Machine Learning Classifier Comparison on Logistic Regression, Decision Tree, and Random Forest

Melody Lin
COGS 118A, Final Project Report

Abstract

The study examines logistic regression, decision tree, and random forest classifiers' performances across three binary classification datasets. Each classifier model involves three train-test partitions (20/80, 50/50, and 80/20) with multiple trials and cross validations to tune the hyperparameters. The model performances are compared and ranked based on testing accuracies. The experimental results suggest that logistic regression provides the most consistent and robust performance across three datasets, closely followed by random forest that also shows high testing accuracy and obvious trend of increasing accuracy with higher training set ratio. The decision tree has the lowest accuracies overall and shows signs of overfitting and low generalization. This study suggests that model performance depends on dataset characteristics, classifier choice, and data partitioning.

1 Introduction

This study conducts comparison of three supervised learning algorithms: logistic regression, decision tree, and random forest. The comparison is performed across three datasets each with three three partition ratios to examine the classifier performance that are mainly evaluated on their testing accuracies. The experiment carries out repeated trials for each partition ratio, hyperparameter tuning with K-fold cross validation, and generates training, testing, and cross validation accuracies. The classifier performance is compared within datasets, across partitions, and ranked overall based on average testing accuracy. The study provides a general view on how classifier choice, dataset characteristics, and training data size contributes to model performance.

2 Method

2.1 Classifiers and experimental setup

This project implements three supervised learning algorithms: Logistic regression, decision tree, and random forest. Each classifier was experimented on three partition ratios: 20/80, 50/50, and 80/20 train-test splits. In each partition, the experiment conducted three independent trials and reported their respective training and testing accuracies. Each classifier's hyperparameters were tuned and selected by K-fold cross validation on the training set with cross-validation accuracies reported.

2.1.1 Logistic regression

The logistic regression classifier uses gradient descent to optimize L2-regularized logistic loss and 5-fold cross validation to tune to the learning rate and regularization. The loss curves are also reported for each dataset.

2.1.2 Decision tree

The decision tree classifier is implemented using DecisionTreeClassifier from sklearn.tree with hyperparameters max_depth and min_samples_split tuned using 5-fold cross validation.

2.1.3 Random forest

The random forest classifier involves number of trees, maximum tree depth, and minimum samples per split as hyperparameters and is tuned using 5-fold cross validation. Compared to the decision tree classifier, random forest reduces overfitting and provides a more stable prediction by using ensemble learning method.

2.2 Datasets

The three datasets used in the experiment are obtained from the UC Irvine Machine Learning Repository, which are the Taiwanese Bankruptcy Prediction, Breast Cancer Wisconsin, and Blood Transfusion Service Center datasets. Feature scalings were applied when conducting logistic regression as the classifier is sensitive to feature magnitude.

2.2.1 Taiwanese Bankruptcy Prediction

The Taiwanese Bankruptcy Prediction dataset contains financial records of 6,819 companies from the Taiwan Economic Journal between 1999 and 2009.No missing values were found in the dataset, and all feature values are continuous. The dataset shows strong class imbalance with 6,599 companies that are non-bankrupt and 220 bankrupt as shown in Figure 1.

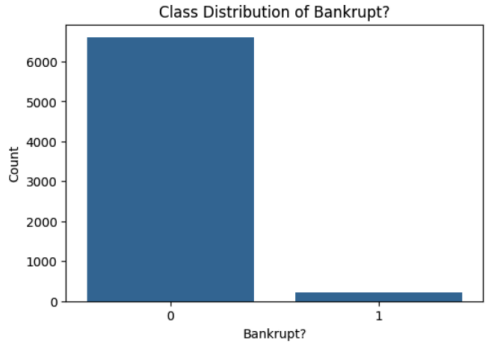

Figure 1: Taiwanese Bankruptcy Prediction dataset distribution

2.2.2 Breast Cancer Wisconsin (Diagnostic)

The Breast Cancer Wisconsin (Diagnostic) dataset contains 569 instances with a binary target variable that indicates the malignant or benign of a tumor. The target variable malignant is indicated as 1 and benign is indicated as 0 during data preprocessing. No missing values were present in the dataset. The class distribution is moderately imbalanced with more benign cases (357) than malignant cases (212).

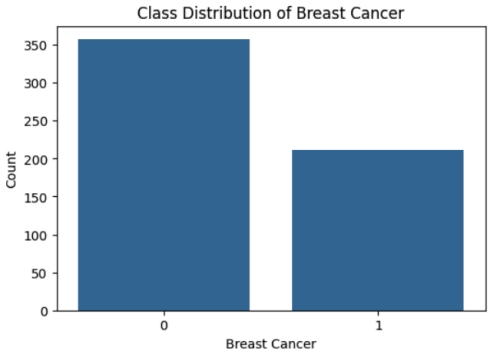

Figure 2: Breast Cancer Wisconsin (Diagnostic) dataset distribution

2.2.3 Blood Transfusion Service Center

The Blood Transfusion Service Center dataset contains 748 instances with binary target variables that indicate whether an individual donated blood in March 2007 in Hsin-Chu City, Taiwan with 0 indicating no donation

and 1 indicating donation. The dataset does not contain any missing values. The dataset shows class imbalance with non-donation cases (570) outnumbering donation cases (178).

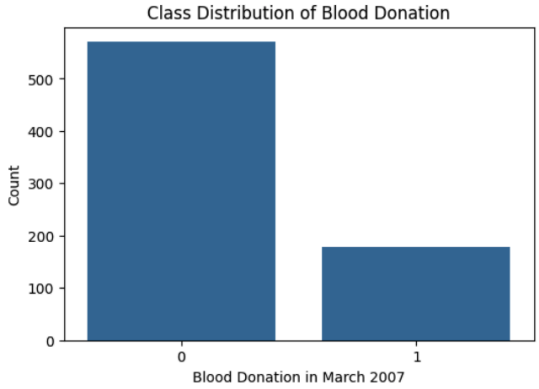

Figure 3: Blood Transfusion Service Center dataset distribution

3 Experiment and results

This section presents the experimental result of implementing the three classifiers across three datasets with three data partitions. The classifiers' performances are compared per dataset and partition and are ranked overall based on the average accuracy. Partition performances are also compared within a classifier to discuss how training size affects the classifier performance.

3.1 Classifier comparison per partition for each dataset

In the classifier comparison per dataset and partition, the best performing classifier is selected based on the highest average testing accuracy. Overall, random forest and logistic regression have the best performance across each dataset's partitions.

3.1.1 Bankruptcy dataset

In the bankruptcy dataset, random forest is the best performing classifier with the highest testing accuracy for both partition 80/20 and 20/80, and logistic regression is the best performing classifier for partition 50/50 according to Table 1. Figure 4 shows that all three classifiers have equally high average testing accuracy that are around 0.96 with logistic regression and random forest performing slightly better.

Table 1: Classifier comparison per partition for bankruptcy dataset

Partition	Classifier	Avg Train Acc	Avg Test Acc	Avg CV Acc	Best classifier
80/20	LOG	0.970975	0.963099	0.969630	
	DT	0.973480	0.965054	0.969264	
	RF	0.987046	0.968475	0.970608	BEST
50/50	LOG	0.968906	0.969501	0.966755	BEST
	DT	0.974773	0.964321	0.966267	
	RF	0.982595	0.968035	0.970960	
20/80	LOG	0.972609	0.964321	0.966980	
	DT	0.979946	0.963038	0.964287	
	RF	0.983859	0.968109	0.970409	BEST

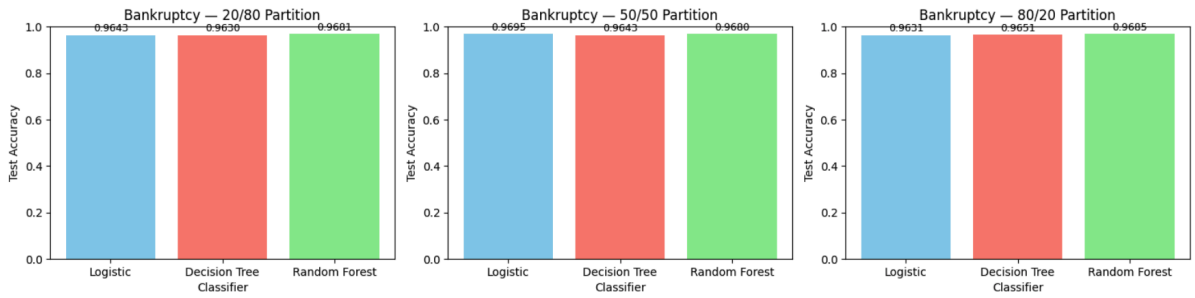

Figure 4: Classifier comparison per partition for bankruptcy dataset

3.1.2 Breast cancer dataset

In the breast cancer dataset, random forest is the best performing classifier with the highest testing accuracy for partition 80/20, and logistic regression is the best performing classifier for both partition 50/50 and 20/80 according to Table 2. Figure 5 shows that the decision tree classifier has notable lower testing accuracies compared to the other two classifiers across the three partitions.

Table 2: Classifier comparison per partition for breast cancer dataset

Partition	Classifier	Avg Train Acc	Avg Test Acc	Avg CV Acc	Best classifier
80/20	LOG	0.982418	0.961988	0.980220	
	DT	0.983883	0.929825	0.939194	
	RF	0.998535	0.964912	0.961905	BEST
50/50	LOG	0.978873	0.976608	0.977694	BEST
	DT	0.994131	0.943860	0.942460	
	RF	0.992958	0.955556	0.951838	
20/80	LOG	0.988201	0.970760	0.967852	BEST
	DT	0.997050	0.920322	0.914229	
	RF	1.000000	0.938596	0.952437	

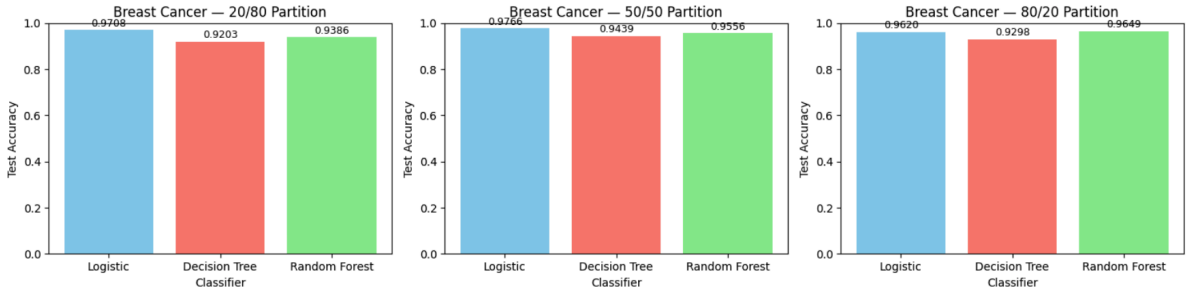

Figure 5: Classifier comparison per partition for breast cancer dataset

3.1.3 Blood donation dataset

In the blood donation dataset, random forest is the best performing classifier with the highest testing accuracy for partition 80/20, and logistic regression is the best performing classifier for both partition 50/50 and 20/80

according to Table 3. Figure 6 shows that the overall average testing accuracies for all three classifiers across the partitions are much lower than the previous two datasets, with the average testing accuracies around 0.76 to 0.79, with the decision tree classifier having notable lower testing accuracies compared to the other two classifiers across the three partitions.

Table 3: Classifier comparison per partition for blood donation dataset

Partition	Classifier	Avg Train Acc	Avg Test Acc	Avg CV Acc	Best classifier
80/20	LOG	0.775362	0.775556	0.776517	
	DT	0.797659	0.786667	0.775359	
	RF	0.812709	0.793333	0.775924	BEST
50/50	LOG	0.762923	0.783422	0.765538	BEST
	DT	0.817291	0.763815	0.775375	
	RF	0.836007	0.780749	0.791459	
20/80	LOG	0.744966	0.782415	0.744904	BEST
	DT	0.843400	0.739566	0.767203	
	RF	0.876957	0.765721	0.798697	

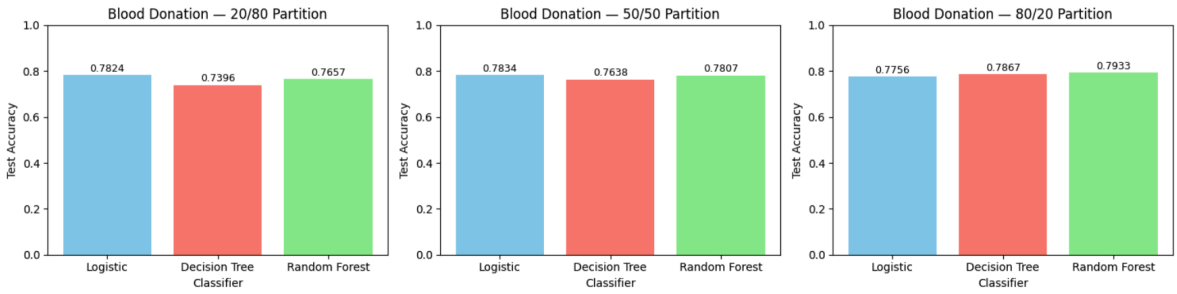

Figure 6: Classifier comparison per partition for blood donation dataset

3.2 Partition comparison per classifier

The partition comparison per classifier looks at how training size affects each classifier's performance on testing accuracy, which inspects whether testing accuracy improves as the training set gets larger. The training size has the most significant impact on random forest as the increase of training size reflects the increase of testing accuracy across all three datasets.

3.2.1 Logistic regression partition comparison

Logistic regression performs the best in 50/50 partition across all three datasets based on Tables 4, 5, 6. The classifier's testing accuracy is significantly lower for the blood donation dataset compared to the other two with 50/50 partition's testing accuracy average of 0.7834. Figure 7 shows that the average testing accuracy remains fairly stable with little variation across partitions and peaks at 50/50 partition.

Table 4: Logistic regression's performance per partition on bankruptcy dataset

Partition	Train Acc (avg)	CV Acc (avg)	Test Acc (avg)
20/80	0.972609	0.966980	0.9643

50/50	0.968906	0.966755	0.9695
80/20	0.970975	0.969630	0.9631

Table 5: Logistic regression's performance per partition on breast cancer dataset

Partition	Train Acc (avg)	CV Acc (avg)	Test Acc (avg)
20/80	0.988201	0.967852	0.9708
50/50	0.978873	0.977694	0.9766
80/20	0.982418	0.980220	0.9620

Table 6: Logistic regression's performance per partition on blood donation dataset

Partition	Train Acc (avg)	CV Acc (avg)	Test Acc (avg)
20/80	0.744966	0.744904	0.7824
50/50	0.762923	0.765538	0.7834
80/20	0.775362	0.776517	0.7756

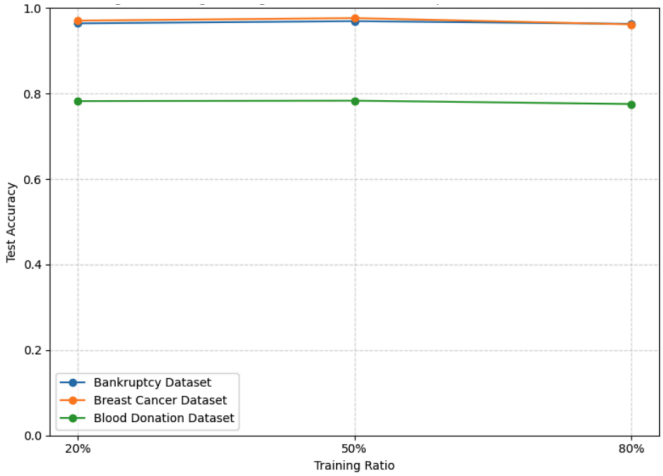

Figure 7: Logistic regression's performance per partition

3.2.2 Decision tree partition comparison

Decision tree performs the best in 80/20 partition in both bankruptcy and blood donation datasets, and performs the best in 50/50 partition in the breast cancer dataset according to Tables 7, 8, 9. The classifier's testing accuracy is significantly lower for the blood donation dataset compared to the other two with 80/20 partition's testing accuracy average of 0.7867. Figure 8 shows the trend of testing accuracy increasing as training set ratio increases for both the bankruptcy and blood donation dataset, and breast cancer dataset's testing accuracy peaks at 50/50 partition followed by slight decrease in 80/20 partition.

Table 7: Decision tree's performance per partition on bankruptcy dataset

Partition	**Train Acc (avg)**	**CV Acc (avg)**	**Test Acc (avg)**
20/80	0.979946	0.964287	0.9630
50/50	0.974773	0.966267	0.9643

80/20	0.973480	0.969264	0.9651

Table 8: Decision tree's performance per partition on breast cancer dataset

Partition	Train Acc (avg)	CV Acc (avg)	Test Acc (avg)
20/80	0.997050	0.914229	0.9203
50/50	0.994131	0.942460	0.9439
80/20	0.983883	0.939194	0.9298

Table 9: Decision tree's performance per partition on blood donation dataset

Partition	Train Acc (avg)	CV Acc (avg)	Test Acc (avg)
20/80	0.843400	0.767203	0.7396
50/50	0.817291	0.775375	0.7638
80/20	0.797659	0.775359	0.7867

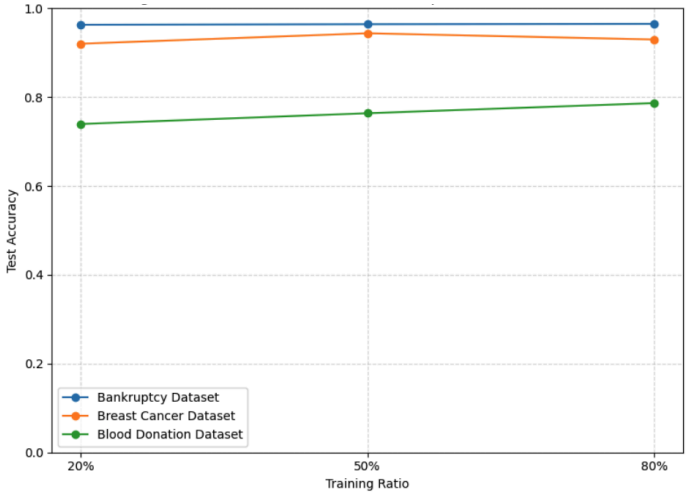

Figure 8: Decision tree's performance per partition

3.2.3 Random forest partition comparison

Random forest performs the best in 80/20 partition for all three datasets according to Tables 10, 11, 12. The classifier's testing accuracy is significantly lower for the blood donation dataset compared to the other two with 80/20 partition's testing accuracy average of 0.7933. Figure 9 shows the obvious trend of random forest's testing accuracy increasing as the training set ratio increases for all three datasets.

Table 10: Random forest's performance per partition on bankruptcy dataset

Partition	Train Acc (avg)	CV Acc (avg)	Test Acc (avg)
20/80	0.983859	0.970409	0.9681
50/50	0.982595	0.970960	0.9680
80/20	0.987046	0.970608	0.9685

Table 11: Random forest's performance per partition on breast cancer dataset

Partition	Train Acc (avg)	CV Acc (avg)	Test Acc (avg)
20/80	1.000000	0.952437	0.9386
50/50	0.992958	0.951838	0.9556
80/20	0.998535	0.961905	0.9649

Table 12: Random forest's performance per partition on blood donation dataset

Partition	Train Acc (avg)	CV Acc (avg)	Test Acc (avg)
20/80	0.876957	0.798697	0.7657
50/50	0.836007	0.791459	0.7807
80/20	0.812709	0.775924	0.7933

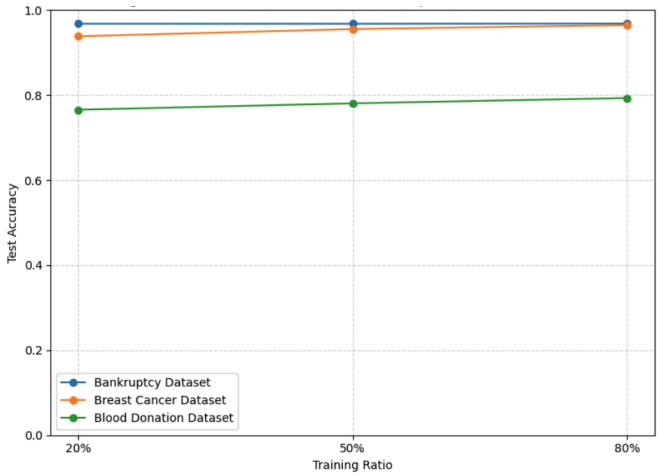

Figure 9: Random forest's performance per partition

3.3 Overall classifier performance ranking for each dataset

The experiment ranks the best performing classifier for each dataset based on their average testing accuracies.

3.3.1 Bankruptcy dataset classifier comparison

Random forest has the best performance for the bankruptcy dataset with the average testing accuracy of 0.968. However, all three classifiers perform equally well with average testing accuracies around 0.96.

Table 13: Classifier ranking on bankruptcy dataset

Rank	Classifier	Avg Test Accuracy
1	Random Forest	0.968206
2	Logistic Regression	0.965640
3	Decision Tree	0.964137

3.3.2 Breast cancer dataset classifier comparison

Logistic regression has the best performance for the breast cancer dataset with the average testing accuracy of 0.969.

Table 14: Classifier ranking on breast cancer dataset

Rank	Classifier	Avg Test Accuracy
1	Logistic Regression	0.969786
2	Random Forest	0.953021
3	Decision Tree	0.931335

3.3.3 Blood donation dataset classifier comparison

Logistic regression has the best performance for the blood donation dataset with the average testing accuracy of 0.78. The testing accuracies for all three classifiers in the blood donation are significantly lower than the other two datasets as the other two datasets have testing accuracies above 0.9.

Table 15: Classifier ranking on blood donation dataset

Rank	Classifier	Avg Test Accuracy
1	Logistic Regression	0.780464
2	Random Forest	0.779934
3	Decision Tree	0.763349

3.4 Overall classifier performance

The experiment provides an overall classifier ranking based on average testing accuracy across all datasets. According to Table 16, logistic regression has the best performance overall with the average testing accuracy of 0.905. Random forest performs equally well with a slightly lower average testing accuracy of 0.900.

Table 16: Overall classifier ranking

Rank	Classifier	Mean Test Accuracy
1	Logistic Regression	0.905297
2	Random Forest	0.900387
3	Decision Tree	0.886274

4 Discussion and conclusion

The experimental results suggest that the classifier performance depends on both the training-testing data partition, classifier characteristics, and dataset characteristics as no single classifier shows significantly high testing accuracy in all settings. One common characteristic for all classifiers is that they all show significantly lower testing accuracy in blood donation dataset compared to the other two datasets.

The Blood Transfusion Service Center dataset yields significantly lower testing accuracies for all three classifiers. The potential cause is its limited feature and weak predictive signal, making it hard for classifiers to distinguish and predict classes effectively. The class imbalance of the dataset may also contribute to this phenomenon as it may bias classifier accuracy.

Random forest shows high testing accuracy across all datasets with an obvious trend of higher accuracy with greater training set ratio. The trend is most obvious in the bankruptcy and breast cancer datasets. This

indicates random forest's strength in non-linear relationships and reducing overfitting with ensemble learning, which also justifies its higher accuracies compared to decision tree classification.

Logistic regression is ranked first among the three classifiers with its strong and stable performance across all datasets and partitions. Although logistic regression does not show a clear trend of increasing testing accuracy with increasing training set ration, it shows a relatively low variance across partitions than the other two classifiers, which make its predictions reliable with its robustness.

Decision tree classifier has the lowest accuracies, making it less well-performing than the other two classifiers. According to the experimental results, decision tree classifier tend to have higher training accuracies yet much lower testing accuracies, especially for the blood donation dataset. This suggests that the decision tree may have overfitted the data during model training and is less robust to datasets with imbalance classes.

One limitation is that since only three trials were executed for each partition, the experiment may not be able to account for full variability. Hyperparameter tuning was also limited for each classifier. In addition, the models may not be widely generalizable since only three datasets were being evaluated, including one dataset that results in significantly lower testing accuracies. Therefore, future projects could experiment on additional datasets, trials, and further hyperparameters optimization to build more confidence and generalizability of the models.

Overall, the experiment demonstrates that logistic regression excels in providing strong and consistent performance across datasets while random forest also shows high testing accuracies overall. The decision tree is more likely to overfit data and is less robust compared to the other two classifiers. Meanwhile, all three classifiers performing significantly worse in blood donation dataset also suggests that the classifier performance is also strongly impacted by the nature of the dataset; therefore, it is important to take dataset characteristics into account when selecting models in order to provide the best predictions.

References

[1] Caruana, R. & Niculescu-Mizil, A. (2006) *An empirical comparison of supervised learning algorithms.* Proceedings of the 23rd International Conference on Machine Learning, pp. 161–168.

AI Assistance Disclosure: Artificial intelligence tool (ChatGPT) was used with code debugging, general implementation guidance, and syntax clarification. All experiment setup, design, and result analysis were conducted independently by the author.