Assignment 4

Student ID: 300318105 Student Name: Snehal Sudhir Bhole

Report:

In this Assignment, I am using scikit-multiflow to work on a stream of data collected from a laser sensor built with low-cost components to remotely capture information about flying insects, in order to aid in intelligent insect trap design. In this assignment, I am using Insects-Abrupt-Balanced, Insects-Incremental-Balanced, and Insects-Gradual-Balanced streams with online stream models Hoeffding Trees, SAM-KNN, Hoeffding Adaptive Trees, AdditiveExpertEnsembleClassifier and Adaptive Random Forest Classifiers with no-change classifier, majority change classifier.

For the first part of the assignment, the no-change and the majority class classifier are initially trained against the first 1000 samples over a window of 1000 instances. No-Change classifier predicts the class label as a previous class label, irrespective of the current target label. In contrast, Majority Class Classifier predicts class labels based on the majority class of 1000 samples in a window. Figure [1],[2],[3] shows the results obtained for Insects-Incremental-Balanced, Insects-Gradual-Balanced and Insects-Abrupt-Balanced, respectively.

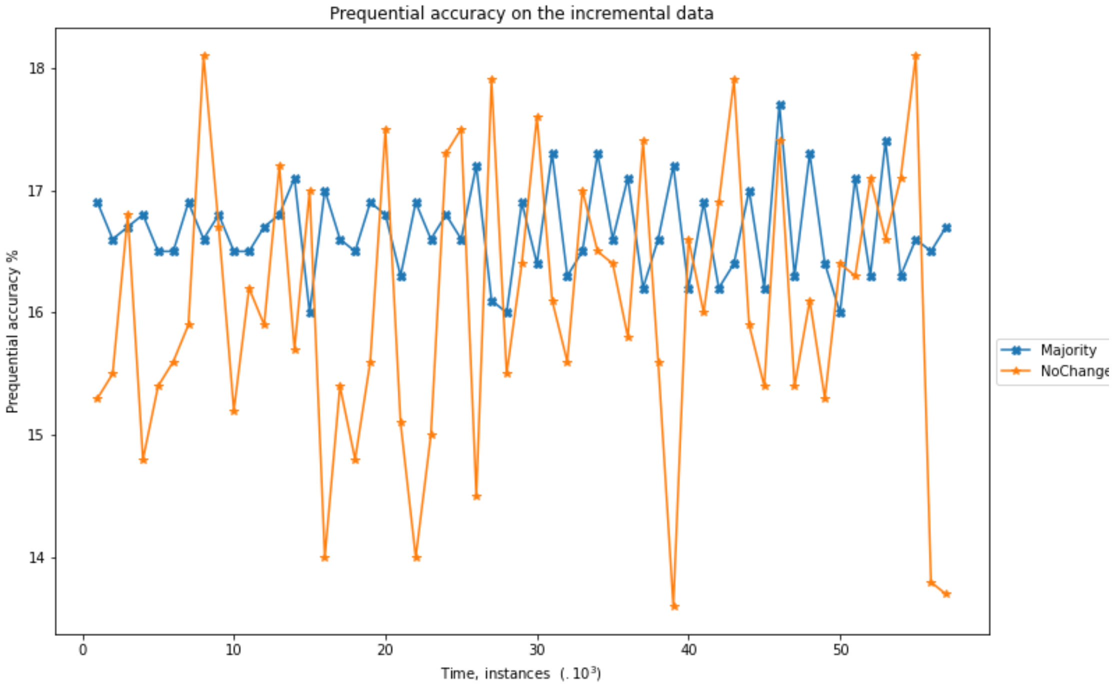


Figure 1: No-Change and Majority change classifier on Insects-Incremental-Balanced data

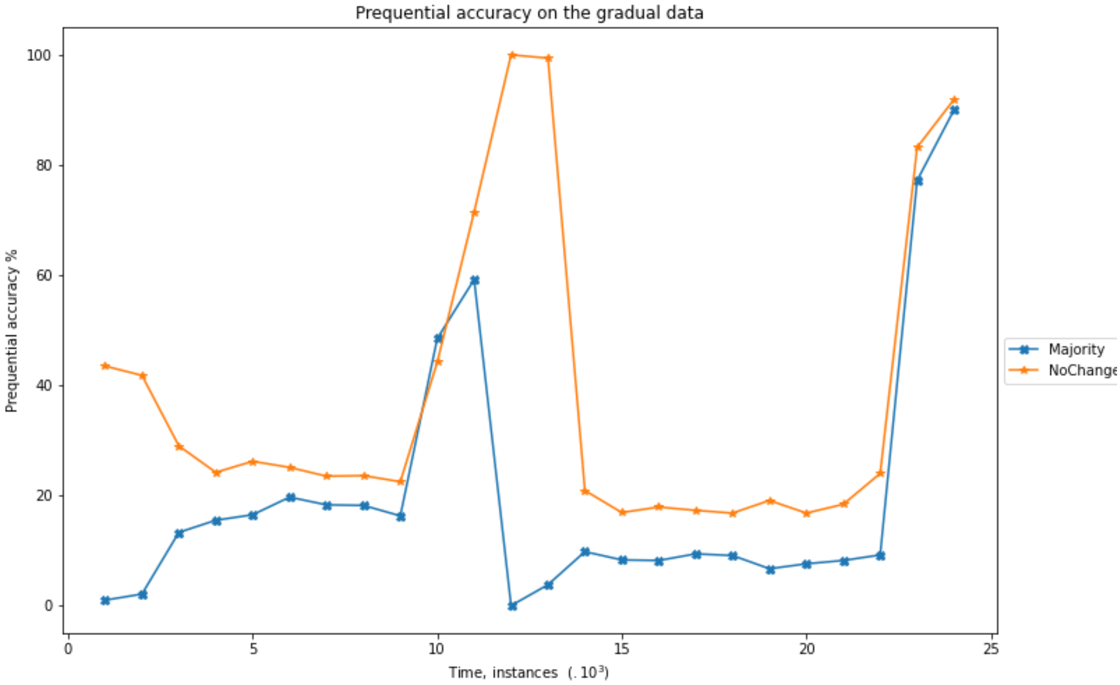


Figure 2: No-Change and Majority change classifier on Insects-Gradual-Balanced data

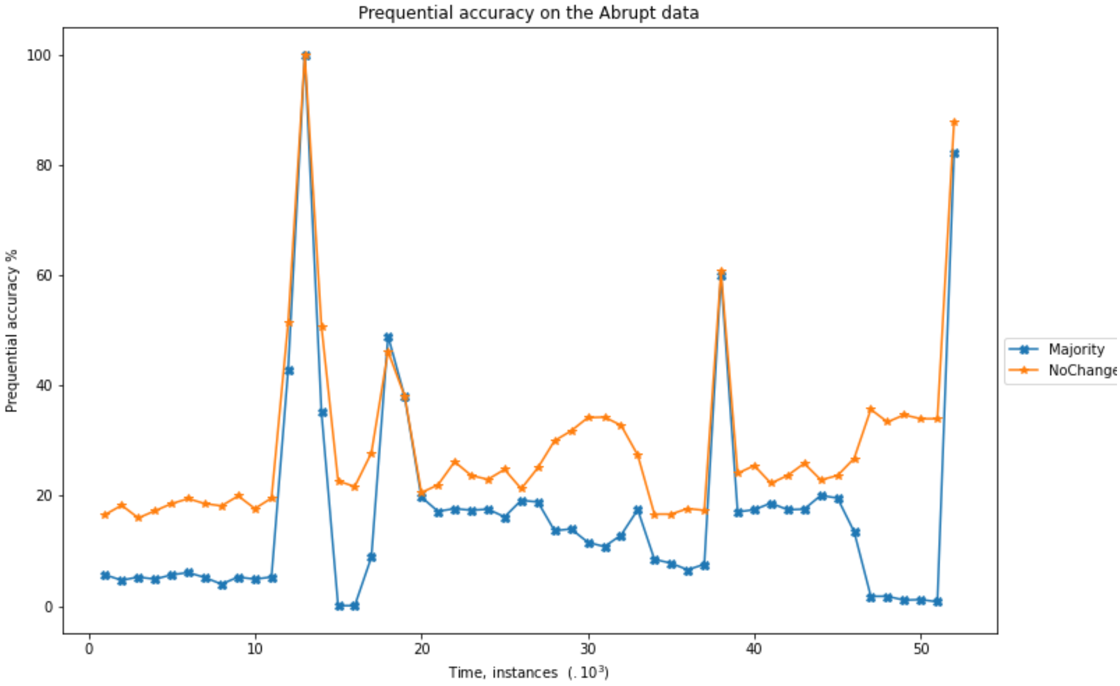
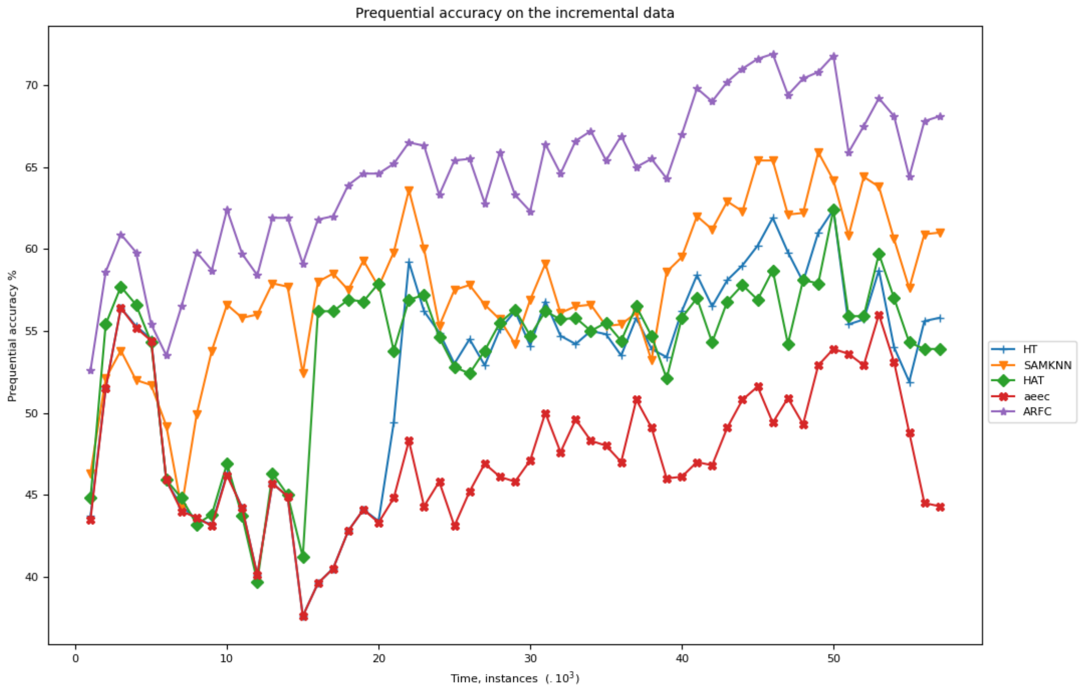
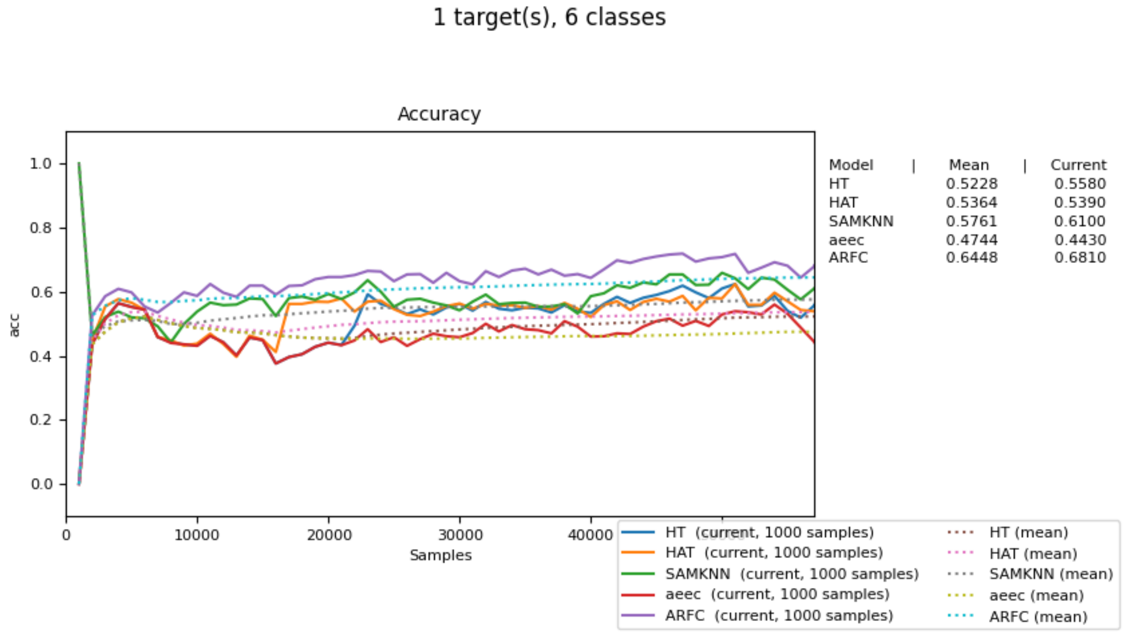


Figure 3: No-Change and Majority change classifier on Insects-Gradual-Balanced data

For the second part of the assignment, I am using 2 ensemble-based models Additive Expert Ensemble Classifier, Adaptive Random Forest Classifiers, Hoeffding Trees, SAM-KNN, and Hoeffding Adaptive Trees models.
We are comparing the prequential accuracies of all the models trained on all 3 datasets.





```
Model     |   Mean    |   Current
HT            0.5228      0.5580
HAT           0.5364      0.5390
SAMKNN        0.5761      0.6100
aeec          0.4744      0.4430
ARFC          0.6448      0.6810
```

Figure 4: Hoeffding Trees, SAM-KNN, Hoeffding Adaptive Trees, AdditiveExpertEnsembleClassifier and Adaptive Random Forest classifier on Insects-Incremental-Balanced data

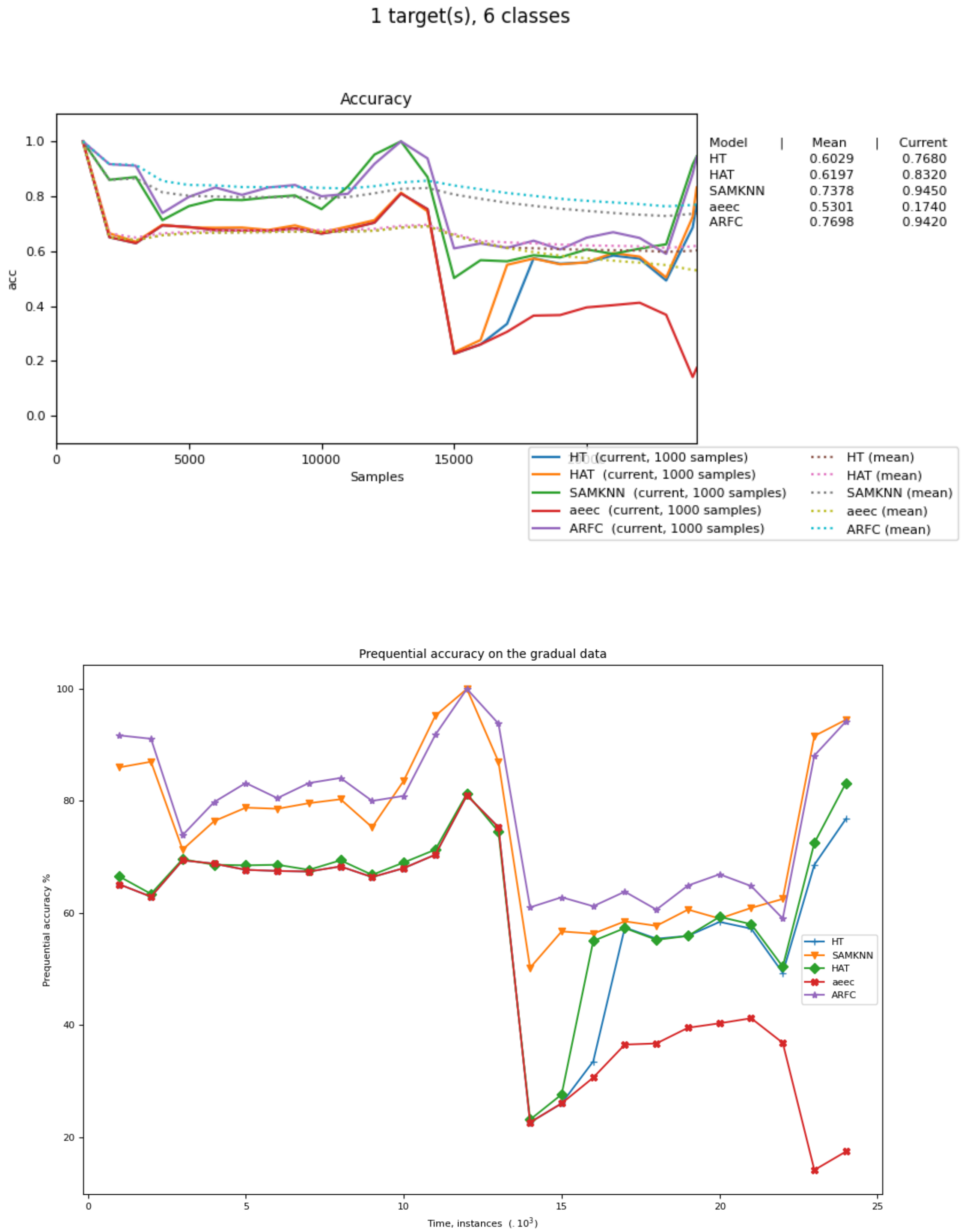
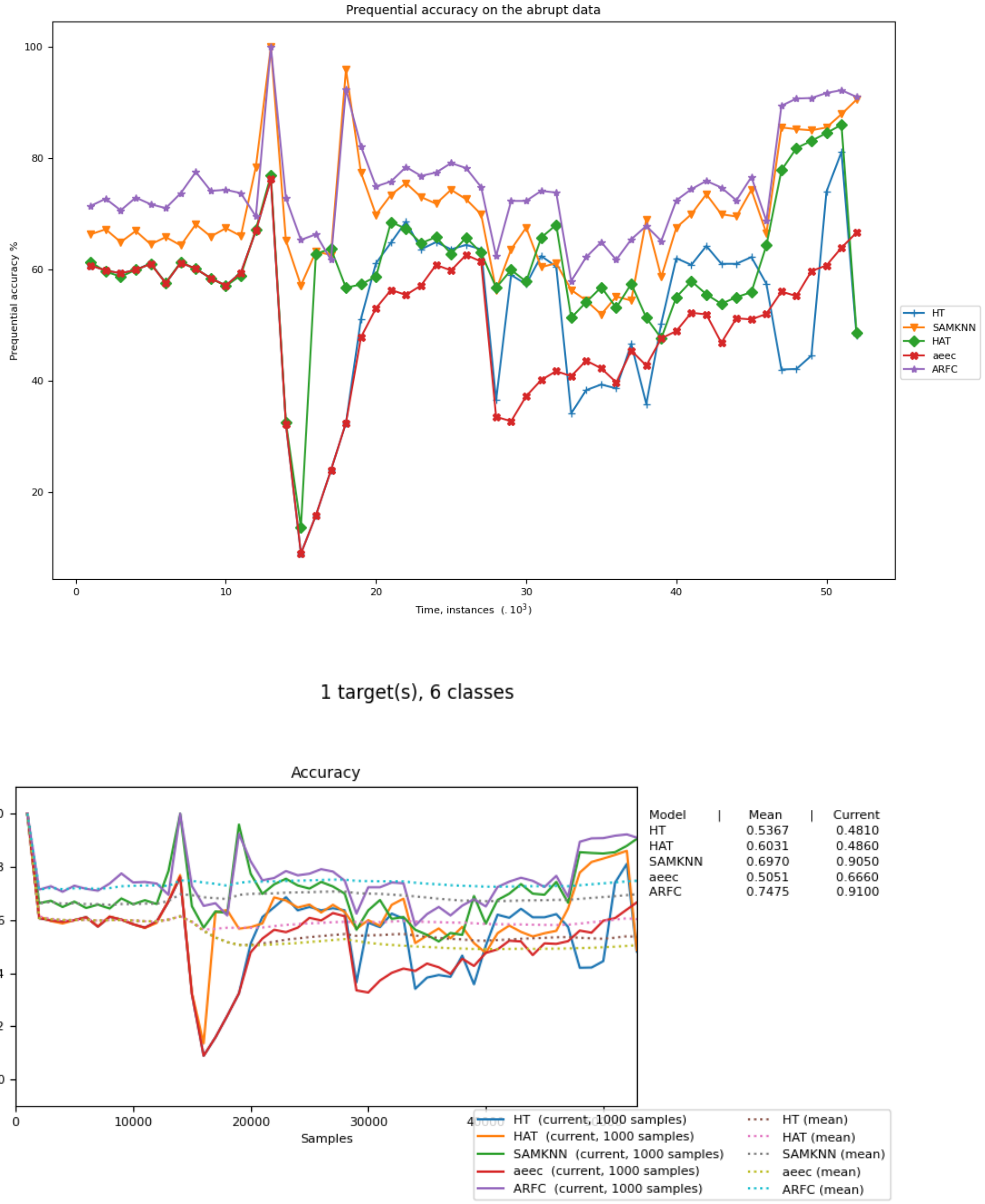


Figure 5: Hoeffding Trees, SAM-KNN, Hoeffding Adaptive Trees, AdditiveExpertEnsembleClassifier and Adaptive Random Forest classifier on Insects-Gradual-Balanced data

In part 3, we are using Adwin Drift Detection with Hoeffding Trees which detects the change in the input stream.ADWIN (ADaptive Windowing) is a statics-based algorithm which monitors the input stream, and analysis the average of statistics over 2 windows at different points. If the absolute difference between 2 surpasses a pre-defined threshold, a change is detected and data before that time is discarded.



Model	Mean	Current
HT	0.5367	0.4810
HAT	0.6031	0.4860
SAMKNN	0.6970	0.9050
aeec	0.5051	0.6660
ARFC	0.7475	0.9100

Figure 6: Hoeffding Trees, SAM-KNN, Hoeffding Adaptive Trees, AdditiveExpertEnsembleClassifier and Adaptive Random Forest classifier on Insects-Abrupt-Balanced data

Figures [4],[5],[6], show the results obtained for Insects-Incremental-Balanced, Insects-Gradual-Balanced and Insects-Abrupt-Balanced, respectively for the models Hoeffding Trees, SAM-KNN, Hoeffding Adaptive Trees, AdditiveExpertEnsembleClassifier and Adaptive Random Forest classifier.

Dataset	**No Change**	**Majority Class Classifier**	**Hoeffding Trees**	**SAM-KNN**	**Hoeffding Adaptive Trees**	**Additive Expert Ensemble**	**Adaptive Random Forest**	**ADWIN with Hoeffding Trees**
Incremental	0.1610	0.1667	0.5228	0.5761	0.5364	0.4744	**0.6448**	0.5296
Gradual	0.3632	0.1733	0.6029	0.7378	0.6197	0.5301	**0.7698**	0.5983
Abrupt	0.2923	0.1691	0.5367	0.6970	0.6031	0.5051	**0.7475**	0.5374

Table 1: Prequential Accuracies Obtained for the models

When we compare the graphs derived from all three datasets, we can see that the change detection for abrupt data with drift detection is more significant when compared with gradual and incremental data. Moreover, Algorithms tend to show more stability while training on gradual and incremental data compared to abrupt data. Figures [7],[8],[9], show the results obtained for Insects-Incremental-Balanced, Insects-Gradual-Balanced and Insects-Abrupt-Balanced, respectively for the models Hoeffding Trees with drift detection method applied to it.

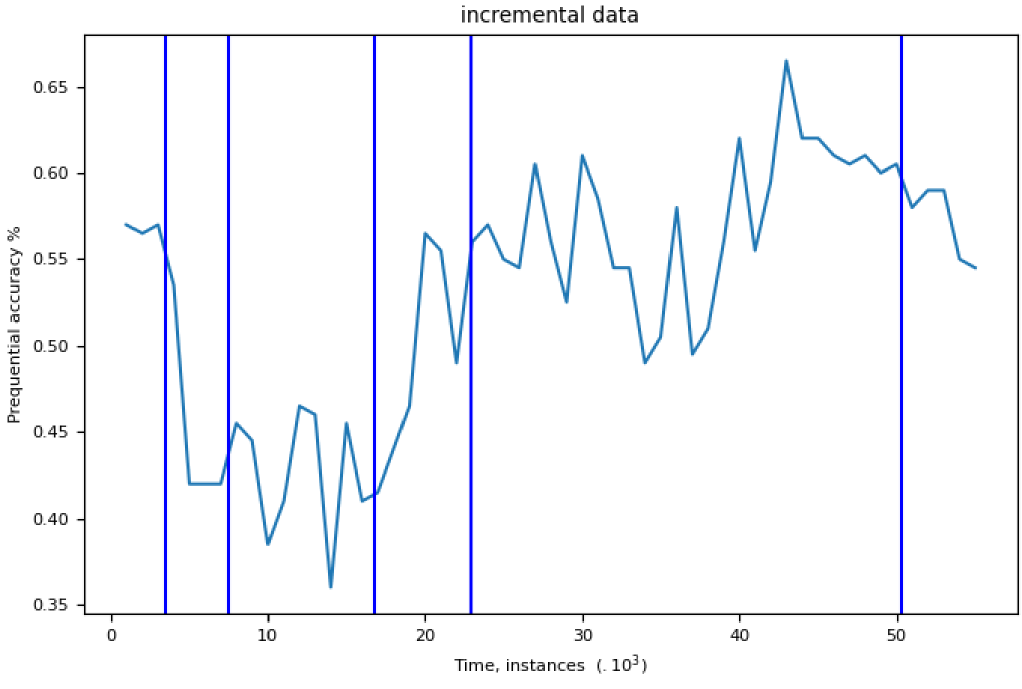


Figure 7: ADWIN Drift Detection with Hoeffding Trees on Insects-Incremental-Balanced data

For the models used, Adaptive Random Forest is performing the best followed by SAMKNN. ARF has obtained 0.6448,0.7698,0.7475 accuracy for incremental, gradual and abrupt data respectively.The poor performance of the majority class classifier and no-change classifier shows that properties like temporal dependence and majority classes are underrepresented in data.

After using the drift detection method to the data stream, the accuracies of the model are decreasing significantly, indicating challenges with the online stream of data play a vital role in model behaviour. Abrupt changes in data lead to changes in the distribution of data resulting in concept drift.

Table 1, shows the accuracies obtained with the gradual, incremental and abrupt datasets using all aforementioned models.

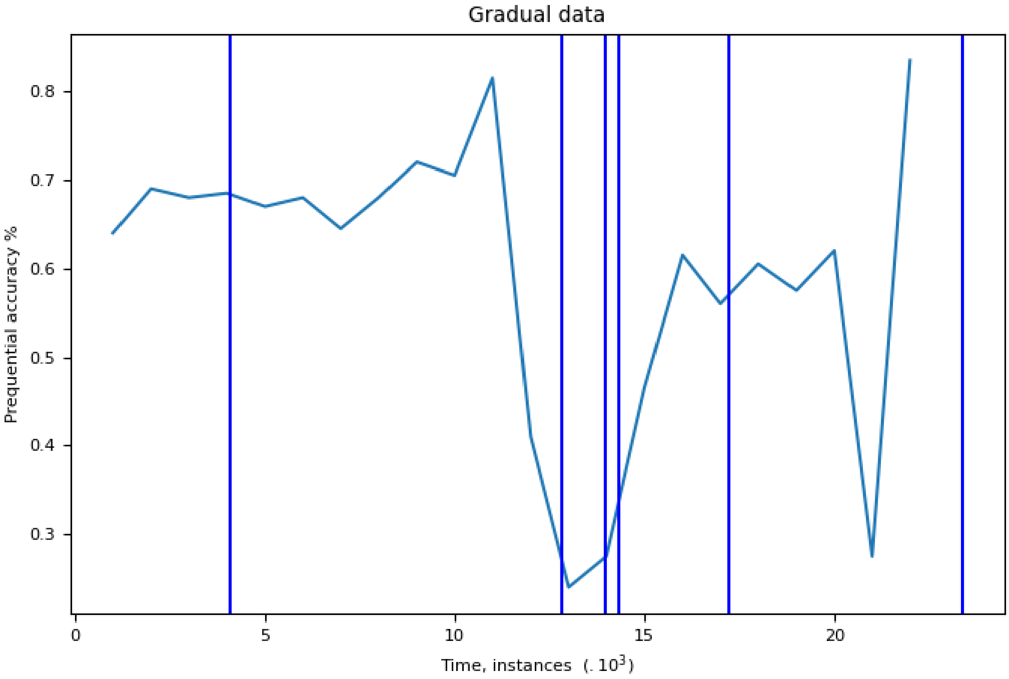


Figure 8: ADWIN Drift Detection with Hoeffding Trees on Insects-Gradual-Balanced data

When we compare the results obtained with the paper, we can see that the authors have used Imbalanced and balanced datasets to compare the model performance with other online stream algorithms. In the paper, authors have used incremental Naive Bayes (NB) Very Fast Decision Trees with Naive Bayes classifiers at the leaves, Leveraging Bagging with 10 VFDT in the ensemble and Adaptive Random Forest algorithms. For drift detection they are using Page-Hinkley, CUSUM, Drift Detection Method, Adaptive Windowing (ADWIN) SEED and Statistical Test of Equal Proportions (STEPD) algorithms.

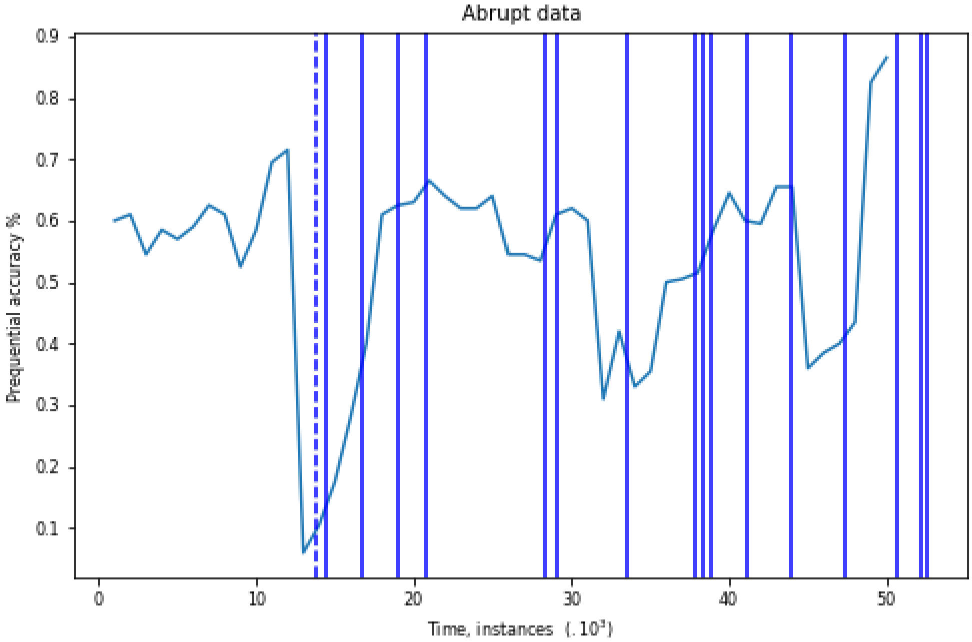


Figure 9:ADWIN Drift Detection with Hoeffding Trees on Insects-Abrupt-Balanced data

For the paper, the Adaptive Random Forest algorithm is the best algorithm followed by Leveraging Bagging where accuracy is around 70-80% for different drift detection patterns. Incremental Algorithms like VFDT and naive Bayes were outperformed by other algorithms which are explicitly designed to handle drifts.