A novel Ensemble Deep Belief Network and Bayesian Adaptive Aggregation for Regression

Saima Hassan*, Mojtaba Ahmadieh Khanesari†, M. Tariq Jan‡ Wali Khan§
*‡§ Kohat University of Science and Technology, Kohat
†Department of Mechanical, Materials and Manufacturing Engineering, Faculty of Engineering, University of Nottingahm, UK

***Abstract*—Ensemble modeling of Neural Networks is a strategy where multiple alternative models (ensemble members) are constructed and then their forecasts are ensembled using various combination approaches. Ensemble of Neural Networks has proved the concept behind this strategy. Deep neural network is a type of neural network that offers potential opportunities to overcome traditional ensemble of neural networks. This paper proposes an ensemble of deep belief networks (DBN). The ensemble members of DBN are constructed with different number of epochs so that the generalization ability can be improved. The outputs of these DBNs are aggregated by a Bayesian model averaging method. The proposed Bayesian adopted ensemble of DBNs is evaluated on two benchmark data sets. Comparison of the proposed model is evaluated with simple averaging and single DBN over a number of forecasting measuring that shows better performance of the proposed model.**

***Index Terms*—Ensemble modeling, deep belief network, Bayesian model averaging, forecast combination.**

I. Introduction

Demand forecasting is essential to control the increasing variety and complexity of planning various operations management. Being an active research area, the selection of a forecasting model is the focus of many researchers over the last few decades. Numerous statistical and computational model are utilized for forecasting demands. In general, the family of Autoregressive moving average (ARMA) have been in practice for forecasting using statistical methods. Their linear structure provides the most effective linear modeling; however, is inadequate in real world problem which are mostly non-linear. In order to overcome this limitation, advanced sophisticated models have been proposed as an alternative for forecasting. Among them artificial neural networks (ANN) has gained a lot of attention in the field of forecasting [1].

TABLE I
Meaning of some of the *abbreviations* used in paper.

abbreviation	Meaning
AMAPE	Average Mean Absolute Percentage Error
ANN	Artificial Neural Networks
ARMA	Autoregressive Moving Average
BMA	Bayesian Model Averaging
DBN	Deep Belief Networks
DNN	Deep Neural Network
MLP	Multilayer Perceptron
RBM	Restricted Boltzmann Machine
RMSE	Root Mean Square Error
SA	Simple Averaging
SMAPE	Symmetric Mean Absolute Percentage Error

Utilization of ANN can be seen extensively in literature for a wide range of application areas. With an ever increasing number of complex issues, the shallow network structure of ANN is deficient for their effective solution. Their popularity starts to decline with the advent of powerful Kernel based approaches like support vector machines. Interest in the use of ANN was revived by substantially better performance of deep neural network that progressively reveal low dimensional nonlinear structure [2]. Since then DNNs have completely revolutionized some fields including demand forecasting [3], [4]. Another approach that improves the forecasting accuracy is the ensemble modeling. Ensemble modeling consists of constructing multiple member models and then combining their output using various aggregation algorithms for better predictive performance. In statistics, forecast combination was pioneered by Bates and Granger [5]. Due to the remarkable performance of this approach, it was adopted by machine learning communities. Ensemble of Neural network was originated [6] to illustrate that the generalization ability of a single NN can be significantly improved through an ensemble of a number of NNs. It has been investigated that a good ensemble Model can make different errors with same dataset [7]. Various application of NN ensemble can be seen in literature [8], [9]. Characteristics of DNN offers potential opportunities to overcome traditional NN ensembles. Ensemble modeling of DNN has also been reported in recent years. An ensemble of DBN was initially proposed for regression and time series forecasting [10]. Another ensemble of DNN based of reconstruction error was presented in [11].

The design of member networks in an ensemble modeling can be broadly categorized into two different approaches. In the first approach, the diverse members can be obtained by varying the architecture of the network models. This can be achieved by selecting different weight functions, network type, number of hidden neurons, learning algorithm and epoch [9], [12]. In the second approach, diverse members of networks are obtained by training them on different training set, such as bagging [13], Boosting [14], cross validation [15]. Apart from these two approaches, member models for an ensemble can be selected from a large number of network models. A highly diverse set of member networks were selected through genetic algorithm [16], Pruning algorithm is utilized to eliminate redundant networks [17]. Best network models are selected based on their best forecasting measure [9]. Extreme forecasts are discarded and the rest of the models are kept for ensemble by trimming the tails symmetrically [9].

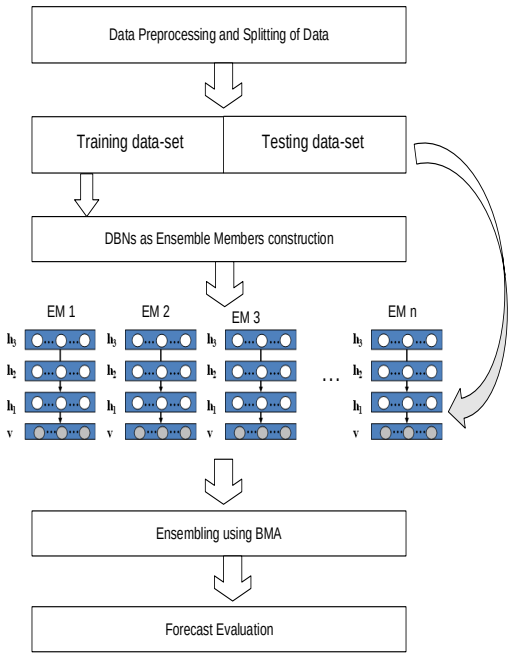

Fig. 1. Flowchart.

As for as output combination is concerned, Bates and Granger [5] in 1969 presented the idea that the performance of combined forecast is better than the single model. This idea has been supported by many researchers [9], [18]. From classical statistical methods to sophisticated machine learning algorithm, various methods have been proposed for combining forecasts. Simple mean and weighted average are the mostly used methods for combining the forecasts. Recently, a DNN model is used to ensemble multi models for cancer prediction [19]. [10] utilized a support vector machine to an ensemble of DBN.

The objective of this paper is to propose a Bayesian adaptive ensembling of DBN for regression. Many components of DBN are generated and a Bayesian model averaging (BMA) is utilized to ensemble them. Specifically the BMA combines the forecasts output of many DBN. Several benchmark datasets are used to demonstrate the performance of the proposed ensemble model.

Rest of the paper is structured as follows. Section II presents the methodology used in this research for forecasting. Section III provides the empirical results and Section IV concludes with some remarks and recommendations.

II. Proposed Methodology of the Deep Belief Network Ensemble Model

An ensemble of multiple DBNs proposed in this research work can be seen in Fig. 1.

A. Data Preprocessing and Splitting

The available data preprocessed and is divided into training and testing datasets. Training dataset is used to train the ensemble members. Trained ensemble members are then used to forecast with the testing dataset.

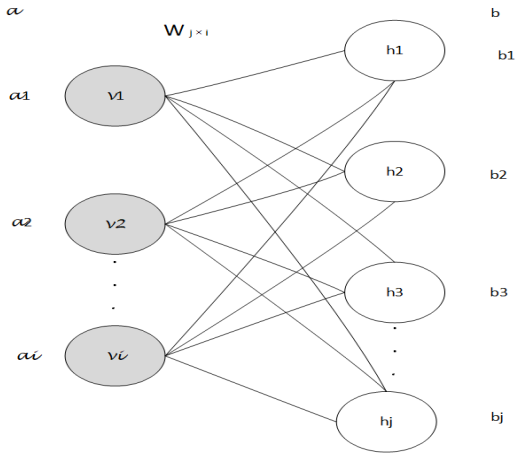

Fig. 2. Schematic Diagram of a RBM.

B. Ensemble construction and training of the Deep Belief Network

DBNs are composed of Restricted Boltzmann Machines (RBMS). RBM is a two layer connected ANN that can learn the probability distribution over the given set of inputs. Structure of an RBM with a visible layer v_i, hidden layer h_j and weights connection matrix $W_{j\times i}$ can be seen in Fig 2. Where a_i and b_j are the bias weights for the visible and hidden units respectively. Stacking RBM on top of each other such that the output of the lowest RBM is used as input to the subsequent RBM, a DBN is formed [2] Fig 3. Each RBM is trained in unsupervised manner. For the proposed work ensemble members of the DBN are constructed in this way. For better generalization ability, the ensemble members are diversified by varying the number of epochs of each DBN. The DBNs are trained with training dataset. Each DBN is evaluated with the testing dataset and forecasts $\hat{y}_{D_T}$ are obtained.

C. Ensembling using Bayesian Model Averaging

The forecasts obtained from DBNs are ensembled using Bayesian setting. Bayesian model averaging (BMA) is a Bayesian approach of combining forecasts. It can be thought of a weighted average model, where the weights of the forecasts are computed based on posterior probabilities of the models. Higher weights are assigned to the member model that fit to the data well. In order to compute the weights of the forecasts for ensembling, the approach used in [20] is followed. They have compared the performance of various model averaging techniques with an application to the growth empirics.

Reference [20] introduced two types of variables: explanatory variables ("focus regressors") and additional variables ("auxiliary regressors") which are of less certain. The research of this paper ignores the additional explanatory variables and uses the focus regressors only. Similar to Bayesian approach, this method combines prior beliefs on the unknown parameters

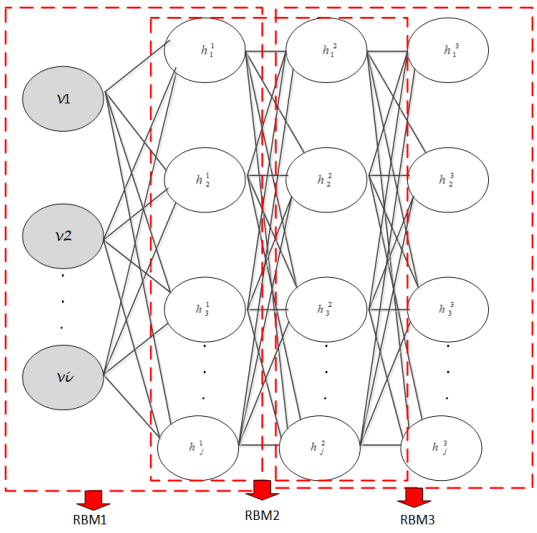

Fig. 3. Deep Belief Network.

of the model with some extra information coming from the data. Some of the key elements of this method are the simple likelihood function, the prior distributions on the regression parameters of ensemble members EM_n and the prior distributions on the model space.

Referring to [20], $k1$ and $k2$ represent the number of focus and auxiliary regressors respectively. Suppose $n_{DBN} = k_1$ are ensemble members of DBN and B is the weights calculated through BMA. As no auxiliary variables are considered here therefore $k2 = 0$ Let $EM_n = \{EM_1, \ldots, EM_{n_{DBN}}\}$ indicates the model space of n_{DBN} ensemble members for forecasting y with training dataset D_T and f_n is the forecasts from n_{th} DBN. In the case y is to be forecast on the basis of D_T then according to the law of total probability the predictive probability density can be given as [21], [22],

$$p(y \mid D_T) = \sum_{n=1}^{n_{total}} w_n p(y \mid EM_n, D_T) \tag{1}$$

Where $p(y \mid EM_n, D_T)$ represents the posterior distributions given by single DBN EM_n and $w_n = p(EM_n \mid D_T)$ is the posterior probabilities. The posterior mean of the BMA forecast can be calculated as:

$$\begin{aligned} E & [y \mid D_T] \\ &= \sum_{n=1}^{n_{DBN}} p(EM_n \mid D_T).E[y \mid EM_n, D_T] \\ &= \sum_{n=1}^{n_{DBN}} w_n f_n \end{aligned} \tag{2}$$

Following the assumption made by Magnus [20] and proposed by Zellner and Fernandez [23] and [24] the prior variance V_n excluding auxiliary variable can be given as:

$$V_n^{-1} = g_i EM1 \tag{3}$$

where $g = 1/max(L, k_2^2)$ is a constant scalar for each ensemble member EM_n. Since it is assumed that $k2 = 0$, then no model selection takes place [20] and $M1 = X1(X1^T X1)^{-1} X1^T$ where $X1$ is $L \times k1$ and L is the length of testing data outputs. A vector of calculated weights (B) and standard errors associated with these weights is generated from these calculations. The forecasts $\hat{y}_{D_T}$ obtained in section II-B are combined with these weights to get the BMA combined forecast as:

$$\hat{y}_{BMA} = \hat{y}_{D_T} \times B \tag{4}$$

D. Forecast Evaluation

Four evaluation indexes are chosen to evaluate the forecasting performance of the proposed model. These indexes are AMAPE, RMSE, SMAPE and a normalized cost function of error called J. The equations describing these evaluation indexes are as follows.

$$AMAPE(\%) = \frac{1}{n} \sum_{t=1}^{n} \left| \frac{A_t - F_t}{\frac{1}{n}\sum_{t=1}^{n} A_t} \right| \times 100 \tag{5}$$

$$SMAPE = \frac{1}{n} \sum_{t=1}^{n} \frac{|A_t - F_t|}{(|F_t| + |A_t|)/2} \tag{6}$$

$$RMSE = \sqrt{\frac{1}{n} \sum_{t=1}^{n} (A_t - F_t)^2} \tag{7}$$

$$J = \frac{\sqrt{\sum_{t=1}^{n} (A_t - F_t)^2}}{\sum_{t=1}^{n} (A_t - mean(A_t))^2} \tag{8}$$

where A_t and F_t are the real and forecasted value at time t and n is the total number of test samples.

III. Results and Comparisons

A. Data and Experimental Setup

Forecasting accuracy of the proposed structure is analyzed and tested against prediction of Mackey Glass chaotic system [25] and Friedman Artificial Domain [26]. A number of comparisons are done to illustrate improvement obtained using the proposed architecture.

The Mackey-Glass models the blood cell regulation and because of its chaotic dynamic equations, it is widely investigated in papers concerning prediction and identification. The dynamic equation describing this system is as follows.

$$\frac{dx}{dt} = \beta \frac{x(t-\tau)}{1 + x^{10}(t-\tau)} - \alpha x(t), \;\; \alpha, \beta, \tau > 0 \tag{9}$$

The number of samples generated for this dataset is equal to 9000 from which 75% is taken for training and 25% is used for test data.

Friedman Artificial Domain is the second dataset used in experiences. This dataset was generated in [26] for the first time and later described in [13], [27].

For chaotic Mackey-Glass data set the input values are selected as $\big(x(t-18), x(t-12), x(t-6), x(t)\big)$ that is used to predict the single output as $x(t+6)$. On the other hand, Friedman Artificial Domain is a static dataset that has

$(x_1, x_2, ..., x_{10})$ as input data and y as a single output. Train and test data in both cases are normalized to the interval of $[0, 1]$ as follows.

$$x_n = \frac{x - x_{min}}{x_{max} - x_{min}} \tag{10}$$

where x_n represents the normalized dataset.

B. Results and Discussions

In order to predict the Mackey-Glass time series, 2 hidden layers for the DBN are selected. The number of nodes for the input layer is equal to 4 nodes and 4 nodes are selected for each of hidden layers. 30 DBNs are trained and their results are combined using BMA. The number of epochs considered for DBNs varies from 440 to 1600 with increments equal to 40. In order to illustrate the efficacy of adding BMA algorithm, the results obtained are compared with that of the case when simple averaging method is used (DBN-SA) and the best results obtained from the single DBN used.

Table II reported that the proposed method (DBN-BMA) outperforms other methods in all performance indexes. Furthermore, as can be seen from Figs 4 and 5 in their zoomed figures although all 30 individual DBNs are far from the real data, by adding BMA the predictions made become very close to real data.

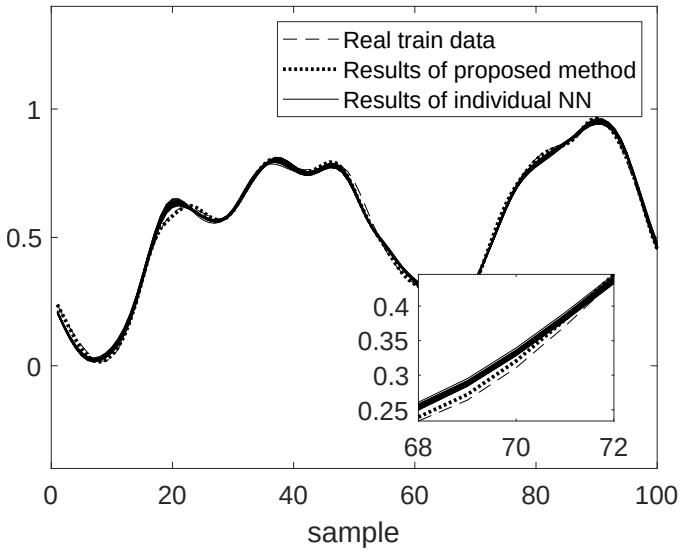

Fig. 4. Prediction performance for Mackey-Glass time series train data.

Figure 6 illustrates RMSE obtained from DBN-BMA, DBN-SA and from all individual DBNs. It can be seen that the best result is obtained with the proposed approach. The minimum RMSE obtained with the single DBN is 0.0161 and that with the DBN-BMA is 0.0091. DBN-SA gave RMSE of 0.0442%. That means that the individual DBN produces better forecasting result than the DBN-SA.

Similar to Mackey-Glass case, in order to estimate Friedman Artificial Domain dataset 2 hidden layers for the DBN are selected. A total of 10 nodes for the input layer is taken and 10 nodes are selected for each of hidden layers. The number

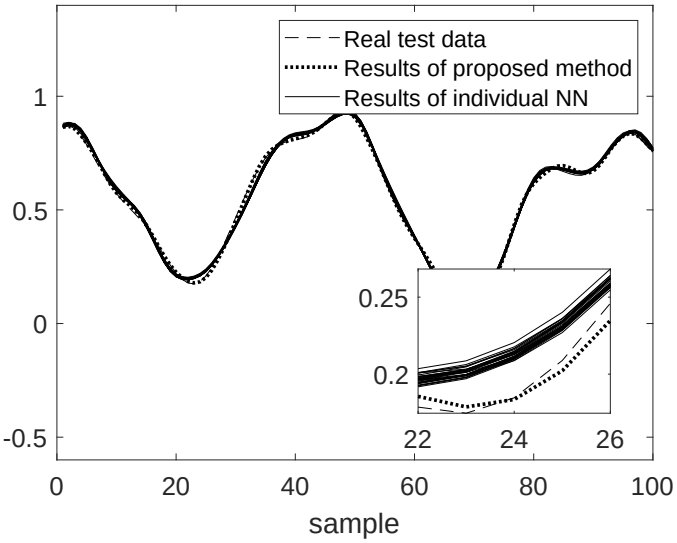

Fig. 5. Prediction performance for Mackey-Glass time series test data.

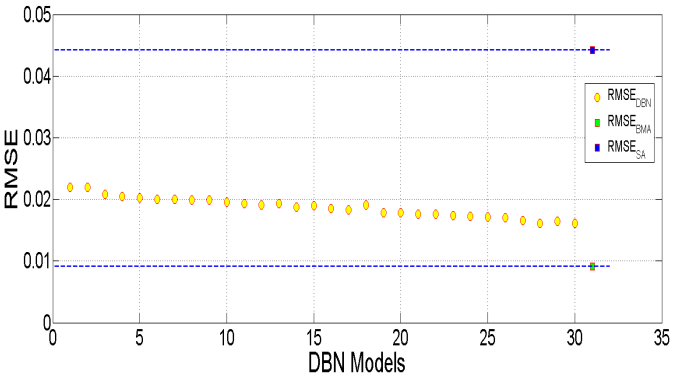

Fig. 6. RMSE obtained from models.

of individual DBN as well as the combination method and the ranges of epochs are considered to be exactly the same as Mackey-Glass case.

It can also be observed from Table III that better forecasting results obtained belong to the proposed method namely DBN-BMA. The zoomed figure in Fig. 7 further illustrates the performance. The addition of BMA makes it possible to obtain the closest results to the real data. The "Best DBN" results reported in these tables are the minimum forecasts obtained among the individual 30 DBNs.

IV. Conclusions

An ensemble deep learning structure is proposed in this paper. The proposed structure benefits from multiple DBNs with different number of epochs. BMA is used to combine the results obtained from 30 different DBNs. It is shown that the addition of BMA highly improves the results. The results obtained are compared with that of DBN-SA and the best results obtained from individual DBN. The benchmark functions used are Mackey-Glass chaotic time series and Friedman Artificial Domain dataset. The comparisons are made

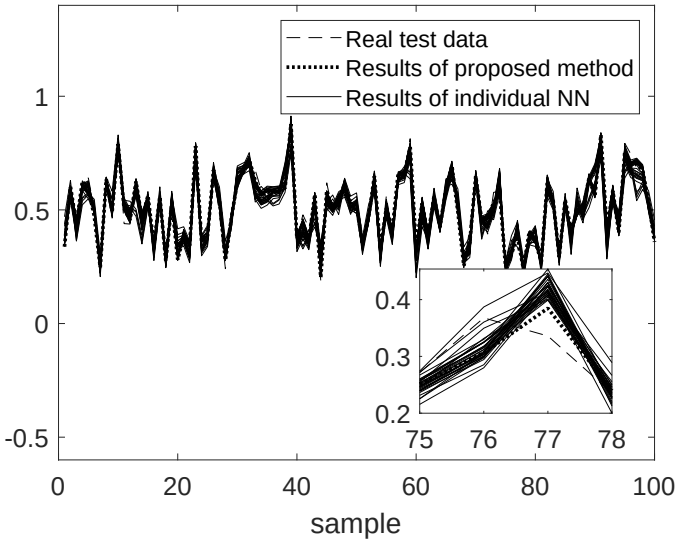

Fig. 7. Prediction performance for Friedman Artificial Domain dataset test data

TABLE II
TEST AND TRAIN RESULTS OBTAINED FOR MACKEY-GLASS CHAOTIC TIME SERIES

	DBN-BMA	DBN-SA	Best DBN
AMAPE test	1.28%	7.1714%	2.3932%
RMSE test	0.0091	0.0442	0.0161
SMAPE test	0.0316	0.1319	0.0529
J of test	0.0353	0.1715	0.063
AMAPE train	1.2734%	7.2218%	2.3813%
RMSE train	0.0091	0.0443	0.016
SMAPE train	0.0324	0.1327	0.0539
J of train	0.0351	0.1719	0.062

in terms of AMAPE, SMAPE, RMSE and a normalized cost function of error. It is shown that addition of BMA improves the results considerably. It is further shown that BMA works better than finding the simple average of all DBNs.

One of the most challenging parts in the training of DBN is finding appropriate value for its epoch number. The proposed method is a solution to such challenge as the information obtained during training using different number of epochs is not lost and they are all used to predict the real output.

TABLE III
TEST AND TRAIN RESULTS OBTAINED FOR FREEDMAN ARTIFICIAL DOMAIN

	DBN-BMA	DBN-SA	Best DBN
AMAPE test	5.1789%	8.9061%	5.8379%
RMSE test	0.0319	0.0552	0.0362
SMAPE test	0.0599	0.1040	0.0677
J of test	0.2	0.3458	0.2269
AMAPE train	5.22%	8.7943%	5.8623%
RMSE train	0.0323	0.0547	0.0363
SMAPE train	0.0590	0.1002	0.0662
J of train	0.2076	0.3514	0.2334

V. ACKNOWLEDGMENT

REFERENCES

[1] K. Hornik, M. Stinchcombe, and H. White, "Multilayer feedforward networks are universal approximators," *Neural Networks*, vol. 2, no. 5, pp. 359–366, Jul. 1989. [Online]. Available: http://dx.doi.org/10.1016/0893-6080(89)90020-8

[2] G. E. Hinton, S. Osindero, and Y.-W. Teh, "A fast learning algorithm for deep belief nets," *Neural Comput.*, vol. 18, no. 7, pp. 1527–1554, Jul. 2006. [Online]. Available: http://dx.doi.org/10.1162/neco.2006.18.7.1527

[3] T. Young, D. Hazarika, S. Poria, and E. Cambria, "Recent trends in deep learning based natural language processing," *CoRR*, vol. abs/1708.02709, 2017. [Online]. Available: http://arxiv.org/abs/1708.02709

[4] J. Lago, F. D. Ridder, and B. D. Schutter, "Forecasting spot electricity prices: Deep learning approaches and empirical comparison of traditional algorithms," *Applied Energy*, vol. 221, pp. 386 – 405, 2018. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S030626191830196X

[5] J. Bates and C. Granger, "The combination of forecasts," *Operations Research*, vol. 20, no. 4, pp. 451 – 468, 1969.

[6] L. K. Hansen and P. Salamon, "Neural network ensembles," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 12, no. 10, pp. 993–1001, Oct 1990.

[7] G. Brown, J. Wyatt, R. Harris, and X. Yao, "Diversity creation methods: a survey and categorisation," *Information Fusion*, vol. 6, no. 1, pp. 5 – 20, 2005. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1566253504000375

[8] A. A. Masaud, B. Stewart, and S. McMeekin, "Application of an ensemble neural network for classifying partial discharge patterns," *Electric Power Systems Research*, vol. 110, pp. 154 – 162, 2014. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0378779614000145

[9] S. Hassan, A. Khosravi, and J. Jaafar, "Examining performance of aggregation algorithms for neural network-based electricity demand forecasting," *International Journal of Electrical Power & Energy Systems*, vol. 64, pp. 1098 – 1105, 2015. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0142061514005511

[10] X. Qiu, L. Zhang, Y. Ren, P. N. Suganthan, and G. Amaratunga, "Ensemble deep learning for regression and time series forecasting," in *2014 IEEE Symposium on Computational Intelligence in Ensemble Learning (CIEL)*, Dec 2014, pp. 1–6.

[11] W. Huang, H. Hong, K. Bian, X. Zhou, G. Song, and K. Xie, "Improving deep neural network ensembles using reconstruction error," in *2015 International Joint Conference on Neural Networks (IJCNN)*, July 2015, pp. 1–7.

[12] Y. Zhao, J. Gao, and X. Yang, "A survey of neural network ensembles," in *2005 International Conference on Neural Networks and Brain*, vol. 1, Oct 2005, pp. 438–442.

[13] L. Breiman, "Bagging predictors," *Machine Learning*, vol. 24, no. 2, pp. 123–140, Aug. 1996. [Online]. Available: http://dx.doi.org/10.1023/A:1018054314350

[14] R. E. Schapire, "The strength of weak learnability," *Machine Learning*, vol. 5, no. 2, pp. 197–227, Jun 1990. [Online]. Available: https://doi.org/10.1007/BF00116037

[15] A. Krogh and J. Vedelsby, "Neural network ensembles, cross validation, and active learning," in *Advances in Neural Information Processing Systems*, vol. 8. MIT Press, 1995, pp. 231–238.

[16] Z.-H. Zhou, J.-X. Wu, Y. Jiang, and S.-F. Chen, "Genetic algorithm based selective neural network ensemble," in *Proceedings of the 17th International Joint Conference on Artificial Intelligence - Volume 2*, ser. IJCAI'01. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2001, pp. 797–802. [Online]. Available: http://dl.acm.org/citation.cfm?id=1642194.1642200

[17] A. Lazarevic and Z. Obradovic, "Effective pruning of neural network classifier ensembles," in *IJCNN'01. International Joint Conference on Neural Networks. Proceedings (Cat. No.01CH37222)*, vol. 2, July 2001, pp. 796–801 vol.2.

[18] K. F. Wallis, "Combining forecasts - forty years later," *Applied Financial Economics*, vol. 21, no. 1-2, pp. 33–41, 2011.

[19] Y. Xiao, J. Wu, Z. Lin, and X. Zhao, "A deep learning-based multi-model ensemble method for cancer prediction," *Computer Methods and Programs in Biomedicine*, vol. 153, pp. 1 – 9, 2018. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0169260717304947

[20] J. Magnus, O. Powell, and P. Prufer, "A comparison of two model averaging techniques with an application to growth empirics," *Journal of E*, vol. 154, pp. 139 – 153, 2010.
[21] A. E. Raftery, F. Balabdaoui, T. Gneiting, and M. Polakowski, "Using bayesian model averaging to calibrate forecast ensembles," *Monthly Weather Review*, vol. 133, pp. 1155–1174, 2005.
[22] G. Li, J. Shi, and J. Zhou, "Bayesian adaptive combination of short-term wind speed forecasts from neural network models," *Renewable Energy*, vol. 36, no. 1, pp. 352 – 359, 2011.
[23] A. Zellner, *Bayesian Inference and Decision techniques: Essay in Honor of Bruno de Finetti*. North-Holland, Amesterdam, 1986, ch. On assessing prior distributions and Bayesian regression analysis with g-prior distributions, pp. 233 – 243.
[24] C. Fernandez, E. Ley, and M. F. Steel, "Benchmark priors for bayesian model averaging," *Journal of Econometrics*, vol. 100, no. 2, pp. 381 – 427, 2001.
[25] M. Mackey and L. Glass, "Oscillation and chaos in physical control system," *Science*, vol. 197, pp. 287–289, 1977.
[26] J. FRIEDMAN, "Multivariate adaptative regression splines," *Annals of Statistics*, vol. 19, 1991.
[27] "Regression datasets." [Online]. Available: http://www.dcc.fc.up.pt/~ltorgo/Regression/DataSets.html