CS 7641 Supervised Learning

Hyesung Ji (hji61@gatech.edu)

Abstract— The supervised learning is a good way to estimate and predict result by train labeled data for classification and regression. In this project, five supervised learning methods (Decision Tree, Neural Networks, Support Vector Machines, AdaBoost, and K-Nearest Neighbors) of are studied in terms of accuracy and training time.

I. Introduction

A. *Determination of datatsets*

In order to analyze and investigate the classification performance of each method, it is very necessary to select datasets. The standard of dataset selection is: first the dataset consists of numbers, not letters. Second, the dataset has to be classified clearly. Also, to compare different datasets, small number with binary-classification dataset and relatively large number with multiclass-classification dataset were selected.

First dataset is 'Breast cancer dataset [1]. Nowadays, a lot of people suffer from disease such as cancer. Therefore, prediction for cancer can be very important to prevent human from cancer. The dataset comes from Fine Needle Aspirate (FNA) image of breast mass. The attributes are 30 and 569 samples are in the dataset. The second dataset 'WiFi-localization' dataset [2]. Nowadays, most people are using personal smartphone and a lot of public spaces provide their customers with WiFi service. The data is signal strength observed by a smartphone of seven WiFi signals from four rooms. Considering those seven signal strengths, a smartphone determines one of the indoor space. This dataset is multi-classification problem and it has 2,000 samples in the dataset.

B. *Pre-processing*

For better classification, two datasets are normalized and the range is transformed the value between 0 and 1. The function 'MinMaxScaler' is used for this normalization. In the case of the dataset 'Breast Cancer', each column of data has different range values from other columns. So, each column is normalized the value between 0 and 1 based on the maximum and the minimum value of the corresponding column. In the case of WiFi-localization dataset, all numbers consist of integer and negatives. The same process was applied to this dataset again.

II. General algorithm

In this project, two or three hyperparameters of each algorithm were investigated. The process is the following. First, the classifier is set with default value and 'random_state' and the performance of the classifier is investigated according to the first hyperparameter by plotting it. Second, the best value of the hyperparameter is applied to the classifier. Third, the second hyperparameter is investigated along the same process as the second process. Finally, the performance of the classifier with two best hyperparameters is investigated by using test sets. The pseudo-code is the following.

Algorithm 1: How to compute the accuracy of each algorithm

```
Initiate classifier with default value, random_state = 42
Set the first ($\alpha$) and the second ($\beta$) hyperparameter range
For accuracy of classifier() with $\alpha = \{\alpha_1, \alpha_2, \dots, \alpha_n\}$
    Best_index = $np.argmax$ (accuracy of classifier($\alpha_i$))
For accuracy of classifier($\alpha_{best}$) with $\beta = \{\beta_1, \beta_2, \dots, \beta_n\}$
    Best_index = $np.argmax$(accuracy of classifier($\alpha_{best}, \beta_i$))
Best_classifier = classifier($\alpha_{best}, \beta_{best}$)
$Y_{predict}$= Best_classifier($X_{test}$)
Accuracy = $\frac{1}{N}\sum_{i=1}^{N} \mathbf{1}(Y_{test} = Y_{predict})$  (1)
```

Considering the number of datasets, the first dataset (Breast Cancer) is split into 70% training-set and 30% test-set. And the second dataset is split into 60% training-set and 40% test-set. In the training-set, the ten cross-validations function were applied for each algorithm and the number of cross-validation is 10. It means that the training-set is divided by 10 validation-folds and each fold plays a role as a validation set. Therefore, the training and validation score are equal to the mean of 10 training and validation accuracies. Additionally, some algorithms are significantly sensitive and changed the performance according to some hyperparameters. Therefore, some trials and errors are done to find the best performance.

III. Decision Tree

For decision Tree, the function 'DecisionTreeClassifier' is used. The first hyperparameter is 'ccp_alpha' and it is used for prunning. In the 'DecisionTreeClassifier' function, the minimal cost-complexity pruning (CCP) is used for pruning to minimize overfitting. This parameter α plays a role as a regularization parameter. The maximu number of depth is another hyperparameter. It controls the maximum depth of the tree.

A. *Breast cancer wisconsin data*

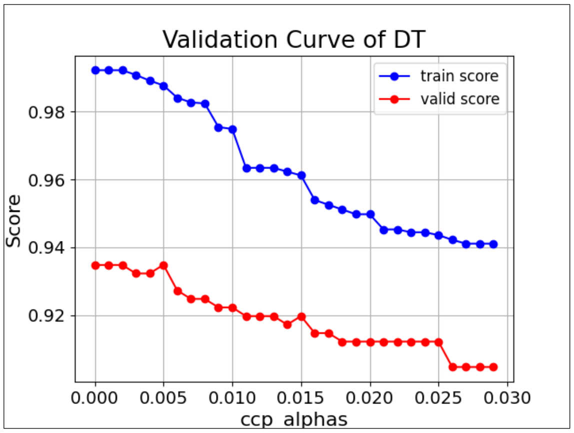

Figure 1. Validation curve according to CCP-α

The parameter $\alpha = \frac{R(t)-R(T_t)}{|T_t|-1}$ (where $R(t), R(T_t)$: a loss function across the t leaves and except for T_t tree respectively. $|T_t|$ is the number of leaves in tree T_t.) According to the Figure 1, as the CCP-α is increasing, the train and validation accuracy tend to be decreased. However, when the CCP-α value is 0.005, the validation score has the best performance. To avoid overfitting and to achieve the best performance, 0.005 is selected as a CCP-α.

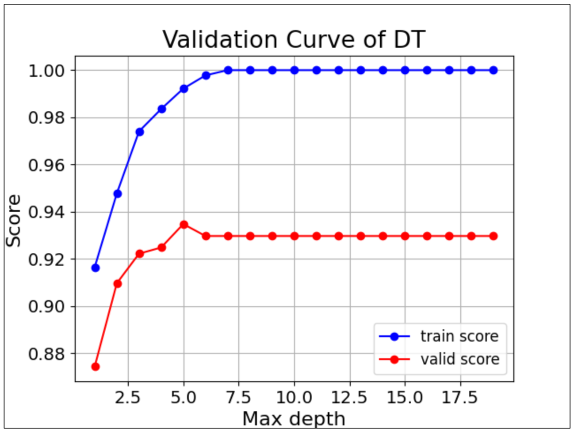

Figure 2. Validation curve according to maximum depth

According to the Figure 2, when the maximum depth is greater than 6, there is no change in train score and validation score. Also, when max depth is 7, the model becomes overfitting. When the maximum depth is 5, the accuracy has the highest value.

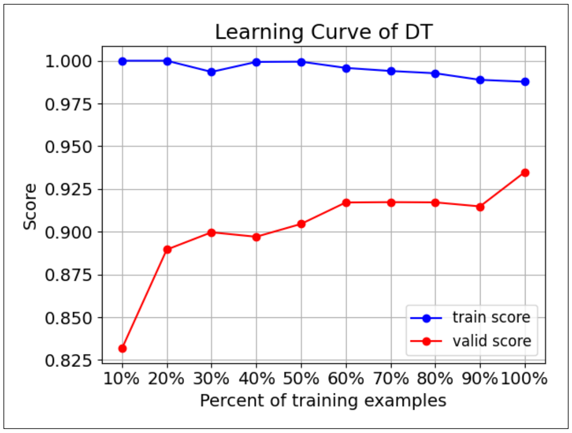

Figure 3. Learning curve

Figure 3 shows the learning curve in accordance with the number of examples. As the number of examples is increased, the train accuracy is mostly stable. On the other hand, validation accuracy is increased as the number of samples is increased. Even though the gap between train and validation accuracy is getting closer, however there is still a large gap. It explains that there are high variance in this classifier. Normally, if the more instances(examples) are added, the variance will be decreased. As discussed in Introduction, the total sample of dataset is 569 and it has 30 features. Therefore, 569 samples may not be sufficient value for Decision Tree method. In summary, there are a high variance and the number of training examples is not sufficient for the Decision Tree.

B. WiFi-Localization

As we can see the above, the WiFi localization data is used for this Decision Tree model.

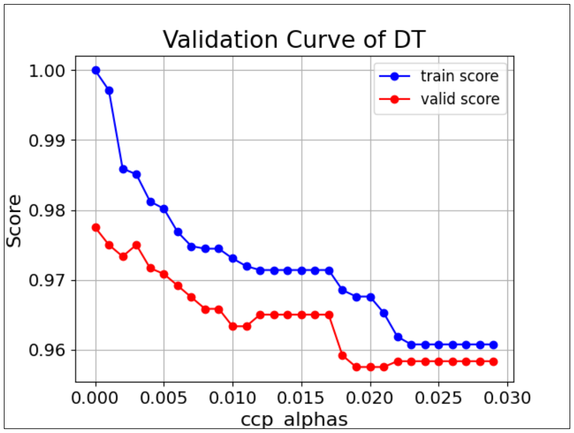

Figure 4. Validation curve according to CCP-α

According to the figure 4, the best CCP-α value is 0 and it is applied to the classifier. It indicates that this data does not need to be pruned.

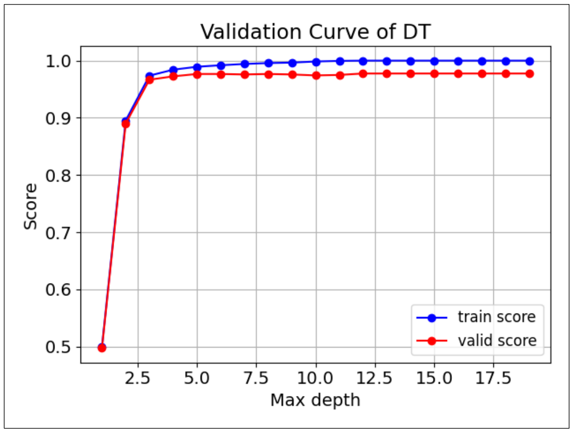

Figure 5. Validation curve according to maximum depth

When the max-depth is 12, the validation accuracy has the highest value. Also, the max-depth is greater than 12, the train accuracy becomes 1.0.

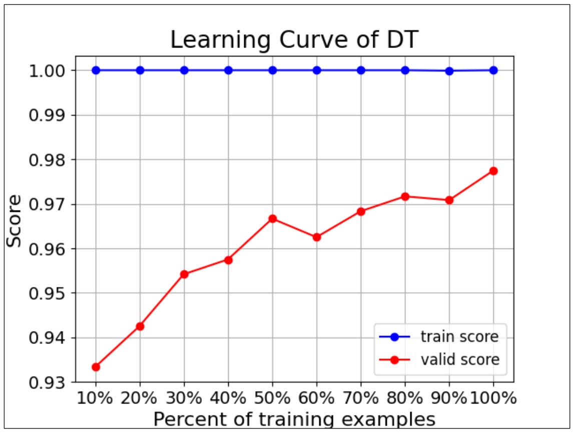

Figure 6. Learning curve

As explained above, when the maximum depth is greater than 12, the training accuracy becomes 1.0. Therefore, the train learning curve is 1.0, and the validation curve is converging to 1.0 as the number of training examples is increased. As we can see, there is still a gap between the training accuracy and the validation accuracy. Compared to the 'Breast Cancer' data, the variance is significantly lower than the previous variance. It also indicates that if more instances are added, the model will become more accurate.

IV. NEURAL NETWORK

For Neural Network, the function 'MLPClassifier' was used and there are a lot of parameters to adjust. In this project, three hyperparameters (learning rate, regularization parameter (α), and number of nodes) were chosen for analysis. For convenience, Other parameters are set to the following.

- The maximum iterations: 2,000
- The optimizer: 'Adam'
- The activation function: 'RELU'

The reason to select the Adam optimizer and RELU function is that these parameters generally have the best performance in many cases. Also, the learning rate is set to constant.

A. *Breast Cancer dataset*

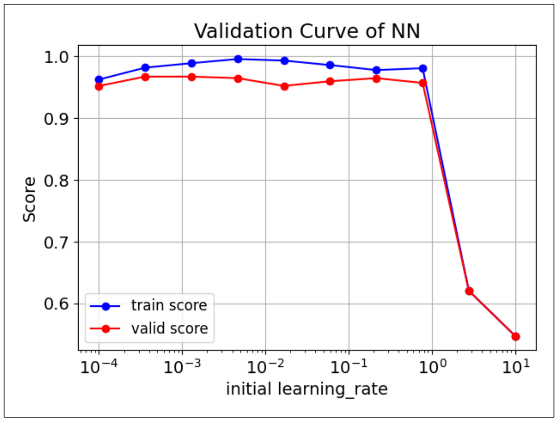

Figure 7. Learning rate

As we can see Figure 7, when the learning rate is approximately 0.00129, the validation accuracy has the highest value. This is applied to the classifier and the regularization value(α) was calculated.

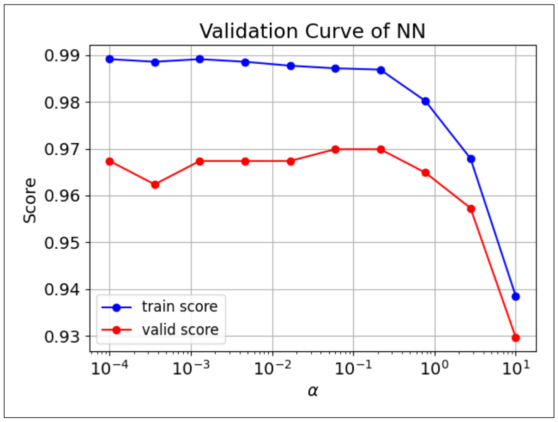

Figure 8. Regularization value(α)

As we can see the Figure 8, the validation score has the highest value when α is approximately 0.2154. So, after fixing the initial learning rate is 0.00129 and the alpha is 0.2154, the number of node for a single and double hidden layer are calculated.

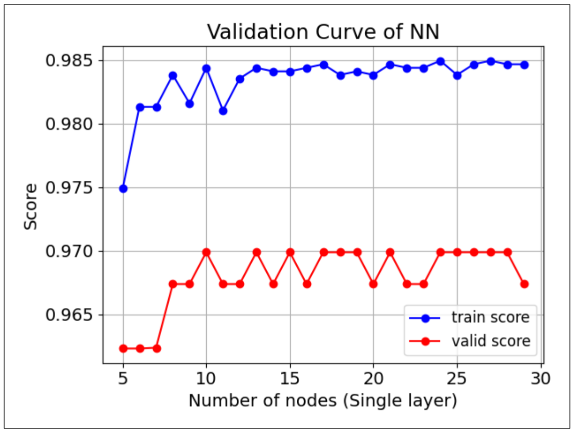

Figure 9. Number of nodes in single Hidden layers

When the number of nodes is greater than 10, the training and validation accuracies are not increased anymore. Compared to the initial learning rate and α, the number of hidden layer is not very sensitive factor to compute the training and validation accuracy in this experiment. Normally, smaller number of layers can be preferred to avoid overfitting. So, the 10 layers are selected. Also, this is single-layer algorithm and the double-layer algorithm can be compared to the single-layer algorithm. The range of hidden nodes is from 5 to 30. So, the total number of hidden nodes are similarly selected.

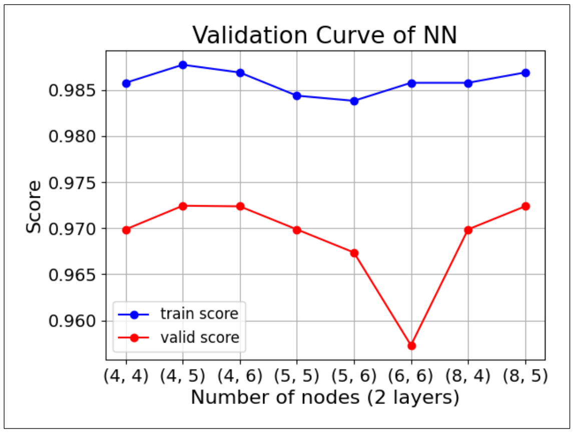

Figure 10. Number of two Hidden layers

The hidden layers were selected similar numbers of the single-layer algorithm. After some trials and erros, some exceptional nodes are eliminated to compare the the accuracy score more easily. (Some cases such as 4× 2 and 4× 3) The best performance is 0.972 when the number of hidden layer is 4×5 and 4×6. So, the So, 4×5 hidden layer is selected. Finally, the best parameters (initial learning rate: 0.00129, α : 0.2154, number of layers: 4×5) are applied to the classifier.

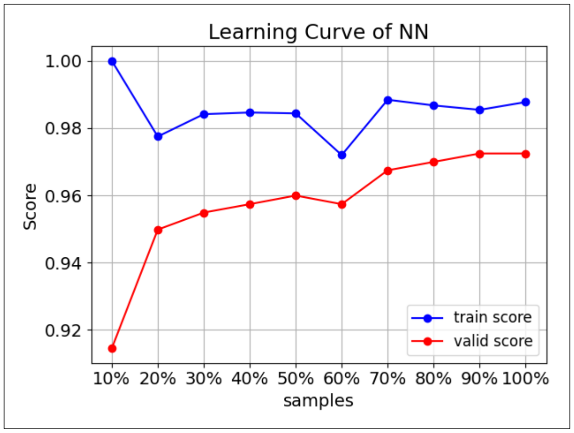

Figure 11. Learning Curve

Compared to the Decision Tree model, the learning curve of the Neural Network has significantly better result. When the number of samples are increased, the gap between the training and validation accuracy is closer and the gap is approximately 0.01. So, the NN model for this dataset has low variance. On the other hand, the Neural Network classifier has a little bias. As the number of samples is increased, the accuracy of classifier does not become 1.0. In summary, even though the model has a little bias and variance, but gap is not too much. So, the model is mostly trained well, but it is necessary to correct the bias.

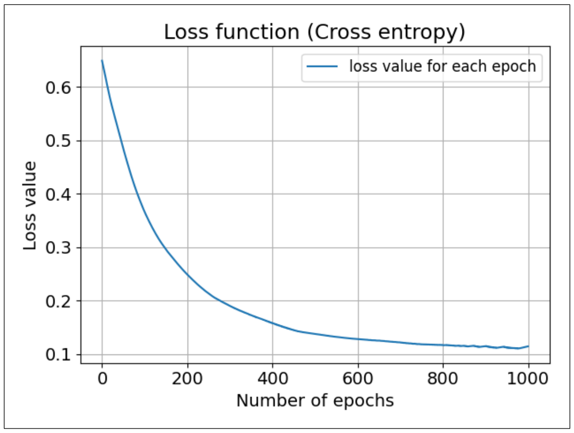

Figure 12. Loss function

This project is to compute classification problem, not regression. So, the cross entropy Loss function is used for this loss function. ($y, \hat{y}$ are real result and predictive result by using test set respectively. N is the total number of rows in the test set)

$$Loss(y, \hat{y}) = \frac{1}{N}\sum_{i=1}^{N} y_i log\hat{y_i} + (1 - y_i)\log(1 - \hat{y_i}) \quad (2)$$

When the number of epochs is increased, the loss function value is stabilized and this model can be a well-trained model.

B. *WiFi-Localization*

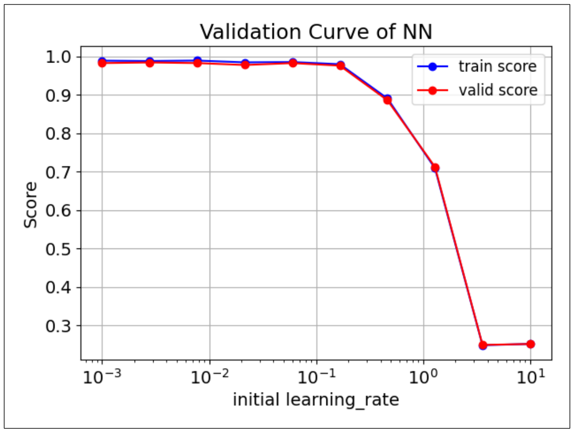

Figure 13. Initial Learning rate

As we can see Figure 13, when the learning rate is approximately 0.00278, the validation accuracy has the highest value. This is applied to the classifier and the regularization value(α) was calculated.

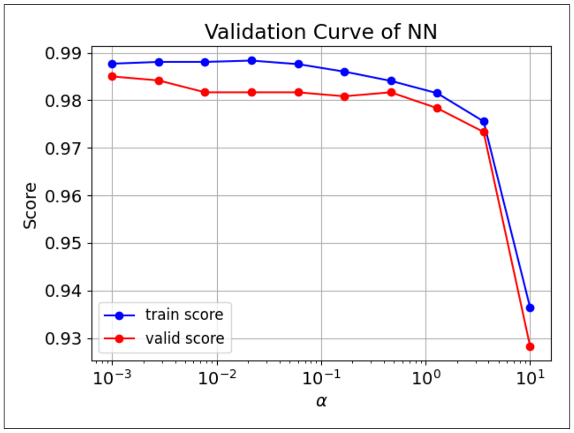

Figure 14. Regularization value(α)

As we can see the Figure 14, the validation score has the highest value when α is 0.001. So, after fixing the initial learning rate is 0.00278 and the alpha is 0.0001, the number of nodes for a single and double layer is calculated.

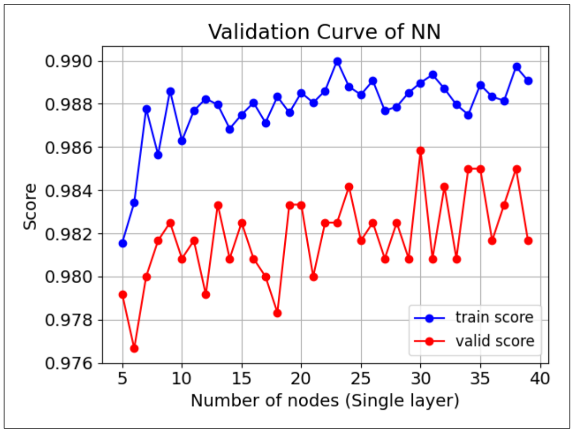

Figure 15. Number of nodes in single Hidden layers

The accuracy of each number of nodes is very similar. It means that the accuracy will not be improved even if the number of nodes is increased. This phenomenon is similar to the previous dataset case.

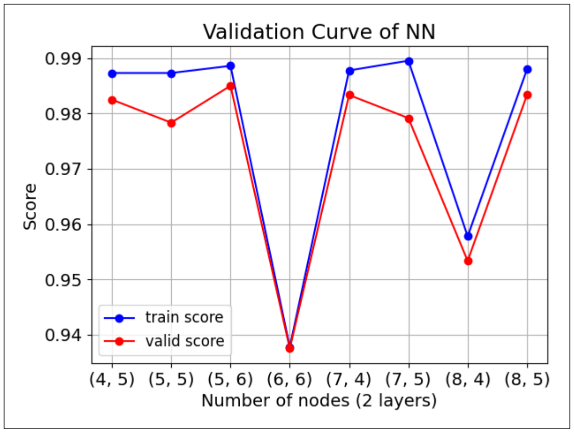

Figure 16. Number of nodes in double Hidden layers

The selection of hidden node size is similar to the Breast Cancer case. As mentioned above, after some trials and erros, some exceptional nodes are also eliminated. (Such as 4× 4 and 4×6) The plot shows that the accuracy results are pretty similar to the single-layer cases. However, when the hidden layer is single layer and 30 nodes are in the layer, the best result is occurred. Finally, the single-layer with 30 nodes is selected in this experiment. Based on these experiments the Learning curve and Loss function are calculated. The best parameters (initial learning rate: 0.00278, α: 0.001, number of layers: 30) are applied to the classifier.

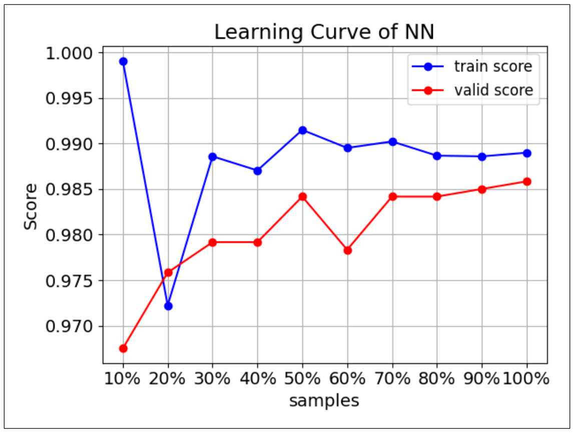

Figure 17. Learning Curve

According to the Learning curve, the model has a little bias. The train curve does not become 1.0 and it seems to converge to 0.99. On the other hands, the difference between train and validation accuracy is small. It means that the model has a lower variance than the previous dataset case. This model is also mostly well-trained.

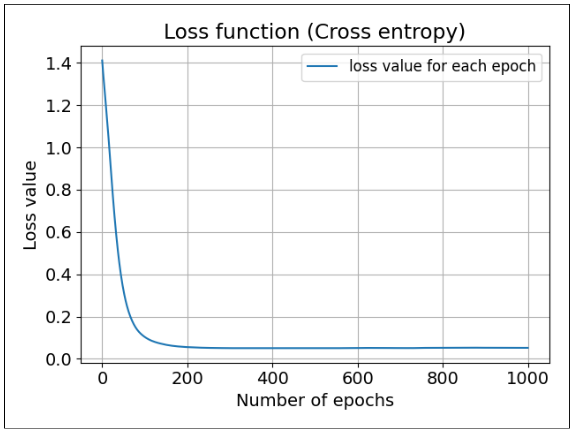

Figure 18. Loss function

The cross entropy is applied for Loss function and it is well converged. In summary, this model seems to be well-trained.

V. SUPPORT VECTOR MACHINE

For Support Vector Machine, the function 'SVC' was used. In this experiment, the type of kernel, the C value, and the gamma value are selected as the hyperparameters. The Kernel is a kind of shape parameter. And the C value plays a role as a regularization parameter and the gamma is the radius (size) of the kernel.

A. Breast Cancer dataset

First, the three kernel types(RBF, Linear, and Sigmoid) are selected for analysis.

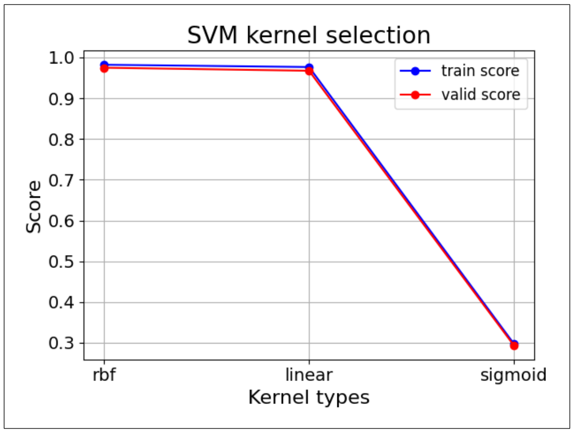

Figure 19. Kernel type

As we can see, the RBF kernel has the best performance and the linear kernel has similar result. But the sigmoid kernel's performance is not proper kernel in this project. Normally, the RBF kernel can be a good performance in most cases. The reason that sigmoid kernel has a bad performance is not identified, one possible reason is the sigmoid kernel is not flexible than other kernels [3]. The sigmoid kernel uses tanh function and it just has +1 or -1 value. The RBF and linear kernel can be adjusted more flexible than the sigmoid function.

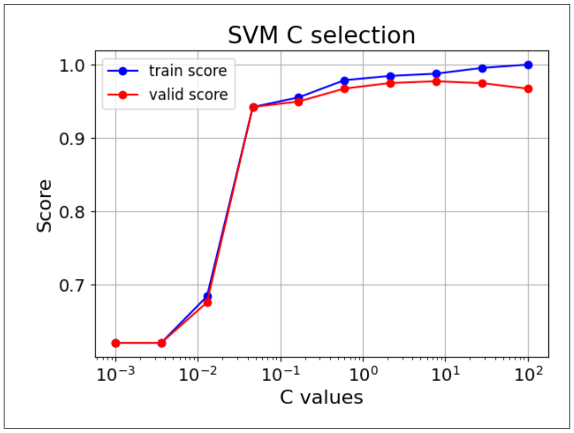

Figure 20. C value

In RBF kernel function, the regularization parameter C is defined the following, where x_n is support vector data and x_i is dataset.

$$K(x_n, x_i) = \exp\left(-\gamma \left|\left|x_n - x_i\right|\right|^2 + C\right) \quad (3)$$

The parameter C adjusts the number of support vectors. So, as we can see Figure 20, when regularization parameter C value is too small, it means the number of support vectors is small, so the accuracy is too low. In the case of large C value, it means the model has a lot of number of support vectors and it makes the model have small margin. So, it shows that the dataset has complex data and the margin of the model should be small to fit the dataset. In this case, the best C value is 7.7426. After fixing the kernel type and C value, the gamma value is calculated.

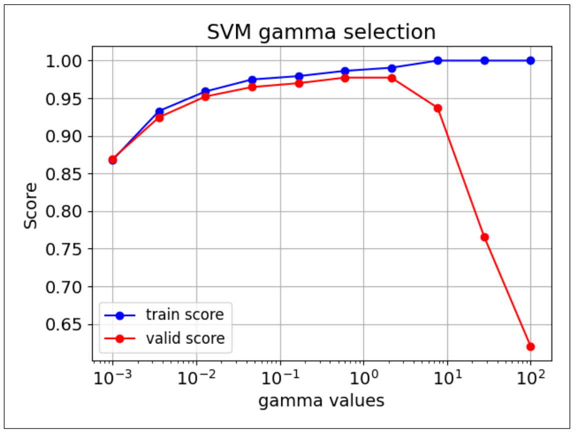

Figure 21. Gamma value

According to the formula (3), the gamma value means the radius of the RBF unit. So, if the gamma is small or large, the radius of each RBF unit becomes large or small. ($\gamma = \frac{1}{2\sigma^2}$)

As a result of the gamma value, when the training accuracy becomes 1.0, the validation accuracy becomes lower. This is typical overfitting. Therefore, searching a proper gamma is very important for SVM classifier.

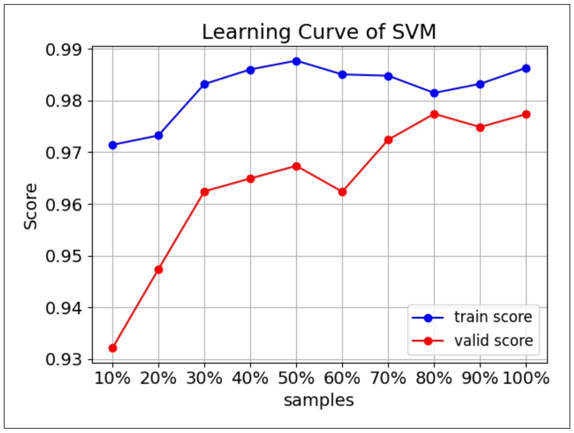

Figure 22. Learning Curve

The difference between train accuracy and validation accuracy is getting smaller as the number of samples is increased. However, the train accuracy does not become exactly 1.0 even though the final value is very close to 1.0. The learning curve shows that this model has a low bias and it also has a still low variance. In summary, the SVM model for this dataset is mostly well-trained.

B. WiFi-localization

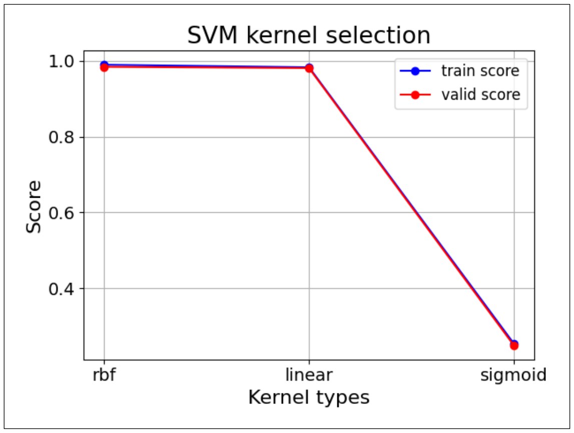

Figure 23. Kernel type

Similarly, the sigmoid kernel has the lower value. RBF and Linear kernel have the similar accuracy, but the RBF kernel has slightly higher value than Linear kernel.

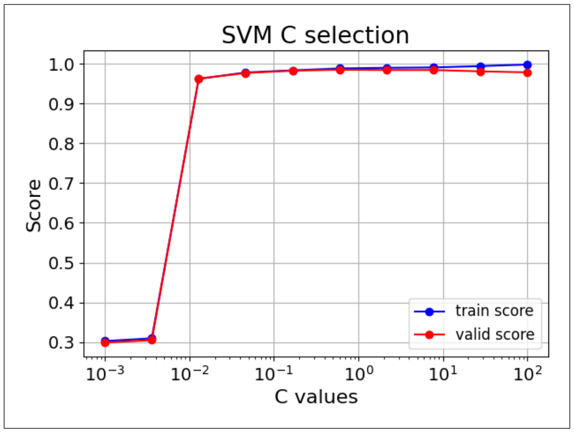

Figure 24. C value

When C value is lower than 0.01. the accuracy is too low. This is similar to the Breast Cancer case. The best C value is 0.5994 and it is smaller than the previous case. It means that the model for WiFi dataset has a greater margin than the model for Breast Cancer dataset.

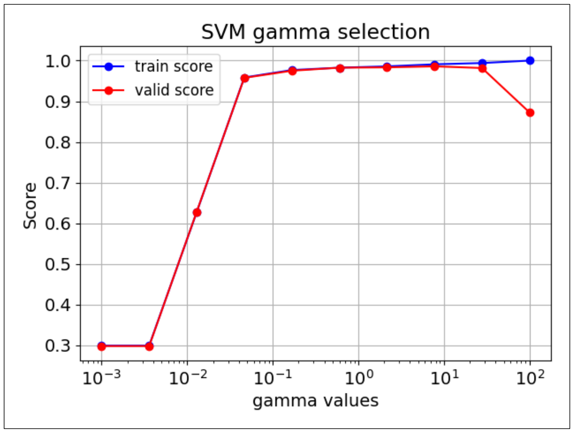

Figure 25. Gamma value

The gamma value is the radius of RBF unit. In this case, the gamma has the best value when gamma is 7.7426. Compared to the previous dataset case, the radius of each RBF unit is very small. This dataset is four class classification problem. Normally, the smaller RBF can be necessary for multi classification than binary classification.

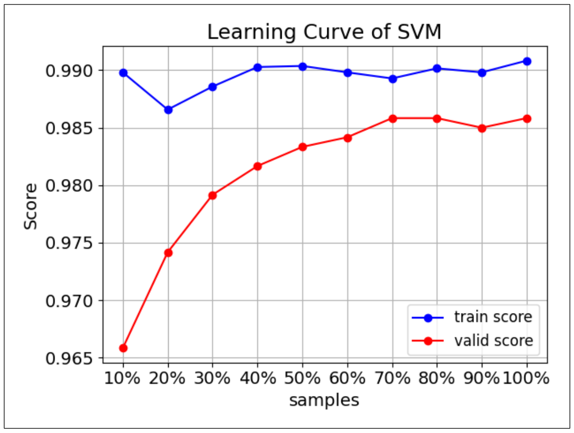

Figure 26. Learning Curve

When the number of samples is increased, the difference between the train accuracy and the validation accuracy is decreased. Also, the train accuracy does not become 1.0. Even though, the gap between the train and the validation accuracy is lower than the previous dataset case, the model has still a small bias and a small variance. In summary, the SVM model for this dataset is mostly well-trained.

VI. ADABOOST TRAINING

Adaboost classifier is an algorithm which makes ‘weak classifier’ and focuses on incorrect classified instances. And then the algorithm adaptively upgrade the weight of weak classifiers to prevent misclassification. So, the linear combination of weak classifiers can be a strong classifiers. In this case, the learning rate and number of estimators are two hyperparameters to be investigated.

A. Breast Cancer

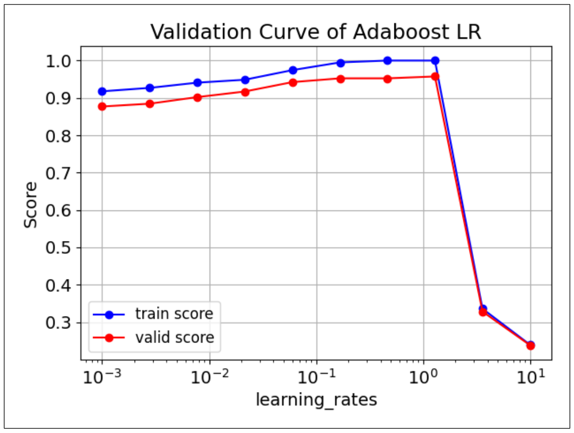

Figure 27. Learning rate

When the learning rate is increased until 1.0, the train and the validation accuracy is increased accordingly. However, when the learning rate is greater than 1.0, the accuracy is significantly decreased.

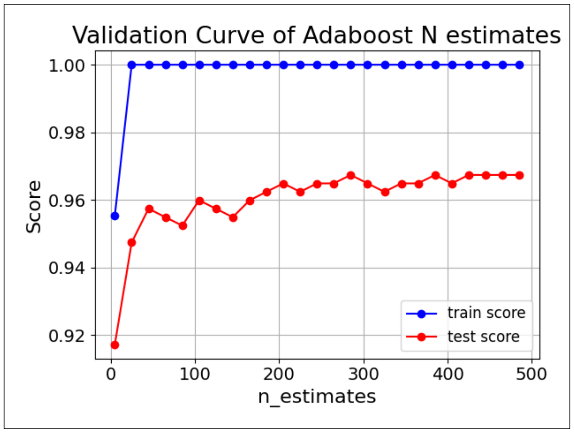

Figure 28. n estimates

When the learning rate is greater than 20, the train accuracy becomes 1.0. However, the validation accuracy is extremely slowly increased as the number of estimates in increased. The best number of estimates is 425.

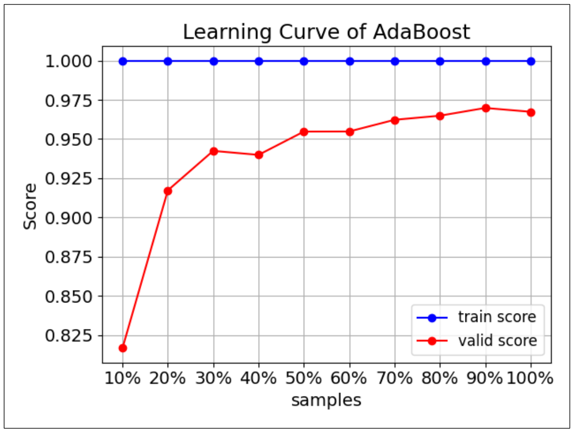

Figure 29. Learning Curve

The shape of Figure 29 is very similar to the Figure 28. It means that the model has relatively large variance even though the model has no bias. As we discussed in Decision Tree case, the number of samples are not sufficient for AdaBoost classifier.

B. *WiFi-localization*

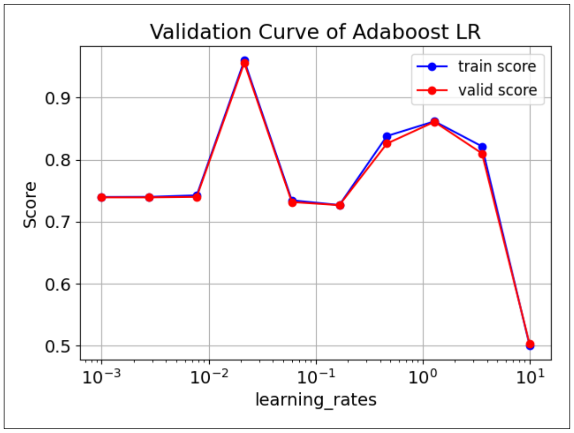

Figure 30. Learning rate

This figure shows very interesting shape. When the learning rate is 0.02154, the train and the validation accuracy has the highest value and the other values has significantly lower values than other models. Basically, the AdaBoost is basically designed for the binary classification and there is an extended version of AdaBoost algorithm for the multi-class classification [4]. So, Even though most learning rate values have good performance for the binary classification, however, a lot of learning rate values do not have good performance for the second dataset problem.

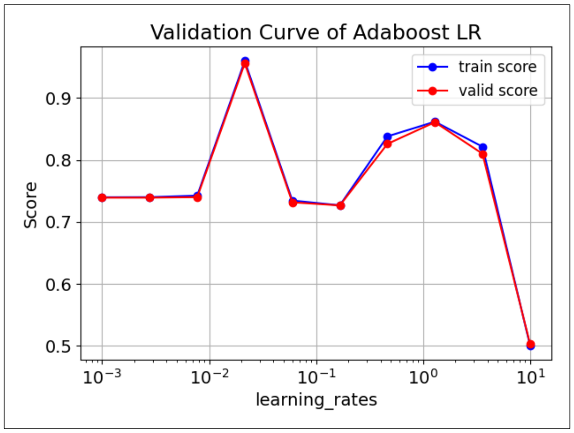

Figure 31. Number of estimates

This figure also shows very interesting shape. When the number of estimates is 51, the train and the validation accuracy are very high. Otherwise, the other values are significantly lower than the highest value. The reason is described above and this problem is multi-class classification problem. So, some specific numbers of estimators can be a good solution of this experiment.

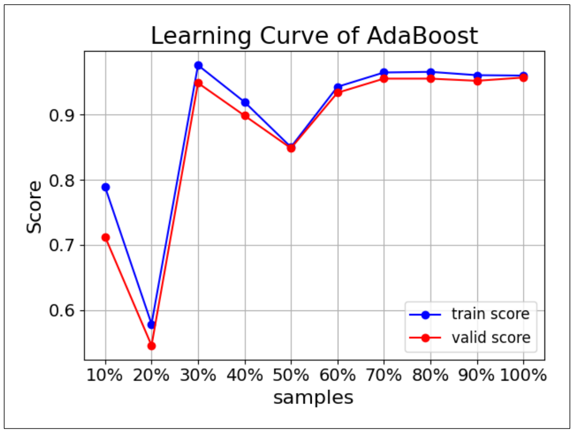

Figure 32. Learning Curve

This learning curve is quite different from the curve from Breast Cancer dataset. This shows significantly low variance, however it has a high bias. Three models were investigated so far and the AdaBoost for this dataset has the highest bias. The difference between training score between 1.0 is not negligible and the reason may be that this dataset has four class classification problem. Thus, it indicates that the AdaBoost classifier cannot achieve very high performance without other adjustment.

VII. K-NEAREST NEIGHBORS

For K-Nearest Neighbors, the P-value and the number of neighbors are selected as a hyperparameters. The P-value is the Minkowski distance and defined as: $(\sum_{i=1}^{n}|x_i - y_i|^p)^{\frac{1}{p}}$. And the number of neighbors is a kind of number of elements in a group to get a class of the corresponding points. So, the odd number of neighbors is preferred to avoid tie.

A. *Breast Cancer*

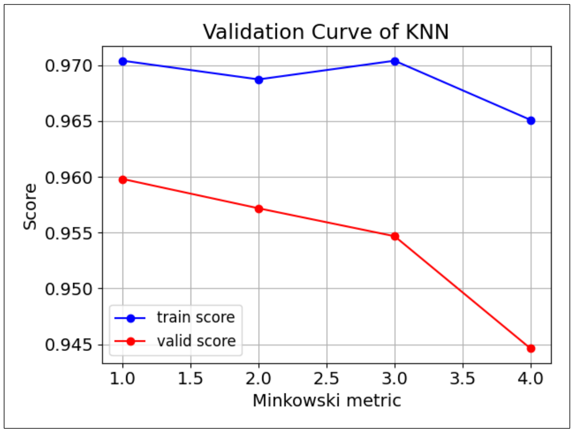

Figure 33. P-value

When the p-value (Minkowski metric) is 1.0, the validation accuracy has the highest value. The validation accuracy is decreased when the p value is increased. After fixing the p-value as 1, the number of neighbors is investigated.

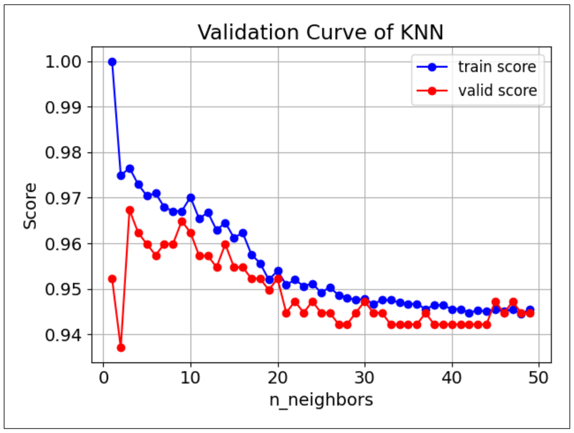

Figure 34. Number of neighbors

When the number of neighbors is 3, the validation accuracy has the highest value.

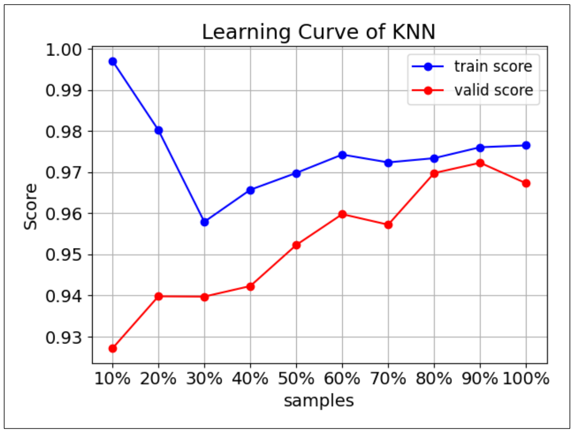

Figure 35. Learning curve

The Figure 35 shows that the KNN model has a relatively high bias, but it has a low variance. The train accuracy does not become 1.0 and the difference between the train accuracy and validation accuracy is smaller than 0.01. The more samples is needed for this model.

B. *Breast Cancer*

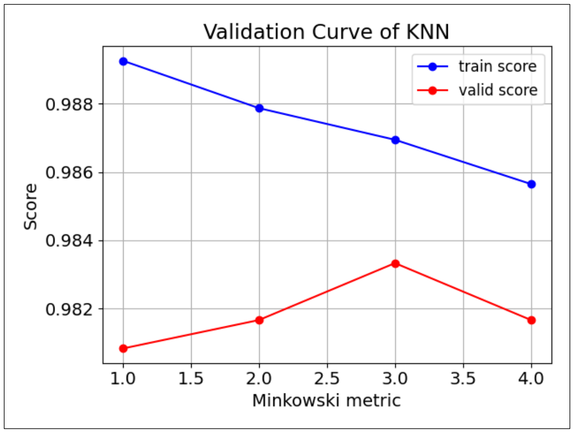

Figure 36. P-value

As we can see the Figure 36, when the Minkowski metric is 3, the validation accuracy has the best performance.

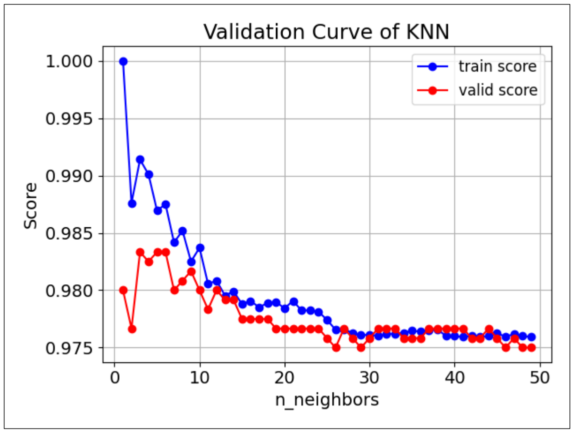

Figure 37. Number of neighbors

Figure 37 shows that the number 3 is the best number of neighbors.

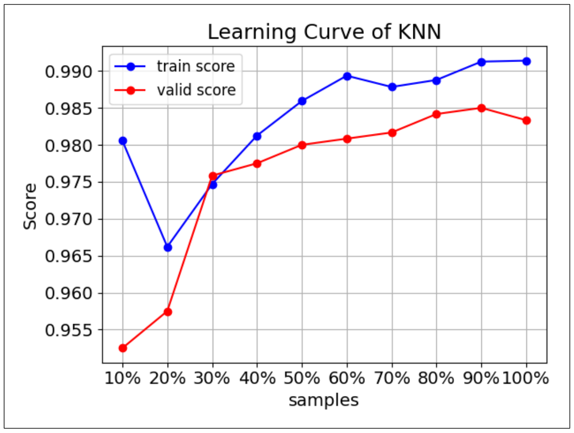

Figure 38. Learning curve

Compared to the Breast Cancer case, the train accuracy becomes 0.99 and the difference between the train accuracy and validation accuracy is similar to the previous case. Because, this dataset has larger data than the previous case. So, the KNN model for WiFi is relatively smaller variance than the KNN model for Breast Cancer. If the number of samples is added, the performance of KNN can be improved.

VIII. Result

The results from five classifiers are investigated and summarized in terms of the training time and accuracy. The training time is measured when the best classifier of each algorithm fits the train-sets. The time to calculate best hyperparameter of each algorithm is not included in the measured time. Also, the accuracy is based on the formula (1). There is no other metrics used for computing the accuracy.

A. *Training time*

1. Breast Cancer dataset

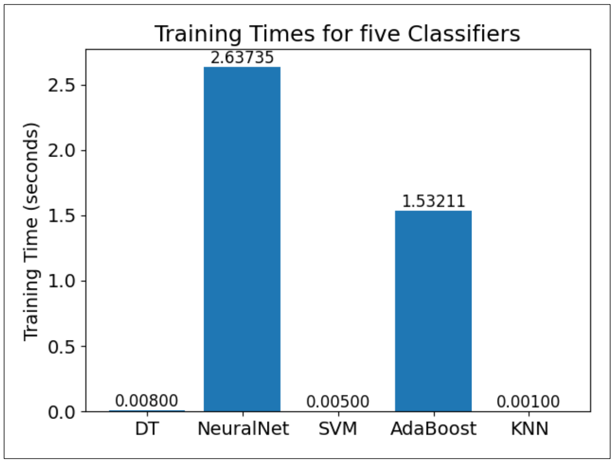

Figure 39. Training Time

In this experiment, the Neural Network and the AdaBoost classifier take relatively longer time than other classifiers. The Neural Network algorithm includes the optimization algorithm to find the best weights for each nodes. Also, the AdaBoost algorithm has to calculate and update weights of each weak classifiers. On the other hand, the Decision Tree, Support Vector Machine, and K-Nearest Neighbos are deterministic algorithm, thus, these can take significantly shorter time than the Neural Networks and AdaBoost.

2. WiFi-localization dataset

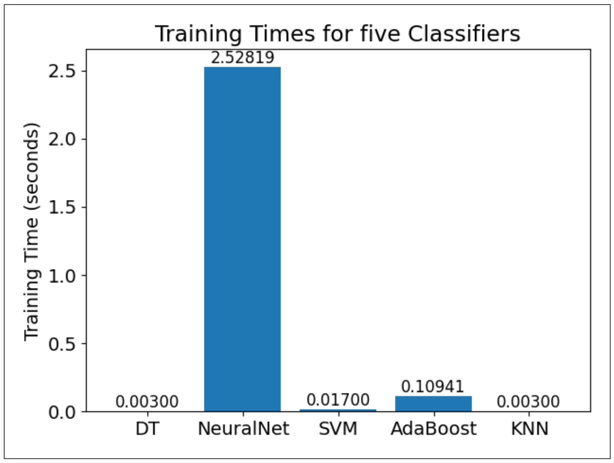

Figure 40. Training Time

In this experiment, the MLP and the AdaBoost classifier take relatively longer time than other classifiers. Also, the Adaboost has 425 estimators and the MLP has 10 nodes. This can be a reason that the Adaboost model takes longer time than the MLP model.

B. *Accuracy for each model*

1. Breast Cancer dataset

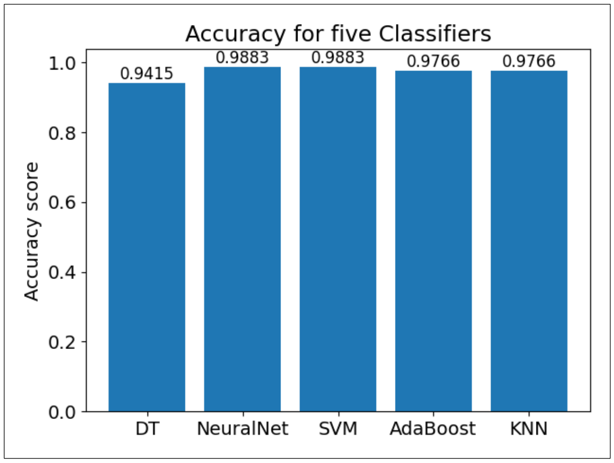

Figure 41. The Accuracy of each classifier

In this experiment, the Neural Network and Support Vector Machine have the same accuracy and the Decision Tree has a slightly lower accuracy than those classifier. As we discussed in Figure 3, it has a high variance and the more samples should be added for a better performance.

2. WiFi-localization dataset

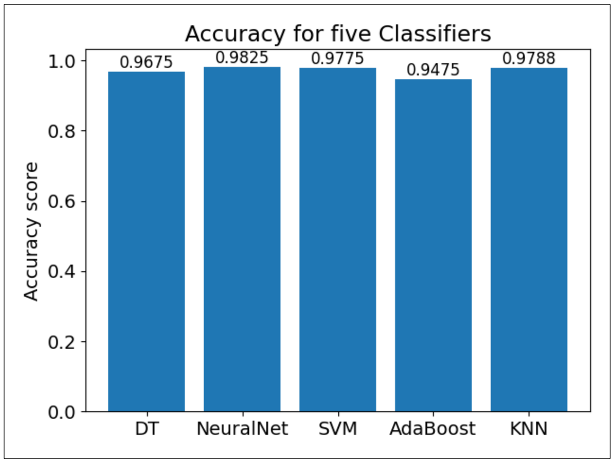

Figure 42. The Accuracy of each classifier

In this experiment, the Neural Network has the best performance and MLP and KNN have very similar accuracy. Also, the AdaBoost has the lowest accuracy among the five classifiers. It may cause that the AdaBoost is especially well-designed for binary classification. If the extended version of AdaBoost is applied, the performance will be highly improved.

IX. Conclusion

In this project, two datasets are investigated to compare the performances of five classifier. The summary of results are described the below table.

Classifier	Training time		Accuracy	
	Breast	WiFi	Breast	WiFi
Decision Tree	0.00800	0.00300	94.15%	96.75%
Neural Network	2.63735	2.52819	98.83%	98.25%
SVM	0.00500	0.04100	98.83%	97.75%
AdaBoost	1.53211	0.10941	97.66%	94.75%
KNN	0.00100	0.00300	97.66%	97.88%

In terms of accuracy, the Neural Network is the best classifier. The accuracy is greater than 98% in both cases. The worst classifier is the Decision Tree. As we can see the Figure 3 and 6, the decision tree classifier has high variance, so the accuracy is relatively lower than other classifiers. To improve the performance of the Decision Tree classifier, more data should be collected. The interesting classifier is the K-Nearest Neighbors and AdaBoost. In the case of KNN, when the number of samples in a dataset is small, the accuracy is relatively lower. However, when the number of samples is 2,000 (WiFi-localization), the accuracy is significantly improved. On the other hand, in the case of AdaBoost, the accuracy of multiclass classification is not very effective. This is the reason that the AdaBoost is basically well-designed algorithm for binary classification. In the case of multi-class classification, the extended version of AdaBoost should be used for this project.

In terms of training time, the Neural Network is the worst performance among the five classifiers. It takes at least 300 times of time to fit the train-set than other deterministic algorithm. If the dataset is small and the accuracy is very significant, the Neural Network should be definitely considered as a model. However, the dataset is very large and the accuracy is not very significant, other fast models can be a good model. Also, as we discussed, the Decision Tree model has still a relatively large variance. It means that there is a possibility to be improved.

Additionally, compared to the deterministic models, the Neural Network is relatively sensitive and finding a proper range of hyperparameter is significant. Considering the dataset, the range of hyperparameters should be determined carefully and it takes some time to consider it.

In summary, the NN is the best model in this project. However, the NN will not always be the best model for all cases. Considering the size of samples and the situation that Machine Learning algorithm is used for analysis, each model can be a strong candidate as a Machine Learning tool.

In this project, some topics should be improved in depth. First, the reason of AdaBoost application in the multi-class classification is not proved clearly. Second, the relationship between Minkowski metric and K-Nearest Neighbors should be investigated more.

References

[1] UCI Machine Learning Repository: Wireless Indoor Localization Data Set

[2] Breast Cancer Wisconsin (Diagnostic) Data Set | Kaggle

[3] Fanghui Liu, Xiaolin Huang, Chen Gong, Jie Yang, and Li Li. Nonlinear pairwise layer and its training for kernel learning. In Proceedings of the AAAI Conference on Artificial Intelligence, pages 3659–3666, 2018.

[4] Freund, Y. and Schapire, R. (1997). A decision theoretic generalization of on-line learning and an application to boosting. Journal of Computer and System Sciences 55 119–139. MR1473055