Cost-Aware Ensemble Learning Approach for Overcoming Noise in Labeled Data

Abdulrahman Gharawi, Jumana Alsubhi and Lakshmish Ramaswamy
School of Computing, University of Georgia, Athens, U.S.A.

Keywords: Machine Learning, Deep Learning, Ensemble Learning, Label Noise, Class Label Noise, Labeling Cost Optimization, Mislabeled Data.

Abstract: Machine learning models have demonstrated exceptional performance in various applications as a result of the emergence of large labeled datasets. Although there are many available datasets, acquiring high-quality labeled datasets is challenging since it involves huge human supervision or expert annotation, which are extremely labor-intensive and time-consuming. Since noisy datasets can affect the performance of machine learning models, acquiring high-quality datasets without label noise becomes a critical problem. However, it is challenging to significantly decrease label noise in real-world datasets without hiring expensive expert annotators. Based on extensive testing and research, this study examines the impact of different levels of label noise on the accuracy of machine learning models. It also investigates ways to cut labeling expenses without sacrificing required accuracy.

1 INTRODUCTION

Machine learning has shown outstanding performance in a variety of applications since the recent emergence of large-scale datasets. This success depends on the availability of large amounts of labeled data (Krizhevsky et al., 2017) and (Li et al., 2017), which is both expensive and time-consuming (Nguyen et al., 2015). There are several techniques presented in the literature to reduce the high labeling cost by using non-expert annotators on Amazon's Mechanical Turk (Nguyen et al., 2015); however, the use of non-experts often results in erroneously labeled data, commonly referred to as noisy labels (Nguyen et al., 2015). The percentage of incorrect labels has been observed to range from 5% to 38% in real-world datasets (Song et al., 2019).

Training supervised machine learning algorithms are known to be sensitive to noisy labels since it is assumed that the training dataset is correctly labeled (Krizhevsky et al., 2017),(Li et al., 2017), and (Song et al., 2019). Noisy labels can negatively affect the performance of ML models more than any other type of noise (Song et al., 2019). Furthermore, they can impact the structure of the models and the time needed to train the classifiers (Garcia et al., 2015). In addition, ML models can even learn on corrupted labels, and it fails to generalize the model (Krogh and Hertz, 1991). Batch normalization, dropout, and data augmentation (Perez and Wang, 2017) have been utilized to overcome the overfitting issue (Perez and Wang, 2017), but with noisy datasets, they did not completely overcome the overfitting issue (Lee et al., 2018).

Since real-world datasets contain label noise by its nature and it is challenging to eliminate label noise completely, utilizing an expert annotator is considered to minimize label noise (Nguyen et al., 2015) and (Aslan et al., 2017). Unfortunately, in most domains, expert annotators are expensive, limited, and many projects cannot afford them (Aslan et al., 2017). Therefore, some proposals include designing annotating techniques with non-experts to acquire high-quality labeled datasets at a considerably lower cost (Nguyen et al., 2015). Other recent studies try to overcome label noise by removing noisy records (Huang et al., 2019). Even so, it is challenging to predict whether or not a record label has noise. Other literature introduces machine learning models that are more robust to a noisy dataset (Xiao et al., 2015) and (Song et al., 2019). Despite the techniques introduced in the literature to obtain high-quality labeled datasets at a cheaper cost, obtaining a clean and accurately annotated dataset remains challenging and costly (Aslan et al., 2017). In addition, reducing the label noise may not help the machine learning model to generalize or

Gharawi, A., Alsubhi, J. and Ramaswamy, L.
Cost-Aware Ensemble Learning Approach for Overcoming Noise in Labeled Data.
DOI: 10.5220/0011675200003393
In *Proceedings of the 15th International Conference on Agents and Artificial Intelligence (ICAART 2023) - Volume 3*, pages 381-388
ISBN: 978-989-758-623-1; ISSN: 2184-433X
Copyright © 2023 by SCITEPRESS – Science and Technology Publications, Lda. Under CC license (CC BY-NC-ND 4.0)

acquire higher accuracy, but it just increases the labeling cost.

Based on extensive experimentation and analysis, this study investigates the influence of various degrees of label noise on the accuracy of machine learning models. Also, we examine the the trade off between labeling cost and accuracy. Furthermore, we propose a cost-effective method to improve the accuracy when there is high label noise. To mimic a noisy real-world dataset. Firstly, we adopt a common approach for adding synthetic label noise to a training set of available benchmark datasets and keeping the test set clean. Next, we evaluate each machine learning algorithm's robustness to noise with distinct levels of label noise. Finally, we discuss cost-effective learning and the best way to select the annotator level of expertise depending on the machine learning task and the budget. This paper makes the following contributions:

- Investigate the effect of class labels noise on some ML models to minimize the cost of annotation by determining the robustness of different ML models towards label noise.
- Explore the trade off between labeling cost and accuracy when the size of dataset and the level of noise change.
- Introduce a cost-effective method using ensemble learning to improve accuracy even in the presence of significant label noise.

The rest of this paper is organized as follows: Section 2 surveys related work and Section 3 discusses the design of the experimental study. The results and limitations are explored in Section 4 and Section 5, respectively. In Section 6, we conclude and discuss our plans for future work.

2 MOTIVATION AND RELATED WORK

Correctly labeled datasets are essential for supervised learning to build a reliable model. Machine learning algorithms typically assume that the dataset is labeled correctly. However, obtaining a reliable labeled dataset is expensive and difficult, leading to a noisy dataset and unreliable models. Crowdsourcing from non-experts and web annotations are two low-cost yet ineffective options for gathering annotations on a big scale. These two options were widely used for image data, where tags and online search terms are recognized as acceptable labels. Unfortunately, both of these options frequently introduce incorrect or noisy labels. Furthermore, the inclusion of noise in a dataset's class label degrades its quality and may reduce classifier prediction accuracy. Therefore, there is an increasing interest in obtaining high-quality datasets with minimum amount of noises for image and text processing applications. In the literature, many methods for training machine learning in the presence of noisy labels have been proposed. However, the effect of mislabeled data and class label noise did not receive appropriate consideration in terms of ML structure and annotator cost. This paper explores the effect of label noise on machine learning and deep learning algorithms with different datasets sizes and labeling costs.

Various strategies and techniques recently introduced in the literature to handle class label noise in datasets (Huang et al., 2019). These techniques try to identify the noisy label and remove the record from the dataset before the training. However, it is hard to identify the noise label. The other suggested technique in literature is to design a noise tolerant machine learning models (Xiao et al., 2015) and (Garcia et al., 2015). On the contrary, in this paper, we study the effect of the label noise on different machine learning algorithms to optimize the cost of labeling. We study the effect of label noise on different machine learning models such as SVM, KNN, MLP, CNN, DT, and RF.

Support vector machines (SVM) are capable of handling both classification and regression problems. The decision boundary for this method is the hyperplane, which must be determined. A decision plane is required to divide a collection of objects into their classes. In other words, SVM searches for a hyperplane in a high-dimensional feature space that has the greatest potential distance between two classes of data. The goal of SVM is to correctly identify the objects using examples from the training data set (Sitawarin and Wagner, 2019).

K-nearest neighbors (KNN) is a straightforward supervised machine learning algorithm that can be used to address classification and regression problems. KNN attempts to identify the k training samples that are closest to a new element before predicting the label of that new element using the k-nearest points. Calculating the distance can be done using any similarity function. Despite being straightforward, KNN frequently performs well in classification scenarios with very irregular decision boundaries (Sitawarin and Wagner, 2019).

The Multilayer perceptron (MLP) is fully connected multi-layer neural network. It is a model for the nonlinear mapping of an input vector to an output vector. The weights and output signals connecting the nodes are a function of the total of the node's inputs, as adjusted by a straightforward nonlinear activation function. MLP can learn through training. A collection of training data is needed, which is made up of input and corresponding output vectors. The training data is repeatedly fed into the multilayer perceptron, and the weights in the network are changed until the desired input—output mapping is achieved (Gardner and Dorling, 1998).

Convolutional neural networks (CNNs), which fall under the Deep Neural Networks category, are widely used in conjunction with two-dimensional inputs like such as images. Millions of photos can be used as inputs to enable them to learn about thousands of different objects. By altering the depth and breadth of the model, CNN's learning capacity can be changed. The convolutional layer is a crucial component of the CNN architecture that comprised of learnable filter banks that are active when a certain feature is discovered (Durga et al., 2019).

Decision trees (DT) are a form of supervised machine learning in which the training data is continually divided based on a particular parameter, by describing the input and the associated output. Decision nodes and leaves are the two components that can be used to explain the tree. The leaves represent the final decisions whereas the decision nodes are where the data is split (Su and Zhang, 2006).

The random forest algorithm (RF) is made up of a variety of independent decision trees. RF employs two random selection processes to build a single decision tree: the first is the random selection of training samples, and the second is the random selection of the sample's characteristic features. When all the decision trees have been built, equal-weight voting is used to determine the final classification decision (Ren et al., 2017).

Recently, many machine learning problems are successfully solved using ensemble techniques. Such techniques include training multiple models and integrating their predictions to increase the predictive performance of a single model. In ensemble learning (EL), an inducer or base-learner uses a series of labeled instances as input to create a model such as a classifier that generalizes these examples. Decision tree, neural network, and linear regression model are some examples of machine learning algorithms that can work as inducers. The fundamental idea behind ensemble learning is that by merging many models, the faults of one inducer will most likely be made up for by other inducers, improving the ensemble's overall prediction performance. A straightforward but efficient method for creating an ensemble of independent models is bagging, in which each inducer is trained using a replacement sample of instances drawn from the original dataset. Each sample typically has the same number of examples as the original dataset in order to guarantee a sufficient number of cases per inducer. The final prediction of an unknown instance is decided by majority voting among the inducers' predictions (Sagi and Rokach, 2018). In this paper, we study the performance of ensemble learning in the presence of label noise by exploring different combinations of the aforementioned machine learning models. The results indicate that ensemble learning is a cost-effective approach for overcoming noise in labeled data.

3 DESIGN OF EXPERIMENTAL STUDY

In this empirical study, we add the noise to the class label in two different ways: (i) random class flipping and (ii) flipping to one of the most three classes human labelers will get confused with. To optimize each model's hyperparameters Ray Tune with HyperOpt-Search was used (Krogh and Hertz, 1991). All models were trained with a batch size of 64. For each time we increase the label noise, we train separate models with different learning rates ranging from 0.01 to 0.9 and pick the learning rate that results in the best performance. Also, generally, we observe that the higher the label noise, the lower the optimal learning rate.

We used MNIST, Flowers, Adult, Breast Cancer Wisconsin, and IMDB reviews datasets in these experiments to show the effect of the label noise on different machine learning algorithms and to optimize the labeling cost.

Adult: The adult dataset, also known as the Census Income dataset from the UCI Machine Learning repository (Dua and Graff, 2017), consists of 48,842 entries extracted from the US Census database with 16 columns and 14 attributes. After cleaning the dataset, 7 percent was deleted because it contained missing values. The remaining 44,5222 entries contain 24.78 percent for the income less than 50K and 75.22 percent for equal or more than 50K. The task is to predict income levels based on the individual's personal information.

The Breast Cancer Wisconsin (WBC): This dataset also was retrieved from the UCI Machine Learning Repository (Dua and Graff, 2017). It contains 569 instances with ten features computed from digitized images of a breast mass. Each feature value is recorded with four significant digits. There are two classes, which are benign with 357 instances and 212 malignant instances.

MNIST: This dataset is being used as a benchmark for classification algorithms (Deng, 2012). The MNIST dataset contains 70,000 images, 28x28 pixels each of handwritten digits between 0 and 9. The training set contains 60,000 images, and the test set includes 10,000 images.

IMDB Reviews: It contains 50,000 reviews for movies. The review in a textual format and class labels are 0 for negative or 1 for positive reviews.

Flowers: This dataset contains 4242 images of 5 different types of flowers chamomile, tulip, rose, sunflower, and dandelion. Each class contains almost 800 images. The size of each image is about 320x240 pixels, but some images have lower resolution than the 320x240 pixels, so we resized all images to 300x200.

3.1 Trade off Between Labeling Cost and Accuracy

This section aims to answer the following question: What are the trade off between labeling cost and accuracy? Table 3 shows the cost of labeling by humans with various expertise levels ranging from non-experts to experts in the field. The cost of labeling increases as the level of expertise increases. Labeling the data with expert labelers would result in less noise and higher accuracy. However, we can determine a suitable ML model, which results in a likable accuracy even if the data has some noise. Some ML models' accuracy can be considerably impacted by label noise (Song et al., 2019). On the other hand, there are some ML models that can cope with label noise because they are more resilient and less sensitive to label noise (Krizhevsky et al., 2017).

Multiple experiments have been conducted using several ML models with different datasets in order to investigate the resilience of various ML models towards label noise. Through intensive investigation of the robustness of various ML models toward mislabeled data, it has been shown that there are some ML models that are more resilient towards label noise.

We studied the effect of label noise on different machine learning models such as SVM, KNN, MLP, CNN, DT, and RF using MNIST, Flowers, Adult, Breast Cancer Wisconsin, and IMDB reviews datasets. To estimate the accuracy when hiring different labelers, we can assume that the dataset with different levels of label noise is obtained from different labelers ranging from expert labelers with 0-5% of noise to non-expert labelers with around %5-30% of noise. Based on the accuracy obtained in the experiments we conducted, we can determine the optimal labeler for each dataset based on the cost and desired accuracy.

In each experiment, the degree of noise is increased from 0 to 30% to demonstrate how the ML responds as the noise level increases. We execute each experiment with 100%, 50%, and 20% of the dataset in order to determine how well ML models perform when there is label noise. Additionally, each experiment was run 30 times, with an average taken to obtain a rough estimate. Since knowledge experts are expensive, obtaining a huge, precisely labeled dataset on a low budget is difficult. However, not all machine learning models need a perfect dataset to achieve higher accuracy. The results in section 4.1 show how well ML models perform when trained on a big, noisy dataset against a smaller, clean dataset. Moreover, in section 4.1, we will discuss the impact of various levels of label noise on 100, 50, and 20 percent of the same dataset, and we will assess the trade-off between dataset size and cleanness to compare the labeling cost for each size.

3.2 Using Ensemble Learning to Reduce Labeling Cost

Ensemble learning combines several individual models to obtain better performance. In other words, an ensemble can be considered a learning technique where many models are joined to solve a problem. This is done because an ensemble tends to perform better than singles improving the generalization ability. In ensemble learning, predictions from various neural network models are combined to lower prediction variance and generalization error (Wang et al., 2014), (Alsubhi et al., 2021), and (Wang et al., 2014). This machine learning paradigm, where multiple learners are trained to solve the same problem, have shown its ability to make an accurate prediction from multiple machine learning in classification problem (Sagi and Rokach, 2018).

This section aims to answer the following question: can ensemble learning perform better than a single model in the presence of label noise? Also, which machine learning model combinations can cope better with mislabeled? Ensemble learning is being used in the literature to increase task prediction accuracy in many fields like Health Care, Speech, Image Classification, Forecasting, and Others (Wang et al., 2014), (Moon et al., 2020), (Dong et al., 2020), and (Yu et al., 2008).

Our ensemble approach has only three classifiers which are combined to operate the ensemble. The models that we used in this experiment are CNN, MLP, DT, RF, KNN, and SVM using MNIST, Flowers, Adult, Breast Cancer Wisconsin, and IMDB reviews datasets. The best three combinations of these machine learning models are presented in section 4.2.

The amount of noise introduced increases from 0 to 30% in each experiment to show how the ensemble learning performs each time noise increases.

4 RESULTS AND DISCUSSION

In this section, we empirically examine the effect of noisy datasets on ML performance, accuracy, and the labeling cost associated with each annotator. Firstly, we simulate class label noise by randomly flipping the class label from its actual label to any other class label. Secondly, we gradually increase the label noise. Then, we compare the annotators cost with the labeling noise.

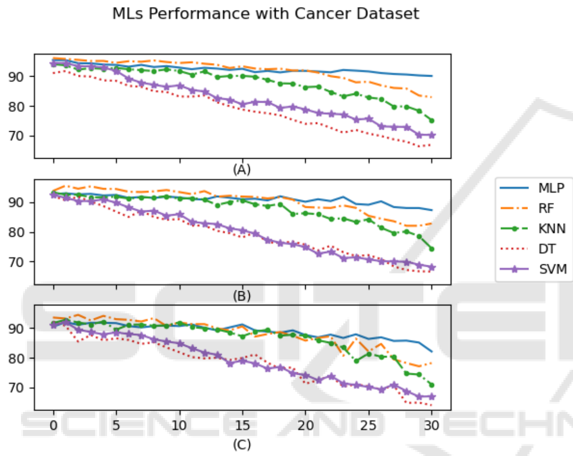

Figure 1: Effect of label noise on MLP, RF, SVM, DT, and KNN with WBC dataset (A) 100% of dataset, (B) 50% of dataset and (C) 20% of the dataset.

The result in this experiment indicates that ML models can be affected differently depending on the dataset. For instance, random forest performance dropped just 0.14% for each 1 percent of noise only with MNIST, but with other datasets, the accuracy dropped around 0.5% , Table 1. Furthermore, with different size of the same dataset the effect is significant as Figure 1 shows. This indicates that the dataset's size and type play an essential role in terms of accuracy and generalizing the model (Lee et al., 2018). Therefore, some machine learning models show more resilience to noise than others, depending on the dataset type and the task.

4.1 Experiment 1

Many machine learning and natural language processing tasks require human-labeled data. It is critical for the ML model to have a high-quality dataset with less label noise on the training set, since falsely labeled data affects machine learning models. Unfortunately, obtaining such a high-quality dataset with low labeling noise is usually quite expensive and necessitates the assistance of a domain expert (Krizhevsky et al., 2017).

Table 1: Average Accuracy Drop.

Dataset	MLP	RF	DT	CNN	SVM	KNN
MNIST	0.59	0.14	1	0.41	0.70	0.19
WBC	0.17	0.43	0.87	-	0.74	0.63
Adult	0.73	0.38	0.94	-	0.70	0.61
Flower	0.91	0.57	-	0.51	-	-

This experiment shows that we can simultaneously optimize the cost of labeling and obtain desirable accuracy. We used the Human Intelligent Tasks (HITs) prices presented in (Krizhevsky et al., 2017) and (Lee et al., 2018) to acquire human knowledge for labeling datasets using Amazon Mechanical Turk (MTurk). According to (Nguyen et al., 2015) and (Feng et al., 2009), the price can be between $1.5 for a non-expert to $150 knowledge expert, which is $0.015 to $1.5 per image. By evaluating the performance of each machine learning models with labeler price shown in Table 3, we can identify the cost needed to achieve the desired accuracy in each machine learning models.

Precisely, we measure: (1) the performance of each classifier with acquired labels from the annotators on the Table 3 on the test set, and (2) the total labeling cost by each annotator. Figure 1 presents the performance of each classifier depending on the accuracy of the labeling for the Cancer dataset and MNIST dataset, respectively. For the most part, when the level of noise increases as a result of using non-expert annotators, we can see that most machine learning models' performance decrease. In other words, the performance of machine learning models increases when the price of labeling increases. However, some machine learning can have better accuracy than others with the same or even more label noise. For instance, SVM can perform better than MLP with the $12,600 labeling cost, but with $14,000 MLP performs better than SVM. Likewise, Random Forest outperforms CNN when the labeling cost is $1,400, but after increasing the labeling cost to $13,000, CNN exceeds Random Forest's performance.

Total Cost = Number of images × Annotator Cost per image

This indicates that with low labeling cost (labeled by non-expert where the data can include label noise) and by choosing the right ML model, we can achieve desirable performance. Figures 1 shows that with up to 10% of added noise, some ML is slightly affected in some datasets than other. For instance, the effect

SCITEPRESS SCIENCE AND TECHNOLOGY PUBLICATIONS

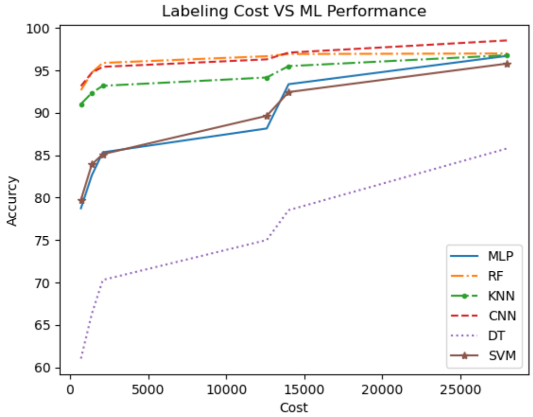

Figure 2: Machine learning algorithm cost with different labeling accuracy using MNIST dataset.

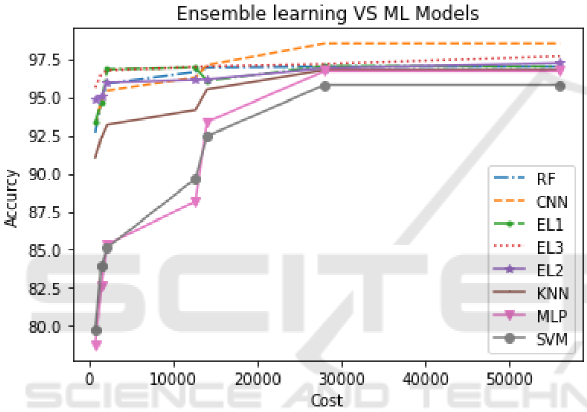

Figure 3: Voting VS Machine learning algorithm costs.

of 10 percent label noise on CNN with the MNIST dataset is just 3 percent. On the other hand, the effect on MLP with the same dataset and percentage of noise is more than 6 percent. Let us assume that the MNIST dataset is unlabeled, and it needs a human annotator. Choosing an expert annotator is very expensive as Figure 2 shows. The cost of labeling the dataset is 28,000 dollars if we choose the experts with only 5 percent noise. In this case, the accuracy is 96 percent when using CNN and 94 percent using MLP. On the contrary, if we labeled the same dataset by an expert with 10 percent noise, the cost is only 14,000 dollars. The accuracy using CNN is 95 percent, and MLP is 92 percent, 2% less than when the label noise is 5%. This shows that we can trade off 2% accuracy and decrease the labeling cost dramatically.

Therefore, depending on the ML task, we can determine if we need to invest more money in the labeling to increase accuracy. In a manufacturing classification problem, for example, assuming that the task is to label the product images as damaged or good. The importance of eliminating the noise will be different depending on the cost of the product. If the cost of the product is 1 dollar, then 10 percent of noise would cost much less than a 50-dollar product. If the cost of misclassification is not crucial, then we can cost-effectively label the data even if it has some noise.

Additionally, by comparing the performance of ML models, with more miniature dataset sizes, the accuracy is lower with the same level of noise. When we deduce the dataset's size to 20%, the accuracy dropped dramatically even with the same level of noise as shown in Figure 1. Nevertheless, expert annotators are expensive, and many projects cannot afford them. For example, the cost of expert annotators can be 30 times more than the cost of non-expert annotators, as Figure 2 shows. By analyzing the results in Figure 2, ML models with large noisy datasets perform better than smaller datasets. For instance, with 20% of the MNIST dataset labeled by an expert labeler costing $11,200, ML models can only achieve 90% accuracy. On the other hand, if we used non-expert labeler to label the entire dataset, we can obtain similar accuracy with only $224. Therefore, while it is essential to have accurately labeled data, the size of the dataset is more valuable to generalize the machine learning model and have an accurate classification.

4.2 Experiment 2

As experiment 1 shows, some algorithms showed more resilience to label noise than others. In this section, we analyze the performance of ensemble learning with different configurations against different levels of label noise. We used three different machine learning algorithms presented in section 2 and combined them together in the voting structure to find the most robust combination. Each model is trained separately and then combined for a hard voting ensemble. For consistency, we use the same hyperparameter optimization framework to get the best configurations in each model. Correspondingly, we implement the same level of in each model. We increase the noise from 0 to 30 percent with each ensemble learning to investigate the model robustness and the sensitivity by examining the average accuracy drop with each 1 percent class noise.

The experimental results presented in Table 2 show the labeling cost associated with different levels of label noise as well as the accuracy of various ML models, including three combinations of ensemble learning models, which are (RF, KNN, and MLP), (CNN, KNN, and SVM), and (CNN, KNN, and MLP). We can see that the labeling cost for the MNIST dataset that has 70,000 images is $56,000,

Table 2: Comparison between ensemble learning and traditional machine learning in terms of the labeling cost.

Cost/item	**Noise**	**Cost**	**EL 1** [a]	**EL 2**[b]	**EL 3**[c]	**CNN**	**DT**	**RF**	**KNN**	**SVM**	**MLP**
0.01	30	700	93.32	94.89	95.63	93.16	61.03	92.67	91.04	79.74	78.75
0.02	25	1400	94.7	95.04	96.54	94.73	66.32	94.75	92.36	83.93	82.65
0.03	20	2100	96.83	95.97	96.77	95.43	70.3	95.89	93.19	85.13	85.35
0.18	15	12600	96.95	96.15	96.98	96.31	75.01	96.68	94.17	89.65	88.16
0.2	10	14000	96.09	96.18	97.01	97.1	78.53	96.94	95.51	92.45	93.38
0.4	5	28000	97.04	96.91	97.17	98.54	85.81	97	96.8	95.81	96.72
0.8	3	56000	97.01	97.23	97.69	98.54	85.81	97	96.8	95.81	96.72

[a] EL1: RF + KNN + MLP.

[b] EL2: CNN + KNN + SVM.

[c] EL3: CNN + KNN + MLP.

Table 3: Labeling Price VS Accuracy.

	Cost/Label	Noise%	Total Cost
Non-Expert	0.01	30	700
Non-Expert	0.02	25	1400
Non-Expert	0.03	20	2100
Non-Expert	0.18	15	12600
Expert	0.2	10	14000
Expert	0.4	5	28000
Expert	0.8	3	56000

which is the most expensive and accurate labeling with only 3% of noise. On the other hand, the cost of labeling is only $700 in total with 30 percent of noise. It can be seen that using ensemble learning outperforms all other machine learning models when used individually. In fact, using a combination of CNN, KNN, and MLP yields the highest accuracy of 95.63 even with the high presence of label noise. We got a desirable accuracy with the lowest cost even with up to 30% of noise. This shows that we can use ensemble learning as a cost effective method that can cope with label noise. Through extensive experimentation, we can see that using ensemble learning can save us $55,300 when using non-expert to label the MNIST dataset with 70,000 images while maintaining relatively similar accuracy. In other words, paying an extra $55,300 for an expert annotator to reduce the label noise from 30 to 3 percent will only increase the accuracy by 2%. Therefore, ensemble learning is considered one of the best method that copes with label noise with low cost to maintain the desirable accuracy.

Unlike using single ML models where the accuracy can drop dramatically when we increase the level of label noise, using an ensemble method maintain high accuracy regardless of the level of noise in the dataset. For instance, the accuracy of DT dropped by roughly 25% when the level of noise increases from 3% to 30% as shown in Figure 2. Even with ML models that are more robust to label noise such as CNN, RF, and KNN, there are a noticeable decrease in the accuracy. These models perform better with cleaner dataset, which is more expansive to obtain. On the contrary, using an ensemble learning with any combinations of ML models would result in a better accuracy with lower cost as illustrated in Figure 3.

5 LIMITATIONS

Although we used a representative distribution of ML models, there are still some ML algorithms that can be investigated. It is challenging to conduct the same experiments with all ML models. However, based on the results of the experiments in this research paper, we can see that the results can be generalized. Furthermore, even though we used various datasets in terms of size and types of data, there are more data that needs to be explored such as sound datasets. However, it is beyond the scope of this research to conduct the same experiments with every data type. Thus, we used various data such as numerical and image data. Also, there are different combinations of ML models that can be used in ensemble learning. Although we explored various combinations of ML algorithms to come up with the best ensemble method in terms of robustness to label noise, there are still other options to explore that may result in a better accuracy.

6 CONCLUSION

In this paper, we explored the impact of class label noise on machine learning algorithms' performance and accuracy. We proposed two simple settings for labeling cost optimization. We explored both settings by analyzing: the tradeoff between the size and cleanliness of the dataset, as well as the tradeoff between labeling cost and machine learning performance. Depending on the budget, we can choose the best ML model. We have shown that machine learning models

SCITEPRESS
SCIENCE AND TECHNOLOGY PUBLICATIONS

have different performance even with the same level of noise. To have the desired accuracy with low-cost, we have also shown that it is more important to have a huge dataset even with high level of noise as opposed to small clean dataset. This is because most ML algorithms need bigger data to train in order to perform well. We have further shown that desirable ML performance can be achieved with a low labeling cost using ensemble learning since it is more resilient and robust to label noise.

REFERENCES

Alsubhi, J., Gharawi, A., and Alahmadi, M. (2021). A performance study of membership inference attacks on different machine learning algorithms. *Journal of Information Hiding and Privacy Protection*, 3(4):193.

Aslan, S., Mete, S. E., Okur, E., Oktay, E., Alyuz, N., Genc, U. E., Stanhill, D., and Esme, A. A. (2017). Human expert labeling process (help): towards a reliable higher-order user state labeling process and tool to assess student engagement. *Educational Technology*, pages 53–59.

Deng, L. (2012). The mnist database of handwritten digit images for machine learning research. *IEEE Signal Processing Magazine*, 29(6):141–142.

Dong, X., Yu, Z., Cao, W., Shi, Y., and Ma, Q. (2020). A survey on ensemble learning. *Frontiers of Computer Science*, 14(2):241–258.

Dua, D. and Graff, C. (2017). UCI machine learning repository.

Durga, S., Nag, R., and Daniel, E. (2019). Survey on machine learning and deep learning algorithms used in internet of things (iot) healthcare. In *2019 3rd international conference on computing methodologies and communication (ICCMC)*, pages 1018–1022. IEEE.

Feng, D., Besana, S., and Zajac, R. (2009). Acquiring high quality non-expert knowledge from on-demand workforce. In *Proceedings of the 2009 Workshop on The People's Web Meets NLP: Collaboratively Constructed Semantic Resources (People's Web)*, pages 51–56.

Garcia, L. P., de Carvalho, A. C., and Lorena, A. C. (2015). Effect of label noise in the complexity of classification problems. *Neurocomputing*, 160:108–119.

Gardner, M. W. and Dorling, S. (1998). Artificial neural networks (the multilayer perceptron)—a review of applications in the atmospheric sciences. *Atmospheric environment*, 32(14-15):2627–2636.

Huang, J., Qu, L., Jia, R., and Zhao, B. (2019). O2u-net: A simple noisy label detection approach for deep neural networks. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3326–3334.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2017). Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90.

Krogh, A. and Hertz, J. (1991). A simple weight decay can improve generalization. *Advances in neural information processing systems*, 4.

Lee, K.-H., He, X., Zhang, L., and Yang, L. (2018). Cleannet: Transfer learning for scalable image classifier training with label noise. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5447–5456.

Li, W., Wang, L., Li, W., Agustsson, E., and Van Gool, L. (2017). Webvision database: Visual learning and understanding from web data. *arXiv preprint arXiv:1708.02862*.

Moon, W. K., Lee, Y.-W., Ke, H.-H., Lee, S. H., Huang, C.-S., and Chang, R.-F. (2020). Computer-aided diagnosis of breast ultrasound images using ensemble learning from convolutional neural networks. *Computer methods and programs in biomedicine*, 190:105361.

Nguyen, A. T., Wallace, B. C., and Lease, M. (2015). Combining crowd and expert labels using decision theoretic active learning. In *Third AAAI conference on human computation and crowdsourcing*.

Perez, L. and Wang, J. (2017). The effectiveness of data augmentation in image classification using deep learning. *arXiv preprint arXiv:1712.04621*.

Ren, Q., Cheng, H., and Han, H. (2017). Research on machine learning framework based on random forest algorithm. In *AIP conference proceedings*, volume 1820, page 080020. AIP Publishing LLC.

Sagi, O. and Rokach, L. (2018). Ensemble learning: A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 8(4):e1249.

Sitawarin, C. and Wagner, D. (2019). On the robustness of deep k-nearest neighbors. In *2019 IEEE Security and Privacy Workshops (SPW)*, pages 1–7. IEEE.

Song, H., Kim, M., and Lee, J.-G. (2019). Selfie: Refurbishing unclean samples for robust deep learning. In *International Conference on Machine Learning*, pages 5907–5915. PMLR.

Su, J. and Zhang, H. (2006). A fast decision tree learning algorithm. In *Aaai*, volume 6, pages 500–505.

Wang, X.-Z., Xing, H.-J., Li, Y., Hua, Q., Dong, C.-R., and Pedrycz, W. (2014). A study on relationship between generalization abilities and fuzziness of base classifiers in ensemble learning. *IEEE Transactions on Fuzzy Systems*, 23(5):1638–1654.

Xiao, T., Xia, T., Yang, Y., Huang, C., and Wang, X. (2015). Learning from massive noisy labeled data for image classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2691–2699.

Yu, L., Wang, S., and Lai, K. K. (2008). Forecasting crude oil price with an emd-based neural network ensemble learning paradigm. *Energy economics*, 30(5):2623–2635.