**Deep Learning Approach**

The limitation of WBCB dataset is having missing data, small dataset size, various ranges of data values, unbalanced data and skewed data. The new method of deep learning adopts multiple (three) hidden layers of deep neuron network [16]. The inputs which are the attribute values computed from digitized image of FNA, while the outputs from the model are the two classification diagnosis classes, B(benign) or M(malignant) [16].

The first hidden layers used 9 neurons and the last hidden layer uses 1 neuron to match the input and output numbers, respectively [16]. The middle layer consists of 5 neurons. The dropout layers with dropout rate of 50% are added between hidden layers to avoid overfitting and improve the model’s performance in accuracy [16].

The rectified linear unit (ReLU) activation function relu is applied in hidden layer 1 and 2 [16]. The sigmoid activation function is applied in the final hidden layer 3 to create continuous output, a threshold of 0.5 is used [16]. The binary cross entropy loss function is selected because the WBCD classification is a binary classification problem [16]. The Adam optimizer function is selected as it is a replacement optimization algorithm for stochastic gradient descent (SGD) [16]. The optimization function is used to train the deep learning model [16].

The unbalanced data issue is coped by producing new data samples which labeled as ‘malignant’. Also, ‘malignant’ sample set is doubled by introducing normal distributed random noise to the existing ‘malignant’ samples [16]. The noise generation is introduced by applying the *Numpy.random* number generation library which is *Numpy.random.normal(mean, sigma, feature. shape)* to produce an array of normal distributed random numbers with the same attribute dimensions, where mean = 0 and standard deviation sigma = 0.1 [16]. Then, the created array of random numbers is incorporated into the dataset element by element to produce a new dataset [16]. The column of “class” labels is reused to label the samples (feature vectors) in the new produced dataset. The labels are reused by combining the new produced array of feature values with “class” column from the given dataset into a new dataset [16].

The original dataset of 699 samples is expanded to a new dataset of 1398. 10-fold cross-validation is performed with 500 epochs in order to verify the robustness of the new deep learning model.

The table below shows that the performance in accuracy of the model is stable in the aspect of dataset changes.

|  |  |
| --- | --- |
| **Cross Validation No** | **Prediction Accuracy** |
| 1 | 100.00 |
| 2 | 100.00 |
| 3 | 97.14 |
| 4 | 91.43 |
| 5 | 96.19 |
| 6 | 98.10 |
| 7 | 94.29 |
| 8 | 98.10 |
| 9 | 98.08 |
| 10 | 98.08 |
| Average Score: | 97.14 (+/- 2.48) |

Table: Prediction Accuracy score by cross validation using Deep Learning Neural Network

**Model Performance Visualization**

To verify the performance in accuracy of the new supervised deep learning model, in the case of splitting the dataset into two parts (i.e., 75% for model training and 25% for model testing), Figure A shows the prediction accuracy history of the new model in model training, while Figure B shows the loss history of the new model in model training. Figures A indicates that the performance in accuracy of the new model quickly increased and stabilized to about 85-90%, while the loss history decreased and stabilized to about 0.25.

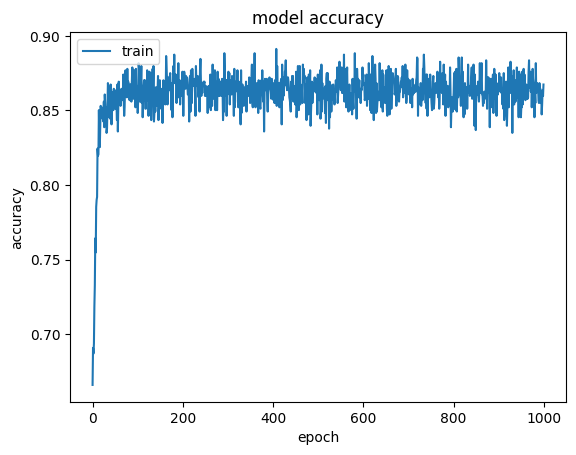


Figure A: Accuracy history of the new deep learning model in model training

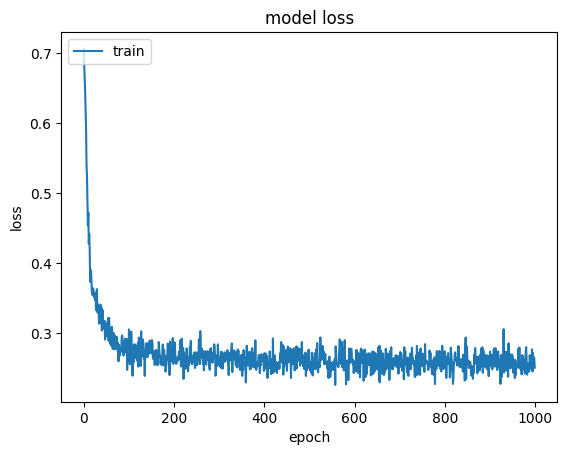


Figure B: Loss history of the new deep learning network model in model training

Reference

1. https://towardsdatascience.com/deep-learning-in-winonsin-breast-cancer-diagnosis-6bab13838abd