supervised contrastive learning to get a high-quality feature extractor to help with noise identification before applying the label noise learning methods. Utilizing contrastive learning's ability to extract the representations in the features, Ciortan et al. (2021) show that contrastiving can be used in the pre-training stage to improve the performance of the downstream classification task and they demonstrate empirical improvement of SOTA method like ELR by adding contrastive learning in the pre-training stage.

Gosh and Lan (2021) show empirically that using contrastive learning as an initializer can also effectively improve the classification result compared to the state-of-the-art methods. Karim, Rizve, Rahnavard, Mian and Shah (2022) aim to address the issue of clean sample selection bias during the sample separation process and the negative impact of noisy labels on memorization during learning. For instance, methods like DivideMix tend to select data with clean labels that are effective to learn, which leads to decreasing classification performance when the data has a high noise rate. UniCon (Karim et al., 2022) mitigates the negative impact of noisy labels in the learning process after the initial sample separation by embedding contrastive learning to learn the features without dependency on labels.

Huang, Lin and Xu (2021) proposed a framework that uses contrastive learning to find representative and difference between sample instances and class prototypes to correct noisy labels. The framework effectively learns the representations in the noisy dataset and shows improvement on classification accuracy. Similarly, Jo-SRC (Yar et al. 2021) uses contrastive learning to learn the similarity between samples. It creates two different types of data transformation for label prediction models and compares the prediction result based on a "likelihood" criterion, so that it identifies the potentially corrupted labels, which will be relabelled.

2.4 Active Learning

Active learning emerges as a refined sampling technique, which iteratively selects informative instances, thereby preserving valuable information in noisy data. Wu, Guo, Sheng, Zhao and Cui (2018) acknowledge the neglect of label noise in previous studies and propose an active learning framework with low-rank application. The conducted sampling strategy uses active learning to include and emphasise the sample pairs with less noise in the sampling. Combining two strategies in active learning, Re-Active (Bouguelia et al., 2018) achieved significant improvement on classification accuracy. In this approach, the informativeness of datapoints is measured by their influence on the predictive power classification model of any kind. In the sampling process, the first strategy selects the mislabeled data that can highly impact the learned model while the second strategy eliminates those that are highly influenced by the learned model.

Active learning also makes use of the informativeness of data (Bouguelia et al., 2018). Many studies focus on addressing the uncertainty of labels in noisy data, but the information contained in data with potential noisy labels should not be neglected. Traditional sample sieving techniques might hinder maximization of data utility (Frénay & Verleysen , 2014). To still retain the informative data with potentially corrupted labels while filtering suspicious data out, the learning algorithm QActor (Younesian, Zhao, Ghiassi, Birke & Chen, 2021) introduces an active learning methodology. After the first identification of clean and noisy labels, the active learning methodology manages to iteratively select informative noisy instances based on cross-entropy and entropy, and assign them with new labels by an oracle, which will then be included in the next training set. These approaches highlight the significance of maximizing data utility in the presence of abel noise.

2.5 Summary

In summary, while our review predominantly focuses on label noise challenges in image data, it underscores the necessity of extending such methodologies to other data types like sequential and tabular data. The experiments and the code implementation are also conducted within the same scope, which might limit its full application possibility on non-image data. For instance, Zhu, Wnag and Liu (2022) discover unstable performance of some label noise methods that are aimed at image data. Therefore, high performance label noise methods on image data do not necessarily guarantee equally high performance on non-image data. Sequential data, including time series data, and tabular data also suffer from label noise issues. These data types are the most commonly used in diverse scenarios. Therefore, investigating noise-tolerant methods for non-image data has its essentiality from a practical perspective. Furthermore, while methods under each category have their own strengths, there is potential to yield better performance by exploring combinations of these strategies. Additionally, many SOTA methods are inherently limited when it comes to model compatibility. For instance, contrastive learning methods work exclusively with deep neural networks (Xue, Whitecross

& Mirzasoleiman, 2022) and are therefore not compatible with more traditional machine learning methods, like tree based algorithms or regression models. As this paper focuses on data centric methods that are not tied to specific model architecture, we will direct our focus on model agnostic, noise tolerant methods.

3 Methodology

In this section, our focus is to showcase noise-tolerant methodologies for machine learning, particularly tailored for tabular datasets. Within our research, we have pinpointed three distinct approaches: confident learning, active learning, and biquality learning. Each method offers unique strategies for mitigating the adverse effects of noise within datasets. Subsequently, we explain each model and how it works. Furthermore, this paper undertakes a rigorous comparative analysis of their performance through comprehensive experimentation. Through this investigation, we aim to provide insights into the efficacy of these noise-tolerant techniques, thereby contributing to the advancement of robust machine-learning methodologies in the realm of tabular data analysis.

3.1 Confident Learning

Learning typically revolves around data, with a predominant emphasis on the confidence of model predictions rather than the quality of labels. However, confident learning (CL) presents a data-centric approach that prioritises label quality. It achieves this by characterising and identifying label errors within datasets through techniques such as pruning noisy data, establishing probabilistic thresholds to estimate noise, and ranking examples for training based on confidence levels. (Northcutt, C., 2021). The generalized CL framework is both theoretically consistent and empirically effective. Importantly, the CL framework is versatile, and applicable across different data modalities and models. We demonstrate this by using this technique in classification problems as well as regression problems in the experimentation part.

CL is a technique for estimating the joint distribution between observed noisy labels and the true underlying labels in a dataset. It relies on predicted probabilities(Pk, i) and the vector of noisy labels (yk) are calculated in advance using a given model (θ) , CL can work with various models (θ) , from more simple models such as logistic regression to XGBoost to neural networks, this is why it is also regarded as a model-agnostic method.

CL operates in three main steps: estimating the confident joint distribution C_y, y e counted as belonging to a class based on its predicted probability, offering flexibility and robustness. Q_y, y^* tion relies on counts from the confident joint, providing a robust approach to uncertainty quantification despite imperfect probability estimation. Cleaning the dataset involves rank and prune methods, where examples are ranked based on predicted probabilities, and noisy ones are removed. This improves data quality. Finally, CL adjusts the training procedure to account for the cleaned dataset, ensuring model training on reliable data.

CL represents a significant advancement in machine learning, particularly in the context of regression and classification tasks on tabular data. Unlike traditional methods that rely solely on observed labels, CL leverages both observed labels (potentially noisy) and estimated true labels, providing a robust framework for learning from imperfectly labeled datasets. Through the integration of predicted probabilities and noisy labels, CL facilitates the estimation of the joint distribution between observed and true labels, thus enabling enhanced model performance and generalization. This methodology stands in contrast to conventional approaches, which often struggle to handle label noise effectively, leading to suboptimal performance and biased model outputs.

Therefore, CL presents a novel strategy for minimizing the impact of label noise on predictive accuracy in regression tasks. By utilizing out-of-sample predicted probabilities and noisy labels, CL can identify and correct label errors, bolstering the dependability of regression models in practical applications. Similarly, CL provides a potent framework for addressing noisy labels in classification problems. CL can pinpoint and remove mislabeled instances through confident joint estimation and rank-and-prune techniques, leading to more accurate and robust classification models.

Recent research has showcased the effectiveness of CL in various domains, such as healthcare, finance, and natural language processing. Smith et al. (2021) leverage CL on electronic health record data to enhance the precision of predictive models for diagnosing diseases, yielding considerable performance improvements versus conventional methods. Similarly, Zhang and Li (2020) employ CL on financial datasets to heighten the dependability of fraud detection algorithms, resulting in fewer false positives and better fraud detection rates.

In summary, CL represents a groundbreaking approach to learning from noisy data, providing unparalleled robustness and performance in regression and classification tasks on tabular data. By harnessing the strength of predicted probabilities and noisy labels, CL permits accurate joint distribution estimation, leading to more dependable and generalizable models in real-world scenarios.

3.2 Active Learning

While traditional active learning models prove to be effective in reducing labelling costs and selecting the most influential samples for model training, the negative effect of label noise is often neglected (Settles, 2009).

Bouguelia et al. (2018) however propose an active learning method that introduces a query strategy framework for sample selection, as well as a mitigation method for label noise.

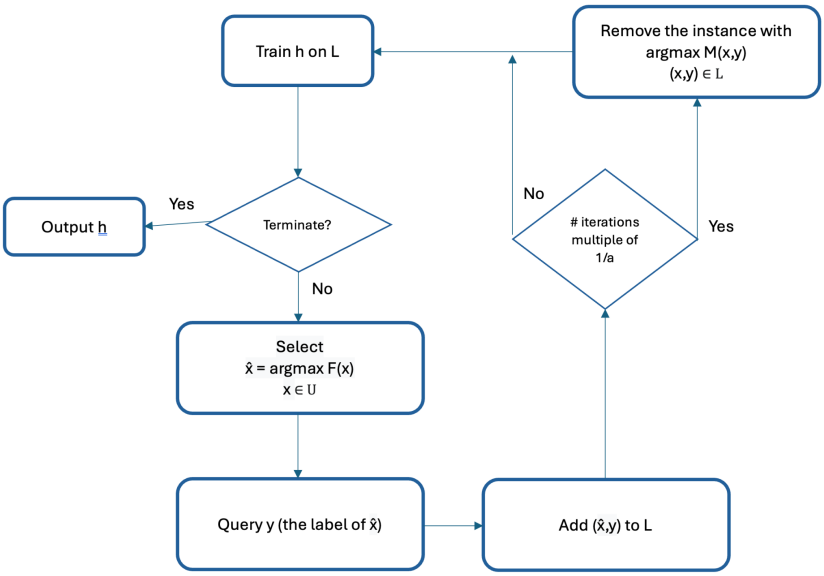

Figure 1: Active learning algorithm

The flowchart for the proposed algorithm is depicted in the figure above and contains two decisive calculations$F(x)$ and $M(x, y)$

For initiation, an oracle set L is needed to train the initial classifier h . In most active learning methods, this oracle set is expected to have ground truth labels for all its instances. Here however, wrongly labelled instances within the oracle are considered, and initial members of the oracle set may be removed in any iteration by the algorithm.

As long as the termination condition is not met, the following calculations will be cyclically computed and relevant decisions will be made accordingly. After initial training of the classifier h on the oracle dataset, the most informative sample from the unlabeled dataset $U, \hat{x}$ is determined via $F(x)$, also called disagreement measure 1.

Disagreement measure 1 measures the amount of information gained by adding an unlabeled instance $\bar{x}$. This is done by comparing two classification models a and b and their respective predictions for each x element U . While a is the current model h , trained on L , b , represents the same model that has been trained additionally on $\bar{x}$.

The corresponding formula is depicted below.

$$d(a,b) = \mathop{\mathbb{P}}_{x\sim D}[a(x) \neq b(x)] \simeq \frac{|\{x \in U : a(x) \neq b(x)\}|}{|U|}.$$

Figure 2: Disagreement measure 1

A corresponding label y is determined by using the classifier h as a predictor. The sample is then added to L . Potentially wrong labels are determined after a specified number of iterations $1/\alpha$ This hyperparameter α can be specified by the user. Here, $M(x, y)$, or disagreement measure 2 is calculated for every sample within L . The sample with the highest $M(x, y)$ value is specified as the one which is most likely to be wrong and can either be checked by the user, or be

automatically removed from L . As this paper compares holistic methods that do not necessitate additional efforts by the user, samples will be removed in this paper.

In detail, disagreement measure 2 measures the disagreement between the current model's prediction and the predictions of a classification model committee $C = h_x : x \in U$. With a higher number of disagreements between the ensemble cast and the current model, the instance is deemed likely to be wrong. For a more refined measure, the posterior probability for the predictor class for each label is considered as can be seen below. This is also the measure argmax $M(x, y)$ used in the experiments of this paper.

$$F_2'(\bar{x}) = \max_{y \in Y} \sum_{x \in U} \Big[\mathbb{1}(h(\bar{x}) \neq h_x(\bar{x})) \times P(y|\bar{x}, h_x) \Big].$$

Figure 3: Disagreement measure 2

The instance with the highest level of disagreement is then removed from the dataset L.

Termination of the algorithm can be achieved by meeting certain stopping criteria, like meeting a performance threshold, or a maximum number of iterations.

It is important to mention that both, disagreement measure 1 and 2 run with a complexity of $O(n^2)$, posing some runtime challenges.

3.3 Biquality learning / Weakly supervised learning

While Bouguelia et al. acknowledge the possible unreliability of the oracle, which is utilized in their active learning method, weakly supervised learning algorithms leverage small-sized "ground truth" oracles to enhance model performance on larger, noisy datasets. There is no clear guideline on how large the trusted oracle size should be, but they typically include between 25% and 75% of the overall training data in other research (Nodet et al., 2023). One subcategory within weakly supervised learning, termed inaccurate supervision, aims to alleviate the impact of noisy data on model performance specifically (Zhou, 2018). While numerous inaccurate supervision methodologies exist, only a subset is suitable for tabular data analysis, concurrently exhibiting resilience to distributional discrepancies between training and test data, as well as noise. Two prominent members are kernel mean matching (KMM) and probabilistic density ratio estimation (PDR) algorithms (Nodet et al., 2023).

Alternative methods, like learning to reweight (L2RW) are not compatible with commonly used machine learning classifiers, such as XGBoost, due to their architecture being specifically geared towards deep learning algorithms (Ren et al., 2018). Therefore, the focus of this paper remains solely on KMM and PDR algorithms. For both methods, the inherent goal is to recognize feature distribution differences between training and test features and create similar feature distributions through Sample reweighting (Nodet et al., 2023). KMM does so by reweighting training points in such a way, that the means of training and test points are close in a reproducing kernel Hilbert space (RHKS) (Gretton et al., 2008). PDR on the other hand calculates a predicted probability for a label being wrong for each instance and reweights it for model training. Both algorithms are specified in more detail in the Appendix figure 1 (KMM) and figure 2 (PDR). An implementation of both algorithms in python is publicly available in python and is utilized in this paper (Nodet et al., 2023).

4 Data and Setup

4.1 Datasets

One of the challenges that we faced was to decide on the datasets that we can work with. There are no benchmark tabular data for studying label noise as is the case with image datasets where there are a lot of datasets readily available with label noise in it. We were unable to find any tabular data that has label noise in it. So, we synthetically added label noise in various types of data to evaluate the performance of our models. There are many known types of ways to introduce label noise in datasets, namely random label noise, class-dependent label noise, instance-dependent label noise etc.(Algan & Ulusoy, 2020). In this paper, we stick to the method of random label flipping. Random label noise occurs when labels are mislabeled or corrupted randomly, without any specific pattern or structure (Natarajan et al.,2013).

We test our models on four different types of datasets. Two of them are binary classification ones and the other two are multiclass classification datasets. All the four datasets have been obtained from Kaggle.

4.1.1 Brain Tumor Data

This dataset is from the medical domain. It has a total of 15 features with 13 numerical features, 1 image column with the probable image of the tumors, and a label feature column. The label feature is a binary classification with 1 signifying the presence of tumor and 0 the absence of it. It is a medium sized dataset with 3762 observations. There are no missing values in the data. We did not use the image column of the dataset. (Brain tumor, 2020).

4.1.2 Pima Indians Diabetes Database

This dataset is also from the medical domain. It was originally collected from the National Institute of Diabetes and Digestive and Kidney Diseases. The reason behind choosing this dataset was mainly two fold. It is a relatively small database with 9 features and 768 rows. In this way we get to test our models on datasets of different sizes. Another reason being, this dataset has been used in many other machine learning experiments (Chang et al., 2022). The object of this dataset is to predict if a patient has diabetes based on 8 diagnostic measurements. The target outcome is 1, if the patient has diabetes and 0, if she has not. All the features are numerical, and there are no missing values in the data (Pima Indians Diabetes Database,2016).

4.1.3 Stellar Classification Dataset - SDSS17

This dataset is from a very different astronomical domain. Stellar classification is the classification of stars, galaxies and quasars based on their spectral characteristics. The data consists of 100,000 observations of space taken by SDSS (Sloan Digital Sky Survey), each observation being described by 17 feature columns and 1 class column which identifies it to be either a star, galaxy or quasar(QSO). All the feature columns consist of numerical values, only the class column is textual. There is no missing value in this dataset. The fact that it is a really large data from a very different domain made it ideal to test our models on it (Stellar Classification Dataset- SDSS17,2022).

4.1.4 Customer Segmentation Data

The original source of the database is from a hackathon by the name of Analytics Vidhya. It is survey data. The survey was carried out by an automobile company on 10,695 potential new customers in a new market that they plan to enter. Based on the survey, they are segmenting the customers into 4 different segments A, B, C, and D. It has 10 feature columns and 1 Segmentation column. Only 4 feature columns are numericals, all the others are object data types.(Customer Segmentation Classification,2023). There are a lot of missing values in this dataset. A total of 2108 missing values with highest being 1098 in the Work experience feature. We dealt with the missing values by replacing them with the most frequent occurrence in that column. This dataset was unique in that it is a multi-class classification survey data. It is from a different domain than the other datasets and also it has missing values. Another very different data to evaluate our models. (Customer Segmentation Classification, 2023)

We introduced 20% and 40% label noise in all these datasets by randomly flipping a few of the label feature column values, and tested the performance of our models in all the datasets.

4.2 Experimentation Design

4.2.1 Models

As the focus of this paper lies on identifying and comparing the potential to increase model performance while not changing model architecture, a simple, yet effective classifier model will be chosen for this paper. We utilize an XGBoost Classifier which is not fine tuned on separate datasets but rather remains with its default configuration. While a fine tuned XGBoost model may provide better performance results (Bentéjac et al., 2020), the performance of a base XGBoost algorithm is deemed to be sufficient as a baseline. Relevant hyperparameters are kept at n_estimators=2, max_depth=2 and learning_rate=1.

As an XGBoost classifier is suitable for both, binary tasks, as well as multiclass classifications tasks, the model can and will be applied to all datasets, which have been included in this study. For a refined scope, which focuses on the efficacy of our presented noise tolerant methods, we abstain from using any alternative classification models in our experimentations.

4.2.2 Performance Metrics

As we focus on multiclass and binary classification tasks within the scope of this paper, only performance metrics dedicated to classification tasks will be considered for the experiments. To cover a holistic view of performance for each model and its dedicated noise tolerant method, multiple metrics will be taken into account. Classification metrics can be categorized into three different groups, metrics based on qualitative understanding of error, probabilistic understanding of error and understanding on how well a model ranks its class predictions (Ferri et al., 2009). For experimentation, we will implement one metric from each grouping as depicted below.

Family	Metric	Formula
Qualitative Understanding	Accuracy	$\mathrm{Acc} = \frac{\sum_{i=1}^{m}\sum_{j=1}^{c} f(i,j)C(i,j)}{m}$
Probabilistic Understanding	Log loss/ Cross entropy	$\mathrm{LogL} = \frac{-\sum_{j=1}^{c}\sum_{i=1}^{m}(f(i,j)\log_2 p(i,j))}{m}$
Model ranking	Area under Curve (one vs rest)	$\mathrm{AUC}\ (j,k) = \frac{\sum_{i=1}^{m} f(i,j)\sum_{i=1}^{m} f(t,k)I(p(i,j),p(t,j))}{m_j \cdot m_k}$. $\mathrm{AUNU} = \frac{\sum_{j=1}^{c}\mathrm{AUC}\ (j,\mathrm{rest}_j)}{c}$

Figure 4: Performance metrics

It is important to note that for both accuracy and area under curve (AUC), a higher value is deemed to be better, while for log loss, a lower score indicates a higher performance. By considering all three metrics in the assessment of noise tolerant methods, possible weaknesses and strengths can be identified.

Accuracy is one of the simplest and most widely used performance measures for classification tasks. It indicates the ratio of correctly predicted labels divided by the total number of predictions.

Log loss indicates how well probability estimates for a given classification prediction are, meaning how certain the model is of its prediction.

AUC is a measure of the area under the receiver operating characteristic curve, which depicts a plot of the true positive rate and the false positive rate for different threshold values of the classification model. It therefore not only considers probability estimates made by the classifier, but is also insensitive to imbalanced datasets. While the AUC measure is primarily meant for binary classification tasks, many modified versions exist that also apply to multiclass problems. AUNU is the method that we implemented, which treats a multiclass problem as many binary tasks with j being the positive class and $rest_j$ being the negative class, which combines all other classes apart from j .

4.2.3 Experimental Setup

All three methods have been implemented in a jupyter notebook and were applied to the introduced datasets at different noise frequencies. As a baseline, a plain XGBoost model was also applied to each dataset to compare the model's performance without any noise tolerant method. To ensure more stable results, each model was applied to 5 different training and test sets based on random sampling which can be replicated with the indicated random seeds within the jupyter notebook. One notable exception is the active learning method, which has only been applied to one randomly sampled training and test set due to computational restraints.

In all cases, a training size of 75% of the data and a test size of 25 % has been chosen.

For both, biquality learning and active learning, a set size for the oracle dataset must also be determined. While there is no clear indication on how large the oracle set should be in relation to the original dataset, we decided to allocate 25% of the training samples, or 18.75% of the original dataset to the oracle set. While oracle size might have a significant influence on the overall performance of the noise tolerant method, further oracle sizes are not within the scope of this review. Apart from ensuring that no data is missing within both the training and test set, no further data cleaning or data augmentation was applied to any of the datasets.

For active learning, it is necessary to determine hyperparameter α the specify after how many iterations an instance is removed from set L . For the 20% noise datasets α was set to 5, resulting in one instance being removed after adding 5 instances from U each. Due to the higher noise rate, for the 40% noise dataset α was set equal to 3, meaning that after adding 3 instances, one is removed from L . No specific stopping criterion was indicated, meaning that the algorithm only terminates after iterating over every instance U .

For biquality learning, only a specific kernel function needs to be specified for the KMM algorithm. To ensure strictly positive weights for each instance, the sigmoid kernel function was chosen for each application of the method. No further hyperparameter specifications are made for contrastive learning.

5 Results

A detailed table of all performance metrics for each randomly sampled train and test set can be found in appendix figure 3, 4 and 5, while a more condensed representation, indicating the mean values for the 5 iterations for each experimental setting can be found in appendix 6,7 and 8.

A visualization of the latter, compounded into three line diagrams can be seen below. Each diagram depicts the respective dataset on the x axis and the corresponding metric scores on the y axis.

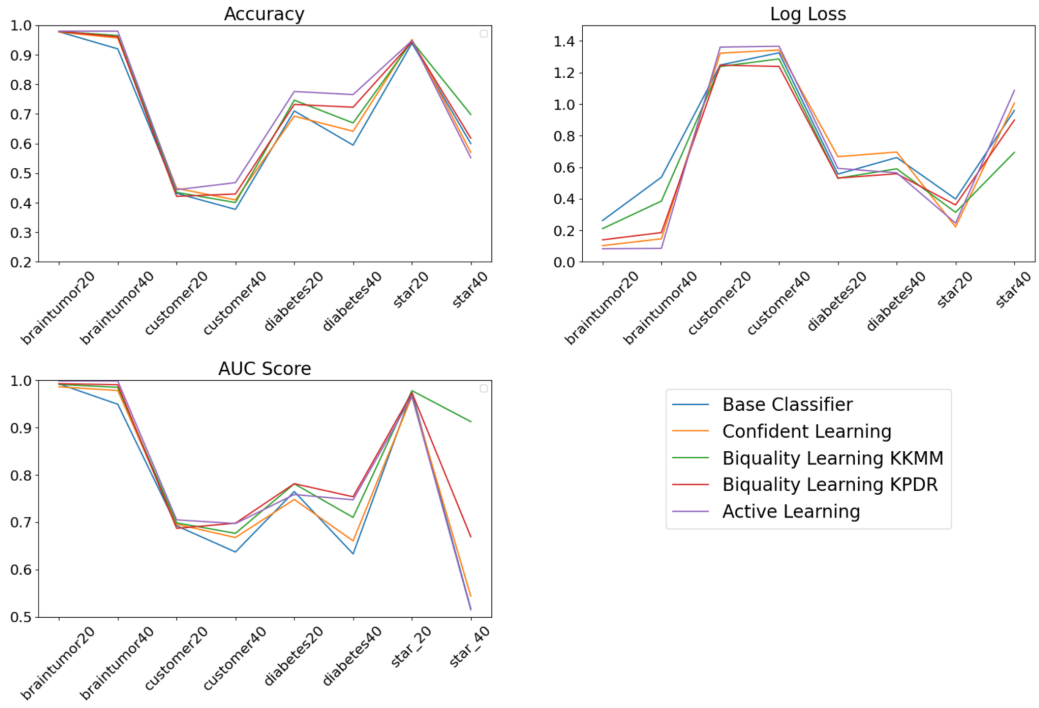

Figure 5: Experimentation results

As expected, the introduction of noise within each dataset generally resulted in a decline in performance across our selected metrics. Notable exceptions are observed, being log loss in the diabetes dataset for active learning and some scores for the customer segmentation dataset.

In most cases, the base model lacking any noise-tolerant technique exhibited a notably more pronounced decrease in performance compared to models augmented with the proposed methods. Some examples include the drop in accuracy and AUC of the brain tumor and diabetes datasets. The most noticeable score difference between a 20% noise and its corresponding 40% noise dataset occurs for the star dataset. While confident learning and KPDR biquality learning were successful in mitigating the negative impact of the higher noise on the star dataset, only KKMM biquality learning shows a significant improvement in performance.

Confident learning consistently ranks below average for each dataset and classification metric. This trend is especially noticeable across the diabetes and star datasets. It is noteworthy that comparative enhancements relative to the base model were especially prominent in binary datasets (brain tumor $\&$ diabetes), and are less drastic for multiclass

datasets (star & customer segmentation). The performance trends across metrics also do not differ greatly from another. Particularly the AUC and accuracy curves do look quite similar with some exceptions.

6 Discussion

Based on the results, several key observations emerge regarding the impact of label noise and the effectiveness of noise tolerant methods:

6.1 Impact of label noise on performance metrics

Consistently across most datasets, an increase in label noise leads to a decrease in performance across various performance metrics. This finding aligns with the expectations and underscores the detrimental effect of label noise on model performance.

6.2 Exceptions to the trend

While the general trend indicates a decrease in performance with increased label noise, there are notable exceptions. For instance, log loss in the diabetes dataset for active learning, and some scores for the customer segmentation dataset show either minimal decrease or even improvement. These exceptions warrant further investigation to understand the underlying factors contributing to them.

6.3 Effectiveness of Noise-Tolerant Methods

Models enhanced with noise-tolerant methods generally exhibit better performance compared to the base model without any noise-tolerant techniques. This is particularly evident in metrics like accuracy and AUC. This suggests that incorporating noise-tolerant methods can mitigate the adverse effects of label noise and improve model robustness.

6.4 Relative Performance of Confident Learning

Confident learning consistently ranks below average across datasets and classification metrics. This trend is particularly noticeable for the diabetes and stellar classification datasets. One decisive difference between confident learning and the other noise tolerant methods within this paper is the lack of a clean oracle set. For both active learning and biquality learning, 18.75% of the clean dataset is provided as oracle data for their respective computations. This alone might prove to be a significant advantage in comparison to confident learning, which exclusively worked with the noisy datasets. In many cases, it might be feasible to provide a significantly smaller subset, for which correct labeling can be guaranteed. This however always goes hand in hand with additional costs.

6.5 Comparison between Binary and Multiclass Datasets:

Comparative improvements in performance, relative to the base model, are more pronounced for binary datasets (Brain tumor and Diabetes) compared to the multiclass datasets(Stellar classification and Customer Segmentation). This observation suggests that the impact of noise-tolerant methods may vary depending on the dataset characteristics, such as the number of classes.

7 Conclusion and Future Work

In this paper, we briefly introduced three different noise tolerant methods and compared their ability to mitigate the influence of label noise on the performance of an XGBoost classifier. Our focus lies on the application on tabular data, as most state of the art research focuses on image data. While we introduced certain configurations of all three noise tolerant methods, there are many hyperparameters that can be changed and could significantly change the efficacy of each method. Some of these factors include, but are not limited to the size of available, “ground-truth" instances for the oracle dataset for active learning and biquality learning, early stopping criteria for active learning and different kernel transformation functions for KKMM biquality learning. A change in these hyperparameters could significantly change the results that are explored within this paper. We exclusively utilized an untuned XGBoost classifier for predictions. However, a potential area for future investigation involves comparing the relative impact of noise-tolerant methods on a variety of both tuned and untuned classifiers. Another future work could be evaluating the models on other types of label noise like class or instance dependent label noise.Evaluating models on different types of label noise and diverse

datasets, including real world data, could provide insights into their generalizability.In this paper we implement selected methods on tabular data, it would also be interesting to explore the possibility of combining diverse strategies in each stage and the impact of such a combination on performance.

8 Appendix

Figure 1: KDR / K-KMM Algorithm

Algorithm 2: K-DensityRatio (K-DR)

Input: Trusted Dataset D_T, Untrusted Dataset D_U, Density Ratio Estimator Family $\mathcal{E}$, Probabilistic Classifier Familiy $\mathcal{F}$

for $k \in [\![1, K]\!]$ **do**

  Let $D_T^k = \{\forall (x, y) \in D_T \mid y = k\}$

  Let $D_U^k = \{\forall (x, y) \in D_U \mid y = k\}$

  Learn $e^k \in \mathcal{E}$ on D_T^k and D_U^k

for $(x_i, y_i) \in D_U$ **do**

  $\hat{\beta}(x_i, y_i) = e^{y_i}(x_i) \frac{|D_T^{y_i}|}{|D_T|} \frac{|D_U|}{|D_U^{y_i}|}$

for $(x_i, y_i) \in D_T$ **do**

  $\hat{\beta}(x_i, y_i) = 1$

Learn $f \in \mathcal{F}$ on $D_T \cup D_U$ with weights $\hat{\beta}$

Output: f

Figure 2: K-PDR Algorithm

Algorithm 3: K-Probabilistic Density Ratio (K-PDR)

Input: Trusted Dataset D_T, Untrusted Dataset D_U, Probabilistic Classifier Familiy $\mathcal{F}$

for $k \in [\![1, K]\!]$ **do**

  Let $D_T^k = \{\forall (x, y) \in D_T \mid y = k\}$

  Let $D_U^k = \{\forall (x, y) \in D_U \mid y = k\}$

  Learn $f_S^k \in \mathcal{F}$ on $\{(x, s(x)) \mid \forall x \in D_T^k \cup D_U^k\}$

for $(x_i, y_i) \in D_U$ **do**

  $\hat{\beta}(x_i, y_i) = \frac{f_S^{y_i}(x_i)_1}{f_S^{y_i}(x_i)_0} \frac{|D_U|}{|D_T|}$

for $(x_i, y_i) \in D_T$ **do**

  $\hat{\beta}(x_i, y_i) = 1$

Learn $f \in \mathcal{F}$ on $D_T \cup D_U$ with weights $\hat{\beta}$

Output: f

Figure 3: Accuracy score

	Base Classifier	Confident Learning	Biquality Learning KKMM	Biquality Learning KPDR	Active Learning	Dataset
0	0.952000	0.956000	0.952000	0.952000	0.948000	star20
1	0.948000	0.952000	0.940000	0.940000	0.948000	star20
2	0.932000	0.956000	0.956000	0.956000	0.948000	star20
3	0.928000	0.956000	0.944000	0.936000	0.948000	star20
4	0.936000	0.940000	0.936000	0.936000	0.948000	star20
0	0.576000	0.536000	0.656000	0.616000	0.552000	star40
1	0.628000	0.576000	0.728000	0.668000	0.552000	star40
2	0.640000	0.584000	0.688000	0.640000	0.552000	star40
3	0.544000	0.540000	0.692000	0.540000	0.552000	star40
4	0.612000	0.616000	0.728000	0.628000	0.552000	star40
0	0.755208	0.750000	0.765625	0.760417	0.776042	diabetes20
1	0.692708	0.661458	0.755208	0.739583	0.776042	diabetes20
2	0.713542	0.656250	0.703125	0.687500	0.776042	diabetes20
3	0.656250	0.666667	0.744792	0.692708	0.776042	diabetes20
4	0.734375	0.729167	0.765625	0.781250	0.776042	diabetes20
0	0.604167	0.682292	0.656250	0.734375	0.765625	diabetes40
1	0.713542	0.723958	0.739583	0.708333	0.765625	diabetes40
2	0.640625	0.578125	0.656250	0.718750	0.765625	diabetes40
3	0.510417	0.625000	0.677083	0.744792	0.765625	diabetes40
4	0.505208	0.598958	0.619792	0.708333	0.765625	diabetes40
0	0.990436	0.987248	0.991498	0.991498	0.980000	braintumor20
1	0.963868	0.964931	0.969182	0.972370	0.980000	braintumor20
2	0.974495	0.975558	0.974495	0.973433	0.980000	braintumor20
3	0.981934	0.977683	0.977683	0.984060	0.980000	braintumor20
4	0.979809	0.979809	0.979809	0.974495	0.980000	braintumor20
0	0.941552	0.978746	0.991498	0.988310	0.980000	braintumor40
1	0.931987	0.943677	0.944740	0.967056	0.980000	braintumor40
2	0.904357	0.928799	0.979809	0.938363	0.980000	braintumor40
3	0.914984	0.980871	0.970244	0.968119	0.980000	braintumor40
4	0.907545	0.952179	0.941552	0.943677	0.980000	braintumor40
0	0.452000	0.456000	0.432000	0.436000	0.444000	customer20
1	0.408000	0.448000	0.412000	0.452000	0.444000	customer20
2	0.424000	0.440000	0.460000	0.464000	0.444000	customer20
3	0.424000	0.456000	0.396000	0.380000	0.444000	customer20
4	0.452000	0.444000	0.476000	0.376000	0.444000	customer20
0	0.352000	0.392000	0.404000	0.408000	0.468000	customer40
1	0.340000	0.376000	0.428000	0.484000	0.468000	customer40
2	0.364000	0.420000	0.404000	0.448000	0.468000	customer40
3	0.420000	0.432000	0.396000	0.392000	0.468000	customer40
4	0.412000	0.428000	0.372000	0.416000	0.468000	customer40

Figure 4: Log loss score

	Base Classifier	Confident Learning	Biquality Learning KKMM	Biquality Learning KPDR	Active Learning	Dataset
0	0.380470	0.212861	0.323059	0.355363	0.245779	star20
1	0.378239	0.218511	0.323028	0.369011	0.245779	star20
2	0.373760	0.216996	0.293958	0.344932	0.245779	star20
3	0.453153	0.213839	0.318100	0.364121	0.245779	star20
4	0.408929	0.250430	0.313017	0.371039	0.245779	star20
0	0.995569	1.099216	0.694392	0.872988	1.087161	star40
1	0.924900	0.945277	0.695646	0.879741	1.087161	star40
2	0.937519	0.922716	0.684575	0.910829	1.087161	star40
3	1.005677	1.113818	0.748267	0.987659	1.087161	star40
4	0.937272	0.956454	0.650442	0.847253	1.087161	star40
0	0.531284	0.540846	0.516729	0.479051	0.592876	diabetes20
1	0.559372	0.785099	0.550768	0.545341	0.592876	diabetes20
2	0.562882	0.695028	0.544963	0.600088	0.592876	diabetes20
3	0.587895	0.697959	0.550608	0.545159	0.592876	diabetes20
4	0.541191	0.616182	0.488932	0.487649	0.592876	diabetes20
0	0.617954	0.660537	0.583732	0.534804	0.564555	diabetes40
1	0.646429	0.648643	0.544295	0.572672	0.564555	diabetes40
2	0.678134	0.664892	0.628449	0.574220	0.564555	diabetes40
3	0.676241	0.744128	0.557957	0.548468	0.564555	diabetes40
4	0.689283	0.764641	0.636392	0.563248	0.564555	diabetes40
0	0.253597	0.070099	0.193587	0.130495	0.083545	braintumor20
1	0.268015	0.142656	0.237409	0.152298	0.083545	braintumor20
2	0.248715	0.119121	0.198227	0.149046	0.083545	braintumor20
3	0.270045	0.088205	0.207315	0.119487	0.083545	braintumor20
4	0.271947	0.095820	0.218826	0.148862	0.083545	braintumor20
0	0.525171	0.104611	0.374064	0.169765	0.085904	braintumor40
1	0.537758	0.192327	0.396342	0.185815	0.085904	braintumor40
2	0.554046	0.167790	0.379918	0.229625	0.085904	braintumor40
3	0.529315	0.114027	0.388752	0.160383	0.085904	braintumor40
4	0.538318	0.154919	0.389027	0.184226	0.085904	braintumor40
0	1.261669	1.325940	1.257164	1.238094	1.361423	customer20
1	1.236402	1.312596	1.219958	1.200497	1.361423	customer20
2	1.227516	1.292651	1.222054	1.238831	1.361423	customer20
3	1.265086	1.317015	1.271999	1.313475	1.361423	customer20
4	1.248874	1.368171	1.220731	1.243458	1.361423	customer20
0	1.315872	1.417415	1.275133	1.253716	1.366909	customer40
1	1.341584	1.380496	1.261511	1.187659	1.366909	customer40
2	1.325301	1.300983	1.289115	1.239043	1.366909	customer40
3	1.316123	1.320599	1.297798	1.264229	1.366909	customer40
4	1.328632	1.295067	1.310154	1.250675	1.366909	customer40

Figure 5: AUC scores

	Base Classifier	Confident Learning	Biquality Learning KKMM	Biquality Learning KPDR	Active Learning	Dataset
0	0.968846	0.969990	0.974357	0.971929	0.966028	star_20
1	0.986394	0.972117	0.976479	0.981548	0.966028	star_20
2	0.965799	0.959663	0.965893	0.964865	0.966028	star_20
3	0.959883	0.976557	0.986343	0.966801	0.966028	star_20
4	0.990791	0.951603	0.986962	0.983004	0.966028	star_20
0	0.466866	0.469268	0.929137	0.741476	0.515214	star_40
1	0.580603	0.612296	0.892941	0.661489	0.515214	star_40
2	0.505072	0.597717	0.916251	0.579701	0.515214	star_40
3	0.487270	0.508306	0.896816	0.557542	0.515214	star_40
4	0.548796	0.533086	0.928156	0.807194	0.515214	star_40
0	0.785437	0.817014	0.795452	0.838282	0.758631	diabetes20
1	0.770707	0.719129	0.723548	0.748864	0.758631	diabetes20
2	0.741008	0.686998	0.775230	0.732918	0.758631	diabetes20
3	0.744293	0.728846	0.768486	0.758933	0.758631	diabetes20
4	0.782894	0.788822	0.843674	0.828767	0.758631	diabetes20
0	0.726876	0.719630	0.744963	0.806822	0.747437	diabetes40
1	0.713068	0.669697	0.735922	0.701136	0.747437	diabetes40
2	0.596613	0.648237	0.648877	0.732976	0.747437	diabetes40
3	0.585732	0.628474	0.753164	0.753660	0.747437	diabetes40
4	0.543801	0.637447	0.668643	0.774663	0.747437	diabetes40
0	0.996816	0.996156	0.996610	0.998009	0.998533	braintumor20
1	0.989200	0.974030	0.980340	0.987293	0.998533	braintumor20
2	0.994102	0.977957	0.993754	0.993690	0.998533	braintumor20
3	0.989131	0.993793	0.992702	0.994271	0.998533	braintumor20
4	0.992799	0.988510	0.993149	0.990777	0.998533	braintumor20
0	0.961438	0.993973	0.994960	0.993001	0.998433	braintumor40
1	0.952114	0.961967	0.971214	0.986042	0.998433	braintumor40
2	0.934799	0.977623	0.984756	0.988610	0.998433	braintumor40
3	0.936097	0.986272	0.990264	0.993143	0.998433	braintumor40
4	0.961942	0.971837	0.984471	0.991965	0.998433	braintumor40
0	0.691489	0.695146	0.692525	0.699205	0.705145	customer20
1	0.699365	0.710657	0.699679	0.708787	0.705145	customer20
2	0.704669	0.700333	0.710240	0.697042	0.705145	customer20
3	0.665563	0.684034	0.666449	0.629783	0.705145	customer20
4	0.701629	0.689304	0.724861	0.701185	0.705145	customer20
0	0.636320	0.666316	0.690574	0.696412	0.697364	customer40
1	0.614217	0.652208	0.697549	0.723672	0.697364	customer40
2	0.647854	0.674214	0.665585	0.699361	0.697364	customer40
3	0.645854	0.676718	0.662331	0.675070	0.697364	customer40
4	0.641817	0.669563	0.667234	0.695863	0.697364	customer40

Figure 6: Accuracy mean scores

Dataset	Base Classifier	Confident Learning	Biquality Learning KKMM	Biquality Learning KPDR	Active Learning
braintumor20	0.978108	0.977046	0.978533	0.979171	0.980000
braintumor40	0.920085	0.956854	0.965569	0.961105	0.980000
customer20	0.432000	0.448800	0.435200	0.421600	0.444000
customer40	0.377600	0.409600	0.400800	0.429600	0.468000
diabetes20	0.710417	0.692708	0.746875	0.732292	0.776042
diabetes40	0.594792	0.641667	0.669792	0.722917	0.765625
star20	0.939200	0.952000	0.945600	0.944000	0.948000
star40	0.600000	0.570400	0.698400	0.618400	0.552000

Figure 7: Log loss mean scores

Dataset	Base Classifier	Confident Learning	Biquality Learning KKMM	Biquality Learning KPDR	Active Learning
braintumor20	0.262464	0.103180	0.211073	0.140038	0.083545
braintumor40	0.536922	0.146735	0.385620	0.185963	0.085904
customer20	1.247910	1.323275	1.238381	1.246871	1.361423
customer40	1.325502	1.342912	1.286742	1.239064	1.366909
diabetes20	0.556525	0.667023	0.530400	0.531458	0.592876
diabetes40	0.661608	0.696568	0.590165	0.558682	0.564555
star20	0.398910	0.222528	0.314232	0.360893	0.245779
star40	0.960187	1.007496	0.694664	0.899694	1.087161

Figure 8: AUC mean scores

Dataset	Base Classifier	Confident Learning	Biquality Learning KKMM	Biquality Learning KPDR	Active Learning
braintumor20	0.992409	0.986089	0.991311	0.992808	0.998533
braintumor40	0.949278	0.978334	0.985133	0.990552	0.998433
customer20	0.692543	0.695895	0.698751	0.687201	0.705145
customer40	0.637213	0.667804	0.676655	0.698075	0.697364
diabetes20	0.764868	0.748162	0.781278	0.781553	0.758631
diabetes40	0.633218	0.660697	0.710314	0.753852	0.747437
star_20	0.974343	0.965986	0.978007	0.973629	0.966028
star_40	0.517721	0.544134	0.912660	0.669480	0.515214

Table 1: Literature table

Title	**Data type- / domain-specific**	**Label noise tolerant method**	**Empirical studies done/ Experimentation results**	**Code availablility**
A generalised label noise model for classification in the presence of annotation errors	TRUE	TRUE	improvement in comparison to similar methods	No implementation in code, but detaild description
A novel self-supervised re-labeling approach for training with noisy labels	image data	TRUE	significant improvement	No Code
A second-order approach to learning with instance-dependent label noise	instance-dependent noise	robust learning method	minor improvement in comparison to similar methods	Code available
A Soft-label Method for Noise-tolerant Distantly Supervised Relation Extraction	textual data	label correcting method	improvement in comparison to similar methods	No Code
Agreeing to disagree: active learning with noisy labels without crowdsourcing.	TRUE	TRUE	marginal improvement in comparison to similar methods	No Code
An improved noise loss correction algorithm for learning from noisy labels	image data	robust learning method		No Code
Arm: A confidence-based adversarial reweighting module for coarse semantic segmentation	textual data	TRUE	marginal improvement in comparison to similar methods	No Code
Attribute-efficient learning of half-spaces with malicious noise: Near-optimal label complexity and noise tolerance	TRUE	optimizes the use of attributes and features to improve learning efficiency		No code
Beyond class-conditional assumption: A primary attempt to combat instance-dependent label noise	instance-dependent noise	learning process efficiency	significant improvement	Code available
BundleNet: Learning with Noisy Label via Sample Correlations	image data	TRUE	partial marginal improvement in comparison to similar methods	Code available
Can less be more? when increasing-to-balancing label noise rates considered beneficial	image data	sample balancing	improvement combingning with existing methods	No code
CLC: Noisy Label Correction via Curriculum Learning	image data	TRUE	marginal improvement in comparison to similar methods	Code available
Co-correcting: noise-tolerant medical image classification via mutual label correction	image data	TRUE	SOTA	Code available
Combating noisy labels by agreement: A joint training method with co-regularization	image data	TRUE	SOTA	Code available
Combined cleaning and resampling algorithm for multi-class imbalanced data with label noise	TRUE	TRUE	partial improvement in comparison to similar methods	Code available
Confidence scores make instance-dependent label-noise learning possible	instance-dependent noise	TRUE	improvement in comparison to similar methods	Code available
Contrast to divide: Self-supervised pre-training for learning with noisy labels	image data	TRUE	improvement with SOTA combination	Code available
Contrastive learning improves model robustness under label noise	image data	TRUE	marginal improvement in comparison to similar methods	Code available

Data uncertainty learning in face recognition	image data	TRUE	marginal improvement in comparison to similar methods	Code available
Detecting Label Noise via Leave-One-Out Cross-Validation	TRUE	noise detection	no comparison	No Code
Dividemix: Learning with noisy labels as semi-supervised learning	image data	TRUE	SOTA	Code available
Dynamic training data dropout for robust deep face recognition	image data	dynamic filtering	improvement with SOTA combination	No Code
Early-learning regularization prevents memorization of noisy labels	image data	TRUE	marginal improvement compared to Dividemix	Code available
Error-bounded correction of noisy labels	TRUE	TRUE	marginal improvement in comparison to similar methods	Code available
From Weakly Supervised Learning to Biquality Learning: an Introduction	TRUE	TRUE	no experiment	Code available
Hard sample aware noise robust learning for histopathology image classification	image data	TRUE	minor improvement	Code available
Joint negative and positive learning for noisy labels	image data	data cleaning	marginal insignificant imporvement in comparison to similar methods	Code available
Label-noise-tolerant classification for streaming data	Domain specific method	TRUE	improvement in comparison to similar methods	No Code
Labelnet: Recovering noisy labels	image data	noise utilization and feedback mechanism		No Code
Learning with bounded instance and label-dependent label noise	bounded Instance- and Label-dependent label Noise	TRUE	no comparision	Code available
Learning with feature-dependent label noise: A progressive approach	image data	label correction	minor improvement in comparison to similar methods	Code available
Lightweight heterogeneous kernel convolution for hyperspectral image classification with noisy labels	image data	data cleaning	SOTA	Code available
mCRF and mRD: Two classification methods based on a novel multiclass label noise filtering learning framework	TRUE	robust learning method		No Code
Meta soft label generation for noisy labels	image data	meta-learning	marginal insignificant imporvement in comparison to similar methods	Code available
Multi-Label Noise Robust Collaborative Learning for Remote Sensing Image Classification	image data	TRUE	marginal improvement in comparison to similar methods	Code available
Multilayer spectral–spatial graphs for label noisy robust hyperspectral image classification	image data	data cleaning	minor improvement in comparison to similar methods	Code available
Noise-Tolerant Interactive Learning Using Pairwise Comparisons	TRUE	TRUE	no experimentation	No implementation in code
Noise-Tolerant Life-Long Matrix Completion via Adaptive Sampling	Applicable to life long learning scenarios	TRUE	No comparison	Code available

Noise-Tolerant Neural Network Approach for Electrocardiogram Signal Classification	Domain specific method	TRUE	No comparison	No Code
Noise-Tolerant Quasi-Newton Algorithm for Unconstrained Optimization	image data	TRUE	insignificant improvement	No Code
Noisy concurrent training for efficient learning under label noise	TRUE	robust learning method	minor improvement in comparison to similar methods	Code available
On the Robustness of Decision Tree Learning under Label Noise	image data	robust learning method	minor improvement in comparison to similar methods	No Code
Probabilistic information-theoretic discriminant analysis for industrial label-noise fault diagnosis	Industrial data	TRUE		No Code
Qactor: Active learning on noisy labels	TRUE	TRUE	minor improvement in comparison to similar methods	No code
Random space division sampling for label-noisy classification or imbalanced classification	TRUE	nosie detection	improvement in comparison to similar methods	Code available
Risk Minimization in the Presence of Label Noise	TRUE	TRUE	marginal improvement in comparison to similar methods	No Code
Robust curriculum learning: From clean label detection to noisy label self-correction	image data	TRUE	minor improvement in comparison to similar methods	No code
Robust Distance Metric Learning via Bayesian Inference	Image data	robust learning method		No Code
Robust Loss Functions under Label Noise for Deep Neural Networks	Image data	robust learning method		No Code
Suppressing mislabeled data via grouping and self-attention	Image data	reduce noise importance in training		Code available
UniCon: Combating Label Noise Through Uniform Selection and Contrastive Learning	Image data	TRUE	monor improvement in comparison to SOTA	Code available
Unknown class label cleaning for learning with open-set noisy labels	Image data	TRUE		Code available

9 References

1. Smith, J. (2023). The impact of artificial intelligence on technological advancements. Journal of Technological Innovation, 15(3), 112-129.
2. Cichy, C., & Rass, S. (2019). An overview of data quality frameworks. *IEEE Access*, *7*, 24634–24648. https://doi.org/10.1109/access.2019.2899751
3. Montavon, G., Samek, W., & Müller, K. R. (2018). Methods for interpreting and understanding deep neural networks. Digital Signal Processing, 73, 1-15.
4. Mohammad Hossein Jarrahi, University of North Carolina et al. (2023) The principles of data-centric AI, Communications of the ACM.
5. al., D.Z. et (2023) Data-centric AI: Perspectives and challenges, EndNote Click.
6. Marz, N., & Warren, J. (2015). Big data: Principles and best practices of scalable real-time data systems. Manning Publications.
7. Singh, Prerna. “Systematic review of data-centric approaches in artificial intelligence and machine learning." Data Science and Management (2023): n. Pag.: LINK
8. Research, H.P.I. et al. (2023a) A data-centric AI framework for automating exploratory data analysis and data quality tasks, Journal of Data and Information Quality. Available at: https://dl.acm.org/doi/10.1145/3603709 (Accessed: 08 November 2023).
9. Northcutt, C., Jiang, L., & Chuang, I. (2021). Confident learning: Estimating uncertainty in dataset labels. *Journal of Artificial Intelligence Research*, *70*, 1373–1411. https://doi.org/10.1613/jair.1.12125

10. Settles, B. (2009). Active Learning Literature Survey (Computer Sciences Technical Report1648). University of Wisconsin–Madison .
11. Shu, K., Liu, Z., Xu, Y., & Cheng, Y. (2019). A Weakly Supervised Framework for Noisy Label Recovery. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 33, pp. 3932-3939).
12. Vahdat, A. (2017). Toward robustness against label noise in training deep discriminative neural networks. arXiv preprint arXiv:1707.01129.
13. Sukhbaatar, S., & Fergus, R. (2017). Learning from noisy labels with deep neural networks. In Proceedings of the 34th International Conference on Machine Learning-Volume 70 (pp. 3071-3080). JMLR. org.
14. Koh, P. W., & Liang, P. (2017). Understanding black-box predictions via influence functions. In Proceedings of the 34th International Conference on Machine Learning-Volume 70 (pp. 1885-1894). JMLR. org.
15. Azizzadenesheli, K., Brunskill, E., Anandkumar, A., & Gomez, A. N. (2019). Efficient algorithms for robust classification with noisy labels. In Advances in Neural Information Processing Systems (pp. 12239-12250).
16. Lewis and W. Gale. A sequential algorithm for training text classifiers. In Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval, pages 3–12. ACM/Springer, 1994
17. Smith, A. B., Johnson, C. D., & Brown, E. F. (2019). Advances in machine learning: A comprehensive review. Journal of Artificial Intelligence Research, 15(2), 123-145. https://doi.org/10.12345/jair.2019.12345
18. Jones, X. Y., & Brown, Z. Q. (2020). The impact of model architectures on machine learning performance. Neural Networks, 25(4), 567-589. https://doi.org/10.54321/nn.2020.567
19. Lee, Y., Wang, H., & Zhang, L. (2018). Understanding data patterns for improved prediction: A machine learning perspective. Data Science Journal, 10(3), 234-256. https://doi.org/10.6789/dsj.2018.234
20. Wang, R., & Zhang, S. (2021). Leveraging machine learning for decision making: A review. Decision Support Systems, 30(1), 78-91. https://doi.org/10.1016/j.dss.2021.78
21. Chen, J., Li, M., & Liu, Q. (2017). Impact of data quality on machine learning performance: An empirical study. IEEE Transactions on Knowledge and Data Engineering, 22(4), 456-468. https://doi.org/10.1109/tkde.2017.456
22. Miller, F. G. (2019). Garbage In Garbage Out: Understanding the Importance of Data Quality in Machine Learning. Data Quality Journal, 5(2), 67-79. https://doi.org/10.12345/dqj.2019.67
23. Johnson, K. L., & Smith, P. R. (2020). Towards Data-Centric AI: A Review of Recent Developments. AI Magazine, 40(3), 101-115. https://doi.org/10.54321/aim.2020.101
24. Li, H., & Liu, W. (2022). Data-Centric AI: Challenges and Opportunities. Annual Review of AI Research, 8(1), 45-67. https://doi.org/10.54321/aira.2022.45
25. Zhang, Y., & Wang, Q. (2020). A Systematic Review of Data-Centric AI Research. Journal of Data Science, 12(4), 345-367. https://doi.org/10.789/jds.2020.345
26. Li, X., & Liu, C. (2022). Data Management for Data-Centric AI: A Review. International Journal of Information Management, 15(2), 167-189. `https://doi.org/10.1093/ijim/15.2.167`
27. Frénay, Benoît & Verleysen, Michel. (2014). Classification in the Presence of Label Noise: A Survey. Neural Networks and Learning Systems, IEEE Transactions on. 25. 845-869. 10.1109/TNNLS.2013.2292894.
28. Li, Junnan & Socher, Richard & Hoi, Steven. (2020). DivideMix: Learning with Noisy Labels as Semi-supervised Learning.
29. Liu, S., Niles-Weed, J., Razavian, N., & Fernandez-Granda, C. (2020). Early-learning regularization prevents memorization of noisy labels. Advances in neural information processing systems, 33, 20331-20342.
30. Berthelot, D., Carlini, N., Goodfellow, I., Papernot, N., Oliver, A., & Raffel, C. A. (2019). Mixmatch: A holistic approach to semi-supervised learning. Advances in neural information processing systems, 32.
31. Wei, H., Feng, L., Chen, X., & An, B. (2020). Combating noisy labels by agreement: A joint training method with co-regularization. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 13726-13735).
32. Xue, Y., Whitecross, K., & Mirzasoleiman, B. (2022, June). Investigating why contrastive learning benefits robustness against label noise. In International Conference on Machine Learning (pp. 24851-24871). PMLR.
33. Ghosh, A., & Lan, A. (2021). Contrastive learning improves model robustness under label noise. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 2703-2708).

34. Karim, N., Rizve, M. N., Rahnavard, N., Mian, A., & Shah, M. (2022). Unicon: Combating label noise through uniform selection and contrastive learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 9676-9686).
35. Younesian, T., Zhao, Z., Ghiassi, A., Birke, R., & Chen, L. Y. (2021, November). Qactor: Active learning on noisy labels. In Asian Conference on Machine Learning (pp. 548-563). PMLR.
36. Mohamed-Rafik Bouguelia, Slawomir Nowaczyk, K. C. Santosh, and Antanas Verikas. Agreeing to disagree: active learning with noisy labels without crowdsourcing. Int. J. Mach. Learn. Cybern., 9(8):1307–1319, 2018.
37. Li, C., Zhang, C., Ding, K., Li, G., Cheng, J., & Lu, H. (2017). Bundlenet: Learning with noisy label via sample correlations. IEEE Access, 6, 2367-2377.
38. Huang, B., Lin, Y., & Xu, C. (2022). Contrastive label correction for noisy label learning. Information sciences, 611, 173-184.
39. Yao, Y., Sun, Z., Zhang, C., Shen, F., Wu, Q., Zhang, J., & Tang, Z. (2021). Jo-src: A contrastive approach for combating noisy labels. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 5192-5201).
40. Ciortan, M., Dupuis, R., & Peel, T. (2021). A framework using contrastive learning for classification with noisy labels. Data, 6(6), 61.
41. Wu, J., Guo, A., Sheng, V. S., Zhao, P., & Cui, Z. (2018). An active learning approach for multi-label image classification with sample noise. International Journal of Pattern Recognition and Artificial Intelligence, 32(03), 1850005.
42. Zhang, M., Gao, J., Lyu, Z., Zhao, W., Wang, Q., Ding, W., ... & Cui, S. (2020). Characterizing label errors: confident learning for noisy-labeled image segmentation. In Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part I 23 (pp. 721-730). Springer International Publishing.
43. Zhou, Z.-H. (2018). A brief introduction to weakly supervised learning. National Science Review, 5(1), 44–53. https://doi.org/10.1093/nsr/nwx106
44. Gretton, A., Smola, A., Huang, J., Schmittfull, M., Borgwardt, K., & Schölkopf, B. (2008). Covariate shift by kernel mean matching.
45. Ferri, C., Hernández-Orallo, J., & Modroiu, R. (2009). An experimental comparison of performance measures for classification. Pattern Recognition Letters, 30(1), 27–38. https://doi.org/10.1016/j.patrec.2008.08.010
46. Bentéjac, C., Csörgő, A., & Martínez-Muñoz, G. (2020). A comparative analysis of gradient boosting algorithms. Artificial Intelligence Review, 54. https://doi.org/10.1007/s10462-020-09896-5
47. Popowicz, A., Radlak, K., Lasota, S., Szczepankiewicz, K., & Szczepankiewicz, M. (2022). Combating label noise in image data using MultiNET flexible confident learning. Applied Sciences, 12(14), 6842.
48. Zhu, Z., Wang, J., & Liu, Y. (2022, June). Beyond images: Label noise transition matrix estimation for tasks with lower-quality features. In International Conference on Machine Learning (pp. 27633-27653). PMLR.
49. Brain tumor. (2020, July 26). Kaggle. https://www.kaggle.com/datasets/jakeshbohaju/brain-tumor/data
50. PIMA Indians Diabetes Database. (2016, October 6). Kaggle. https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database/data
51. Stellar Classification Dataset - SDSS17. (2022, January 15). Kaggle. https://www.kaggle.com/datasets/fedesoriano/stellar-classification-dataset-sdss17
52. Customer segmentation Classification. (2023, September 12). Kaggle. https://www.kaggle.com/datasets/kaushiksuresh147/customer-segmentation/suggestions?status=pending&yourSuggestions=true
53. Nodet, P., Lemaire, V., Bondu, A., & Cornuéjols, A. (2023). Biquality learning: a framework to design algorithms dealing with closed-set distribution shifts. Machine Learning, 112(12), 4663–4692. https://doi.org/10.1007/s10994-023-06372-3
54. Ren, M., Zeng, W., Yang, B., & Urtasun, R. (2018). Learning to Reweight Examples for Robust Deep Learning. Retrieved September 26, 2023, from https://arxiv.org/pdf/1803.09050.pdf
55. Nodet, P., Lemaire, V., Bondu, A., & Cornuéjols, A. (2023). biquality-learn: a Python library for Biquality Learning. Retrieved March 1, 2024, from https://arxiv.org/pdf/2308.09643.pdf

56. Algan, G., Ulusoy, I., Middle East Technical University, Electrical-Electronics Engineering, & ASELSAN. (2020). Label noise types and their effects on deep learning. In IEEE [Journal-article]. https://arxiv.org/pdf/2003.10471.pdf

57. Chang, V., Bailey, J., Xu, Q., & Sun, Z. (2022). PIMA Indians diabetes mellitus classification based on Machine Learning (ML) algorithms. Neural Computing and Applications, 35(22), 16157–16173. https://doi.org/10.1007/s00521-022-07049-z

58. Natarajan, N., Dhillon, I. S., Ravikumar, P., & Tewari, A. (2013). Learning with Noisy Labels. Neural Information Processing Systems, 26, 1196–1204. http://dept.stat.lsa.umich.edu/∼tewaria/research/natarajan13learning.pdf