On the Impact of Noisy Labels on Supervised Classification Models*

Rafał Dubel[1], Agata M. Wijata[2,3] $^{[0000-0001-6180-9979]}$, and Jakub Nalepa[1,3] $^{[0000-0002-4026-1569]}$

[1] Faculty of Automatic Control, Electronics and Computer Science, Department of Algorithmics and Software, Silesian University of Technology, Akademicka 16, 44-100 Gliwice, Poland
jnalepa@ieee.org
[2] Faculty of Biomedical Engineering, Silesian University of Technology, Roosevelta 40, 41-800 Zabrze, Poland
awijata@ieee.org
[3] KP Labs, Konarskiego 18C, 44-100 Gliwice, Poland
{awijata, jnalepa}@kplabs.pl

Abstract. The amount of data generated daily grows tremendously in virtually all domains of science and industry, and its efficient storage, processing and analysis pose significant practical challenges nowadays. To automate the process of extracting useful insights from raw data, numerous supervised machine learning algorithms have been researched so far. They benefit from annotated training sets which are fed to the training routine which elaborates a model that is further deployed for a specific task. The process of capturing real-world data may lead to acqurring noisy observations, ultimately affecting the models trained from such data. The impact of the label noise is, however, under-researched, and the robustness of classic learners against such noise remains unclear. We tackle this research gap and not only thoroughly investigate the classification capabilities of an array of widely-adopted machine learning models over a variety of contamination scenarios, but also suggest new metrics that could be utilized to quantify such models' robustness. Our extensive computational experiments shed more light on the impact of training set contamination on the operational behavior of supervised learners.

Keywords: Supervised machine learning · binary classification · label noise · robustness

1 Introduction

The amount of acquired data grows tremendously in virtually all domains, spanning across medical imaging [22, 15], text analysis and categorization [7], speech

* AMW was supported by the Silesian University of Technology, Faculty of Biomedical Engineering grant (07/010/BK_23/1023). JN was supported by the Silesian University of Technology Rector's grant (02/080/RGJ22/0026).

recognition [20], predictive maintenance [8], and many others. Gathering such enormous amounts of data of different modalities, however, poses new practical challenges concerned with its automated analysis and exploitation using data-driven techniques. In *supervised machine learning* (ML), we benefit from the acquired training data coupled with ground-truth labels to build models that are deployed to process incoming observations in a plethora of classification and regression tasks. Although deep learning—which benefits from the automated representation learning paradigm—established the state of the art in a multitude of fields, classic ML techniques are still widely used and researched due to their simplicity, resource frugality (which is especially important while deploying them on e.g., edge devices [6]), enhanced interpretability [13], and reduced requirements on the amount of training data necessary to elaborate well-generalizing models, effectively dealing with the unseen data.

Independently of the type of an ML model, we need to face the problem of noise which may easily affect the training data (also, training sets may be weakly-labeled [12]). Such noise may have different sources—it can be a result of a human or a sensor error, incorrectly designed data acquisition process, wrongly interpreted data or even hostile actions [10]. In general, we distinguish two types of the data noise, being random or systematic, and influencing supervised learners: the (*i*) attribute (feature) noise, and (*ii*) the label noise [2]. In the former case, the features, corresponding to the observed objects in the training set are contaminated, whereas in the latter scenario, the class labels are mistakenly assigned to training examples, due to e.g., an incorrect data annotation process or human errors/bias. Here, the label noise commonly leads to more severe consequences, as it can directly mislead the learning process [21] resulting in random predictions [9], and it cannot be compensated by other (not noisy) features if others are contaminated with noise. Also, noisy training data commonly leads to overly complex models.

Understanding its impact on the capabilities of ML techniques, thus robustifying them against such unexpected data-level contaminations is of paramount practical importance. The vast majority of works concerning this issue focus on developing more comprehensive and computationally-intensive processing pipelines [14], coping with noisy ML datasets, through the identification or reduction of noise, and pruning such contaminated training samples [1, 2]. On the other hand, there are significantly less studies investigating the influence of noise on the "vanilla" versions of ML algorithms. Capturing the empirical behavior of supervised learners and understanding their intrinsic robustness against data-level noise can, however, influence the selection of an ML model for implementation, given the characteristics of the data acquisition and operation environment [17].

In this work, we address the above-discussed research gap, and thoroughly investigate an array of widely-adopted supervised ML classifiers trained from the datasets contaminated with different levels of label noise, in a variety of contamination scenarios (Section 2). Our extensive computational experiments (Section 3), performed over artificially-generated and benchmark datasets shed more light on the impact of the training set imbalance, cardinality and char-

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36021-3_8

acteristics on the overall performance of ML models trained from noisy labels. We belive that the results reported in our study can constitute an interesting point of departure for further research on robustifying classic ML learning models elaborated from large, imbalanced and (potentially) noisy training sets, hence on enhancing their practical utility in real-life data acquisition environments.

2 Materials and Methods

In this study, we aim at quantifying the impact of the class-label noise which contaminates the training set $\boldsymbol{T}$ (the contaminated training set is denoted as $\boldsymbol{T}'$) on the generalization capabilities of supervised learners, calculated over the unseen test set Ψ that is *not* affected by the label noise (Fig. 1). As we target the binary classification problems, we flip the class labels of randomly selected training examples to the opposite one (i.e., the positive-class label is swapped to the negative-class label, or vice versa) in our simulation process. To capture various real-world scenarios, we investigate the following simulations:

- **Uniform label noise**, in which the same percentage η of class labels are flipped in both classes (positive and negative).
- **Positive-class label noise**, in which a given percentage η of training vectors originally belonging to the positive class are swapped and become the negative-class examples. For simplicity, we assume that the positive class corresponds to the majority class.
- **Negative-class label noise**, in which a given percentage η of training vectors originally belonging to the negative class are swapped and become the positive-class examples. For simplicity, we assume that the negative class corresponds to the minority class.
- **Random label noise**, in which the class labels of a given percentage η of all training vectors are swapped. Here, the contaminated vectors are randomly drawn, without considering their original class labels.

The following levels of the label noise contamination are considered in this study: $\eta \in \{0\%, 5\%, 10\%, 20\%, 30\%, 40\%, 50\%, 70\%, 90\%\}$ (note that we target uncontaminated $\boldsymbol{T}$'s, as well as extremely noisy training sets). To understand the robustness of the most popular learners, we deploy the following models in our pipeline (which is independent of the ML model): k-Nearest Neighbors ($k = 5$), linear support vector machines ($C = 0.025$), Gaussian Process classifiers with the radial basis function kernel, Decision Trees (with the Gini impurity measuring the quality of the split, max. depth of the tree: 7), Random Forests (max. depth of the tree: 7, max. number of trees: 50), Multi-Layer Perceptron (MLP) classifier (with rectified linear unit activations), AdaBoost classifier (max. number of estimators: 100), and the Quadratic Discriminant Analysis-based models [18].

To investigate the classification capabilities of the learners, we quantify their performance over the unseen (uncontaminated) Ψ's using classic metrics such as accuracy (Acc), sensitivity (Sen), specificity (Spe), F1-score [19] and the

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36021-3_8

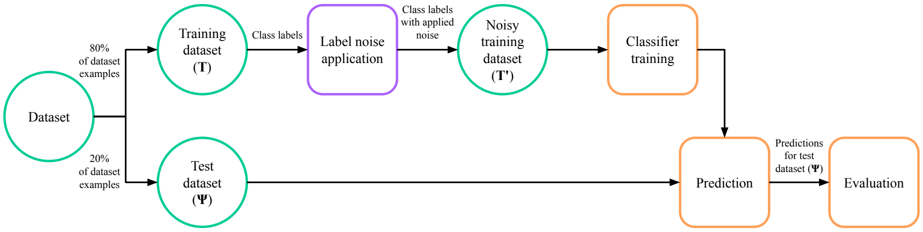


Fig. 1. A high-level flowchart of our computational experiments, in which a dataset is split into training and test subsets ($\boldsymbol{T}$ and Ψ, respectively, containing e.g., 80% and 20% of all training examples). The ML models are trained using a contaminated training set ($\boldsymbol{T}'$), and their performance is quantified over the uncontaminated Ψ.

Matthews correlation coefficient (MCC), with MCC commonly used for imbalanced classification, as the most robust quality metric [5]. The MCC values range from -1 (very strong negative relationship between ground-truth labels and prediction) to 1 (very strong positive relationship between them), whereas all other metrics range from zero to one, with one corresponding to the perfect classification performance. Additionally, we propose two auxiliary metrics:

- **D1**—it quantifies the stability of the ML model trained from contaminated $\boldsymbol{T}$'s ($\eta = 0\% - 40\%$). Here, the "robustness" is calculated as the mean standard deviation of the obtained MCC scores (the smaller, the better).
- **D2**—it is calculated as the mean MCC score obtained by the ML model trained from contaminated $\boldsymbol{T}$'s with $\eta = 0\% - 40\%$ (the larger, the better).

The models were trained and validated on (i) synthesized and (ii) almost 40 benchmark datasets, the latter acquired from the KEEL and sklearn repositories, and manifesting different imbalance ratio across positive- and negative-class examples (Table 1). The synthetic datasets were generated using the `make_moons`, `make_circles`, `make_blobs` and `make_classification` sklearn functions with various parameterizations, concerning the number of training examples (ranging from 100 up to 5000), and the number of features (up to 60 features, with and without redundant ones). The majority of the KEEL/sklearn benchmark datasets include below 1000 training examples (with a few having above 5000 of them), with the mean and median of 1976 and 611 training vectors, respectively. The majority of benchmarks have up to 50 features (attributes). For full details of the investigated datasets, see the supplement available at `https://gitlab.com/agatawijata/impact-of-noisy-labels`.

Table 1. The aggregated imbalance ratio of all benchmarks (KEEL and sklearn).

Minimum	Mean	Median	Maximum
1.00	1.37	1.22	1.88

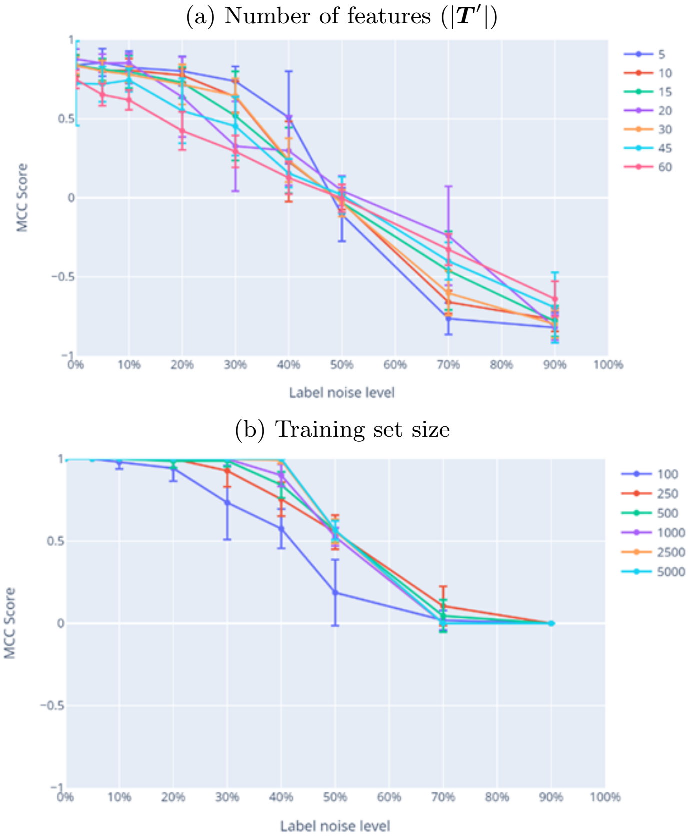
ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36021-3_8

3 Experimental Results

The experiments were split into those focused on investigating all models over the (i) synthetic and (ii) benchmark (KEEL/sklearn) datasets (all experiments, for all ML models and datasets were executed in 10 independent runs, and the results were aggregated). The results obtained over the simulated datasets were consisted across all ML models, and showed that their robustness increases with the larger number of training vectors, as the uncontaminated examples were able to effectively compensate those affected by the label noise (see Fig. 2; for brevity, we present the MCC scores—all other metrics are available in the supplementary material). Similar, albeit not as obvious observations may be drawn for the models trained over $\boldsymbol{T}'$'s with varying numbers of features. Here, the robustness of the models tends to increase for larger dimensionalities of the dataset, but—especially for smaller $\boldsymbol{T}'$'s, the generalization capabilities can drop, due to the inherent curse of dimensionality issues (as the effective number of "correct" training examples is further decreased once they are contaminated with the class-label noise). We can hypothesize, however, that increasing the number of training examples could make the models more robust against feature-level noise here—this requires further investigation. Overall, the experiments over the synthetic data showed that Gaussian Process, AdaBoost, MLP and SVM classifiers offer the best generalization while trained over contaminated $\boldsymbol{T}'$'s.

The results obtained over all benchmark datasets are rendered in Fig. 3, presenting the D1 and D2 metrics quantifying the "robustness" of the models against different noise contamination scenarios (D1 should be minimized, whereas D2—maximized; note the reversed D1 axis). We can observe that the MLP classifier elaborates the best aggregated prediction quality for the uniformly and randomly (across classes) applied class-label noise (the largest D2), with the linear support vector machines offering the best stability of predictions (reflected in the lowest D1 values). It is of note that the label noise applied to separate classes (either the positive or negative, corresponding to the majority and minority ones) led to consistent behavior of the investigated machine learning models. However, for the datasets with a larger imbalance ratio (commonly larger than 1.8), contaminating the minority classes triggered a visible drop in the MCC scores. This can be attributed to the fact that contaminating the minority class with noise makes the learning process much more challenging, as the intrinsic properties of the minority-class examples may be lost or severly modified through the noise injection. On the other hand, contaminating the majority class with noise may indeed act as an additional regularizer—there are techniques which exploit the noise injection to synthetically generate training examples, e.g., in the context of augmenting training sets for deep learning models targeting the hyperspectral image classification and segmentation [16]. In their recent work, Beinecke and Heider showed that deploying Gaussian Noise Up-Sampling, which effectively selects the minority-class examples and adds noise to the data points in order to smooth the class boundary can indeed reduce overfitting in some clinical decision making tasks as well [3]. Also, it worked on par with well-established the Synthetic Minority Over-sampling [4] and Adaptive Synthetic Sampling [11] ap-

DOI: 10.1007/978-3-031-36021-3_8


(a) Number of features ($|\boldsymbol{T}'|$)

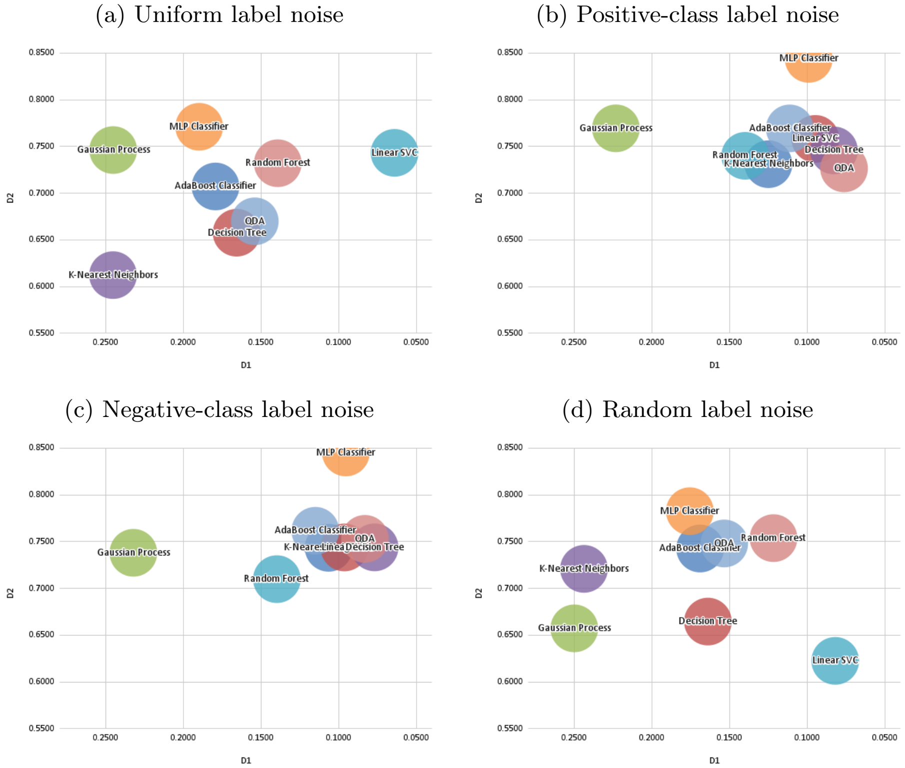
(b) Training set size

Fig. 2. The MCC scores obtained by the Gaussian Process classifier once the training set was contaminated by (a) the uniform label noise (for different numbers of features simulated using the `make_classification` dataset, and by (b) the positive-class label noise for different numbers of training set examples generated using the `make_circles` function. The number of features and training set examples is given in the legend.

proaches, and even outperformed those algorithms on selected datasets, showing the potential of utilizing noise simulations in training set augmentation routines.

4 Conclusions

Training supervised learners from noisy data has become an important practical issue, given the amount of data acquired on a daily basis. Such data may be contaminated with feature and class label noise due to various reasons, ranging from the operator bias, incorrect acquisition process or malfunctioning of the sensory system. Such noisy data, however, directly affects supervised learners trained from such data. Understanding the robustness of ML models against class label


(a) Uniform label noise

(b) Positive-class label noise

(c) Negative-class label noise

(d) Random label noise

Fig. 3. The scatter plots of our D1 and D2 metrics quantifying the robustness of all investigated models over all benchmark datasets and noise contamination scenarios (D1 should be minimized, whereas D2—maximized).

noise remains under-researched—we tackled this research gap, and thoroughly investigated an array of established models, following a variety of noise contamination scenarios. On top of that, we proposed new metrics that can be utilized to quantify the robustness of ML models against various levels of noise. Our extensive experiments, performed over synthetic and benchmark datasets revealed that there are indeed ML models which are more robust against label noise (the robustness directly depends on the size of the training set). We believe that the results reported in this work can constitute an exciting departure point for further research focused on developing the noise-robust models, and on objectively quantifying their robustness against real-life data acquisition conditions.

References

1. Awasthi, P., Balcan, M.F., Haghtalab, N., Urner, R.: Efficient learning of linear separators under bounded noise (2015)

ICCS Camera Ready Version 2023
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-36021-3_8

2. Balcan, M.F., Haghtalab, N.: Noise in classification (2020)
3. Beinecke, J., Heider, D.: Gaussian noise up-sampling is better suited than SMOTE and ADASYN for clinical decision making. BioData Mining **14**(1), 49 (Nov 2021)
4. Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P.: SMOTE: Synthetic Minority over-Sampling Technique. J. Artif. Int. Res. **16**(1), 321–357 (jun 2002)
5. Chicco, D., Jurman, G.: The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. BMC Genomics **21** (2020)
6. Dhar, S., Guo, J., Liu, J.J., Tripathi, S., Kurup, U., Shah, M.: A survey of on-device machine learning: An algorithms and learning theory perspective. ACM Trans. Internet Things **2**(3) (2021)
7. Duarte, J.M., Berton, L.: A review of semi-supervised learning for text classification. Artificial Intelligence Review (2023)
8. Es-sakali, N., Cherkaoui, M., Mghazli, M.O., Naimi, Z.: Review of predictive maintenance algorithms applied to hvac systems. Energy Reports **8**, 1003–1012 (2022)
9. Frenay, B., Verleysen, M.: Classification in the presence of label noise: A survey. IEEE TNNLS **25**(5), 845–869 (2014)
10. Gupta, S., Gupta, A.: Dealing with noise problem in machine learning data-sets: A systematic review. Procedia Computer Science **161**, 466–474 (2019)
11. He, H., Bai, Y., Garcia, E.A., Li, S.: ADASYN: Adaptive synthetic sampling approach for imbalanced learning. In: Proc. IEEE WCCI. pp. 1322–1328 (2008)
12. Kawulok, M., Nalepa, J.: Towards robust SVM training from weakly labeled large data sets. In: Proc. IAPR ACPR. pp. 464–468 (2015)
13. Kotowski, K., Kucharski, D., et al.: Detecting liver cirrhosis in computed tomography scans using clinically-inspired and radiomic features. Computers in Biology and Medicine **152**, 106378 (2023)
14. Leung, T., Song, Y., Zhang, J.: Handling label noise in video classification via multiple instance learning. In: Proc. IEEE ICCV. pp. 2056–2063 (2011)
15. Nalepa, J., Kotowski, K., et al.: Deep learning automates bidimensional and volumetric tumor burden measurement from MRI in pre- and post-operative glioblastoma patients. Computers in Biology and Medicine **154**, 106603 (2023)
16. Nalepa, J., Myller, M., Kawulok, M.: Training- and test-time data augmentation for hyperspectral image segmentation. IEEE Geoscience and Remote Sensing Letters **17**(2), 292–296 (2020)
17. Nettleton, D.F., Orriols-Puig, A., Fornells, A.: A study of the effect of different types of noise on the precision of supervised learning techniques. Artificial Intelligence Review **33**(4), 275–306 (2010)
18. Pedregosa, F., et al.: Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research **12**(85), 2825–2830 (2011)
19. Powers, D.: Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation (2020)
20. Pradana, W.A., Adiwijaya, Wisesty, U.N.: Implementation of support vector machine for classification of speech marked hijaiyah letters based on mel frequency cepstrum coefficient feature extraction. Journal of Physics: Conference Series **971**(1), 012050 (2018)
21. Sáez, J.A., Galar, M., Luengo, J., Herrera, F.: 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 (2014)
22. Wijata, A.M., Nalepa, J.: Unbiased validation of the algorithms for automatic needle localization in ultrasound-guided breast biopsies. In: Proc. IEEE ICIP. pp. 3571–3575 (2022)

DOI: 10.1007/978-3-031-36021-3_8