

Neural Networks for Breast Cancer Detection on Limited Featuresets

Lars Bogner

University of Oslo

<https://github.uio.no/larsbog/FYSSTK-Project2>

(Dated: November 5, 2025)

We study an approach to classification on the Wisconsin Breast Cancer Diagnostic dataset using neural networks to expand a limited set of physical features to the full feature set. Using only the radius and area of cell nuclei, we train a feedforward neural network to predict the remaining features with a mean squared error of $3.15\sigma^2$. Using out-of-fold prediction we obtain independent predictions for the entire dataset. Using these predictions as input for a logistic regression classifier, we achieve an accuracy of 94.74% and an AUC score of 0.99105, comparable to results using the full dataset.

Contents

I. Introduction	1
II. Methods	1
A. Neural Networks	1
B. Activation Functions	2
C. Loss Functions	2
1. Other Evaluation Metrics	3
D. Regularization	3
E. Optimization Algorithms	3
F. Data Splitting Techniques	4
1. Train-Test Splitting	4
2. Out-of-Fold prediction	4
G. Implementation	4
H. Use of AI tools	5
III. Results and Discussion	5
A. Introduction to the Dataset	5
B. Regression on Limited Feature Set	6
1. Hyperparameter Tuning	6
2. Comparison against Linear Regression	7
C. Classification on Regression Output	7
1. Using Logistic Regression for Classification	8
IV. Conclusion	8
A. Potential Future Work	9
References	9

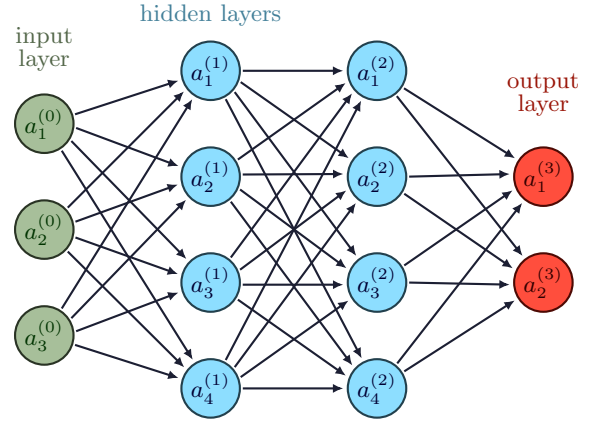


Figure 1: Visualization of a neural network, modified work of [4].

Starting from the features of radius and area from the Wisconsin Breast Cancer Diagnostic dataset [1, 3] we obtain the full set of features as presented in the dataset using an optimized regression model. Using this regression model it will be possible in the future to use standard classifiers while only needing to measure the radius and area of the cell nuclei in diagnosis. In section II we present the methods necessary to implement the corresponding neural networks and how to optimize them. The results and a discussion thereof is presented in section III. Finally, our findings are concluded in section IV.

I. INTRODUCTION

Sampling the cell nuclei of tissue extracted using a fine needle aspirate from the breast allows for a very easy separation between benign and malignant tissue[1]. While very good performance is already shown with straightforward approaches, like decision trees [2], we try to improve the efficiency of the detection process in this paper.

Using a very small subset of features which can be obtained using basic image recognition, we enhance them to a larger set of nuclei features using a regression model, based on which the straightforward classification can be continued using the large set of features.

II. METHODS

A. Neural Networks

Neural networks are the continuation of a linear combination by introducing nonlinearity. While linear regression uses a linear combination of input features to create an output, neural networks add a nonlinear function to the output of the linear combination. This is the entire secret behind the powerful possibilities neural networks have shown over the last years [5, 6].

A feedforward neural network (FFNN) uses a layered

structure as shown in figure 1, where each node represents a separate linear combination of its inputs. A general node uses the inputs x_i to create a weighted sum z , which is then activated using a so called activation function $g(z)$. More detail on the activation function is provided in section II B. The output of the general node can then be calculated as

$$y = g(z) = g\left(\sum_i (w_i x_i) + b\right), \quad (1)$$

where b is the so-called bias of the node which in turn leads to a shift of the weighted sum, irrespective of the inputs. The linear combination $z = \sum_i (w_i x_i) + b$, can also be rewritten as a vector-vector multiplication of vectors $\vec{a}_j^{(k)} = (b, \vec{w})$ and $\vec{X} = (1, \vec{x})$. This leads to a simple expression for the output of a node of $y = g(\vec{a} \cdot \vec{X})$. All parameters of a single node in layer k and position j in the layer are contained within $\vec{a}_j^{(k)}$ [5, 7].

We use the term layer, as the output of each layer will be used as the input of the following layer, thus presenting a clear hierarchy. This hierarchy is also useful for calculating the gradient of the final output of the neural network, with respect to all the weights in the layers, as we can use chained differentiation. Calculating the gradient with respect to all the weights is necessary as the optimization techniques presented in section II E rely on the knowledge of the gradients to find the optimal set of weights. The gradient of a single node given the input $\vec{x}$ is given by

$$\frac{\partial y}{\partial \vec{a}} = \vec{X} g'(\vec{a} \cdot \vec{X}), \quad (2)$$

where g' is the derivative of the activation function with respect to its input. Then given the input of a node in the previous layer we can calculate the gradient with respect to its weights using the chain rule of derivatives.

B. Activation Functions

Activation functions provide a source of nonlinearity in the otherwise linear combination of inputs of a neural network. This nonlinearity is key, as it provides the basis for the famous Universal Approximation theorem, which states, that FFNN can approximate any arbitrary real function given a large enough number of nodes [5, 7]. While one might come to the conclusion, that if this nonlinearity is of uttermost importance, the right choice of activation is complicated, it is proven by many practical applications, that already very simple activation functions may provide superb performance.

a. Rectified Linear Unit For simplicity this paper will therefore focus on a set of three activation functions. A very simple non-linear function is provided by the Rec-

tified Linear Unit (ReLU), which shall be defined as

$$\text{ReLU}(z) = \begin{cases} 0 & z < 0 \\ z & z \geq 0 \end{cases}. \quad (3)$$

While this function provides all necessary properties of an activation function, we will provide an explanation, why it is not well suited for the optimization procedure of FFNN. The issue with the ReLU is that for $z < 0$ the gradient of the function vanishes. This poses a problem, as mentioned above, the optimization algorithm relies on the gradient of the output to adjust and optimize the weights of the neural network. A vanishing gradient disables a possible optimization. This problem is further amplified by the fact, that the gradient of weights in the first layers of the network depends on the product of the gradients with the previous layers, thus a vanishing gradient in one of the last layers has adverse effects on the optimization of weights in the previous layers[8].

b. Leaky Rectified Linear Unit As the ReLU is otherwise quite optimal due to its simplicity, we will mainly use a slightly modified version of the ReLU for the majority of nodes in the studied neural networks, namely the Leaky ReLU (LReLU). The leaky ReLU modifies the standard ReLU by introducing a very small but non-vanishing gradient for $z < 0$. It is defined as

$$\text{LReLU}(z) = \begin{cases} \epsilon z & z < 0 \\ z & z \geq 0 \end{cases}, \quad (4)$$

with ϵ being the leak coefficient, a small number, set to $\epsilon = 10^{-5}$ for all studies. Due to the step in the derivative at $z = 0$ our function is still nonlinear, but providing a positive derivative over the whole domain[5].

c. Softmax The last activation to introduce for our studies will be the Softmax function. It shall be defined as [8]

$$\text{Softmax}(z_j^{(k)}) = \frac{\exp z_j^{(k)}}{\sum_{j'} \exp z_{j'}^{(k)}}. \quad (5)$$

Additionally, to providing a source of nonlinearity, the softmax function has the special property, that the total output of a layer, i.e. the sum of its outputs is regularized to a value of 1. This property is especially useful in the case, where it is used as a classifier, where each of the layers outputs corresponds to a probability of being member in a certain class. While there is no guarantee, that the output of such a layer corresponds to a frequency in the frequentist interpretation, it is very useful in combination with the Cross Entropy introduced in the next section, as we have a nice behaving gradient, useful for optimization [8].

C. Loss Functions

The function of the loss function is to quantize the discrepancy between the prediction of a numerical model

and the true values [5, 6, 8]. It is also sometimes called cost function. It must be a differentiable function which is minimal when, the prediction aligns with the true values. The choice of loss function needs to be specific to the task for which the model should be optimized for. For the regressional part of this paper we will use the mean squared error (MSE) as a loss function. The MSE is defined using the prediction $\hat{y}$ and the true values y as

$$\text{MSE} = \frac{\sum_i (\hat{y}_i - y_i)^2}{N}. \quad (6)$$

Its derivative is therefore defined as

$$\frac{\partial \text{MSE}}{\partial \hat{y}_i} = \frac{2(\hat{y}_i - y_i)}{N}. \quad (7)$$

The advantage of the MSE as a loss function is, that the loss value can be directly interpreted in the context of the metric on which we optimize [5]. Furthermore, it gives the advantage, that compared to the mean absolute error, which uses the absolute deviation between the target and the prediction, that it is continuous differentiable and that large deviations are penalized more than small deviations. Especially in the context of small measurement uncertainties this is important.

The second loss function used is the cross entropy loss. It is used for the case of classification tasks and is defined as [5]

$$\text{CE} = \sum_i y_i \log \hat{y}_i. \quad (8)$$

The cross entropy loss is inspired by the definition of the physical entropy of stochastic systems and is useful, as it is minimal when the probabilities $\hat{y}$ and y align. Combined with the softmax activation function from section II B, the derivative simplifies to a proportional relation with respect to $\hat{y} - y$.

1. Other Evaluation Metrics

Additional to the metrics on which we optimize the model, there are other scores than can be important in the context of evaluating the quality of a models performance. The relative amount of true predictions of a classification model can be judged using the accuracy. The accuracy is defined as

$$\text{Accuracy} = \frac{\sum_{c \in C} N(y = c, \hat{y} = c)}{N_{\text{total}}}, \quad (9)$$

where C contains all the possible classes of the dataset and N is the number of predictions with certain conditions. A second metric for evaluating the performance of a classifier is the area-under-curve (AUC) score of the receiver-operating-characteristic (ROC). It is defined as

$$\text{AUC} = \int_0^1 \mathcal{T}(\mathcal{F}) d\mathcal{F}, \quad (10)$$

where $\mathcal{T}$ is the true-positive-rate and $\mathcal{F}$ is the false-positive-rate. A score of 1 indicates perfect classification, while a score of 0.5 is equivalent to a random classifier, as in that case $\mathcal{T}(\mathcal{F}) \equiv \mathcal{F}$.

D. Regularization

As the number of parameters in a neural network grows very quickly, special attention needs to be paid to avoid overfitting the training dataset. Overfitting means, that the available degrees of freedom of the model will be used to represent all the datapoints in the training dataset perfectly. This will lead to very poor generalization ability of the model. This phenomenon is part of the bias-variance trade-off, which is ubiquitous in machine learning. With increasing numbers of parameters in the model, i.e. due to an increase in layer size and/or increase in the number of layers, the number of degrees of freedom also increases. With a larger number of degrees of freedom the variance of the model increases. If there are too few parameters, the bias of the model will increase. The goal is therefore to find the middle point, where neither the bias nor variance are increased.

A useful tool to control the bias variance trade-off is regularization in neural networks. Using regularization we can decrease the effective number of degrees of freedom, therefore decreasing the tendency to overfit the data [5]. The decrease is thereby controlled by the regularization coefficient λ . In the limit of $\lambda \rightarrow 0$, the solution approaches the unregularized model, while there is a steady decrease of the number of effective degrees of freedom with an increase of λ . Regularization works by adding a term to the cost function of the form

$$C_{\text{regularized}} = C + \lambda \sum_k \sum_j \|a_j^{(k)}\|_n^n. \quad (11)$$

Equation (11) is the generalized expression for a L_n -regularization, where n is the degree of the norm in equation (11). The most common types of regularization are L1- and L2-regularization. Using a 1-norm for regularization leads to the benefit that it has the tendency to force an ever-increasing number of parameters of the model towards zero. It thus effectively reduces the number of weights that influence the final result [7]. Especially in highly optimized architectures this may be advantageous, as it reduces the number of computations needed for each evaluation. On the other hand it is known to produce less stable results, as the L2-regularization [8].

E. Optimization Algorithms

The goal of an optimization algorithm is to tune the weights of a FFNN in a way, so that the cost function will be minimized, i.e. the predictions and the targets align. Most algorithms use a way of gradient descent to

update the weights and find the optimum via iterative optimization. Gradient descent uses the derivatives of the cost function

$$\delta_j^{(k)} \equiv \frac{\partial C}{\partial a_j^{(k)}} \quad (12)$$

and update the weights given a learning rate α using [8]

$$a_j^{(k)} \rightarrow a_j^{(k)} - \alpha \cdot \delta_j^{(k)}. \quad (13)$$

To decrease the computational resources needed we will use a more advanced version of this algorithm in this paper. The Adam optimizer updates the weights according to [9]

$$\begin{aligned} m_j^{(k)} &\rightarrow \beta_1 m_j^{(k)} + (1 - \beta_1) \delta_j^{(k)}, \\ v_j^{(k)} &\rightarrow \beta_2 v_j^{(k)} + (1 - \beta_2) (\delta_j^{(k)})^2, \\ \hat{m}_j^{(k)} &= \frac{m_j^{(k)}}{1 - \beta_1^k}, \\ \hat{v}_j^{(k)} &= \frac{v_j^{(k)}}{1 - \beta_2^k}, \\ a_j^{(k)} &\rightarrow a_j^{(k)} - \alpha \frac{\hat{m}_j^{(k)}}{\sqrt{\hat{v}_j^{(k)} + \epsilon}}, \end{aligned} \quad (14)$$

with the optimization parameters $\beta_{1/2}$. As shown in our previous article [10], Adam tends to converge faster towards an optimal solution. It is also thus widely used in machine learning optimizations and will be used for all training procedures throughout the studies.

The most important constant to tune in the optimization procedure is the learning rate α . It controls the step size of each update of the weights. If it is chosen too large, the optimization may overshoot the optimal solution and diverge. If it is chosen too small, the optimization will take unnecessarily long to converge. A good choice of learning rate is therefore crucial for an efficient optimization procedure. Choosing the learning rate in this study is done by judging the convergence behavior at different learning rates graphically. We then choose a rate which provides fast convergence without divergence and an additional margin of safety.

F. Data Splitting Techniques

1. Train-Test Splitting

As we mentioned before, there is the possibility for our model to overadjust for the datapoints present during the optimization process. To still get an accurate measure for the models performance we will only use a subset of the dataset for the process of training. A random subset of 20% is reserved for testing the model only. This means, that all performance metrics are evaluated in the same

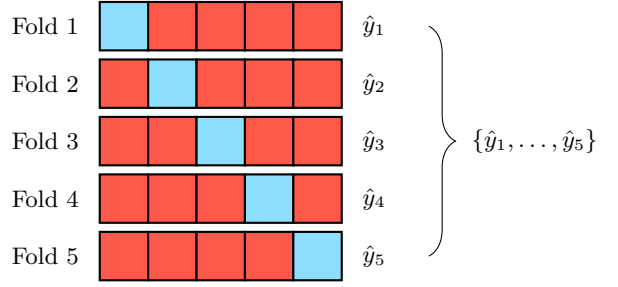


Figure 2: Illustration of predicting the full dataset using out-of-fold prediction to avoid contamination of training data. Modified from [10]. Red represents the training dataset, while the features from the blue part are used to make the prediction. Finally, all predictions are combined into one dataset of the same size as the input dataset.

way as new unseen data would perform [8]. Due to the random nature of the subset, the performance between runs may differ stochastically. This may be counteracted by redoing the tests with different choices of training and testing datasets and averaging the metrics.

2. Out-of-Fold prediction

A further hindrance in our approach with using the output of a regression model to make classifications is, that for an efficient workflow, we need a full but independent prediction of the dataset. To achieve this, we use a workflow inspired by cross-validation. In cross-validation you use multiple folds of the dataset, whereby for every fold, a different part of the dataset is used as testing data, and the rest of the dataset is used for training purposes. This provides independence of training and testing data while being able to test the models general performance for the entire dataset. For out-of-fold prediction, we use the same approach of changing the training data to leave out a specific part each time, but furthermore we use the model trained on each training set, to make predictions for the remaining set. Combining all the predicted sets into one set, allows for having independent predictions for the entire dataset. Having the predictions be independent is important, as otherwise no statement can be made on the generalization ability of the workflow. Using this approach we can expect the same performance on new and yet unseen data.

G. Implementation

All the code necessary to use a FFNN is implemented from scratch in Python3 as the `easynn` library. The source code therefore is located in the `/src/easynn/` folder of the accompanying git repository [17]. It is separated into two main modules, called `schedulers` and `feedforward`. The scheduler modules use modified ver-

sions of our previous work on using gradient descent techniques for ordinary least squares techniques. It is also structured in a modular way, so that scheduling tasks like early stopping and dynamic learning rate adjustment are easy to implement in further versions. The base scheduler implements an update method of signature `def update(self, gradients: gradient_type) -> gradient_type:`, with `gradient_type = List[Tuple[np.ndarray, np.ndarray]]`. This can be overwritten to implement any optimization algorithm. The method takes the gradients of the cost functions with respect to the weights and biases of each layer. Each layer is assigned a tuple of two arrays of the gradients. It should then return the updates for each of the weights and biases in the same format. The training process which fetches new updates after each evaluation of the loss and gradients will continue with the `cont` property of the base class is true.

The main structures of the neural network are implemented as classes in the `feedforward` module. Again it is structured in a way, where it is easy to be extended in the future. The FFNN can be constructed using the arguments `class FFNN: def __init__(self, layers: List[Layer], scheduler: Scheduler, loss_fn: LossFunction)`. The layers are constructed using `class Layer: def __init__(self, input_dim: int, num_nodes: int, activation_function: ActivationFunction, regularization: Optional[Regularization] = None)`. Using this structure it is easy to control the hyperparameters of each hidden layer carefully and for example choose the activation and regularization of each layer precisely. The weights and biases of each layer are initialized from a uniform random distribution in the bounds $[-s, s]$ with

$$s = \frac{1}{\sqrt{I \cdot N}}, \quad (15)$$

where I is the input dimension and N is the number of nodes. This initialization is taken from the standard implementation in the widely used `pyTorch` library.

For the entire numerical computation of the neural network, the `numpy` [11] library is used. Numpy is well-known for its performance in linear algebra performance and vectorization of array computations. Other libraries are solely used for plotting purposes. The libraries used during the plotting procedure include `pandas` [12], `scikit-learn` [18] [13], `matplotlib` [14] and `seaborn` [15].

To check the functionality of the developed code, in a first stage, the loss history is evaluated visually showing a clear convergence after a high enough number of epochs. Furthermore, the quality of the models is evaluated using toy-datasets created using `scikit-learn`. In detail, both a regression toy-dataset as well as a classification dataset are validated. A MSE of 0.000 631 on a dataset created from `make_regression(n.samples=1000, n.features=5,`

`n.informative=3, noise=0.1)` is achieved. For classification the `make_classification(n.samples=1000, n.features=5, n.informative=3, n.classes=2)` function is used achieving a testing accuracy of 0.944. For the classification task a one-hot encoding is used. No hyperparameter tuning was used for the toy-datasets. The performance is assessed to be more than sufficient. Similar performance would be expected using a widespread machine learning framework.

For the comparison against ordinary least squares regression, we use the framework presented before by the same author [10]. It uses the analytical solution to the ordinary least squares regression to find the optimal parameters.

As a further validation of the functionality, we use the Runge function

$$f(x) = \frac{1}{1 + 25x^2}, \quad (16)$$

with added Gaussian noise of standard deviation 1.0 to create a regression dataset. On this dataset of size $N = 100\,000$ and 80 %-20 % train-test-splitting we achieve a MSE of 0.9136 to 0.9178, using between one and two hidden layers with either 50 or 100 nodes per hidden layer. This exceeds the performance of our previous work on least squares regression using different regularization and optimization techniques [10], where a MSE of 0.92 was achieved using the same dataset [19] and train-test-splitting. This motivates further the use of neural networks for regression tasks, as possible non-linearities in the data can be modeled more easily. Least squares methods on the other hand provide only linear combinations of the input features.

H. Use of AI tools

During the writing of this paper artificial intelligence in the form of large-language models (LLMs) has been used. Two different LLMs have been used for a faster writing of the code used. Mainly for short rewrites of code structure and brain-storming the model GPT-5 by OpenAI and the model DeepSeek by Hangzhou DeepSeek Artificial Intelligence have been used. At no point was artificial intelligence used for writing the report. All output by LLMs has been checked by the author.

III. RESULTS AND DISCUSSION

A. Introduction to the Dataset

For the entire study we use the dataset presented by Wolberg [3]. It includes features recorded from nuclei in breast tissue and their geometric properties. Each physical feature is present three times, where the means and standard deviations of the measured physical features are used as separate features. Furthermore, the mean of the

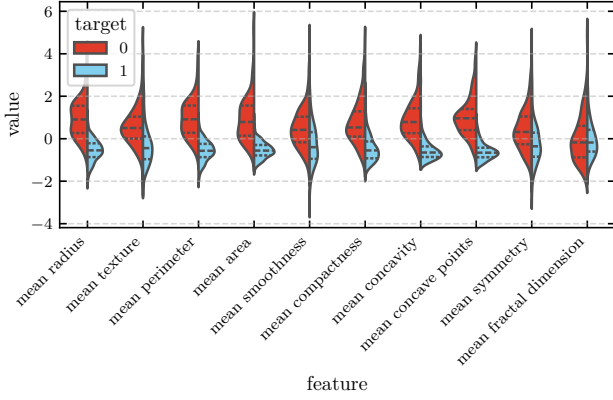


Figure 3: Violin plot of the distribution of all features in the dataset, separated by benign and malignant samples. A target value of 0 corresponds to malignant samples, while a target value of 1 corresponds to benign samples.

three largest measurements of each physical feature is present in the dataset. The physical features include: the radius, the standard deviation of the images gray-scale values (called texture), the perimeter, the area and the local variation in the radius length, called smoothness. Additionally, the concavity, the number of concave points, the compactness, the symmetry and the fractal dimension are present. For normalization, all features are rescaled to be centered around 0 and have a standard deviation of 1. The normalization is achieved using the `StandardScaler` class of the `scikit-learn` library. A quantitative overview on the average of each of the physical features is presented in figure 3. The violinplot shows the feature distribution for both classes of benign and malignant cell nuclei.

The separability of the two classes is easily visible given all features in the class. On the other hand, there is no single feature which allows for very good separation. Thus reproducing separability using only a limited set of features is the challenge of this study.

B. Regression on Limited Feature Set

We use a FFNN with six input nodes and hidden layers of equal node count. For the output we use a layer using linear activation $g(z) = z$, i.e. no activation function or in other words a linear regression. To measure the performance of the model, the MSE of the models outputs is used as a loss metric. The model is fitted using the Adam optimizer algorithm with a learning rate of $\lambda = 10^{-2}$ and decay parameters $\beta_1 = 0.9$ and $\beta_2 = 0.999$. The mean squared error per target variable is evaluated. For a simple model without any hyperparameter optimization the MSE is in the range of $2 \times 10^{-4} \sigma^2$ to $2 \times 10^1 \sigma^2$, where σ signifies the standard deviation of the feature values.

For all studies we use train-test-splitting with 20% of

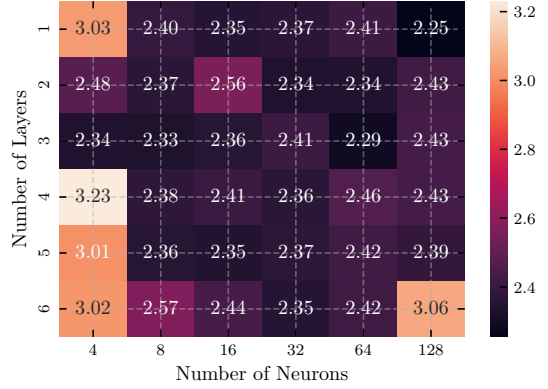


Figure 4: MSE for different numbers of hidden layers in the regression network and nodes per hidden layer.

the data reserved for testing. This is important as to get an unbiased estimate of the models performance on unseen data.

1. Hyperparameter Tuning

As mentioned in section II B, the Leaky ReLU combines the disadvantages of the ReLU while conserving non-vanishing gradients on the entire domain. Thus, it is not surprising, that they perform similar in our studies with the LReLU providing a more stable training process. Due to this finding we will use the LReLU for all the hidden layers. Figure 4 shows the resulting MSE for different numbers of hidden layers and numbers of nodes in the layers. We can observe a general trend of decrease in MSE with a decrease in the number of hidden layers and an increase in the number of neurons per hidden layer. This general trend is to be expected as with an increase in the number of hidden layers the gradients in the last layers can converge towards zero with deep networks due to the multiplicative nature of the gradients. Additionally, can an increase number of nodes per layer better capture even complicated relationships in the data. Due to the adverse effects of an increase in node number, we chose a conservative set of parameters with 32 neurons per hidden layer and two hidden layers.

In a second sub-study we observe the influence of the regularization constant with and L2-regularization on all layers. Further studies on variable regularization per layer were not carried out. The studies show a steady MSE for a range of 0 to 10^{-3} with a sharp increase by a factor of 10 with the step to $\lambda = 10^{-2}$. Thus, we choose to use $\lambda = 10^{-3}$ for the regressional model. This provides a stable insurance against overfitting of the training data while at the same time not negatively affecting the models' generalization ability.

The final set of hyperparameters is then again tested for MSE on all the features of the dataset it is optimized

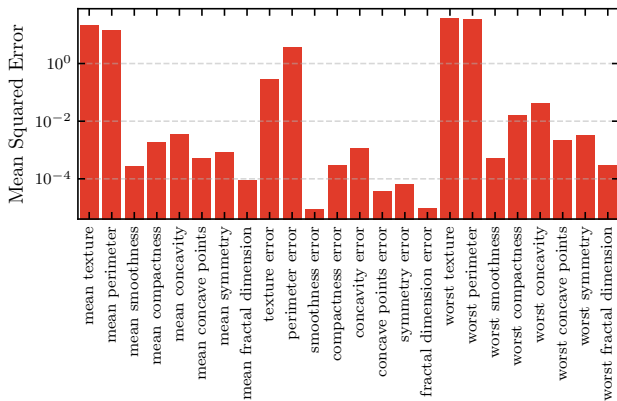


Figure 5: Average MSE of the predicted features using the regression model by feature label.

to predict. The results of the per target MSE are displayed in figure 5, while we see an increase in the upper bound of the per target MSE, to $10^1 \sigma^2$, in comparison to the unoptimized model, the general trend for the MSE shows a lower baseline. The lowest accuracy show the predictions for the texture and perimeter. Both show an MSE in excess of $5\sigma^2$ in the mean and worst values. Especially the perimeter values are surprising, as one would expect the perimeter to follow from a simple relation with the nuclei’s radius. All the other physical features behave similar with respect to the prediction error of $10^{-3} \sigma^2$ to $10^{-4} \sigma^2$. In general the prediction uncertainties are smaller on the error values of the physical features. The total MSE over all features amounts to $3.15 \sigma^2$. While the performance is not perfect, it is sufficient to create a classification model based on the regression output, as not all features are equally important for the classification task.

Based on these hyperparameters we trained models using 5-fold out-of-fold prediction as introduced in section IIF 2, to create a full version of the dataset using only radii and areas as inputs. For the classification task we will use this dataset.

2. Comparison against Linear Regression

Figure 6 shows the performance comparison in the MSE between an ordinary least squares (OLS) regression and the optimized neural network. Both show very similar performance with only slightest differences in the MSE. In general ordinary least squares underperform, compared to the FFNN by a small margin. The very good results of the ordinary least squares approach is surprising.

Ordinary least squares regression is only able to capture linear relationships between the input features and the target features. Thus, one would expect that the more complex FFNN would be able to outperform the

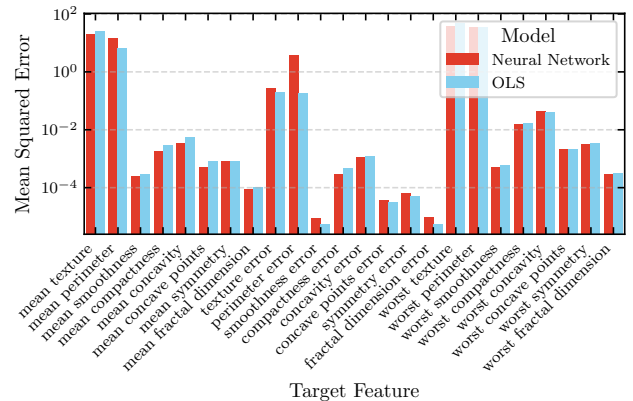


Figure 6: Comparison of the MSE per target feature between the neural network regression and ordinary least squares regression.

OLS regression by a larger margin. The small difference in performance indicates, that the relations between the input features and the target features are mostly linear in nature. Using OLS for describing linear relations is also advantageous, as it provides a closed form solution for the optimal weights. Thus, no iterative optimization procedure is necessary, which saves computational resources and time. The closed form solution also guarantees that the found solution is the global optimum of the cost function, while iterative optimization may get stuck in local minima. Even if using regularized version of linear regression, that do not possess closed form solution, the optimization of such models as the gradients are simpler to compute and the models contain fewer degrees of freedom, leading to faster convergence.

The excellent performance of the linear regression suggests that the additional challenges posed by using the FFNN are not worth the extra development costs. We thus recommend using this approach only for datasets with complex relations.

C. Classification on Regression Output

Similar to the structure of the regression model, we use a FFNN using a variable number of nodes per hidden layer in configurations with different numbers of hidden layers to create a classifier. To use the model as a classifier, the loss function of the model is changed to the cross-entropy loss introduced in section II. Additionally, we change the activation of the output layer to the softmax function. The leaky rectified linear unit is again used for the purpose in all other neurons, following the results from the previous study. The parameters are optimized using loss minimization using the Adam optimization algorithm.

As above, a parameter scan is used to find the best parameter combination for accurate prediction, whether

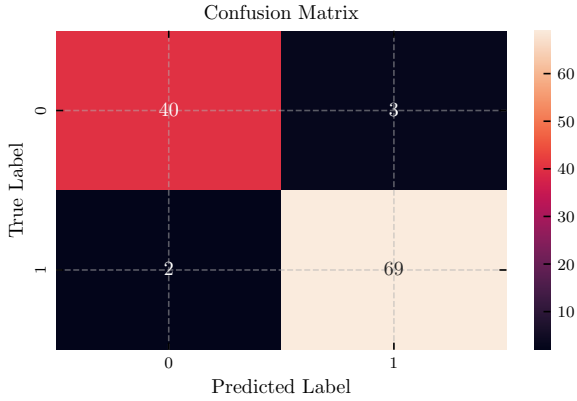


Figure 7: Confusion Matrix of the classifier for the regression output. A target value of 0 indicates malignant nuclei, while 1 is indicative of benign samples.

the nuclei are part of benign or malignant tissue. The results of the parameter scan are displayed in figure 9. For the number of hidden layers and neurons, there is a clear maximum in the area-under-curve (AUC) score for the case of 64 neurons in either a single or two hidden layers. Using the full dataset in the form without regression AUC scores of 0.9962 have been shown using logistic regression [16]. To account for stochastic variations in the training success the AUC is averaged over a total of 50 runs for each entry in figure 9. The regularization constant is chosen to be $\lambda = 10^{-6}$ according to the figure. During 50 runs this combination of hyperparameters is able to achieve an average accuracy of 92.70% and an AUC score of 0.9781. While this performance is not as good as the performance using logistic regression on the full dataset [16], it is to be expected, that the performance decreases when using only a limited set of features to create the regression output and following that the classification.

The performance is further evaluated using the confusion matrix in figure 7. Potential asymmetric biases for false identification, i.e. whether type 1 or 2 errors are more pronounced, can be detected using a confusion matrix. It shows which relative amount of each class (malignant or benign) is identified as which by the classifier. No significant asymmetry is present in our FFNN based classifier.

1. Using Logistic Regression for Classification

Thus, we also study the use of logistic regression to create a classification based on the outputs of the regression model. The logistic regression model uses the same framework code as the FFNN for comparability. For creating a logistic regression classification, a singular layer with two neurons and a softmax activation is used. The cross-entropy-loss is used as an optimization metric,

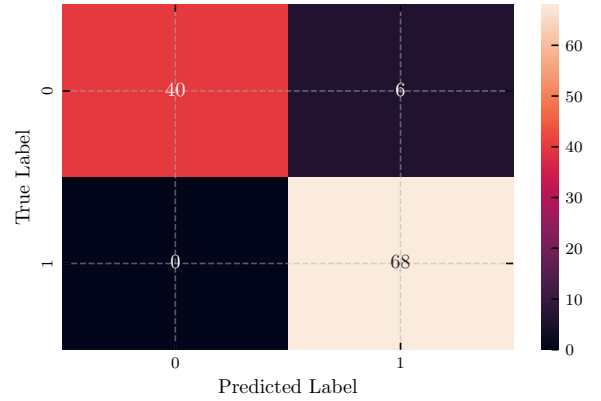


Figure 8: Confusion Matrix of the logistic regression for the FFNN regression output. A target value of 0 indicates malignant nuclei, while 1 is indicative of benign samples.

which is then minimized using the Adam optimizer.

After the minimization a test split is evaluated for accuracy and receiver-operating-characteristic leading to an accuracy value of 94.74% and an AUC score of 0.99105. The logistic regression thus outperforms the classification models by far. It is especially noteworthy, that nearly the same performance as for full original dataset data can be reached. While it might seem counterintuitive, that the more complex models are not able to match the performance of the arguably very simple logistic regression is due to the fact, that even with regularization, the more complex models have a higher tendency to overfit simple relations between the features and the classification.

The good performance of the linear regression also helps for future diagnostic purposes, as we have shown, that with the correct combination of neural networks and logistic regression it is possible to use a very small, easy to extract subset of the physical features to achieve nuclei identification at the highest level.

Additionally, we compare the confusion matrix in figure 8 against our findings for the FFNN. A bias towards type 2 errors, where malignant tissue is predicted as benign, is evident. Such a misclassification is especially important in the context of diagnostic use.

IV. CONCLUSION

It is evident, that a classification performance close to the one using the full dataset can be achieved using only a limited set of physical features. Using a FFNN based regression model followed by a logistic regression classifier, an accuracy of 94.74% and an AUC score of 0.99105 is achieved. This performance is comparable to the performance of logistic regression on the full dataset [16]. Using only six features derived from two physical measurements, the radius and area of cell nuclei, this

performance is possible.

It is therefore possible to reduce the data acquisition for future diagnostic purposes to a small set of physical features, while at the same time conserving a high classification accuracy. This is especially important in the context of fast and cheap diagnostic procedures.

Additionally, we have shown, that for the regression task on the limited feature set, the performance of ordinary least squares regression is comparable to the one of a carefully optimized FFNN. Thus, for similar tasks in the future, it may be possible to use ordinary least squares regression due to its simplicity and analytical solution.

A. Potential Future Work

While subpar performance of the FFNN regression with a high MSE on certain features is evident, future

work may include modifying the optimization metric, to better judge the importance to certain features for the classification task. This may lead to better performance of the regression model on the important features, while at the same time allowing for a higher error on unimportant features.

Furthermore, the use of a large set of features in training of a FFNN to make predictions on fewer features during application, is a promising field of study. Using such an approach, it may be possible to create models, which can make accurate predictions using only a limited set of easily acquirable features, while at the same time leveraging the information contained in a larger set of features during training.

-
- [1] W. N. Street, W. H. Wolberg, and O. L. Mangasarian, in *IS&T/SPIE's Symposium on Electronic Imaging: Science and Technology*, edited by R. S. Acharya and D. B. Goldgof (San Jose, CA, 1993), pp. 861–870.
 - [2] K. Bennett, Proceedings of the 4th Midwest Artificial Intelligence and Cognitive Science Society Conference, Utica, Illinois (1992).
 - [3] O. M. William Wolberg, *Breast Cancer Wisconsin (Diagnostic)* (1993).
 - [4] *Neural networks – TikZ.net* (2024).
 - [5] M. Elstner and T. Kubar, *Lecture: Machine Learning for Chemistry* (2025).
 - [6] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning* (MIT Press, 2016).
 - [7] C. M. Bishop, *Pattern Recognition and Machine Learning*, Information Science and Statistics (Springer, New York, 2006), ISBN 978-0-387-31073-2.
 - [8] J. Kiesler, *Lecture: Modern Methods of Data Analysis* (2025).
 - [9] D. P. Kingma and J. Ba, *Adam: A Method for Stochastic Optimization* (2017), 1412.6980.
 - [10] L. Bogner, FYS-STK4155 Reports **2025** (2025).
 - [11] C. R. Harris, K. J. Millman, S. J. van der Walt, R. Gommers, P. Virtanen, D. Cournapeau, E. Wieser, J. Taylor, S. Berg, N. J. Smith, et al., *Nature* **585**, 357 (2020).
 - [12] T. pandas development team, *Pandas-dev/pandas: Pandas*, Zenodo (2025).
 - [13] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, et al., *Journal of Machine Learning Research* **12**, 2825 (2011).
 - [14] J. D. Hunter, *Computing in Science & Engineering* **9**, 90 (2007).
 - [15] M. L. Waskom, *Journal of Open Source Software* **6**, 3021 (2021).
 - [16] O. Mostafa, *Breast Cancer Prediction ($f1 = 0.99$, $AUC = 0.99$)*, <https://kaggle.com/code/omarmostafataha/breast-cancer-prediction-f1-0-99-auc-0-99>.
 - [17] <https://github.uio.no/larsbog/FYSSTK-Project2>
 - [18] Solely used for data splitting and performance metrics.
 - [19] The x values were expanded into multiple of features containing polynomial multiples x^n of themselves. For optimal performance $n \in \{0, 1, \dots, \geq 10\}$ were required.

Pairwise Hyperparameter Relationships (AUC)

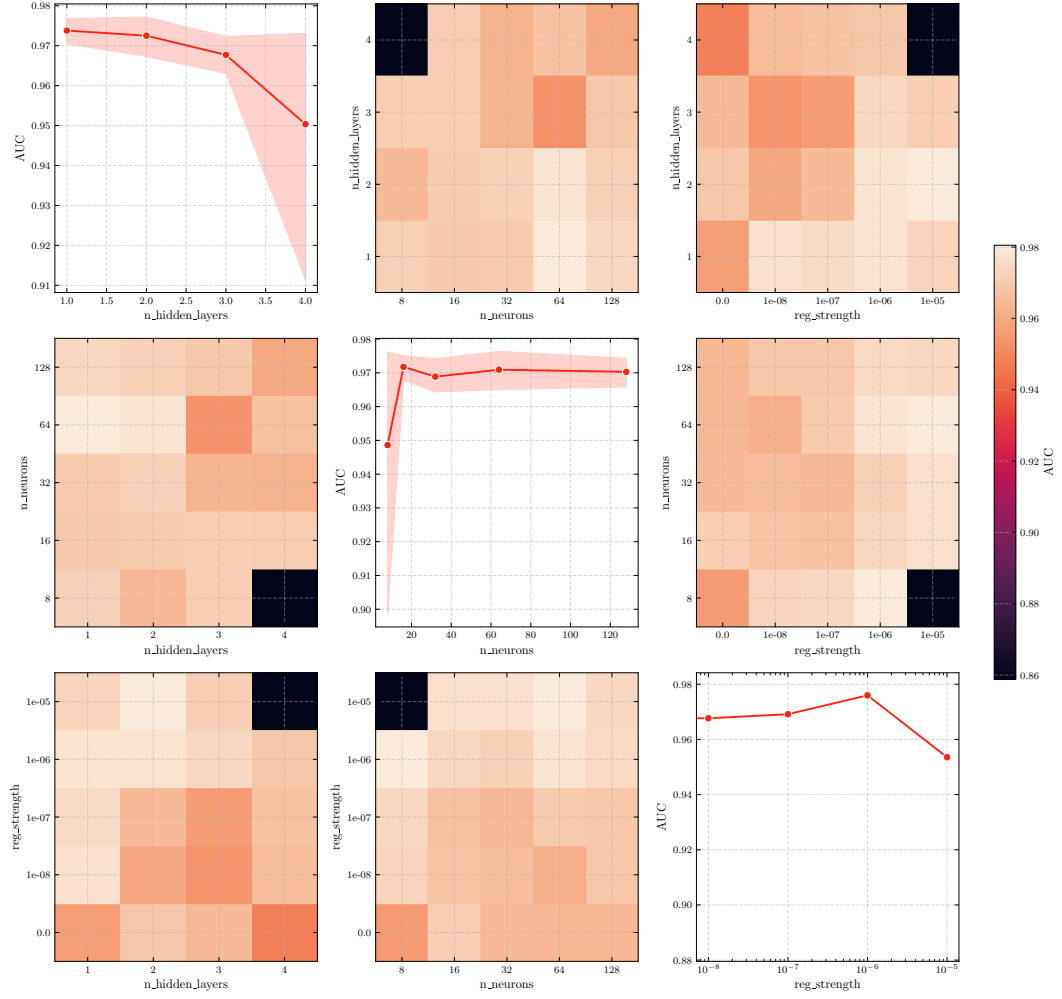


Figure 9: Area-under-Curve score dependance on the number of hidden layers, the number of nodes per hidden layer and the regularization strength λ for a L2 regularization. The off-diagonal plots show the average AUC per combination of two parameter values. The diagonal values show the average score per single parameter value. Optimization using Adam with $\alpha = 0.05$ and 500 epochs. LReLU activation in the hidden layers, softmax activation in the output layer. Optimized on training cross-entropy loss.