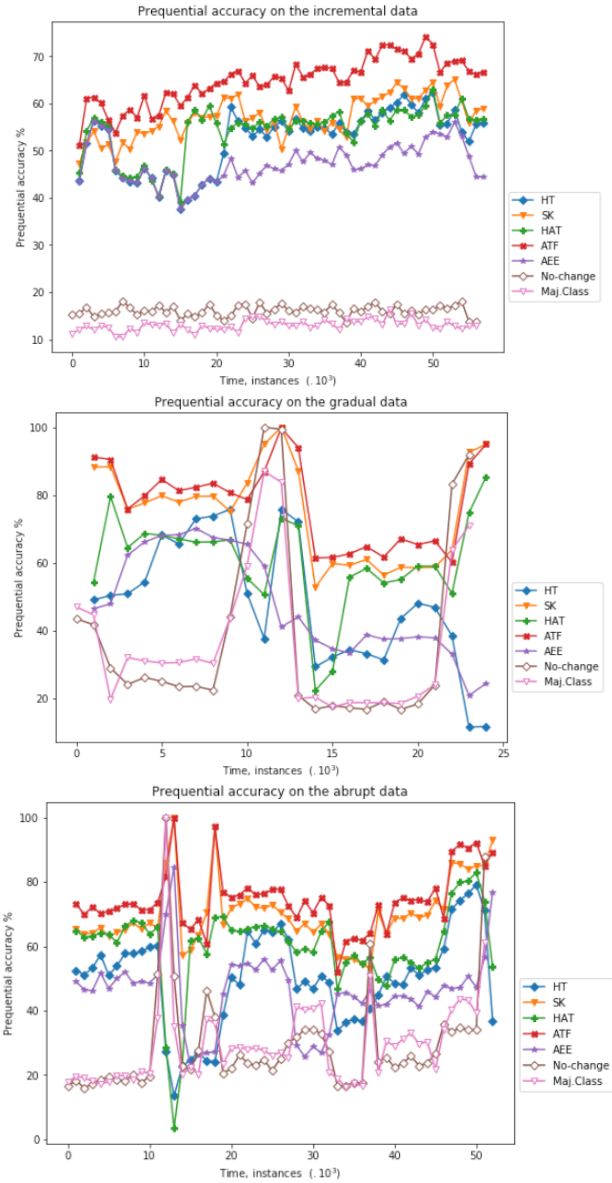
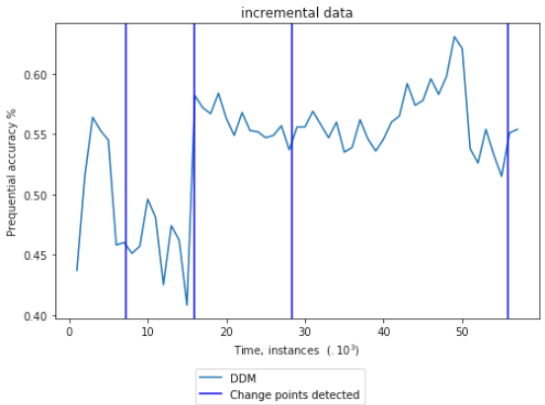
CSI 5155 Assignment 4 Report

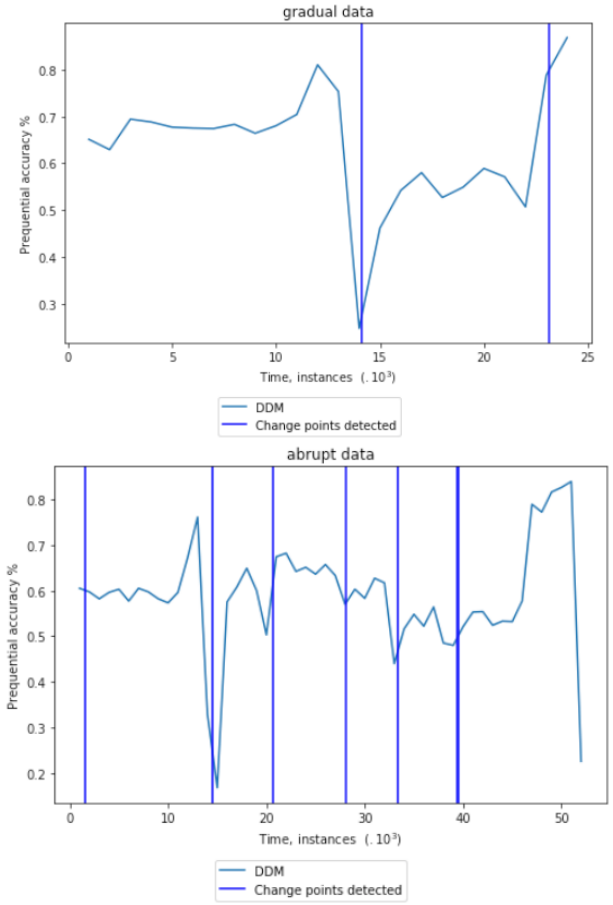
Name: Zhikun (Sam) Yuen Student ID: 300323972

Q3



Q5





Q6

	No-change	Maj.Class	Hoeffding Trees	SAM-KNN	Hoeffding Adaptive Trees	Adaptive Radom Forest	Additive Expert Ensemble	Hoeffding Tree + DDM
Incremental	0.16	0.13	0.51	0.58	0.53	0.64	0.47	0.54
Gradual	0.38	0.37	0.46	0.76	0.62	0.79	0.46	0.63
Abrupt	0.29	0.30	0.49	0.71	0.61	0.75	0.45	0.59

Q7

I try 8 approaches for the three datasets. For the models provided by Scikit-Multiflow, I use the default parameters except the pretrained = 1000 and window_size for prequential accuracy = 1000. For the majority class classifier, I use a deque with max length = 1000 as a buffer to save the last 1000 data as training data.

In the three datasets, models on Incremental data usually have lower prequential accuracy than the other two datasets. No-change classifier and majority classifier are always have the lowest performance. Sam-KNN and Adaptive Random Forest are usually have the best accuracy. However, another ensemble model, additive expert ensemble, is very bad. Hoeffding Adaptive Trees is better than Hoeffding Trees in three datasets. Using DDM drift detection with Hoeffding Tree can boost the performance, which are better than the one without drift detection in all three datasets.

The all the models in incremental dataset except for no-change and majority class classifier, their prequential accuracy gradually with respect to time. Also, at the beginning, there is a sudden increasing for the prequential accuracy around 10 – 20 % but then drops rapidly. The

performance of no-change classifier and majority class classifier is remain stable in incremental data, keeping in 1x%. The performance of Hoeffding adaptive trees is similar to

In gradual data, there are two huge jump for the accuracy for all models in around 12,000 time instances and at the end, even though no-change classifier and majority class classifier achieve > 90% accuracy at these time step. If we add a drift detector for the Hoeffding tree classifier, the accuracy will increase after the change is detected. However, the original Hoeffding tree classifier does not.

In abrupt data, the prequential accuracy of all models increase gradually, including the no-change and majority class classifier. There is also some sudden change for the prequential accuracy in the data around 12,000 time instances.

For the drift detection, I use the DDM(). Followed by the document, I input either 0 or 1 into the detector. 0 means correct classification, 1 means miss-classification. There is a detected_warning_zone() function in the detector. I use it to detect the warning zone. When the warning zone occurs, I will training a new Hoeffding Tree Classifier. Once the drift is detected, I replace the current classifier with the new pretrained one and reset the detector. This can improve the classifier's prequential accuracy. However, not all shift can be detected, only some very obvious shift can be detected by DDM().

Q8

Similar to the paper, the performance of no-change classifier and majority class classifier are also the worst comparing to other models. I use Hoeffding Trees, SAM-KNN, Hoeffding Adaptive tree, RAT and Additive Expert Ensemble. The paper uses NB, VFDT, Lev.Bag and ARF. The best models in the paper are usually Leveraging Bagging with 10 VFDT in the ensemble and Adaptive Random Forest. Both of them are ensemble methods. In my result, the best models are also Adaptive Random Forest and Sam-KNN.

For the result in incremental data, most of my models have a gradually increase in prequential accuracy except the no-change and majority class classifier. In the paper, only two models Leveraging Bagging with 10 VFDT and Adaptive Random Forest have gradually increase of the performance. Both no-change classifier and majority class classifier in my result and the paper are the lowest accuracy in this dataset all the time and no change of the accuracy.

For the abrupt data, our results have some similar point. In around 12,000 time instance, most our models have a sudden increase in the performance. All my models in the abrupt data has gradually increase in the performance, but only no-change classifier, VFDT, ARF and Lev.Bag have a gradually increase in the performance.

For the gradual data, we have several similar points. At time instance 12,000 there is a rapid increase for most of our models in the prequential accuracy. This is also happen at the end again. For other part of the stream, the performance of all models are very stable.

For the drift detection part, I use DDM but the best drift detector in the paper are ADWIN and STEPD. In the paper, there are many points detected by the STEPD drift detector, but my DDM can only detect some obvious changes. In the paper, they also used DDM, in their result in

Table 6 (in the paper), DDM also detects very few changed detected in the three datasets (< 10 times). However, it used a naïve bayes classifier instead of a Hoeffding Tree Classifier.