International Journal on Emerging Technologies **11**(3): 816-820(2020)

ISSN No. (Print): 0975-8364
ISSN No. (Online): 2249-3255

Implementation of Stacked Autoencoder with RBM for Predicting and Monitoring Aquatic Biodiversity

V. Murugan[1], J. Jeba Emilyn[2] and M. Prabu[3]
[1]Assistant Professor, Department of Computer Science and Engineering,
Trichy Engineering College, Trichy (Tamil Nadu), India.
[2]Associate Professor, Department of Information Technology,
Sona College of Technology, Salem (Tamil Nadu), India.
[3]Assistant Professor, Department of Computer Science and Engineering,
National Institute of Technology Calicut, Calicut (Kerala), India.

(Corresponding author: V. Murugan)
(Received 16 March 2020, Revised 13 May 2020, Accepted 16 May 2020)
(Published by Research Trend, Website: www.researchtrend.net)

ABSTRACT: Water pollution causes significant effects on water quality parameters. Aquatic biodiversity depends on water quality parameters for survival. Our research focuses on monitoring natural water resources for the management of aquatic biodiversity and water quality. Deep learning with its unique self-learning feature is capable of learning layer by layer and creating a new model for a given dataset. We developed a new semi-supervised framework for predicting and monitoring aquatic life by using a stacked auto-encoder with Restricted Boltzmann Machine (RBM) training. The unlabeled data is learned by stacked autoencoder and labeled data is learned by softmax classifier. The new combined network evaluates the available biodiversity. The simulation and prediction is done by analysis of various water quality parameters (pH, Dissolved Oxygen, Nitrate Content and Turbidity). The new model has better accuracy and less mean squared error for predicting aquatic biodiversity. Results from experiments show that this new framework is robust and accurate with good prospects for practical applications.

Keywords: Deep Learning, Restricted Boltzmann Machines, Stacked Autoencoder, Water Quality, Biodiversity.
Abbreviations: RBM, restricted boltzmann machines; DO, dissolved oxygen; pH, potential of hydrogen; ppm, parts per million.

I. INTRODUCTION

Sea, river, ponds, wetlands and aquatic living organisms form a major part of aquatic biodiversity. The aquatic biodiversity can be broadly classified as freshwater and marine biodiversity. Human needs are placed on aquatic biodiversity which leads to loss of aquatic biodiversity. Water quality and aquatic biodiversity is affected mainly due to overuse of water resources, water pollution and other alien species [16, 24]. Conservation and protection of aquatic biodiversity is very important as humans rely on this biodiversity for food, medicine and other materials. Not only humans, animals and the entire world ecosystem depends on aquatic biodiversity and any harmful effect on this will lead to serious issues. There are several water quality monitoring systems available to monitor water quality parameters [2, 3, 12, 14, 15, 20]. But a proper monitoring system using modern tools should be developed to monitor, protect and conserve aquatic biodiversity.

The existence of aquatic living organisms is governed by the physical and chemical water quality parameters [18]. Factors like concentration of ammonia, oxygen, nitrate, etc. contribute towards chemical properties of water. Factors like taste, odour, temperature are some of the physical properties of water. The concentration of hydrogen ions is termed as pH. The pH values vary from 0 to 14. For pure water pH is 7. When the pH value is less than 7 it is considered to be acidic and when greater than 7 it is alkaline. Different aquatic organisms live at different pH levels.

Excess amounts of nitrate in water will lead to the growth of algae and other water plants. When these plants or algae grow beyond a certain limit it will become hard for sunlight to enter into the water. Due to this type of plant growth, the excess amount of oxygen is generated which affects the equilibrium of DO in water. When equilibrium is affected it causes stress to fishes which in turn affects the reproduction cycle of fishes. Likewise, it causes a similar effect to organisms that take in oxygen. The habitat of fishes is greatly affected since they cannot find space for reproduction as most of the space is occupied by plants. High nitrate content can even cause direct illness to fishes and other water species.

Dissolved oxygen plays an important role in the survival of aquatic organisms. Dissolved oxygen is measured in ppm (Parts per Million). For a healthy aquatic biodiversity 5-6 ppm is required. The oxygen content in water is produced due to photosynthesis by plant bodies. Fishes and other aquatic organisms consume oxygen for survival. Normally during day, the time DO content will be high and during night time DO will be low. A constant DO equilibrium should be maintained for healthy aquatic life.

Turbidity of water is caused mainly due to suspended particles. The effects can be classified as lethal and sub-lethal effects. The lethal effects are more dangerous as they can kill the fishes and other life. The sub-lethal effects are less dangerous as it causes only tissue damage to water organisms. Due to industrial wastes high toxic products gets into water. The physical parameter temperature also plays a vital role as different temperatures suite different organism growth. Mainly due to this fishes are classified as hot and cold water fishes. So it is clear that there is a relationship that exists between different water quality parameters.

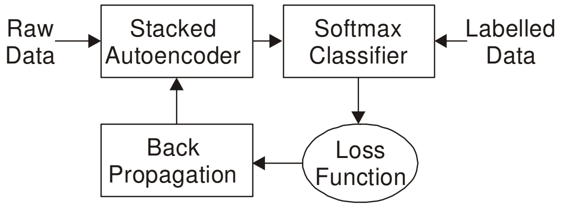

Fig. 1. Semi-Supervised Learning.

The relationship that exists between different water quality parameters can be studied using support vector machines, deep belief networks [21, 25], and many other neural networks. A neural network can be trained using different approaches. The two most common methods are supervised and unsupervised learning. In supervised learning the data input is labeled. The output is compared with input for accuracy and if it does not meet expectations it is trained again to get the desired output. In unsupervised learning the input is not labeled. The main goal of unsupervised learning is to find patterns inside the data and to represent in a useful manner. Semi-supervised learning is combination of supervised and unsupervised learning.

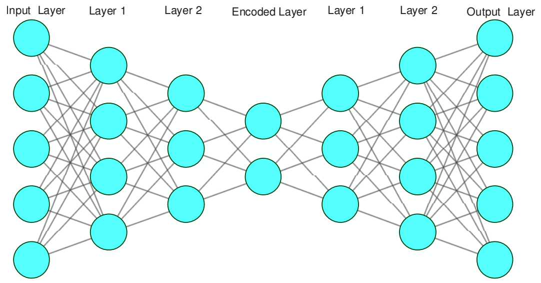

Fig. 2. Stacked Autoencoder.

An autoencoder is a type of neural network which has three layers. In an autoencoder, output units are connected back to the input units through hidden units. The input is given to the input layer which is then passed to the hidden layer. The hidden layer tries to reconstruct the original input as output.

A stacked autoencoder is a type of neural network in which the output of each layer is connected as the input of the successive layer as shown in Fig. 2. The stacked autoencoder has multiple hidden units connected together. The main aim of the stacked autoencoder is to minimize the loss and to reconstruct the accurate original input.

II. RELATED WORK

Tan and Eswaran (2008) compared the performance of stacked autoencoder and stacked autoencoder with RBM. Through their research, they were able to find that stacked autoencoder with RBM training has better performance than traditional stacked autoencoder. The experiment was conducted using an image dataset and the same hyperparameters were used for both models [19]. Yuan and Jia (2015) used a sparse autoencoder for the assessment of water quality parameters. A semi-supervised learning model comprising of a sparse autoencoder and a softmax classifier was used to evaluate the water quality by comparing the labeled and unlabeled data [21]. Zhang *et al.*, (2018) clearly explains about Restricted Boltzmann Machines, training methods and application areas of RBM. They also explain the construction of deep neural networks using RBM and how RBM can be combined with a convolutional neural network [23].

Zhou *et al.*, (2018) developed a water quality prediction technique using Improved Grey Relational Analysis (IGRA) algorithm and Long-Short Term Memory (LSTM). IGRA was used for feature selection and LSTM was used for prediction using a very large amount of historical data from Tai Lake and Victoria Bay [25]. Demetillo *et al.,* (2019) developed a system for monitoring water quality at a low cost using wireless sensor network which can cover a large area. Real-time data was sent to the web using GSM technology [4]. The previous studies were mainly focusing on water quality monitoring and prediction only. But there exists a relationship between water quality and biodiversity, so we propose a new system for the prediction and assessment of aquatic biodiversity using water quality parameters.

III. METHODOLOGY

A deep learning network is a generative graphical model. It is represented as a solution of vanishing gradient problem. The proposed methodology is to train the stacked autoencoder with RBM. A deep belief network is constructed by stacking more than one autoencoder. The output of the first autoencoder with RBM is fed as input to the next single hidden layer autoencoder. While training the autoencoder, four hyperparameters has to be considered namely code size, number of layers, number of nodes in each layer and finally the loss function. Since the input is not binary MSE is used as loss function here.

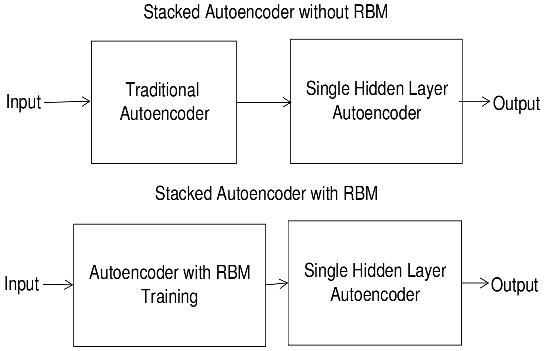
Stacked Autoencoder without RBM
Input
Traditional Autoencoder
Single Hidden Layer Autoencoder
Output
Stacked Autoencoder with RBM
Input
Autoencoder with RBM Training
Single Hidden Layer Autoencoder
Output

Fig. 3. Autoencoder with and without RBM.

ReLU is used as an activation function since the gradient is independent of the variable size. The number of stacking depends on the requirement. The back propagation is used to fine-tune the overall weights.

As stated earlier the first stack is trained as RBM. An RBM can extract the features and reconstruct the input. The training approach used here is Contrastive Divergence. During the first step of greedy training the first stack of autoencoder with RBM is created. The inputs to the RBM are mapped to the first layer (V_1, V_2,

…… V_n). The classes or outcomes are mapped to the first hidden layer (H_1,H_2, ….. , H_n). Normally training of RBM means training the optimal weights (W_1,W_2, …., W_n). The optimal weight is represented by W_1 for the first layer.

A. RBM Training by Contrastive Divergence

Step 1: Choose the dataset for the training. States for visible units are set for training dataset.

Step 2: Parallel hidden units are updated i.e. positive statistics for each edge is computed for E_{ij} Pos(E_{ij}), which is given by P(H_j=1|V). This is called positive phase. Activation probability of each hidden layer is given by

$$\mathrm{P(H_j=1|V)}=\sigma(B_j+\sum_{i=1}^{m} W_{ij}V_i) \qquad (1)$$

Step 3: Using similar technique, the visible units are reconstructed which is called negative phase. Negative statistics for each edge is computed for E_{ij} Neg(E_{ij}) which is given by P(V_i=1|H).

Activation probability of each visible layer is given by

$$\mathrm{P(V_i=1|H)}=\sigma(A_i+\sum_{j=1}^{n} W_{ij}H_j) \qquad (2)$$

Step 4: Next step is to update the weights of edges. The weight of the edge W_{ij} is updated to new weight

$$\mathrm{Upd(W_{ij})= W_{ij}}+L*(Pos(E_{ij})-Neg(E_{ij})) \qquad (3)$$

Here L is called as learning rate.

Step 5: The above steps are repeated for all training set till the threshold is achieved.

B. Greedy Technique

Step 1: Set parameters for W_1 of the first layer RBM.

Step 2: Vector W_1 defines the first layer features and Samples H_1 from P(H_1|V) = P(H_1|V,W_1) is fed as input for the next corresponding layer.

Step 3. Vector W_2 defines the second layer features and Samples H_2 from P(H_2|H_1) = P(H_2|H_1,W_2) is fed as input for the next corresponding layer.

Step 4. Recursive approach is used for further layers for the desired outcome.

IV. EXPERIMENTS AND DISCUSSION

A. Data Collection

Data collection forms the first and very important step of research process. The data is collected from the Indian Ministry (Forest and Climatic Change) official website. The data consist of water quality parameters of several water bodies all over India [5]. The collected data contains information about pH, Dissolved Oxygen, Nitrate content, turbidity and temperature. The data about biodiversity in a particular location is got through other internet resources, physical visits, etc.

Table 1: Water Quality Range.

Water Quality Parameter(Unit)	Range	Species
pH	3.6 – 10.0	Fish
Dissolved Oxygen(ppm)	2.1 – 3.9	Insects
Turbidity(NTU)	300 - 500	Worms
Nitrate(ppm)	3.3 – 4.5	Crab
Temperature (^{o}C)	30 - 32	Crab
	18-25	Fish

Table 1 gives information about the different water quality parameters and their range in which an aquatic organism survives. The unlabeled data contains only water quality parameters whereas the labeled data has both water quality parameters and available biodiversity as shown in Table 2 and 3. The available biodiversity data is labeled manually using Table 1 values. It is to be noted that the table gives only information about the favorable environment for the existence of certain organisms and it doesn't imply that other species cannot exist in the given environment with a given range. For example, in the range of (2.0-4.0) ppm of nitrate fishes can live and in (3.3 – 4.5) ppm of nitrate crabs can live. It means that in (3.3 – 4.0) ppm of nitrate both fishes and crabs can survive.

Table 2: Sample Unlabeled Data.

Site	pH	DO	Ni	Turbidity	Temp.
101	7.13	5.09	0.21	144	20.1
102	6.98	6.11	0.12	143	19.8
103	6.88	7.12	3.22	149	19.2
104	6.83	5.98	0.24	263	22.4
105	8.11	6.01	0.13	353	24.2
116	7.65	5.99	0.23	241	20.3
107	7.55	2.02	0.20	300	22.1
108	7.66	6.22	0.11	302	21.2
109	7.98	3.90	0.21	149	20.0
110	8.88	5.66	0.21	263	25.0
111	7.60	5.96	0.15	353	25.2
112	6.98	6.02	0.17	241	23.1
113	6.99	5.02	2.20	353	20.5
114	7.33	5.67	2.21	241	18.6

Table 3: Sample Labeled Data.

Site	pH	DO	Ni	Temp.	Species
101	7.13	5.09	0.21	20.1	Fish
102	6.98	6.11	0.12	19.8	Fish
103	6.88	7.12	3.22	29.9	Crab
104	6.83	5.98	0.24	22.4	Fish
105	8.11	6.01	0.13	24.2	Fish
116	7.65	5.99	0.23	20.3	Fish
107	7.55	2.02	0.20	22.1	Insect
108	7.66	6.22	0.11	21.2	Fish
109	7.98	3.90	0.21	20.0	Insect
110	8.88	5.66	0.21	25.0	Fish
111	7.60	5.96	0.15	25.2	Fish
112	6.98	6.02	0.17	23.1	Fish
113	6.99	5.02	2.20	20.5	Worm
114	7.33	5.67	2.21	18.6	Fish

B. Data Preprocessing

Before processing the data, it is necessary to remove noise or any unwanted data. This can be done by removing replicated data, empty data or any other unnecessary data. The data collected contains maximum and minimum values, so mean is considered as the final value.

C. Results and Discussion

We conducted experiments to predict the biodiversity available on aquatic environments using two models. The dataset contained 200 distinct sites. Each site's dissolved oxygen, pH, nitrate content, turbidity, temperature and available biodiversity is taken as contents of the labeled dataset and site's dissolved oxygen, pH, nitrate content, turbidity and temperature is taken as contents unlabeled dataset. The softmax classifier predicts the biodiversity by comparing labeled and unlabeled data. The same learning rate of 1.0 was used for both models. A total of 250 epochs were used and results are shown after 200 epochs. For RBM without the stacked autoencoder random weights were assigned initially.

For stacked autoencoder the output of the previous hidden layer act as the input for the next layer. The stacked autoencoder was trained using backpropagation method. The hidden layers were again trained for the next 50 epochs. This process was iterative and more hidden layers were attached onto the autoencoder. A total number of 250 epochs were used in this case also. Fine-tuning of the stacked autoencoder with RBM done in phases by stochastic method as stated earlier.

Fig. 4 shows the plot of MSE vs Epochs for stacked autoencoder without RBM. We can explicitly see that MSE is high during the initial stage and gets reduced after some epochs. It is also to be noted that there is no much difference in MSE during the training and testing phases. Fig. 5 shows the plot of MSE vs Epochs for stacked autoencoder with RBM. It can be noted that MSE is high during the initial stage and gets reduced after some epochs. There is a lot of difference in MSE during the training and testing phases. The autoencoder with RBM outperforms the autoencoder without RBM in accuracy. The performance comparison of both encoders during training and testing phases are shown in Table 4.

Table 4: Comparison of Mean Squared Errors.

Autoencoder	**Mean Squared Error**	
	Training	**Testing**
Stacked autoencoder without RBM	5.98	6.73
Stacked autoencoder with RBM	3.81	4.22

Considering the time and space complexity of the two models, Stacked autoencoder with RBM consumes more time and occupies more space than stacked autoencoder without RBM, since an additional layer of RBM is added to the model. We have also tried with experimenting by reducing the number of features like excluding ph, DO, nitrate content, turbidity, etc., one by one in each experiment but the model gives higher accuracy only when all the features are supplied.

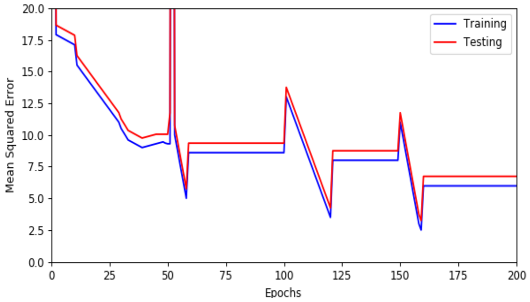

Fig. 4. MSE of Stacked Autoencoder without RBM

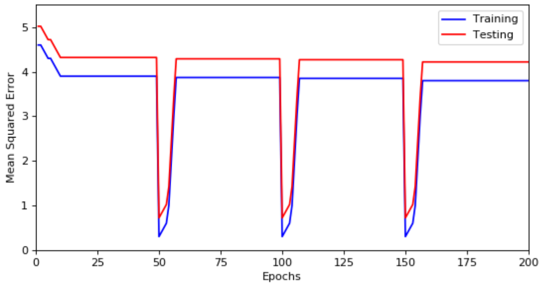

Fig. 5. MSE of Stacked Autoencoder with RBM.

V. CONCLUSION AND FUTURE SCOPE

We propose a framework for predicting the aquatic biodiversity through semi-supervised learning methodology using stacked autoencoder with RBM. Many aquatic species are becoming extinct due to pollution, global warming and other reasons. This methodology helps in identifying the suitable habitat for aquatic biodiversity through learning of water quality parameters and their relationship with aquatic biodiversity. This intelligent system can be extended for investigating the reason for the reduction in the population of a particular aquatic species and protect them from becoming extinct or endangered species. The system can also be integrated with an IoT environment. The data for the current system is collected through an internet repository. In future we propose to develop a system to collect real-time water quality parameters using sensors and IoT infrastructure.

Conflict of Interest. No.

REFERENCES

[1]. Babić, G., Vukovic, M., Voza, D., Takic, L., & Mladenovic-Ranisavljevic, I. (2019). Assessing Surface Water Quality in the Serbian part of the Tisa River Basin. *Polish Journal of Environmental Studies*, 28(6), 4073-4085.

[2]. Bannikoppa, S., & Bhairi, V. (2015). Water Quality Monitoring Over Wireless Sensor Network and Purification. *International Journal on Emerging Technologies*, 6(2), 1-6.

[3]. Brown, R. S., & Hussain, M. (2003). *The Walkerton tragedy—issues for water quality monitoring*. The Analyst,*128*(4), 320-322.

[4]. Demetillo, A. T., Japitana, M. V., & Taboada, E. B.(2019). A system for monitoring water quality in a large aquatic area using wireless sensor network technology. *Sustainable Environment Research*, *29*(1), 1-9

[5]. ENVIS Centre on Control of Pollution Water, Air and Noise. Retrieved from http://www.cpcbenvis.nic.in

[6]. Freshwater-Aquaculture and Water Quality in Aquaculture. Retrieved from https://freshwater-aquaculture.extension.org/

[7]. Jin, J., Jiang, P., Li, L., Xu, H., &Lin, G. (2019). Water quality monitoring at a virtual watershed monitoring station using a modified deep extreme learning machine. *Hydrological Sciences Journal*, *65*(3), 415–426.

[8]. Kupe, L., Schanz, F., & Bachofen, R.(2008). Biodiversity in the Benthic Diatom Community in the Upper River Töss Reflected in Water Quality Indices. *CLEAN – Soil, Air, Water*, *36*(1), 84–91.

[9]. Li, W., Fu, H., Yu, L., Gong, P., Feng, D., Li, C., andClinton, N. (2016). Stacked Autoencoder-based deep learning for remote-sensing image classification: a case study of African land-cover mapping. *International Journal of Remote Sensing*, *37*(23), 5632–5646.

[10]. Li, Z., Peng, F., Niu, B., Li, G., Wu, J., &Miao, Z. (2018). Water Quality Prediction Model Combining Sparse Auto-encoder and LSTM Network. *IFAC-PapersOnLine,51*(17), 831–836.

[11]. Muharemi, F., Logofătu, D., & Leon, F. (2019). Machine learning approaches for anomaly detection of

water quality on a real-world data set. *Journal of Information and Telecommunication*, *3*(3), 294–307.

[12]. Nagaoka, H., and Sanda, K., (2005). Simulation of turbulence and dissolved oxygen concentration profiles over biofilm using k–ε turbulence model. *Water Science and Technology*, *52*(7), 181–186.

[13]. Qiu, Y., Liu, Y., & Huang, D. (2016). Date-Driven Soft-Sensor Design for Biological Wastewater Treatment Using Deep Neural Networks and Genetic Algorithms. *Journal of Chemical Engineering of Japan*, *49*(10), 925–936.

[14]. Ross, M. R., Topp, S. N., Appling, A. P., & Yang, X. (2019). AquaSat: A Data Set to Enable Remote Sensing of Water Quality for Inland Waters. *Water Resources Research*, *55*(11), 10012-10025.

[15]. Shah, K. A. & Joshi, G. S. (2017). River Water Quality Modelling for the Assessment of the Impact of Urbanization. *International Journal on Emerging Technologies*, *8*(1), 196-201.

[16]. Sofi, I. R., Chuhan, P. P., Sharma, H. K., & Manzoor, J. (2018). Assessment of Physico-Chemical Properties of Water and Sediments of Asan Lake Dehradun, India. *International Journal of Theoretical and Applied Sciences*,*10*(1), 68-76.

[17]. Solanki, A., Agrawal, H., & Khare, K. (2015). Predictive Analysis of Water Quality Parameters using Deep Learning. *International Journal of Computer Applications*, *125*(9), 29–34.

[18]. Steve, O. N., Phillip, O. R., & Alfred, A. (2014). The impact of water quality on species diversity and richness of macroinvertebrates in small water bodies in Lake Victoria Basin, Kenya. *Journal of Ecology and The Natural Environment*, *6*(1), 32–41.

[19]. Tan, C. C., & Eswaran, C. (2008). Performance Comparison of Three Types of Autoencoder Neural Networks. *2008 Second Asia International Conference on Modelling and Simulation (AMS)*, 213–218.

[20]. Vica, M. L., Popa, M., Matei, H. V., Glevitzky, I., &Siserman, C. V.(2018). Study of groundwater quality in urban area of alba iulia, Romania. *Journal of Environmental Protection and Ecology*, *19*(4), 1481-1489.

[21]. Yuan, Y., & Jia, K. (2015). A water quality assessment method based on sparse autoencoder. *2015 IEEE International Conference on Signal Processing, Communications and Computing (ICSPCC)*, 1-4

[22]. Zhang, C., Chen, Y., Xu, B., Xue, Y., & Ren, Y. (2019). How to predict biodiversity in space? An evaluation of modelling approaches in marine ecosystems. *Diversity and Distributions*, *25*(11), 1697-1708.

[23]. Zhang, N., Ding, S., Zhang, J., & Xue, Y. (2018). An overview on Restricted Boltzmann Machines. *Neurocomputing*, *275*, 1186–1199.

[24]. Zhi, Z., & Fang, L. (2019). Research on the water pollution monitoring and rapid decision-making system based on artificial intelligence agent. *Journal of Environmental Protection and Ecology*, *20*(3),1565-1573.

[25]. Zhou, J., Wang, Y., Xiao, F., Wang, Y., & Sun, L. (2018). Water Quality Prediction Method Based on IGRA and LSTM. *Water*, *10*(9), 1-9.

How to cite this article: Murugan, V., Emilyn, J. J. and Prabu, M. (2020). Implementation of Stacked Autoencoder with RBM for Predicting and Monitoring Aquatic Biodiversity. *International Journal on Emerging Technologies*, *11*(3): 816–820.