Identifying Label Noise in Time-Series Datasets

Gentry Atkinson
gma23@txstate.edu
Texas State University
San Marcos, Texas

Vangelis Metsis
vmetsis@txstate.edu
Texas State University
San Marcos, Texas

ABSTRACT

Reliably labeled datasets are crucial to the performance of supervised learning methods. Time-series data pose additional challenges. Data points lying on borders between classes can be mislabeled due to perception limitations of human labelers. Sensor measurements may not be directly interpretable by humans. Thus label noise cannot be manually removed. As a result, time-series datasets often contain a significant amount of label noise that can degrade the performance of machine learning models. This work focuses on label noise identification and removal by extending previous methods developed for static instances to the domain of time-series data. We use a combination of deep learning and visualization algorithms to facilitate automatic noise removal. We show that our approach can identify mislabeled instances, which results in improved classification accuracy on four synthetic and two real publicly available human activity datasets.

CCS CONCEPTS

• **Computing methodologies** → *Supervised learning by classification.*

KEYWORDS

Label cleaning, neural networks, time-series data, CNN, accelerometer, human activity recognition, label noise

ACM Reference Format:
Gentry Atkinson and Vangelis Metsis. 2020. Identifying Label Noise in Time-Series Datasets. In *Adjunct Proceedings of the 2020 ACM International JointConference on Pervasive and Ubiquitous Computing and Proceedings of the 2020 ACM International Symposium on Wearable Computers (UbiComp/ISWC'20 Adjunct), September 12–16, 2020, Virtual Event, Mexico.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3410530.3414366

UbiComp/ISWC '20Adjunct, September 12–16,2020, Virtual Event, Mexico

ACM ISBN 978-1-4503-8076-8/20/09…$15.00
https://doi.org/10.1145/3410530.3414366

1 INTRODUCTION

Noise is most commonly associated with errors in the data collection process. Sources of errors can include faulty sensors, atmospheric disturbances, corrupted files, and others. There are many methods to decrease the noise in collected data. But errors in the labels assigned to data points are often overlooked. Label noise frequently occurs in manually annotated datasets, especially when researchers rely on observation or self-reporting from research subjects. Such situations are especially common, but not limited to, bio-signal data.

Removing or re-labeling mislabeled data points can be a powerful pre-processing step. Unfortunately, label noise reduction is often not possible after the initial label assignment due to the fact that sensor measurements are captured in a format that is not directly interpretable by humans.

Several packages have been introduced to identify mislabeled data points in large datasets. A common approach to the task of label noise detection is the single learning method [6], which trains a classifier on a full dataset and then assumes that the data points classified with the least certainty are the ones most likely to be mislabeled. However, most single learning systems do not include classifiers that are appropriate for time-series data, nor do they have the ability to recognize data as being time-series. Instead, they are applied to static instances, where each instance is a fixed-size feature vector or a matrix, in the case of images. This makes it of limited use in bio-signal analysis.

Time-series data present a challenge over static numerical data because the significance of a label difference between two points in an instance is not fixed. A difference between two points is relevant if they are close in time and potentially irrelevant if they are distant in time. In this way, time-series data are similar to image data and benefit from many of the same advances in machine learning, particularly Convolutional Neural Networks (CNNs). Activity recognition presents additional challenges over other varieties of time-series data [14]. The classes are often very poorly distinguished and often depend on self-reporting by research subjects. This makes the likelihood of instances being mislabeled even higher.

The goal of this project is to develop a new single learning [6] label noise detector using a 1-dimensional CNN (1D-CNN). Our work extends the capabilities of an existing tool

for label noise removal, namely Labelfix [13], by adapting it to process time-series data. Like its predecessor our system works with minimal input from the end user. The only required input from the user is the parameter α, i.e. the percentage of the dataset to be flagged for review. User's can also pass in hyper-parameter for the classifier but are not required to.

In order to demonstrate the effectiveness of the system in removing label noise from time-series data, a pair of classifiers are trained on both uncleaned and cleaned data. Our experiments show that the accuracy of the classifiers improves by as much as 7.5% and the F1 score by as much as 4.2% when trained and evaluated on the cleaned data. These tests were performed on four synthetic datasets of different sizes and label counts, and two publicly available real datasets, the UniMiB SHAR dataset [12], and the Sussex-HuaWei Locomotion hand and torso training data [5]. The precision of the mislabel detection was also measured by adding artificial label noise to the datasets to allow a precise count of true- and false-positives. The experiments show that the precision of mislabeled point detection ranges from 0.53 to 0.98 depending on the data and the applied α, ranging from 1 to 3% in our experiments.

This paper is organized as follows. In section 2 we provide an overview of the previous work related to the methods used in this work. Section 3 elaborates on the our methodology for time-series label noise removal. The datasets used to evaluate the effectiveness of our method are described in section 4. In section 5, we present and discuss our experimental results. Finally, section 6 summarizes our findings and concludes this paper.

2 BACKGROUND

As explained in [4], any process that pollutes labels can be called label noise. This could be caused by misreporting by research volunteers, transcription errors when data is copied, inconsistent data encoding, or mis-annotation. Subjectivity is another common cause of data being mislabeled [1]. Noise represents a mismatch between the true class of an instance and the recorded label of the instance. Noise can be grouped as: completely at random, at random, or not at random depending on the dependencies between the feature space and occurrences of mislabeling [4]. Likewise noise detectors can be grouped into: Local Learning, Ensemble Learning, and Single Model Learning based on the tools used to comb data for noise [6].

Labelfix was introduced in 2019 as a technique for identifying mislabeled instances in large datasets [13]. The team developed a robust system for identifying a certain percentage of the instances in a full dataset as requiring human review for mislabeling. Furthermore, the system was able to function without user-defined hyper-parameters other than the percentage of instances to flag, which the authors call the α value. Labelfix is a good example of the single learning [6] method of label noise detection. This class of noise detectors train a model on the questionable data and then assume that the points which were classified with the least certainty are the most likely to be mislabeled. One major advantage of Labelfix over other single learning methods is the ability of the system to select a model and a set of hyper-parameters without further input from the user.

Several approaches have been tested to relabeling or cleaning activity recognition datasets. Zhao et al. developed a technique [15] that used human workers through Amazon's mechanical turk to relabel points close to a decision boundary. This system had the workers view a short segment of video and then assign a label to the corresponding segment of inertial data. Another approach, as demonstrated in [8] is to create a classifier which is robust by assuming the presence of some label jitter in training data. Our system does not require video data and detects label noise that is broader than the "temporal label misalignment" used in [8].

Convolutional Neural Networks (CNNs), which were first introduced in 1998 [9], use small clusters of nodes referred to as kernels whose inputs are transposed over the full range of an instance from a dataset. These kernels learn low-level features from the data. High layers in the network can process and combine the output of the low level features to learn higher level features. The ability of CNNs to learn features from raw data based on comparisons of small collections of data points in an instance makes them well suited to signal processing. 1D-CNNs have been previously used for human activity recognition [10], among other applications.

These activities are generally broad classes of movement such as: walking, running, climbing stairs, sitting down, etc. Many sensors can be employed in this classification task but the 3-axis accelerometer is the most common choice [14] and is used in this work.

3 METHODOLOGY

The synthetic data used in the described experiments were generated using a technique described in 1999 [7]. Arrays are initialized to random values in a user-defined range and then processed by a series of generators, which add recognizable characteristics to the random arrays. Collections of initialization variables to pass to the data generators were used to generate synthetic signals into classes which are reliably similar to each other. This process is described in detail in Section 4.

Our current implementation assumes that numerical time-series data appear as a single channel, or can be converted into single channel stream. Furthermore we assume that an instance of time-series data will have a greater vector length (dimensionality, listed as "d" in Table 2) than an instance

Table 1: The list of features extracted from signals.

• Mean • Standard Deviation • Absolute Energy • Sum of Changes • Auto Correlation • Count of Values Above Mean • Count of Values Below Mean • Kurtosis • Longest Strike Above Mean • Zero Crossing Rate • Number of Peaks • Sample Entropy • Welch Spectral Density (6 coefficients)

of extracted features. For these experiments, datasets with more that 150 samples per instance are processed as time-series data. Time series data and numerical data are both normalized to a range of -1 to 1.

Features were extracted when necessary using the TSFresh Python library [2]. A fixed set of 13 features were chosen from the many offered by this library and are shown in Table 1. These features were chosen for their usefulness in classification tasks, for the relative ease of computation, and for being portable through a broad variety of signals.

Data that have been identified as time-series are fit by a 1-D CNN. This network employs 4 convolutional layers with a kernel size of 16 samples. The output of each 2 convolutional layers is maxed pool and a 0.25 dropout is applied. 2 dense layers follow the convolutional layers with the output of the second identifying the label of the input data sample.

The output of the CNN is sorted by the certainty of classification, with the most uncertain points at the top of the list. The points at the head of the list will now be the samples which are most likely to be mislabeled. The user can review or remove as many of the indexes as is appropriate with one to three percent being common[13], by setting the α value.

Some varieties of data lend themselves more readily to hand review than others. It's easy to see a picture of a shirt labeled as a dress and understand that the label is incorrect. But time-series data are substantially more difficult for human viewers to interpret. Because of this experiments have been devised for this project which do not rely on human interpretation. The assumption was made that a classifier will perform more accurately on a dataset which does not contain mislabeled data.

In the first experiment an SVM is trained and evaluated on uncleaned data and then compared to an SVM trained and evaluated on cleaned data. The training and test sets were generated with an 80/20 shuffled split of the full dataset. This comparison was repeated 5 times, with the dataset being reshuffled and re-split for each repetition. Averaged results are presented in Section 5. $\alpha = 0.02$ was used as the cleaning percentage for these trials, meaning that 2% of the instances will be flagged.

The second experiment is similar to the first but employs a CNN as a classifier rather than an SVM. Because feature extraction is not necessary or beneficial for CNNs, all training and testing was done using only the raw datasets. The CNN used in this experiment is a more compact model than the one used in the data cleaning steps with only 2 convolutional layers feeding into 2 dense layers after max pooling. The hope is to show improved performance of a "lightweight" model by cleaning data. Again, an α value of 0.02 was used.

The final experiment introduces label noise into the eight datasets by altering three percent of the labels. Three percent was chosen based on the observation that most real-world datasets have approximately 5% label noise [3] with the modification down to 3% suggested by [13]. This noise was introduced in random fashion with an altered label being equally likely to be any label other than the correct one. This type of noise is patterned after the "Noisy Completely at Random" class of label noise introduced in [4]. The four synthetic datasets already have a 3% noise introduced during their creation. The label sets for the UniMib SHAR and Sussex-HuaWei datasets were altered for only this experiment by changing labels to an incorrect class with a 3% probability. The indexes of the altered labels are stored and used to compute the precision and recall of label cleaning at rates of α = 1, 2, and 3% of labels being marked as mislabeled.

We hypothesize that cleaning label noise out of a time-series dataset using our approach will improve the performance of a classifiers trained and tested with cleaned data as compared to trained and tested on uncleaned. This will be demonstrated by a consistent improvement in the measured precision, accuracy, and recall of the classifiers that were trained on the cleaned time-series data.

4 DATA

Eight datasets were employed in this project. Four of them were synthetically generated using parameters that make them well suited to the techniques being developed for this project. Another two were taken from the UniMiB SHAR dataset [12], which is offered as a two label fall detection data set or as a 17 label activity detection dataset. The final two sets are the "hand" and "torso" train sets from the Sussex-Huawei locomotion dataset [5].

The synthetic datasets were generated using a technique proposed in [7]. Arrays are filled with random values and then processed using generators which add particular distinct features (cylinders, bells, and funnels) to the data in ways that mimic real-world time-series data. Since data generation depends on user-defined parameters it is possible to create sets of these parameters that reliably generate data into distinct classes. All four of the synthetic datasets were given roughly 3% label noise by randomly altering some of the labels to represent one of the incorrect classes.

The UniMiB SHAR dataset [12], was collected at University of Milano Bicocca in 2017 using commercial smart phones. The phones were carried in a front trouser pocket of the participants. The accelerometer sampling rate was 50Hz. Two sets of labels are provided for the data used in

Table 2: A summary of the eight datasets used in this project.

	instances	d	classes
Synthetic 1	1000	500	2
Synthetic 2	1000	500	5
Synthetic 3	5000	1000	2
Synthetic 4	5000	1000	5
Sussex-HuaWei 1	196072	500	8
Sussex-HuaWei 2	196072	500	8
UniMiB SHAR 1	11771	151	2
UniMib SHAR 2	11771	151	17

these experiments. One labels instances as fall/not-fall. The other labels instances with 17 activities of daily life including: standing from laying, standing from sitting, lying down from standing, running, sitting down, going downstairs, going upstairs, walking, jumping, jogging, falling back, falling forward, falling left, falling right, and hitting an obstacle.

The Sussex-Huawei Locomotion dataset [5], was collected in 2018 with subjects wearing several devices on their bodies engaged in 8 activities of locomotion: standing still, walking, running, bicycling, riding in a bus, riding in a car, riding in a train, and riding in a subway. Collected sensor modalities included accelerometer, GPS, gyroscope, and video data. We chosen to analyze only the accelerometer data in these experiments as they are interpretable, proven in reliable for activity recognition, and well suited to our label noise cleaning technique.

Table 2 summarizes the size of each dataset. The value "d" is the dimensionality of each instance, which is the number of samples in each instance. All eight datasets are single channel. Where necessary, a total acceleration as calculated as the sum of accelerations as the vector norm of the x, y, and z axes.

5 RESULTS

The first experiment run on our data was to train and evaluate an SVM classifier on the original, "uncleaned" data, then retrain and evaluate using the "cleaned" dataset, on which instances identified as possibly mislabeled by our system were removed. The indexes to be removed were identified using only the raw data. The SVM was trained using the extracted feature set as described in Section 3. The accuracy and F1 score averaged over 5 runs are presented in Table 3. These figures indicate that the SVM demonstrated an average improvement of 2.47% accuracy and 2.54% in the F1 score when classifying cleaned data on the synthetic datasets, 0.20% accuracy and 0.37% in the F1 score on the Sussex-HuaWei datasets, and 0.70% accuracy and 1.69% in the F1 score.

The second experiment is similar to the first in that a classifier is trained and evaluated on both cleaned and uncleaned data. The classifier used in this process is a small CNN. The accuracy and F1 score of the evaluated CNN are presented in Table 4. The CNN showed an average 2.12% improvement in accuracy and a 2.09% improvement in F1 score when classifying cleaned synthetic data. The Sussex-Huawei datasets averaged a 5.12% improvement in accuracy and a 3.38% improvement in F1 score. The performance of the CNN on the UniMiB datasets was improved by an average of 1.29% in accuracy and 1.65% in the F1 score.

Table 3: The average accuracy and F1 score is presented for 5 runs of classification using an SVM on cleaned and uncleaned data based on an 80/20 shuffled split on the full dataset.

	Uncleaned		Cleaned	
	Acc	F1	Acc	F1
Synthetic 1	0.968	0.968	0.991	0.991
Synthetic 2	0.835	0.833	0.862	0.862
Synthetic 3	0.972	0.972	0.989	0.989
Synthetic 4	0.909	0.910	0.943	0.943
Sussex-HuaWei 1	0.559	0.580	0.564	0.586
Sussex-HuaWei 2	0.618	0.652	0.616	0.654
UniMiB SHAR 1	0.908	0.899	0.908	0.899
UniMiB SHAR 2	0.523	0.388	0.538	0.417

Table 4: The average accuracy and F1 score is presented for 5 runs of classification using a CNN on cleaned and uncleaned data based on an 80/20 shuffled split on the full dataset.

	Uncleaned		Cleaned	
	Acc	F1	Acc	F1
Synthetic 1	0.976	0.976	0.995	0.995
Synthetic 2	0.835	0.834	0.860	0.862
Synthetic 3	0.970	0.970	0.992	0.992
Synthetic 4	0.909	0.910	0.969	0.969
Sussex-HuaWei 1	0.515	0.545	0.557	0.543
Sussex-HuaWei 2	0.531	0.606	0.607	0.631
UniMiB SHAR 1	0.998	0.997	0.999	0.999
UniMiB SHAR 2	0.805	0.728	0.829	0.759

The third experiment performed in this project was to introduce 3 percent label noise into each of the eight datasets. This allowed lists of true mislabeled points to be collected and compared to the lists of potential mislabels generated by the system. The precision and recall were measured with 1, 2, and 3% of the dataset flagged. Precision values are presented in Table 5 and recall in Table 6.

A visualization of the UniMiB Fall dataset is presented in Figure 1. t-SNE [11] has been used to reduce the dimensionality of the feature set down to two dimensions, in a way that preserves the relative distance of instances. The

Table 5: The precision of instance identification as mislabeled in 8 datasets with 3% label noise introduced into the data. Each value is averaged over five runs.

	Precision		
	α=0.01	α=0.02	α=0.03
Synthetic 1	0.962	0.850	0.783
Synthetic 2	0.933	0.913	0.835
Synthetic 3	0.988	0.942	0.858
Synthetic 4	0.989	0.919	0.828
Sussex-HuaWei 1	0.534	0.534	0.539
Sussex-HuaWei 2	0.572	0.586	0.604
UniMiB SHAR 1	0.986	0.984	0.931
UniMiB SHAR 2	0.813	0.815	0.799

Table 6: The recall of instance identification as mislabeled in 8 datasets with 3% label noise introduced into the data. Each value is averaged over five runs.

	Recall		
	α=0.01	α=0.02	α=0.03
Synthetic 1	0.688	0.781	0.837
Synthetic 2	0.641	0.769	0.824
Synthetic 3	0.669	0.804	0.865
Synthetic 4	0.680	0.806	0.856
Sussex-HuaWei 1	0.512	0.523	0.538
Sussex-HuaWei 2	0.525	0.559	0.606
UniMiB SHAR 1	0.663	0.822	0.927
UniMiB SHAR 2	0.595	0.691	0.770

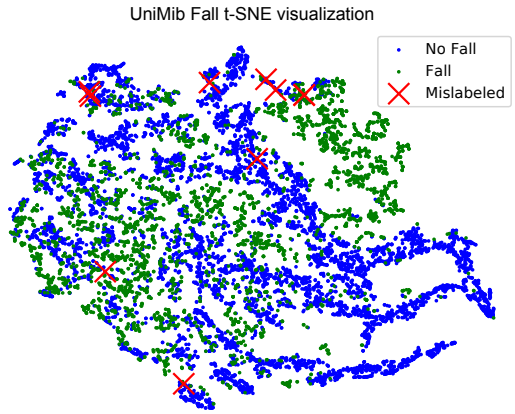

Figure 1: Likely mislabeled instances in the UniMiB fall dataset are marked in red.

Sussex-HuaWei Torso dataset is shown using the same technique in Figure 2. The points that our system has identified as the most likely to be mislabeled have been marked. For comparison Synthetic Set 4 is plotted in Figure 3 with its mislabeled instances marked.

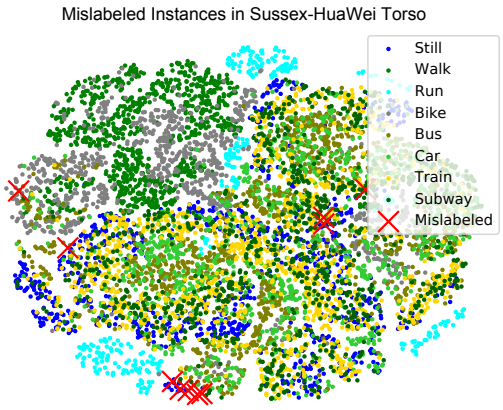

Figure 2: Likely mislabeled instances in the Sussex-HuaWei Torso dataset are marked in red.

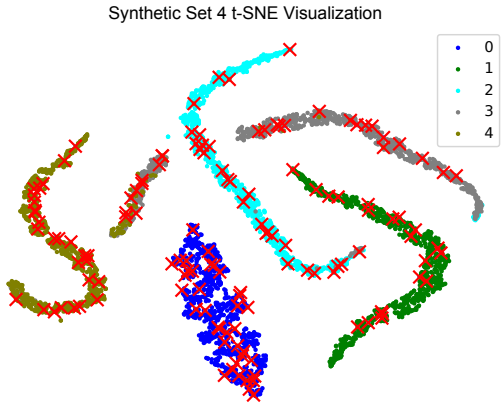

Figure 3: Mislabeled instances in the Synthetic 4 dataset are marked in red.

Discussion

Human Activity Recognition (HAR) remains one of the most complicated classification problems in bio-signal analysis [14]. Because activity classes may be very similar to one another and because several activities might blend into one another, activity recognition is particularly susceptible to label noise. This is why it is particularly important to have access to data cleaning methods based of the most effective technologies, e.g. deep and convolutional networks.

We have presented our findings based on cleaning both the train and test data. As mentioned in Section 5 it is also possible to clean only the training data for a classification model but with far less efficacy. This follows the assumption that the primary use of our system will be "offline" cleaning of large datasets. In this case, an analyst will have access to labeled train and test data. Some work has shown that removing label noise in a training set can have a harmful effect if the noise is still present in the test set [1].

We have approached this project assuming that the UniMiB SHAR and Sussex-Huawei datasets were mostly cleanly labeled to begin with. This is why 3% noise was artificially introduced during the third experiment. This decision may account for the relatively low precision seen in experiment three on the Sussex-HuaWei datasets. If noisy labels existed in the dataset before our manipulation, then they would appear as false positives in the experiment and decrease the F1 score. The relative absence of label noise in the UniMiB datasets is also a factor in the excellent performance of the CNN on the uncleaned fall detection dataset. Although the cleaned data still produced a more reliable classifier, there was very little room left for improvement.

Noise removal is not without risks. It is possible that edge cases or ambiguous classes will be cleaned entirely out of the system, which could degrade rather than improve the performance of classifiers working on the cleaned data. Flagged instances should be reviewed rather than automatically treated as being mislabeled. It is also helpful to start by only flagging a small percentage of the datasets. This project followed this approach by only flagging 2% of the data in experiments one and two, and 1, 2, or 3% in experiment three.

As a final note, anyone who wishes to repeat these experiments will need to remember that the output of the mislabeled instance finding function is not purely deterministic as it depends on the probability of a multi-class prediction by a trained CNN, which can vary between runs. Therefore results derived from rerunning this project's code will be very similar but not identical.

6 CONCLUSION

Our system has demonstrated consistent improvements in classifiers trained on cleaned data as judged by both the accuracy and F1 score. Furthermore we were able to correctly identify intentionally mislabeled points in real-world datasets with up to 98.6% precision. Labelfix demonstrated similar precision under the same circumstances [13], so our results are consistent with the state of the art but have incorporated time-series data.

Although our focus was on human activity recognition for this project, nothing about this work is limited to functioning exclusively in the arena of HAR. Any time-series dataset could be improved by our approach to label-noise removal. This is supported by the strong results we achieved on the synthetic datasets, which were not constructed to mimic activity data specifically.

ACKNOWLEDGMENTS

The authors would like to thank Müller et al. for providing a well documented codebase with Labelfix, Micucci et al. for the UniMiB SHAR dataset, Gjoreski et al. for the Susex-HuaWei dataset, Sarkar for providing the implementation of an data generator, and finally their fellow researchers in the Texas State University IMICS lab.

REFERENCES

[1] Carla E Brodley and Mark A Friedl. 1999. Identifying mislabeled training data. *Journal of artificial intelligence research* 11 (1999), 131–167.

[2] Maximilian Christ, Nils Braun, Julius Neuffer, and Andreas W Kempa-Liehr. 2018. Time series feature extraction on basis of scalable hypothesis tests (tsfresh–a python package). *Neurocomputing* 307 (2018), 72–77.

[3] Benoît Frénay, Ata Kabán, et al. 2014. A comprehensive introduction to label noise.. In *ESANN*.

[4] Benoît Frénay and Michel Verleysen. 2013. Classification in the presence of label noise: a survey. *IEEE transactions on neural networks and learning systems* 25, 5 (2013), 845–869.

[5] Hristijan Gjoreski, Mathias Ciliberto, Lin Wang, Francisco Javier Ordonez Morales, Sami Mekki, Stefan Valentin, and Daniel Roggen. 2018. The university of sussex-huawei locomotion and transportation dataset for multimodal analytics with mobile devices. *IEEE Access* 6 (2018), 42592–42604.

[6] Donghai Guan and Weiwei Yuan. 2013. A survey of mislabeled training data detection techniques for pattern classification. *IETE Technical Review* 30, 6 (2013), 524–530.

[7] Mohammed Waleed Kadous. 1999. Learning Comprehensible Descriptions of Multivariate Time Series.. In *ICML*, Vol. 454. 463.

[8] Hyeokhyen Kwon, Gregory D Abowd, and Thomas Plötz. 2019. Handling annotation uncertainty in human activity recognition. In *Proceedings of the 23rd International Symposium on Wearable Computers*. 109–117.

[9] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. 1998. Gradient-based learning applied to document recognition. *Proc. IEEE* 86, 11 (1998), 2278–2324.

[10] Song-Mi Lee, Sang Min Yoon, and Heeryon Cho. 2017. Human activity recognition from accelerometer data using Convolutional Neural Network. In *2017 ieee international conference on big data and smart computing (bigcomp)*. IEEE, 131–134.

[11] Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of machine learning research* 9, Nov (2008), 2579–2605.

[12] Daniela Micucci, Marco Mobilio, and Paolo Napoletano. 2017. Unimib shar: A dataset for human activity recognition using acceleration data from smartphones. *Applied Sciences* 7, 10 (2017), 1101.

[13] Nicolas M Müller and Karla Markert. 2019. Identifying Mislabeled Instances in Classification Datasets. In *2019 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 1–8.

[14] Sreenivasan Ramasamy Ramamurthy and Nirmalya Roy. 2018. Recent trends in machine learning for human activity recognition—A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 8, 4 (2018), e1254.

[15] Liyue Zhao, Gita Sukthankar, and Rahul Sukthankar. 2011. Incremental relabeling for active learning with noisy crowdsourced annotations. In *2011 IEEE Third International Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third International Conference on Social Computing*. IEEE, 728–733.

A ONLINE RESOURCES

Code and data generators for this project can be found at https://github.com/imics-lab/identifying_label_noise