Leveraging Repeated Unlabelled Noisy Measurements to Augment Supervised Learning

Birk Martin Magnussen
birk.magnussen@biozoom.net
biozoom services GmbH
Kassel, Germany

Claudius Stern
Claudius.Stern@fom.de
FOM Hochschule für Oekonomie & Management
Kassel, Germany

Bernhard Sick
bsick@uni-kassel.de
Universität Kassel
Kassel, Germany

ABSTRACT

Often, producing large labelled datasets for supervised machine learning is difficult and expensive. In cases where the expensive part is due to labelling and obtaining ground truth, it is often comparably easy to acquire large datasets containing unlabelled data points. For reproducible measurements, it is possible to record information on multiple data points being from the same reproducible measurement series, which should thus have an equal but unknown ground truth. In this article, we propose a method to incorporate a dataset of such unlabelled data points for which some data points are known to be equal in end-to-end training of otherwise labelled data. We show that, with the example of predicting the carotenoid concentration in human skin from optical multiple spatially resolved reflection spectroscopy data, the proposed method is capable of reducing the required number of labelled data points to achieve the same prediction accuracy for different model architectures. In addition, we show that the proposed method is capable of reducing the negative impact of noisy data when performing a repeated measurement of the same sample.

CCS CONCEPTS

• **Computing methodologies** → **Semi-supervised learning settings**; **Neural networks**; *Supervised learning by regression*; • **Applied computing** → Consumer health.

KEYWORDS

noisy data, inhomogeneous labels, neural networks, semi-supervised learning, reflection spectroscopy

ACM Reference Format:
Birk Martin Magnussen, Claudius Stern, and Bernhard Sick. 2023. Leveraging Repeated Unlabelled Noisy Measurements to Augment Supervised Learning. In *2023 the 6th International Conference on Computational Intelligence and Intelligent Systems (CIIS) (CIIS 2023), November 25–27, 2023, Tokyo, Japan.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3638209.3638210

1 INTRODUCTION

In machine learning, large datasets are required to correctly learn to differentiate noise from the actual data of interest. However, the acquisition of a sufficient number of training samples with associated ground truth is typically problematic. Depending on the task to be learned, the lack of sanples with ground truth can have different causes. One possible cause can be the difficulty or high cost of correctly assigning ground truth labels to the available training samples. In this case, it may still be possible to acquire a large number of unlabelled data points cheaply. In a lot of cases, the acquisition of training samples can be reproducible, allowing for the capture of different, noisy samples of the same situation, for example with a sensor repeatedly measuring the same test sample. By modifying the training method, it is possible to utilize these unlabelled data points from reproducible measurements to augment the labelled datasets.

In addition to a labelled dataset, this article proposes the curation of a known-equal dataset. A known-equal dataset contains unlabelled data points. From these unlabelled data points, multiple data points should be from a reproducible data source, such as measurements of the same test sample. Next, information about which unlabelled training samples are from the same test sample is stored. Even without expensively labelling these data points with ground truth, since these data points are from the same test sample, we know that they are expected to have the same ground truth if labelled. This article also proposes a method to use the information stored in a known-equal dataset to improve the trained model's noise resilience and reduce the number of labelled training samples required to roughly half, while achieving a prediction accuracy comparable to a model trained on the full labelled dataset. The method facilitates end-to-end training of existing model architectures by combining the model architecture's loss function with a loss function quantifying the homogeneity of samples known to be equal.

On the example of predicting the concentration of carotenoids in the human skin based on multiple spatially resolved reflection spectroscopy [3] (MSRRS), the proposed method is shown to be capable of reducing the required number of labelled samples by half without a significant drop in prediction accuracy, while being able to reduce the unadjusted sample variance of noisy, known-equal samples by up to half when the full dataset is available. In addition, the proposed method is shown to be effective for different model architectures without the need to manually adapt or tune the model architecture, making it ideal for retrofitting into existing model architectures.

The article will first discuss related work in the field of semi-supervised learning and utilizing unlabelled data, as well as related work on multi-objective optimization. The article will then introduce the proposed method and discuss different alternatives. In addition, the proposed method will be evaluated in the context of

CIIS 2023, November 25–27, 2023, Tokyo, Japan

improving the task of learning to predict the carotenoid concentration in human skin from optical MSRRS data.

2 RELATED WORK

Previous works for incorporating similar heterogeneous training sets fall into multiple categories. Some methods are designed for true semi-supervised learning settings, where in addition to the labelled dataset, only a dataset of completely unlabelled data is available.

These methods include the denoising autoencoders [14] (DAE). A DAE is pre-trained on the unlabelled dataset, intending to learn how to remove noise from the training samples to reduce the required complexity of models operating on the denoised training samples. The DAE is trained by artificially adding noise to clean samples in order to simulate real-world noise, and training the DAE to restore the correct sample. When a known-equal dataset is available, it is possible to train a DAE on only real noisy samples to restore a clean sample [7]. However, this introduces complexity, as the DAE is employed as a separately trained model instead of being able to be end-to-end trained in combination with the target task.

Another method is to pre-cluster both the labelled and unlabelled data [5]. Classifiers are then trained on each labelled cluster, and utilized to classify the unlabelled clusters. This type of cluster-then-label algorithms however requires that the data forms clusters or manifolds in the input data domain, which may not always be the case, and is thus not always applicable.

A different category of methods for heterogeneous training sets relies on having noisy or unreliable labels available in a second dataset in addition to the ground truth dataset. Some of these methods attempt to model or clean the noisy labels of the training dataset in order to clean or compensate for the labels during training [15, 13]. However, the known-equal dataset is not sufficient in order to effectively utilize these types of methods, as no information about the ground truth of an individual sample is available. Thus, these methods can not be used in conjunction with known-equal datasets.

Another approach to augment learning by knowing which samples are close together or far apart is called contrastive learning. The basic idea is that, given two samples, their distance in the output space is maximized if they are from different classes, whereas their distance is minimized if they are from the same classes [1]. Further enhancements in contrastive learning utilize triplets of input samples, where one sample is any input sample, the second sample is an input sample of the same class, and the third sample is a sample of a different class [10]. The main issue of contrastive learning method, is that they require negative samples that have different class labels. Because the known-equal dataset is only capable of yielding positive samples that are expected to be equal, contrastive learning methods cannot be effectively applied. In addition, randomly sampling other data points to use as negative samples [12] can be unsuitable for certain datasets where it can be assumed that many samples are close to each other in the output space.

Previous work for multi-objective optimization suggests combining optimization criteria using weighted sums. Similarly, it is suggested to rephrase multi-objective optimization into constrained single-objective optimization [4]. Loss combination in machine learning is often used for multi-task learning. One approach in multi-task learning is to minimize uncertainties from multiple factors by modelling the likelihoods of each uncertainty to weigh each task's loss [2].

3 INCORPORATING KNOWN-EQUAL DATA POINTS

Wanting to leverage the availability of a known-equal dataset for an existing, arbitrary model architecture implies that instead of the model architecture, the training itself must be modified to incorporate the additional data.

With normal, supervised regression learning, a labelled data point is input into the neural network. Then, the output is compared to the ground truth using some kind of loss function, such as the mean square error. The weights of the neural network are then updated to minimize this loss function.

In the proposed method, in addition to the loss yielded from the labelled samples, the known-equal samples are leveraged. To do this, each data sample in a tuple X of n known-equal samples is input into the neural network as well. This is possible, as the unlabelled samples in the known-equal dataset are the same type of samples as in the labelled dataset, just without attached ground truth. As no attached ground truth is available, the loss function to be minimized is the unadjusted sample variance between the results of the neural network for each of the n samples of the tuple of known-equal samples. When minimizing this loss, the neural network is trained to minimize the difference in result for samples which are known to be equal.

The unadjusted sample variance, as defined by [11], can be calculated with Equations 1 and 2, where $f(x_n)$ denotes the result of the neural network to be trained on the training sample $x_n \in X$:

$$\overline{f(x)} = \frac{1}{n} \cdot \sum_n f(x_n) \tag{1}$$

$$\sigma^2 = \frac{1}{n} \cdot \sum_n \left(f(x_n) - \overline{f(x)} \right)^2 \tag{2}$$

The resulting unadjusted sample variance σ^2 is then used as a quality metric for the homogeneity of the results of the neural network for known-equal samples. Thus, it can be used as the loss for training the neural network.

The proposed method suggests to sequentially input multiple data samples into the neural network for each training step. That means, that within one training step, the model is inferred multiple times instead of just once as in traditional supervised learning. The multiple results of the neural network are then calculated into two separate losses. One for the prediction accuracy of labelled sample, and one for the homogeneity of the results of the neural network for the known-equal samples. In order to reduce these to one loss, from which backpropagation can be performed, both losses need to be combined by a loss combination kernel. See section 4 for a discussion on loss combination kernels.

In the proposed method, as each labelled sample during training is augmented by the one tuple of n known-equal data points, training is performed using a total ratio of 1 to n labelled samples to known-equal samples. The potential effect of changing the ratio of labelled samples to tuples of n labelled samples is interesting and will be investigated in future research.

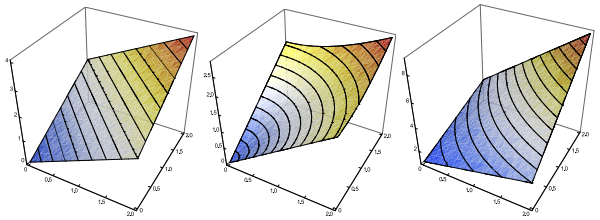

Figure 1: 3D plots of loss combination kernels. Left shows an additive kernel, center shows a euclidean kernel, right shows a multiplicative kernel.

4 LOSS COMBINATION KERNELS

The method as described in section 3 yields two losses, one from the prediction accuracy of labelled data and one from the homogeneity of the results of the neural network for known-equal samples. In order to train the neural network to minimize both losses, first the two losses are combined into one loss. This is possible with functions that take each loss as a parameter, and output one loss. In the case of the proposed method, that means the function is a simple mapping of $\mathbb{R}^2 \mapsto \mathbb{R}$, as two losses need to be combined.

In order to serve as a loss combination kernel, a function needs to fulfill a set of requirements. First, any loss combination kernel must be well defined for the value range of its input losses, usually $\mathbb{R}_{\geq 0}$, which make up the loss combination kernel's domain. In addition, with its domain restricted to the value range of its input losses, a loss combination kernel should have a global minimum where all input losses are minimal. Finally, any loss combination kernel must be strictly monotonous, that is to say: If any of the input losses decrease, with the other losses less or equal, the resulting loss will also decrease. Formally, a combination kernel $f(l_1, ..., l_n)$ with individual losses $l_1, ..., l_n$ must fulfill Equation 3:

$$l_i < l'_i, 1 \leq i \leq n \wedge l_1 \leq l'_1 \wedge ... \wedge l_n \leq l'_n \implies f(l_1, ..., l_n) < f(l'_1, ..., l'_n) \tag{3}$$

Figure 1 shows the combination kernels that were investigated for the proposed method. The losses used for the proposed method are a mean square error for the prediction accuracy and the unadjusted sample variance for the homogeneity of the results of the neural network for known-equal samples. The value range for both of these losses is $\mathbb{R}_{\geq 0}$. All three of the functions shown in Figure 1 fulfill the conditions, as they are well defined for $l_1, l_2 \in \mathbb{R}_{\geq 0}$, have a global minimum at $(0|0)$, and are strictly monotonous.

The first loss combination kernel, as shown on the left in Figure 1, is an additive kernel, defined as $f(l_1, l_2) := l_1 + l_2$. In most literature, such additive kernels have the individual losses weighted [4], and correctly weighting the losses is considered a difficult task with careful tuning required [2]. However, the prediction accuracy and the homogeneity of the results of the neural network for known-equal samples are both the same unit (the unit of the prediction target squared) as well as comparable in magnitude. In addition, the importance of both losses is similar to the end quality of the predictions of the neural network to be trained. Whether an inaccurate prediction was caused by a generally unreliable model or by excessive noise between samples expected to be equal is often of little importance to a user of the model to be trained. Combined with the similar magnitude of the two losses, it can be assumed that the weights of the two losses need to be of similar magnitude as well. Hence, using 1 as equal weights for each loss is sufficient as a first approximation.

The second loss combination kernel, as shown in the center in Figure 1, is a euclidean kernel, defined as $f(l_1, l_2) := \sqrt{\left(l_1^2 + l_2^2\right)}$. The partial derivative of this kernel towards either input loss, representing the importance of a change of that input loss for the total loss, is given by Equation 4:

$$\frac{\partial f(l_i, l_j)}{\partial l_i} = \frac{l_i}{\sqrt{\left(l_i^2 + l_j^2\right)}} \tag{4}$$

From Equation 4, we can see that the partial derivative, and thus the importance of a change in input loss, will decrease for a decreasing input loss, assuming the other loss is constant. The implication for gradient descent-based learning algorithms is that overly optimizing for one of the input losses without also optimizing for the other input loss is discouraged. This can help to avoid local minimum cases where one loss' global minimum prevents the other loss from lowering. One example of this would be the homogeneity of the results of the neural network for known-equal samples being 0 when all samples are assigned an identical result. The homogeneity would be ideal and the corresponding loss 0, but the correlation to the ground truth would be low.

The third loss combination kernel, as shown on the right in Figure 1, is a multiplicative kernel, defined as $f(l_1, l_2) := (l_1 + 1) \cdot (l_2 + 1)$. Unlike a simple multiplication of the input losses, this multiplicative kernel still fulfills the strict monotony criterion, even when one of the input losses is 0. Opposite to the euclidean kernel, the multiplicative kernel encourages optimizing the lower input loss further over-optimizing the higher loss, as the partial derivative of this kernel decreases for one loss with the other loss decreasing, as shown in Equation 5.

$$\frac{\partial f(l_i, l_j)}{\partial l_i} = l_j + 1 \tag{5}$$

The property of this kernel to prioritize a lower loss enables the neural network to utilize the knowledge gained from the lower loss to better abstract the correlation described by the higher loss, provided the losses do not describe counterproductive correlations.

5 EXPERIMENTS

The proposed method is evaluated on the task of predicting the concentration of carotenoids in the human skin from optical data measured using a sensor based on MSRRS. The measurement device, as described in [3], consists of multiple light detectors, and multiple light emitters, including emitters of different nominal wavelengths. The resulting data of one measurement, available to a neural network, consists of a measured brightness for each emitter-detector pair. In addition, the difference between actual and nominal wavelength is known for each light emitter.

5.1 Datasets

Two datasets are available for evaluation. One dataset contains labelled data points, while the other datasets is a known-equal

dataset used to augment the training of the neural network with the method proposed in this article. Both datasets are split into a training and a validation set each. The first dataset contains 2 000 MSRRS measurement samples of approximately 500 test subjects in a controlled test environment. For each test subject, the concentration of carotenoids in the skin is measured in vivo using a reference sensor. These reference measurements for each test subject serve as ground truth labels for the dataset. The second dataset is available with approximately 32 000 MSRRS measurements. These measurements are collected from users of the measurement device in a real-world environment. When measuring, the users are prompted to repeat the measurement four times to receive a more accurate result. From this, the 32 000 measurements are grouped into approximately 8 000 tuples of four measurements each, for which the result is unknown, but expected to be equal. These measurements are taken in an uncontrolled, real-world environment. Because of this, it is to be expected that the data contains unwanted noise. Possible error causes include for example incorrect use of the sensor, or that one or more of the repeated measurements has been taken by a different user, causing the results to no longer be equal as expected. Certain cases of misuse of the device are located and filtered by comparing the optical MSRRS data to approximately known limits when measuring human tissue.

5.2 Evaluation Setup

For evaluation, two different network architectures suitable for processing MSRRS data will be used. The first architecture is a continuous feature network (CFN), which has been shown to be effective for the type of available optical data [8]. For comparison, the other network is a simpler multi-layer feed-forward network (MLFF). As the multi-layer feed-forward network has significantly fewer parameters, it is expected to require fewer samples to be trained.

Both evaluation models and the denoising autoencoder were trained using the Adam optimizer [6] and implemented using the LibTorch bindings of the PyTorch framework [9].

In addition to the proposed method, an autoencoder is trained on the known-equal dataset as a comparison. Normally, autoencoders are trained on clean data, which is artificially inflicted with noise to be removed by the autoencoder [14]. The MSRRS datasets available however do not contain clean data for which such a method can be performed. However, instead of generating noisy samples from a clean sample, it is possible to utilize the available noisy samples [7]. From this, the denoising autoencoder is trained to output a clean sample corresponding to the tuple of known-equal samples for each sample within the tuple. The mean square weighted error is utilized as the loss function. Weighing is required here, as the different emitter-detector pairs in one measurement sample have different magnitudes. To compensate, the magnitude of each emitter-detector pair is computed by taking the average brightness of the emitter-detector pair over the entire dataset. From the computed magnitude, the prediction error of the autoencoder is weighted accordingly.

5.3 Results

Figure 2 shows the results of applying the method to the continuous feature network. The first graph shows the squared pearson correlation of the neural network prediction to the ground truth, relative to a baseline network trained on the labelled dataset without the proposed method. Higher is better.

The data shows that when the full dataset of 2 000 labelled samples is available, the proposed method is able to slightly increase the achieved prediction accuracy depending on the combination kernel used. When only half the dataset is available, a significant drop in prediction accuracy can be observed for the baseline network, reducing the squared pearson correlation to 85% compared to the baseline. However, when an additive kernel is used, a prediction accuracy almost as high as with the full dataset can be achieved, at 96% of the baseline value. While the multiplicative kernel performs only slightly below the additive kernel, the euclidian kernel performs subpar, even below the baseline network. This is presumably due to the optimizer being forced to reduce both prediction loss and known-equal homogeneity loss at the same time in order to reduce the overall loss, causing local minima which the optimization process is unable to pass. When only 37.5% of the 2 000 samples from the dataset are available, the prediction accuracy of the baseline network can be seen to sharply drop to only 35% of the squared

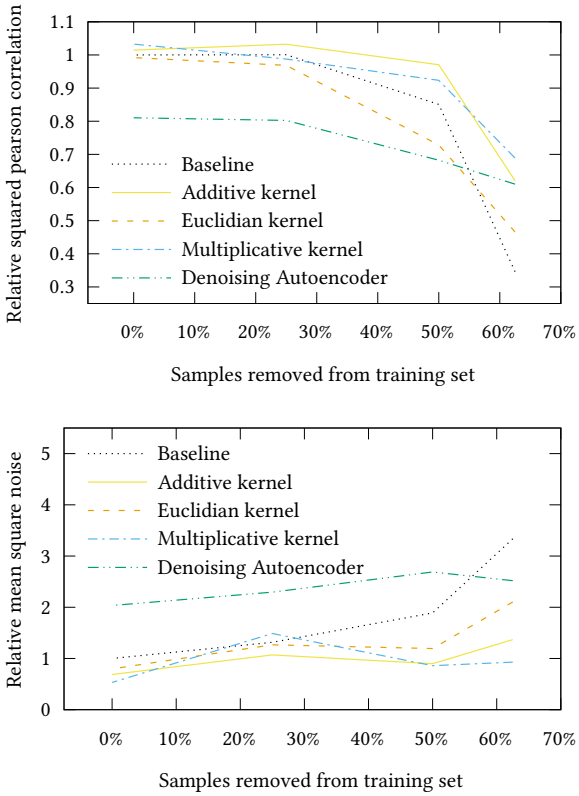

Figure 2: Results from applying the proposed method to a continuous feature network.

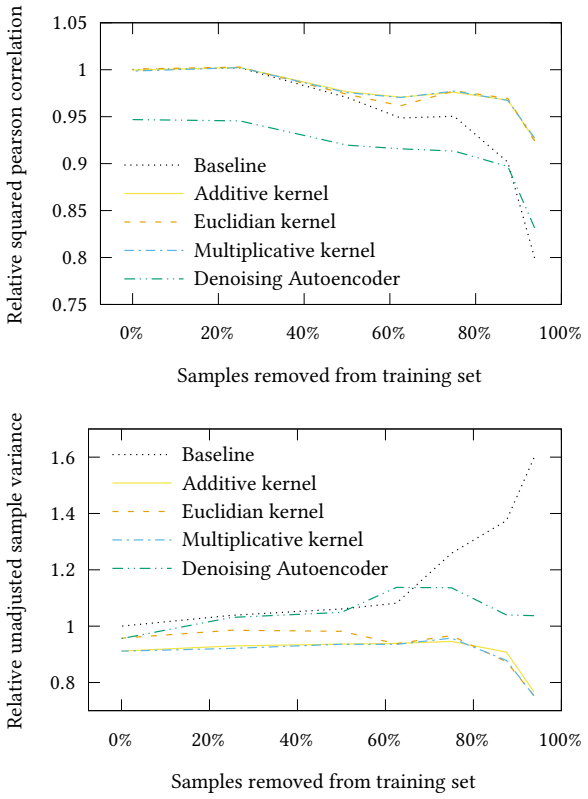

Figure 3: Results from applying the proposed method to a multi-layer feed-forward network.

Scenario	Base.	Add.	Eucl.	Mult.	DAE
CFN full dataset	1.00	1.01	0.99	**1.03**	0.81
CFN 60% removed	0.35	0.62	0.47	**0.69**	0.61
MLFF full dataset	1.00	0.99	**1.01**	0.99	0.95
MLFF 94% removed	0.80	0.93	0.92	**0.93**	0.83

(a) Relative squared pearson correlation. Higher is better.

Scenario	Base.	Add.	Eucl.	Mult.	DAE
CFN full dataset	1.00	0.69	0.80	**0.53**	2.03
CFN 60% removed	3.34	1.37	1.19	**0.93**	2.52
MLFF full dataset	1.00	0.91	0.96	**0.91**	0.96
MLFF 94% removed	1.60	0.77	0.75	**0.75**	1.03

(b) Relative unadjusted sample variance. Lower is better.

Table 1: Summary of measurement results, each relative to the value of their model's corresponding baseline measurement with the full training set available.

pearson correlation. With so little labelled data available, all models trained using the proposed method, regardless of combination kernel, are outperforming the prediction accuracy of the baseline model, still retaining up to 69% of the squared pearson correlation. Models trained on the output of the denoising autoencoder show a lower prediction accuracy compared to the baseline model, even at the full dataset being available. However, models trained on the output of the denoising autoencoder are much less subject to a drop in prediction accuracy when less data is available, allowing the model trained on the output of the denoising autoencoder to outperform the baseline when only 37.5% of the 2 000 samples are available.

The second graph shows the unadjusted sample variance of known-equal outputs of the model, relative to the baseline network, where lower is better. The data shows that in addition to increasing the prediction accuracy of the model when the full dataset is available, the proposed method is able to reduce the unadjusted sample variance. When the full dataset is available, the proposed method is able to reduce the unadjusted sample variance to 53%. Similarly to the prediction accuracy, the proposed method is able to keep the unadjusted sample variance lower, compared to the baseline network which sees an increase to 330% of the unadjusted sample variance of the full dataset. While the network trained on the denoising autoencoder has a higher unadjusted sample variance compared to the baseline when the full dataset is available, it is able to outperform the baseline when only 37.5% of the 2 000 samples are available.

Figure 3 shows the results of applying the proposed method to the comparison multi-layer feed-forward network. As the multi-layer feed-forward network has significantly fewer learnable parameters compared to the continuous feature network, the drop in prediction accuracy of the baseline model is not as steep as for the continuous feature network when less labelled training data is available. However, a drop in prediction accuracy can be observed when the amount of labelled data points is reduced below 12.5% of the 2 000 samples. As the data shows, the models trained using the proposed methods are able to improve the prediction accuracy for these cases to a level close to when the full dataset is available. When used with a simple model with few parameters as the multi-layer feed-forward network used, the type of combination kernel is shown to be of little importance for the achieved prediction accuracy. For the multi-layer feed-forward network, the model trained on the output of the denoising autoencoder is only able to achieve an improvement over the baseline for cases with very few data points available.

The graph of the unadjusted sample variance of the different models shows that the proposed method is able to reduce the observed unadjusted sample variance for all investigated amounts of labelled data points. While all kernels show an improvement over the baseline, the euclidian kernel performs worse than the other kernels for a larger amount of data points available. For very few data points, the proposed method will achieve a lower unadjusted sample noise in the neural network compared to when applied with many data points, at the cost of reduced prediction accuracy. This effect is observable regardless of the loss combination kernel used. While the autoencoder-trained network is only able to achieve an improvement in prediction accuracy compared to the baseline for very few data points available, it is capable of keeping the unadjusted sample variance low. While the unadjusted sample variance of the baseline network increases once less than 40% of the 2 000 labelled samples are available, the autoencoder is able to keep the

unadjusted sample variance roughly comparable to the baseline with the full dataset available.

6 CONCLUSION AND FUTURE WORK

This article proposes a novel method to utilize datasets containing known-equal but otherwise unlabelled datapoints in order to augment supervised learning. When such a known-equal dataset exists, the proposed method can be used to improve learning such that the number of labelled samples required to train a neural network to a similar level of prediction accuracy is reduced. In addition, the method can reduce the fluctuations in the predictions on data from repeated measurements of the same sample. The nature of the proposed method only requires modification of the training loop and the loss calculation to utilize known-equal datasets. As no modification of the actual model or its inputs is required, it is thus easily retrofittable into existing model architectures.

The proposed method is evaluated on the use case of predicting the carotenoid concentration in human skin using an optical MSRRS-based sensor. It was shown that, for multiple different neural network architectures, the method enabled a retainment of 95% of the squared pearson correlation with a reduction in dataset size for which the baseline model drops to 80% of the squared pearson correlation. Similarly, the proposed method is capable of reducing the unadjusted sample variance of measurements taken from the same sample to only 53% of the baseline, while being able to keep the unadjusted sample variance close to the baseline value when fewer labelled data points are available, while the same neural network without the proposed method saw an increase of the unadjusted sample variance to up to 330% of the baseline value.

As the proposed method has been evaluated using a regression task, evaluating the efficacy of the proposed method for classification-based tasks may be of interest.

ACKNOWLEDGMENTS

Supported by the state Hessen through funding as part of the Distr@l research grant 22_0041_2B.

REFERENCES

[1] Sumit Chopra, Raia Hadsell, and Yann LeCun. 2005. Learning a similarity metric discriminatively, with application to face verification. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*. Vol. 1, 539–546 vol. 1. DOI: 10.1109/CVPR.2005.202.

[2] Roberto Cipolla, Yarin Gal, and Alex Kendall. 2018. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE Computer Society, Los Alamitos, CA, USA, (June 2018), 7482–7491. DOI: 10.1109/CVPR.2018.00781.

[3] Maxim E Darvin, Björn Magnussen, Jürgen Lademann, and Wolfgang Köcher. 2016. Multiple spatially resolved reflection spectroscopy for in vivo determination of carotenoids in human skin and blood. *Laser Physics Letters*, 13, 9, (Aug. 2016), 095601. DOI: 10.1088/1612-2011/13/9/095601.

[4] Matthias Ehrgott. 2006. *Multicriteria Optimization.* (2nd ed.). Springer Berlin, (Jan. 2006). ISBN: 978-3-540-27659-3. DOI: 10.1007/3-540-27659-9.

[5] Andrew Goldberg, Xiaojin Zhu, Aarti Singh, Zhiting Xu, and Robert Nowak. 2009. Multi-manifold semi-supervised learning. In *Proceedings of the Twelth International Conference on Artificial Intelligence and Statistics* (Proceedings of Machine Learning Research). David van Dyk and Max Welling, (Eds.) Vol. 5. PMLR, Hilton Clearwater Beach Resort, Clearwater Beach, Florida USA, (Apr. 2009), 169–176. https://proceedings.mlr.press/v5/goldberg09a.html.

[6] Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*. Yoshua Bengio and Yann LeCun, (Eds.) DOI: 10.48550/arXiv.1412.6980.

[7] Jaakko Lehtinen, Jacob Munkberg, Jon Hasselgren, Samuli Laine, Tero Karras, Miika Aittala, and Timo Aila. 2018. Noise2noise: learning image restoration without clean data. DOI: 10.48550/arXiv.1803.04189.

[8] Birk Martin Magnussen, Claudius Stern, and Bernhard Sick. 2023. Utilizing continuous kernels for processing irregularly and inconsistently sampled data with position-dependent features. In *Proceedings of The Nineteenth International Conference on Autonomic and Autonomous Systems*. IARIA. ThinkMind, (Mar. 2023), 49–53. http://www.thinkmind.org/index.php?view=article%5C&articleid=icas%5C%5F2023%5C%5F1%5C%5F90%5C%5F20043.

[9] Adam Paszke et al. 2019. Pytorch: an imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*. H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, (Eds.) Vol. 32. Curran Associates, Inc. DOI: 10.5555/3454287.3455008.

[10] Florian Schroff, Dmitry Kalenichenko, and James Philbin. 2015. Facenet: a unified embedding for face recognition and clustering. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 815–823. DOI: 10.1109/CVPR.2015.7298682.

[11] Marco Taboga. 2021. Lectures on probability theory and mathematical statistics. In chap. Unadjusted sample variance. https://www.statlect.com/glossary/unadjusted-sample-variance.

[12] Yonglong Tian, Dilip Krishnan, and Phillip Isola. 2020. Contrastive multiview coding. In *Computer Vision – ECCV 2020*. Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, (Eds.) Springer International Publishing, Cham, 776–794. ISBN: 978-3-030-58621-8.

[13] Andreas Veit, Neil Alldrin, Gal Chechik, Ivan Krasin, Abhinav Gupta, and Serge Belongie. 2017. Learning from noisy large-scale datasets with minimal supervision. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE Computer Society, Los Alamitos, CA, USA, (July 2017), 6575–6583. DOI: 10.1109/CVPR.2017.696.

[14] Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. 2008. Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th International Conference on Machine Learning* (ICML '08). Association for Computing Machinery, Helsinki, Finland, 1096–1103. ISBN: 9781605582054. DOI: 10.1145/1390156.1390294.

[15] Tong Xiao, Tian Xia, Yi Yang, Chang Huang, and Xiaogang Wang. 2015. Learning from massive noisy labeled data for image classification. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2691–2699. DOI: 10.1109/CVPR.2015.7298885.