Correlation Analysis and Feature Transformation for Breast Cancer Malignancy Diagnostics using Neural Networks.

Melek Mizher†

November 19, 2022

† Address to which correspondence should be addressed:

melekmizher@me.com

Abstract

Advancements in computational power and technologies have brought about a revolution in modern medical diagnostics. This study compares the usage of multiple statistical learning techniques to model Breast Cancer malignancy using cellular nuclei measurements. Providing medical professionals with tools to diagnose breast cancer malignancy with high accuracy can help improve prognosis. We compare multiple classical statistical classification methods and Deep Neural Networks (DNN) to improve patient outcomes. The highest accuracy was produced by careful parameter tuning to design Neural Network.

1.0 Introduction and Problem Statement

This study applies feature squaring and logarithmic transformations to the Breast Cancer Wisconsin dataset to create a system for Breast Cancer classification. Early classification of Breast Cancer is key to improved patient outcomes by ensuring that the necessary care is prompt and effective. A breast cancer diagnosis is made through a breast exam, mammogram, ultrasound, MRI, or through the extraction of breast mass through a biopsy (Mayo Clinic 2022). Extraction of a breast mass through a fine needle aspirate procedure can confirm the malignancy of the breast cancer. The process requires sample extraction, preparation, and cytological inspection to decide if the extracted mass is benign or malignant. The goal of this study is to identify the best possible classification model to diagnose Breast Cancer malignancy while eliminating the introduction of human error caused by qualitative cytological inspection of the samples. Multiple classification models were used to identify the highest classification accuracy of Breast Cancer malignancy using cellular nuclei measurements of breast masses extracted through fine needle aspirate

procedures.

2.0 Literature Review

The medical field contains multiple obstacles to developing and implementing Artificial Intelligence and Machine Learning programs. The most significant obstacle is interpretability. Professionals in the field demand simple explanations for classification algorithms. Geoffrey Hinton (2018) describes the irregularity of Neural Networks as they can be trained multiple times, and those different neurons can represent different learned features in the intermediate layers of the model as initial weights change. The lack of interpretability creates ethical and legal challenges that professionals and regulators have to face.

Artificial intelligence applications have grown due to the significant increase in computing power and the size of datasets being used to train them (Boris Roberts 2022). Multiple studies have been performed on the Wisconsin Diagnostic Breast Cancer (WBDC) dataset. Ashutosh Dubey et al. (2016) proposed a K-means Clustering approach to the Wisconsin dataset with an accuracy of 92%. Reddy Anuradha (2021) proposed the usage of a Support Vector Machine Classifier achieving an accuracy of 96.06% on the Wisconsin dataset. Roseline Ogundokun et al. (2022) created two models, one using SVMs with an accuracy of 96.5% and a Multi-Layer Perceptron with 97.2% accuracy.

Liu et al. (2018) propose a fully connected layer first (FCLF) CNN ensemble model of four base models. Liu et al. achieved an accuracy of 98.71% on the WBDC dataset. Al Shayeji et al. (2022) propose an ANN model that achieved 99.47% accuracy, only misclassifying three samples. Their best model used five-fold cross-validation and 100 neurons in a single hidden layer while maintaining all 30 features from the data.

3.0 Data

The data used for this study was the Breast Cancer Wisconsin Diagnostic data set obtained from Kaggle (2016) containing 569 samples, of which 357 are benign, and 212 are malignant. These samples were gathered using a fine needle aspirate of a breast mass as previously described. The data includes 30 features, including ten different cellular nuclei measurements, their standard errors, and the averages of the three most significant values, called the "worst" values. The features measured included nuclei radius, texture, perimeter, area, smoothness, compactness, concavity, concave points, symmetry, and fractal dimension. We performed essential feature selection by removing any values with a correlation lower than 25% to the diagnosis. This resulted in the removal of five features: Fractal Dimension Mean, Texture SE, Smoothness SE, Symmetry SE, and Fractal Dimension SE.

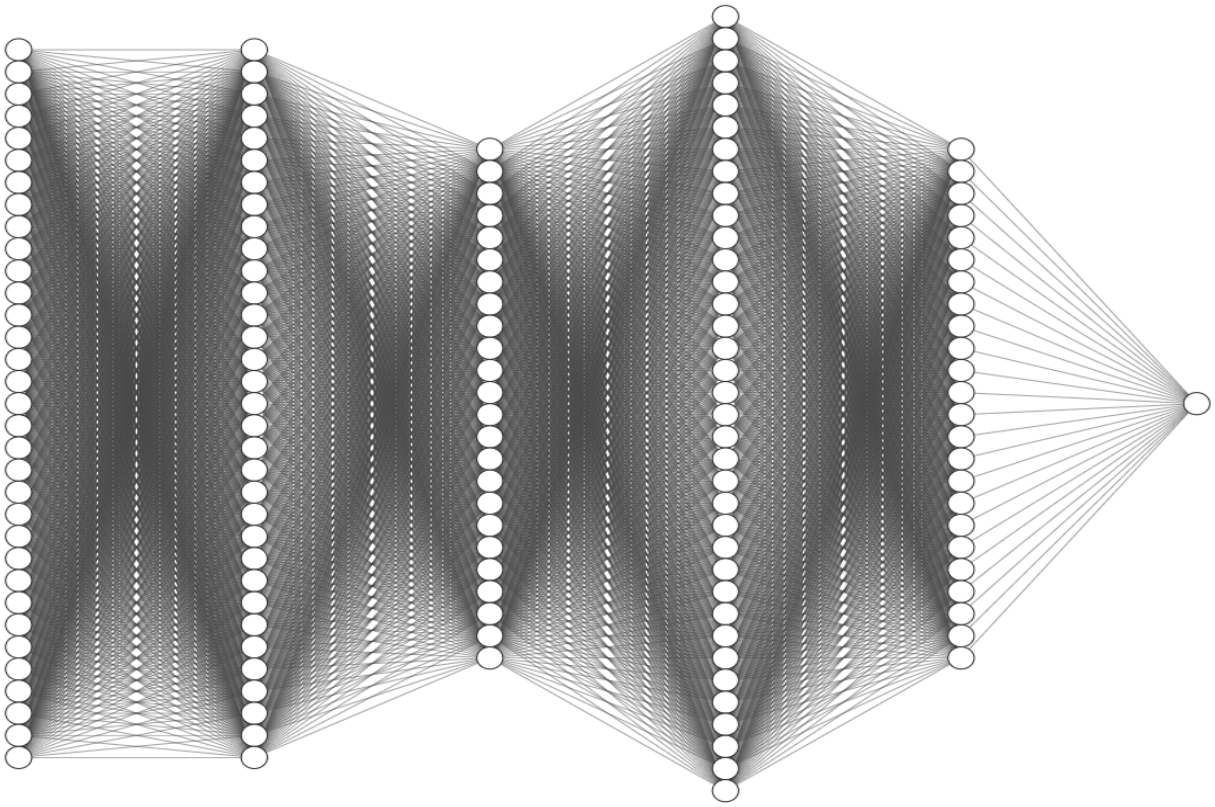
For the second model created, the data was transformed by finding the square and log-transformed values for all. Data that had values of 0 was replaced with 0.00001 as to allow the transformation of all values turning the dataset from 30 to 50 total features. All values that had correlations below 40% were then removed (including the features removed in the first model), creating a dataset with a total of 33 features.

4.0 Research Design and Modeling Methods

We removed some features based on the initial exploratory data analysis to simplify the model and provide a more straightforward input sequence for the classifiers. Model simplification led to a total dataset of 25-features. The dataset was standardized using MinMax Scaling of values from 0 to 1. The data was then split between training and test data using a 20% test data size. Multiple classification models were evaluated and compared, including K-Nearest Neighbors (KNN), Linear Support Vector Machines (LSVM), RBF Support Vector Machines (RBF SVM), Gaussian Processes, Decision Trees,

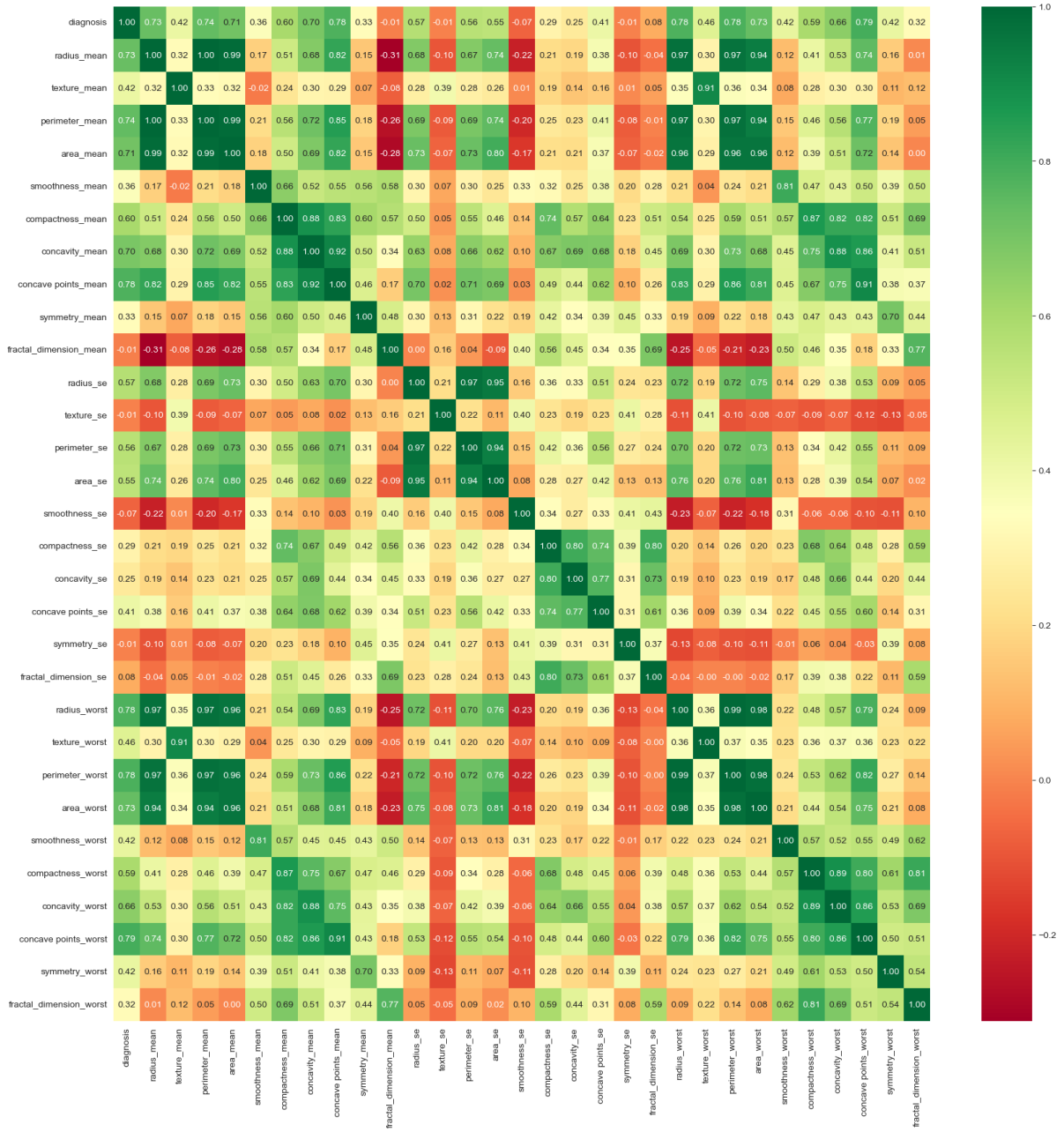
Random Forest (RF), Multi-Layer Perceptron Neural Networks (MLP-NN), AdaBoost Trees, Naive Bayes (NB), Quadratic Discriminant Analysis (QDA) and Deep Neural Networks (DNN) were scored to identify the highest classification accuracy for the breast cancer malignancy data. These models provide different advantages, some of which are best for a linear classification of data; hence they produced lower scores. During the data exploration, Neural Networks, Random Forests, AdaBoost Trees, and K-Nearest Neighbors were decided to be the best models to represent this problem due to their non-linearity.

Figure 1: Final Neural Network Design including 4 hidden layers. 33 input vectors and Dense layers with 33, 24, 36, and 24 neurons with ReLU activations leading to a final sigmoid activation for binary classification reached the best consistent results for a Deep Neural Network at 98.25% accuracy on the test data. Total of 151 neurons and 3633 weights.



5.0 Results

Figure 2: Data Correlation Matrix before low correlation feature removal from the initial model.



The preliminary results obtained by testing different classifiers showed accuracy results ranging from 93.86% for Decision Trees to the best results attained by Random Forest with a max depth of 5 and 10 estimators and a simple Multi-Layer Perceptron with an alpha value of 1, both of which achieved 98.25% accuracy. The proposed Deep Neural

Network achieved an average accuracy of 98.25%, which generates a model comparable to the best models explored in Table 1.

Figure 3: Preliminary results of accuracy scores for different classifiers.

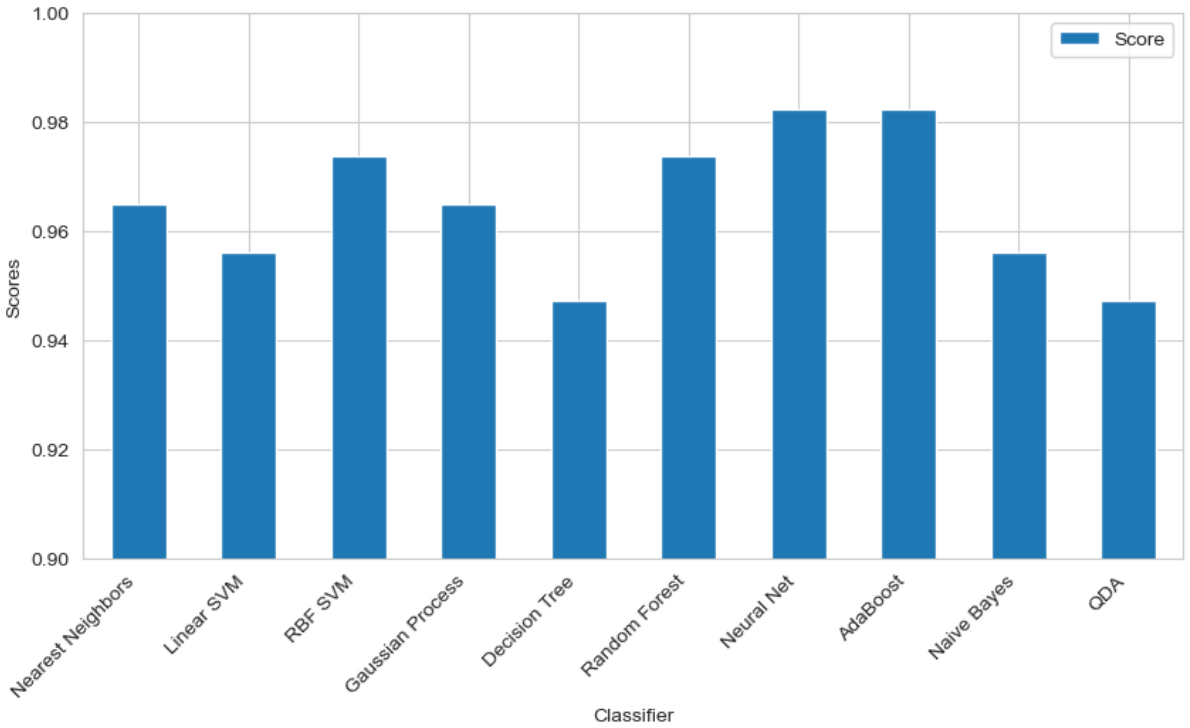


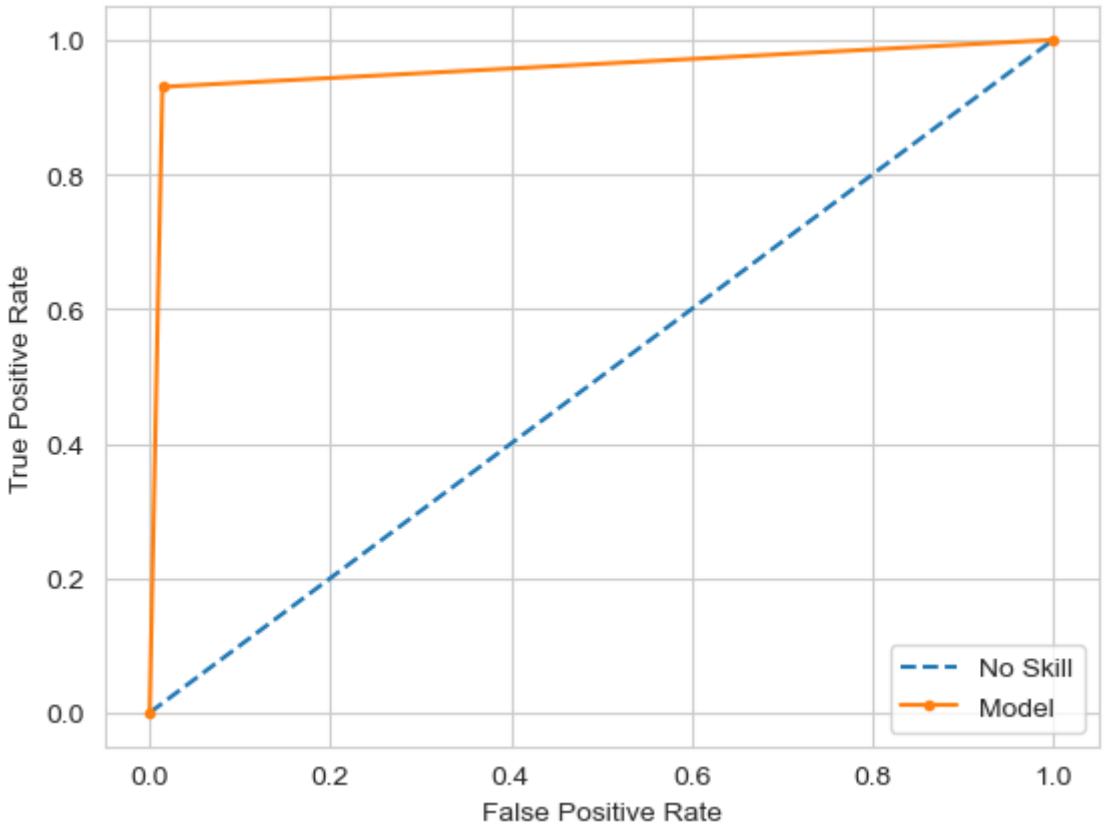
Table 1: Accuracy scores using different classifiers.

Classifier	Score
Nearest Neighbors	0.964912
Linear SVM	0.956140
RBF SVM	0.973684
Gaussian Process	0.964912
Decision Tree	0.947368
Random Forest	0.973684
Neural Net	0.982456
AdaBoost	0.982456
Naive Bayes	0.956140
QDA	0.947368

We evaluated the Deep Neural Network by running through epochs ranging from 30 to 40 with very similar accuracy scoring for each run. It is significantly accurate, but at the levels required to ensure better patient outcomes, it can fall short of expectations, understanding

that other, relatively more straightforward methods achieved similar and better results.

Figure 4: Initial Deep Neural Network ROC curve as compared to the no-skill model.



6.0 Analysis and Interpretation

We found the Concave Points Worst data to be the most significant factor in defining the difference between identifying benign and malignant tumors. Other significant factors include Area Mean and Texture Mean. Concavity Standard Error was the feature with the lowest significance in classifying sample malignancy.

Table 2: Important Cellular Nuclei measurement values as related to the type of Diagnosis in the dataset. Benign data points are 0, while Malignant data points are 1.

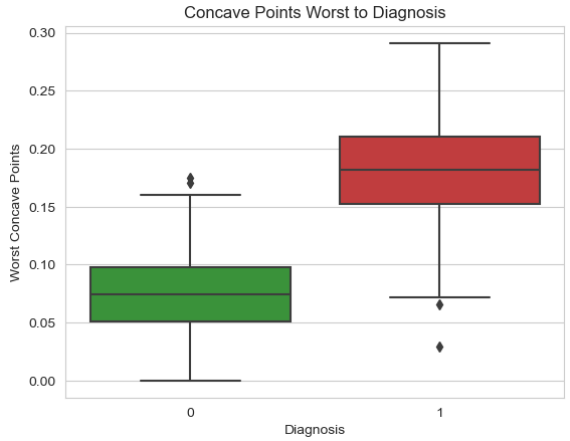
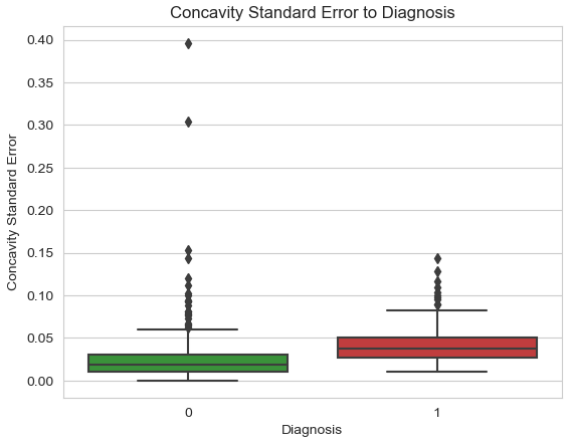
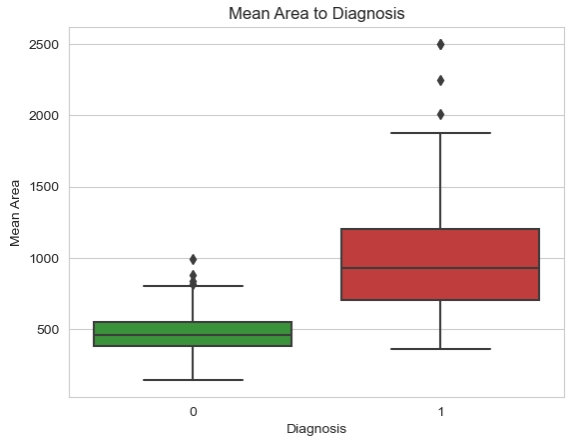
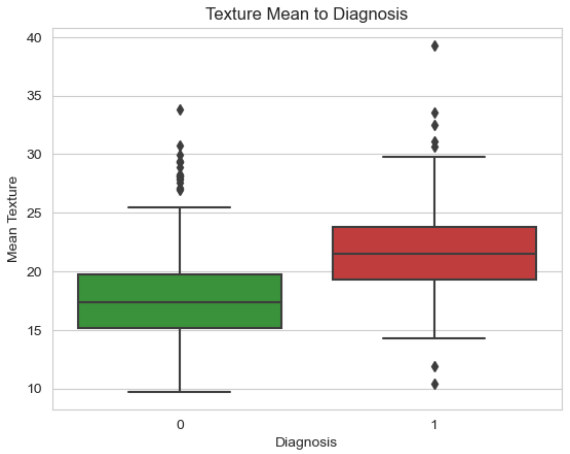
Sample Cellular Nuclei Evaluation as Related to Diagnosis	
Concave Points Worst	**Concavity Standard Error**
	
Mean Area	**Texture Mean**
	

Figure 7, included in the Appendix, displays a clustered heatmap that shows three main regions of interest. All the values related to the Perimeter, Radius, and Area of the samples show close similarities that become overrepresented in the algorithms. Concave Points Worst and Concave Points Mean were the most closely related values to the Diagnosis. Lastly, the five removed features are all confirmed by their associations to have a low relationship to the Diagnosis as they cluster together in the results while showing stark

differences to the Texture values, which were more significant to the Diagnosis although they formed similar clusters. Texture and Smoothness values are similar in their definitions, but they are disconnected and do not show any prevalent trends, as shown in Figure 5 in the Appendix. We can modify the network architecture based on the results from the proposed DNN. However, changes in the design will not generate significant improvements as other slight modifications of the architecture.

7.0 Conclusion

The most significant challenge when assessing the Breast Cancer diagnostic dataset is the relatively. The results from the different classification methods show a slight difference between linear and non-linear methods. We can improve by using more profound methods for feature selection, outlier removal, and data transformation before testing other DNN methods. These methods can represent data transformations that may affect data quality. The small differences between methods highlight the power of ANNs to generate weight combinations that exceed expectations, as in Al Shayeji et al. (2022), even if the data preparation steps are not extensive.

8.0 Directions for Future Work

The most important consideration for future work would be outlier evaluation and removal to ensure that the most extreme samples do not sway the models. Further hyperparameter tuning can also be performed to ensure optimal values for the Deep Neural Network. The dataset can be further simplified to ensure that other features that do not significantly affect diagnosis are removed or refined.

Appendix

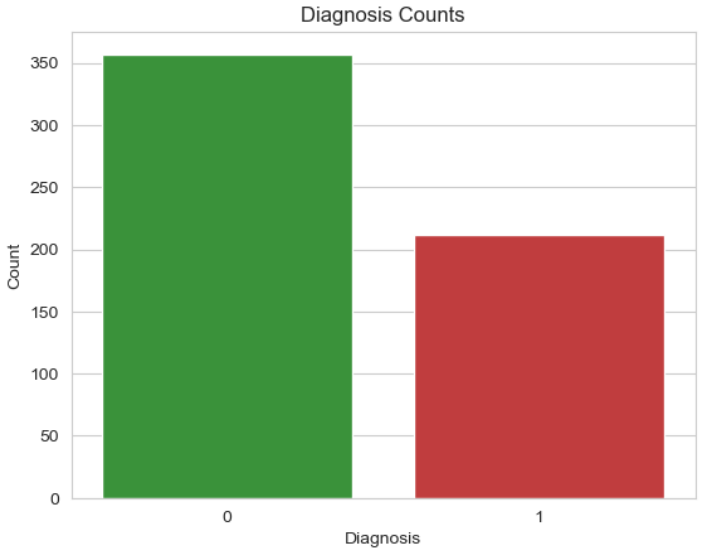


Figure 5: Dataset Diagnosis Counts. Benign data points are denoted as 0 while Malignant data points are denoted as 1.

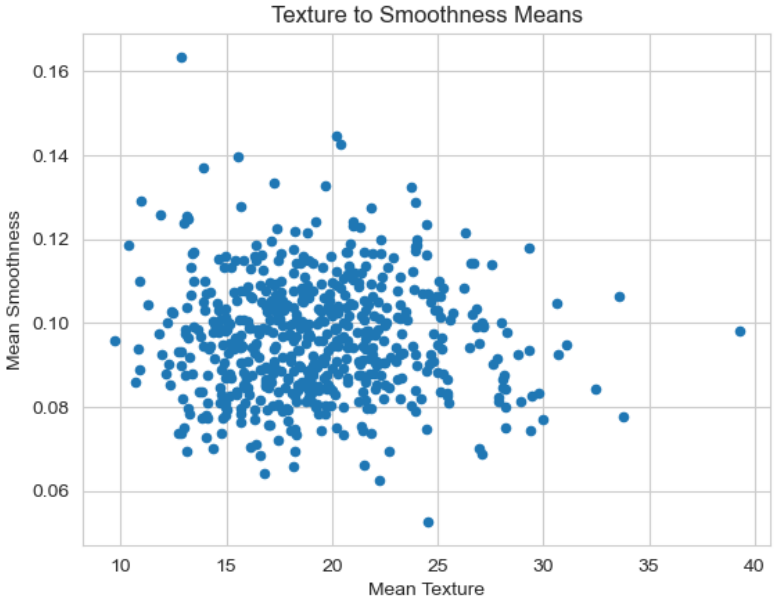


Figure 6: Comparison of mean data for nuclear texture and smoothness values.

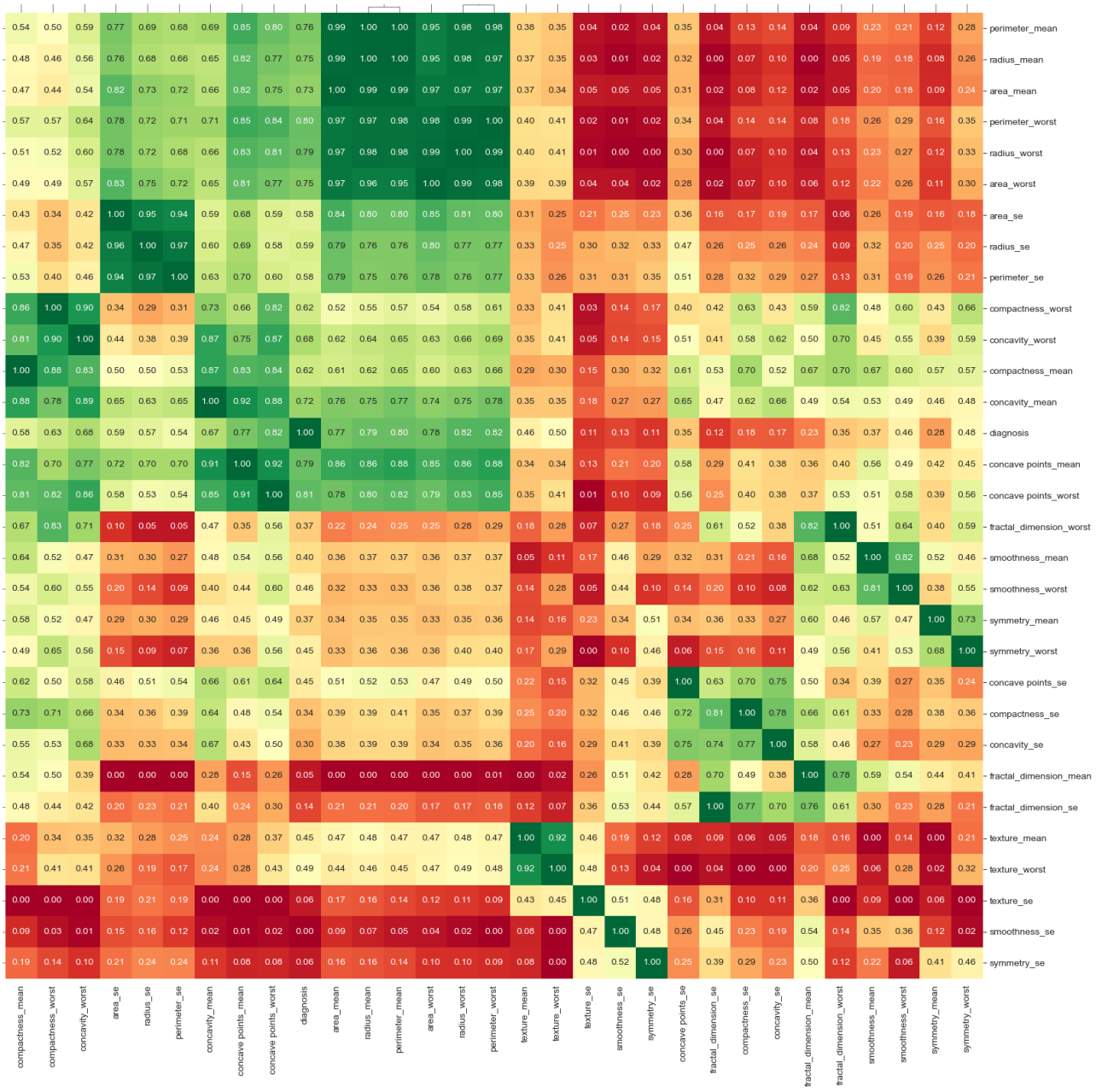


Figure 7: Cluster Heatmap Matrix of values for first model showing features grouped by trends.

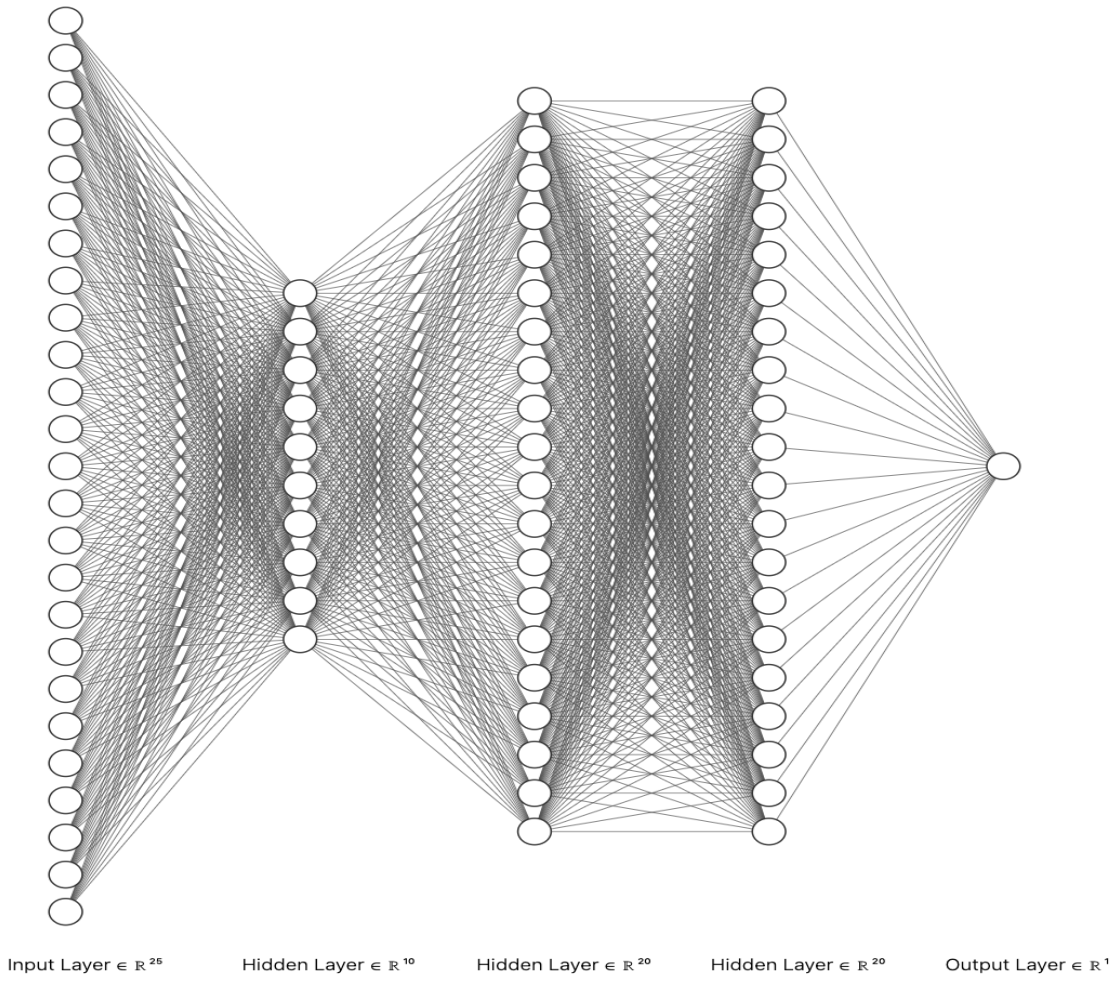


Figure 8: Initial Deep Neural Network architecture with 25 input features computed through three hidden layers of 10, 20, and 20 neurons using ReLu activation and a single output node. The proposed model had a total of 76 neurons with 870 weights.

```
visible = Input(shape=(len(final_data.columns),))
hidden1= Dense(33, activation='relu')(visible)
hidden2 = Dense(24, activation='relu')(hidden1)
hidden3 = Dense(36, activation='relu')(hidden2)
output = Dense(1, activation='sigmoid')(hidden3)
model = Model(inputs=visible, outputs=output)
# summarize layers model.summary()
# compile model
model.compile(loss='binary_crossentropy', optimizer='rmsprop', metrics='accuracy')
```

Figure 9: Final Neural Network design structure with compilation parameters. The optimizer was changed to Root Mean Squared Propagation (rmsprop) while the Initial Neural Network used the Adaptive Moment Estimation (adam) optimizer.

	score	epochs	batch size	validation split	validation steps
0	0.982456	100	20	0.05	4
1	0.982456	100	25	0.05	4
2	0.982456	100	30	0.05	4
3	0.982456	100	35	0.05	4
4	0.982456	100	40	0.05	4

Figure 10: Final Neural Network results are consistent after Neural Network parameter tuning over a set of batch sizes.

Bibliography

Anuradha, Reddy. “Support Vector Machine Classifier for Prediction of Breast Malignancy Using Wisconsin Breast Cancer Dataset.” ASIAN JOURNAL OF CONVERGENCE IN TECHNOLOGY 7, no. 3 (2021): 57–60. https://doi.org/10.33130/ajct.2021v07i03.010.

Alshayeji, Mohammad H., Hanem Ellethy, Sa'ed Abed, and Renu Gupta. “Computer-Aided Detection of Breast Cancer on the Wisconsin Dataset: An Artificial Neural Networks Approach.” Biomedical Signal Processing and Control 71 (2022): 103141. https://doi.org/10.1016/j.bspc.2021.103141.

Dubey, Ashutosh Kumar, Umesh Gupta, and Sonal Jain. “Analysis of K-Means Clustering Approach on the Breast Cancer Wisconsin Dataset.” International Journal of Computer Assisted Radiology and Surgery 11, no. 11 (2016): 2033–47. https://doi.org/10.1007/s11548-016-1437-9.

Hinton, Geoffrey. “Deep Learning—a Technology with the Potential to Transform Health Care.” JAMA 320, no. 11 (2018): 1101. https://doi.org/10.1001/jama.2018.11100.

Liu, Kui, Guixia Kang, Ningbo Zhang, and Beibei Hou. “Breast Cancer Classification Based on Fully-Connected Layer First Convolutional Neural Networks.” IEEE Access 6 (2018): 23722–32. https://doi.org/10.1109/access.2018.2817593.

Mayo Clinic. “Breast Cancer.” Mayo Clinic. Mayo Foundation for Medical Education and Research, April 27, 2022. https://www.mayoclinic.org/diseases-conditions/breast-cancer/diagnosis-treatment/drc-20352475#:~:text=A%20biopsy%20is%20the%20only,tissue%20from%20the%20suspicious%20area.

Ogundokun, Roseline Oluwaseun, Sanjay Misra, Mychal Douglas, Robertas Damaševičius, and Rytis Maskeliūnas. “Medical Internet-of-Things Based Breast Cancer Diagnosis Using Hyperparameter-Optimized Neural Networks.” Future Internet 14, no. 5 (2022): 153. https://doi.org/10.3390/fi14050153.

Roberts, Daniel A., Sho Yaida, and Boris Hanin. The Principles of Deep Learning Theory: An Effective Theory Approach to Understanding Neural Networks. Cambridge: Cambridge University Press, 2022.

UCI Machine Learning. “Breast Cancer Wisconsin (Diagnostic) Data Set.” Kaggle, September 25, 2016. https://www.kaggle.com/datasets/uciml/breast-cancer-wisconsin-data.