ENHANCED BREAST CANCER PREDICTION WITH MACHINE LEARNING ALGORITHMS, ARTIFICIAL NEURAL NETWORK AND CONVOLUTIONAL NEURAL NETWORK

A report by:
Ayomikun Mosaku u2127901
Ilir Krijezi u1954234
Jithen Kranth Reddy Nareddy u2194105
Wendy Brown u1514082

The University of East London

Abstract: This study delves into the integration of advanced computational methods for early detection and diagnosis of breast cancer, focusing on invasive ductal carcinoma (IDC), the most common form of breast cancer in women. Utilising the Breast Cancer Wisconsin (Original) dataset and Breast Histopathology Images, the research explores machine learning (ML) algorithms, artificial neural networks (ANNs), and Convolutional Neural Networks (CNNs) for breast cancer prediction. These methods excel in analysing complex medical data and mammogram imaging. The study encompasses data preprocessing, implementation of non-parametric models like Decision Trees and KNN, and parametric models like ANNs and CNNs using the MobileNetV2 architecture. Figures such as distribution plots, box plots, correlation heatmaps, and pair plots provide insights into the dataset's characteristics. The models' performance is depicted through accuracy scores, confusion matrices, and loss plots. The results highlight a high accuracy rate in breast cancer prediction with the ANN reaching an accuracy of 98.5%, demonstrating the significant potential of computational methods in early cancer detection, crucial for effective treatment.

1. Introduction:

Breast cancer is the most common form of cancer in women, and invasive ductal carcinoma (IDC) is the most common form of breast cancer. Accurately identifying and categorising breast cancer subtypes is an important clinical task, and automated methods can be used to save time and reduce errors.

This topic delves into the integration of advanced computational methods for the early detection and diagnosis of breast cancer, which is crucial in improving patient outcomes.

Machine Learning algorithms have been increasingly used in healthcare for their ability to learn from and make predictions based on data. In breast cancer prediction, these algorithms can analyse complex medical data, identifying patterns that might be indicative of cancerous growths. Techniques such as decision trees, support vector machines, and logistic regression are commonly explored.

Artificial Neural Networks, a subset of ML, mimic the functioning of the human brain to process information. ANNs are particularly effective in pattern recognition, making them ideal for analysing medical imaging data such as mammograms. They can discern subtle differences in tissue density and structure, which might indicate the presence of tumours.

Convolutional Neural Networks, a specialised neural network, are mighty in image processing and analysis. CNNs are adept at handling the vast amounts of data present in medical images, extracting features that are essential for accurate breast cancer diagnosis. This includes identifying tumour size, shape, and other morphological features.

The integration of these technologies enhances the accuracy of breast cancer predictions and aids in early detection, which is critical for effective treatment. The application of ML, ANNs, and CNNs in breast cancer prediction represents a groundbreaking approach that combines the fields of medicine and computer science to combat one of the most prevalent cancers globally.

2. Description of Dataset:

The Breast Cancer Wisconsin (Original) dataset is a collection of samples provided by Dr. Wolberg, who reports his clinical cases periodically. This dataset is unique as it chronologically groups the data based on the time of reporting. Below is a summary of the dataset's chronological grouping, which is not included in the data itself:

Group 1: 367 instances (January 1989)
Group 2: 70 instances (October 1989)
Group 3: 31 instances (February 1990)
Group 4: 17 instances (April 1990)
Group 5: 48 instances (August 1990)
Group 6: 49 instances (Updated January 1991)
Group 7: 31 instances (June 1991)
Group 8: 86 instances (November 1991)
Total: 699 data points (as of the donation date on 15 July 1992)

Note: The original Group 1 had 369 instances, but 2 were removed, leaving 367 instances. The dataset underwent revisions, including replacing and removing specific data points and modifying values in certain fields.

Columns Description:
The dataset comprises several features, each rated on a scale of 1 to 10, except for the class labels:

- Clump Thickness: Measures the thickness of the clump of breast cancer cells.
- Uniformity of Cell Size: Evaluates the uniformity in the size of the cancer cells.
- Uniformity of Cell Shape: Assesses the uniformity in the shape of the cancer cells.
- Marginal Adhesion: Measures cell adhesion at the edges.
- Single Epithelial Cell Size: Indicates the size of individual epithelial cells.
- Bare Nuclei: Refers to nuclei not surrounded by cytoplasm. This column contains missing values.
- Bland Chromatin: Describes the appearance of the nucleus in cells.
- Normal Nucleoli: Nucleoli are small structures in the nucleus.
- Mitoses: Measures the rate of cell division.

Class: Classifies the breast cancer case as benign (2) or malignant (4).
Class Labels:

2: Benign - Indicates a less aggressive breast cancer case that's less likely to invade surrounding tissues or spread.
4: Malignant - Signifies a more aggressive breast cancer case with a higher likelihood of spreading or recurring after treatment.

For the CNN, Breast Histopathology Images were used. The original dataset consisted of 162 whole-mount slide images of Breast Cancer (BCa) specimens scanned at 40x. From that, 277,524 patches of size 50 x 50 were extracted

(198,738 IDC negative and 78,786 IDC positive). Each patch’s file name is of the format: u_xX_yY_classC.png — > example 10253_idx5_x1351_y1101_class0.png.

Where u is the patient ID (10253_idx5), X is the x-coordinate of where this patch was cropped from, Y is the y-coordinate of where this patch was cropped from, and C indicates the class where 0 is non-IDC and 1 is IDC.

The original files are located here: http://gleason.case.edu/webdata/jpi-dl-tutorial/IDC_regular_ps50_idx5.zip

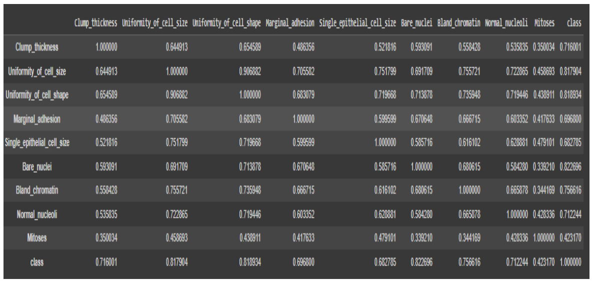
2.1. Data Analysis

The descriptive statistics of the dataset reveal the central tendency and dispersion of each feature. Most features have a scale from 1 to 10, indicating varying degrees of severity or prominence in cell characteristics. There are 16 missing values in the 'Bare Nuclei' column, which must be addressed in further preprocessing steps.

The target variable 'class' has two unique values: 2 (benign) and 4 (malignant). There are 458 benign cases and 241 malignant cases in the dataset. This distribution indicates that the dataset is somewhat imbalanced, with more benign cases than malignant ones. From this analysis, we can define the nature of the problem:

The problem is a classification problem, as the goal is to predict whether a given set of breast cancer cell characteristics corresponds to a benign or malignant tumour. There are two classes in the dataset: benign (class label 2) and malignant (class label 4).

This analysis was done with Google Colab and the link to the code is attached below: https://colab.research.google.com/drive/1H7NLoEopNhXAXBvxxRhjuC5I24qloVcp?usp=sharing



	Clump_thickness	Uniformity_of_cell_size	Uniformity_of_cell_shape	Marginal_adhesion	Single_epithelial_cell_size	Bare_nuclei	Bland_chromatin	Normal_nucleoli	Mitoses	class
Clump_thickness	1.000000	0.644913	0.654589	0.486356	0.521816	0.593091	0.558428	0.535835	0.350034	0.716001
Uniformity_of_cell_size	0.644913	1.000000	0.906882	0.705582	0.751799	0.691709	0.755721	0.722865	0.458693	0.817904
Uniformity_of_cell_shape	0.654589	0.906882	1.000000	0.683079	0.719668	0.713878	0.735948	0.719446	0.438911	0.818934
Marginal_adhesion	0.486356	0.705582	0.683079	1.000000	0.599599	0.670648	0.666715	0.603352	0.417633	0.696800
Single_epithelial_cell_size	0.521816	0.751799	0.719668	0.599599	1.000000	0.585716	0.616102	0.628881	0.479101	0.682785
Bare_nuclei	0.593091	0.691709	0.713878	0.670648	0.585716	1.000000	0.680615	0.584280	0.339210	0.822696
Bland_chromatin	0.558428	0.755721	0.735948	0.666715	0.616102	0.680615	1.000000	0.665878	0.344169	0.756616
Normal_nucleoli	0.535835	0.722865	0.719446	0.603352	0.628881	0.584280	0.665878	1.000000	0.428336	0.712244
Mitoses	0.350034	0.458693	0.438911	0.417633	0.479101	0.339210	0.344169	0.428336	1.000000	0.423170
class	0.716001	0.817904	0.818934	0.696800	0.682785	0.822696	0.756616	0.712244	0.423170	1.000000

Table 2.1 Description of the dataset

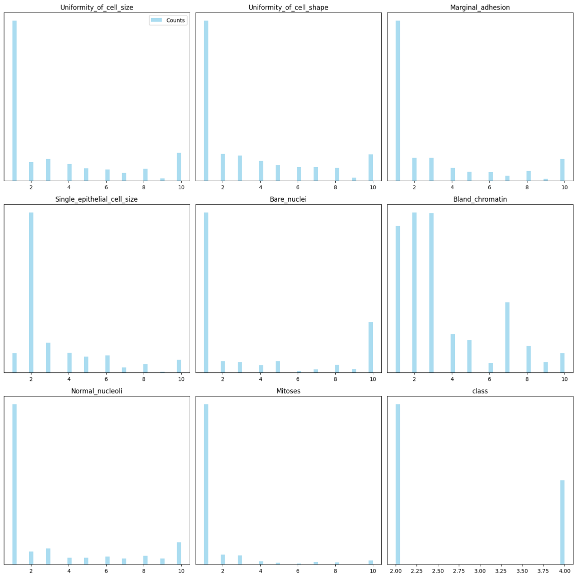


Fig 2.1: Plot of the distribution across all columns

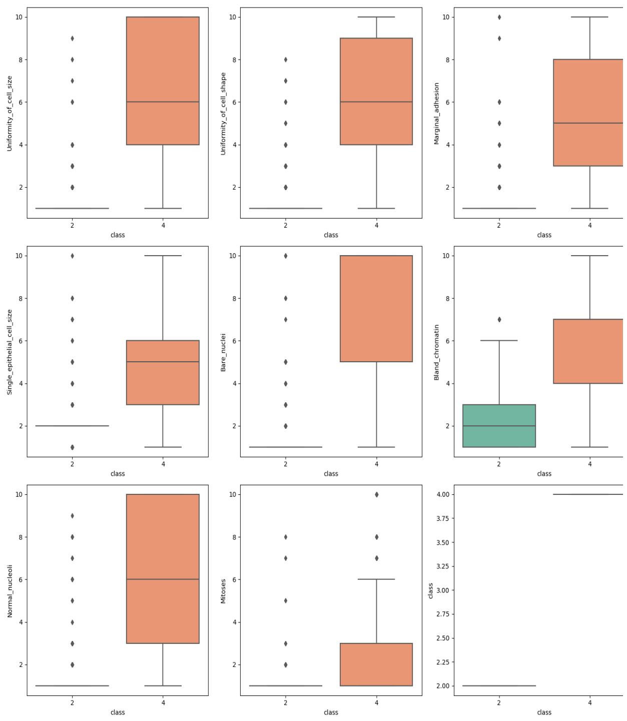


Fig 2.2: Box Plot of all the variables in the dataset

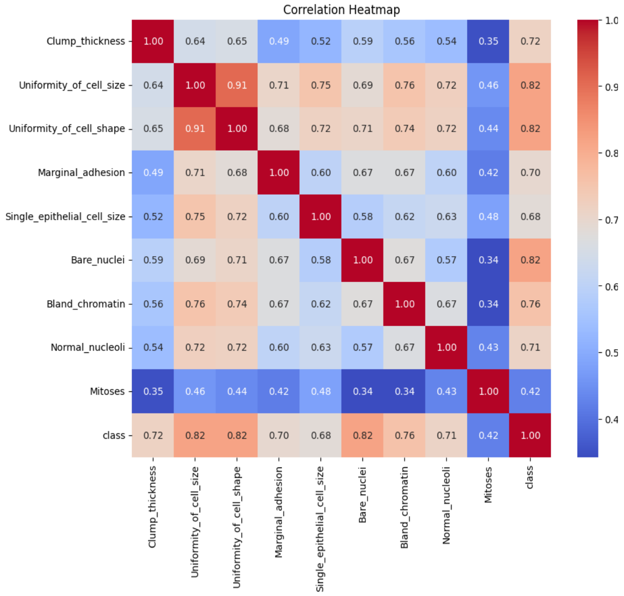


Fig 2.3: Correlation HeatMap for the Dataset

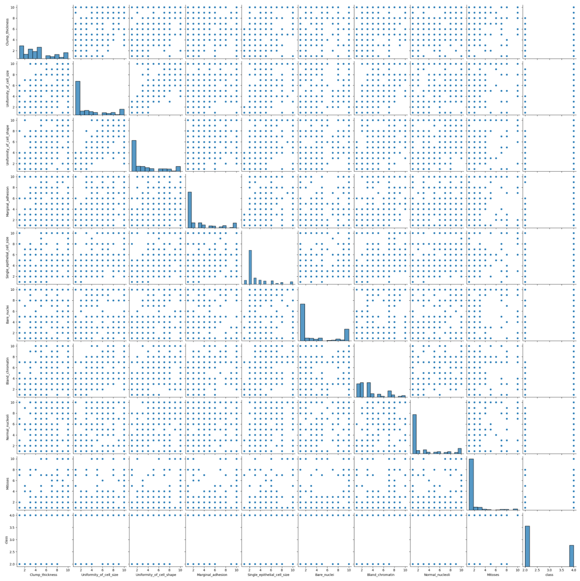


Fig 2.4: Pairplot for the Dataset

3. Model Building

In the endeavour to predict breast cancer using the provided dataset, a comprehensive and diverse range of models were employed, each with its unique strengths and capabilities. The models used encompass a wide spectrum from simple linear classifiers to more complex ensemble and deep learning methods.

Data Preprocessing

The data was preprocessed by handling all the missing values, changing the datatypes where necessary and splitting into train and test sets with over 70% of the data used for training.The links to all the codes is as follows:

Model building:https://colab.research.google.com/drive/1shOuat_knCP5TqYMsZ_Wf755hq_vs6qB#scrollTo=qrj1gKFau3it

CNN:

https://colab.research.google.com/drive/17ztjmkuYeemRzyjB7LQbW9Y6pDxg7Wiu#scrollTo=TU20nIXsyN3p

3.1 Non-Parametric Model

A non-parametric model is a type of model used in statistics and machine learning that does not assume any specific form for the relationship between independent and dependent variables. Unlike parametric models, which are characterised by a finite set of parameters and a predetermined functional form, non-parametric models are more flexible as they do not require the underlying distribution or structure to be defined in advance.

Decision Tree and KNN was used and they got an accuracy of 98% and 97% respectively.

3.2 Parametric Model: Artificial Neural Network.

The Artificial Neural Network (ANN) for the breast cancer prediction task was built in a Python Notebook using TensorFlow and Keras, and its construction involved several key decisions regarding its architecture and configuration. Here’s a detailed description of the process:

Number of Layers and Neurons:

The ANN comprises three layers: two hidden layers and one output layer.

First Hidden Layer: It consists of 32 neurons. The choice of 32 neurons is a balance between having enough capacity to learn complex patterns in the data and avoiding overfitting. This layer size is often sufficient for a dataset of moderate complexity, like the one used here.

Second Hidden Layer: Also equipped with 32 neurons, this layer helps in capturing deeper interactions between features which might not be possible with a single hidden layer. The same number of neurons as the first layer maintains consistency and provides ample model capacity.

Output Layer: The network concludes with a single neuron, as it’s a binary classification task (predicting benign or malignant). The single neuron outputs a value between 0 and 1, representing the probability of the input being classified as one of the two classes.

Activation Functions:

The hidden layers use the ReLU (Rectified Linear Unit) activation function. ReLU is chosen for its efficiency and effectiveness in non-linear transformations without being computationally expensive.

The output layer uses the sigmoid activation function. This is because sigmoid is suitable for binary classification, transforming the neuron's output into a probability score (between 0 and 1).

Initial Weights Configuration:

In this ANN, the initial weights were not explicitly set, meaning they were automatically initialised by Keras. By default, Keras uses a variant of the 'Glorot uniform' initializer, also known as Xavier uniform initializer.

This initializer draws weights from a distribution with a zero mean and a specific variance. For ReLU activation functions, a variant called 'He initialization' is often preferred, but the default initializer generally works well for moderate-sized networks.

Additional Configuration:

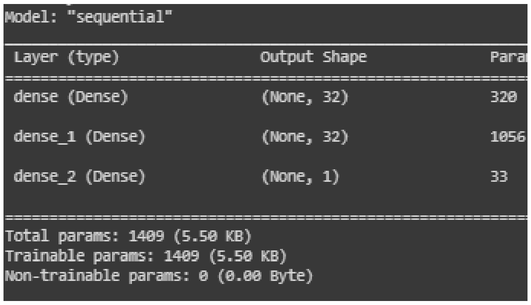
Optimizer: The Adam optimizer was used. Adam is widely used due to its adaptive learning rate capabilities, making it more efficient in converging to a minimum.

Loss Function: T he binary_crossentropy loss function was chosen, which is standard for binary classification problems.

Metrics: Accuracy was the metric for evaluating the model's performance.

Training Strategy:

An Early Stopping callback was implemented to prevent overfitting. This stops training when the validation loss hasn't improved for a set number of epochs (10 in this case), making the training process more efficient.



```
Model: "sequential"
________________________________________________________________
 Layer (type)                Output Shape              Para
================================================================
 dense (Dense)               (None, 32)                320

 dense_1 (Dense)             (None, 32)                1056

 dense_2 (Dense)             (None, 1)                 33

================================================================
Total params: 1409 (5.50 KB)
Trainable params: 1409 (5.50 KB)
Non-trainable params: 0 (0.00 Byte)
```

Fig. 4.1 Model Architecture

3.3 Convolutional Neural Network

This CNN utilises the MobileNetV2 architecture as a base model and performs additional customizations for the specific task. Here’s a detailed explanation of each key step:

Data Loading and Preprocessing:

The dataset, which contains images, is organised into separate folders for training (train_dir), validation (val_dir), and testing (test_dir).

ImageDataGenerator is used for loading and preprocessing images. It rescales the images by normalising pixel values to the range [0, 1]. Additional data augmentation could be included here, but it's not in the current code.

The data generators for training, validation, and testing sets are created with a target size of 224x224 pixels (standard input size for MobileNetV2) and a batch size of 32. The class mode is set to 'binary', indicating a binary classification task.

MobileNetV2 as Base Model:

MobileNetV2, a lightweight and efficient deep learning model, is loaded with pre-trained weights from ImageNet (a large image database). This technique, known as transfer learning, leverages pre-learned patterns on a large dataset to improve performance on a smaller dataset.

The include_top=False argument is used to exclude the top (final) layers of the MobileNetV2, as these layers are specific to ImageNet classes and not relevant for the breast cancer classification task.

Custom Top Layers:

After the base MobileNetV2 layers, custom layers are added for the specific task:

AveragePooling2D reduces the spatial dimensions (height and width) of the output from the base model.

Flatten transforms the 2D pooled features into a 1D vcector.

Dropout layers are inluded to reduce overfitting by randomly setting a fraction of input units to 0 during training.

A Dense layer with 128 units and ReLU activation is used to learn non-linear combinations of features.

The final Dense layer has 1 unit with a sigmoid activation function, suitable for binary classification.

Model Compilation:

The model is compiled with the Adam optimizer and a learning rate of 0.001.

The loss function is binary_crossentropy, ideal for binary classification tasks.

The metric used for performance evaluation is accuracy.

Training with Callbacks:

ModelCheckpoint is used to save the model after each epoch where there is an improvement in validation accuracy.

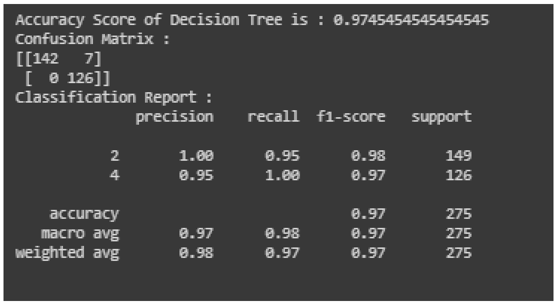
EarlyStopping is employed to stop training if the validation accuracy doesn't improve for 5 consecutive epochs, preventing overfitting.

The model is trained for up to 20 epochs.

Model Evaluation and Saving:The model’s performance is evaluated on the test set.

The CNN had an accuracy of 87% after 17 epochs.

4. Results:

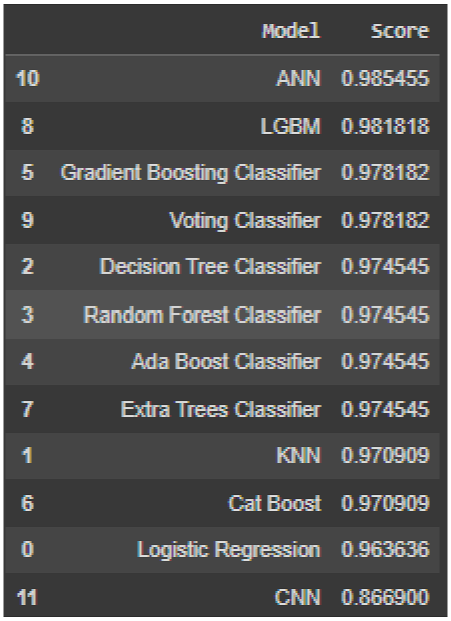


```
Accuracy Score of Decision Tree is : 0.9745454545454545
Confusion Matrix :
[[142   7]
 [  0 126]]
Classification Report :
              precision    recall  f1-score   support

           2       1.00      0.95      0.98       149
           4       0.95      1.00      0.97       126

    accuracy                           0.97       275
   macro avg       0.97      0.98      0.97       275
weighted avg       0.98      0.97      0.97       275
```

Table 4.1. Confusion Matrix for Decision Trees



	Model	Score
10	ANN	0.985455
8	LGBM	0.981818
5	Gradient Boosting Classifier	0.978182
9	Voting Classifier	0.978182
2	Decision Tree Classifier	0.974545
3	Random Forest Classifier	0.974545
4	Ada Boost Classifier	0.974545
7	Extra Trees Classifier	0.974545
1	KNN	0.970909
6	Cat Boost	0.970909
0	Logistic Regression	0.963636
11	CNN	0.866900

Table 4.2. Accuracy scores of the Models

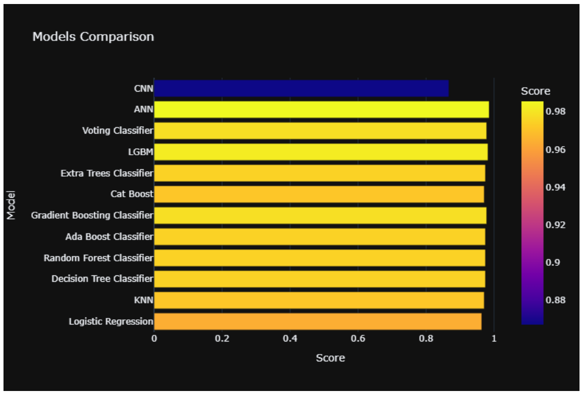


Fig 4.1. Plot of Accuracy scores of the Models

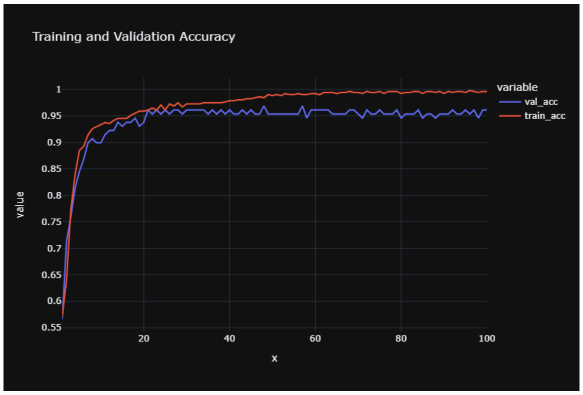


Fig 4.1. Accuracy curve of the ANN

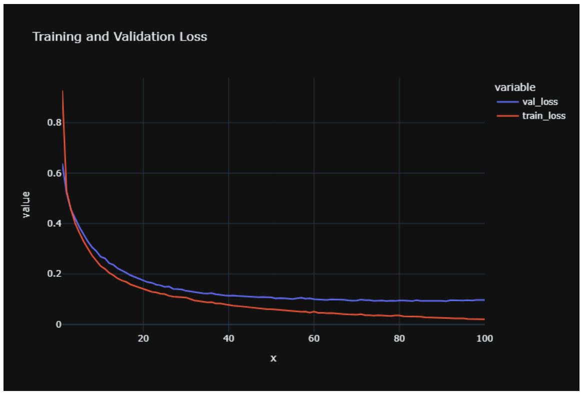


Fig 4.1. Loss plot of the ANN.

The models had an accuracy of 85% and above with the ANN having the highest accuracy of 98.5%.

5. Conclusion

This research demonstrates the effectiveness of machine learning and neural network models in the early detection and diagnosis of breast cancer. The main challenges faced included handling missing data, balancing the dataset, and optimising model parameters to avoid overfitting. If provided more time, further exploration into more complex models, additional feature engineering, and incorporating larger datasets for a more comprehensive analysis would be pursued. The study employed various computational methods, including Decision Trees, KNN, ANNs, and a CNN with MobileNetV2 architecture, achieving high accuracy, especially with the ANN model at 98.5%. These findings underscore the importance of technological advancements in healthcare, particularly in breast cancer subtype identification and diagnosis accuracy. The application of CNNs in medical imaging analysis reaffirms the potential of deep learning techniques in medical diagnostics. Overall, this research contributes significantly to medical technology, aiming to enhance patient outcomes through early and precise cancer detection. Figures depicting the dataset characteristics and models' performance metrics are integral to understanding the study's thoroughness and efficacy.

References:

Zuo, D., Yang, L., Jin, Y., Qi, H., Liu, Y., & Ren, L. "Machine learning-based models for the prediction of breast cancer recurrence risk." BMC Medical Informatics and Decision Making, Volume 23, Article number: 276, 2023. URL: BMC Medical Informatics and Decision Making.

Shen, L., Margolies, L. R., Rothstein, J. H., Fluder, E., McBride, R., & Sieh, W. "Deep Learning to Improve Breast Cancer Detection on Screening Mammography." Scientific Reports, Volume 9, Article number: 12495, 2019.

URL: Scientific Reports.

Hu, Q., Whitney, H. M., & Giger, M. L. "A deep learning methodology for improved breast cancer diagnosis using multiparametric MRI." Scientific Reports. Volume 10, Article number: 10536 (2020).