ATENEO DE MANILA UNIVERSITY

Comparative Analysis of Attention-Free Transformer, MLP Mixer, and Vanilla MLP Architectures in Binary Classification Tasks

A FINAL PROJECT SUBMITTED TO
THE GRADUATE FACULTY OF
THE SCHOOL OF SCIENCE AND ENGINEERING
FOR THE PARTIAL COMPLETION OF
CSCI 214: PATTERN RECOGNITION

DEPARTMENT OF INFORMATION SYSTEMS
AND COMPUTER SCIENCE

BY
BENJAMIN LOUIS L. ANG
JEREMY MARCUS S. TAN

QUEZON CITY, PHILIPPINES
DECEMBER 2024

TABLE OF CONTENTS

CHAPTER

CHAPTER I

INTRODUCTION

1.1 Introduction

Breast cancer is one of the most lethal and heterogeneous diseases, causing a significant number of deaths among women worldwide [11]. It is the second leading cause of cancer-related mortality among women, highlighting the urgent need for accurate and efficient diagnostic tools [11]. Breast cancer originates from malignant tumors, where uncontrolled cell growth leads to the formation of abnormal masses in the fatty and fibrous tissues of the breast [6]. As these cancer cells spread, they contribute to the progression of various stages and types of breast cancer [6].

The complexity of breast cancer diagnosis lies in its multivariate nature and the intricate relationships between biological features such as cell radius, texture, and smoothness. Traditional statistical methods and early machine learning algorithms, while useful, often fail to capture these complexities [4]. Advanced techniques, including Convolutional Neural Networks (CNNs) and transformers, have shown promise but come with challenges such as high computational costs and the risk of overfitting when applied to smaller datasets [6].

One of the most foundational neural network architectures used in machine learning tasks is the vanilla multilayer perceptron (MLP)

[5]. A vanilla MLP consists of fully connected layers that process input features to learn patterns and relationships [5].

Recent innovations in neural network architectures, such as MLP-Mixer and Attention-Free Transformers, offer a promising alternative. MLP-Mixer efficiently captures multivariate dependencies through feature mixing [12], while Attention-Free Transformers eliminate the need for computationally intensive attention mechanisms [16]. These architectures strike a balance between computational efficiency and predictive accuracy, making them well-suited for breast cancer datasets.

This paper aims to evaluate the performance of MLP-Mixer and Attention-Free Transformers in predicting breast cancer outcomes. Specifically, it explores their ability to handle multivariate dependencies, improve classification accuracy, and maintain computational efficiency compared to the vanilla multilayer perceptron.

1.2 Research Questions

1. How do MLP-Mixer and AFT compare to a vanilla MLP in terms of classification accuracy, precision, and recall for breast cancer prediction?

2. Do MLP-Mixer and AFT demonstrate improved capability in capturing multivariate dependencies and non-linear relationships compared to a vanilla MLP?

CHAPTER II

REVIEW OF RELATED LITERATURE

Breast cancer is one of the leading causes of cancer-related mortality among women worldwide, underscoring the critical importance of early and accurate diagnosis. The disease is highly heterogeneous, with various types such as Ductal Carcinoma in Situ (DCIS), Invasive Ductal Carcinoma (IDC), and Lobular Breast Cancer (LBC), each presenting unique diagnostic challenges [11, 6]. Accurate diagnosis is essential for timely treatment and improved patient outcomes, particularly in distinguishing between benign and malignant tumors.

The complexity of breast cancer lies in the interplay of multiple biological features, such as cell radius, texture, smoothness, and symmetry, which are used to identify abnormalities [4]. These features exhibit non-linear relationships and interdependencies, making traditional diagnostic methods insufficient for capturing the full scope of information. Advances in machine learning have opened new possibilities for analyzing these multivariate datasets, offering the potential to improve diagnostic accuracy and efficiency.

Machine learning algorithms have become essential tools in breast cancer diagnosis, enabling automated analysis of large datasets and providing insights that support clinical decision-making [4]. However, the choice of algorithm plays a critical role in determining the success of these methods. While traditional approaches, such as statistical models and early machine learning algorithms, have shown promise, they often require extensive preprocessing and feature engineering [8]. These limitations highlight the need for advanced neural network architectures capable of handling the inherent complexities of breast cancer data.

This review discusses the evolution of machine learning in breast cancer diagnosis. First, traditional machine learning methods are explored, including their strengths and limitations in handling multivariate datasets. Next, the focus shifts to neural network-based approaches, starting with the vanilla multilayer perceptron (MLP) as a baseline and progressing to more advanced architectures like MLP-Mixer and Attention-Free Transformers. Finally, the gaps in existing research and opportunities for further exploration are identified, setting the foundation for this study.

2.1 Traditional Machine Learning Methods

The application of machine learning in breast cancer diagnosis has been transformative, enabling automated analysis of complex datasets. Traditional machine learning methods, such as logistic regression, decision trees, support vector machines (SVMs), and random forests, have demonstrated effectiveness in predicting breast cancer outcomes by leveraging multivariate feature sets [11].

Logistic regression, as one of the most widely used statistical models, has been applied to classify breast cancer cases into benign or malignant categories. Its simplicity and interpretability make it a popular choice for medical applications. However, logistic regression assumes linear relationships among features, limiting its effectiveness in capturing the non-linear dependencies prevalent in breast cancer data [6].

Support vector machines (SVMs) have been employed to identify optimal decision boundaries for classification tasks. By using kernel functions, SVMs can model non-linear relationships between features. Despite their effectiveness, SVMs require careful tuning of hyperparameters and are computationally expensive for large datasets [4].

Random forests, an ensemble learning method, have also been widely applied in breast cancer diagnosis. By aggregating predictions from multiple decision trees, random forests improve robustness and reduce overfitting. However, they lack interpretability, which is critical for understanding the underlying factors influencing predictions in medical contexts [6].

While these methods have contributed significantly to the development of predictive models for breast cancer, they rely heavily on feature engineering and manual preprocessing. Additionally, traditional machine learning approaches often struggle to capture the complex, non-linear interactions between features, highlighting the need for more advanced neural network architectures.

2.2 Vanilla Multilayer Perceptron (MLP)

The multilayer perceptron (MLP) is one of the most fundamental architectures in neural networks and serves as a baseline for predictive tasks [2]. An MLP consists of an input layer, one or more hidden layers, and an output layer, with all layers fully connected [5]. Each neuron in the network processes a weighted sum of its inputs, applies an activation function, and passes the result to the next layer. This enables the network to learn complex, non-linear mappings from input features to target labels [5].

The simplicity of the vanilla MLP makes it computationally efficient and easy to implement, which has contributed to its widespread use in smaller datasets like the Breast Cancer Wisconsin dataset. The architecture can be adapted for classification or regression tasks by modifying the output layer's activation function, such as softmax for classification or linear for regression. Moreover, MLPs are theoretically capable of approximating any continuous function with sufficient hidden layers and neurons, making them broadly applicable across domains [5].

In breast cancer diagnosis, MLPs classify tumors as benign or malignant based on features such as cell radius, texture, smoothness, and symmetry. These features are fed into the network, and the output layer produces a probability score indicating the likelihood of malignancy. Despite this, vanilla MLPs face several challenges:

- **Inability to Capture Complex Interactions:** MLPs treat input features independently, which may limit their ability to model intricate dependencies in breast cancer data.

- **Overfitting:** With small medical datasets, MLPs are prone to overfitting, especially when the number of parameters is high relative to the amount of training data. Techniques such as dropout and weight decay are often required to address this.

- **Scalability:** The fully connected nature of MLPs makes them computationally expensive for datasets with many features, as the number of trainable parameters increases quadratically with input size.

Despite these limitations, the vanilla MLP remains an essential benchmark for evaluating the performance of more advanced architectures. Its simplicity provides a baseline for assessing the effectiveness of new models, such as the feature-mixing capabilities of MLP-Mixer or the computational efficiency of Attention-Free Transformers.

2.3 MLP Mixer

The MLP-Mixer is a groundbreaking architecture introduced by Tolstikhin et al. [12], designed to simplify neural network architectures while maintaining competitive performance. Originally developed for image classification tasks in computer vision, the MLP-Mixer challenges conventional approaches by replacing convolutional and attention mechanisms with multi-layer perceptrons (MLPs) [12]. Its innovative design has since been adapted for structured datasets, including those used in medical diagnostics such as breast cancer classification.

In computer vision, MLP-Mixer emerged as a response to the complexity of transformers and convolutional neural networks (CNNs) [12]. While transformers, such as the Vision Transformer (ViT), introduced attention mechanisms to replace convolutions, they demanded significant computational resources, particularly for large image resolutions.

MLP-Mixer addressed these limitations by relying solely on MLPs, proving that sophisticated tasks could be achieved with a simpler, more efficient architecture [12]. The design divides image patches into tokens and processes them through alternating layers of token mixing and channel mixing.

The MLP-Mixer architecture consists of two primary components:

Token-Mixing MLPs, which enable communication between different spatial locations or data points, capturing dependencies across features or samples. Channel-Mixing MLPs, which independently process each feature channel, allowing the model to learn intricate relationships between features. This separation of token- and channel-level processing allows MLP-Mixer to handle both spatial and feature-based interactions effectively. It also employs layer normalization, skip connections, and GELU activation functions to improve training stability and efficiency. While originally tailored for computer vision, the MLP-Mixer's simplicity and adaptability make it an excellent candidate for structured medical datasets like Breast Cancer Wisconsin [12].

MLP-Mixer offers several advantages over traditional and advanced architectures [12]:

- **Simplicity**: By eliminating convolutional and attention layers, MLP-Mixer reduces architectural complexity and computational overhead.

- **Efficiency**: Its linear computational complexity makes it ideal for smaller datasets, which are common in medical applications.

- **Flexibility**: The architecture can be generalized to domains beyond vision, including tabular data.

In the context of breast cancer diagnosis, MLP-Mixer's ability to model multivariate dependencies is particularly valuable. Token-mixing layers capture correlations across features such as cell radius, texture, and smoothness, while channel-mixing layers facilitate deeper insights into feature interactions. This dual capability enhances predictive accuracy while maintaining computational efficiency.

2.4 An Attention Free Transformer

The Attention-Free Transformer (AFT), introduced by Zhai et al. [16], is a groundbreaking architecture that reimagines the traditional transformer by eliminating dot-product self-attention. This innovation addresses the computational inefficiencies of conventional transformers, which suffer from quadratic time and space complexity with respect to sequence length [16]. By reducing this complexity to linear, AFT achieves scalability and efficiency while maintaining robust performance, making it a compelling choice for structured datasets like breast cancer diagnosis [16].

AFT retains the foundational components of the transformer—queries (Q), keys (K), and values (V)—but processes them differently to streamline computation [16]. Instead of relying on dot-product attention, AFT employs element-wise operations to aggregate keys and values, significantly reducing computational overhead. The architecture incorporates learned position biases, which encode the relationships between sequence elements, enabling the model to maintain global dependencies. Variants such as AFT-local and AFT-conv further enhance the architecture's adaptability by introducing locality and spatial weight sharing for tasks requiring specific feature interactions [16].

The efficiency of AFT is one of its defining features. By replacing the computationally intensive attention mechanism with a simpler yet effective aggregation method, AFT scales seamlessly to larger inputs and models [16]. This reduction in complexity is particularly valuable in medical applications, where datasets like Breast Cancer Wisconsin require efficient yet accurate models. AFT's modular design also allows it to adapt to variable-sized inputs, ensuring versatility across different domains.

In the context of breast cancer diagnosis, AFT's ability to model global dependencies between features such as cell radius, texture, and smoothness is particularly advantageous. The learned position biases ensure that the architecture captures meaningful relationships without relying on extensive computational resources. Moreover, its variants, such as AFT-local, can be optimized for smaller datasets, reducing the risk of overfitting while maintaining predictive accuracy.

CHAPTER III

METHODOLOGY

3.1 Data Collection and Pre-processing

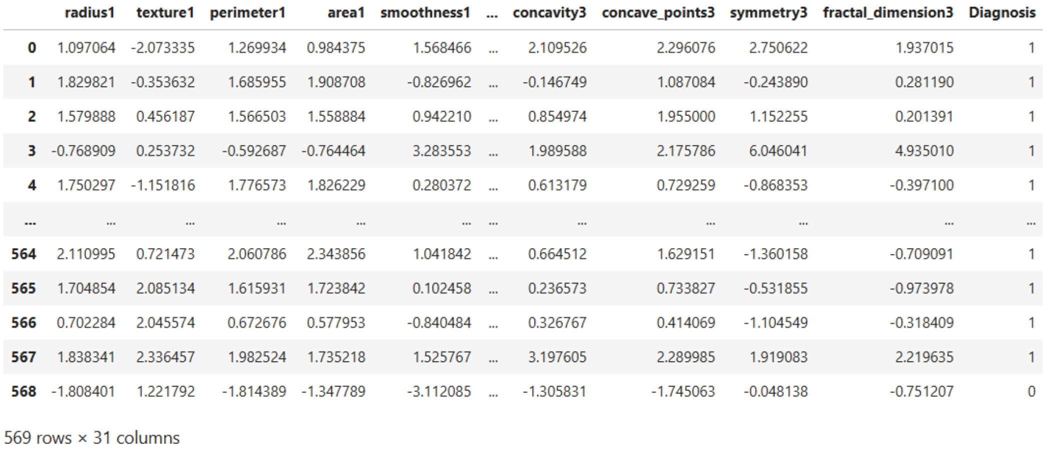
3.1.1 Classification Dataset

A breast cancer database of patients from Wisconsin was used as the training and testing dataset for all the models used in this study. The dataset was downloaded from the UC Irvine Machine Learning Repository [15]. There are a total of 569 instances in the dataset, each corresponding to a patient. Moreover, the dataset contained 30 continuous, real-valued features and one categorical variable for the label. For each patient, three cell nuclei from the breast were analyzed. In each cell nucleus, ten metrics were measured: radius, texture, perimeter, area, smoothness, compactness, concavity, number of concave points, symmetry, and fractal dimension. On the other hand, the label for each instance was the patient's diagnosis, either "M" for malignant or "B" for benign.

3.1.2 Data Pre-processing

Before the dataset was input into the models, the features were first normalized. Z-score normalization was done on all the features, resulting in each feature having a mean of 0 and a standard deviation of 1. On the other hand, the categorical variable was binary-encoded, where 'M' (malignant) was mapped to 1 while 'B' (benign) was mapped to 0.

The processed dataset can be seen in figure 3.1.



	radius1	texture1	perimeter1	area1	smoothness1	...	concavity3	concave_points3	symmetry3	fractal_dimension3	Diagnosis
0	1.097064	-2.073335	1.269934	0.984375	1.568466	...	2.109526	2.296076	2.750622	1.937015	1
1	1.829821	-0.353632	1.685955	1.908708	-0.826962	...	-0.146749	1.087084	-0.243890	0.281190	1
2	1.579888	0.456187	1.566503	1.558884	0.942210	...	0.854974	1.955000	1.152255	0.201391	1
3	-0.768909	0.253732	-0.592687	-0.764464	3.283553	...	1.989588	2.175786	6.046041	4.935010	1
4	1.750297	-1.151816	1.776573	1.826229	0.280372	...	0.613179	0.729259	-0.868353	-0.397100	1
...	...	...	...	...	...	...	...	...	...	...	...
564	2.110995	0.721473	2.060786	2.343856	1.041842	...	0.664512	1.629151	-1.360158	-0.709091	1
565	1.704854	2.085134	1.615931	1.723842	0.102458	...	0.236573	0.733827	-0.531855	-0.973978	1
566	0.702284	2.045574	0.672676	0.577953	-0.840484	...	0.326767	0.414069	-1.104549	-0.318409	1
567	1.838341	2.336457	1.982524	1.735218	1.525767	...	3.197605	2.289985	1.919083	2.219635	1
568	-1.808401	1.221792	-1.814389	-1.347789	-3.112085	...	-1.305831	-1.745063	-0.048138	-0.751207	0

569 rows × 31 columns

Figure 3.1: The Normalized Breast Cancer Dataset

3.1.3 Data Splitting

The entire dataset was split $50\%-25\%-25\%$ into the training, validation, and testing sets, respectively. Given the limited number of instances in the dataset, careful allocation of data across these sets was crucial to ensure sufficient samples for model evaluation. The dataset was randomly split using the *train_test_split* function from **Scikit**. The presence of the validation set is important for the hyperparameter tuning of the models.

3.2 Classification Model Architectures

This study compared the classification performance of three neural networks: the basic multilayer perceptron (MLP), the multilayer perceptron mixer (MLP Mixer), and the Attention-Free Transformer (AFT). The study focused on determining the learning effectiveness of the models in classifying tabular data.

All three models were constructed, trained, and evaluated using the **PyTorch** library from Python. PyTorch is a machine learning library that provides a flexible way of building various neural networks [1]. The normalized features from the breast cancer dataset were used as inputs in all the models. Moreover, batching was implemented in all the neural network training. A batch size of 32 was used for all three models.

3.2.1 Multilayer Perceptron

This study implemented a basic MLP model to serve as a baseline model of reference against the more advanced MLP Mixer and AFT models. In this study, the basic MLP architecture contained two dense hidden layers and one output layer. The hidden layers used the ReLU activation function since it is more computationally efficient compared to exponential functions like sigmoid and tanh [9]. The output layer used the sigmoid activation function. The sigmoid function is commonly used in binary classification tasks because it results in an output between 0 and 1.

3.2.2 MLP-Mixer

The MLP-Mixer was introduced as a more computationally efficient alternative to traditional convolutional neural networks and attention-based transformers in image classification tasks [12]. The MLP-Mixer is composed entirely of blocks of multilayer perceptrons and fully-connected layers. The input to the MLP-Mixer is a sequence of image patches or tokens, where each image patch has several channels. For example, a colored image contains three separate channels: red, blue, and green (RGB). The original study introduced two different MLP blocks in the MLP Mixer architecture: a channel-mixing block and a token-mixing block. The channel-mixing MLP block allows the model to find relationships between different channels while operating on tokens individually. On the other hand, the token-mixing MLP blocks allow for communication between channels. In essence, image patches can be transformed using simple feed-forward layers in order to extract more complex relationships between tokens. A figure of the MLP Mixer architecture taken from the original paper from Google can be seen in figure 3.2 [12].

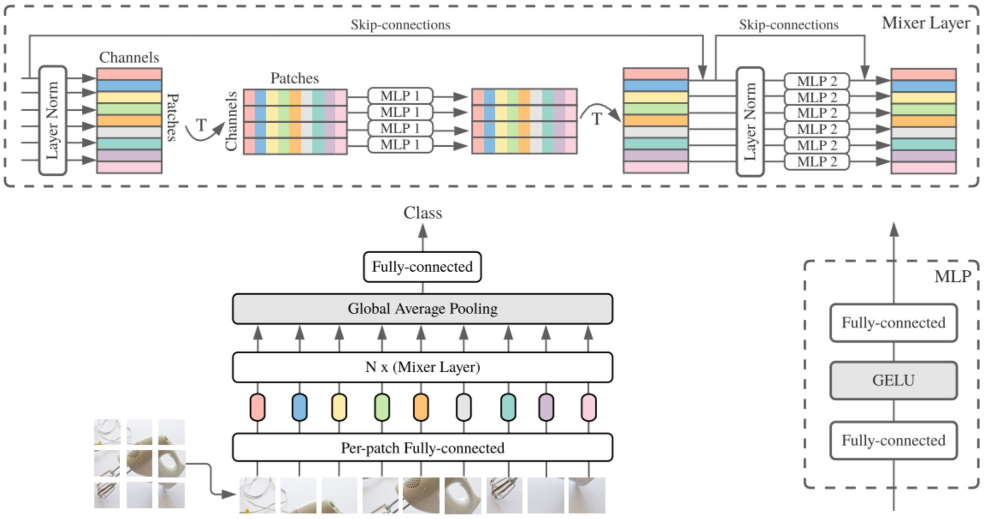


Figure 3.2: MLP Mixer Architecture from [12]

This study extends the use case of the MLP-Mixer from image classification to binary classification on tabular data. The code used in this study was adapted from a Github repository, "MLP Mixer for MNIST Classification", which used the same architecture as the MLP Mixer in the original paper [3]. The input shape of the MLP Mixer for image classification had to be adjusted to input the tabular data into the feed-forward layer of the MLP Mixer.

Given tabular data, one instance in the dataset is a single-dimensional array with elements equal to the values of the features. Since the MLP block expects the input to be a two-dimensional array of features and tokens, it is necessary to add another dimension to the input data for the token. Since each input only represents one token, a 1 is set as the number of tokens in the input shape. Since batching was implemented in model training, the shape of the input data is changed from (batch size, features) to (batch size, 1, features). For this study, batch size was set to 32 and the number of features was set to 30. Now, that the input data contains both token and channel dimensions, the input can be fed through both the channel-mixing and token-mixing MLP blocks.

3.2.3 Attention-Free Transformer

Similar to the MLP-Mixer, the Attention-Free Transformer (AFT) was introduced by Apple as a more efficient alternative architecture to regular attention-based transformers [16]. The AFT is less computationally expensive than the regular transformer models, which learn relationships between any two inputs at the cost of computationally expensive dot-product self-attention. Instead of doing scaled dot product operations, the AFT operation utilizes element-wise products, which are less computationally expensive.

This study utilized a simpler version of the AFT layer introduced in the original paper from Apple. The AFT-simple layer does not consider position biases between inputs in the dataset. For the breast cancer database, the position of the instances in the dataset does not provide any information for training the models since each patient is not related to any of the other patients.

The implementation of the AFT was adapted from the Github site created by one of the authors of the original AFT paper [7]. Since the input data for the AFT layers was made for 2-D image data, it was necessary to change the input shape of the instances in the tabular dataset. Similar to the MLP Mixer implementation, the input data shape was changed from (batch size, features) to (batch size, 1, features).

3.3 Bayesian Optimization

The **Optuna** Python package was used in this study to optimize the hyperparameters in the model's architecture. **Optuna** is an efficient hyperparameter optimization framework that can be used with any machine learning framework like PyTorch. It uses the Tree-structured Parzen Estimator (TPE) algorithm, a Bayesian optimization method, to automatically find the best hyperparameter values [14]. Bayesian optimization uses probabilistic models to approximate the objective value function based on past evaluations [13]. TPE is a specific type of Bayesian optimization that is particularly effective in high-dimensional or feature-intensive training sets.

For hyperparameter tuning, only the training and validation sets were utilized. The testing data was kept for the final model evaluation.

Parameter	Possible Values
Number of Neurons in 1st hidden layer	$8,\ 10,\ 16$
Number of Neurons in 2nd hidden layer	$6,\ 7,\ 14$
Learning Rate	$(0.00001, 0.01)$

Table 3.1: Hyperparameters in the Vanilla MLP

The validation loss of the model was the objective function, which was to be optimized or minimized. Binary cross entropy was used as the loss function for the three neural networks. Binary cross entropy is designed for binary classification problems as it penalizes the model for assigning low predictions to the correct class [10]. For all three model architectures, 50 different hyperparameter configurations were evaluated.

Tables 3.1, 3.2, and 3.3 show the list of hyperparameters that were optimized in each of the three models.

Parameter	Possible Values
Number of MLP Mixer Blocks	$1,\ 2,\ 3,\ ...8$
Dimensionality of Token Mixing Block	$32,\ 64,\ 128$
Dimensionality of Channel Mixing Block	$32,\ 64,\ 128$
Learning Rate	$(0.00001, 0.01)$

Table 3.2: Hyperparameters in the MLP-Mixer

Parameter	Possible Values
Dimensionality of AFT Block	$8,\ 10,\ 16$
Number of Neurons in 1st hidden layer	$32,\ 64,\ 128$
Number of Neurons in 2nd hidden layer	$16,\ 32,\ 64$
Learning Rate	$(0.00001, 0.01)$

Table 3.3: Hyperparameters in the Attention-Free Transformer

3.4 Evaluation Metrics

Since the sigmoid function was used in the output layer of the models, the resulting output would be a range from 0 to 1, which represents the probability of the positive class. A threshold value of 0.5 was used for the classification decision. The performance of the models on the validation and testing set was measured using the following commonly used classification metrics:

1. Accuracy measures the total number of correct predictions compared to the total number of predictions. The threshold value of 0.5 was used here.

   $$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN},$$

   where TP is the number of true positives, TN is the number of true negatives, FP is the number of false positives, and FN is the number of false negatives.

2. The ROC-AUC score, the area under the receiver operating char-

acteristic curve, is the probability that the model will classify a random positive sample over a random negative sample. The true positive rate and false positive rate are plotted at various thresholds. The true positive rate is the proportion of actual positive cases correctly classified, while the false positive rate is the proportion of negative cases incorrectly classified as positive. The model with a greater area under the curve is considered the better model.

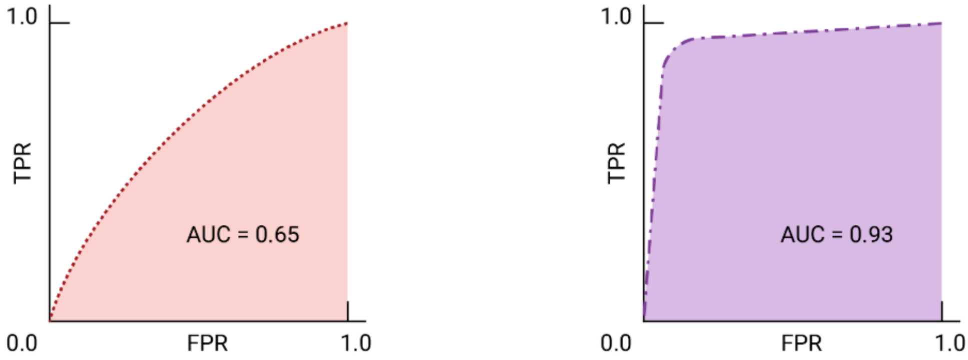


Figure 3.3: Sample ROC curves from Google

Lastly, confusion matrices were produced when the best models were evaluated on the testing set. The confusion matrix summarizes the model's predictions compared to the real labels. Other model performance metrics like precision, recall, and F1 score can be derived from the model's confusion matrix.

CHAPTER IV

Results and Analysis

4.1 Bayesian Optimization Results

The optimal hyperparameters for the Vanilla MLP were 14 neurons for the first hidden layer, 10 neurons for the second hidden layer, and a learning rate of 0.0018. On the other hand, the optimal hyperparameters for the MLP-Mixer were 8 MLP Mixer blocks, with each block having a token-mixing MLP dimension of 128 and a channel-mixing MLP dimension of 32, and a learning rate of 0.0002. Lastly, the Attention-Free Transformer had optimal parameters of 32 neurons in the AFT Block, 128 neurons for the first hidden layer, 32 neurons for the second hidden layer, and a learning rate of 0.0003. The evaluation metrics of the best-performing model from each architecture on the validation set can be seen in Table 4.1. Overall, all three models performed well on the validation set. The MLP-Mixer performed the best as it had both the highest accuracy and the highest ROC-AUC score.

4.2 Test Results

After the hyperparameter tuning, the best-performing models from each architecture were evaluated on the testing set. The evaluation metrics can be seen in Table 4.2. Similar to the evaluation on the validation

Model	Accuracy	ROC-AUC score
Vanilla MLP	0.9737	0.9857
MLP-Mixer	0.9824	0.9912
Attention-Free Transformer	0.9649	0.9863

Table 4.1: Performance of the best-performing models on the validation set

set, the MLP-Mixer had the best overall results on the test set with the highest accuracy and the best ROC AUC score.

Model	Accuracy	ROC-AUC score
Vanilla MLP	0.9474	0.9928
MLP-Mixer	0.9737	0.9954
Attention-Free Transformer	0.9561	0.9921

Table 4.2: Performance of the best-performing models on the test set

4.2.1 Confusion Matrices

The confusion matrices for the best-performing models on the testing set given a classification threshold of 0.5 are given in figures A.1, A.2, A.3. Note that the true negatives (benign) are in the upper left quadrant of the confusion matrix, while the true positives (malignant) are in the lower right quadrant.

A key metric for breast cancer classification that can be gleaned from the confusion matrices is the true positive rate or recall. This is the proportion of all malignant cases that were classified as malignant by the classification model. The Vanilla MLP had a recall of 0.91 as it correctly classified 39 out of 43 malignant cases. The MLP-Mixer had a slightly higher recall of 0.93, correctly classifying 40 malignant cases. The Attention-Free transformer had a perfect recall of 1.00 as it correctly classified all 43 malignant cases in the test set.

CHAPTER V

Conclusion

This study was able to successfully apply three MLP-based neural networks for the binary classification of tabular data in the domain of breast cancer detection. After the models from each architecture were trained and its parameters were optimized, the best-performing models all achieved an accuracy and ROC-AUC score of above 0.94 on the testing set data. The results demonstrate the flexibility of both the MLP-Mixer and Attention-Free Transformer models beyond their original use in image classification. The MLP Mixer and AFT blocks are capable of capturing non-linear relationships between features in high-dimensional tabular datasets such as breast cancer detection.

Overall, the MLP-Mixer demonstrated the best overall results in both the validation and test sets. When compared to the vanilla or regular MLP, the MLP-Mixer was able to find more nuanced relationships between the features using the innovative channel-mixing and token-mixing MLP blocks. When compared to the Attention-Free Transformer, this may be due to the nature of the breast cancer dataset, where capturing the relationship between features like cell radius and smoothness is more important than finding positional relationships between the tokens or instances. The main advantages of using the full AFT model, such as learning position biases in the input sequences, are not as relevant for the breast cancer dataset as data instances are not inherently sequential. However, the MLP mixer may not necessarily be the preferred model in different applications. In the context of cancer detection, where identifying positive cases is crucial, the AFT model demonstrated superior performance in recall, outperforming the other two MLP models on the test set.

The scope of this comparative analysis of MLP model advancements can be expanded to other types of machine learning tasks and domains. Future researchers can explore the performance of the MLP models on multiclass classification or regression. The superiority of the MLP-Mixer as the best-performing MLP model can be further validated on more feature-intensive datasets with a greater number of interconnected features or on less feature-intensive datasets with a fewer number of features.

BIBLIOGRAPHY

[1] Pytorch: An open source machine learning framework, 2024.

[2] DESAI, M., AND SHAH, M. An anatomization on breast cancer detection and diagnosis employing multi-layer perceptron neural network (mlp) and convolutional neural network (cnn). *Artificial Intelligence in Medicine* (2020), Available online 24 November 2020. Open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

[3] ELMAN. Mlp-mixer for mnist classification. GitHub, 2024.

[4] FATIMA, N., LIU, L., HONG, S., AND AHMED, H. Prediction of breast cancer, comparative review of machine learning techniques, and their analysis. *IEEE Access 8* (2020), 150360–150376.

[5] G, S., AND RAMKUMAR, G. Identification and classification of breast cancer using multilayer perceptron techniques for histopathological image. In *2023 Second International Conference on Electronics and Renewable Systems (ICEARS)* (2023), pp. 1633–1639.

[6] LU, Y., LI, J.-Y., SU, Y.-T., AND LIU, A.-A. A review of breast cancer detection in medical images. In *Proceedings of the IEEE*

Visual Communications and Image Processing (VCIP) (December 2018), pp. 1–4.

[7] RISH 16. aft-pytorch. GitHub, 2021.

[8] SARASWAT, S., KESWANI, B., AND SARASWAT, V. Classification of breast cancer using machine learning: An in-depth analysis. In *Proceedings of World Conference on Artificial Intelligence: Advances and Applications* (Singapore, 2023), A. K. Tripathi, D. Anand, and A. K. Nagar, Eds., Springer Nature Singapore, pp. 191–203.

[9] SCHMIDHUBER, J. Deep learning in neural networks: An overview. *Neural Networks 61* (2015), 85–117.

[10] SHUKLA, A., AND ARORA, D. Deep learning model for identification and classification of web based toxic comments. *2023 International Conference in Advances in Power, Signal, and Information Technology (APSIT)* (2023), 274–279.

[11] SUN, Y.-S., ZHAO, Z., YANG, Z.-N., XU, F., LU, H.-J., ZHU, Z.-Y., AND ET AL. Risk factors and preventions of breast cancer. *International Journal of Biological Sciences 13* (2017), 1387.

[12] TOLSTIKHIN, I., HOULSBY, N., KOLESNIKOV, A., BEYER, L., ZHAI, X., UNTERTHINER, T., YUNG, J., STEINER, A., KEYSERS, D., USZKOREIT, J., LUCIC, M., AND DOSOVITSKIY, A. Mlp-mixer:

An all-mlp architecture for vision. *arXiv preprint arXiv:2105.01601* (2021).

[13] WANG, X., JIN, Y., SCHMITT, S., AND OLHOFER, M. Recent advances in bayesian optimization. *ACM Computing Surveys 55* (2022), 1 – 36.

[14] WATANABE, S. Tree-structured parzen estimator: Understanding its algorithm components and their roles for better empirical performance, 2023.

[15] WOLBERG, W., MANGASARIAN, O., STREET, N., AND STREET, W. Breast Cancer Wisconsin (Diagnostic). UCI Machine Learning Repository, 1993. DOI: https://doi.org/10.24432/C5DW2B.

[16] ZHAI, S., TALBOTT, W., SRIVASTAVA, N., HUANG, C., GOH, H., ZHANG, R., AND SUSSKIND, J. An attention free transformer. *arXiv preprint arXiv:2105.14103* (2021).

APPENDIX A

Confusion Matrices

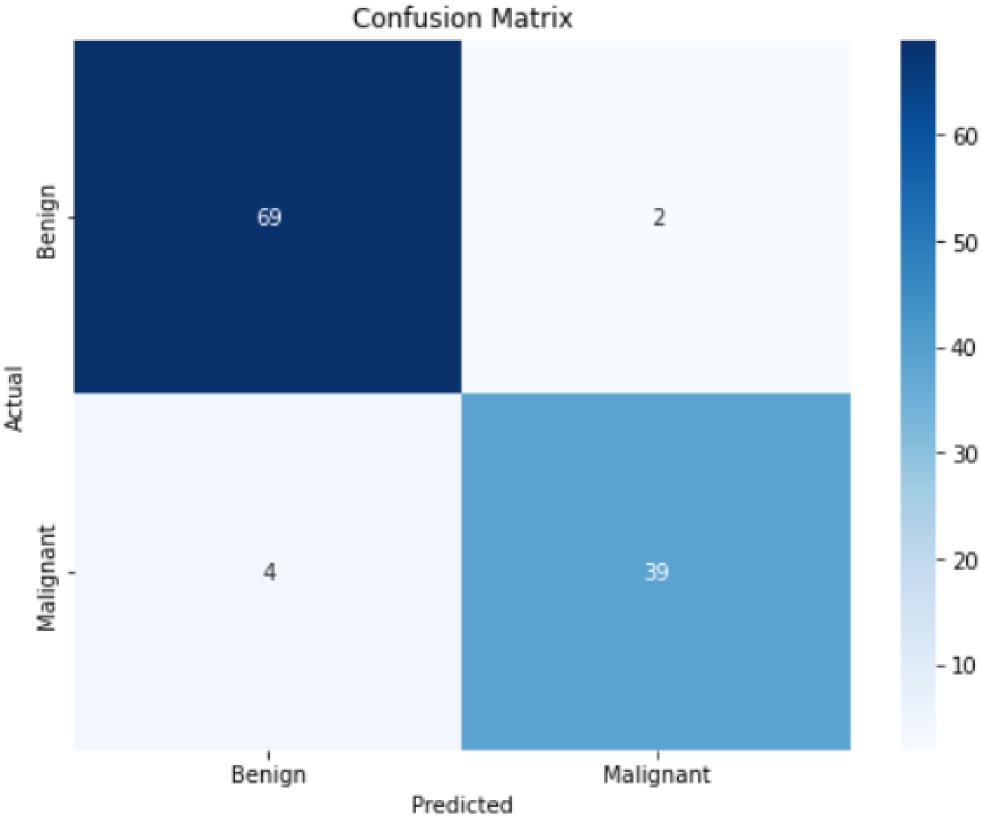


Figure A.1: Confusion Matrix of the Vanilla MLP

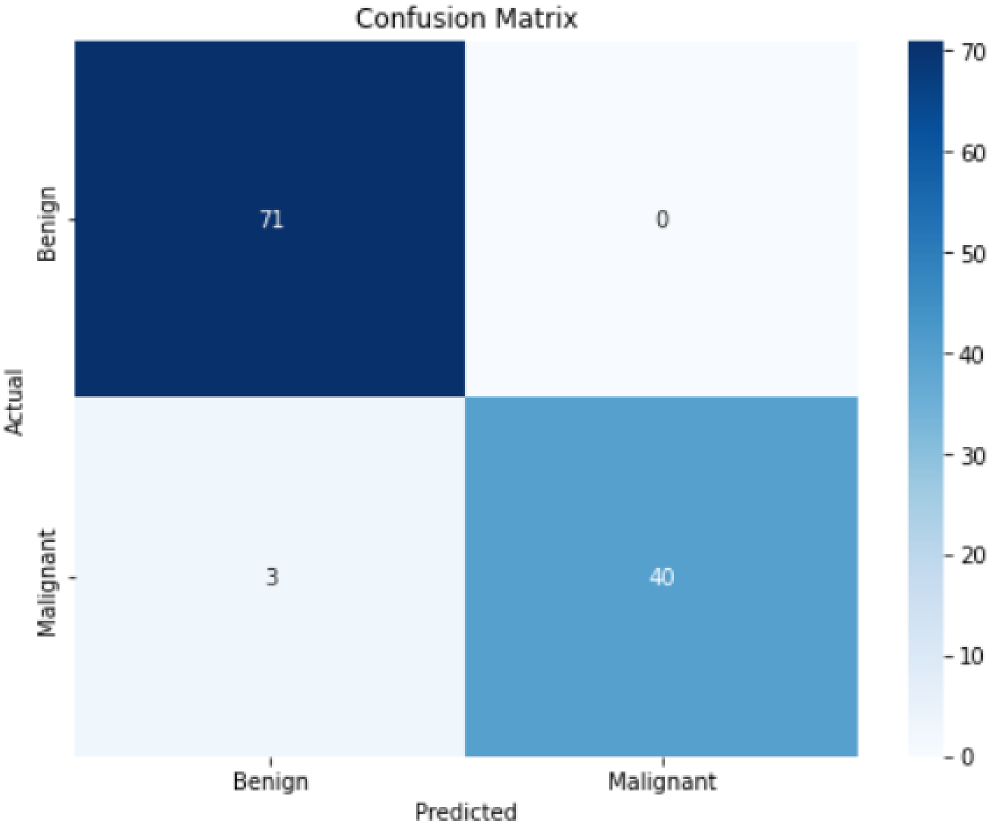


Figure A.2: Confusion Matrix of the MLP-Mixer

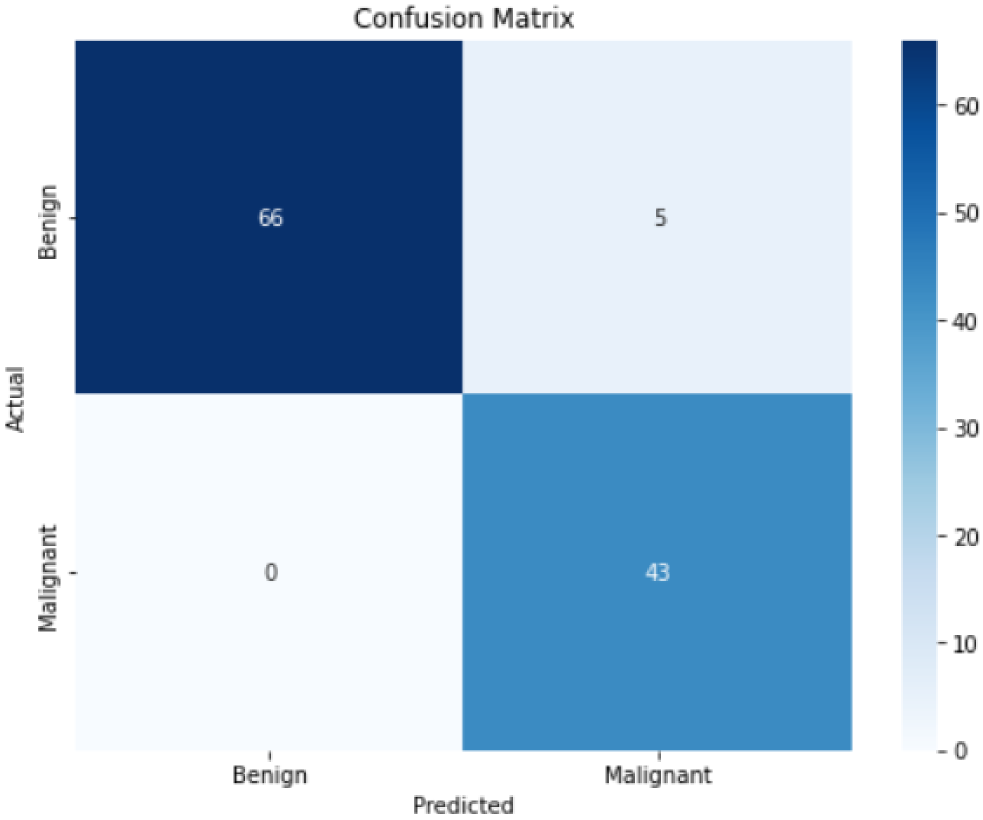


Figure A.3: Confusion Matrix of the Attention-Free Transformer