Intro to AI analasis

Zhentao Wei

June 27, 2025

Contents

1 What the papers evalueated **3**
1.1 Volurabilities in code 3
1.2 Improved tumor analasis by image fusion 4

2 Limited data model vs full data model **5**
2.1 Method 5
2.2 Results 5
2.3 Discussion 6
2.4 Appendix 7
2.4.1 Learnig curves 7
2.5 Boxplots 7

1 What the papers evalueated

Below will briefly describe the X and Y papers. What they did, and what they should have considered.

1.1 Volurabilities in code

The paper titled "Just another copy and paste? Comparing the vulnerabilities of ChatGPT generated code and StackOverflow answers" by Sivana Hamer and Marcelo d'Amorim and Laurie A. Williams [**?**], is based around volurabilities that may be preasent in code you find online. This paper choose do a statistical analasis on the amount of CWEs (Common Weakness Enumeration) volurabilities there are from 2 sources.

They choose to compare between OpenAI's ChatGPT and the information sharing website called Stack overflow. Each test sample is a question from StackOverflow, which was fed into ChatGPT, and the result was the StackOverflow answer and the ChatGPT answer. Since the same question was given to both StackOverflow and ChatGPT, the samples were independent and paired. The sample results were collected in a categorirized way of either having at least one volurability or no volurability. Thus, with a paired and categorirized sample, then the perfect fit to check if there was a difference between ChatGPT and StackOverflow, is a Chi-square test, which their paper did. The researchers found that the Chi-square test received a p-value of 0.87, which showed no evidence of statistical significance of either ChatGPT or StackOverflow had more or less volurabilities.

The paper later continues with the statistical testing but this time check if either ChatGPT's or StackOverflow's code snippets had more volurabilities in total. They choose to do a paired another Chi-square test, which is the correct choice, since this is just a count of occurances and nothing that can be averaged. The test resulted in a p-value of 0.02, which is statistical evidence and can reject the null-hypothisis. Which means that either ChatGPT or StackOverflow have more volurabilities in their code snippits.

1.2 Improved tumor analasis by image fusion

In the paper titled "Investigating the Role of Image Fusion in Brain Tumor Classification Models Based on Machine Learning Algorithm for Personalized Medicin"[**?**], they experiment with merging multiple images before predicting tumors in the images. Their theory is merging multiple images of the same subject, provides more data for the model, which in turn would provide better prediction results. This paper calculated various model performance metrics, such as accuracy, precision, recall, specificity, and F1 score. Which were later used to create a confusion matrix. This paper contains various statistical analasis flaws and conflicting information.

The first inconsistancy is "Features are extracted from 200 SPECT images collected from a medical database, and these features are given as input to SVM, KNN, and decision tree classifiers. The results of classifiers are tabulated in Table 3." but in table 3, the total of $TP + FP + TN + FN = 400$, which is inconsistant to the supposed 200 images. The second flaw is missing paired tests between the methods. Usually there would be some paired test to compare the performance between the methods, but this paper seems to be missing one of such paired test. An example of a paired test for the 3 methods is a McNemar test, but it seems that this paper lacks one of such tests and just shows some numbers.

Since they are proposing their own model, they are missing some kind of cross-validation, or at least it is not included in the paper. Without cross-validation, it is likely possible for the model to get overfitted to the data, and thus not perform well on new data. With an accuracy of 96.8%, it is likely that the training data might have included some testing data.

2 Limited data model vs full data model

There have been made 2 models, one with only 3 input features, and the other with all features. They were then statistically analized, and compared against eachother.

2.1 Method

The dataset provided is a subset of the EmoPairCompute dataset. This dataset was processed by one-hot encoding the catecorical data. The data was then divided into 10 chunks or folds and each fold was then further divided into 5 inner... Using kfold cross-validation ensures less bias in the data used for training and testing a model. For each outer fold, each permutation of [1, 2, 4, 8, 16, 32] of amount of neurons in the hidden layer was tested. The model's loss function is defined by

$$(insertfnctionhere)$$

The best model with lowest loss for each outer fold is chosen as the final model for the outer fold. By going through each permutation, the best amount of neurons will be choosen for each outer fold. This model permutation was done twice, where the first model only had the input features of "Puzzler", "Phase_phase1", "Phase_phase2", and "Phase_phase3", and the second model had all features in the dataset provided to it.

Now with 2 lists of fold scores, based on mean square error, where lower is better, the 2 models can be statistically compared. 2 boxplots will be made to briefly visualize the means and distribution of both the models kfold scores. The parametric paired t-test and the non parametric wilcoxon test, will be used to test for statistal significance of wether one or the other model is better.

2.2 Results

The scores producted from the kfolds is table 1 and table 2.

Index	Limited features	Full features
1	0.809	0.490
2	0.973	0.847
3	0.274	0.448
4	0.707	0.605
5	0.814	0.589
6	0.703	0.826
7	0.604	0.581
8	0.942	0.903
9	0.600	0.606
10	0.733	0.774

Table 1: Features comparison

Index	Limited hidden units	Full hidden units
1	2	1
2	2	1
3	2	1
4	4	1
5	1	2
6	1	1
7	1	1
8	4	1
9	8	2
10	2	1

Table 2: Hidden-unit settings

The mean for the limited model and the full model are 0.716 and 0.667, respectively. For the standard deviation, it is 0.199 for the limited model and 0.158 for the full model. It shows

that the limited model has slightly more error than the full model. Nevertheless, a t-test and/or a wilcoxon has to be done to find if this difference has any statistical significance. In this case both a parametric paired t-test and a non-parametric wilcoxon test was performed.

The result of the pared t-test is a p-value of "0.3339". At a default significance of 5%, the null null-hypothisis cannot be rejected and no significance was found.

As for the wilcoxon test, the p-value 0.4316, which is slightly more significant than the test with the t-test, though not by much. This p-value also fails to reject the null and therefore there is no evidence of either model performing better.

2.3 Discussion

Eventhough, the limited model was provided

2.4 Appendix

2.4.1 Learnig curves

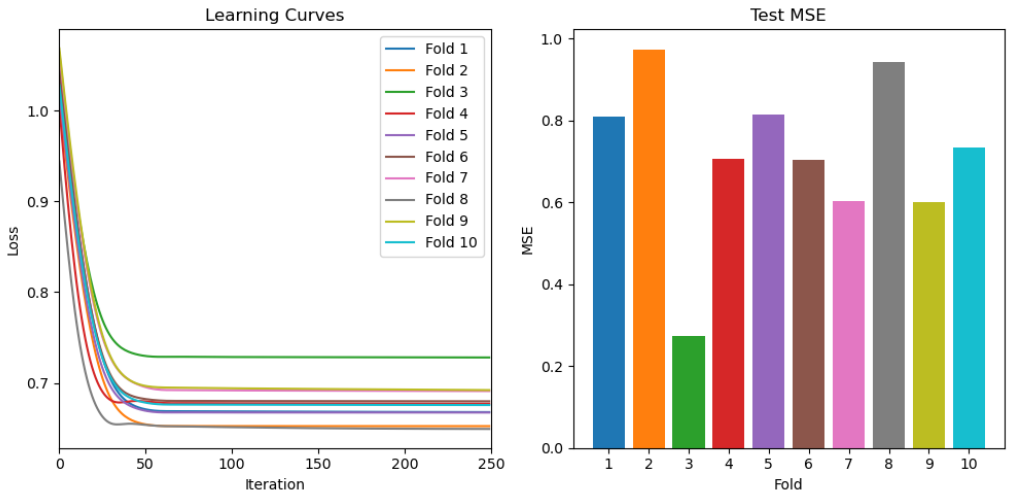

Figure 1: Limited model's learnung curves

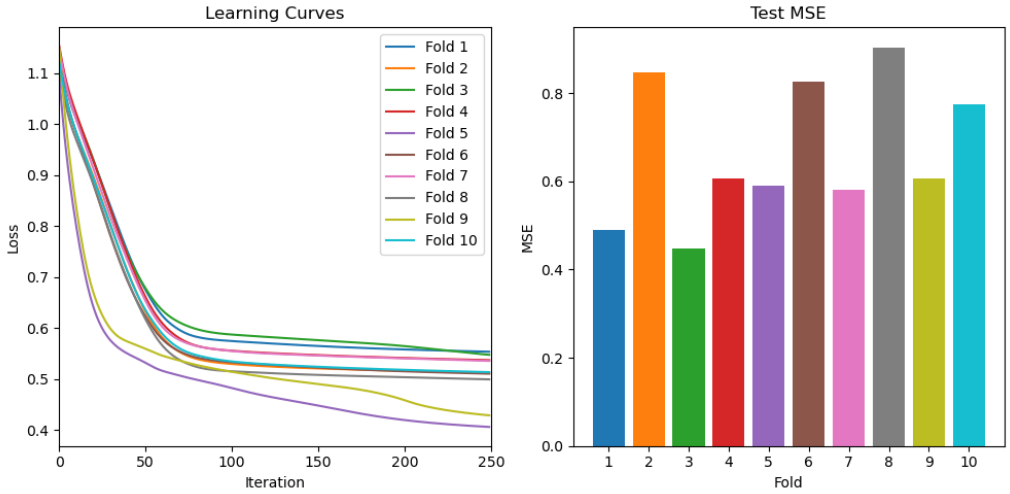

Figure 2: Full model's learnung curves

2.5 Boxplots

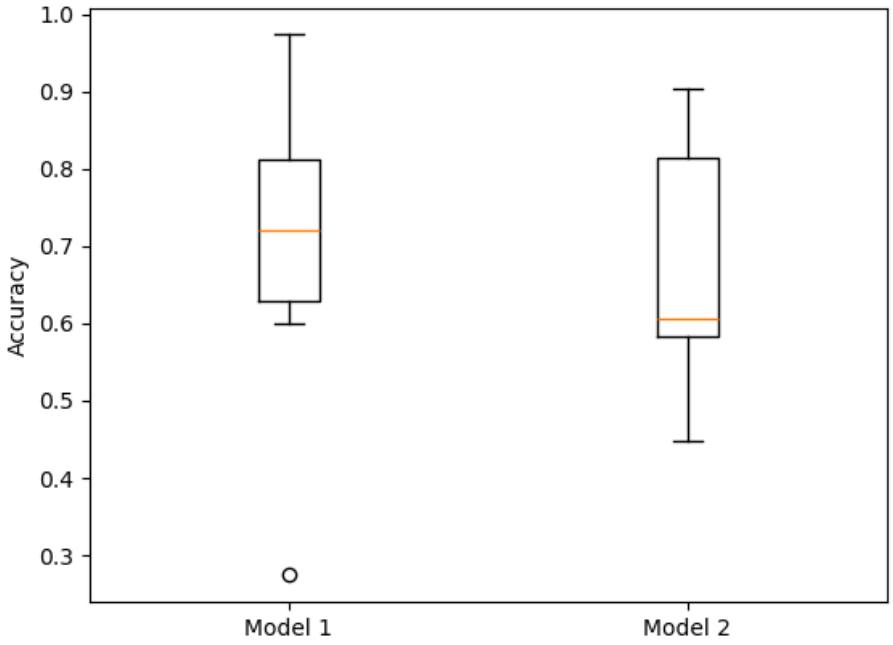

Figure 3: Boxplots of both model's fold scores