Breast Cancer Detection using Shallow Neural Network

Introduction

The aim of this assignment was to develop a Shallow Neural Network model which can be used to classify patients into two categories: those diagnosed with breast cancer and those who are not. TensorFlow and Keras were used to create a shallow neural network comprised of one hidden layer for a binary classification model. The model was trained on the Breast Cancer Wisconsin (Diagnostic) dataset and evaluated based on accuracy, precision, recall, and visualization of the confusion matrix.

Methodology

Dataset Description

The Breast Cancer Wisconsin (Diagnostic) dataset contains samples from patients with features computed from a digitized image of fine needle aspirate (FNA) of a breast mass. It includes 569 samples with 30 features that describe characteristics of the cell nuclei present in the image.

Data Pre-processing

To improve accuracy and reliability of the model preprocessing steps were completed before bringing the dataset into the model. The features were standardized to have zero mean and unit variance.

```
scaler = StandardScaler()
X_trans = scaler.fit_transform(X)
```

The labels “Malignant” and “Benign” were then encoded to 1 and 0 respectively to allow for binary classification.

```
label_encoder = LabelEncoder()
y_trans = label_encoder.fit_transform(y)
```

Finally, the data was split into training and test sets with an 80 to 20 ratio to allow for evaluation and validation cycles as well as a random state to allow for reproducibility.

```
X_train, X_test, y_train, y_test = train_test_split(
    X_trans, y_trans, test_size=0.2, random_state=17)
```

Model

The model used is a shallow neural network comprised of one hidden layer made using TensorFlow and Keras. The single layer was chosen due to more simplistic nature of a binary classification task as well as the small data set. The hidden layer contains the Rectified Linear Unit (ReLu) activation function which was chosen to introduce non-linearity and express any complexities available within the data. For the output layer the sigmoid activation function was selected to as it produces probability scores which can be used for facilitating binary classification.

```
model = Sequential()
model.add(Dense(16, input_dim=X_train.shape[1], activation='relu'))
model.add(Dense(8, activation='relu'))
model.add(Dense(1, activation='sigmoid')) # as it is binary classification
model.add(Dropout(rate=0.1))
```

The model was compiled with binary cross-entropy loss and the Adam optimizer.

```
model.compile(optimizer='adam', loss='binary_crossentropy', metrics=['accuracy'])
```

Hyperparameters

Hyperparameters are crucial for tuning and improving model performance. The parameters of interest to this model are the number of neurons in the hidden layer, number of epochs, and the batch size. As well as evaluation metrics and validation split. The values for the hyperparameters are below.

Hyperparameter	**Value**
Number of neurons in the hidden layer	16
Number of epochs	150
Batch size	32
Evaluation Metric	Accuracy
Validation Split	0.1

The number of neurons was largely based on the original model discussed in the Breast Cancer paper. However, this relatively small number can be useful for preventing overfitting and promotes better generalization. The number of epochs was selected largely based on conversations had in class, where there was a statement regarding the ineffectiveness of anything below 150. The performance of the model did increase when raising the epochs from 50 to 150 so there is empirical evidence to support the selection. The batch size was chosen as a balance between the frequent updates but greater noise of a smaller value and the more stable updates but slower convergence of larger values. Accuracy was chosen as the evaluation metric as we are concerned with the diagnostic ability of the dataset. The validation split is a standard value used for training and validation.

```
model.fit(X_train, y_train,
          epochs=150, batch_size=32,
          validation_split=0.1, verbose=1)
```

Evaluation Metrics

To evaluate the efficacy of the model several metrics were used. Accuracy was used to identify the percentage of correctly classified samples. Precision to identify the percentage of true positive predictions among all positive predictions. Recall to identify the proportion of true positive predictions among all actual positive samples. A confusion matrix was also employed as a visual representation of the model's performance, showing the true positive, true negative, false positive, and false negative predictions.

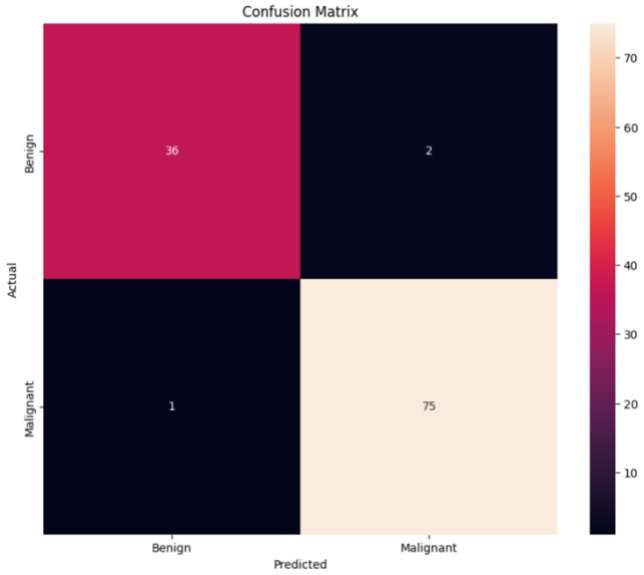
Results

The model results exhibit very good performance on the dataset, demonstrating high accuracy, precision, and recall. We can see below that all 3 metrics were 97% or greater.

Metric	**Score**
Accuracy	97.37%
Precision	97.40%
Recall	98.68%

These metrics indicate that the model accurately identifies patients with breast cancer and effectively minimizes false positives and false negatives. The high accuracy score reflects the overall correctness of the model's predictions, while precision highlights the model's ability to correctly identify true positive cases among all positive predictions. Additionally, the high recall score signifies the model's capability to capture a large proportion of true positive cases among all actual positive samples.

The confusion matrix shows there are 36 true negative benign cases, 2 false positive malignant cases, 41 true positive malignant cases, and 1 false negative malignant case.



Conclusion

In conclusion, the developed Shallow Neural Network model exhibits very good performance in detecting and diagnosing breast cancer from the Breast Cancer Wisconsin dataset. With accuracy, precision, and recall scores all exceeding 97%, the model is capable of providing trustworthy binary classification of breast cancer diagnosis. As this is a diagnosis tool, the only improvement I would look towards is the eradication of any false negatives, even at the risk of more false positives. The identification of more cancer is more important than the accidental false negatives.

References

- Breast Cancer Wisconsin (Diagnostic)
- A Beginner's Guide to Shallow Neural Network
- Determining the Right Batch Size for a Neural Network to Get Better and Faster Results