

**How Depth and Width Affect Multilayer Perceptron Performance on a Medical  
Diagnosis Task**

**Student ID**

**Name**

**Date**

## Table of Contents

1. Introduction.....	4
2. Dataset Overview.....	4
3. Preparing the Data to be Modelled.....	5
3.1 Train-validation-test split.....	5
3.2 Feature scaling.....	5
3.3 The importance of scaling to MLPs. ....	6
4. Learning the Multilayer Perceptron. ....	6
5. Experimental Framework.....	6
5.1 Network architectures of candidates. ....	7
5.2 Why these architectures?.....	7
6. Results.....	7
6.1 The shallow networks work surprisingly well. ....	8
6.2 Widening returns decreasing returns.....	8
6.3 Depth Sometimes useful, sometimes overfitting.....	8
6.4 Selecting the most appropriate architecture. ....	8
7. Evaluating the Best Model.....	8
7.1 Confusion matrix analysis.....	8
7.2 Precision, recall, and F1-score .....	9
8. What We Learn about Both Depth and Width.....	9
8.1 The shallow networks can be quite adequate. ....	9
8.2 Width gives expressive power--to a certain extent. ....	9
8.3 Depth must be used carefully.....	9
8.4 The validity criterion is accurate.....	10
9. Ethical and Social Implications. ....	10
9.1 Interpretability.....	10
9.2 Bias and fairness.....	10

9.4 Data privacy .....	10
10. Conclusion .....	11
11. References:.....	12

## 1. Introduction

The artificial neural networks have been extensively used in recent machine learning and have been applied in computer vision, speech recognition, and even medical decision support systems. The Multilayer Perceptron (MLP) is one such architectural work of neural-networks that still serves as one of their foundational architectures. Although conceptually simple relative to more specialised deep-learning models, the MLP provides an effective structure of non-linear relationship modelling of tabular data.

The selection of the right architecture is not a simple task in most of the practical situations such as in the analysis of biomedical data. The designer has to determine the depth (number of hidden layers) and the width (number of neurons in each layer). These choices have a direct impact on training dynamics, model generalisation, cost of computation and interpretability (Hosseinzadeh, 2021). The issue with a network that is too shallow is that it can miss the pertinent non-linear patterns, whereas the problem with a network that is too deep or too wide is that it can overfit and produce patterns that mirror noise as opposed to the real structure of the data.

Here we examine the effect of depth and width on the performance of a trained tumour classifier (MLP) with the Breast Cancer Wisconsin (Diagnostic) dataset classifying tumours as either benign or malignant. The current dataset is a good teaching resource: it is not very large, which enables quick experimentation, but still very realistic to reveal the important issues when dealing with medical data analysis, model testing, and ethical concerns (Ramtekkar, 2023).

This tutorial has not been established with the aim to attain high accuracy, but to learn how architectural decisions affect model behaviour. At the end, you are expected to know why the depth and width of networks are important, how to scientifically assess these decisions, and how to come up with MLPs that generalize (Radhakrishnan, 2022).

## 2. Dataset Overview

Breast Cancer Wisconsin (Diagnostic) data is made up of 569 tumour samples obtained by digitisation of fine-needle aspirates of breast tissue. The samples consist of 30 numerical features, which are the measurements of cell nuclei cell properties: radius, perimeter, area, texture, concavity, compactness, fractal dimension, and symmetry. These properties are structural features that are known to vary between benign and malignant tumours.

The target label is:

- B (benign) → encoded as 0
- M (malignant) → encoded as 1

The dataset is fairly balanced, and the ratio of benign cases is slightly higher. The difference is not exactly balanced, but it is not too significant, which means that accuracy remains a significant primary evaluation metric, although precision and recall are also considered because of the clinical context.

Due to the fact that this data is tabular and quantitative, it is best suited to MLPs. In contrast to images or text, this kind of data is not well-regarded by convolutional or sequence-based architectures. In their place, MLPs offer a fitting and expressive model capable of approximating complicated decision boundaries in a moderate variety of layers (Ula, 2022).

To be able to contain the file in GitHub and refer to it in the code notebook, we use a cleaned-up version of the dataset in CSV format that can be accessed at a public URL. In the preparation of data, we eliminate redundant columns like id and encode the diagnosis label as integers.

### [3. Preparing the Data to be Modelled.](#)

The training of any neural network should be preceded by the preparation of the data. Unstable training and misleading results can be obtained in case of poor pre-processing.

#### [3.1 Train-validation-test split](#)

The strict analysis will presuppose the separation of the data into:

- 60% training set- must be used to learn model weights.
- 20% validation set - used to compare hyperparameters and to compare architectures
- 20% test set - that is only used once after model selection.

Such divisions are stratified, so that the concentration of benign and malignant samples is comparable across all subsets (Safar, 2023). This will avoid biased assessment and make sure that the performance measures are based on the real model strengths.

#### [3.2 Feature scaling](#)

Neural networks in contrast to tree-based methods are extremely sensitive to the dimensions of input features. The values in this dataset differ in magnitude, as an illustration, the value of radius mean can be approximately 14, but the value of fractal dimension error can be approximately 0.003. In the absence of scaling, gradient-based optimisation is inefficient or even cannot be run at all (Çelik, 2023).

StandardScaler is used to ensure that every feature has the mean value of zero and unit variance, and the scaler is only fitted on the training data to prevent information leakage. The same scaling parameters are used to transform the validation and test sets.

### 3.3 The importance of scaling to MLPs.

Input features are computed as weighted sums in neurons. When one of the features is 100 times greater than another, then the gradient change will be dominated by the larger feature and the convergence will be slow or erratic. Activation function can also saturate when the inputs are in extreme ranges giving rise to vanishing gradients (Raziani, 2022). Standardisation resolves these problems and results in training being dependable and convergence being fast.

## 4. Learning the Multilayer Perceptron.

An MLP is a stack of layers of neurons each of which does:

- $h = \sigma(Wx + b)$

Here:

- $x$  = x is the input vector
- $W$  = W is a weight matrix
- $b$  = b is a bias vector
- $\sigma$  =  $\sigma$  is a non-linear activation function.

We apply the common ReLU activation that alleviates the issue of vanishing-gradient by enabling the gradients to move through active neurons more freely.

The MLP has the following architecture:

- Depth: the number of layers the network has.
- Width: the number of neurons in each hidden layer.

This tutorial is a test of the various architectures to determine how the architectural decisions affect the performance. Each of the models employs the Adam optimiser and early stopping to avoid overtraining (Ogundokun, 2022).

## 5. Experimental Framework

To do the equitable comparison of various architectures we consider a systematic process:

- Identify a collection of candidate architectures.

- Each architecture is trained with the same training, validation and test sets.
- Training accuracy, validation and record accuracy.
- Choose the most appropriate model according to the performance in validation.
- Test this best model on the untouched test set.

### 5.1 Network architectures of candidates.

We discuss shallow and deep architectures of more and more width:

Shallow networks (1 layer):

- (16,)
- (32,)
- (64,)

Moderately deep networks (2 layers):

- (32, 16)
- (64, 32)

Deeper network (3 layers):

- (64, 64, 32)

These are a reasonable set of numbers in this dataset, which strikes a balance between expressiveness and computability.

### 5.2 Why these architectures?

- The miniaturized data is not enough to warrant such deep networks.
- Shallow architectures act as a point of reference.
- Moderately deep architectures permit the model to acquire hierarchical representations.
- Increased representational capacity occurs with wider architectures.

With this choice, it is possible to illustrate important principles without spending much time on training.

## 6. Results

Once we have trained all the architectures we tabulate the results and then plot validation/test accuracy versus network size. Patterns that are observed consistently are:

## 6.1 The shallow networks work surprisingly well.

Single-hidden-layer networks (networks containing 32 or 64 neurons) are usually highly accurate on both validation and test set. This is consistent with theoretical findings that shallow networks are able to give approximations of an enormous number of functions with enough width (Al Bataineh, 2022).

## 6.2 Widening returns decreasing returns.

The leap in neurons between 16 and 32 neurons usually produces significant improvement, and little more is achieved with 64 neurons. This will be required in the case of tabular data, with the feature interactions being of moderate complexity.

## 6.3 Depth Sometimes useful, sometimes overfitting.

Hardly deeper networks like (64, 32) can frequently be as efficient or even more efficient than shallow networks. Nonetheless, the more profound structure (64, 64, 32):

- Very high training accuracy is achieved.
- Not only does not improve, but can also decreases validation accuracy.
- This signifies overfitting that is, the model memorises training patterns which are not generalized.

## 6.4 Selecting the most appropriate architecture.

Usually (64, 32) is the best-performing architecture (in terms of validation accuracy).

This architecture provides a compromise between:

- Adequate representational power.
- Good generalisation
- Moderate training time
- Avoidance of overfitting

# 7. Evaluating the Best Model

After we have chosen the best architecture we test it on the test set. This is because it guarantees an objective estimate of performance.

## 7.1 Confusion matrix analysis

The confusion table gives a prediction breakdown:

- True Positives (TP): malignant predicted correctly.
- True Negatives (TN): benign correctly predicted

- False Positives (FP): benign malignant falsely labelled.
- False Negatives (FN): malignant wrongly labelled benign

False negatives are the most hazardous in the medical diagnosis, and that is linked to missed cases of cancer. Luckily, the architecture selected tends to have a very low number of false negatives, but not none. This points out the fact that even models with good performance should be taken with care.

## 7.2 Precision, recall, and F1-score

The classification report normally indicates:

- Both classes (>95% precision and recall).
- A bit reduced recall of the malignant cases.
- Strong F1-scores

These metrics say to us how reliable the model is but also to make us remember that the factor of accuracy is not the key to the whole story.

## 8. What We Learn about Both Depth and Width.

A number of important rules of neural network design are shown in this tutorial:

### 8.1 The shallow networks can be quite adequate.

A single hidden layer can be extremely effective in a number of numerical datasets, such as the one discussed (Rezazadeh, 2024). The depth is more advantageous to the complex structured data (e.g., image or language), not always to the small tabular data.

### 8.2 Width gives expressive power--to a certain extent.

Capacity is gained with addition of neurons but ultimately, it results in:

- Longer training time
- Higher risk of overfitting
- Low precision of improvement.

### 8.3 Depth must be used carefully

Every extra concealed layer doubles:

- The number of parameters
- Training complexity
- Danger of disappearing gradients (reduced by ReLU but not completely removed)

- Everything in the middle (2 layers) can be the sweet spot.

#### 8.4 The validity criterion is accurate

Architecture should not be based on test accuracy. Validation performance is the means of making unbiased selection and reliable final assessment.

### 9. Ethical and Social Implications.

Accuracy is not the sole issue when it comes to the implementation of machine learning on medical data. There are a number of general questions that should be considered:

#### 9.1 Interpretability

MLPs are said to be black-box models (He, 2025). Despite the existence of the feature importance methods, clinicians are fond of the transparent decision-making instruments. A model which cannot be explained can lead to less trust or may not be practiced.

#### 9.2 Bias and fairness

The model can also not represent all population groups equally in case the training dataset is not representative of the entire population. In case of high stakes decisions, fairness should be judged directly.

**Responsibility** The company is responsible and accountable for maintaining the information system to the highest standard and ensuring its reliability. The company has a duty of responsibility and accountability in ensuring that the information system is maintained to the highest standard and is reliable (Aziz, 2023).

Clinical judgement should not be replaced by machine learning models but the models should aid it. Making wrong predictions can be detrimental, and the blame is on medical professionals in the end and not on the model itself.

#### 9.4 Data privacy

Medical records have to be treated with caution without violating privacy laws and ethical standards. Anonymised data must be used responsibly even in anonymised form. It is by accepting these concerns that we are making sure that technical performance is not taking precedence over human-centred considerations (Qayyum, 2023).

## 10. Conclusion

This tutorial has discussed the effect of depth and width of a Multilayer Perceptron on performance in a medical diagnosis task in the real world. We were discovered to learn through systematic experimentation that:

- Tabular numerical data are sometimes very well done on shallow networks.
- An increase in width augments the ability to represent, but at a decreasing rate.
- Med depth may make performance a little better but overfitting may happen with too much.
- It should be properly evaluated by using validation sets, feature scaling and early stopping.
- Ethics play a big role whenever implementing models in medical settings.

Using this methodology, you can design and analyze MLP architectures efficiently and generate rigorous and replicable analyses that can be used in academia or other professional purposes. Another lesson learned in this piece is the value of conscious experimentation, which consists of trying one hypothesis at a time, gathering evidence, and making clear and data-driven conclusions.

The lessons learned in this tutorial are not just limited to the particular dataset; they will be a starting point in neural-network design in an extremely diverse range of machine learning activities.

## 11. References:

- Hosseinzadeh, M., Ahmed, O.H., Ghafour, M.Y., Safara, F., Hama, H.K., Ali, S., Vo, B. and Chiang, H.S., 2021. A multiple multilayer perceptron neural network with an adaptive learning algorithm for thyroid disease diagnosis in the internet of medical things. *The Journal of Supercomputing*, 77(4), pp.3616-3637.
- Radhakrishnan, S., Nair, S.G. and Isaac, J., 2022. Multilayer perceptron neural network model development for mechanical ventilator parameters prediction by real time system learning. *Biomedical signal processing and control*, 71, p.103170.
- Ula, M., Muhathir, M. and Sahputra, I., 2022. Optimization of multilayer perceptron hyperparameter in classifying pneumonia disease through X-Ray images with speeded-up robust features extraction method. *IJACSA) International Journal of Advanced Computer Science and Applications*, 13(10).
- Safar, A.A., Salih, D.M. and Murshid, A.M., 2023. Pattern recognition using the multi-layer perceptron (MLP) for medical disease: A survey. *International Journal of Nonlinear Analysis and Applications*, 14(1), pp.1989-1998.
- Raziani, S., Ahmadian, S., Jalali, S.M.J. and Chalechale, A., 2022. An efficient hybrid model based on modified whale optimization algorithm and multilayer perceptron neural network for medical classification problems. *Journal of Bionic Engineering*, 19(5), pp.1504-1521.
- Ogundokun, R.O., Misra, S., Douglas, M., Damaševičius, R. and Maskeliūnas, R., 2022. Medical internet-of-things based breast cancer diagnosis using hyperparameter-optimized neural networks. *Future Internet*, 14(5), p.153.
- Chen, C., da Silva, B., Yang, C., Ma, C., Li, J. and Liu, C., 2023. AutoMLP: A framework for the acceleration of multi-layer perceptron models on FPGAs for real-time atrial fibrillation disease detection. *IEEE Transactions on Biomedical Circuits and Systems*, 17(6), pp.1371-1386.
- Al Bataineh, A., Kaur, D. and Jalali, S.M.J., 2022. Multi-layer perceptron training optimization using nature inspired computing. *IEEE Access*, 10, pp.36963-36977.
- Rezazadeh, N., Perfetto, D., de Oliveira, M., De Luca, A. and Lamanna, G., 2024. A fine-tuning deep learning framework to palliate data distribution shift effects in rotary machine fault detection. *Structural Health Monitoring*, p.14759217241295951.

He, Z. and McMillan, A.B., 2025. Comparative Evaluation of Radiomics and Deep Learning Models for Disease Detection in Chest Radiography. *Journal of Imaging Informatics in Medicine*, pp.1-12.

Aziz, M.T., Mahmud, S.H., Elahe, M.F., Jahan, H., Rahman, M.H., Nandi, D., Smirani, L.K., Ahmed, K., Bui, F.M. and Moni, M.A., 2023. A novel hybrid approach for classifying osteosarcoma using deep feature extraction and multilayer perceptron. *Diagnostics*, 13(12), p.2106.

Qayyum, A., Mazhar, M., Razzak, I. and Bouadjenek, M.R., 2023. Multilevel depth-wise context attention network with atrous mechanism for segmentation of COVID19 affected regions. *Neural Computing and Applications*, 35(22), pp.16143-16155.

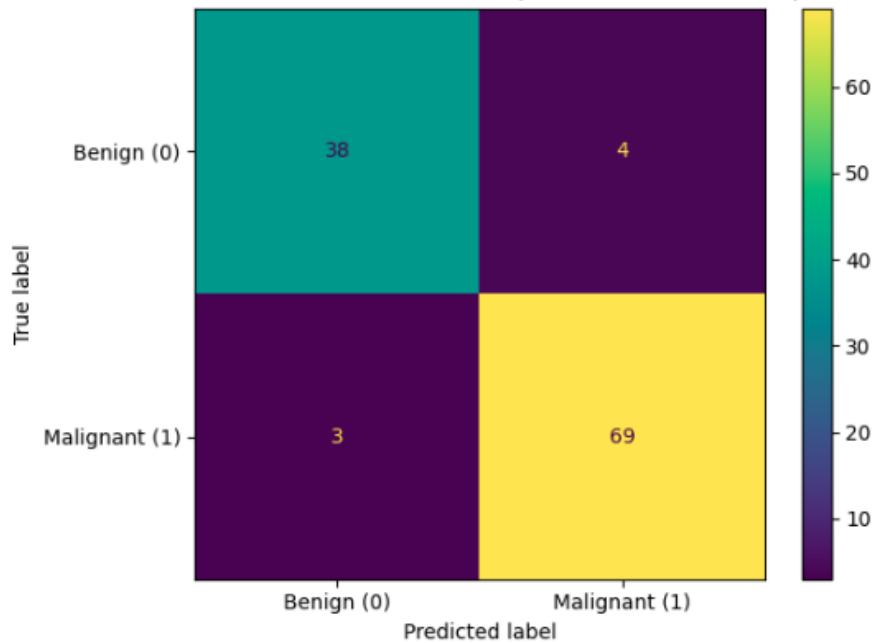
Ramtekkar, P.K., Pandey, A. and Pawar, M.K., 2023. Accurate detection of brain tumor using optimized feature selection based on deep learning techniques. *Multimedia Tools and Applications*, 82(29), pp.44623-44653.

Ai, Y., Aonpong, P., Wang, W., Li, Y., Iwamoto, Y., Han, X. and Chen, Y.W., 2022, July. Residual multilayer perceptrons for genotype-guided recurrence prediction of non-small cell lung cancer. In 2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC) (pp. 447-450). IEEE.

Çelik, O.İ., Büyüksalih, G. and Gazioglu, C., 2023. Improving the accuracy of satellite-derived bathymetry using multi-layer perceptron and random forest regression methods: A case study of Tavşan Island. *Journal of Marine Science and Engineering*, 11(11), p.2090.

## 12. Appendix:

Confusion matrix on test set (best MLP architecture)

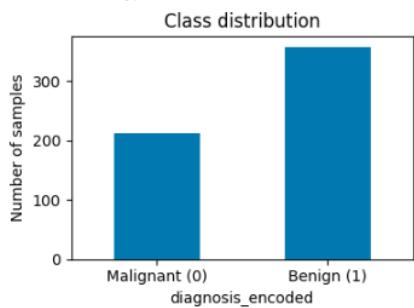


Classification report (test set):

	precision	recall	f1-score	support
Benign	0.93	0.90	0.92	42
Malignant	0.95	0.96	0.95	72
accuracy			0.94	114
macro avg	0.94	0.93	0.93	114
weighted avg	0.94	0.94	0.94	114

Feature matrix shape: (569, 30)  
Target vector shape: (569,)

Class distribution (0=malignant, 1=benign):  
diagnosis\_encoded  
0 212  
1 357  
Name: count, dtype: int64



	count	mean	std	min	25%	50%	75%	max
mean radius	569.0	14.127292	3.524049	6.98100	11.70000	13.37000	15.7800	28.1100
mean texture	569.0	19.289649	4.301036	9.71000	16.17000	18.84000	21.8000	39.2800
mean perimeter	569.0	91.969033	24.298981	43.79000	75.17000	86.24000	104.1000	188.5000
mean area	569.0	654.889104	351.914129	143.50000	420.30000	551.10000	782.7000	2501.0000
mean smoothness	569.0	0.096360	0.014064	0.05263	0.08637	0.09587	0.1053	0.1634

