Prediction of Abalone Age Using Regression-Based Neural Network

Conference Paper · September 2019

DOI: 10.1109/AiDAS47888.2019.8970983

7 authors, including:

Muhammad Faiz Misman
University of Technology Malaysia
18 PUBLICATIONS **154** CITATIONS
SEE PROFILE

Azurah A Samah
University of Technology Malaysia
41 PUBLICATIONS **319** CITATIONS
SEE PROFILE

Hairudin Abdul Majid
University of Technology Malaysia
45 PUBLICATIONS **261** CITATIONS
SEE PROFILE

Zuraini Ali Shah
University of Technology Malaysia
35 PUBLICATIONS **227** CITATIONS
SEE PROFILE

Prediction of Abalone Age Using Regression-Based Neural Network

Muhammad Faiz Misman, Azurah A Samah, Nur Azni Ab Aziz, Hairudin Abdul Majid, Zuraini Ali Shah, Haslina Hashim, Muhamad Farhin Harun
Artificial Intelligent and Bioinformatics Group (AIBIG), Department of Information System,
School of Computing, Faculty of Engineering, Universiti Teknologi Malaysia,
81310, Johor Bahru, Johor, Malaysia
faizmisman@gmail.com, azurah@utm.my, nurazniaziz@gmail.com, hairudin@utm.my, aszuraini@utm.my, haslinah@utm.my, farhin1993@gmail.com

***Abstract*— Artificial neural networks (ANN) has been widely used to speed up data prediction operations with over thousands of features available. In this paper, we propose a regression-based ANN model with three hidden layers to predict the age of abalones. It is salient to predict abalone age as it helps farmers and sellers to determine the market price of abalones. The economic value of abalone is positively correlated with their respective ages. The age of the abalone can be estimated by measuring the number of layers of shell rings The model was built based on a dataset obtained from the UCI Machine Learning Repository. Before developing and training the model, a pre-processing methodology was applied to the dataset. Parameters tuning, which involves modifications in the number of hidden layers as well as the number of epochs, were done to obtain the best result. The finalised results were analysed and the results show that physical measurements of abalone can predict its respective age with less time consumption. This study has shown a result of low root mean-squared error, obtained from the proposed model in comparison with other methods stated in this study. Finally, the proposed model was validated using test dataset, and the results reveal a lower root-mean-squared error value in contrast to the value obtained during model training.**

Keywords-artificial neural network; prediction; regression problem; abalone age

I. INTRODUCTION

Abalone is a type of single-shelled marine snails. It is also known as the snail of the sea with a large fleshy body including broad muscular food, which has become an expensive gourmet delicacy due to its short fishing seasons [1-2]. The value of abalone is economically related to its age. Therefore, the age of the abalone plays a vital role for both farmers and consumers to determine its price [3].

Determining its age is required in scientific studies such as marine biology on abalone. Conventionally, the numbers of layers of shell rings are measured to estimate the abalone age [2]. The rings formed corresponds to the growth of the abalone [3]. The process of determining the abalone age starts with cutting down the shell of an abalone. The shell cone is then stained so that the rings will be more prominent to count [4].

Abalone age can be estimated based on the number of dark bands on the right-hand side of the section [5]. Due to the uncertainty of completely staining all rings, researchers have decided to add a value of 1.5 during the ring count to improve the approximation of the abalone ages. This conventional method affects farmers in terms of costs and time processing [3]. Therefore, to improvise this process, machine learning can be implemented to help researchers using a large number of data containing physical measurements of abalone to predict its age in a short amount of time.

II. EXPERIMENTAL FRAMEWORK

The experimental framework is as shown in Fig. 1.

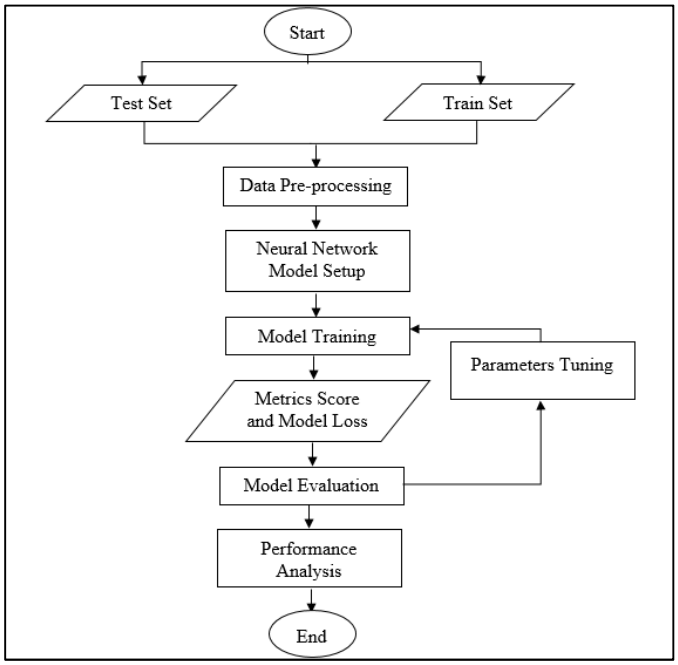

Fig. 1. Experimental Framework

The framework is initiated with data pre-processing, to be continued with model development along and model evaluation phase, as shown in Fig. 1.

A. Data Collection and Pre-processing

The data obtained from the UCI Machine Learning Repository [4], consists of 4177 samples with nine attributes. Each sample represents eight numerical attributes and one categorical attribute. The target data is fixed according to the number of rings for each sample of abalone. After the files have been created, the test dataset contained 1,253 samples while train dataset comprised of 2,923 samples. We randomly split the datasets into training and testing sets according to a ratio of 70:30. The dataset described is shown in Table I.

978-1-7281-3041-5/19/$31.00 ©2019 IEEE

TABLE I. DATASET DESCRIPTION

Abalone Dataset		
No.	***Attributes***	***Values***
1	Sex	0,1
2	Length	0.0750 – 0.8150
3	Diameter	0.0550 – 0.6500
4	Height	0.0000 – 1.1300
5	Whole Weight	0.0020 – 2.8255
6	Shucked Weight	0.0010 – 1.4880
7	Viscera Weight	0.0005 – 0.7600
8	Shell Weight	0.0015 – 1.005
9	Rings	1.000 – 29.000
Output Values		
1	Age	Real value (float)

"Sex" attribute was converted into binary format, where 1 represent male while 0 represent female.

Next, the dataset undergoes the data cleaning process, where all the missing values were removed. A sample had to be removed from the dataset due to the abundance of null values in "Height" attribute. Next, the datasets were standardised to range each attribute within 0 to 1. Finally, the correlation matrix was calculated and visualised through the heat map to observe the rate of dependency between variables in the dataset.

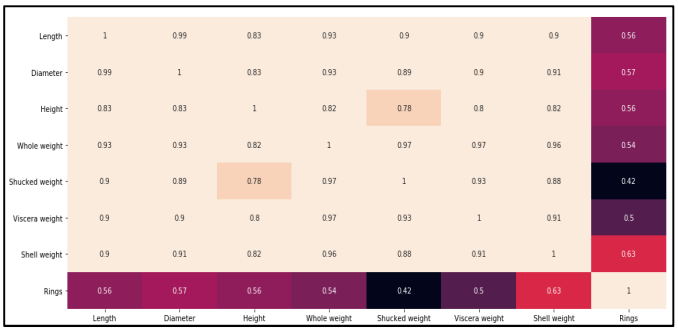

Fig. 2. Heat map representing Data Correlation

According to the heat map above, we can conclude that the features are highly correlated with each other. Majority of the correlation values range from 0.8 to 1.0, as this indicates that our dataset is good enough to be used for the specified case study.

B. *Artificial Neural Network Modelling*

Artificial Neural Network (ANN) is made up of several components including (1) input layer (2) hidden layer(s) (3) output layer (4) weights, (5) bias, and (6) activation functions. Data flows according to feedforward neural network as input layer plays the receiver's role, receiving data and sending it to the hidden layer (processing unit), which sends processed data to the output layer. Weights are capable of modifying inputs to impact its respective output [7]. Bias are values that are added to sums that are calculated at each node excluding input nodes, during the feedforward phase [8]. Activation of neurons or nodes in ANN architecture depends on activation function, which decides whether the information received by the nodes are relevant or otherwise [9].

Later on, the network architecture will be set up. This involves the number of nodes, number of layers, and types of activation function used in each layer. Fig. 3 shows the network architecture for the model that has been proposed.

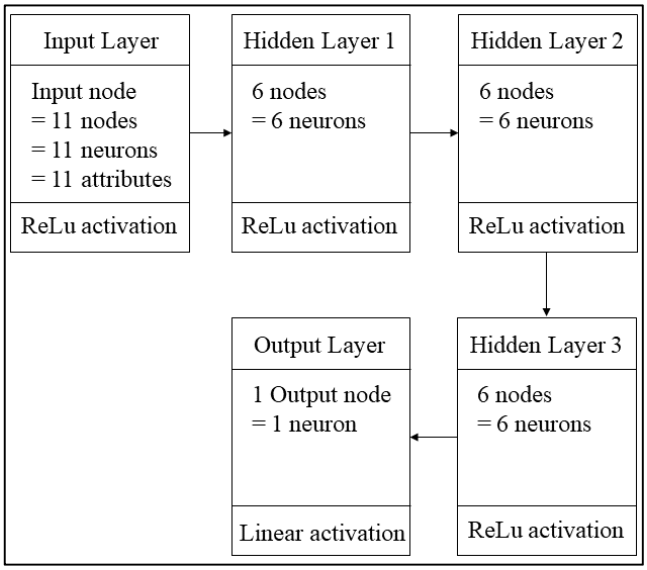

Fig. 3. Network architecture for the proposed model

The input layer consists of 11 neurons or nodes referring to the number of the attributes from the dataset. The hidden layers were set into 6 nodes each, while one node is assigned at the output layer. The network was trained using training data. Rectified Linear Unit (ReLU) activation function was used in all hidden layers to ensure the maximum value of the variables can be passed through the network during back-propagation [10]. Apart from that, ReLU turns negative variables into zero [11]. Since this study is based on the linear regression problem, linear activation function was applied in the output layer.

III. MODEL IMPLEMENTATION AND ASSESSMENT

The training dataset was utilised to perform model implementation for the prediction of abalone age. The training set was split into a ratio of 60:40, assigned into training and validation sets respectively. The model assessment was done by measuring model performance based on the results. After the results were obtained, further analysis will be done to analyse whether the model could be improvedbased on the size of parameters.

A. *Performance Metrics*

There are two metrics used for this study: (1) Mean Absolute Error (MAE) and (2) Root Mean Squared Error (RMSE). These two metrics were considered as this study is based on the regression problem. MAE is defined as the absolute average difference between each predicted values and observed values [12] and was set to be the loss function in this study. MAE Equation (1) shows the formula of MAE:

$$\text{MAE} = \frac{1}{n}\sum_{j=i}^{n} |y_j - \hat{y}_j| \quad (1)$$

RMSE is defined as Root Mean Squared Error as it has a greater amount of similarity to mean absolute error in terms of

contribution. However, RMSE calculates the square root of the average difference while MAE calculates the absolute average difference [12]. Equation (2) shows the formula of RMSE:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{j=i}^{n} |y_j - \hat{y}_j|} \quad (2)$$

The $\hat{y}$ in both equations represent the predicted target value while y represents the actual target value. The value of RMSE is generally bigger or may be equal to MAE since it squares the error.

B. Parameters Tuning

Table II shows the list of parameters used in the model.

TABLE II. PARAMETERS IN PROPOSED MODEL

Parameters	*Description*	*Value*
Number of hidden layers	Layers between the input layer and the output layer. Less number of layers may generate high bias. However, many layers may cause overfitting [13].	3
Batch size	Sub sample is taken from the data every time update in parameters occurs [14].	32
Number of epochs	The number of times the data set is passed forward and backward in the network [15].	400
Dropout	Noise added on a certain layer to make the model more robust [14].	0.5

To improve the model efficiency in predicting abalone age, adjustments in parameter sizes may be needed to get the best result from the model

IV. RESULTS AND ANALYSIS

The model was trained using the training set. Table III shows the result of the model training using a different number of epochs with two hidden layers. We measured the model fit by comparing training and validation loss. When training loss is far lower than the validation loss, the model is considered as over fit. "loss" keyword in the tables indicates the training loss while "val_loss" represents validation loss obtained from the validation set. "rmse" indicates RMSE training value while "val_rmse" indicates the RMSE value for the validation set.

TABLE III. 2 HIDDEN LAYERS WITH BATCH SIZE 32

Epochs	*Loss*	*RMSE*	*Model Fitting*
100	loss: 0.0631 val_loss: 1.0431	rmse: 0.0861 val_rmse: 1.3584	**Over fit**
200	loss: 0.0101 val_loss: 1.0001	rmse: 0.0114 val_rmse: 1.3039	**Over fit**
300	loss: 0.0094 val_loss: 1.0000	rmse: 0.0103 val_rmse: 1.3040	**Over fit**
400	loss: 0.0080 val_loss: 1.0026	rmse: 0.0091 val_rmse: 1.3072	**Over fit**

From Table III, we can say that the model for each epoch over fit since the training loss values are much lower compared to validation loss. Another observation from Table III is the number of loss becomes lower as the number of epochs increases. Table IV shows the result of using the same parameters as in Table III, differ by the addition of dropout layers in between hidden layers with batch size assigned 32.

TABLE IV. 2 HIDDEN LAYERS WITH DROPOUT LAYERS

Epochs	*Loss*	*RMSE*	*Model Fitting*
100	loss: 3.4817 val_loss: 2.5014	rmse: 4.5097 val_rmse: 3.0178	**Good fit**
200	loss: 1.7208 val_loss: 1.6360	rmse: 2.5110 val_rmse: 2.1534	**Good fit**
300	loss: 1.5258 val_loss: 1.5846	rmse: 2.2931 val_rmse: 2.1255	**Good fit**
400	loss: 1.5104 val_loss: 1.4988	rmse: 2.3323 val_rmse: 2.0622	**Good Fit**

The model showed a good fit after dropout layer has been applied. Despite the overall loss and RMSE value higher than the ones obtained from experiments without using dropout layers, the results from Table IV were preferred as it demonstrated a good fit model. Over fitting the model is not desired since it might give out less performance for test dataset later.

Next, the number of hidden layers were added to see whether the model can get a lower error rate or otherwise. This is because when the network is deeper, the test set accuracy might increase too. Table V shows the error results after using three hidden layers with batch size 32.

TABLE V. 3 HIDDEN LAYERS WITH BATCH SIZE 32

Epochs	*Loss*	*RMSE*	*Model Fitting*
100	loss: 0.0103 val_loss: 1.0012	rmse: 0.0114 val_rmse: 1.3048	**Over fit**
200	loss: 0.0087 val_loss: 1.0058	rmse: 0.0095 val_rmse: 1.3094	**Over fit**
300	loss: 0.0095 val_loss: 1.0003	rmse: 0.0096 val_rmse: 1.3033	**Over fit**
400	loss: 0.0042 val_loss: 0.9969	rmse: 0.0050 val_rmse: 1.3020	**Over fit**

From table V, the model for each epoch still over fit, but the number of loss became lower as the number of epochs increased. Also, the overall loss and RMSE value slightly decreased compared to the results in Table III. Therefore, dropout layers were applied once again to see whether this model can overcome overfitting or not. Table VI shows the result of using the same parameters as in Table V, differing by the addition of dropout layers in between hidden layers.

TABLE VI. 3 HIDDEN LAYERS WITH DROPOUT LAYERS

Epochs	*MAE*	*RMSE*	*Model Fitting*
100	loss: 3.1525 val_loss: 2.1126	rmse: 4.1429 val_rmse: 2.7104	**Good fit**
200	loss: 2.0707 val_loss: 1.7881	rmse: 2.9272 val_rmse: 2.3067	**Good fit**
300	loss: 1.5258 val_loss: 1.5846	rmse: 2.2931 val_rmse: 2.1255	**Good fit**
400	loss: 1.2903 val_loss: 1.3669	rmse: 2.0691 val_rmse: 1.8552	**Good fit**

From Table VI, the model has achieved a good fit when using three hidden layers with dropout layers inserted in between the fully connected layers. The value of training loss has become higher compared to validation loss. In addition, the overall loss and errors has decreased as the number of hidden layers increased. Therefore, models representing good fit were selected to undergo fine-tuning as the lowest error possible is being considered in which, the results produced

through the implementation of 400 epochs that were selected for our case.

A. *Results Visualization*

Fig. 3 shows a graph illustrating overall error value and loss value corresponding to the number of hidden layers with dropout layers.

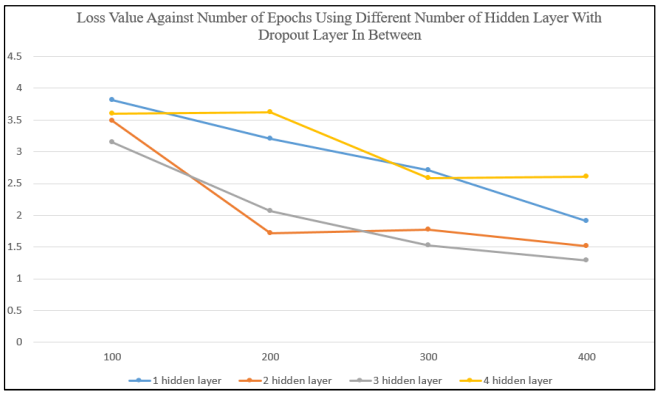

Fig. 4. Loss value against the number of epochs using a different number of hidden layers with Dropout Layer in between

According to Fig. 3, almost all hidden layer lines show a decreasing value of loss. This concludes that higher amount of hidden layer would result in a lower rate of loss value. Among all the 4 lines in the graph, the line (grey) that represent 3 hidden layers shows inversely descending graph.

Based on Table VI, the experiment with 400 epochs was chosen as the best possible model to predict the age of abalones. Fig. 4 shows the loss graph. From the figure, the training loss is slightly higher than the validation loss.

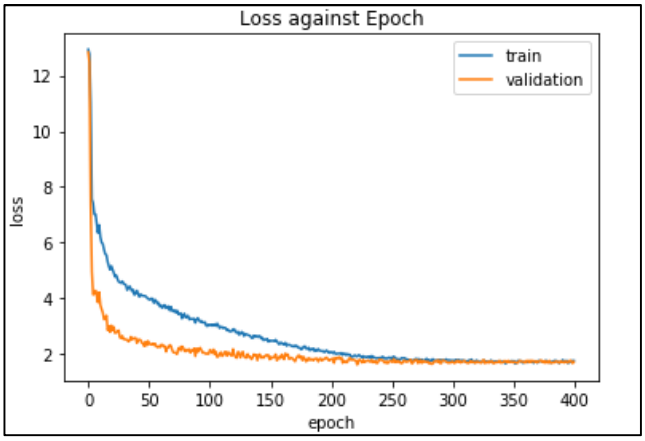

Fig. 5. Graph of loss against the number of epochs

Another measurement which is considered to evaluate the performance of the model is RMSE. The lower the RMSE, the better the performance of the model [12]. Fig. 5 shows the errors against the number of epochs. The graph has shown that the training error is significantly higher than the validation error. Therefore, the model is considered not over fit.

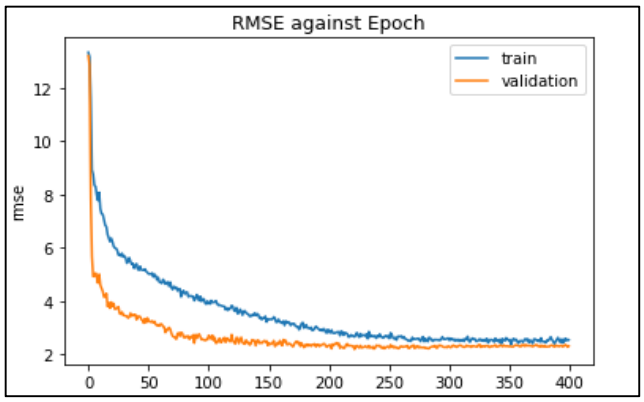

Fig. 6. Graph of RMSE against the number of epochs

After that, the predicted value for each sample was obtained. The prediction values were then kept in a submission file which was created after the prediction function is executed. Then, each predicted value was compared to the actual value. Fig. 6 below shows the predicted value of abalone age and the actual value of abalone age retrieved from the dataset. From this figure, the prediction value is considerably close to the actual value.

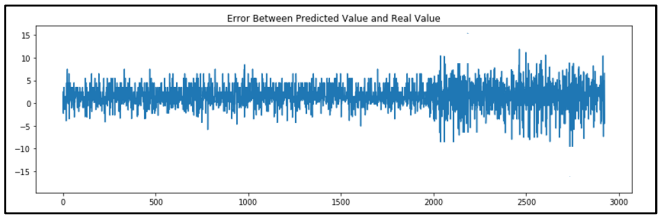

Fig. 7. Graph of the difference of predicted abalone age and actual abalone age

B. *Discussion of Proposed Model Results with Other Methods*

To justify the evaluation of the proposed model performance, comparisons between models are made in terms of error rate assessment. Therefore, results from previous studies were taken into consideration as other methods will use the same dataset to perform comparative analysis. Table VII shows the list of proposed model performance and the previous models in RMSE value.

TABLE VII. LIST OF PROPOSED MODEL PERFORMANCE AND PREVIOUS MODELS IN RMSE VALUES

Method	***ANN + Fuzzy Logic [16]***	***Random Forest [17]***	***ANN + Principal Component Analysis [18]***	***Proposed Model (Regression-based Neural Network Within Keras Library)***
RMSE value	8.300	2.6700	4.3500	2.0691
Improvement	6.2309	0.6009	2.2809	
Improvement Percentage	75.07%	12.50%	52.43%	

As shown in Table VII, the proposed method demonstrates an improvised version of the previous researcher's model in terms of error rate. Implementation of ANN with Fuzzy Logic method illustrates that the researcher has used Fuzzy Logic rules based on the neural network model built with two hidden layers and with only 50 epochs followed by one Fuzzy rule

layer. Fuzzy Logic is a logical system that mimics human reasoning which comes from various aspects such as uncertainty and judgements. Moreover, computers can only manipulate precise valuations. Fuzzy logic provides a combined solution to these aspects as such a model is then called an Adaptive-Networks-based Fuzzy Inference System (ANFIS) [19]. From the paper, Fuzzy inference system will be utilised along with the guidance conducted by a training algorithm, which enables parameters of the fuzzy system to be tuned [16]. The system starts with the fuzzification process, which translates inputs into truth value, followed by Fuzzy rule evaluation, which computes output truth value, and lastly defuzzification which transfer truth value to output [20]. Based on the study discussed, the RMSE value obtained for prediction of abalone age using ANFIS model is 8.300.

For regression random forest method, a new ensemble method for feature selection with neural networks called Random Brains was introduced. Random Brains is derived from the random forest algorithm. Random forest is made up of multiple decision trees that are merged together to get a stable and accurate prediction. The neural network is set up with only two hidden layers constantly. Then, the dataset is fed into the network without using Random Brains at the initial phase. After that, the dataset is compared to the ones that used Random Brains. However, RMSE value obtained from random forest method (with Random Brains) only managed to get anywhere near the RMSE value compared to the previous network which is without Random Brains at 2.6700 for abalone dataset [17].

For ANN with Principal Component Analysis (PCA) method, PCA is being implemented on the dataset before feeding them into the neural network. The aim of this method is to reduce the dimensionality of the dataset. Due to the deficient number of attributes in abalone dataset, the method obtained only 4.3500 RMSE value. In this study, the hidden layers were set to one with decay being assigned into 0.01 to avoid overfitting[18].

Based on the discussion for each of the method according to previous studies, the number of hidden layers might play a role in assuring low RMSE value. Among all the available methods in the table, the proposed model used three hidden layers. It also obtained the lowest RMSE value compared to any other methods considered throughout our case study. Therefore, the proposed method shows that this method is suitable to be used for a dataset with over thousands of samples.

C. Testing Proposed Model Using Test Dataset

After the model has been trained, a new dataset is fed into the model to see if the model is capable to do prediction on unseen data. The new dataset used is the test dataset that has been split from abalone dataset earlier before the network modelling was initiated. By implementing the exact model parameters used during training, the network analyses the new dataset. The test dataset contains 1253 samples. Table VIII shows the result of the new dataset in RMSE value and loss value.

TABLE VIII. RESULTS OF NEW DATASET

Epochs	***MAE***	***RMSE***
400	loss: 1.5277	rmse: 1.7447

From the result, we can conclude that RMSE value and loss value shows a significant decrease compared to the ones during training. Fig. 7 and Fig. 8 shows the visualisation of the result using test dataset.

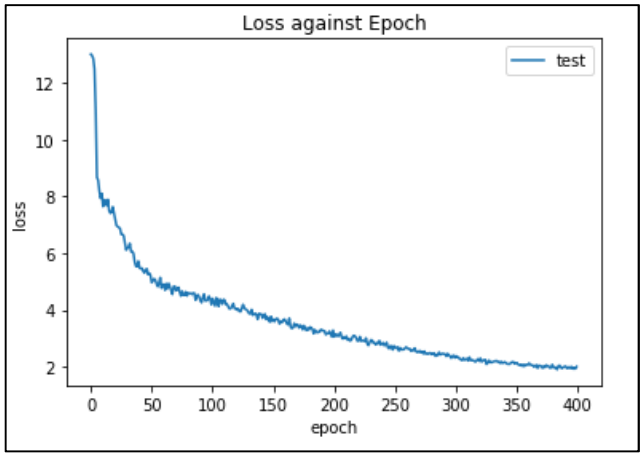

Fig. 8. Loss value using the test dataset

The loss value has considerably decreased compared to the ones during training dataset. The same result goes to RMSE value. Figure 8 also illustrates the decrease of RMSE value for the test dataset.

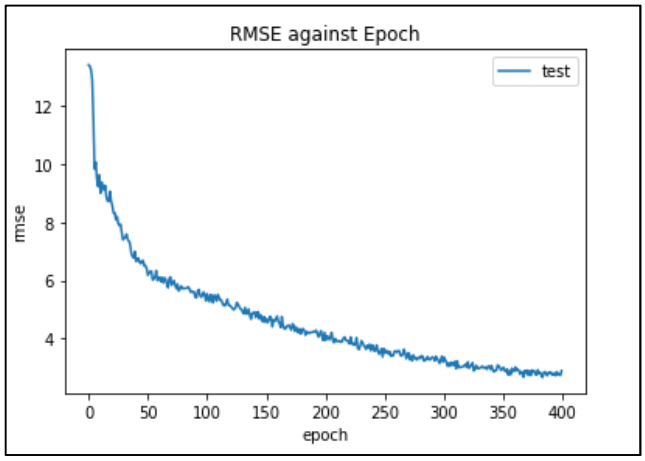

Fig. 9. RMSE value using the test dataset

Furthermore, each predicted value was compared to the actual value altogether. Fig. 9 below shows the predicted value of abalone age and the actual value of abalone age given from the test dataset.

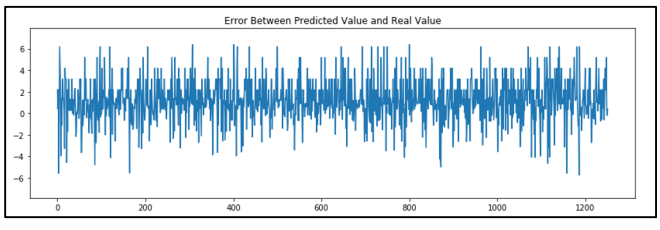

Fig. 10. Graph of the difference of predicted age and actual age for the test dataset

From Fig. 9, the difference between prediction value is also very close to the actual value as the highest difference value is 6 while Fig. 6 illustrates more than 10. This concludes that the model used is suitable for abalone dataset. Therefore, by only using physical measurements of an abalone, it is age

can be predicted computationally with reduced time consumption.

V. Conclusion

This study was conducted to develop a regression-based neural network to predict abalone age according to its physical measurements. The performance of the developed model was analysed and validated using RMSE performance measurement. We also compared the performance of other methods from previous studies. We chose three hidden layers with 400 epochs, 32 batch size, and 6 nodes in each hidden layer, as the best model with prediction result obtained 2.0691.

VI. Future Outlook

Artificial neural network is still evolving from time to time. In addition, new methods such as deep learning that can handle even larger data than abalone dataset is being rapidly tested by researchers. This study has done its role in contributing a way to help farmers or sellers to reduce their time on predicting abalone age. It has also gained an ability to predict a continuous value from a given dataset. Therefore, this study may well create a benchmark for future researchers to improve neural network efficiently. The model developed (regression-based neural network), should be improved, especially in finding and optimising the best parameter sizes so that the result in future might be better than the current result.

Acknowledgment

The authors would like to express gratitude to the reviewers and editors for helpful suggestions and Universiti Teknologi Malaysia for sponsoring this research by the GUP Research Grant (Grant Number: Q.J130000.2528.20H49) and FRGS Research Grant (Grant Number:R.J130000.7851.5F110). The research is also sponsored by the School of Computing, Faculty of Engineering UTM.

References

[1] Chen, Li, and John Ryan. “Abalone in Diasporic Chinese Culture: The Transformation of Biocultural Traditions through Engagement with the Western Australian Environment.” Heritage, vol. 1, no. 1, 2018, pp. 122–41, doi:10.3390/heritage1010009.

[2] Mayukh, Hiran. *Age of Abalones Using Physical Characteristics : A Classification Problem*. no. Ml, 2010, pp. 1–4Hossain, M. and Chowdhurry, N. M. (2019) ‘Econometrics Ways to Estimate the Age and Price of Abalone’, Munich Personal RePEC Archive, pp. 1–18.

[3] Md. Mobarak, Hossain, and Md Niaz Murshed Chowdhury. Econometric Ways to Estimate the Age and Price of Abalone. no. 7068, 2008..

[4] Dong, Jiaxu, et al. “Abalone Muscle Texture Evaluation and Prediction Based on TPA Experiment.” Journal of Food Quality, vol. 2017, 2017, doi:10.1155/2017/2069470.

[5] Officer, Rick Andrew. ISSN 1441-8487 Number 17 SIZE LIMITS AND YIELD FOR BLACKLIP. no. January 2003, 2014.

[6] Dua, Dheeru, and Casey Graff. "UCI Machine Learning Repository". *Archive.Ics.Uci.Edu*, 2017, https://archive.ics.uci.edu/ml/citation_policy.html.

[7] Ciaburro, Giuseppe, and Balaji Venkateswaran. *Neural Networks With R*. 1st ed., Packt Publishing Ltd., 2017, p. 314.

[8] Bonato, Vanderlei, et al. “Applied Reconfigurable Computing: 12th International Symposium, ARC 2016 Mangaratiba, RJ, Brazil, March 22-24, 2016 Proceedings.” Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 9625, no. April, 2016, doi:10.1007/978-3-319-30481-6.

[9] Montavon, Grégoire, et al. “Methods for Interpreting and Understanding Deep Neural Networks.” Digital Signal Processing: A Review Journal, vol. 73, Elsevier Inc., 2018, pp. 1–15, doi:10.1016/j.dsp.2017.10.011.

[10] Ramachandran, Prajit, et al. Searching for Activation Functions. 2017, pp. 1–13, http://arxiv.org/abs/1710.05941.

[11] Agostinelli, Forest, et al. *Learning Activation Functions to Improve Deep Neural Networks*. no. 2013, 2015, pp. 1–9.

[12] Smola, Alex, and S.V.N. Vishwanathan. *Introduction To Machine Learning*. 1st ed., The Press Syndicate Of The University Of Cambridge, 2008, p. 234.

[13] Hinton, Geoffrey. *Dropout : A Simple Way to Prevent Neural Networks from Overfitting*. Vol. 15, 2014, pp. 1929–58.

[14] Radhakrishnan, Pranoy. "What Are Hyperparameters ? And How To Tune The Hyperparameters In A Deep Neural Network?". *Medium*, 2017, https://towardsdatascience.com/what-are-hyperparameters-and-how-to-tune-the-hyperparameters-in-a-deep-neural-network-d0604917584a.

[15] LISA lab, University of Montreal. *Deep Learning Tutorial*. 2015.

[16] Pariyani, Manisha and Kavita Burse. “Age Prediction and Performance Comparison by Adaptive Network based Fuzzy Inference System using Subtractive Clustering.” .

[17] Embrechts, Mark J., Jonathan D. Linton, and Jorge M. Santos. "Random Brains: An ensemble method for feature selection with neural networks." *ESANN*. 2013.

[18] Duyck, James, et al. *Modified Dropout for Training Neural Network*. pp. 1–9.

[19] Dernoncourt, Franck. "Introduction to fuzzy logic." *Massachusetts Institute of Technology* 21 (2013).

[20] Reyes, “Fuzzy Logic: Introduction to Fuzzy Inference Systems,” 2018.