Evaluating the multi-task learning approach for land use regression modelling of air pollution

Andrzej Dulny
University of Würzburg
Germany
andrzej.dulny@stud-mail.uni-wuerzburg.de

Michael Steininger
University of Würzburg
Germany
steininger@informatik.uni-wuerzburg.de

Florian Lautenschlager
University of Würzburg
Germany
lautenschlager@informatik.uni-wuerzburg.de

Anna Krause
University of Würzburg
Germany
anna.krause@informatik.uni-wuerzburg.de

Andreas Hotho
University of Würzburg
Germany
hotho@informatik.uni-wuerzburg.de

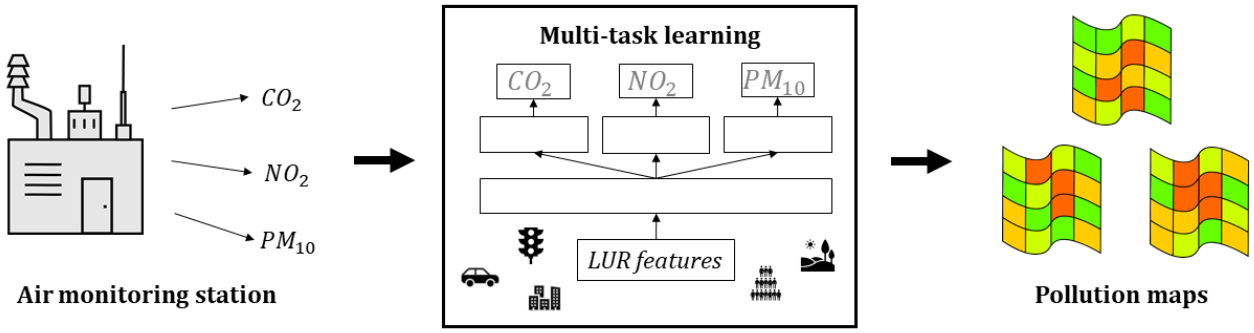


Figure 1: Air monitoring stations collect data from different pollutants. This opens up the possibility of modelling multiple pollutants simultaneously using multi-task learning. We evaluate this approach using multilayer perceptron models. The resulting multi-task learning models provide more accurate pollution maps than similar single task-learning models.

ABSTRACT

Air pollution has been linked to several health problems including heart disease, stroke and lung cancer. Modelling and analyzing this dependency requires reliable and accurate air pollutant measurements collected by stationary air monitoring stations. However, usually only a low number of such stations are present within a single city. To retrieve pollution concentrations for unmeasured locations, researchers rely on land use regression (LUR) models. Those models are typically developed for one pollutant only. However, as results in different areas have shown, modelling several related output variables through multi-task learning can improve the prediction results of the models significantly.

In this work, we compared prediction results from single-task and multi-task learning multilayer perceptron models on measurements taken from the OpenSense dataset and the London Atmospheric Emissions Inventory dataset. LUR features were generated from OpenStreetMap using OpenLUR and used to train hard parameter sharing multilayer perceptron models. The results show multi-task learning with sufficient data significantly improves the performance of a LUR model.

CCS Concepts

•**Applied computing** → **Environmental sciences;**
•**Computing methodologies** → **Multi-task learning;** *Neural networks;* Supervised learning by regression;

Keywords

Multi-task learning, Land-use regression, LUR, Environmental modeling, Air pollution

1. INTRODUCTION

Evidence suggests that air pollution has adverse effects on health [1, 2, 3] and the environment [4]. In order to research, assess and prevent these effects, availability of high quality measurements of air pollutants is necessary. Official authorities maintain stationary air quality monitoring networks, mostly equipped with sensors for multiple pollutants [5, 6]. However, such stations only offer data from a limited number of locations, as usually only few stations are present within a single city and land-use regression (LUR) models have been used successfully to account for spatial variability within cities and for epidemiological analyses [7].

While LUR models are trained on only one pollutant, several research areas have shown the potential of training on multiple related target variables, so called multi-task learning [8].

Air pollution monitoring stations often measure concentrations of more than one pollutant and the high spatial and temporal correlation of air pollutant concentrations [9] suggests that the emissions of different pollutants depend on the same set of factors. Thus, the tasks of modelling the pollutants should be highly related and modelling them with a multi-task learning model might improve the accuracy of the predictions. However, this approach has not been assessed yet in the context of modelling air pollution. To evaluate it, the performance of a multilayer perceptron LUR model is compared between single-task and multi-task learning on two different datasets - measurements taken from the OpenSense project collected by low-cost, portable sensors in the city of Zurich and modelled concentrations taken from the London Atmospheric Emissions Inventory.

The contribution of this work is twofold: We (i) propose a new approach to developing LUR models using concentration data from multiple pollutants, which takes advantage of the available measurements as shown in Figure 1 and (ii) demonstrate the potential of the multi-task learning approach compared to traditional single-task models for LUR.

The work is structured as follows: In Section 2 we summarize the related work. Section 3 describes the air pollution datasets and LUR features used to develop the models. The selected multi-task learning framework is described in Section 4 and the experimental approach and the models in Section 5. In Section 6 we present our results and in Section 7 we discuss the advantages and limitations of the multi-task learning approach. Section 8 provides a conclusion and outlook for future work.

2. RELATED WORK

2.1 Land-Use Regression

LUR models are an active field of research as the public awareness of the health and environmental effect of air pollution grows. Such models have been developed for numerous large cities worldwide. A 2008 review collected models developed for several cities in Europe as well as the USA [10] and a recent review from 2017 includes LUR models for 16 different cities worldwide [11]. Within the European ESCAPE-Project aimed at assessing the long-term effects of air pollution on human health, models have been developed for 36 cities in Europe using a standardized approach of model selection for a linear regression [1].

Traditionally, linear regression has been used for LUR [10, 11], however, several other machine learning models have been proposed to increase the accuracy of the predictions and model non-linear relationships between the variables. Generalized additive models are one such example and they have been used to improve prediction scores in LUR models of nitrogen oxides (NO_x) in Southern California [12] and $PM_{2.5}$ models in Beijing-Tianjin-Hebei (BTH) region in China [13]. Brokamp et al. used random forest regression to improve concentration predictions of in the urban city of Cincinnati, Ohio [14], and in [15] this approach is used to model NO_2 concentration in a metropolitan area of Japan.

Models using neural networks for LUR have also been proposed. For example, Alam and McNabola compare linear LUR models with multilayer perceptron models, achieving better results with the latter [16], while Adams and Kanaroglou use multilayer perceptron LUR models to construct real-time air pollution health risk maps [17].

Steininger et al. use a deep learning neural network to model air pollutant concentrations directly from globally available map images [18] and Lautenschlager et al. use features generated from geographical information available in the OpenStreetMap databank [20] to develop models performing better than similar models using features from local or closed sources [19].

In all of these studies, only one pollutant is predicted with a single model. It has been shown that different air pollutants show high temporal and spatial correlation patterns [21] and thus, the tasks of modelling different air pollutants can be highly related. The goal of this work is to explore the possibility of achieving better prediction results using a multi-task learning framework.

2.2 Multi-task Learning

Multi-task learning is a machine learning paradigm in which several related tasks are modelled simultaneously. A shared representation is used to guide the models to the most relevant features, thus potentially improving generalization and performance [8, 22]. It has been shown to increase effectiveness of machine learning models in a wide range of fields.

Collobert and Weston proposed a multi-task learning approach for natural language processing, in which several speech related predictions are made using a single neural network [23]. Gibert et al. use a multi-task learning framework to automatically detect anomalies for railway track inspections using machine vision. The multi-task model performs with increased accuracy as compared to single-task detectors [24].

Ramsundar et al. use a multi-task framework to develop large-scale models in the field of drug discovery. The results show increasing prediction accuracy when additional tasks are added to the model and the shared representation learned by the models can be transferred to other tasks, which were not used during training [25].

Caruana [8] explored the direct comparison of single-task models and multi-task models, the latter achieving better results on problems including autonomous driving simulations, recognizing knobs on images of doors and predicting the severity of pneumonia.

Multi-task learning has been applied in several fields where multiple related tasks are modelled, performing better than using single-task models separately. However, its application in the context of air pollution modelling has not been assessed. This work is aimed at filling this gap, by comparing single-task LUR models with models used to predict several air pollutants at the same time.

We used a multilayer perceptron hard parameter sharing model for multi-task learning, as it is the most commonly used approach in other applications and because it allows for a direct comparison within a single framework.

3. MATERIALS

In this section, the data sources used for the evaluation of the multi-task learning approach are introduced: the OpenSense dataset collected during a mobile sensing campaign in Zurich [26] and the London Atmospheric Emission In-

ventory [27], which contains a dataset developed using an atmospheric dispersion model and is published by London authorities. Furthermore, the LUR features which have been used to develop single-task and multi-task learning models are discussed in this section.

3.1 OpenSense Dataset

The OpenSense Project collected pollution data over the period of several years between 2012 and 2016 from mobile, low-cost sensing units equipped with an ultrafine particle (UFP) sensor, carbonmonoxide (CO) sensor and ozone (O_3) sensor placed on top of ten street cars, travelling on regular routes within the city of Zurich. The particulate matter pollution was sampled every 5 s and the O_3 and CO concentrations every 20 s [28]. A GPS signal receiver provided spatial information about the measurements. The gas sensors were equipped with water and dust covers to minimize possible interference [28].

3.1.1 Data selection

For creating the LUR models, measurements from the year 2014 have been selected from the dataset. Although there are certainly LUR models being developed for smaller time scales using additional weather information as features ([29], [30]), the most common approach is to consider a long time period for averaging the measurements. This removes any possible seasonal trends, which have a considerable influence on air pollution [31, 32]. Additionally, aggregated means are important from a regulatory perspective, as for example the European Commision enforces limits on annual averages for emissions of air pollutants [33]. We used the concentration data collected during the year 2014, with the exception of CO, where measurements were not available for the first two months of 2014. Instead, to maintain a comparable representation of all seasons, CO was averaged using the period of one year starting from 03/2014. Table 1 summarizes the data that has been selected from the OpenSense dataset to develop the models.

Table 1: Subset of the OpenSense dataset considered.

Pollutant	Start	End	Samples
UFP	01/2014	12/2014	11.9 Mio
O_3	01/2014	12/2014	5.3 Mio
CO	03/2014	02/2015	19.7 Mio

3.1.2 Data preprocessing

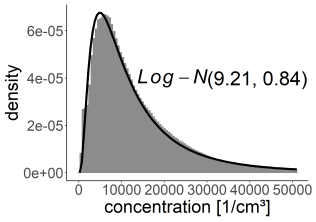
The UFP dataset has been properly calibrated and filtered and thus contains accurate measurements [34]. Reference measurements and internal variables of the sensors are not available for the ozone and carbonmonoxide datasets and thus a null-offset calibration cannot be done. The concentration measurements for those two pollutants are therefore taken from the factory pre-calibrated sensors without additional calibration. Following Hasenfratz et al., an initial GPS-filter is applied to assure an accurate geo-tagging of the concentrations, based on the horizontal dilution of precision (HDOP) [29]. Hasenfratz et al. discarded all measurements with a HDOP of smaller than 3. To obtain an even more accurate positioning we used the threshold of 2. Additionally, all measurements taken outside of the boundaries of the routes taken by the street cars are also discarded. The remaining samples (97.1% of the UFP measurements, 96.8% of the CO measurements and 96.9% of the O_3 measurements) were used for further processing.

3.1.3 Data aggregation

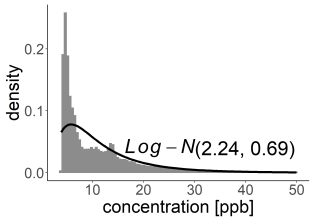
Following [29], where a 100 m × 100 m grid was used to develop a LUR model with the OpenSense UFP data, annual averages for the same spatial resolution of 100 m were calculated. Because the measurements were taken by mobile sensors, there was a considerable variation in the number of observations in each cell, ranging from 1 to over 300 000 for the 100 m grid. To ensure that the mean annual concentrations are reliable and to exclude possible outliers due to positioning errors, cells with less than 50 measurements were discarded.

3.1.4 Data evaluation

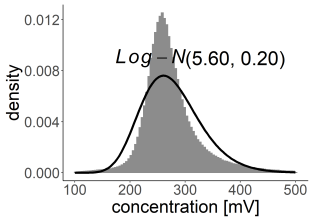
Data quality for air pollutants can be assessed using a well established observation that the measurements of air pollutants approximately follow a log-normal distribution [35]. Figure 2 shows the empirical distribution of the raw measurements in comparison to a theoretical log-normal distribution with the same mean and standard deviation. The relatively close fit of both distributions for the UFP indicates that the measurements are reliable. This does not hold true for the CO and O_3 data, where there are substantial deviations from the log-normal distribution. The poor fit of the log-normal distribution function to the concentration measurements for CO and O_3 indicates poor data quality, which can be attributed to the sensors not being adequately calibrated. A proper calibration would require accurate reference data for a wide range of humidity and weather conditions, which is not available for this dataset. The variability of atmospheric conditions in which the mobile sensors have been used can thus result in the measurements not being accurate [34].



(a) Ultrafine particles



(b) Ozone



(c) Carbonmonoxide

Figure 2: Distribution of the concentrations of ultrafine particles, ozone and carbon monoxide in the year 2014 as compared to a log-normal distribution. Based on the OpenSense dataset.

3.1.5 Summary

The previous analysis of the OpenSense dataset suggests that the measurements of CO and O_3 might be noisy and the question of whether an evaluation of multi-task learning can be done on such a dataset should be addressed. While a dataset containing measurements from well calibrated sensing units would serve this purpose better, to our best knowledge there is no such a dataset that also contains enough samples to enable training multilayer perceptron models. The measurements of UFP from the OpenSense dataset have been used successfully to develop LUR models before [29, 19] and for this reason, while accepting the limitations of using noisy measurements of the other pollutants, we decided to include the OpenSense dataset in our analysis.

The OpenSense datasets that have been used here can be accessed online: [36] for the UFP dataset and [37] for the CO and O_3 datasets.

3.2 London Atmospheric Emissions Inventory

The London Atmospheric Emissions Inventory (LAEI) is a data collection containing estimates of pollutant emissions and their sources for a given year in the city of London. The input factors include traffic data from road and rail networks, domestic and commercial fuel consumption, aviation, and pollution from individual industrial sites. The emission data is used to model ground-level average yearly concentrations of air pollutants on a $20\,\text{m} \times 20\,\text{m}$ grid using a atmospheric dispersion model.

In this work, we used the 2013 version of the inventory to develop LUR models for multiple pollutants: nitrogen dioxide (NO_2), nitrogen oxides (NO_x), particulate matter of diameter less than $10\,\mu\text{m}$ (PM_{10}), number of days with a daily mean PM_{10} concentration greater than $50\,\mu\text{g}\,\text{m}^{-3}$ (PM_{10d}), and particulate matter of diameter less than $2.5\,\mu\text{m}$ ($PM_{2.5}$). It is important to stress that the LAEI contains modelled annual mean concentrations and not measurements from air monitoring stations. However, the inventory has been used for LUR modelling, as for example Steininger et al. developed deep learning LUR models using the concentration data for NO_2 from the LAEI [18].

3.3 Features

LUR features where generated using the OpenLUR approach [19]. Starting from a given point, the total area of commercial, industrial and residential buildings within a pre-defined radius can be computed using geographical information from OpenStreetMap. Additionally, the total length of roads of any type and the distance to the closest traffic signal, motorway, primary road and industrial area can be computed using such data. In total 244 features were generated this way, which are summarized in Table 2.

3.3.1 Selection

Feature selection is a systematic method of selecting the variables upon which to build the model. Selecting only the relevant features ensures that the model is easily interpretable and improves the performance of the model by enhancing generalization [38]. We used a selection method based on the best performing features on linear models, similar to [1].

The features are selected iteratively from the pool of all features based on the best improvement. First, a simple linear regression model is fitted for every feature. The feature with the best average R^2 score over all pollutants is selected. Next, for each of the remaining features, a multiple linear regression model containing this feature and all previously chosen features is evaluated and the average R^2 score is calculated. Following [1], the score is then compared to the average R^2 score of the linear model containing only the previously selected features. If the score improvement of the linear model by including the feature is larger than 1%, the feature with the biggest improvement is added to the pool of selected features and the procedure is repeated. If the condition is not met, the feature is not included and the feature selection ends.

Features were selected separately for the OpenSense and LAEI. In total 13 features have been selected for the OpenSense data at and 3 features for the London Atmospheric Inventory as shown in Table 3. The order in which the features are listed corresponds to the order in which features have been selected by the procedure.

Table 2: Features generated for each dataset using the OpenLUR approach.

Feature type	Measurement	Radii (in 50 m steps)
commercial area	total area	50 m - 3000 m
industrial area	total area	50 m - 3000 m
residential area	total area	50 m - 3000 m
large road	total length	50 m - 1500 m
small road	total length	50 m - 1500 m
traffic signals	distance	-
motorway	distance	-
primary road	distance	-
industrial area	distance	-

Table 3: Results of feature selection for both datasets in the order in which features have been selected by the procedure.

(a) Features selected for the OpenSense data

Features selected
residential area within 1550 m
length of large roads within 1500 m
distance to the closest primary road
residential area within 700 m
length of large roads within 850 m
length of large roads within 100 m
distance to the closest industrial area
residential area within 3000 m
residential area within 2950 m
industrial area within 1750 m
industrial area within 3000 m
commercial area within 3000 m
industrial area within 2550 m

(b) Features selected for the LAEI data

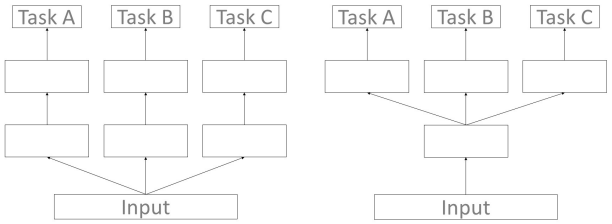
Features selected
length of large roads within 50 m
residential area within 2150 m
distance to the closest traffic signal

4. METHOD

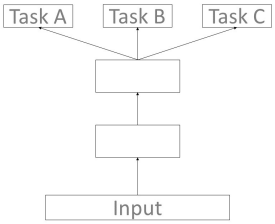
In this work, a hard-parameter sharing multi-task learning approach is implemented using a multilayer perceptron model with two hidden layers. Multilayer perceptron models have been applied successfully in LUR to model single pollutants [16]. It allows for a straightforward translation into a multi-task learning framework by providing additional outputs and thus a relatively direct comparison. When hidden layers are shared between outputs, the network is forced to learn a shared representation between the tasks which reduces the risk of overfitting [39, 22], possibly improving the performance of the model.

4.1 Network structure

For a direct comparison between multi-task learning and single-task learning, the model's overall structure is kept constant while varying the number of shared layers. By changing the number of shared layers, it is possible to manipulate the degree of multi-task learning. This enables defining a fully multi-task learning model when all of the hidden layers are shared between pollutants, as well as a single-task model if all of the layers are task-specific. Additionally, a model with one shared layer and one task-specific layer can be defined. The structure of the three different models used for the evaluation is displayed in Figure 3.



(a) 0 shared layers (b) 1 shared layer



(c) 2 shared layers

Figure 3: Models with different numbers of shared layers considered for the comparison of multi-task learning and single-task learning.

4.2 Training

All training of the multilayer perceptron models was performed using the Keras library version 2.3.1 [40]. The Adam optimizer was used for the weight updating with the default learning rate of 0.001. The mean squared error is used as the loss function and additionally the R^2-score is monitored. An early stopping condition is used to determine the end of the training before the maximum number of epochs set to 2000. After each epoch, the improvement of the mean squared error on the validation data is checked and if the score did not improve during the last 20 epochs the training stops and the best performing weights are restored. The final R^2-score of the model on the validation data is calculated and averaged over all cross-validation sets. The same cross-validation division was used for training and scoring all models. This score is then used to select the best performing model.

5. EXPERIMENTS

In order to evaluate the multi-task learning approach the data is split into several training and evaluation sets and used to train the baseline and experimental models. This procedure and the baseline models are described in the following.

5.1 Data selection

In total, 929 cells with annual mean concentrations were available from the OpenSense dataset and 5 851 915 from the LAEI. However, only 4500 cells were sampled from the large dataset for training and evaluating the models and the features have been calculated only for those measurements. The decision to only include a limited number of data points was made due to the high computational cost of obtaining LUR features from OpenStreetMap and to increase the generalizability of our evaluations, as datasets usually used for LUR only contain limited number of samples [29]. For a comprehensive evaluation of the multi-task learning approach, training sets of different sizes were included. For the OpenSense dataset samples of 100, 300 and 500 measurements were sampled uniformly as training sets to investigate the influence of the size of the training data on the performance of the multi-task learning models. The models for the London Atmospheric Emissions were trained using sample sizes of 100, 300, 500 and 3000. The measurements not included in the training set were used to evaluate the resulting models to obtain the final score. All model types (including the baseline models) were trained using the same training set and evaluated using the same test set and used the same cross-validation division for all models. In total 7 training sets were created as summarized in Table 4.

Table 4: Training and test sets on which all models were evaluated.

Source	Training Samples	Test Samples
OpenSense	100	829
OpenSense	300	629
OpenSense	500	429
LAEI	100	4400
LAEI	300	4200
LAEI	500	4000
LAEI	3000	1500

5.2 Baseline

In order to evaluate the multi-task learning model and put the observed differences in context, the LUR models for the available datasets are first developed using traditional approaches - linear regression and random forest regression. The details on how models were trained and evaluated using both approaches are provided here.

5.2.1 Linear Model

Linear regression has been traditionally used in LUR models [41], it is therefore a good baseline to consider for the performance of other models. For each dataset, a linear model was fitted on the training set using the features selected with the algorithm described in Section 3.3.1. Each of the pollutants was modelled separately. The resulting models were then evaluated on the available test samples.

5.2.2 Random Forest

Random forest regression has been used for LUR models yielding good prediction results [14], which is the reason why it was included as a baseline. For each dataset, a random hyperparameter search is performed to find the optimal model. Table 5 shows the hyperparameters included in the search. The remaining hyperparameter for the model use the default values provided by the scikit-learn library in version 0.22.1. The mean R^2-score from a ten-fold cross-validation was used to select the best performing model, which was then fitted on the whole training set and evaluated using the test set.

Table 5: Hyperparameters searched for optimizing the random forest model

Hyperparameter	Min	Max
Number of trees	10	2000
Fraction of features	0	1
Fraction of data	0	1
Minimum samples	2	21

5.3 Hyperparameters

To evaluate the multi-task learning models, a hyperparameter optimization procedure was implemented for each of the training sets to find the best performing models of each structure. All models have two hidden layers, each of which contains the same number of neurons. They differ only in the number of layers shared between the different pollutants.

During each step of the hyperparameter optimization, a set of hyperparameters was sampled from a predefined hyperparameter space shown in Table 6. For the other hyperparameters, the default values provided by the Keras library version 2.3.1 were used during the training [40]. All three models with different degrees of multi-task learning were trained using this set of hyperparameters and evaluated using a ten-fold cross-validation method, similarily to the training of the random forest regression models. The subsample of the training set left-out by the given cross-validation iteration is used as validation data for monitoring the performance of the model during training and for calculating the final score.

Table 6: Hyperparameters searched for optimizing the multilayer perceptron models

Hyperparameter	Min	Max	Distribution
Neurons per layer	5	200	uniform
Dropout rate	0	0.8	uniform
L_2-regularization	0.0001	1	log-uniform

6. RESULTS

In this chapter we present the results of the different LUR models. This includes the results of the baseline models and the comparison of different multi-task learning models and single-task learning models. The models were trained using the best found hyperparameters and evaluated using the test dataset which was not used before. The same ten-fold cross-validation division of the training set used during hyperparameter search was applied to keep track of the R^2-score during training for the purpose of early stopping. Each cross-validation was performed 30 times in total. This was a compromise between the high computational cost of fitting the models and the requirements for an accurate estimate of the scores.

6.1 OpenSense dataset

6.1.1 Multi-task learning

Table 7 shows the average R^2-scores of models using different degrees of multi-task learning and single-task learning on the OpenSense dataset. Zero shared layers correspond to single-task learning, while with two or one shared layers features and activations of hidden layers are shared between pollutants thus corresponding to multi-task learning.

The bold scores in Table 7 indicate the best model for each training sample. The results show an improvement of the R^2-scores by using at least some shared representation as compared to single-task learning for all training samples considered.

Table 7: Average R^2-scores on the test samples from the OpenSense dataset using multilayer perceptron models with different numbers of shared layers. The increase is calculated between the single-task learning model (zero shared layers) and the best performing multi-task learning model (at least one shared layer).

	Shared layers			
Samples	0	1	2	Increase
100	0.224	0.169	**0.224**	+0.41%
300	0.410	**0.448**	0.391	+9.23%
500	0.463	**0.474**	0.379	+2.26%

The optimal structure of the model varies with the training sample considered as does the amount of improvement as shown in the increase percentage of the R^2-scores in Table 7. The one-way ANOVA performed for each training set individually shows that the modelling approaches differ significantly ($p < .001$).

6.1.2 Comparison

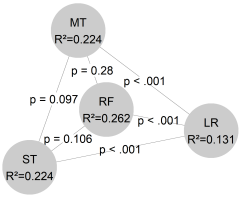
The comparison of the results of both baseline models, single-task learning models and the best multi-task learning models for the OpenSense dataset is shown in Table 8. The best performing model with at least one shared layer has been taken to represent the multi-task learning approach.

To check whether the resulting R^2-scores were significantly different, for each training sample the models were tested pairwise using the Mann-Whitney U-test. The resulting p-values are shown in Figure 4.

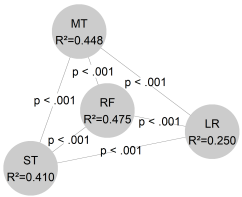
For 300 and 500 training samples, the random forest model performs significantly better than any other model. The linear models perform significantly worse than non-linear

Table 8: Average R^2-scores on the test samples from the OpenSense dataset using linear regression (LR), random forest regression (RF), as well as single-task learning (ST) and multi-task learning (MT) using a multilayer perceptron (MLP).

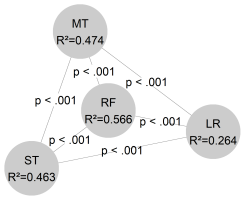
			MLP	
Samples	LR	RF	ST	MT
100	0.131	0.262	0.224	0.224
300	0.250	0.475	0.410	0.448
500	0.264	0.566	0.463	0.474



(a) 100 training samples



(b) 300 training samples



(c) 500 training samples

Figure 4: Pairwise Mann-Whitney U-tests between linear regression models (LR), random forest models (RF) and multilayer perceptron models using single-task (ST) and multi-task learning (MT) on the Opensense dataset.

models. For all considered samples, the multi-task learning model performs better than a similar multilayer perceptron single-task model. The difference is significant for the training samples of size 300 and 500. For the training sample of size 100, the difference is not statistically significant.

6.2 LAEI dataset

6.2.1 *Multi-task learning*

Table 9 shows the average R^2-scores of LUR models using different degrees of multi-task learning and single-task learning on the LAEI dataset. The best performing model for each training sample is in bold type.

Table 9: Average R^2-scores on the test samples from the LAEI dataset using multilayer perceptron models with different numbers of shared layers.

	Shared layers			
Samples	0	1	2	Increase
100	0.489	**0.490**	0.476	+0.32%
300	**0.506**	0.468	0.490	−3.09%
500	0.514	**0.515**	0.507	+0.18%
3000	0.522	0.528	**0.534**	+2.25%

The results show an increase of the R^2-scores when using multi-task learning for models trained with 100, 500 and 3000 samples, while for 300 samples the single-task model performs better.

Similarly to the OpenSense dataset, the results show that there is no one-fits-all optimal structure of the model, with the optimal amount of shared layers varying with the training sample considered. A clear increase of the R^2-score with increasing degree of multi-task learning can however be seen when using a large training set of 3000 samples. The one-way ANOVA performed for each training set individually shows, that the modelling approaches differ significantly ($p < .001$).

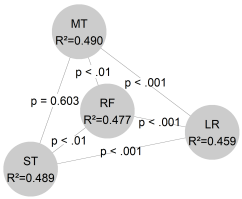
6.2.2 *Comparison*

The comparison between the different models for the LAEI dataset can be seen in Table 10.

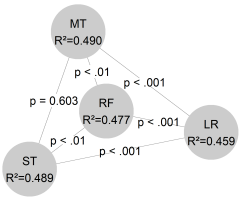
Table 10: Average R^2-scores on the test samples from the LAEI dataset using linear regression (LR), random forest regression (RF), as well as single-task learning (ST) and multi-task learning (MT) using a multilayer perceptron (MLP).

			MLP	
Samples	LR	RF	ST	MT
100	0.459	0.477	0.489	0.490
300	0.488	0.527	0.506	0.490
500	0.499	0.537	0.514	0.515
3000	0.505	0.572	0.522	0.534

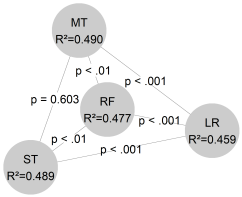
Similarly to the OpenSense dataset, the models have been compared using pairwise Mann-Whitney U-tests. The results are shown in Figure 5.



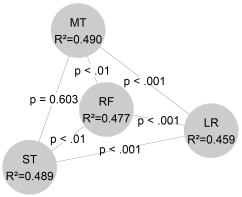
(a) 100 training samples



(b) 300 training samples



(c) 500 training samples



(d) 3000 training samples

Figure 5: Pairwise Mann-Whitney U-tests between linear regression models (LR), random forest models (RF) and multilayer perceptron models using single-task (ST) and multi-task learning (MT) on the London Atmospheric Emissions Inventory dataset

The comparison shows that the random forest model performs significantly better than other models and linear re-

gression offers the statistically significant worse fit.

When comparing single-task and multi-task learning multilayer perceptron models, the results show an increased fit of the models using a multi-task learning approach compared to single-task models when trained using 100, 500 and 3000 samples. However, the difference is only statistically significant when trained with 3000 samples. When using 300 training samples, the multi-task learning model performs worse than the single-task learning model.

7. DISCUSSION

The comparison of all the single-task models, including the baselines, shows a clear advantage of the random forest models over all other model types. This holds true for both datasets and all examined sample sizes. It is not an unexpected result, as previous comparisons have shown that random forest models provide high accuracies in the context of LUR [14, 42].

When comparing multi-task learning with the single-task learning approach on the multilayer perceptron models the results for both the OpenSense and LAEI datasets indicate a possible increase in performance of the models when using a shared representation. However, the increase in performance is not large enough to surpass the random forest baseline model, which still outperforms the multi-task learning model.

In this section we discuss possible reasons for this limitation, what can be done to increase the benefits of multi-task learning and why it can still be a promising approach.

7.1 Task relatedness

Caruana [8] argues that multi-task learning helps improve generalization when using related tasks. Two tasks are defined to be related if they use the same variables to predict the outcome and if they use those variables in the same way [8].

Using this definition, it is possible to explore the relatedness of the tasks by comparing the relative feature importance between different pollutants. If two tasks (modelling two different pollutants) depend stronger on the same set of features and less so on different features, the tasks is considered highly related.

We used the permutation variable importance measure introduced by Fisher et al. [43] on the baseline random forest regression models to calculate the feature importance for all of the training samples. Figure 6 shows the feature importance calculated for the OpenSense dataset with 500 training samples and Figure 7 for the LAEI dataset with 3000 samples. For all the other training samples the calculations show a very similar pattern of feature importance.

For the OpenSense dataset, the tasks appear to be less related, as the feature importance values vary strongly between the pollutants (fig. 6). Figure 7 shows that all pollutants, except $PM_{2.5}$, depend similarly on the features. It is therefore reasonable to assume that the tasks of modelling different pollutants in this dataset are highly related.

The feature importance analysis does not paint a clear picture of how task relatedness translates into performance gain from shared representation. In our experiments, models on both datasets benefit from the multi-task approach even though the task relatedness, as measured by feature importance, is higher for the LAEI dataset.

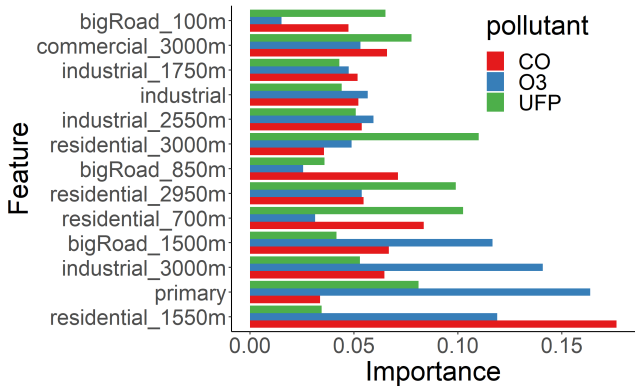


Figure 6: Feature importance for the OpenSense dataset with 500 samples.

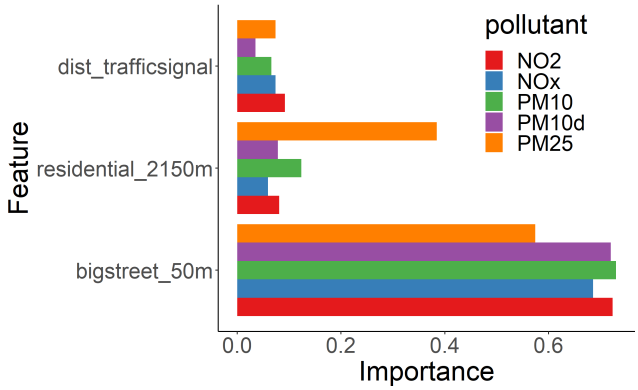


Figure 7: Feature importance for the London Atmospheric Emissions Inventory dataset with 3000 samples.

7.2 Feature selection

One possible explanation for this unclear relationship between task relatedness and the advantage of multi-task learning could be the used feature selection procedure. As described in Section 3.3.1, variables used for training the models have been selected from a large pool of 244 features generated from OpenStreetMap. The selection procedure involved comparing the R^2-scores of linear models built using each of the features and including the best one.

We used an average over all pollutants to calculate the score for each feature. An alternative approach, which can be explored in further research, would be to select important features for each pollutant individually and then consider an aggregate of those features for the multi-task learning. However, because a shared metric was used only features that could on average predict all pollutant concentrations reasonably well were included in the pool of variables to be used for the multi-task learning models.

This selection procedure had two important consequences: First, it introduced a bias to the feature importance metric as calculated in Section 7.1. All features that would offer very accurate models for only one of the pollutants, but not for the others, are not selected. Since only those features were selected that on average predicted concentrations for all pollutants reasonably well, the tasks are more related when comparing their feature importance than if a different

feature selection method was used.

Second, selecting features that on average predict all pollutants well is in itself a form of multi-task learning. In fact, sparsity-enforcing regularization techniques have been used for linear models in the context of multi-task learning [22, 44]. Arguably feature selection is also one of the core mechanisms how multi-task learning improves prediction scores in multilayer perceptron models [8]. While the multi-task learning models considered for modelling pollutants still benefit from a shared hidden representation, the single-task models are not truly independent, as they all depend on features that have been selected using a multi-task method, possibly decreasing the observed difference.

7.3 Data quality

As discussed in Section 3.1.4 measurements for two of the pollutants within the OpenSense dataset are possibly noisy and only one pollutant offers high quality measurements. In contrast, the LAEI offers estimated concentrations of air pollutants which are not directly measured, but instead modelled using a atmospheric dispersion model.

As neither of the datasets offers high-quality data from physical monitoring stations of air pollutant concentrations, the question arises of how well the findings would generalize to such a hypothetical dataset. While a definitive answer can only be given by examining multi-task learning on such a dataset, there are some arguments that can be made on why our approach would still work.

As can be seen on the OpenSense dataset, multi-task learning increases the fit of the model compared to a similar single-task learning model for all pollutants, including ultrafine particles for which high-quality measurements are available. Thus, since including noisy measurements can improve the prediction accuracy of high-quality data when modelling in a multi-task learning context, it is reasonable to believe that a similar effect would be observed if high-quality data was used as the additional tasks.

The LAEI dataset, on the other hand, offers only modelled concentrations. While air dispersion models will always offer a simplified model of the emissions and spread of air pollution, they generate accurate general trends, especially when only long averaging periods of one year or more are considered. Thus, a similar benefit of multi-task learning can be expected when accurately measured data is used.

Both single- and multi-task models are trained on equal data quality. The results show that multi-task learning models offer better prediction performance than similar multilayer perceptron single-task models. It is, however, unclear how the difference would manifest when comparing the predictions from models trained on accurate air pollution data from high-end monitoring stations. Especially when the sources of error are not independent, the multi-task models might only learn the noise patterns in the data. While the findings on the OpenSense data indicate this not to be the case, additional experiments using poor-quality data with independent sources of error could rule out this possibility.

7.4 Sample size

It is a known observation in machine learning that small sample sizes often lead to overfitting, especially when using complex models like artificial neural networks as compared to traditional models (e.g. linear regression) [45, 46]. This limitation makes applying complex models in the context of LUR difficult, since high-quality measurements are often a limited resource as mentioned in Section 1.

Our results clearly confirm this pattern, as more training samples lead to better prediction scores for both datasets and all considered model types. When comparing multi-task and single-task learning models, the advantage of a shared representation approach only becomes apparent with sufficient training data. For the LAEI dataset the largest positive effect appears for 3000 training samples, while with less data single-task models do not differ significantly from multi-task models or perform even better. Multi-task learning shows a significant advantage for the 300 and 500 samples subset of the OpenSense dataset.

Large data requirements make the application of multi-task learning models for LUR difficult, as datasets containing air pollutant concentrations usually contain limited numbers of samples.

8. CONCLUSIONS AND FUTURE WORK

In this work, we assessed multi-task learning for LUR. As pollutants are often monitored together, the potential dependence on the same set of factors makes modelling several pollutants simultaneously an attractive possibility. The results do indeed show that multi-task learning models perform significantly better than similar multilayer perceptron single-task learning models when using a large enough training set.

However, for both datasets that have been considered - the London Atmospheric Emissions Inventory and the OpenSense dataset, random forest regression still outperforms the multi-task learning models for all training samples. A possible direction for future research is the application of multi-task learning using tree-based ensemble methods [47]. Non-parametric ensemble models might overcome the large data requirements of multilayer perceptron models while still benefiting from shared information between different pollutants.

In order to decrease the data requirements for multilayer perceptron models it might be worthwhile to explore pre-training with weak labels. Interpolation methods may be used to produce dense maps of pollution estimates from measurements which can be used as weak labels. These labels can be used to train the model. Thereafter, the model can be fine-tuned using only labels from real measurements. This training procedure might improve multi-layer perceptron model performance for LUR, where there are often relatively few data points.

Another promising research direction is the application of multi-task learning for deep-learning based LUR models like MapLUR [18]. This model has shown better performance than random forests on the dataset of the London Atmospheric Emissions Inventory for single-task learning and our results suggest that multi-task learning can further increase performance.

Future work should also explore different feature selection methods, as more liberal selection procedures might allow for higher variability in feature dependence between different pollutants and consequently multi-task learning might benefit even more from a shared representation. Especially sparsity-enforcing regularization techniques used for multivariate linear regression [22, 44] might be a promising approach to building LUR models using multi-task learning.

As high-quality air pollution datasets mostly contain only a limited number of measurement locations, the experiments

have been performed on data obtained from low-cost sensors in the case of the OpenSense dataset and modelled air concentrations using an atmospheric dispersion model in the case of the London Atmospheric Emissions Inventory. An important direction for future research would be to compare multi-task learning and single-task learning on a large-scale dataset containing high-quality measurements.

Overall, the multi-task learning method using multilayer perceptrons shows better performance than similar single-task models, while still being outperformed by Random Forest models. However, this work demonstrates the potential of learning shared representations for better air pollution prediction performance, which can be explored in further research using different model types.

References

[1] R. Beelen *et al.*, "Effects of long-term exposure to air pollution on natural-cause mortality: an analysis of 22 European cohorts within the multicentre ESCAPE project," *Lancet*, vol. 383, no. 9919, pp. 785–795, Mar 2014.

[2] M. Bauer *et al.*, "Urban particulate matter air pollution is associated with subclinical atherosclerosis: results from the HNR (Heinz Nixdorf Recall) study," *J. Am. Coll. Cardiol.*, vol. 56, no. 22, pp. 1803–1808, Nov 2010.

[3] A. P. C. Takano *et al.*, "Pleural anthracosis as an indicator of lifetime exposure to urban air pollution: An autopsy-based study in Sao Paulo," *Environ. Res.*, vol. 173, pp. 23–32, 06 2019.

[4] E. E. Agency, *Air Quality in Europe - 2019 Report*. Luxembourg: Publications office of the European Union, 2019.

[5] P. Hystad *et al.*, "Creating national air pollution models for population exposure assessment in canada," *Environmental health perspectives*, vol. 119, no. 8, pp. 1123–1129, Aug 2011.

[6] BAFU and EMPA, "Technischer bericht zum nationalen beobachtungsnetz für luftfremdstoffe (nabel)," 2018.

[7] P. H. Ryan *et al.*, "A comparison of proximity and land use regression traffic exposure models and wheezing in infants," *Environmental Health Perspectives*, vol. 115, no. 2, pp. 278–284, 2007.

[8] R. Caruana, "Multitask learning," *Machine Learning*, vol. 28, no. 1, pp. 41–75, 1997.

[9] Y.-H. Dai and W.-X. Zhou, "Temporal and spatial correlation patterns of air pollutants in chinese cities," *PLOS ONE*, vol. 12, no. 8, pp. 1–24, 08 2017.

[10] P. H. Ryan and G. K. LeMasters, "A review of land-use regression models for characterizing intraurban air pollution exposure," *Inhalation toxicology*, vol. 19 Suppl 1, no. Suppl 1, pp. 127–133, 2007.

[11] H. Amini, M. Yunesian, V. Hosseini, C. Schindler, S. B. Henderson, and N. Künzli, "A systematic review of land use regression models for volatile organic compounds," *Atmospheric Environment*, vol. 171, pp. 1 – 16, 2017.

[12] L. Li, J. Wu, M. Wilhelm, and B. Ritz, "Use of generalized additive models and cokriging of spatial residuals to improve land-use regression estimates of nitrogen oxides in southern california," *Atmospheric Environment*, vol. 55, pp. 220 – 228, 2012.

[13] B. Zou, J. Chen, L. Zhai, X. Fang, and Z. Zheng, "Satellite based mapping of ground pm2.5 concentration using generalized additive modeling," *Remote Sensing*, vol. 9, no. 1, p. 1, Dec 2016.

[14] C. Brokamp, R. Jandarov, M. Rao, G. LeMasters, and P. Ryan, "Exposure assessment models for elemental components of particulate matter in an urban environment: A comparison of regression and random forest approaches," *Atmospheric Environment*, vol. 151, pp. 1 – 11, 2017.

[15] S. Araki, M. Shima, and K. Yamamoto, "Spatiotemporal land use random forest model for estimating metropolitan no2 exposure in japan," *Science of The Total Environment*, vol. 634, pp. 1269 – 1277, 2018.

[16] M. S. Alam and A. McNabola, "Exploring the modeling of spatiotemporal variations in ambient air pollution within the land use regression framework: Estimation of pm10 concentrations on a daily basis," *Journal of the Air & Waste Management Association*, vol. 65, no. 5, pp. 628–640, 2015.

[17] M. D. Adams and P. S. Kanaroglou, "Mapping real-time air pollution health risk for environmental management: Combining mobile and stationary air pollution monitoring with neural network models," *Journal of Environmental Management*, vol. 168, pp. 133 – 141, 2016.

[18] M. Steininger, K. Kobs, A. Zehe, F. Lautenschlager, M. Becker, and A. Hotho, "Maplur: Exploring a new paradigm for estimating air pollution using deep learning on map images," 2020.

[19] F. Lautenschlager, M. Becker, K. Kobs, M. Steininger, P. Davidson, A. Krause, and A. Hotho, "Openlur: Off-the-shelf air pollution modeling with open features and machine learning," *Atmospheric Environment*, vol. 233, p. 117535, 2020.

[20] OpenStreetMap contributors, "Planet dump retrieved from https://planet.osm.org ," https://www.openstreetmap.org, 2017.

[21] H. Guo, Y. Wang, and H. Zhang, "Characterization of criteria air pollutants in beijing during 2014–2015," *Environmental Research*, vol. 154, pp. 334 – 344, 2017.

[22] S. Ruder, "An overview of multi-task learning in deep neural networks," *CoRR*, vol. abs/1706.05098, 2017.

[23] R. Collobert and J. Weston, "A unified architecture for natural language processing: Deep neural networks with multitask learning," in *Proceedings of the 25th International Conference on Machine Learning*, ser. ICML '08. New York, NY, USA: Association for Computing Machinery, 2008, p. 160–167.

[24] X. Gibert, V. M. Patel, and R. Chellappa, "Deep multitask learning for railway track inspection," *IEEE Transactions on Intelligent Transportation Systems*, vol. 18, no. 1, pp. 153–164, Jan 2017.

[25] B. Ramsundar, S. Kearnes, P. Riley, D. Webster, D. Konerding, and V. Pande, "Massively multitask networks for drug discovery," 2015.

[26] K. Aberer, S. Sathe, D. Chakraborty, A. Martinoli, G. Barrenetxea, B. Faltings, and L. Thiele, "Opensense: Open community driven sensing of environment," in *Proceedings of the ACM SIGSPATIAL International Workshop on GeoStreaming*, ser. IWGS '10. New York, NY, USA: Association for Computing Machinery, 2010, p. 39–42.

[27] A. Q. T. G. L. Authority)., "London atmospheric emissions inventory (laei)." https://data.london.gov.uk/dataset/london-atmospheric-emissions-inventory-2013, 2013.

[28] J. J. Li *et al.*, "Sensing the air we breathe: The opensense zurich dataset," in *Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence*, ser. AAAI'12. AAAI Press, 2012, p. 323–325.

[29] D. Hasenfratz, O. Saukh, C. Walser, C. Hueglin, M. Fierz, and L. Thiele, "Pushing the spatio-temporal resolution limit of urban air pollution maps," in *2014 IEEE International Conference on Pervasive Computing and Communications (PerCom)*, March 2014, pp. 69–77.

[30] M. Johnson *et al.*, "Development of temporally refined land-use regression models predicting daily household-level air pollution in a panel study of lung function among asthmatic children," *Journal of Exposure Science & Environmental Epidemiology*, vol. 23, no. 3, pp. 259–267, 2013.

[31] D. Roberts–Semple, F. Song, and Y. Gao, "Seasonal characteristics of ambient nitrogen oxides and ground–level ozone in metropolitan northeastern new jersey," *Atmospheric Pollution Research*, vol. 3, no. 2, pp. 247 – 257, 2012.

[32] R. D. Peng, F. Dominici, R. Pastor-Barriuso, S. L. Zeger, and J. M. Samet, "Seasonal Analyses of Air Pollution and Mortality in 100 US Cities," *American Journal of Epidemiology*, vol. 161, no. 6, pp. 585–594, 03 2005.

[33] Council of European Union, "Directive 2008/50/ec of the european parliament and of the council," http://data.europa.eu/eli/dir/2008/50/2015-09-18, 2008.

[34] D. Hasenfratz, O. Saukh, and L. Thiele, "On-the-fly calibration of low-cost gas sensors," in *Wireless Sensor Networks*, G. P. Picco and W. Heinzelman, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 228–244.

[35] W. R. Ott, "A physical explanation of the lognormality of pollutant concentrations," *Journal of the Air & Waste Management Association*, vol. 40, no. 10, pp. 1378–1383, 1990.

[36] B. Maag, D. Hasenfratz, O. Saukh, Z. Zhou, C. Walser, J. Beutel, and L. Thiele, "Ultrafine particle dataset collected by the opensense zurich mobile sensor network," https://zenodo.org/record/3298842, Sep 2018.

[37] ——, "Ozone and carbon monoxide dataset collected by the opensense zurich mobile sensor network," https://zenodo.org/record/3355208, Jul 2019.

[38] M. L. Bermingham *et al.*, "Application of high-dimensional feature selection: evaluation for genomic prediction in man," *Scientific reports*, vol. 5, pp. 10 312–10 312, May 2015.

[39] J. Baxter, "A bayesian/information theoretic model of learning to learn via multiple task sampling," *Machine Learning*, vol. 28, no. 1, pp. 7–39, Jul 1997.

[40] F. Chollet *et al.*, "Keras," https://keras.io, 2015.

[41] G. Hoek, R. Beelen, K. de Hoogh, D. Vienneau, J. Gulliver, P. Fischer, and D. Briggs, "A review of land-use regression models to assess spatial variation of outdoor air pollution," *Atmospheric Environment*, vol. 42, no. 33, pp. 7561 – 7578, 2008.

[42] X. Hu, J. H. Belle, X. Meng, A. Wildani, L. A. Waller, M. J. Strickland, and Y. Liu, "Estimating pm2.5 concentrations in the conterminous united states using the random forest approach," *Environmental Science & Technology*, vol. 51, no. 12, pp. 6936–6944, 2017.

[43] A. Fisher, C. Rudin, and F. Dominici, "All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously," 2018.

[44] A. Argyriou, T. Evgeniou, and M. Pontil, "Convex multi-task feature learning," *Machine Learning*, vol. 73, no. 3, pp. 243–272, Dec 2008.

[45] A. Vabalas, E. Gowen, E. Poliakoff, and A. J. Casson, "Machine learning algorithm validation with a limited sample size," *PloS one*, vol. 14, no. 11, pp. e0 224 365–e0 224 365, Nov 2019.

[46] S. J. Raudys and A. K. Jain, "Small sample size effects in statistical pattern recognition: recommendations for practitioners," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 13, no. 3, pp. 252–264, 1991.

[47] J. SIMM, I. M. D. ABRIL, and M. SUGIYAMA, "Tree-based ensemble multi-task learning method for classification and regression," *IEICE Transactions on Information and Systems*, vol. E97.D, no. 6, pp. 1677–1681, 2014.