Evaluating the Impact of Neural Network Layers on Performance in Classification and Regression Tasks

Seyedmojtaba Mohasel MOJTABA.MOHASEL@YAHOO.COM
Department of Mechanical and Industrial Engineering
Montana State University
Montana, USA

Nitasha Fazal NITASHA.FAZAL@MONTANA.EDU
Gianforte School of Computing
Montana State University
Montana, USA

Behzad Karimi BEHKARIMI@GMAIL.COM
Department of Electrical and Computer Engineering
Montana State University
Montana, USA

Editor: -

Abstract

In this project, we implemented and evaluated several neural network training algorithms for classification and regression tasks on six datasets (three for classification and three for regression) from the UCI Machine Learning Repository. Using fully-connected feedforward architectures, We assess the performance of different network configurations, including varying numbers of layers and the use of varying number of nodes in each layer, to enhance network learning and generalization. Our experiments provide insights into the effectiveness of neural network architectures for various types and sizes of data. The findings highlights the importance of network complexity: while adding more layers improves performance (measured by MSE for regression and Macro averaged F1 score for classification) on some datasets, this effect is not consistent across all datasets

Keywords: Neural Networks, Regression, Classification, Convergence

1 Introduction

Artificial Neural Networks have become foundational models in machine learning, with applications spanning classification, regression, and beyond. Densely-connected feedforward neural networks, in particular, offer a flexible and effective approach for various supervised learning tasks due to their fully connected structure and adaptability in learning complex data patterns. This project explores the implementation of feedforward neural networks on six diverse datasets from the UCI Machine Learning Repository, covering both classification and regression tasks. Our focus for this project is on constructing networks with multiple hidden layers, incorporating non-linear activation functions in these layers to enhance model expressiveness. For the output layer in classification networks, we apply softmax activation for multi-class datasets and sigmoid activation for binary classification tasks. In regression

networks, a linear activation function is applied at the output layer. Additionally, we examine the use of bias nodes across layers, which are known to improve model flexibility and support generalization.

2 Problem Statement and Hypotheses

The problem in this project consists of several components: developing a Neural network classifier and regressor, training the classifier on three datasets, and reporting its performance using F1-score, precision, recall, and F1 Score (considering the macro average for all metrics). The regressor is trained on three datasets and its performance is reported using Mean Absolute Error and Mean Squared Error (MSE). The project also addresses issues such as handling missing values, converting categorical attributes into dummy variables, and using stratified 10-fold cross-validation for optimization and hypothesis testing in both classification and regression tasks.

2.1 Hypotheses

In this section, we present our hypothesis. Our rationale for the hypothesis is based on the size of the dataset (number of instances in Table 1). We hypothesize that for datasets with relatively large sample sizes (Breast cancer,Abalone, and Forest Fires), the addition of layers leads to improvement in performance and a faster convergence rate. However, in other datasets (glass, computer hardware (machine), or Soybean) addition of layers leads to decreased performance and a slower convergence rate.

Our first hypothesis is that adding more layers and nodes will achieve higher F1-score in classification tasks and lower mean squared error (MSE) in regression tasks compared to using fewer layers.

Null Hypothesis (H0): Adding more layers and nodes will not achieve a higher F1-score in classification tasks nor a lower MSE in regression tasks compared to using fewer layers.

Alternative Hypothesis (H1): Adding more layers and nodes will achieve a higher F1-score in classification tasks and a lower MSE in regression tasks compared to using fewer layers.

We will use stratified 10-fold cross-validation with paired t-test to test our hypotheses.

We present our second hypothesis as follows:

Null Hypothesis (H0): Adding more layers and nodes will not result in a slower convergence rate.

Alternative Hypothesis (H1): Adding more layers and nodes will result in a slower convergence rate.

We will use 95 percent confidence interval and compare the convergence of models to evaluate our second hypothesis.

3 Experimental Approach and Program Design

3.1 Experimental Approach

This work uses six datasets from the UCI Machine Learning Repository. A brief description of the datasets is given in Table 1. Class imbalance occurs when one class significantly

Data set	# Features	# Instances	Missing Values	Continuous Data	Imbalance	Curse of Dimensionality Risk	Size
Abalone	8	4177	No	Yes	Moderate	Low	Large
Breast Cancer	10	699	Yes	Yes	Moderate	Low	Large
Glass	9	214	No	Yes	High	Moderate	Moderate
Computer Hardware	10	209	No	Yes	Moderate	Moderate	Small
Soybean (small)	35	47	Yes	No	Moderate	High	Small
Forest Fires	12	517	No	Yes	Moderate	Low	Large

Table 1: Dataset Characteristics

outnumbers others in a dataset. This imbalance can lead to biased model performance, where the algorithm tends to favor the majority class, often misclassifying instances from the minority class. To categorize the risk of the curse of dimensionality, we considered both the number of features and training samples. The Soybean dataset, for example, had 35 features but only 46 instances, which we expected to be prone to the curse of dimensionality. We also evaluated the size of the dataset in terms of the number of training samples.

3.2 Program design

The dataset undergoes preprocessing by the implementation of specific classes for each dataset (e.g., BreastCancer and Glass). These classes handle tasks such as dropping irrelevant columns, filling missing values with the mode, and generating dummy variables for categorical attributes. Preprocessing was also performed for numerical values using min-max normalization.

We implemented stratified 10-fold cross-validation separately for both regression and classification tasks. We developed three models for each dataset. Model 1 is a neural network with no hidden layer. Model 2 has 1 hidden layer and the number of nodes was optimized with other hyperparameters using grid search. Model 3 has 2 hidden layers with an optimized number of nodes by a grid search. Stratified 10-fold cross-validation was used once for tuning the model's hyperparameters and then again for statistical testing. The hyperparameters for tuning were the number of nodes in each layer, the activation function (between logistic or hyperbolic tangent), the learning rate, the momentum term, and the batch size.

After tuning the model architecture, we performed stratified 10-fold cross-validation for each model, which provided us with 10 test F1-scores for each. In our experimental module, we first ensured that the 10 test scores followed a normal distribution, as this is a prerequisite for the t-test. Then, we applied a pairwise t-test to test our hypothesis. If the normality assumption was violated, we used the Wilcoxon test, a non-parametric statistical test.

4 Results

In this section, first we present the results obtained for our first hypothesis for both classification task and the regression. Then we provide the results for the second hypothesis.

In Hypothesis 1, we investigated whether adding more layers to neural network models improves the F1 score. Results are summarized in Table 2. Model 1, Model 2, and Model 3

Table 2: Statistical Comparison of Models Across Different Datasets

Comparison	Dataset	t_stat	p_value	Significant
Model 1 vs Model 2	Breast Cancer	-1.355	1.920000e-01	False
Model 1 vs Model 3	Breast Cancer	-1.355	1.920000e-01	False
Model 2 vs Model 3	Breast Cancer	0.000	1.000000e+00	False
Model 1 vs Model 2	Glass	0.314	7.570000e-01	False
Model 1 vs Model 3	Glass	10.358	5.180000e-09	True
Model 2 vs Model 3	Glass	12.024	4.890000e-10	True
Model 1 vs Model 2	Soybean	1.322	2.030000e-01	False
Model 1 vs Model 3	Soybean	4.238	5.000000e-04	True
Model 2 vs Model 3	Soybean	2.070	5.300000e-02	False
Model 1 vs Model 2	Abalone	0.106	9.170000e-01	False
Model 1 vs Model 3	Abalone	-0.246	8.090000e-01	False
Model 2 vs Model 3	Abalone	-0.368	7.170000e-01	False
Model 1 vs Model 2	Machine	-0.753	4.610000e-01	False
Model 1 vs Model 3	Machine	-0.469	6.450000e-01	False
Model 2 vs Model 3	Machine	0.341	7.370000e-01	False
Model 1 vs Model 2	Forest Fire	10.789	4.610000e-02	True
Model 1 vs Model 3	Forest Fire	11.134	3.310000e-02	True
Model 2 vs Model 3	Forest Fire	0.876	3.810000e-01	False

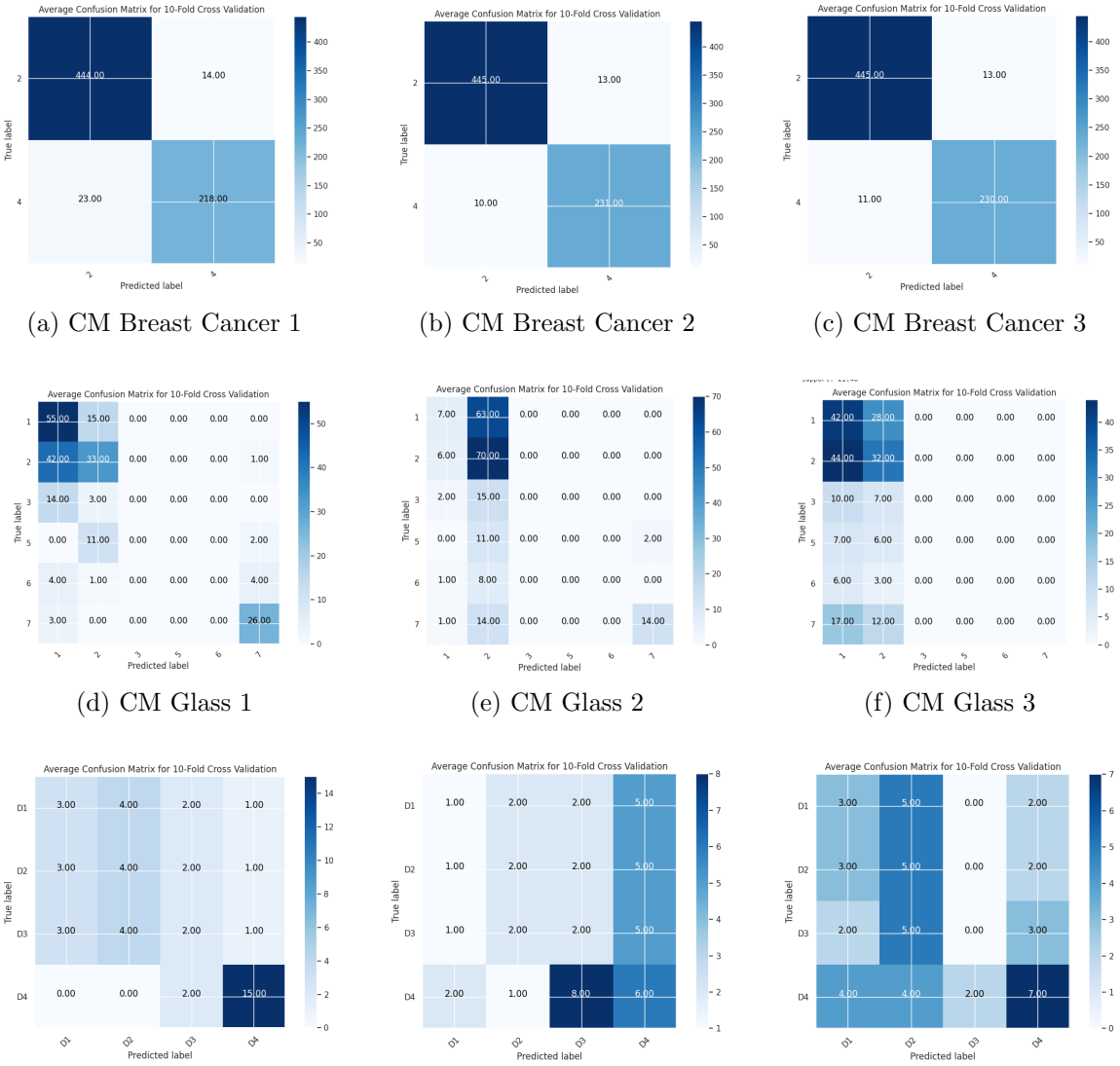
represent neural networks with 0, 1, and 2 layers respectively. The models were compared across multiple datasets: Breast Cancer, Glass, Soybean, Abalone, Machine, and Forest Fires. Paired t-tests were conducted to evaluate the significance of performance differences between models, using a significance level of $\alpha = 0.05$. Details of the results for each dataset are explained as follows:

4.1 Breast Cancer Dataset

For the Breast Cancer dataset, no significant difference in F1 Score was found between any of the models(Model 1 vs Model 2, Model 1 vs Model 3, and Model 2 vs Model 3 in Table 2). This suggests that increasing the number of layers from 0 to 2 did not have a statistically significant impact on model performance for this dataset. The confusion matrices (Figure 1) reveal that all models achieved similar classification results, with minor variations in false positives and false negatives. The overall classification F1 Score remained consistent across the models, suggesting that increasing the number of layers did not significantly alter the model's capability to differentiate between classes (Table 3).

Model	**Accuracy**	**Precision**	**Recall**	**F1-Score**
Breast Cancer Model 1	0.95	0.95	0.94	0.94
Breast Cancer Model 2	0.97	0.96	0.97	0.96
Breast Cancer Model 3	0.97	0.96	0.96	0.96

Table 3: Performance metrics (Macro average) of Breast Cancer Models.



(a) CM Breast Cancer 1 (b) CM Breast Cancer 2 (c) CM Breast Cancer 3

(d) CM Glass 1 (e) CM Glass 2 (f) CM Glass 3

(g) CM Soybean 1 (h) CM Soybean 2 (i) CM Soybean 3

Figure 1: Confusion Matrices for Different Models Across Various Datasets

4.2 Glass Dataset

In the case of the Glass dataset, comparisons between Model 1 vs Model 3 and Model 2 vs Model 3 yielded significant differences (p-values of 5.18×10^{-9} and 4.89×10^{-10}, respectively in Table 2). This indicates that adding more layers to the neural network significantly degraded performance (Table 4). However, no significant difference was found between Model 1 and Model 2 (p-value = 0.757 in Table 2), suggesting that the decrease in F1 Score is most pronounced when moving from 0 or 1 layer to 2 layers. The confusion matrices (Fig. 1) illustrate that all models performed poorly in classifying samples in classes 3 to 6.

Model	Accuracy	Precision	Recall	F1-Score
Glass Model 1	0.58	0.36	0.41	0.37
Glass Model 2	0.52	0.29	0.35	0.30
Glass Model 3	0.35	0.12	0.17	0.14

Table 4: Performance metrics (Macro average) of Glass Models.

4.3 Soybean Dataset

For the Soybean dataset, a significant decrease in the F1 Score was observed when comparing Model 1 to Model 2. F1 score was improved when having 2 layers in comparison to 1 layer. The best result was obtained with 1 layer, indicating that the addition of layers contributed to decreased performance. The confusion matrices (Figure 1) show that Model 2 performed poorly in classifying samples belonging to D3.

Model	Accuracy	Precision	Recall	F1-Score
Soybean Model 1	0.51	0.28	0.45	0.32
Soybean Model 2	0.23	0.10	0.23	0.12
Soybean Model 3	0.32	0.14	0.33	0.18

Table 5: Performance metrics (Macro average) of Soybean Models.

4.4 Abalone Dataset

For the Abalone dataset, none of the pairwise comparisons yielded statistically significant results (Table 2). This suggests that adding more layers did not result in a meaningful increase in the F1 Score for this dataset.

Metric	Mean Absolute Error (MAE)	Mean Squared Error (MSE)	R-squared (R2)
Model 1	1.5320647395133158	4.533684265214422	0.5637658583680532
Model 2	1.4998660493405085	4.400574300511819	0.5765737884747504
Model 3	1.5021145870039387	4.445008995654316	0.5722982522970568

Table 6: Regression Metrics for Abalone Dataset

4.5 Machine Dataset

Similarly, for the Machine dataset, none of the pairwise comparisons showed statistically significant results, with all p-values greater than 0.05 (Table 2). This suggests that increasing the number of layers did not significantly improve performance.

4.6 Forest Fires Dataset

Results of the forest fires dataset indicate that adding layers to the network resulted in improved performance Table 8. This indicated that our hypothesis is supported.

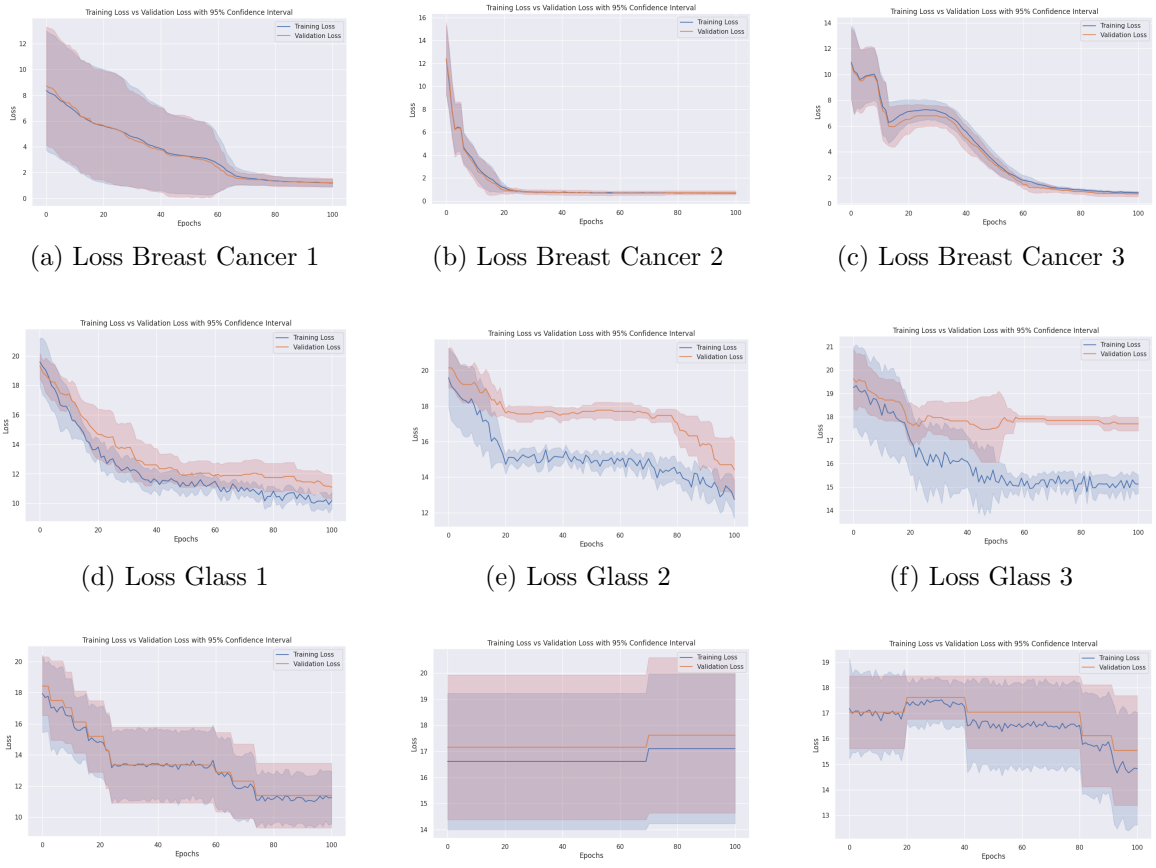
Metric	Mean Absolute Error (MAE)	Mean Squared Error (MSE)	R-squared (R2)
Model 1	37.42400595176949	4321.286585865267	0.8321358569978076
Model 2	40.71974958964391	6982.953900101653	0.7287410707962407
Model 3	40.78015886548629	6358.6601728412	0.7529923046419673

Table 7: Regression Metrics for Machine Dataset

Metric	Mean Absolute Error (MAE)	Mean Squared Error (MSE)	R-squared (R2)
Model 1	35.672	4053.48	0.8413
Model 2	30.281	3500.92	0.8651
Model 3	29.009	3400.67	0.8695

Table 8: Regression Metrics for Forest Fire Dataset

Hypothesis 2

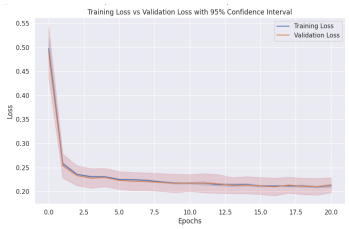


(a) Loss Breast Cancer 1 (b) Loss Breast Cancer 2 (c) Loss Breast Cancer 3

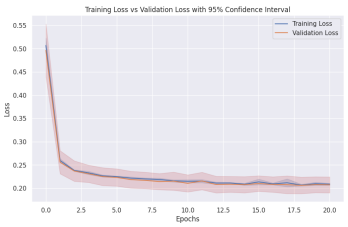
(d) Loss Glass 1 (e) Loss Glass 2 (f) Loss Glass 3

(g) Loss Soybean 1 (h) Loss Soybean 2 (i) Loss Soybean 3

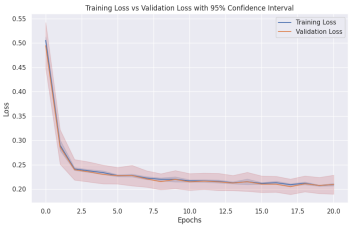
Figure 2: Loss Plots for Different Models Across Various Datasets



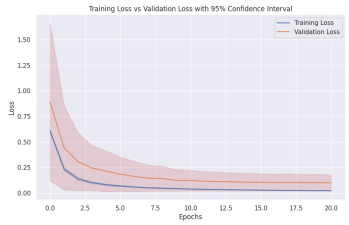
(a) Abalone Loss 1



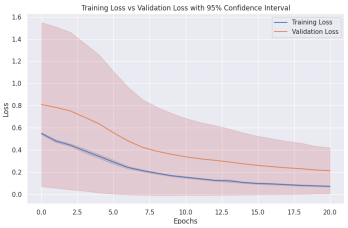
(b) Abalone Loss 2



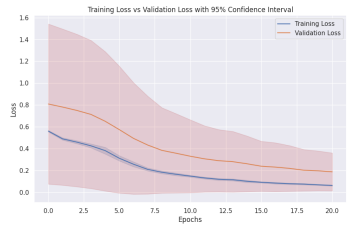
(c) Abalone Loss 3



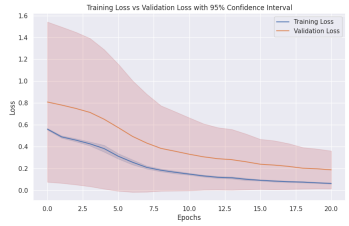
(d) Machine Loss 1



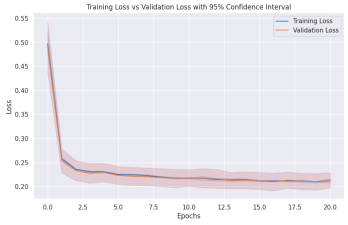
(e) Machine Loss 2



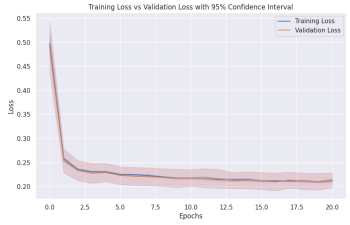
(f) Machine Loss 3



(g) Forest Fires Loss 1



(h) Forest Fires Loss 2



(i) Forest Fires Loss 3

Figure 3: Loss Plots for Abalone, Machine, and Forest Fires Models

We hypothesized that with the increase in layers, the model converges faster for for Abalone and Breast Cancer datasets. Conversely for the remaining datasets we hypothesized that the increase of layers lead to lower convergence rate.

For the Breast Cancer dataset, Figures 2 (a), (b), and (c) also illustrate a fast decrease in loss, particularly in Figures (b), where the convergence occurs more swiftly compared to Figure (a). Figure (c) did not show fast initial convergence however at the end in converged to a good point where the loss is minimum. The training and validation losses stabilize smoothly, and the confidence intervals remain narrow, indicating a well-optimized and generalizing model. These results align with the hypothesis that increasing the model's depth enhances its learning efficiency. The deeper models demonstrate a superior ability to represent the data, which leads to faster and smoother convergence while maintaining good generalization performance.

In the case of the Abalone dataset, the results shown in Figures 3 (a), (b), and (c) provide evidence of rapid initial convergence. In all three figures, both the training and validation losses decrease significantly within the first few epochs and stabilize around epoch 5. The confidence intervals are relatively small, indicating consistency in the learning process.

For the Glass dataset, Figures 3 (d), (e), and (f) reveal a consistent trend of slower convergence as the number of layers increases. In particular, Figures (e) and (f) exhibit pronounced fluctuations in both training and validation losses, and the confidence intervals are notably wider, reflecting greater instability in the training process. The gradual loss reduction, especially in Figure (f), indicates that the deeper models struggle to optimize effectively. These observations confirm the hypothesis that increasing the complexity of the model slows down the convergence rate, likely due to the data's inherent noise or complexity and potential issues like vanishing gradients or becoming stuck in local minima.

The Soybean dataset results, depicted in Figures 3 (g), (h), and (i), further support the hypothesis. Figures (h) and (i) show significant challenges in achieving convergence, with pronounced oscillations in the loss values and wide confidence intervals. The models appear to plateau early, and subsequent training does not lead to notable improvements. This indicates that deeper models are less effective for this dataset, possibly due to overfitting or an inability to generalize from noisy data. The increased complexity of the model exacerbates the learning difficulties, resulting in a slower and more unstable convergence process.

The Machine dataset results, shown in Figures 3 (d), (e), and (f), also illustrate the negative impact of additional layers on convergence. Figures (e) and (f) display extremely wide confidence intervals and a very slow reduction in loss. The models struggle to stabilize, and the validation loss remains significantly higher than the training loss, suggesting poor generalization and overfitting. This confirms the hypothesis that increasing the number of layers slows down convergence and makes the training process less stable. The deep models seem unable to efficiently learn from this dataset, highlighting the limitations of increased complexity when dealing with challenging data structures.

Finally, for the Forest Fires dataset, shown in Figures 3 (g), (h), and (i), also illustrate the positive impact of additional layers on convergence. This confirms the hypothesis that increasing the number of layers accelerates the convergence.

To sum up, the results for the Breast Cancer datasets Abalone and Forest Fires support the hypothesis that increasing the number of layers leads to faster convergence. In contrast, the results for the Glass, Soybean, and Machine datasets indicate that adding more layers slows down convergence, introduces instability, and potentially reduces model performance.

5 Discussion

The findings of this study provide mixed support for our hypotheses regarding the effect of adding more layers to neural networks on classification and regression tasks. While the addition of layers did lead to performance improvements in some cases, it also led to increased model complexity, which affected the model's ability to learn efficiently, particularly for certain datasets.

For the Glass and Soybean datasets, we observed that adding more layers reduced the F1 Score significantly. This suggests that these datasets did not benefit from increased model depth. However, for other datasets like Breast Cancer or Forest Fires, the addition of layers resulted in improvements in performance.

In terms of convergence rate, our results also indicated a varied response to increasing model depth. For larger datasets, such as Breast Cancer, Abalone, and Forest Fires, deeper models converged more quickly. In contrast, for datasets like Glass, Soybean, and Machine,

deeper models led to slower convergence rates. This behavior may be attributed to overfitting. This highlights the importance of considering the characteristics of the data when choosing the depth of a neural network model.

Additionally, the curse of dimensionality appeared to play a role in determining the effectiveness of added layers. For example, the Soybean dataset, with its high number of features relative to the number of samples, posed challenges for deeper models, which struggled to achieve convergence. This further emphasizes the importance of considering dataset characteristics, such as dimensionality, when designing neural network models.

6 Summary

In conclusion, the findings of this project illustrate that increasing the number of layers in a neural network is not a universally optimal strategy for enhancing performance for all datasets. Instead, the dataset's complexity, structure, and feature space must be taken into account to determine the appropriate model depth. These insights align with our hypothesis that deeper models can yield better performance for larger datasets but may also introduce instability and reduced learning efficiency in datasets with an inadequate number of samples.

Appendix A.

Overall the workload was distributed equally between group members.