Deep Learning Versus Traditional Machine Learning Methods for Aggregated Energy Demand Prediction

Nikolaos G. Paterakis, Elena Mocanu, Madeleine Gibescu
Department of Electrical Engineering
Eindhoven University of Technology, Netherlands
{n.paterakis, e.mocanu, m.gibescu}@tue.nl

Bart Stappers, Walter van Alst
Balancing & Sourcing Department
Scholt Energy Control B.V., Netherlands
{bart.stappers, walter.van.alst}@scholt.nl

***Abstract*—In this paper the more advanced, in comparison with traditional machine learning approaches, deep learning methods are explored with the purpose of accurately predicting the aggregated energy consumption. Despite the fact that a wide range of machine learning methods have been applied to probabilistic energy prediction, the deep learning ones certainly represent the state-of-the-art artificial intelligence methods with remarkable success in a spectrum of practical applications. In particular, the use of Multi Layer Perceptrons, recently enhanced with deep learning capabilities, is proposed. Furthermore, its performance is compared with the most commonly used machine learning methods, such as Support Vector Machines, Gaussian Processes, Regression Trees, Ensemble Boosting and Linear Regression. The analysis of the day-ahead energy prediction demonstrates that different prediction methods present significantly different levels of accuracy in the case of a challenging dataset that comprises an interesting mix of consumers, wind and solar generation. The results show that Multi Layer Perceptrons outperform all the eight methods used as a benchmark in this study.**

***Index Terms*—deep learning, energy consumption, energy prediction, forecasting, machine learning**

I. Introduction

A. Motivation

Many approaches aiming to accurately and robustly predict the energy consumption have been proposed. In general, at the building level, two types of approaches are discerned. Models of the first type are based on physical principles to express their thermal dynamics and energy behavior. Depending on the type of building and the number of the considered parameters, these models might include space heating systems, natural ventilation, air conditioning systems, passive solar heating, photovoltaic systems, financial considerations, occupants' behavioral characteristics, etc. The second type of approaches is based on statistical methods. Such methods are used to predict energy consumption by correlating it with influencing variables related to weather and energy costs. Interested readers are referred to [1] and [2] for a more comprehensive discussion on the applications of artificial intelligence in building energy systems, and the more recent reviews [3], [4]. Moreover, to account for the evolution of future building energy systems, hybrid approaches which combine some of the aforementioned models to optimize predictive performance have been developed [5]–[8].

Predicting energy consumption at the building level is important for optimizing several parameters related to energy efficiency and economy in a localized context. From a different perspective, accurately forecasting the consumption of aggregated demand consisting of a portfolio of customers is an equally important problem for energy providers seeking to optimize their position in several market floors, as well as for distribution and transmission system operators in order to improve the short- and long-term management of the power system. Machine learning methods have been applied extensively to such problems too. Nevertheless, deep learning is a relatively new concept and comprises a palette of methods that have clear advantages over the traditional machine learning ones.

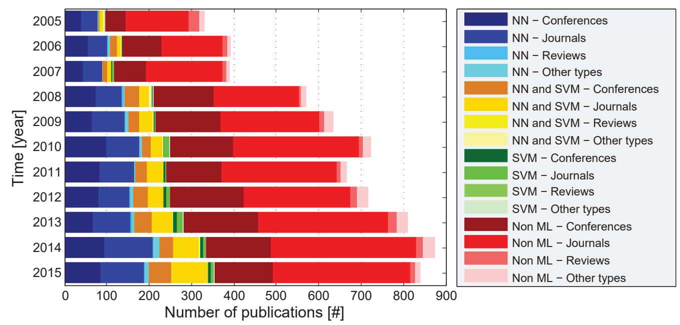

Fig. 1. A summary of the Scopus-indexed publications focusing on electricity prediction in the decade 2005-2015.

In an attempt to determine which machine learning techniques are the most popular and to highlight the integration of deep learning methods in the existing literature, a targeted bibliometric analysis of the collections of publications related to the electricity prediction problem is performed by querying the Scopus database. There are 6613 studies focusing on electrical energy prediction in the last decade, from which 839 where published in 2015. The distribution of the relevant literature based on publication type (conference papers, journal articles, reviews, other) is portrayed in Fig. 1. Moreover, these publications have been classified based on whether they propose the use of *machine learning* methods or not. Neural Networks (NNs) [9], [10] and Support Vector Machines (SVMs) [11] emerge as the most widely used machine learning methods in energy prediction. More specifically, NNs are used in 2380 publications and SVMs in 820, while 703 studies combine

978-1-5386-1953-7/17/$31.00 ©2017 IEEE

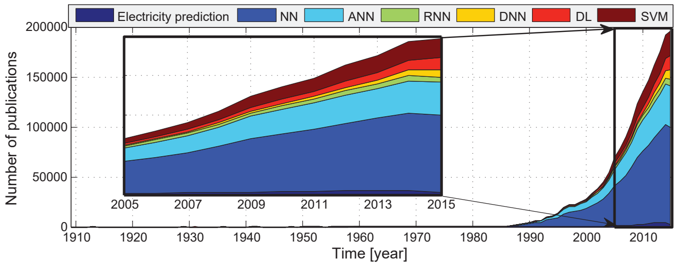

Fig. 2. Prediction - A summary of the Scopus-indexed publications focusing on prediction over the last century (i.e.1909-2015), including a zoom in over the decade 2005-2015.

both methods. Hidden Markov Models (HMM) [12] constitute another popular stochastic model for time series demonstrating good results in different application fields, ranging from bio-informatics to stock market price prediction; however, their use in the context of energy prediction has not been investigated much yet [13].

To provide a broader perspective, in Fig. 2 an overview of the evolution of machine learning methods applied to all types of prediction problems during the last century is presented, with a zoom in on the most productive last decade. This is a starting point to identify the state-of-the-art and at the same time a glimpse into potential future trends in energy prediction, which at the moment is the subject of only a fraction of the literature related to prediction. Evidently, the most popular machine learning methods are based on NNs and their variations, e.g. Artificial Neural Networks (ANN), Recurrent Neural Networks (RNN), Deep Neural Networks (DNN) or Deep Belief Networks (DBN). It is worth mentioning that up to 2016, the collection of publications indexed by Scopus that was related with NNs exceeds one million entries. From Fig. 2, it can be also observed that Deep Learning models (including DL proper and DNN) represent the most important trend in the last years [14].

The operation of modern power systems relies on increasing amounts of measurement data, both historical and real-time. In this brave new era of big data, a promising new technology is deep learning, quickly emerging as the most advanced set of solutions for large-scale applications. Currently, deep learning methods for energy prediction are in an incipient phase and a limited number of relevant studies have been published so far. The use of conditional Restricted Boltzmann Machines (CRBMs) was introduced by [15]. Moreover, Factored Conditional Restricted Boltzmann Machines (FCRBMs) have been used for building energy prediction [16]. When predicting aggregated demand, FCRBMs were used in a price-responsive context based on data from the EcoGrid project. A hybrid architecture was proposed, where the FCRBM was used in combination with a Gaussian Restricted Boltzmann Machine for feature selection [17]. Marino et al. [18] applied a DNN architecture to the short-therm building energy demand forecasting problem. A review of the applications of deep learning to intelligent buildings was presented by Manic et al. [19]. Apart from these studies, a relatively small number of publications in 2016 reinforce the interest in using deep learning methods for energy prediction. For example, Ryu et al. [20] applied the Long Short-Term Memory (LSTM) method to building energy prediction, while in [21] the LSTM was used for prediction as an integrated part of a solution to optimize a decentralized renewable energy system. The effectiveness of the LSTM was also exploited for short-term wind power forecasting [22], [23]. Finally, a Deep Belief Network (DBN) was combined with reinforcement learning to perform unsupervised building energy prediction in [24].

B. *Contribution*

Despite the fact that the number of applications of deep learning methods to address problems in the context of smart grids is increasing, a comparison between the most recently proposed methods has not been performed yet. In this respect the contribution of this study is twofold:

- studying the prediction capabilities of Multi Layer Perceptrons (MLPs) in the context of day-ahead aggregated energy consumption prediction for smart grids;
- analyzing the possibility of improving prediction accuracy when using deep learning methods, by performing a comparison with traditional ML methods, including Support Vector Regression (SVR), Ensemble Methods and Gaussian Processes, as well as the popular Linear Regression and Generalized Linear Regression methods.

The remainder of this paper is organized as follows: in Section II the methods that are compared in this study are listed and the proposed approach to the energy prediction problem is elaborated. In Section III an analysis of the dataset that is used is presented and prediction accuracy metrics are defined. Subsequently, in Section IV the results from the conducted experiments are reported. Finally, further discussion is provided in Section V and conclusions are drawn.

II. Benchmark Methods And Multi Layer Perceptrons

In this study a series of predictions are produced by using a range of benchmark prediction methods. First, the well-known Linear and Multiple Linear Regression methods are employed. Then, traditional machine learning methods including SVR, Regression Trees, Ensemble Trees and their variations such as Ensemble Regression Tree, Ensemble Tree Least Squares Regression Boost, as well as Guassian Processes [25] are used in order to investigate and understand the accuracy estimation performance using the specific database of aggregated demand profiles.

These methods are compared against MLPs [26], recently enhanced with deep learning capabilities. MLPs are classical feed forward ANN models that map a set of input data to the corresponding set of output data and therefore, are used for supervised learning. A MLP is composed of an input layer in which neurons represent the input data, an output layer in which the neurons represent the output data and an arbitrary number of hidden layers in between, with neurons representing the hidden, automatically discovered features of the input data.

The flow of information in MLPs is unidirectional, starting from the input layer towards the output layer. Furthermore, connections exist only between consecutive layers, while any two consecutive layers of an MLP are fully connected. In [27], it was demonstrated that MLPs are universal function approximators. In other words, they can be used to model any type of regression problem, or its particular case – the classification problem.

MLPs were a popular machine learning solution in the 1980s and found applications in diverse fields such as speech and image recognition, machine translation software, etc. However, since the 1990s strong competition was introduced by the much simpler and effective SVMs [28]. More recently, the interest in backpropagation networks has been renewed due to the success of deep learning and deep reinforcement learning.

Working with MLPs (as with most of the ANNs) consists of two phases: 1) the training (or learning) phase in which the weights of the connections between neurons are optimized using various algorithms, e.g., backpropagation combined with stochastic gradient descent (SGD) [29], [30], to minimize a loss function, and 2) the exploitation phase in which the optimized MLP model is used to fulfill its purpose using input data that it has not been exposed to during the training phase.

III. Dataset Visualization And Statistics

A. *Description of the Dataset*

To assess the performance of the various prediction methods a large dataset provided by Scholt Energy Control B.V. was employed. Along with historical measurements of portfolio-level electricity demand, the dataset comprises short-term prices from the Dutch electricity market, numerical weather predictions and aggregated wind power forecast data.

The data spans the period from January 2013 through October 2016. As regards electricity price data, historical day-ahead APX prices and historical imbalance prices (feed and take) obtained from the Dutch TSO, TenneT are included. Meteorological data stem from day-ahead hourly forecasts based on Numerical Weather Prediction (NWP) models. Both the data on wind power forecasts and installed wind capacity are related to a representative subset of the Scholt Energy Control B.V. wind portfolio. Furthermore, historical aggregate load data utilized in this study pertain to one of the company's consumption portfolios, consisting primarily of corporate and industrial customers.

In Figure 3 the heat-map of aggregated electrical demand of Scholt Energy Control B.V. from January 2013 to October 2016 with 15 minutes resolution is presented. The increasing trend that is observed in the aggregated consumption over time is mainly the result of the increasing size of the demand portfolio.

The Pearson Correlation Coefficient (PCC) that is defined later in Section III-B is used to reveal factors that potentially influence the energy demand profiles. In Table I the PCC value of different factors with respect to the aggregated energy consumption, along with basic statistical characteristics of the

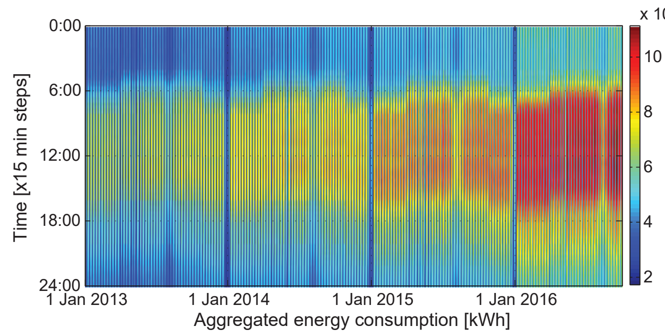

Fig. 3. Demand profile at the aggregation level of the portfolio of Scholt Energy Control B.V. between 1 January 2013 and 1 October 2016

TABLE I
Summary of the Correlation Values For Price, Weather Factors and the Aggregated Electrical Energy Demand

Influencing factors	Min	Max	Mean	Std.dev.	PCC
Energy [kWh]	17606.24	111053.74	55825.43	20367.1	1
APX price [€/MWh]	0.12	130.27	41.34	13.69	0.168
Feed price [€/MWh]	-450.7	737.99	39.92	66.10	0.085
Take price [€/MWh]	-442.2	737.99	46.40	70.03	0.075
Temperature [C°]	-12.9	33.8	11.06	6.63	0.181
Irradiation [$J/cm^2/h$]	0	340	44.39	70.90	0.318
Wind capacity [MW]	156.21	249.6	211.74	30.11	0.241
Wind velocity [m/s]	0	14	3.48	1.91	0.056

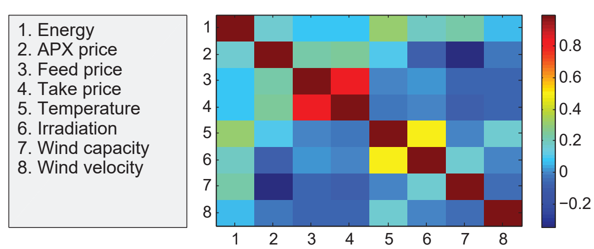

Fig. 4. Correlation results within the Scholt Energy Control dataset, including energy, price and meteorological data between 1 January 2013 and 1 October 2016 with 15 minute resolution.

entire dataset are summarized. Among the different factors irradiation is found to be the most correlated with electricity demand with a corresponding PCC value of 0.318. Moreover, the correlation matrix between all the factors considered in this analysis is shown in Figure 4. The low correlations could be partially justified due to the high levels of uncertainty introduced by a complex mix of various power generation sources at the end-user premises. For this reason pricing and meteorological data will not be fed into the day-ahead prediction model, implying that inputs are based only on historical values of the time series itself.

B. *Prediction Evaluation Metrics*

To quantify the performance of the prediction methods, a variety of standard metrics were used. First, the prediction accuracy is evaluated using three popular metrics capable of penalizing differently the same error, namely the root mean square error (RMSE), the normalized root-mean-square error (NRMSE), and the mean absolute percentage error (MAPE), defined by (1)-(3), respectively.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{x}_i)^2} \tag{1}$$

$$NRMSE[\%] = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\frac{(x_i - \hat{x}_i)^2}{\max(x) - \min(x)}} \cdot 100 \tag{2}$$

$$MAPE[\%] = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{x_i - \hat{x}_i}{\max(x)}\right| \cdot 100 \tag{3}$$

In (1)-(3) n represents the total number of predicted steps, x_i represents the actual aggregated energy consumption for the time-step i and $\hat{x}_i$ represents the value that is estimated by the prediction model for the same time-step. Secondly, the PCC defined in (4) is used to indicate the degree of the linear dependence between the real and the predicted values.

$$PCC(x, \hat{x}) = \frac{\mathbb{E}[(x - \mu_x)(\hat{x} - \mu_{\hat{x}})]}{\sigma_x \sigma_{\hat{x}}} \tag{4}$$

In (4) $\mathbb{E}[\cdot]$ is the expected value operator with means μ_x and $\mu_{\hat{x}}$, and standard deviations σ_x and $\sigma_{\hat{x}}$, for the actual and estimated aggregated energy consumption values, respectively. The PCC value is within the range [-1,1]. The sign of the correlation coefficient defines the direction of the relationship, either positive or negative.

IV. Experiments And Results

The MLPs were implemented in Python using the Keras Deep Learning library [31]. For the classical machine learning methods used in this study, the Machine Learning Toolbox™in Matlab® was used. If not otherwise stated, default settings were used. To predict the aggregated energy consumption for the next day, data spanning one week in the past are used as input features.

A. Day-ahead Aggregated Energy Prediction

In Table II the results for day-ahead energy prediction with 15 minutes resolution for all 9 methods used in this study are summarized in terms of different error metrics. Each value represents an average of the corresponding error metric for one year in order to increase its statistical significance level.

B. MPL Implementation Details

a) MLP hyper-parameter tuning: By carefully selecting the MLP hyper-parameters, the prediction accuracy can be significantly improved. For this reason the prediction accuracy of MLPs with different number of hidden layers and neurons, as well as various activation functions is explored. A summary of the experiments that were conducted is presented in Table III and evidences an improvement in accuracy by successfully tuning the hyper-parameters.

TABLE II
AVERAGE RESULTS OVER ONE YEAR SHOWING THE ACCURACY [%] OF EACH PREDICTION METHOD

Method	RMSE [kWh]	NRMSE [%]	MAPE [%]	PCC
Linear regression	5434.89	25.81	7.39	0.90
Multiple linear regression	22566.99	94.15	29.86	0.57
SVR (RBF)	22452.64	87.08	30.63	0.78
SVR (polynomial)	10617.40	40.94	13.87	0.88
SVR (linear)	4443.11	23.63	6.34	0.91
Regression tree	14140.10	74.39	18.84	0.71
Ensemble regression tree	12791.76	67.53	17.42	0.75
Ensemble Tree LS Boost	12791.76	67.53	17.42	0.75
Ensemble Tree Bag	10285.76	52.50	14.60	0.81
Gaussian process (linear)	5529.98	27.13	7.54	0.87
Gaussian process	4326.72	18.85	5.35	0.93
Multilayer perceptron	2376.38	12.45	3.40	0.96

b) Data normalization: After the first 13 variations, that is, from MLP_1 to MLP_{13}, normalization was applied to the data. For this particular dataset, the most accurate results were observed for a MLP with one hidden layer consisting of 32000 hidden neurons, i.e. MLP_{25}.

c) Optimization algorithms: There are plenty of methods to optimize deep learning architectures. First the MLP models were trained using SGD [29] with various settings. The update rule for SGD parameters θ is given by (5).

$$\theta = \gamma\theta - \beta v_{t-1} + \alpha\nabla_\theta\mathcal{J}(\theta) \tag{5}$$

In (5) θ represents the weights of the connections of the model, γ is the weight decay, α the learning rate, and β is the momentum which is usually set to a value close to 0.9.

Moreover, two different stochastic optimization algorithms, namely the RMSprop [31] and the Adaptive Moment Estimation (ADAM) [32] optimization methods were investigated. RMSprop is an unpublished, adaptive learning rate method that was proposed by Hinton and it is similar to the Adadelta optimization algorithm [33]. The idea behind RMSprop is to divide the learning rate by an exponentially decaying average of squared gradients. Adadelta uses the root-mean-square (RMS) of parameters in the numerator of the update rule. ADAM adds bias-correction and momentum to RMSprop. In general, RMSprop, Adadelta, and ADAM are similar algorithms that perform well under similar circumstances. Kingma et al. [32] showed that due to bias-correction ADAM slightly outperforms RMSprop and Adadelta towards the end of the optimization since the gradients become sparser.

d) Activation functions: Recently, the rectified linear unit (ReLU) became the most common choice of activation function $f(\cdot)$ for deep learning methods [14] and is defined as $f(x) = max(0, x)$. ReLU and its variations, compared to sigmoid function or similar activation functions, allow for faster and effective training of deep neural architectures on large and complex datasets [34]. As it can be seen in Table III the LeakyReLU [35] was used for MLP_{20}, while for the other architectures the S-shaped rectified linear activation unit (SReLU) [36] function was used.

TABLE III
SELECTING THE MULTILAYER PERCEPTRON (MLP) HYPER-PARAMETERS BASED ON THE PREDICTION ACCURACY. ERROR RESULTS ARE AVERAGED OVER A YEAR OF TRAINING DATA WITH 15 MINUTE RESOLUTION FOR EACH MLP VARIATION.

Method	RMSE [kWh]	NRMSE [%]	MAPE [%]	PCC	Hidden[#] [L1/ L2/ L3]	Drop out	Batch size	Number of epochs	Optimization algorithm	Opt. parameters $\alpha = 0.01$, β, γ	Activation function
MLP_1	5696.69	29.56	8.32	0.80	500/500	0.5	2	100	SGD	$\beta = 0, \gamma = 0$	SReLU
MLP_2	9091.47	48.15	13.51	0.81	500/500	0.5	2	100	ADAM	adaptive	SReLU
MLP_3	7596.03	34.34	10.45	0.83	500/500	0.5	2	100	RMSprop	adaptive	SReLU
MLP_4	7156.72	32.36	10.20	0.84	500/500	0.5	2	200	SGD	$\beta = 0, \gamma = 0$	SReLU
MLP_5	7478.44	32.18	10.31	0.85	500/500	0.5	2	200	ADAM	adaptive	SReLU
MLP_6	6140.87	36.66	9.25	0.83	500/500	0.5	2	200	RMSprop	adaptive	SReLU
MLP_7	8026.24	38.50	11.65	0.79	500/500/500	0.5	2	100	SGD	$\beta = 0, \gamma = 0$	SReLU
MLP_8	12158.19	47.63	16.48	0.78	500/500/500	0.5	2	100	ADAM	adaptive	SReLU
MLP_9	10519.51	39.99	14.18	0.81	500/500/500	0.5	2	100	RMSprop	adaptive	SReLU
MLP_{10}	11162.12	42.77	14.95	0.79	100/100	0.5	2	100	SGD	$\beta = 0, \gamma = 0$	SReLU
MLP_{11}	7471.65	33.53	10.38	0.80	300/300	0.5	2	100	SGD	$\beta = 0, \gamma = 0$	SReLU
MLP_{12}	7173.61	34.81	10.38	0.82	600/600	0.5	2	100	SGD	$\beta = 0, \gamma = 0$	SReLU
MLP_{13}	6890.53	32.74	10.12	0.86	500/500	0.3	2	100	SGD	$\beta = 0, \gamma = 0$	SReLU
MLP_{14}	3829.51	23.46	6.09	0.87	500/500	0.5	2	100	SGD	$\beta = 0, \gamma = 0$	SReLU
MLP_{15}	3855.19	21.31	5.51	0.92	2000	0.7	2	100	SGD	$\beta = 0, \gamma = 0$	SReLU
MLP_{16}	3339.68	17.41	4.59	0.91	2000	0.6	1	100	SGD	$\beta = 0, \gamma = 0$	SReLU
MLP_{17}	3034.20	16.59	4.29	0.92	2000	0.6	5	100	SGD	$\beta = 0, \gamma = 0$	SReLU
MLP_{18}	3046.25	17.02	4.35	0.92	2000	0.6	5	200	SGD	$\beta = 0, \gamma = 0$	SReLU
MLP_{19}	2858.10	16.67	4.12	0.93	4000	0.6	5	100	SGD	$\beta = 0, \gamma = 0$	SReLU
MLP_{20}	2634.32	14.14	3.72	0.95	8000	0.6	5	100	SGD	$\beta = 0.9, \gamma = 0.0002$	LeakyReLU
MLP_{21}	2885.20	14.82	4.05	0.95	8000	0.6	5	100	SGD	$\beta = 0.9, \gamma = 0.0002$	ReLU
MLP_{22}	3085.90	16.39	4.44	0.94	8000	0.6	5	100	SGD	$\beta = 0.9, \gamma = 0.0002$	SReLU
MLP_{23}	3013.97	15.63	4.26	0.95	8000	0.7	5	100	SGD	$\beta = 0.9, \gamma = 0.0002$	SReLU
MLP_{24}	2500.45	13.01	3.58	0.96	16000	0.6	5	100	SGD	$\beta = 0.9, \gamma = 0.0002$	SReLU
MLP_{25}	2376.38	12.45	3.40	0.96	32000	0.7	5	100	SGD	$\beta = 0.9, \gamma = 0.0002$	SReLU

e) Convergence criterion: For MLP_5 a maximum limit of 200 iterations was set, while for the other MLP variations the maximum iteration number was set to 100. However, it has been observed that 50 iterations are usually sufficient for the day-ahead energy prediction problem that is addressed in this study.

C. Analysis of the Prediction Error

Although the most common metrics were used in order to assess the quality of the predictions, it is particularly difficult to conclude whether the predictions are consistently accurate or not. For this reason, for the most accurate results reported in Tables II and III an illustration of the error probability density function is depicted in Fig. 5. Probability density functions that are closer to zero produce both accurate and reliable results.

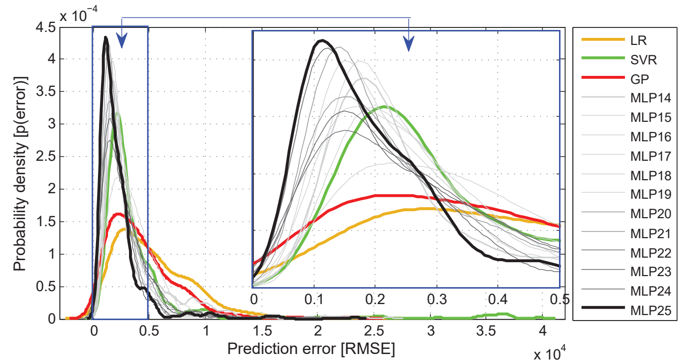

Fig. 5. The error probability density functions in terms of RMSE for Linear Regression, SVR, Gaussian Processes, and MLP with various settings.

V. DISCUSSION AND CONCLUSIONS

In this paper a comparison between nine prediction methods applied to the problem of predicting the aggregated energy demand was performed in terms of four performance metrics. A new deep learning method for the day-ahead aggregate energy prediction problem, namely the Multi Layer Perceptron (MLP) is introduced and analyzed. The prediction capabilities of MLPs are explored under various settings. Experimental validation was performed on a real-world dataset provided by Scholt Energy Control B.V., The Netherlands, which includes information on the Dutch portfolio of the company. The obtained results indicate that MLPs enhanced with deep learning capabilities present better prediction accuracy, with higher accuracy in terms of RMSE, NRMSE, and MAPE in comparison with popular traditional regression and machine learning methods. A visualization of the error distribution function confirms that MLPs outperform the benchmark techniques in terms of accurate and reliable prediction outcomes. However, an open challenge remains as regards the process of finding appropriate parameters for the MLP model when using a given dataset, since it is not trivial to guarantee that optimal values can be found.

One can argue that taking influencing factors into account as additional inputs to the prediction method may improve accuracy. Nevertheless, for the particular case of the database used in this study, poor correlations between pricing, meteorological conditions and aggregated energy consumption data was found. This analysis suggests a challenging mix of net consumption profiles which include local generation sources, with a predominant role of solar energy. Consequently, it can be argued that due to the growing uncertainty at the demand portfolio level, including influencing factors in the prediction methodology does not necessarily improve the prediction accuracy.

Further investigation of the day-ahead aggregated energy prediction problem could focus on additional deep learning methods such as Convolutional Neural Networks, Deep Boltzmann Machines, etc. Possible benefits of deep learning methods in terms of increased prediction accuracy deserve to be more thoroughly investigated for different types of smart grid related data, such as real-time price signals.

ACKNOWLEDGMENT

Research leading to the results presented in this study was funded by Scholt Energy Control B.V., The Netherlands.

REFERENCES

[1] M. Krarti, *Energy Audit of Building Systems: An Engineering Approach, Second Edition*, ser. Mechanical and Aerospace Engineering Series. Taylor & Francis, 2012.

[2] A. I. Dounis, "Artificial intelligence for energy conservation in buildings,"*Advances in Building Energy Research*, vol. 4, no. 1, pp. 267–299, 2010.

[3] A. Foucquier, S. Robert, F. Suard, L. Stephan, and A. Jay, "State of the art in building modelling and energy performances prediction: A review,"*Renewable and Sustainable Energy Reviews*, vol. 23, no. 0, pp. 272 –288, 2013.

[4] H. xiang Zhao and F. Magouls, "A review on the prediction of building energy consumption," *Renewable and Sustainable Energy Reviews*, vol. 16, no. 6, pp. 3586 – 3592, 2012.

[5] M. Aydinalp-Koksal and V. I. Ugursal, "Comparison of neural network, conditional demand analysis, and engineering approaches for modeling end-use energy consumption in the residential sector," *Applied Energy*, vol. 85, no. 4, pp. 271 – 296, 2008.

[6] L. A. Hurtado-Munoz, P. H. Nguyen, W. Kling, and W. Zeiler, "Building energy management systems : optimization of comfort and energy use," in *UPEC 2013 - 48th International Universities' Power Engineering Conference*, 2013.

[7] L. Xuemei, D. Lixing, L. Jinhu, X. Gang, and L. Jibin, "A novel hybrid approach of kpca and svm for building cooling load prediction," in *Knowledge Discovery and Data Mining, 2010. WKDD '10. Third International Conference on*, 2010, pp. –.

[8] B. M. J. Vonk, P. Nguyen, M. Grond, J. Slootweg, and W. Kling, "Improving short-term load forecasting for a local energy storage system," in *47th International Universities Power Engineering Conference (UPEC), 2012*, Sept 2012, pp. 1–6.

[9] S. Wong, K. K. Wan, and T. N. Lam, "Artificial neural networks for energy analysis of office buildings with daylighting," *Applied Energy*, vol. 87, no. 2, pp. 551–557, 2010.

[10] S. A. Kalogirou, "Artificial neural networks in energy applications in buildings," *International Journal of Low-Carbon Technologies*, vol. 1, no. 3, pp. 201–216, 2006.

[11] C. Cortes and V. Vapnik, "Support-Vector Networks," *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, Sep. 1995.

[12] L. E. Baum and T. Petrie, "Statistical inference for probabilistic functions of finite state Markov chains," *Annals of Mathematical Statistics*, vol. 37, pp. 1554–1563, 1966.

[13] T. Zia, D. Bruckner, and A. Zaidi, "A hidden markov model based procedure for identifying household electric loads," in *IECON 2011 - 37th Annual Conference on IEEE Industrial Electronics Society*, 2011, pp. 3218–3223.

[14] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, May 2015.

[15] E. Mocanu, P. H.Nguyen, M. Gibescu, and W. Kling, "Comparison of machine learning methods for estimating energy consumption in buildings," in *Proceedings of the 13th International Conference on Probabilistic Methods Applied to Power Systems, Durham, UK*, 2014.

[16] E. Mocanu, P. H. Nguyen, M. Gibescu, and W. L. Kling, "Deep learning for estimating building energy consumption," *Sustainable Energy, Grids and Networks*, vol. 6, pp. 91 – 99, 2016.

[17] E. Mocanu, E. M. Larsen, P. H. Nguyen, P. Pinson, and M. Gibescu, "Demand forecasting at low aggregation levels using factored conditional restricted boltzmann machine," in *IEEE Power Systems Computation Conference (PSCC)*, June 2016.

[18] D. L. Marino, K. Amarasinghe, and M. Manic, "Building energy load forecasting using deep neural networks," *CoRR*, vol. abs/1610.09460, 2016.

[19] M. Manic, K. Amarasinghe, J. J. Rodriguez-Andina, and C. Rieger, "Intelligent buildings of the future: Cyberaware, deep learning powered, and human interacting," *IEEE Industrial Electronics Magazine*, vol. 10, no. 4, pp. 32–49, 2016.

[20] S. Ryu, J. Noh, and H. Kim, "Deep neural network based demand side short term load forecasting," *Energies*, vol. 10, no. 1, p. 3, 2016.

[21] T. Sogabe, H. Ichikawa, T. Sogabe, K. Sakamoto, K. Yamaguchi, M. Sogabe, T. Sato, and Y. Suwa, "Optimization of decentralized renewable energy system by weather forecasting and deep machine learning techniques," in *2016 IEEE Innovative Smart Grid Technologies - Asia (ISGT-Asia)*, Nov 2016, pp. 1014–1018.

[22] W. Wu, K. Chen, Y. Qiao, and Z. Lu, "Probabilistic short-term wind power forecasting based on deep neural networks," in *2016 International Conference on Probabilistic Methods Applied to Power Systems (PMAPS)*, Oct 2016, pp. 1–8.

[23] Q. Xiaoyun, K. Xiaoning, Z. Chao, J. Shuai, and M. Xiuda, "Short-term prediction of wind power based on deep long short-term memory," in *2016 IEEE PES Asia-Pacific Power and Energy Engineering Conference (APPEEC)*, Oct 2016, pp. 1148–1152.

[24] E. Mocanu, P. H. Nguyen, W. L. Kling, and M. Gibescu, "Unsupervised energy prediction in a smart grid context using reinforcement cross-building transfer learning," *Energy and Buildings*, vol. 116, pp. 646 – 655, 2016.

[25] C. Rasmussen and C. Williams, *Gaussian Processes for Machine Learning*, ser. Adaptive Computation and Machine Learning. Cambridge, MA, USA: MIT Press, Jan. 2006.

[26] F. Rosenblatt, *Principles of neurodynamics; perceptrons and the theory of brain mechanisms.* Washington: Spartan Books, 1962.

[27] G. Cybenko, "Approximation by superpositions of a sigmoidal function,"*Mathematics of Control, Signals and Systems*, vol. 2, no. 4, pp.303–314, 1989.

[28] C. Cortes and V. Vapnik, "Support-vector networks," *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, Sep. 1995.

[29] O. Bousquet and L. Bottou, "The tradeoffs of large scale learning," in *Advances in Neural Information Processing Systems 20*, J. C. Platt,D. Koller, Y. Singer, and S. T. Roweis, Eds. Curran Associates, Inc., 2008, pp. 161–168.

[30] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *Nature*, vol. 323, pp. 533 – 536, 1986.

[31] F. Chollet, "keras," 2015. [Online]. Available: https://github.com/fchollet/keras

[32] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *International Conference for Learning Representations (ICLR)*, 2015.

[33] M. D. Zeiler, "ADADELTA: an adaptive learning rate method," *CoRR*, vol. abs/1212.5701, 2012. [Online]. Available: http://arxiv.org/abs/1212.5701

[34] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, J. Frnkranz and T. Joachims, Eds., 2010, pp. 807–814.

[35] A. Y. H. Andrew L. Maas and A. Y. Ng., "Rectifier nonlinearities improve neural network acoustic models," *ICML Workshop on Deep Learning for Audio, Speech, and Language Processing (WDLASL 2013)*, 2013.

[36] X. Jin, C. Xu, J. Feng, Y. Wei, J. Xiong, and S. Yan, "Deep learning with s-shaped rectified linear activation units," *CoRR*, vol. abs/1512.07030, 2015.