Optimizing Neural Network Efficiency by Neuron Pruning

Amany G. Fadhil*, Hana M. Ali, Zainab A. Khalaf

Department of Mathematics, College of Science, University of Basrah, Basra, Iraq

*Corresponding author E-mail: amany.gf84@gmail.com

Doi:10.29072/basjs.20230302

ARTICLE INFO	**ABSTRACT**
Keywords Optimizing Neural Network, shear wave sonic, dielectric anisotropy, pruning neurons.	This study primarily focuses on two main objectives. The primary objective is to develop an appropriate model for accurately predicting the missing shear wave data in the Volve Oil Field located in the North Sea. By employing the Multi-Layer Perceptron Regression and modifying the neural network structure, the first goal attains a prediction accuracy of 0.943 for missing S-Wave log data. Furthermore, the objective of the study is to enhance the precision of forecasting incomplete S-Wave log data by optimising the structure of the artificial neural network, using neuron pruning techniques based on sensitivity analysis. This optimisation leads to a heightened accuracy rate of 0.9609. The effectiveness of these pruning strategies is clearly evident in their demonstrated capacity to improve the accurate prediction of missing data in the sonic wave log.

Received 19 Nov. 2023; Received in revised form 16 Dec 2023; Accepted 22 Dec 2023, Published 31 Dec 2023

1. Introduction

The foundation of neural networks lies in their topology, a critical determinant of behavior and learning performance. Despite advancements, challenges persist in establishing optimal topologies for real-world applications, necessitating the exploration of factors such as the organization of networks, the number of neurons, hidden layers, and connections. In this pursuit, a Python code utilizing the Tensorflow library has been developed to efficiently determine the optimal number of hidden layers and neurons within a neural network [1]. Various algorithms exist for generating pruned networks, with Han et al. (2015) proposing a method involving initial training, convergence assessment, and subsequent trimming. This approach, recognized as a standard pruning algorithm, has seen refinements and adaptations by subsequent researchers, contributing to the field's complexity [2,3]. Shifting to practical applications, this paper employs open-source petrophysical logs from the Volve Oil Field (2008-2016) to predict synthetic shear wave sonic (S-Wave) logs. In a context where fossil fuel prices rise and the demand for clean energy surges, data-driven science emerges as a cost-effective decision-making tool for the fossil fuel industry. Artificial intelligence (AI), driven by data, finds extensive applications within the oil and gas sector [4]. S-Wave log data proves vital for understanding reservoir geomechanical properties, although obtaining accurate S-wave logs entails significant capital investment and is subject to wellbore conditions [5]. Recognizing the value of new information, AI modeling becomes essential for generating synthetic S-Wave logs. The adoption of dynamic pruning represents a substantial advancement in efficiently determining optimal network configurations [6]. Dynamic pruning, known for its capability to reduce calculations and enhance network performance, introduces adaptability into neural network science. Utilizing open-source petrophysical logs from the Volve Oil Field, this study illustrates the efficacy of AI in addressing challenges within the fossil fuel industry. In the midst of the intricacies of pruning methodologies, this work contributes to unraveling the complexities of the field, emphasizing the dynamic nature of advancements in neural network science. The primary objective of this paper is to determine optimal network configurations, considering factors such as the number of neurons, hidden layers, and connections. By doing so, the aim is to provide insights into the organization and structure of neural networks, facilitating the creation of more effective network configurations and optimizing the prediction of missing S-wave data. The research incorporates mathematical algorithms and techniques for dynamic pruning in neural networks, with a specific exploration of neuron pruning methods to enhance the prediction accuracy of missing data in shear wave sonic logs within the Volve oil field dataset [7]. Leveraging open-source petrophysical logs data, the study utilizes artificial intelligence methods, specifically the Multi-Layer Perceptron (MLP), for predicting synthetic shear wave sonic logs.

2. Proposed System

The proposed system is delineated through a systematic process encompassing five primary phases, as illustrated in Figure 1: Datasets phase, feature selection phase, optimization phase, pruning phase, and evaluation phase. The datasets phase assumes a pivotal role in preparing and organizing data for subsequent phases. Subsequently, the feature selection phase is implemented to extract relevant features. The selected features then proceed to the optimization phase, where optimal parameters and outcomes are determined. Moving forward, the prediction phase utilizes these optimized features to forecast missing values in the S-wave log data. Finally, the pruning phase is applied to refine the neural network model by strategically removing hidden neurons based on their sensitivity.

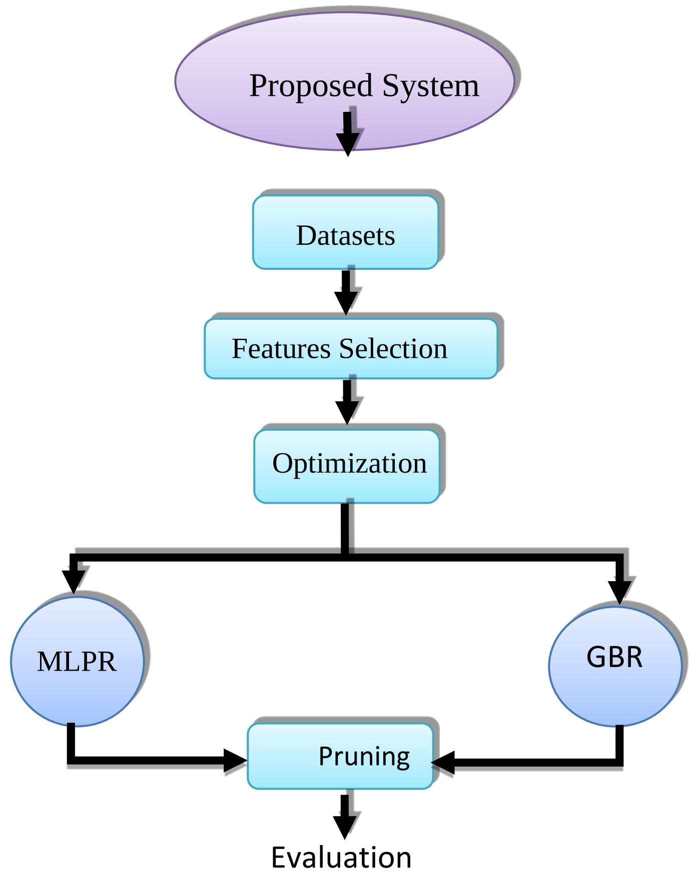


Figure 1: Flowchart of the proposed system

2.1 Dataset Phase

The Volve oil field, located approximately 200 kilometers west of Stavanger in the southern part of the North Sea, underwent extensive development with the drilling of 24 wells. This development resulted in a cumulative production of 63 million barrels of oil during its operational span from 2008 to 2016. In 2018, Equinor released a substantial amount of public data, including comprehensive well data and the geological model of the Volve field, to facilitate study, research, and development [8]. For the purposes of this paper, the well-log data was employed to predict the S-Wave log. Table 1 provides an overview of the dataset, specifically focusing on the five wells for which complete well logs are available.

Table 1: Well logs data summary

<table>
<tr><th>Wells</th><th>Rows Count</th><th>Logs</th></tr>
<tr><td>15/9-F-14</td><td>9822</td><td rowspan="5">MD (M)
TVD (M)
CALI (inches)
DRHO (g/cm3)
DT (us/ft)
GR (API)
NPHI (v/v)
RHOB (g/cm3)
RT (ohm.m)
PEF (b/elec)</td></tr>
<tr><td>15/9-F-1 B</td><td>3262</td></tr>
<tr><td>15/9-F-11 A</td><td>11374</td></tr>
<tr><td>15/9-F-11 T2</td><td>19433</td></tr>
<tr><td>15/9-F-1 A</td><td>10222</td></tr>
</table>

In total, the dataset comprised 21 wells, yielding a substantial well-log data set of 19,342,000 data points. Post data filtering, which involved excluding missing data and retaining wells with S-wave logs alongside essential well log sets, the final well count was reduced to 5. The resulting dataset for training and testing consisted of 481,370 data points. The primary objective of the model is to achieve high accuracy in predicting the missing S-wave data, totaling 2161 data points, within these five trained wells. The well logs retained after the data cleaning process included Measured Depth (MD), True Vertical Depth (TVD), Neutron Porosity (NPHI), Bulk Density (RHOB), Density correction (DRHO), Gamma Ray (GR), Resistivity (RT), Photoelectric factor (PEF), Caliper (CALI), P-Wave (DT), and Shear Wave (DTS). Figure (2) displays the five wells involved in constructing the S-wave predictive model, namely 15/9-F-1 A, 15/9-F-11 A, 15/9-F-14, Well 15/9-F-11 T2, and 15/9-F-1 B. Additionally, Figure 2, track 1, illustrates the GR, representing the total formation radioactivity content measured in API (American Petroleum Institute) units. Track 2 presents RT, indicating the true resistivity of the formation, measured in ohms per meter. Track 3 contains RHOB and NPHI, reflecting the formation's bulk density (measured in grams per cubic centimeters) and porosity (measured in v/v), respectively. Track 4 incorporates DRHO, representing the correction factor for RHOB, measured in grams per cubic centimeters. Lastly, track 5 includes DT and DTS, representing the P-Wave log (measured in microseconds per foot) and the S-Wave log travel time, respectively.

2.2 Feature Selection

To gain insights into the trends of input data and streamline information, the research conducted descriptive analytics. Utilizing hierarchical clustering analysis, relationships between input and output data were approximated by grouping them based on similarity measures [9], [10]. In this analysis, Euclidean distance [11] served as the similarity metric to measure dissimilarity between sets of data, as expressed by the formula:

$$\text{Euclidean Distance} = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (1)$$

The analysis revealed that DT (P-Wave log) and NPHI exhibited greater similarities with DTS (S-Wave log). Subsequently, GR demonstrated the next level of similarity, while RT and RHOB shared minor similarities with DTS. In addition to hierarchical clustering, another approach involved the utilization of the Random Forest (RF) Feature Importance algorithm, specifically the mean decrease impurity (MDI) [5]. RF feature importance was calculated using MDI and permutation importance (PIMP) techniques [6]. The results indicated that NPHI and DT had the most substantial influence, while RT had the least influence on the model.

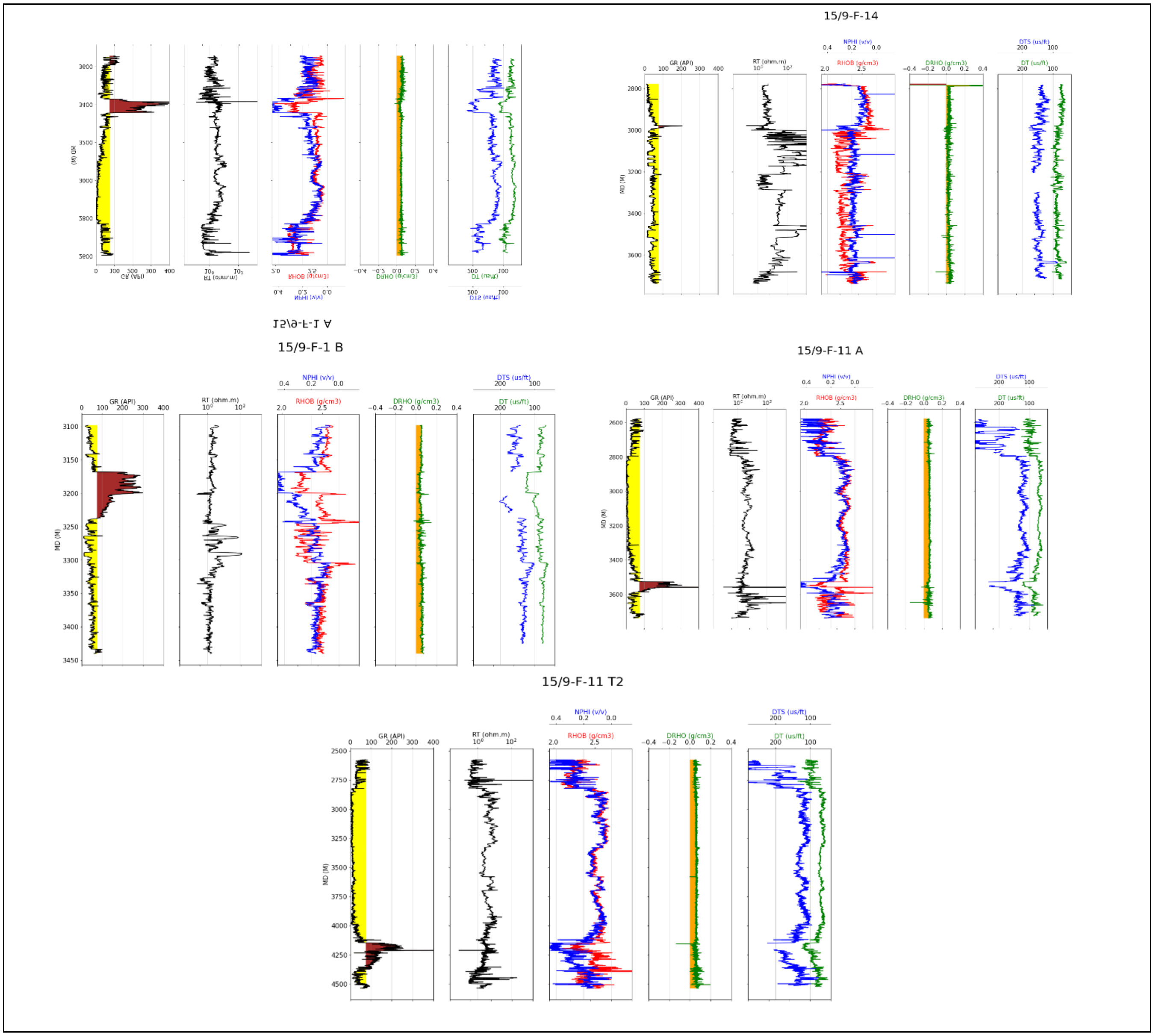


Figure 2: Well logs for five wells used in building the predictive model

2.3 Optimization Phase

This section focuses on the identification of the optimal neural network topology by determining the most suitable number of layers. Selecting the appropriate topology for artificial neural networks is crucial for tailoring them to specific tasks and achieving optimal results. This objective is accomplished by tuning the parameters to discover the optimal topology. TensorFlow/Keras served as the platform to train an Artificial Neural Network (ANN). The process involved configuring the neural network model, training it with early stopping, and preserving the input layer weights for each epoch. A duplicate model, limited to 100 iterations, was created to identify the optimal number of hidden layers based on evaluation metrics. Moreover, heat mapping is used for visualization and analysis which serves as a valuable technique for a comprehensive assessment of the architecture and connections within an artificial neural network (ANN). It facilitates a deeper understanding of the network's structure and allows for the visualization of network traffic flow. In heat maps, colors or intensity values are assigned to linkages based on weights or activation levels, providing a means to identify critical pathways within the network. This visualization method enhances the interpretability of the neural network's internal dynamics.

2.4 Pruning Phase

The optimization of artificial neural networks (ANN) involves a sensitivity-based pruning technique aimed at determining the optimal number of neurons, layers, and connections. The primary aim is to prune and eliminate extraneous neurons, thus enhancing the overall efficiency and accuracy of the neural network. The pruning methodology implemented involves the removal of neurons based on a predetermined percentage, determined through the evaluation of their sensitivity ($sensitivity_{neurons}$). The strategic rationale is to eliminate neurons with the lowest sensitivity, minimizing their impact on network performance or output. This meticulous pruning process results in a more streamlined network, reducing complexity, computational demands, and processing time. However, it is imperative to exercise caution in pruning, as excessive removal of neurons can adversely affect the network's accuracy and overall performance. Thus, achieving a delicate balance is crucial during the optimization process to prevent any loss of critical information. This paper specifically employs sensitivity-based pruning for neuron removal, ensuring that the pruned network maintains accuracy by selectively removing neurons with minimal impact on network performance. To systematically evaluate the impact of removing each neuron on the network's performance, accuracy metrics are rigorously compared. The overarching goal of this optimization approach is to strike a delicate balance between the model's size and performance, thereby potentially enhancing its overall efficiency. Neurons with the lowest levels of sensitivity emerge as candidates for pruning, as they are considered less critical for the current model. Algorithm (1) shows the steps of sensitivity-based pruning Algorithm.

Algorithm (1): Sensitivity-based Pruning Algorithm	
Inputs	***N***: Set of neurons in the neural network ***Neuron*** $_{Input}$ ***(i)***: The input to neuron i. ***p***: Pruning percentage
Outputs: Pruned Network	The neural network after sensitivity-based pruning.
Step 1: Calculate Sensitivity	$Selectivity\ (i) = \lvert \frac{\partial\ Output}{\partial\ Neuron_{Input}(i)} \rvert$
Step 2: Rank Neurons by Sensitivity	$Ranked_{Neuron} = Sort\ (N, Sensitivities, Desending = True)$
Step 3: Determine Neurons to Prune	$Neuron_{Prune} = \lfloor p\ \times \lvert N \rvert \rfloor$
Step 4: Keep (1-p) Neurons to prune	$Neuron_{Keep} = Ranked_{Neuron}[Neuron_{Prune}:]$
Step 5: Pruned Neurons	$Purned_{Network} = PruneNetwork(Original_{Network}, Neuron_{Keep})$

2.5 Evaluation Phase

The performance of the proposed system is systematically assessed. This phase involves applying the trained artificial neural network (ANN) to predict missing S-wave sonic log data based on the known wave sonic log values. The R-squared score is employed to quantify the accuracy and effectiveness of the predictions. **R-squared (R^2)** serves as an indicator of how well the model's predictions align with the actual observed values. A higher R-squared value implies that a larger percentage of the variability in the dependent variable is accounted for by the independent variables, indicating a better fit.

$$R^2 = 1 - \frac{SS_{residual}}{SS_{Total}} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (2)$$

Where:

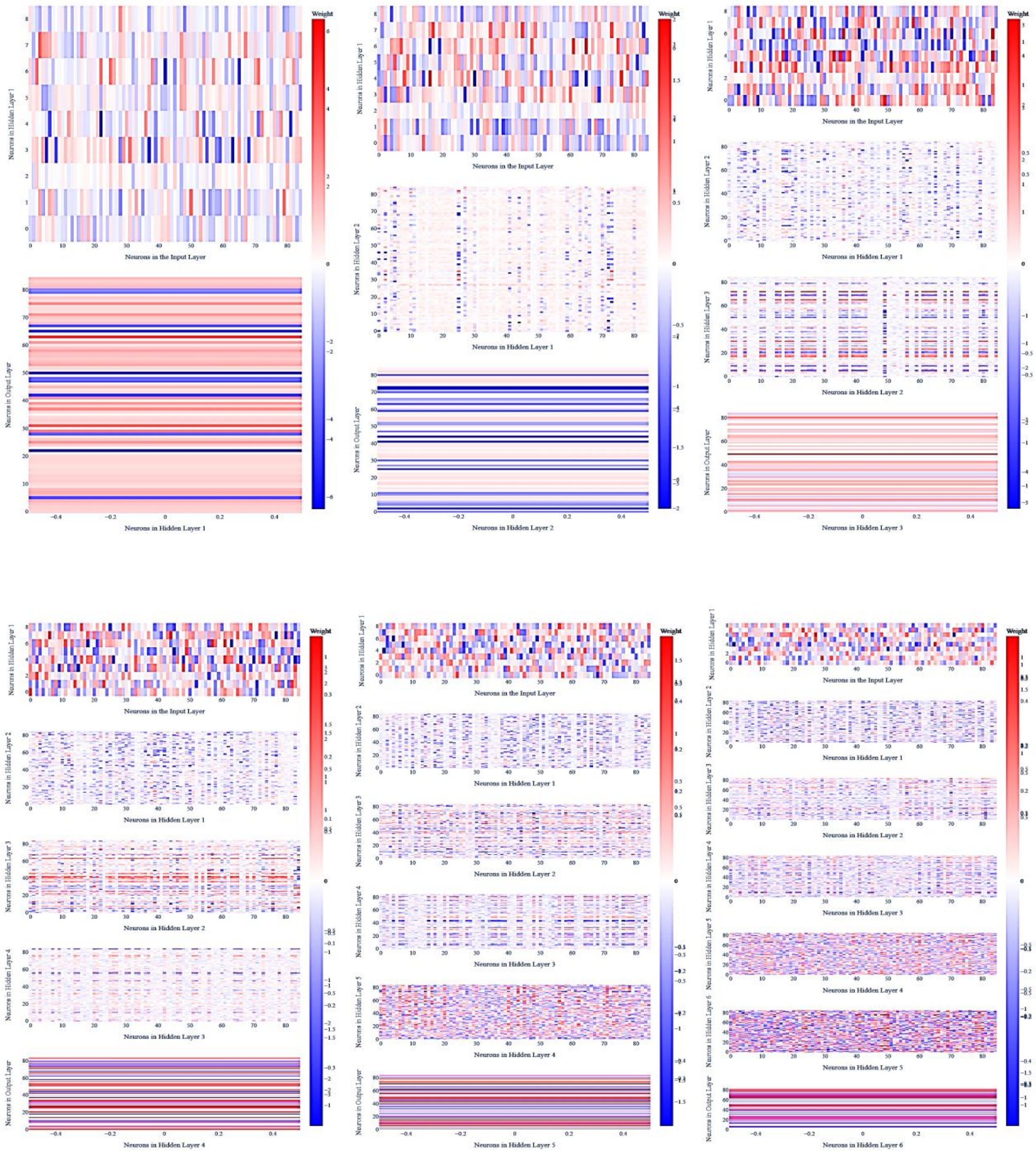
$SS_{residual}$: is the sum of squared residuals (the differences between observed and predicted values), SS_{Total}: is the total sum of squares (the differences between observed values and the mean of the observed values).

3. Results and discussion

The initial step involved the identification of the experimental setup. All algorithms executed in Python 3.11 within the ANACONDA environment. Notably, the scikit-learn library (sklearn) assumed a crucial role in the manipulation and analysis of data. TensorFlow, a prominent deep learning framework, was instrumental in the iterative training of the model, with an additional function of storing information on hidden input layer neurons in dedicated Excel sheets. Furthermore, for visualization and plotting requirements, Matplotlib was employed.

3.1 Optimization phase:

To determine the optimal configuration for hidden neurons and layers, a systematic approach was adopted. An iterative loop was executed, testing eight hidden layers, each with varying combinations of hidden neurons: 35, 45, 55, 65, 75, 85, 95, and 100. These layers were progressively added up to a total of eight, utilizing the 'Adam' solver and 'RELU' activation functions. Incorporating an early stopping mechanism with a 10% validation dataset prevented overfitting, halting the training process when necessary. The maximum number of iterations was capped at 1000. Results revealed that the most effective model featured 7 hidden layers with 85 hidden neurons. TensorFlow/Keras were employed to train an Artificial Neural Network (ANN), using R^2 metric function. Each layer was incrementally added, and the model was trained separately for each configuration to assess performance. The chosen model, with five hidden layers and 85 hidden neurons, yielded R^2 values of 0.9718 for training data, 0.9541 for validation data, and 0.9511 for blind data. Deviating from this configuration resulted in lower R^2 values, impacting prediction performance. Consequently, the optimal configuration was determined as five hidden layers with 85 hidden neurons. A predictive system combining Multilayer Perceptron Regression (MLPR) and Gradient Boosting Regression (GBR) was created to predict missing wave sonic logs. The highest R^2 scores of 0.96 for training and 0.94 for testing were achieved. Additionally, GBR yielded R^2 scores of 0.99 for training and 0.98 for testing. Figure 3, employing heat maps, visually represents the influence of hidden neurons and the strength of connections across the seven hidden layers, providing insights into the network's dynamics.


Weight
Neurons in the Input Layer
Neurons in Hidden Layer 1
Neurons in Hidden Layer 2
Neurons in Hidden Layer 3
Neurons in Hidden Layer 4
Neurons in Hidden Layer 5
Neurons in Hidden Layer 6
Neurons in Output Layer

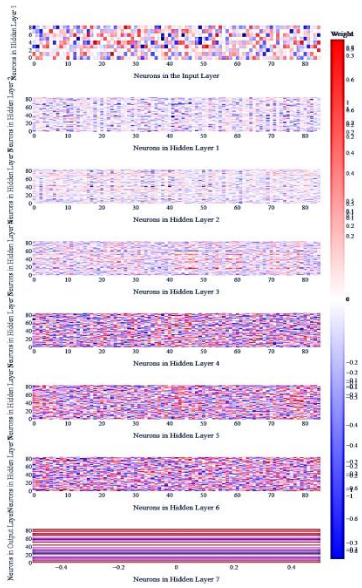


Figure 3: Heatmap of the model for each hidden layer

Figure 4 visually displays the alignment between the model's predictions and the actual values across each dataset, including the training, validation, and blind sets for the original model. Each dataset is presented as a scatter plot, offering a clear visual representation to facilitate the assessment of the model's performance. R-squared (R^2) values are employed to evaluate this performance for varying numbers of hidden layers. Within the figure, the training data is represented by green dots, the validation data by blue dots, and the blind testing data by red dots.

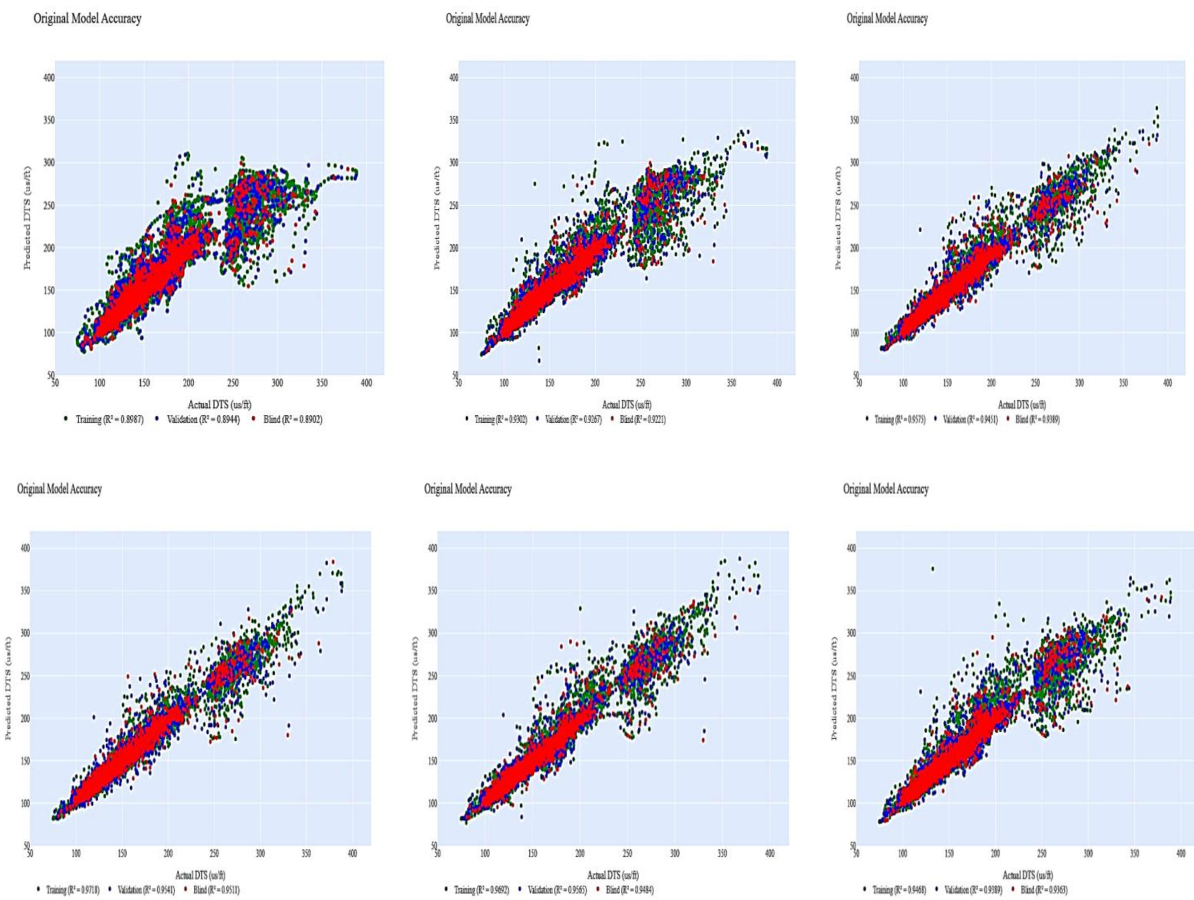


Figure 4: A scatter plot for evaluating the accuracy of the model

3.2 Pruning phase

In this section, the outcomes of the sensitivity-based method for pruning hidden neurons are presented and discussed. During this process, 8 hidden neurons were pruned from each of the hidden layers and their indices are detailed in Table 2, except for the fifth layer, which had no hidden neurons removed. This aids in comprehending the decision-making process of the network and identifying the crucial factors influencing its predictions. The pruning process was thoroughly evaluated and validated, with the findings demonstrating its success and accuracy (R^2) in prediction achieves an accuracy rate (R^2) of 0.9609. Table 3 illustrates the performance results of the pruned model after pruning each hidden layer separately.

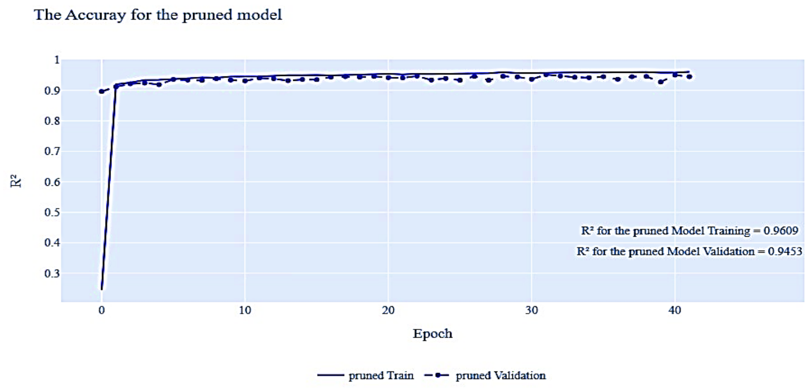


Figure 5: The accuracy of the pruned neurons model (with 5 hidden layers) based on the Sensitivity.

Table 2: The pruned neuron indices for the model with all five hidden layers

Indices. of hidden layers pruning	Original performance	Indices of pruned neurons
1	55.581905364990234	[20 28 40 21 18 5 36 1]
2	1810.6697998046875	[50 82 1 71 66 23 53 78]
3	4648.8779296875	[15 21 9 33 57 12 19 60]
4	10645.498046875	[1 50 12 9 71 69 76 0]
5	16376.8232421875	[ ]

Table 3: Pruned model's performance results after pruning for each hidden layer separately.

Indices of hidden layers	Epoch	Loss function	Val. Loss function	RMS error	Val. RMS error	R^2	Val. R^2
1	69	40.8344	58.5449	6.3902	7.6515	0.9597	0.9477
2	42	46.6496	63.7745	6.8300	7.9859	0.9548	0.9446
3	31	52.5561	71.9773	7.2496	8.4839	0.9494	0.9331
4	42	47.9547	65.7928	6.9249	8.1113	0.9533	0.9418
5	42	45.8783	66.2329	6.7734	8.1384	0.9553	0.9406

4. Conclusions

The study aimed to optimize a neural network model for predicting missing sound log data in the Volve field, utilizing various techniques across different phases. To create an optimal neural network model, two methods (MLP, and GBR) were employed to determine the best configuration of hidden layers and neurons. The first method suggested seven hidden layers with 85 neurons each, while the second method recommended five hidden layers, also with 85 neurons. The second method outperformed the first, showing R^2 of 0.9618 for training and 0.9515 for validation. While the second method achieved R^2 for training reached 0.9620, and for validation, it achieved 0.9516. Hence, the most effective model featured 5 hidden layers, each housing 85 neurons. After prediction, pruning was performed for the original model's hidden neurons and each hidden layer independently. Pruning all layers together proved more accurate, with 0.9606 accuracy for training and 0.9453 for testing. Pruning methods contributed to reducing network complexity, increasing efficiency, and shortening training periods. The proposed system successfully optimized the neural network model, providing valuable insights for accurate predictions in the oil field domain. The combination of the effective model configuration, feature importance analysis, and pruning techniques contributed to improved accuracy and efficiency in predicting missing sound log data. With lower memory consumption during both training and validation.

References

[1] L. M. Alhelfi and H. M. Ali, Using Persistence Barcode to Show the Impact of Data Complexity on the Neural Network Architecture, Iraqi J. Sci., 63 (2022)2262–2278, https://doi1 0.24996/ijs.2022.63.5.37

[2] T. Lin, L. Barba, and M. Jaggi, Dynamic Model Pruning with Feedback, ICLR. 12 (2020) 1-22, https://doi.org/10.48550/arXiv.2006.07253

[3] T. Liang, J. Glossner, L. Wang, S. Shi, and X. Zhang, Pruning and quantization for deep neural network acceleration: A survey, Neurocomputing.,461(2021)370–403, https://doi.org/10.1016/j.neucom.2021.07.045

[4] E. Grossi and M. Buscema, Introduction to artificial neural networks, Eur. J. Gastroenterol. Hepatol., 19(2007) 1046–1054, https://doi.org/10.1097/meg.0b013e3282f198a0

[5] S. Maleki, A. Moradzadeh, R. G. Riabi, R. Gholami, and F. Sadeghzadeh, Prediction of shear wave velocity using empirical correlations and artificial intelligence methods, NRIAG J Astr Geophys, 3 (2014) 70–81, https://doi.org/10.1016/j.nrjag.2014.05.001

[6] R. Reed, Pruning Algorithms A Survey, IEEE Transactions on Neural Networks., 4 (1993) 740–747, doi: https://doi.org/10.1109/72.248452

[7] Y. Wang, X. Zhang, X. Hu, B. Zhang, and H. Su, Dynamic network pruning with interpretable layerwise channel selection, AAAI Conf. Artif. Intell., 34(2020) 6299–6306, https://doi.org/10.1609/aaai.v34i04.6098

[8] Y. M. Chiang, L. C. Chang, and F. J. Chang, Comparison of static-feedforward and dynamic-feedback neural networks for rainfall-runoff modeling, J. Hydrol., 290(2004) 297–311, https://doi:+10.1016/j.jhydrol.2003.12.033

[9] S. C. Johnson, Hierarchical clustering schemes, Psychometrika, 32 (1967) 241–254, https://doi.org/10.1007/BF02289588

[10] Q. Liu, S. Du, B. Jacobus, V. Wyk, and Sun Yanxia, Niching particle swarm optimization based on Euclidean distance and hierarchical clustering for multimodal optimization, Springer Nat. B.V. Nonlinear Dyn., 99(2019) 2459–2477, doi.org/10.1007/s11071-019-05414-7

[11] F. Murtagh, P. Contreras, Algorithms for hierarchical clustering: an overview, Data Min. Knowl. Discov., 2(2012) 86–97, https://doi.org/10.1002/widm.53

تحسين كفاءة الشبكة العصبية عن طريق تشذيب الخلايا العصبية

اماني جمال فاضل[1]*, هناء مرتضى علي[2], زينب علي خلف[3]

قسم الرياضيات، كلية العلوم، جامعة البصرة

المستخلص

تركز هذه الورقة بشكل أساسي على هدفين رئيسيين. أولاً، يهدف هذا البحث إلى صياغة نموذج أمثل للتنبؤ ببيانات موجة القص المفقودة في حقل فولف للنفط، بحر الشمال. باستخدام انحدار Perceptron متعدد الطبقات وإعادة تكوين طوبولوجيا الشبكة العصبية، يحقق الهدف الأول معدل دقة قدره 0.943 في التنبؤ ببيانات سجل S-Wave المفقودة. ثانيًا، تسعى الدراسة إلى تحسين دقة التنبؤ ببيانات سجل S-Wave المفقودة عن طريق ضبط طوبولوجيا الشبكة العصبية الاصطناعية، باستخدام أساليب تقليم الخلايا العصبية القائمة على الحساسية. يؤدي هذا التحسين إلى معدل دقة مرتفع يبلغ 0.9609. تتجلى الفعالية الكبيرة لتقنيات التقليم هذه في قدرتها المثبتة على تعزيز التنبؤ الفعال بالبيانات المفقودة في سجل الموجات الصوتية.