Supplementary Material of OSNN: An Online Semisupervised Neural Network for Nonstationary Data Streams

Rodrigo G. F. Soares, Leandro L. Minku

I. Hyperparameter sensitivity

To analyze the impact of the choice of H and N on OSNN's performance, we plot the prequential accuracy of OSNN on both real and artificial streams, namely, Sine2, Elec, NOAA and Power Supply with nonuniform labeling distribution for different values of these hyperparameters.

In Figure 1, we show the sensitiveness of OSNN to H and N for artificial data: the Sine2 stream with 20% labeled examples and abrupt drift. H controls the complexity of OSNN. It is expected that smaller H produce simpler networks and simpler manifold representations that might not have the sufficient variance to learn the correct decision boundary, however smaller networks might adapt quickly to new concepts.

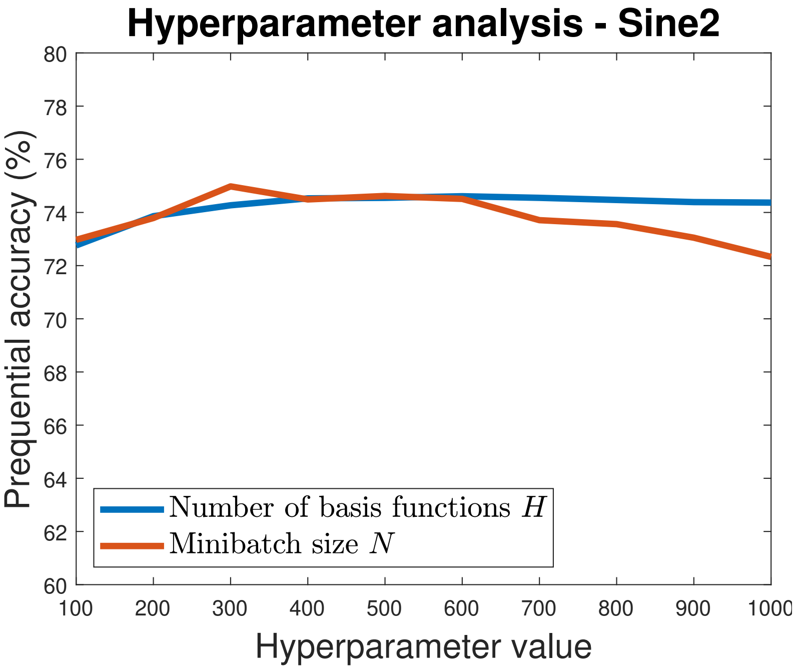

Fig. 1: Prequential accuracy for H and N on Sine2 with 20% labeled examples and abrupt drift.

In contrast, OSNN with larger hidden layers can learn more complex decision boundaries, however it tends to overfit the data and have slower adaptation to concept drifts. Such a trade-off is shown in Figure 1, where the best-performing value for H is 600 and OSNN's performance is slightly degraded with smaller or greater values. Figure 1 indicates that OSNN is fairly robust to the choice of H, though a finer tuning of H can help further improving generalization.

With smaller N, OSNN may be able to more quickly adapt to new concepts. However, with larger N, OSNN may be able to assess more information to induce the manifold and learn the scarce labels at the cost of computational time and speed of adaptation to concept drifts. Such a trade-off is shown in Figure 1, where the tuned value of N is 300. Smaller or greater values indicate worse balances and tend to degrade OSNN's performance. More specifically, since this Sine2 data stream has

Rodrigo G. F. Soares is with the Department of Statistics and Informatics of the Federal Rural University of Pernambuco, Recife, Brazil and with the School of Computer Science, The University of Birmingham, Birmingham B15 2TT, UK. rodrigo.gfsoares@ufrpe.br.

Leandro L. Minku is with the School of Computer Science, The University of Birmingham, Birmingham B15 2TT, UK. L.L.Minku@cs.bham.ac.uk.

abrupt concept drifts (drift length is 1 time step), larger values of N tend to make OSNN learn with data from an outdated concept. That is, the learning process is not able to switch from one concept to another rapidly enough when the previous concept should be forgotten instantly so that the model can learn an entirely different concept. This fact can be verified by the steep descending slope for $N > 300$. Nevertheless, Figure 1 also shows that variation in accuracy for different choices of N is relatively small. This fact demonstrates the robustness of OSNN to the choice of N.

In the context of real-world data streams, Figure 2 depicts the prequential accuracy as a function of H and N for the Elec data stream. In this case, H is recommended at 700. Smaller values lead to smaller neural networks that are not able to properly learn from data in Elec, whereas $H > 700$ tend to produce overfit networks. The hyperparameter N also has an impact on performance. It is clear that the best N is 500. With smaller N, there is a local maximum at 200. OSNN's performance start degrading with $N > 500$. Such a fact indicates that the amount of recent data that OSNN requires depends on the length of the concepts in each stream. Although H and N can be optimized via hyperparameter tuning, Figure 2 shows that the impact in prequential accuracy of different choices of H and N is relatively small. This fact demonstrates the robustness of OSNN to the choice of H and N.

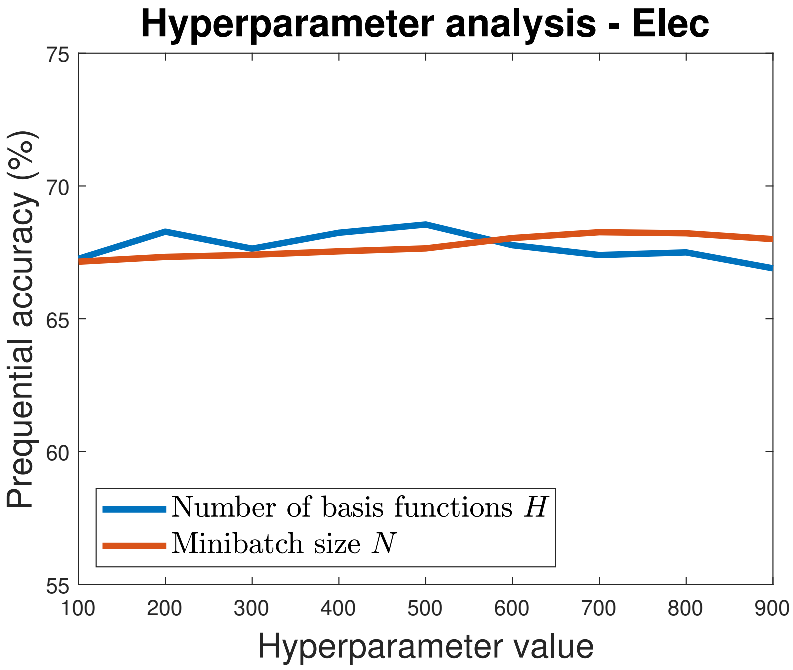

Fig. 2: Prequential accuracy for N and H.

The sensitivity of OSNN to N and H for the NOAA data stream is shown in Figure 3. Variations in accuracy for different choices of H and N are relatively small, except when H is smaller than 100. This fact demonstrates the robustness of OSNN to the choice of H and N. Considering variations with smaller magnitudes, we can notice that hyperparameter tuning can still be beneficial.

In this stream, OSNN tends to need larger neural networks, i.e. larger values for the hyperparameter H, which might denote that the NOAA data stream has overlapping classes with noisy data. The size of the neural network tends to stabilize with $H > 300$. This fact might indicate that new data is not produced in a single unknown space, instead it might come from several spaces that circumvent the current manifold. Therefore, larger H might be necessary in this case to represent the multiple sources of the new concept. The plateau for N indicates that concepts are noisy and overlap severely. In this case, it is unclear when the model should learn a new concept and forget the previous one. For the NOAA stream, a smaller N is advised for time performance.

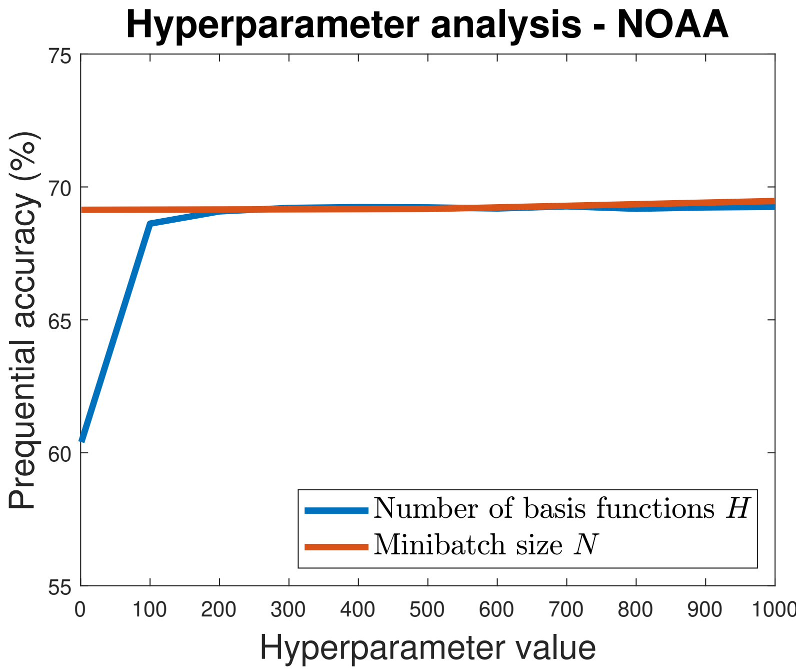

Fig. 3: Prequential accuracy for N and H.

In Figure 4, the prequential accuracy produced by different values of H and N are analyzed with the Power Supply data stream. As in the previous streams, H should be tuned in order to find a good trade-off for the network complexity. In this case, H is best at 300, with smaller or larger values producing inferior predictive performance. The minibatch size N should be set to 100, since OSNN with smaller N are not able to learn the current concept due to the lack of data; and OSNN with $N > 100$ may not migrate rapidly enough from on concept to another. Such a fact indicates that the amount of recent data that OSNN requires depends on the length of the concepts in each stream, the severity of the drift and of the complexity of the decision boundary of each concept. Although H and N can be optimized via hyperparameter tuning, Figure 2 also shows that the impact of different choices of H and N in generalization performance is relatively small. This fact demonstrates the robustness of OSNN to the choice of H and N.

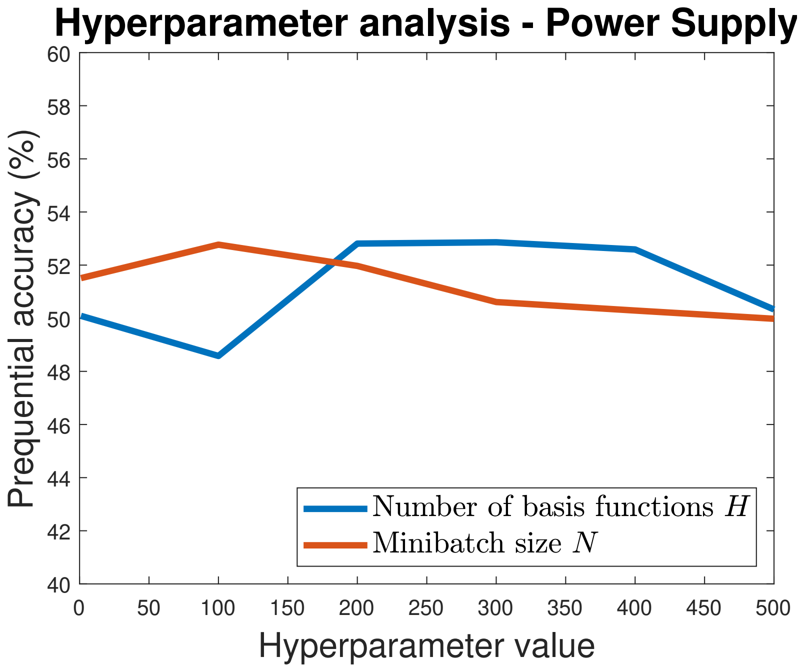

Fig. 4: Prequential accuracy for N and H.

Figures 1, 2, 3 and 4 highlight the robustness of OSNN to the choice of H and N, and also show that properly tuning such hyperparameters can bring some further small improvements to generalization. The recommended values for H depend on the data distribution and noise of each concept; the spatial differences between adjacent concepts; and the severity of the concept drift. It is important to point out that most learning methods in literature have hyperparameters that regulate the trade-off for model complexity and that their tuning is important for improving generalization performance. An adaptive H is a potential alternative to improve OSNN's predictive performance, as each concept might have severely contrasting data distributions with different amounts of noise coming in varying speeds. On the other hand, an approach for adaptive H might introduce new hyperparameters to the method.

As we can see from the above, the hyperparameters H and N have an effect on the behavior of OSNN and its ability to learn different concepts or tackle concept drifts and should ideally be tuned. However, we can also see that most of the time this effect is not large, meaning that the proposed approach is quite robust to hyperparameter choice. In particular, poor choices of hyperparameter values rarely caused large decay in predictive performance. Moreover, our experiments done to answer RQ3 have shown that the strategy of using an initial portion (the initial 10% of the stream) of the data stream for tuning is successful in leading to hyperparameter choices that enabled our proposed approach to achieve top performances compared to other existing approaches.

It is worth noting that the hyperparameter N controls the size of the data chunks. This kind of hyperparameter limits the ability of existing sliding window or chunk-based approaches from the literature to perform well on data streams with sudden drifts. This happens because these existing approaches usually reset their models when new chunks arrive, or create new models from scratch for each new chunk. This means that large chunk sizes are necessary to achieve good performance during stable periods. However, large chunk sizes prevent adaptation to sudden drifts. OSNN overcomes this issue based on the following two strategies:

1) OSNN does not reset its model and does not create new models from scratch to learn new chunks. Therefore, its chunks do not need to be very large for achieving good predictive performance during stable periods. This enables OSNN to deal with both abrupt and gradual drifts while maintaining its ability to perform well during stable periods. The fact that the chunks do not need to be very large is illustrated in our analysis of sensitivity to hyperparameters shown above.
2) The learning rate is automatically adjusted based on the chunk of data, so that an appropriate level of forgetting of the data within the chunk is automatically chosen to tackle different kinds of concept drift or stable concepts. Such adjustment is analyzed in Section VII.C of the main manuscript. The adaptive learning rate also helps our approach to be more robust to different choices of N, as it can increase or decrease the size of the learning steps according to changes in the incoming data distribution regardless the size of the minibatch. Such robustness to different values of N can be observed in the hyperparameter sensitivity analysis shown above.

II. Prequential accuracy figures

In this section, we show all Figures with plots of the prequential accuracy of the compared methods. Figures 5, 6 and 7 present Prequential accuracy on the Agrawal data stream with uniform labeling distribution and abrupt concept drifts with 5%, 10% and 20% of labels, respectively. In Figures 8, 9 and 10, we show the plots for gradual concept drifts in the Agrawal data stream with uniform labeling distribution and 5%, 10% and 20% of labels, respectively. We also show the plots for the real-world data stream (Power Supply) with uniform labeling distribution and 5%, 10% and 20% of labels, respectively, in Figures 11, 12 and 13.

We also show the Figures with plots of the prequential accuracy of the compared methods for nonuniform labeling distribution. Figures 14, 15 and 16 present Prequential accuracy on the Agrawal data stream with nonuniform labeling distribution and abrupt concept drifts with 5%, 10% and 20% of labels, respectively. In Figures 17, 18 and 19, we show the plots for gradual concept drifts in the Agrawal data stream with nonuniform labeling distribution and 5%, 10% and 20% of labels, respectively. We also show the plots of prequential accuracy for the real-world Power Supply stream with nonuniform labeling distribution and 5%, 10% and 20% of labels, respectively, in Figures 20, 21 and 22.

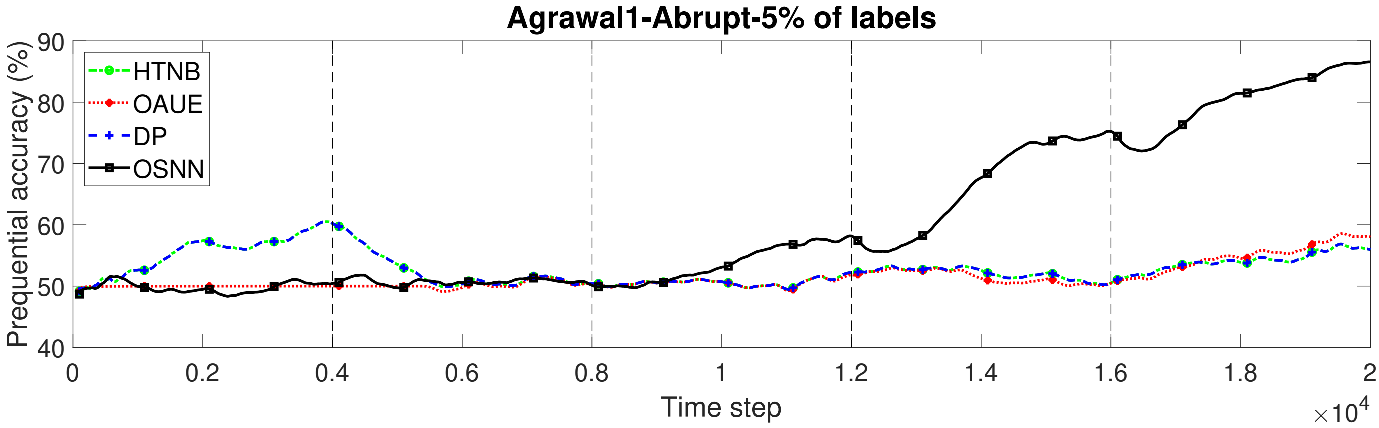

Fig. 5: Prequential accuracy for Agrawal1 with 5% of uniformly distributed labels and abrupt drifts.

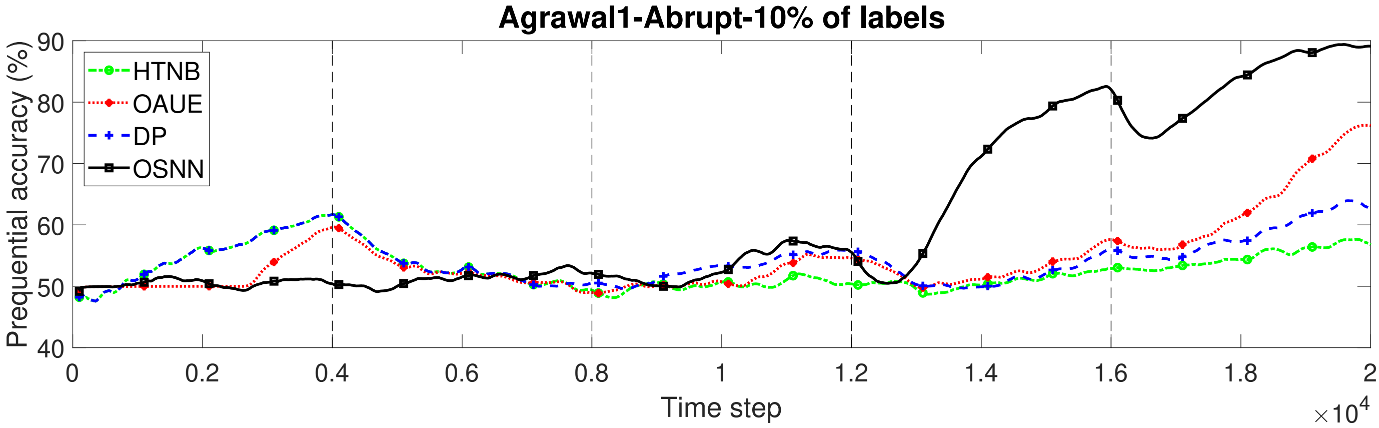

Fig. 6: Prequential accuracy for Agrawal1 with 10% of uniformly distributed labels and abrupt drifts.

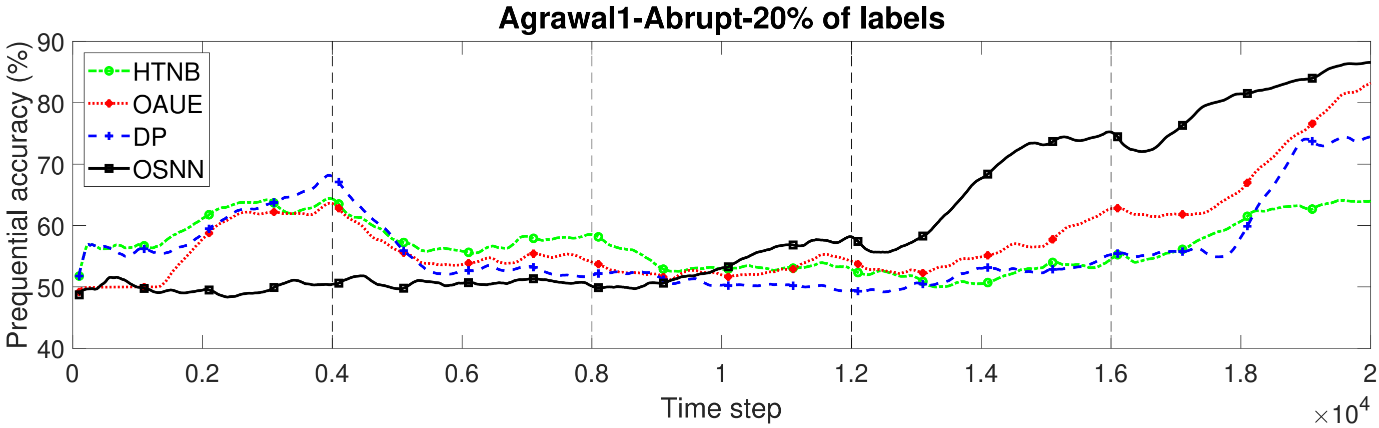

Fig. 7: Prequential accuracy for Agrawal1 with 20% of uniformly distributed labels and abrupt drifts.

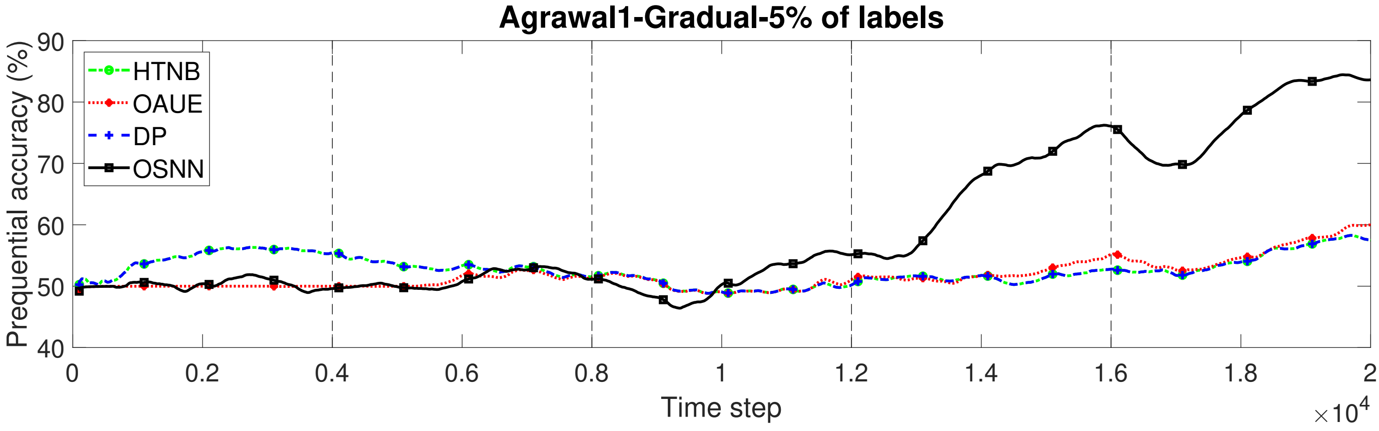

Fig. 8: Prequential accuracy for Agrawal1 with 5% of uniformly distributed labels and gradual drifts.

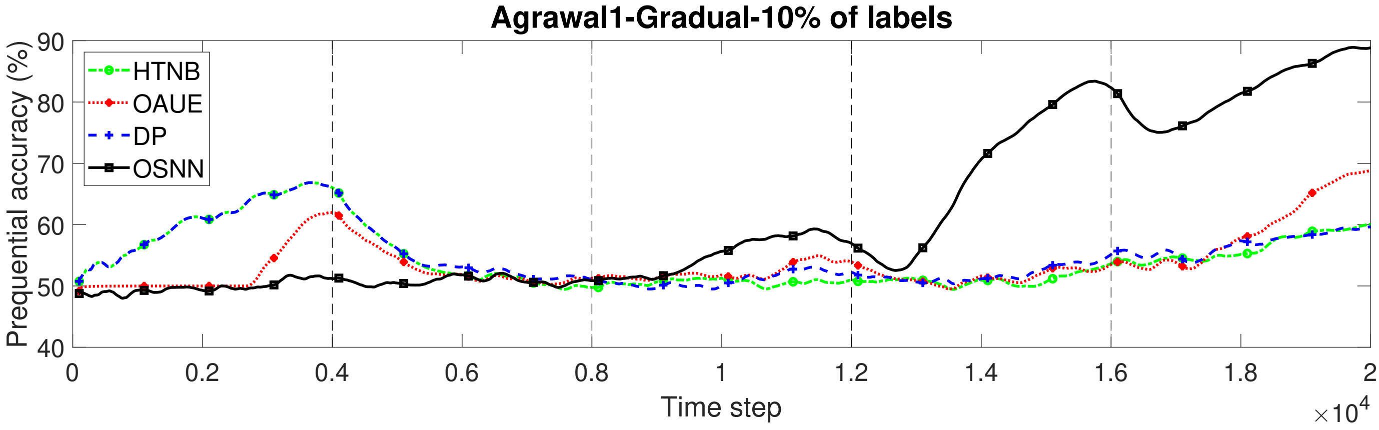

Fig. 9: Prequential accuracy for Agrawal1 with 10% of uniformly distributed labels and gradual drifts.

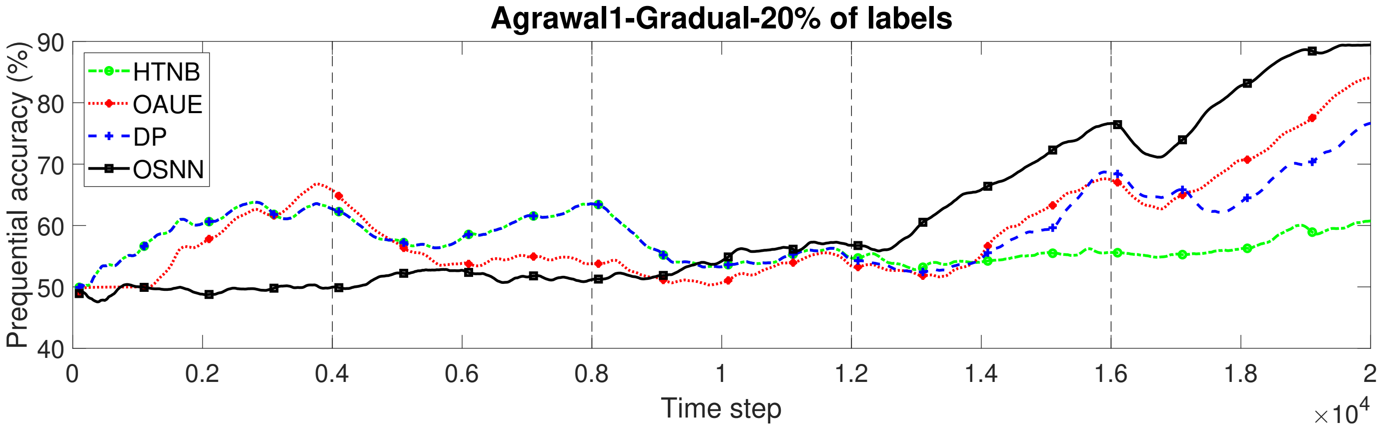

Fig. 10: Prequential accuracy for Agrawal1 with 20% of uniformly distributed labels and gradual drifts.

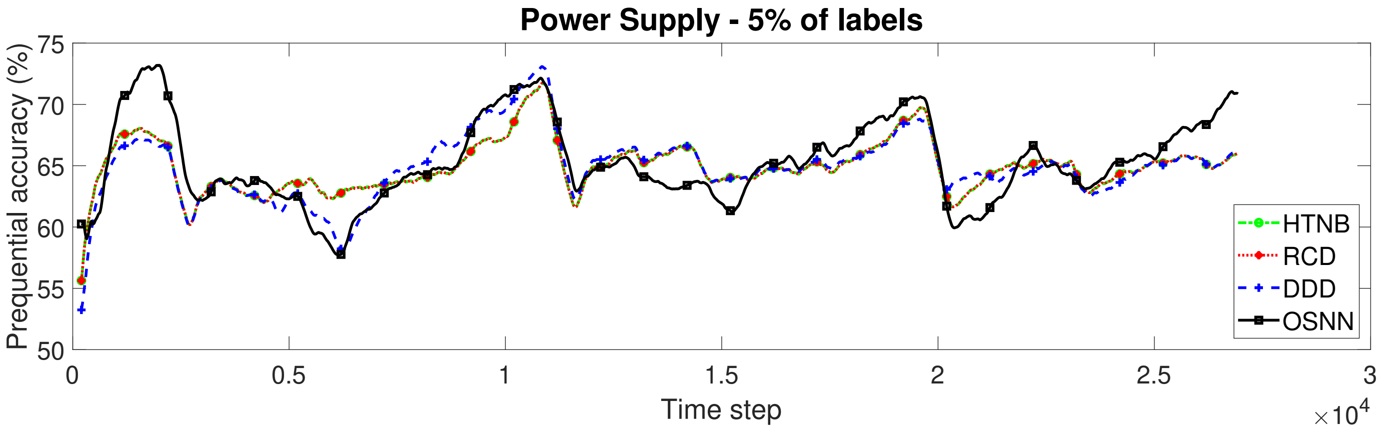

Fig. 11: Prequential accuracy for Power Supply with 5% of uniformly distributed labels.

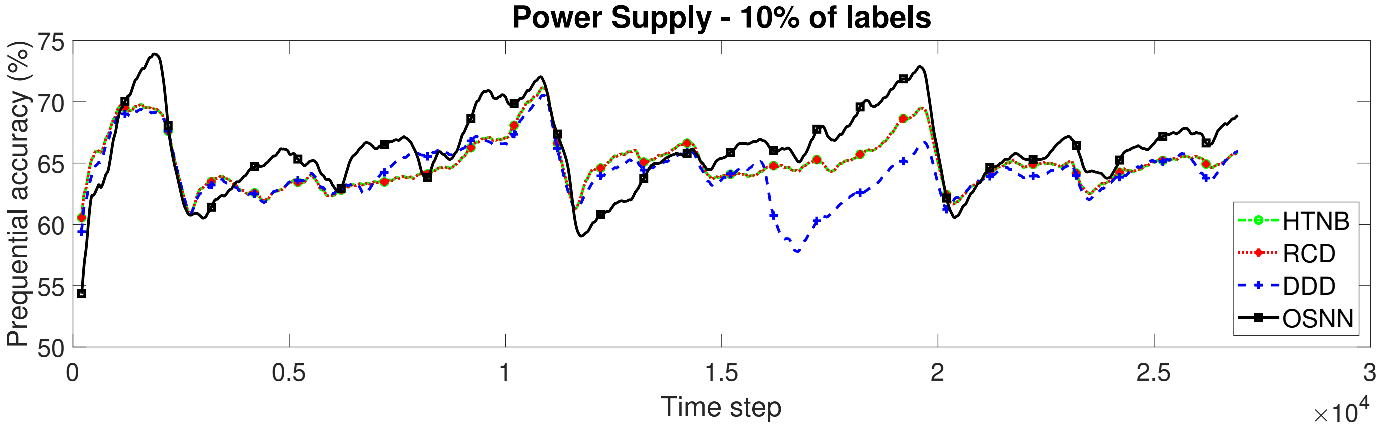

Fig. 12: Prequential accuracy for Power Supply with 10% of uniformly distributed labels.

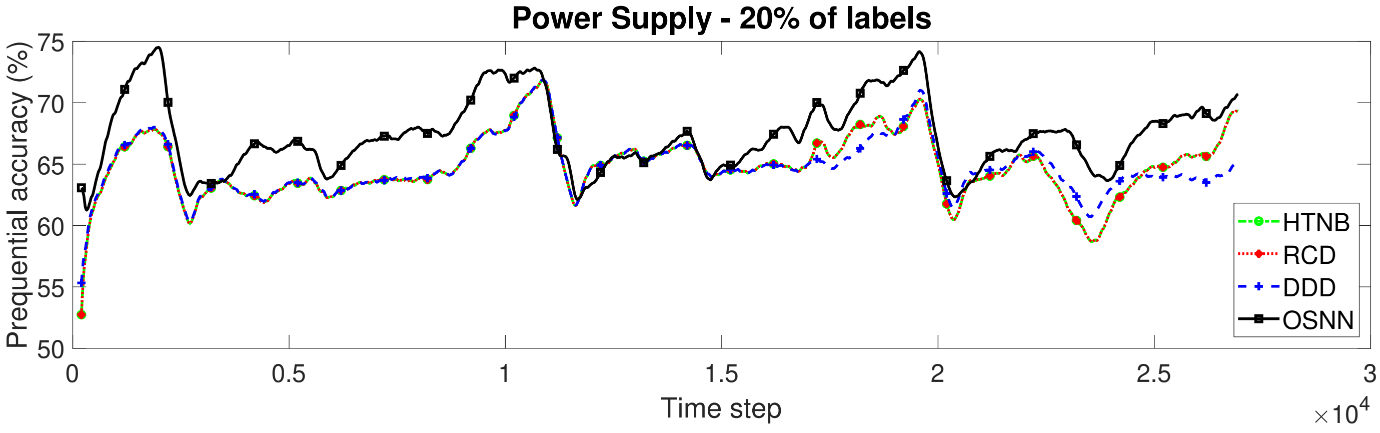

Fig. 13: Prequential accuracy for Power Supply with 20% of uniformly distributed labels.

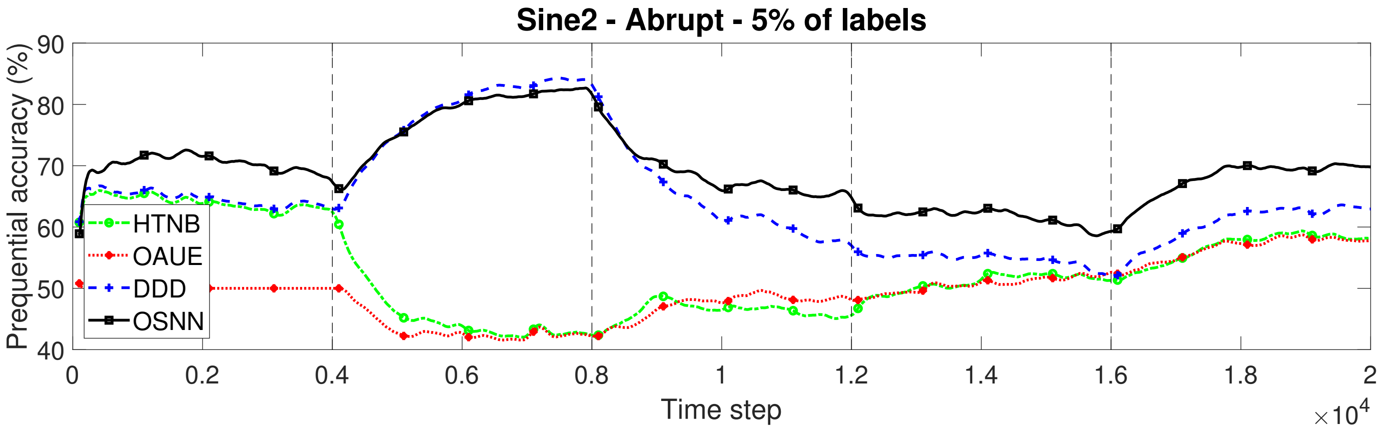

Fig. 14: Prequential accuracy for Sine2 with 5% of nonuniformly distributed labels and abrupt drifts.

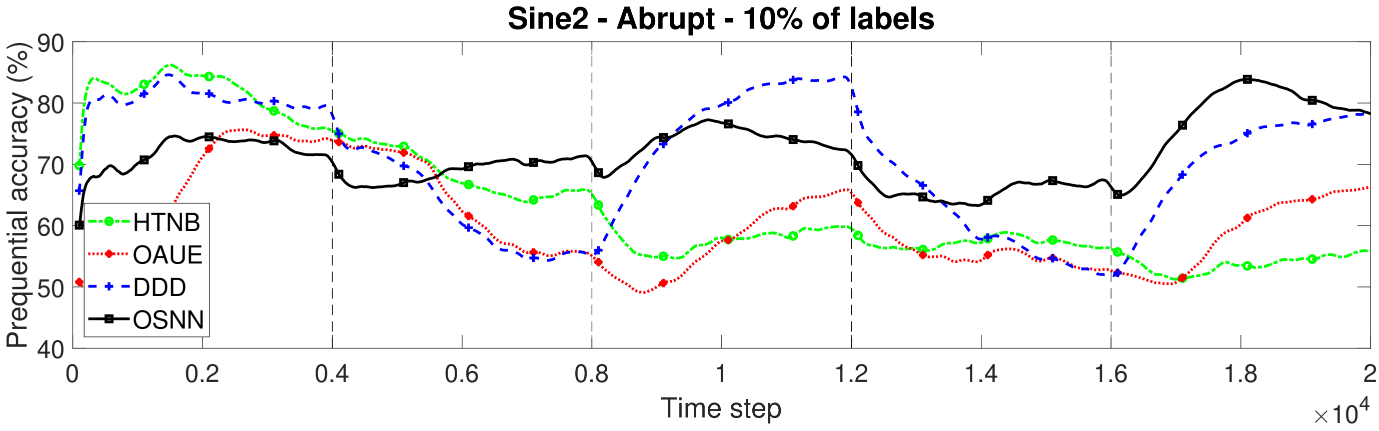

Fig. 15: Prequential accuracy for Sine2 with 10% of nonuniformly distributed labels and abrupt drifts.

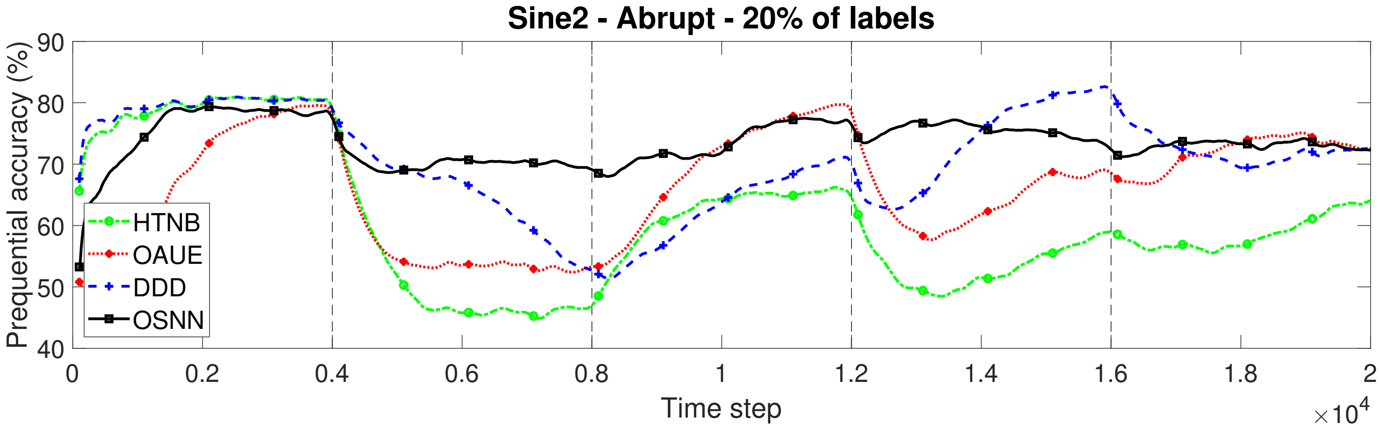

Fig. 16: Prequential accuracy for Sine2 with 20% of nonuniformly distributed labels and abrupt drifts.

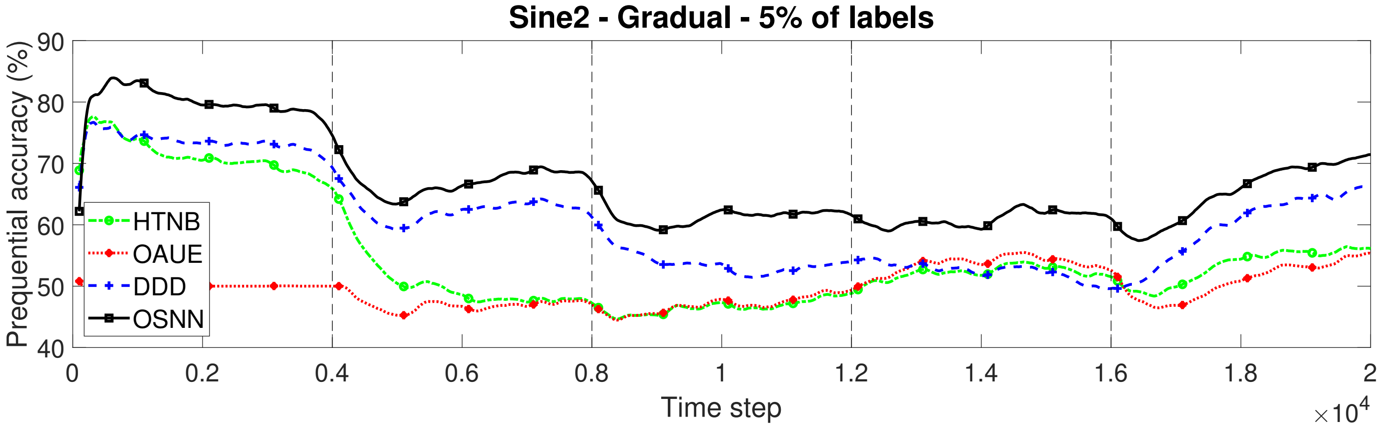

Fig. 17: Prequential accuracy for Sine2 with 5% of nonuniformly distributed labels and gradual drifts.

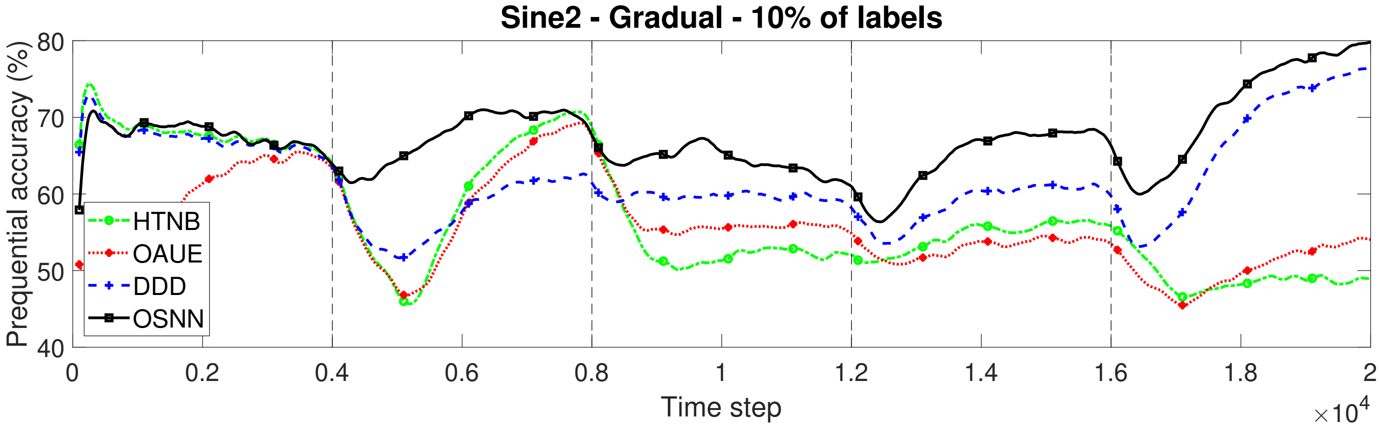

Fig. 18: Prequential accuracy for Sine2 with 10% of nonuniformly distributed labels and gradual drifts.

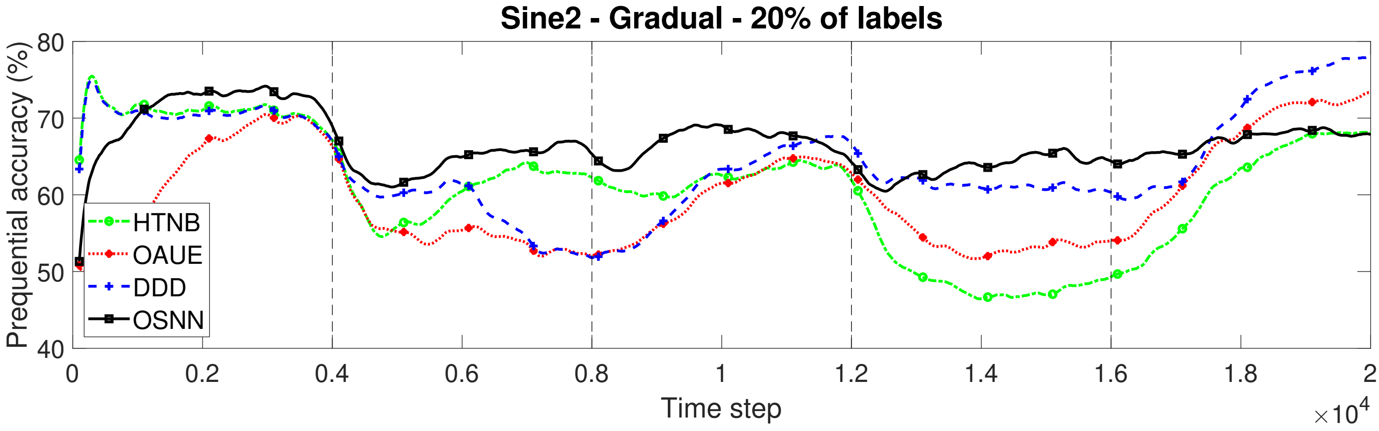

Fig. 19: Prequential accuracy for Sine2 with 20% of nonuniformly distributed labels and gradual drifts.

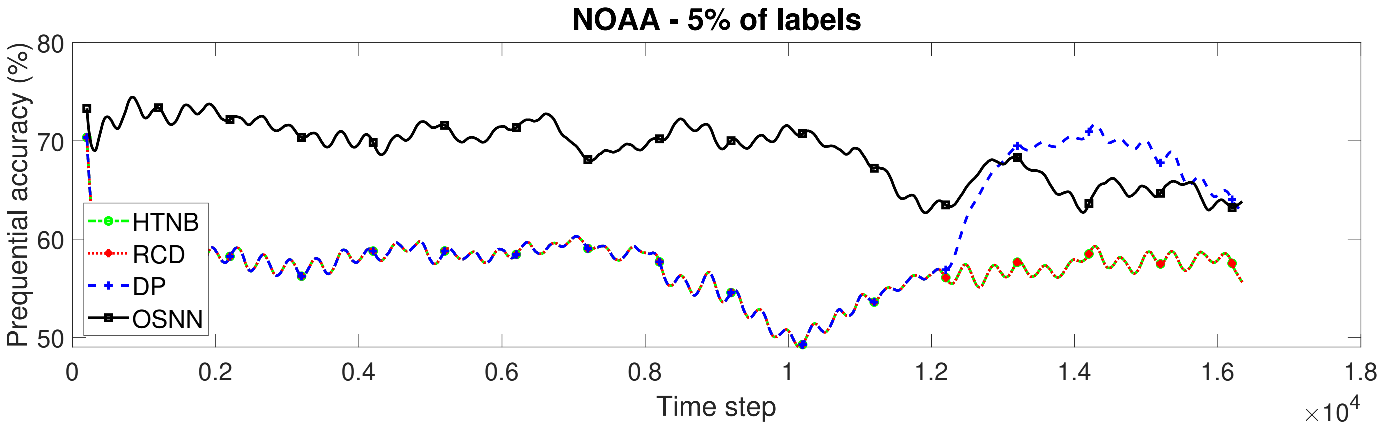

Fig. 20: Prequential accuracy for NOAA with 5% of nonuniformly distributed labels.

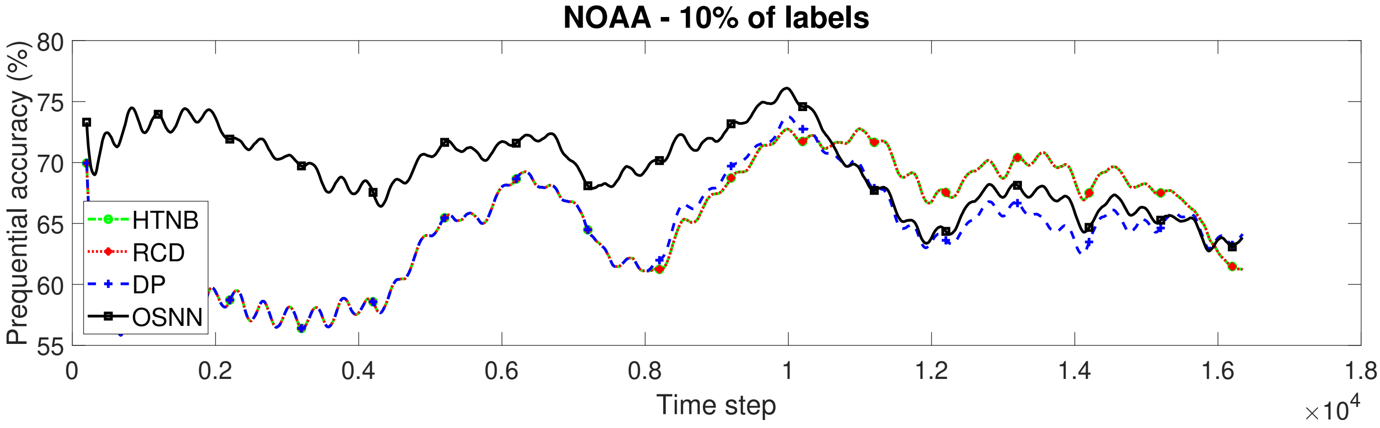

Fig. 21: Prequential accuracy for NOAA with 10% of nonuniformly distributed labels.

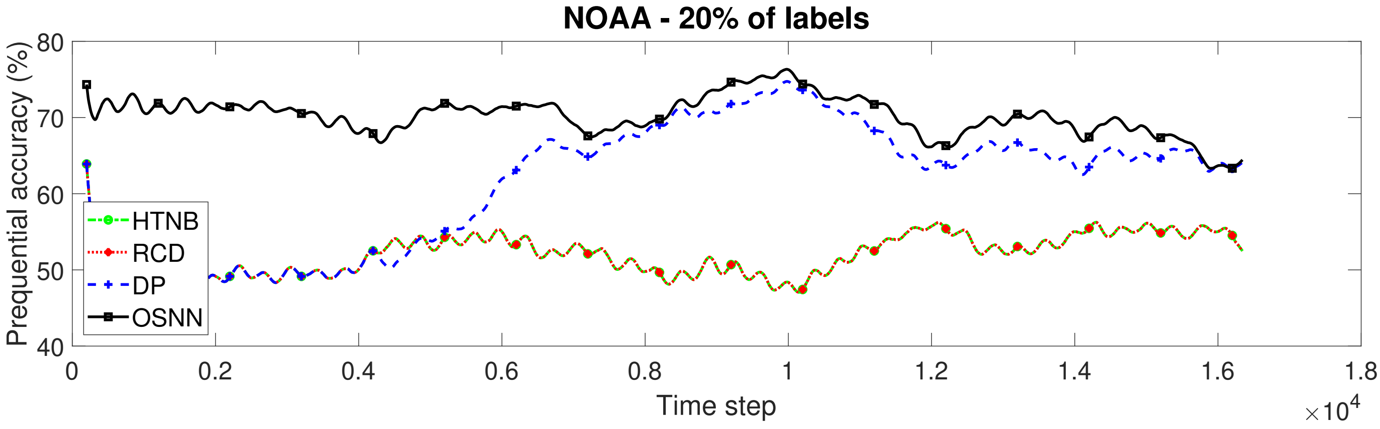

Fig. 22: Prequential accuracy for NOAA with 20% of nonuniformly distributed labels.

III. RQ2 – Result tables

In this section, we show the detailed result tables for answering Research Question (RQ) 2. Subsection III-A presents the results for uniformly distributed labels. And, in Subsection III-B, we show the results for nonuniformly distributed labels.

A. *Result tables for uniformly distributed labels*

Tables I, II and III present the mean and standard deviation of each method in each data stream for abrupt drifts, gradual drifts and real-world streams, respectively, with all amounts of labeled data in the scenario with uniformly distributed labels.

TABLE I: Mean and standard deviation of prequential accuracy of our supervised version of our approach and OSNN on artificial streams with abrupt concept drifts and uniformly distributed labels.

Type of drift: abrupt		
	OSNN (Supervised)	OSNN
Labels at 5%		
Sine1	76.144±6.271	72.946±7.794
Sine2	80.499±9.923	76.854±10.308
Agrawal1	61.289±13.76	60.068±12.358
Agrawal2	61.052±10.825	59.529±8.483
Agrawal3	52.457±2.076	53.618±2.261
Agrawal4	52.754±3.095	53.756±3.427
SEA1	81.104±5.317	81.545±6.804
SEA2	82.101±4.304	79.166±8.157
STAGGER1	93.602±4.76	81.308±8.405
STAGGER2	94.16±6.178	85.489±8.778
Labels at 10%		
Sine1	78.264±6.681	76.453±6.494
Sine2	82.933±10.049	82.893±7.783
Agrawal1	63.76±15.579	61.218±13.950
Agrawal2	64.902±11.493	62.045±8.989
Agrawal3	53.91±2.158	52.793±2.112
Agrawal4	55.716±5.083	54.631±4.836
SEA1	84.213±4.496	82.240±5.930
SEA2	84.145±4.055	82.731±5.266
STAGGER1	95.998±4.388	93.651±4.051
STAGGER2	96.099±6.309	93.358±8.822
Labels at 20%		
Sine1	77.926±9.337	78.505±4.553
Sine2	83.3±10.741	85.644±6.496
Agrawal1	66.95±16.222	61.728±13.926
Agrawal2	68.442±12.379	62.790±9.350
Agrawal3	55.994±3.506	53.484±2.766
Agrawal4	57.634±5.206	53.982±4.399
SEA1	85.618±3.466	84.121±4.109
SEA2	84.863±3.236	83.595±3.863
STAGGER1	96.99±3.699	97.979±2.477
STAGGER2	96.535±4.673	97.278±5.101

TABLE II: Mean and standard deviation of prequential accuracy of our supervised version of our approach and OSNN on artificial streams with gradual concept drifts and uniformly distributed labels.

Type of drift: gradual		
	OSNN (Supervised)	OSNN
Labels at 5%		
Sine1	75.03±7.223	71.878±8.297
Sine2	81.585±9.065	77.630±8.689
Agrawal1	60.169±13.175	59.083±11.734
Agrawal2	60.443±9.317	60.206±7.334
Agrawal3	50.409±1.721	51.848±1.899
Agrawal4	52.616±3.199	52.756±3.197
SEA1	81.695±4.728	80.859±5.696
SEA2	80.731±5.279	80.748±6.046
STAGGER1	93.378±4.459	82.310±9.939
STAGGER2	90.866±7.105	84.390±8.153
Labels at 10%		
Sine1	77.573±6.823	75.539±6.177
Sine2	82.987±9.519	80.289±7.675
Agrawal1	64.318±15.582	61.166±13.522
Agrawal2	65.282±10.918	62.188±8.938
Agrawal3	54.171±2.588	53.829±3.218
Agrawal4	54.648±5.411	53.520±4.654
SEA1	84.285±3.975	82.819±6.495
SEA2	83.262±4.906	79.859±8.554
STAGGER1	95.639±4.561	93.159±4.759
STAGGER2	94.928±5.594	91.277±7.953
Labels at 20%		
Sine1	78.309±8.061	77.582±5.876
Sine2	83.431±10.387	83.650±6.718
Agrawal1	67.24±15.747	60.542±12.687
Agrawal2	67.747±12.224	62.303±9.266
Agrawal3	57.516±3.126	54.412±3.124
Agrawal4	58.265±4.995	54.505±4.499
SEA1	85.081±4.012	83.091±5.091
SEA2	85.06±3.728	82.615±5.734
STAGGER1	96.484±4.291	96.822±3.619
STAGGER2	96.331±4.142	95.746±4.776

TABLE III: Mean and standard deviation of prequential accuracy of our supervised version of our approach and OSNN on artificial streams with real-world streams with uniformly distributed labels.

	OSNN (Supervised)	OSNN
Labels at 5%		
Elec	70.175±5.984	73.765±7.665
NOAA	73.764±3.139	71.264±3.207
Power Supply	63.739±3.194	65.279±3.588
Sensor	73.518±11.066	73.913±12.437
Labels at 10%		
Elec	74.889±4.925	74.187±7.826
NOAA	76.717±2.369	72.083±2.629
Power Supply	64.519±3.371	66.003±3.357
Sensor	79.011±11.642	76.872±12.196
Labels at 20%		
Elec	76.131±5.409	75.201±7.060
NOAA	78.525±2.052	73.123±2.379
Power Supply	64.573±3.225	67.336±3.085
Sensor	82.472±11.345	80.657±11.580

B. Result tables for nonuniformly distributed labels

Tables IV, V and VI present the mean and standard deviation of each method in each data stream for abrupt drifts, gradual drifts and real-world streams, respectively, with all amounts of labeled data in the scenario with nonuniformly distributed labels.

TABLE IV: Mean and standard deviation of prequential accuracy of our supervised version of our approach and OSNN on artificial streams with abrupt concept drifts and nonuniformly distributed labels.

Type of drift: abrupt		
	OSNN (Supervised)	OSNN
Labels at 5%		
Sine1	61.487±6.785	71.058±4.889
Sine2	59.772±6.543	69.194±6.275
Agrawal1	52.919±5.648	53.683±6.270
Agrawal2	52.211±4.859	53.917±6.127
Agrawal3	51.879±2.462	51.777±2.118
Agrawal4	52.294±2.459	51.409±1.901
SEA1	56.198±3.848	65.622±4.595
SEA2	58.111±6.178	67.310±6.592
STAGGER1	59.671±9.880	68.210±10.137
STAGGER2	56.217±7.132	58.756±2.676
Labels at 10%		
Sine1	59.823±5.573	59.840±6.310
Sine2	67.482±7.832	71.583±5.359
Agrawal1	52.409±3.358	55.559±7.257
Agrawal2	53.609±3.856	53.681±4.796
Agrawal3	50.598±1.305	51.058±2.008
Agrawal4	51.359±1.943	51.599±1.360
SEA1	56.785±5.956	63.591±4.972
SEA2	61.046±7.151	67.990±8.656
STAGGER1	62.862±12.878	69.589±8.013
STAGGER2	65.500±9.036	65.260±5.617
Labels at 10%		
Sine1	63.198±3.412	64.436±4.306
Sine2	68.737±8.069	73.422±3.974
Agrawal1	57.496±10.329	57.022±9.180
Agrawal2	54.151±5.069	56.533±7.528
Agrawal3	52.770±2.736	50.355±1.711
Agrawal4	51.450±1.861	52.821±3.441
SEA1	69.236±6.850	68.714±7.047
SEA2	68.623±10.099	73.625±9.589
STAGGER1	66.413±8.716	69.899±8.622
STAGGER2	69.208±13.065	65.532±5.092

TABLE V: Mean and standard deviation of prequential accuracy of our supervised version of our approach and OSNN on artificial streams with gradual concept drifts and nonuniformly distributed labels.

	Type of drift: gradual	
	OSNN (Supervised)	OSNN
	Labels at 5%	
Sine1	58.252±5.199	65.324±3.234
Sine2	57.608±8.374	66.803±7.472
Agrawal1	52.892±5.512	53.233±6.722
Agrawal2	52.687±5.330	53.461±4.067
Agrawal3	51.098±1.991	49.074±3.412
Agrawal4	51.479±1.867	52.165±3.086
SEA1	58.360±6.661	66.475±15.049
SEA2	59.049±6.731	67.990±4.933
STAGGER1	57.146±6.679	65.305±7.059
STAGGER2	55.921±5.459	59.712±4.051
	Labels at 10%	
Sine1	60.830±7.315	61.655±6.573
Sine2	62.581±6.177	67.090±4.905
Agrawal1	53.740±9.287	54.106±6.980
Agrawal2	55.012±8.020	52.362±4.453
Agrawal3	50.693±1.821	51.294±2.337
Agrawal4	51.145±1.560	50.686±1.258
SEA1	60.836±5.839	69.206±7.562
SEA2	58.490±5.591	66.025±5.567
STAGGER1	66.476±10.486	65.978±7.671
STAGGER2	56.253±4.852	57.406±4.190
	Labels at 20%	
Sine1	66.388±5.294	67.997±4.571
Sine2	66.955±4.584	66.480±3.747
Agrawal1	55.988±7.153	56.099±7.592
Agrawal2	53.551±2.746	54.561±3.011
Agrawal3	49.430±1.753	50.070±2.381
Agrawal4	50.423±1.277	51.386±2.492
SEA1	68.191±5.794	70.036±5.484
SEA2	63.793±6.464	65.742±6.721
STAGGER1	65.718±9.652	66.495±4.545
STAGGER2	73.234±8.928	64.202±3.564

TABLE VI: Mean and standard deviation of prequential accuracy of our supervised version of our approach and OSNN on artificial streams with real-world streams with nonuniformly distributed labels.

	OSNN (Supervised)	OSNN
	Labels at 5%	
Elec	55.709±6.941	55.124±13.538
NOAA	44.305±11.053	69.205±3.584
Power Supply	52.469±4.340	58.369±8.474
Sensor	51.211±9.323	60.137±10.583
	Labels at 10%	
Elec	52.695±8.819	51.199±9.622
NOAA	50.893±11.936	69.652±3.697
Power Supply	56.400±6.248	51.880±5.992
Sensor	52.343±9.839	59.993±15.509
	Labels at 20%	
Elec	60.245±11.148	67.668±6.542
NOAA	46.712±10.336	70.429±3.126
Power Supply	58.758±5.659	50.102±7.445
Sensor	54.296±9.758	59.506±11.393

IV. RQ3 – Result tables

In this section, we show the detailed result tables for answering RQ3. Subsection IV-A presents the results for uniformly distributed labels. And, in Subsection IV-B, we show the results for nonuniformly distributed labels.

A. *Result tables for uniformly distributed labels*

Tables VII, VIII and IX present the mean and standard deviation of each method in each data stream for abrupt drifts, gradual drifts and real-world streams, respectively, with all amounts of labeled data in the scenario with uniformly distributed labels.

TABLE VII: Mean and standard deviation of prequential accuracy on artificial streams with abrupt concept drifts and uniformly distributed labels.

Type of drift: abrupt							
	HTNB	OZABAG	OAUE	RCD	DDD	DP	OSNN
Labels at 5%							
Sine1	63.630±13.842	70.100±13.417	55.273±10.852	77.032±7.699	75.757±8.154	77.032±7.699	72.95±7.79
Sine2	63.785±19.531	61.237±18.182	64.597±12.971	80.500±8.246	79.924±8.768	80.500±8.246	76.85±10.31
Agrawal1	52.866±2.692	52.832±2.701	51.363±2.156	52.866±2.692	52.746±2.628	52.866±2.692	60.07±12.36
Agrawal2	52.643±2.357	51.547±2.205	51.343±1.292	52.815±2.306	51.458±2.074	52.643±2.357	59.53±8.48
Agrawal3	52.563±2.153	52.453±2.136	51.090±1.308	52.018±1.264	52.299±2.022	52.563±2.153	53.62±2.26
Agrawal4	52.171±2.779	51.586±2.368	50.712±1.295	52.109±2.809	52.089±2.068	52.171±2.779	53.76±3.43
SEA1	82.076±5.802	81.577±6.083	75.405±14.771	82.076±5.802	81.318±6.359	82.076±5.802	81.55±6.80
SEA2	82.476±4.580	81.875±5.255	74.714±14.053	82.476±4.580	81.852±5.206	82.476±4.580	79.17±8.16
STAGGER1	79.419±16.424	82.225±11.896	67.707±19.763	78.732±16.720	95.835±5.860	96.646±4.901	81.31±8.41
STAGGER2	70.623±24.328	70.932±24.128	62.539±13.129	78.238±17.157	94.918±6.598	95.652±6.013	85.49±8.78
Labels at 10%							
Sine1	61.166±20.151	60.612±19.413	59.112±14.360	80.801±4.385	79.972±5.329	80.801±4.385	76.45±6.49
Sine2	58.080±29.440	62.062±19.744	62.089±16.383	84.641±7.463	84.920±7.376	86.567±6.543	82.89±7.78
Agrawal1	52.789±3.285	52.987±2.935	54.413±5.947	53.920±3.949	53.017±3.133	54.345±3.805	61.22±13.95
Agrawal2	54.249±5.845	54.477±7.304	53.414±3.862	54.536±5.863	54.764±6.067	54.770±6.012	62.04±8.99
Agrawal3	53.630±3.260	53.646±2.740	52.351±2.617	53.630±3.260	53.364±2.631	53.630±3.260	52.79±2.11
Agrawal4	52.656±4.243	52.952±3.768	52.093±2.910	53.128±3.472	51.984±2.413	53.462±3.847	54.63±4.84
SEA1	83.885±4.221	83.775±4.265	80.239±10.876	83.885±4.221	83.831±4.295	83.885±4.221	82.24±5.93
SEA2	83.766±4.049	83.702±4.227	79.368±11.086	83.766±4.049	83.648±4.343	83.766±4.049	82.73±5.27
STAGGER1	73.070±20.078	77.538±18.015	74.731±16.443	76.623±19.861	97.182±4.511	98.228±3.481	93.65±4.05
STAGGER2	68.822±22.760	71.355±21.107	77.580±15.786	79.044±17.487	96.770±6.333	96.822±6.255	93.36±8.82
Labels at 20%							
Sine1	61.020±24.098	61.118±23.407	69.458±13.555	83.219±4.339	82.285±4.719	83.219±4.339	78.51±4.55
Sine2	59.380±18.717	59.445±18.411	71.861±14.471	88.005±4.285	86.935±5.129	88.005±4.285	85.64±6.50
Agrawal1	56.743±4.156	57.289±7.451	57.946±7.321	56.481±4.005	54.639±3.952	56.101±6.591	61.73±13.93
Agrawal2	55.872±6.890	56.850±8.124	57.119±6.412	55.672±4.915	57.613±7.742	57.776±6.738	62.79±9.35
Agrawal3	53.761±2.575	54.649±2.998	54.204±2.471	53.761±2.575	54.974±2.823	53.761±2.575	53.48±2.77
Agrawal4	56.467±3.748	55.206±3.305	54.691±2.748	53.387±3.531	53.473±1.781	56.870±3.313	53.98±4.40
SEA1	83.813±3.793	84.200±3.514	82.182±8.822	83.813±3.793	84.067±3.745	83.813±3.793	84.12±4.11
SEA2	83.713±3.564	83.466±3.643	81.357±8.840	83.713±3.564	83.689±4.080	83.713±3.564	83.59±3.86
STAGGER1	72.452±22.691	75.551±19.867	84.330±13.919	98.733±2.481	98.531±2.678	98.733±2.481	97.98±2.48
STAGGER2	69.202±22.596	79.427±13.388	83.033±13.697	88.872±12.685	98.302±4.372	98.332±3.977	97.28±5.10

TABLE VIII: Mean and standard deviation of prequential accuracy on artificial streams with gradual concept drifts and uniformly distributed labels.

Type of drift: gradual							
	HTNB	OZABAG	OAUE	RCD	DDD	DP	OSNN
Labels at 5%							
Sine1	70.598±8.726	66.568±8.782	63.498±11.762	75.684±5.630	74.909±7.375	75.684±5.630	71.88±8.30
Sine2	56.210±24.421	56.257±21.740	55.177±20.404	81.912±7.235	81.745±7.809	81.912±7.235	77.63±8.69
Agrawal1	52.706±2.372	52.382±1.964	51.710±2.484	52.662±1.929	51.906±2.004	52.706±2.372	59.08±11.73
Agrawal2	52.167±3.702	51.893±3.352	50.455±0.948	52.193±3.143	52.631±2.985	53.620±3.128	60.21±7.33
Agrawal3	52.394±1.591	52.372±1.548	51.385±1.207	51.905±1.297	52.159±1.530	52.394±1.591	51.85±1.90
Agrawal4	52.748±3.370	52.097±2.872	50.859±1.270	51.887±2.044	52.052±2.891	52.748±3.370	52.76±3.20
SEA1	82.074±6.236	82.184±6.134	74.719±14.968	82.074±6.236	82.564±5.887	82.074±6.236	80.86±5.70
SEA2	81.522±6.177	80.992±7.243	74.840±13.803	81.522±6.177	80.850±7.342	81.522±6.177	80.75±6.05
STAGGER1	72.023±19.672	75.028±17.338	61.000±14.604	76.503±18.071	93.890±6.319	94.745±5.641	82.31±9.94
STAGGER2	66.969±21.481	67.175±19.827	62.814±14.135	71.963±17.892	92.650±7.953	91.987±7.694	84.39±8.15
Labels at 10%							
Sine1	61.403±21.888	61.162±21.506	62.374±12.054	80.561±4.594	80.458±4.213	80.561±4.594	75.54±6.18
Sine2	64.800±17.618	63.420±18.191	67.948±12.278	82.531±7.516	82.091±7.486	82.531±7.516	80.29±7.68
Agrawal1	54.372±4.779	53.906±4.008	53.644±4.419	52.886±2.564	54.046±3.887	54.892±4.541	61.17±13.52
Agrawal2	53.147±5.338	53.930±5.202	52.790±3.307	54.430±4.728	54.605±5.141	54.546±4.843	62.19±8.94
Agrawal3	53.187±2.577	53.328±2.693	52.087±2.246	52.602±1.985	51.550±1.664	53.187±2.577	53.83±3.22
Agrawal4	52.782±4.256	52.370±3.754	52.047±2.478	51.725±3.148	52.748±2.353	53.185±4.077	53.52±4.65
SEA1	83.217±4.177	82.338±4.225	79.677±11.519	83.217±4.177	82.959±4.918	83.217±4.177	82.82±6.49
SEA2	82.440±4.792	82.450±5.374	78.858±11.207	82.440±4.792	82.281±5.606	82.440±4.792	79.86±8.55
STAGGER1	70.148±19.858	70.675±20.279	74.882±16.115	91.362±5.645	96.521±4.492	96.083±4.328	93.16±4.76
STAGGER2	67.518±21.689	70.680±19.549	73.762±13.175	88.627±8.978	94.397±6.283	94.072±6.429	91.28±7.95
Labels at 20%							
Sine1	61.870±20.025	61.563±21.418	69.473±13.012	82.637±4.415	81.922±5.017	82.637±4.415	77.58±5.88
Sine2	58.330±22.172	61.623±19.001	69.828±14.049	85.870±5.800	85.027±6.088	85.870±5.800	83.65±6.72
Agrawal1	57.260±3.278	56.931±6.274	59.054±8.201	54.418±4.030	54.260±3.747	59.848±5.563	60.54±12.69
Agrawal2	55.609±8.255	56.223±9.193	56.457±7.280	56.837±7.975	55.671±7.537	57.374±8.213	62.30±9.27
Agrawal3	53.825±2.271	54.389±2.547	53.514±2.051	53.942±1.982	53.469±2.408	53.391±2.526	54.41±3.12
Agrawal4	54.959±4.474	55.737±4.831	54.798±4.362	52.960±4.269	52.836±3.269	54.313±4.691	54.50±4.50
SEA1	83.361±3.902	83.571±3.904	81.318±9.012	83.361±3.902	83.422±3.898	83.361±3.902	83.09±5.09
SEA2	83.181±2.964	82.924±3.580	81.102±8.946	83.181±2.964	82.994±3.740	83.181±2.964	82.62±5.73
STAGGER1	69.985±22.615	77.875±17.252	84.372±13.435	96.585±4.765	97.084±3.905	96.585±4.765	96.82±3.62
STAGGER2	68.808±23.044	72.464±20.215	81.915±13.132	93.299±5.722	95.633±5.023	95.115±4.940	95.75±4.78

TABLE IX: Mean and standard deviation of prequential accuracy on real-world streams with uniformly distributed labels.

	HTNB	OZABAG	OAUE	RCD	DDD	DP	OSNN
Labels at 5%							
Elec	73.368±6.649	73.514±6.843	67.275±15.393	73.696±5.846	73.426±6.800	73.498±6.838	73.77±7.67
NOAA	65.562±2.794	67.955±2.670	67.723±3.375	65.562±2.794	69.194±3.303	65.562±2.794	71.26±3.21
Power Supply	64.877±2.324	64.793±2.301	62.393±5.775	64.877±2.324	64.844±2.863	64.877±2.324	65.28±3.59
Sensor	67.667±15.870	67.597±15.118	69.610±15.915	62.202±13.172	75.369±13.769	73.512±14.509	73.91±12.44
Labels at 10%							
Elec	71.323±6.180	73.573±7.802	69.878±13.386	73.538±6.724	74.454±6.857	74.354±5.828	74.19±7.83
NOAA	69.324±3.132	71.639±2.843	71.599±2.717	69.308±3.148	71.020±3.447	69.309±3.147	72.08±2.63
Power Supply	64.982±2.251	64.981±2.245	63.078±4.760	64.982±2.251	64.279±2.444	64.885±2.353	66.00±3.36
Sensor	71.908±14.075	76.026±13.265	81.003±13.963	68.755±14.536	80.012±14.141	81.591±12.671	76.87±12.20
Labels at 20%							
Elec	75.179±5.292	76.487±5.292	72.423±10.476	73.862±6.881	76.178±5.323	76.064±5.884	75.20±7.06
NOAA	65.154±2.825	65.368±2.807	67.226±3.029	66.110±2.861	69.968±3.530	66.005±2.722	73.12±2.38
Power Supply	64.714±3.005	64.887±3.042	64.366±3.648	64.714±3.005	64.720±2.597	64.714±3.005	67.34±3.09
Sensor	71.876±16.101	86.462±10.322	87.869±11.387	79.301±10.068	90.831±7.045	89.246±6.810	80.66±11.58

B. Result tables for nonuniformly distributed labels

Tables X, XI and XII present the mean and standard deviation of each method in each data stream for abrupt drifts, gradual drifts and real-world streams, respectively, with all amounts of labeled data in the scenario with nonuniformly distributed labels.

TABLE X: Mean and standard deviation of prequential accuracy on artificial streams with abrupt concept drifts and nonuniformly distributed labels.

Type of drift: abrupt							
	HTNB	OZABAG	OAUE	RCD	DDD	DP	OSNN
Labels at 5%							
Sine1	58.294 ±3.969	58.906 ±4.019	53.367 ±4.359	58.294 ±3.969	65.291 ±5.528	58.294 ±3.969	71.058 ±4.889
Sine2	52.629 ±7.370	53.597 ±7.512	49.532 ±4.639	52.629 ±7.370	64.570 ±8.998	59.415 ±10.504	69.194 ±6.275
Agrawal1	50.583 ±1.063	50.219 ±0.804	50.743 ±1.189	50.199 ±0.825	50.80 ±0.905	50.587 ±1.408	53.683 ±6.270
Agrawal2	51.636 ±2.370	51.731 ±2.190	51.610 ±2.426	51.717 ±2.48	50.727 ±1.375	51.766 ±2.691	53.917 ±6.127
Agrawal3	50.683 ±0.761	50.623 ±0.809	50.498 ±0.829	50.555 ±0.793	50.388 ±0.886	50.260 ±0.771	51.777 ±2.118
Agrawal4	50.259 ±1.191	49.919 ±1.331	49.880 ±0.975	50.077 ±1.415	50.256 ±1.077	50.358 ±1.038	51.409 ±1.901
SEA1	65.988 ±3.965	67.888 ±4.337	61.283 ±7.724	65.988 ±3.965	63.355 ±3.291	67.644 ±4.590	65.622 ±4.595
SEA2	72.348 ±6.124	70.884 ±7.147	67.962 ±12.371	72.348 ±6.124	71.988 ±6.665	69.056 ±4.643	67.310 ±6.592
STAGGER1	67.431 ±12.625	72.275 ±12.689	57.648 ±9.022	67.431 ±12.625	70.167 ±12.761	65.772 ±8.217	68.210 ±10.87
STAGGER2	49.900 ±10.548	54.191 ±6.404	46.794 ±8.424	49.900 ±10.548	56.090 ±5.645	62.858 ±10.526	58.756 ±2.676
Labels at 10%							
Sine1	54.664 ±10.711	54.923 ±10.320	53.108 ±5.651	54.664 ±10.711	55.903 ±4.847	55.941 ±4.417	59.840 ±6.310
Sine2	63.734 ±10.447	61.603 ±11.642	60.87 ±7.98	63.734 ±10.447	70.341 ±10.367	64.697 ±11.774	71.583 ±5.359
Agrawal1	50.620 ±2.098	51.435 ±3.406	50.849 ±1.908	50.552 ±1.378	51.081 ±1.845	51.127 ±1.927	55.559 ±7.257
Agrawal2	50.455 ±1.115	50.369 ±1.338	50.322 ±1.756	49.781 ±1.303	50.334 ±1.382	50.529 ±1.312	53.681 ±4.796
Agrawal3	50.670 ±1.017	50.756 ±1.152	50.985 ±1.001	50.462 ±1.082	50.877 ±1.049	50.48 ±0.948	51.058 ±2.008
Agrawal4	50.166 ±0.920	50.223 ±0.954	49.884 ±0.891	50.380 ±0.775	50.269 ±0.751	50.520 ±0.761	51.599 ±1.360
SEA1	64.519 ±5.206	65.211 ±5.197	59.301 ±6.175	64.519 ±5.206	63.342 ±5.242	58.747 ±8.059	63.591 ±4.972
SEA2	66.077 ±9.003	66.232 ±8.801	64.765 ±9.400	66.077 ±9.003	67.682 ±8.815	65.621 ±8.68	67.990 ±8.656
STAGGER1	62.011 ±12.875	64.368 ±12.482	62.798 ±8.631	62.011 ±12.875	66.628 ±10.259	65.759 ±10.599	69.589 ±8.08
STAGGER2	67.834 ±9.168	68.309 ±9.270	70.651 ±10.209	67.834 ±9.168	66.588 ±3.982	64.988 ±3.530	65.260 ±5.617
Labels at 20%							
Sine1	52.163 ±4.840	52.186 ±5.044	52.459 ±3.565	52.163 ±4.840	58.426 ±2.898	56.053 ±4.405	64.436 ±4.306
Sine2	60.667 ±11.014	61.536 ±11.674	65.824 ±9.321	60.667 ±11.014	70.559 ±8.244	58.376 ±11.605	73.422 ±3.974
Agrawal1	51.873 ±2.191	52.570 ±3.280	53.693 ±5.097	51.178 ±1.435	52.081 ±2.437	52.451 ±3.479	57.022 ±9.180
Agrawal2	52.421 ±4.602	52.191 ±2.977	52.391 ±3.846	52.607 ±4.603	52.554 ±4.009	52.163 ±4.614	56.533 ±7.528
Agrawal3	49.245 ±1.623	49.174 ±1.639	49.605 ±1.427	49.922 ±1.018	49.955 ±1.109	50.068 ±1.053	50.355 ±1.711
Agrawal4	50.724 ±1.556	50.935 ±1.961	51.218 ±1.767	50.028 ±1.181	50.763 ±1.689	50.521 ±1.392	52.821 ±3.441
SEA1	65.146 ±5.297	65.624 ±4.474	64.815 ±7.502	65.146 ±5.297	67.817 ±7.038	54.734 ±6.375	68.714 ±7.047
SEA2	73.478 ±4.871	74.658 ±3.923	69.085 ±9.609	73.478 ±4.871	79.030 ±5.654	58.844 ±10.128	73.625 ±9.589
STAGGER1	59.020 ±5.977	61.070 ±5.759	62.197 ±5.587	59.020 ±5.977	66.358 ±5.103	63.198 ±4.606	69.899 ±8.622
STAGGER2	65.777 ±5.84	60.528 ±7.369	60.569 ±9.948	65.777 ±5.84	65.903 ±8.399	61.304 ±6.659	65.532 ±5.092

TABLE XI: Mean and standard deviation of prequential accuracy on artificial streams with gradual concept drifts and nonuniformly distributed labels.

Type of drift: gradual							
	HTNB	OZABAG	OAUE	RCD	DDD	DP	OSNN
Labels at 5%							
Sine1	55.227 ±3.490	55.297 ±3.87	51.997 ±3.011	55.227 ±3.490	61.343 ±3.298	58.688 ±3.961	65.324 ±3.234
Sine2	54.752 ±8.963	55.025 ±9.989	49.692 ±2.982	54.752 ±8.963	60.364 ±7.992	56.658 ±9.841	66.803 ±7.472
Agrawal1	50.575 ±1.669	50.658 ±1.780	50.295 ±0.983	50.345 ±1.238	50.285 ±1.072	50.48 ±1.198	53.233 ±6.722
Agrawal2	49.881 ±0.994	49.937 ±1.102	50.036 ±0.725	49.788 ±1.067	49.992 ±1.037	49.727 ±0.973	53.461 ±4.067
Agrawal3	49.856 ±0.865	49.944 ±0.974	50.223 ±0.891	49.856 ±0.865	49.888 ±0.965	50.002 ±0.850	49.074 ±3.412
Agrawal4	50.723 ±1.351	50.755 ±1.485	50.767 ±1.120	50.723 ±1.351	50.640 ±1.326	50.327 ±1.432	52.165 ±3.086
SEA1	64.739 ±5.793	65.852 ±6.022	62.111 ±9.995	64.739 ±5.793	65.376 ±8.820	65.505 ±5.479	66.475 ±15.049
SEA2	65.376 ±6.919	66.708 ±5.575	56.462 ±5.227	65.376 ±6.919	70.717 ±6.153	65.700 ±6.844	67.990 ±4.933
STAGGER1	60.198 ±7.444	64.656 ±11.986	54.066 ±7.089	60.198 ±7.444	68.127 ±9.394	59.441 ±5.299	65.305 ±7.059
STAGGER2	52.234 ±5.473	51.758 ±6.079	48.822 ±2.616	52.234 ±5.473	57.244 ±6.027	54.952 ±4.184	59.712 ±4.051
Labels at 10%							
Sine1	53.328 ±8.387	53.528 ±8.253	50.749 ±5.987	53.328 ±8.387	54.457 ±3.851	54.304 ±4.273	61.655 ±6.573
Sine2	56.981 ±7.831	56.042 ±7.777	55.445 ±5.787	56.981 ±7.831	62.008 ±5.902	57.925 ±9.338	67.090 ±4.905
Agrawal1	50.683 ±1.241	50.696 ±1.408	50.937 ±1.328	50.624 ±1.251	49.974 ±0.826	50.635 ±1.955	54.106 ±6.980
Agrawal2	50.116 ±1.097	50.320 ±1.110	49.824 ±1.233	50.201 ±1.084	49.841 ±0.994	49.554 ±0.997	52.362 ±4.453
Agrawal3	50.709 ±1.785	50.327 ±1.633	50.917 ±1.622	50.259 ±1.207	51.024 ±1.388	50.998 ±1.749	51.294 ±2.337
Agrawal4	50.451 ±1.294	50.382 ±1.253	50.669 ±1.479	50.296 ±1.022	50.501 ±1.087	50.952 ±1.692	50.686 ±1.258
SEA1	72.740 ±5.892	72.282 ±5.215	66.907 ±7.946	72.740 ±5.892	68.451 ±5.85	66.783 ±6.548	69.206 ±7.562
SEA2	67.212 ±3.716	68.633 ±3.618	63.709 ±8.315	67.212 ±3.716	67.793 ±3.380	65.322 ±5.563	66.025 ±5.567
STAGGER1	53.931 ±11.685	58.017 ±7.890	58.543 ±7.975	53.931 ±11.685	62.609 ±10.100	59.357 ±7.419	65.978 ±7.671
STAGGER2	56.761 ±5.762	59.830 ±6.752	53.217 ±3.800	56.761 ±5.762	59.694 ±6.953	57.060 ±5.190	57.406 ±4.190
Labels at 20%							
Sine1	52.495 ±3.238	52.84 ±3.115	55.191 ±5.363	52.495 ±3.238	60.591 ±6.948	54.431 ±3.186	67.997 ±4.571
Sine2	60.768 ±7.826	60.801 ±7.873	59.612 ±6.832	60.768 ±7.826	64.148 ±6.648	54.856 ±8.367	66.480 ±3.747
Agrawal1	51.569 ±2.210	52.066 ±2.464	51.678 ±3.178	51.809 ±2.210	52.521 ±2.363	52.332 ±2.237	56.099 ±7.592
Agrawal2	52.817 ±2.324	51.454 ±1.487	51.734 ±1.709	51.298 ±1.575	52.256 ±1.641	51.607 ±1.694	54.561 ±3.011
Agrawal3	50.092 ±1.653	50.354 ±1.605	49.832 ±1.156	50.129 ±1.81	49.770 ±1.285	49.912 ±1.083	50.070 ±2.381
Agrawal4	51.691 ±1.602	51.689 ±1.592	51.154 ±1.620	51.329 ±1.450	51.272 ±2.019	51.263 ±1.360	51.386 ±2.492
SEA1	66.956 ±3.424	66.878 ±3.678	64.451 ±6.031	66.956 ±3.424	67.002 ±6.957	64.879 ±5.901	70.036 ±5.484
SEA2	63.994 ±3.394	63.987 ±3.821	60.932 ±5.021	63.994 ±3.394	63.830 ±5.566	60.114 ±7.192	65.742 ±6.721
STAGGER1	57.740 ±6.163	59.231 ±5.570	59.286 ±4.947	57.740 ±6.163	59.153 ±7.029	63.841 ±7.895	66.495 ±4.545
STAGGER2	63.169 ±3.649	66.561 ±6.649	60.994 ±4.812	63.169 ±3.649	67.493 ±6.849	63.766 ±4.499	64.202 ±3.564

TABLE XII: Mean and standard deviation of prequential accuracy on real-world streams with nonuniformly distributed labels.

Type of drift: real							
	HTNB	OZABAG	OAUE	RCD	DDD	DP	OSNN
Labels at 5%							
Elec	57.510 ±17.734	53.952 ±14.072	46.966 ±9.684	57.510 ±17.734	52.422 ±11.539	52.733 ±8.098	55.124 ±8.538
NOAA	57.329 ±4.121	57.485 ±4.140	59.274 ±9.84	57.329 ±4.121	58.817 ±7.145	59.887 ±6.200	69.205 ±3.584
Power Supply	54.528 ±12.735	54.405 ±12.755	58.520 ±10.436	54.528 ±12.735	52.575 ±12.280	55.300 ±8.883	58.369 ±8.474
Sensor	65.451 ±12.051	64.328 ±14.073	58.853 ±12.502	58.484 ±10.212	56.073 ±16.033	54.278 ±10.826	60.87 ±10.583
Labels at 10%							
Elec	55.804 ±8.504	56.809 ±8.473	56.004 ±12.357	55.804 ±8.504	51.889 ±10.580	54.990 ±11.028	51.199 ±9.622
NOAA	65.232 ±5.490	65.208 ±5.485	65.094 ±5.444	65.232 ±5.490	56.786 ±10.096	64.277 ±5.012	69.652 ±3.697
Power Supply	49.737 ±6.98	46.808 ±6.625	51.588 ±7.465	49.737 ±6.98	50.615 ±7.625	48.937 ±7.656	51.880 ±5.992
SensorClasses	63.824 ±12.269	63.007 ±14.211	58.476 ±11.988	59.020 ±10.841	55.099 ±11.947	51.752 ±10.502	59.993 ±15.509
Labels at 20%							
Elec	61.263 ±8.821	62.793 ±14.318	60.547 ±14.226	61.263 ±8.821	61.991 ±8.230	62.48 ±9.008	67.668 ±6.542
NOAA	52.429 ±4.239	52.478 ±4.236	54.436 ±6.405	52.429 ±4.239	52.789 ±4.934	61.767 ±8.466	70.429 ±3.126
Power Supply	60.82 ±5.167	57.864 ±6.228	53.837 ±6.647	60.82 ±5.167	55.654 ±7.201	54.234 ±7.926	50.102 ±7.445
Sensor	69.745 ±11.183	69.287 ±12.83	62.060 ±11.808	65.041 ±11.687	61.176 ±15.668	51.346 ±10.177	59.506 ±11.393

V. Results for F-score, precision and recall

In this Section, we show the detailed result tables for the F-score, precision and recall metrics evaluated in a prequential manner in our experiments. We group 72 streams according to type of drift, amount of labels and data stream in two analyses: with uniform and nonuniform labeling probabilities, shown in Subsections V-A and V-B, respectively. We used the Scott-Knott multiple comparison procedure to evaluate statistical differences in prequential F-score, precision and recall. Best-performing methods are successively assigned ranks $1, 2, \ldots, 7$.

The results for these added metrics, especially for F-score (which measures the trade-off between the number of false positives and false negatives), support the conclusions of the prequential accuracy analysis shown in the main paper. In particular, OSNN delivered significantly better trade-off between false positives and false negatives than all other methods when considering F-score across data streams for the nonuniformly distributed label, while being in the top ranked group in terms of F-score across data streams for the uniformly distributed labels.

A. *Uniformly distributed labels*

The Scott-Knott test was performed for each group of streams with uniform labeling distribution. The rankings of these groups for F-score, precision and recall are shown in Tables XIII, XIV, and XV, respectively. Algorithms with significantly superior predictive performance are highlighted in green.

Table XIII shows the rankings for the F-score metric, which is the harmonic mean between precision and recall. The F-score results indicate that, when the labels are uniformly distributed along the length of the streams grouped by amount of labels, OSNN was among the highest ranking algorithms for all groupings and it outperformed HTNB, OzaBag and OAUE in most cases. The fact that it performed similar to RCD, DDD and DP might indicate that these distributions of labels do not present a useful structure that OSNN can exploit to improve predictive performance over the other methods. Such a result follows the outcome from the groupings by concept drift and streams. The exception is the Agrawal stream, in which a meaningful underlying structure is present in the unlabeled data and is revealed and exploited by OSNN to deliver higher F-score than existing methods. Nevertheless, the F-score results show that OSNN delivered a competitive balance between the amount of false positives and false negatives compared to other approaches. OSNN is consistently among the highest ranked algorithm in most groups, which denotes its ability to use labeled data well when unlabeled data does not help.

TABLE XIII: Statistical ranking of prequential F-score on streams grouped by factors with uniformly distributed labels.

Groups	HTNB	OZABAG	OAUE	RCD	DDD	DP	OSNN
Grouped by amount of labeled data							
5%	1	1	2	1	1	1	1
10%	2	2	3	1	1	1	1
20%	2	2	2	1	1	1	1
Grouped by type of concept drift							
Abrupt	2	2	3	1	1	1	1
Gradual	2	2	3	1	1	1	1
Real-world	1	1	1	1	1	1	1
Grouped by streams							
Sine	3	3	4	1	1	1	2
Agrawal	2	2	3	2	2	2	1
SEA	1	1	2	1	1	1	1
STAG.	3	3	3	2	1	1	2
Elec	2	1	2	1	1	1	1
NOAA	1	1	1	1	1	1	1
Power S.	1	1	1	1	1	1	1
Sensor	1	1	1	1	1	1	1
All streams	2	2	3	1	1	1	1

Highlighted ranks denote significant superior performance.

The precision rankings in Table XIV also indicate that, when the labels are uniformly distributed along the length of the streams grouped by amount of labels, RCD, DDD, DP and OSNN have similar amounts of false positives. However, OSNN is consistently among the highest ranked algorithm in most groups, which again denotes its ability to use labeled data well when unlabeled data does not help.

TABLE XIV: Statistical ranking of prequential precision on streams grouped by factors with uniformly distributed labels.

Groups	HTNB	OZABAG	OAUE	RCD	DDD	DP	OSNN
Grouped by amount of labeled data							
5%	1	1	2	1	1	1	1
10%	2	2	2	1	1	1	1
20%	2	2	2	1	1	1	1
Grouped by type of concept drift							
Abrupt	2	2	3	1	1	1	1
Gradual	2	2	2	1	1	1	1
Real-world	1	1	1	1	1	1	1
Grouped by streams							
Sine	2	2	2	1	1	1	1
Agrawal	2	2	3	2	2	2	1
SEA	1	1	3	1	1	1	2
STAG.	4	4	4	3	1	1	2
Elec	1	1	2	2	1	1	1
NOAA	2	2	2	2	1	2	1
Power S.	1	1	1	1	1	1	1
Sensor	2	1	1	2	1	1	1
All streams	2	2	3	1	1	1	1

Highlighted ranks denote significant superior performance.

The rankings for prequential recall (Table XV) follow the results obtained for prequential precision in Table XIV. RCD, DDD, DP and OSNN deliver statistically similar amounts of false negatives, that is, they are able to recover similar amounts of instances of the positive class. The recall metric also indicates that the uniform labeling distributions do not present an meaningful manifold structure that OSNN can exploit to improve predictive performance over the other methods. However, OSNN is consistently among the highest ranked algorithm in most groups, which denotes its ability to exploit labeled data well when unlabeled data is not useful.

TABLE XV: Statistical ranking of prequential recall on streams grouped by factors with uniformly distributed labels.

Groups	HTNB	OZABAG	OAUE	RCD	DDD	DP	OSNN
Grouped by amount of labeled data							
5%	1	1	2	1	1	1	1
10%	2	2	3	1	1	1	1
20%	2	2	2	1	1	1	1
Grouped by type of concept drift							
Abrupt	2	2	3	1	1	1	1
Gradual	2	2	3	1	1	1	1
Real-world	1	1	1	1	1	1	1
Grouped by streams							
Sine	3	3	4	1	1	1	2
Agrawal	2	2	3	2	2	2	1
SEA	1	1	2	1	1	1	1
STAG.	5	4	6	3	1	1	2
Elec	2	1	2	1	1	1	1
NOAA	1	1	1	1	2	1	2
Power S.	1	1	1	1	1	1	1
Sensor	1	1	1	1	1	1	1
All streams	2	2	3	1	1	1	1

Highlighted ranks denote significant superior performance.

The results for F-score, precision and recall (Tables XIII, XIV and XV) follow the prequential accuracy outcome of the experiment with uniformly distributed labels in main manuscript. The overall performance across all streams was also assessed. OSNN regularly outperformed HTNB, OzaBag, OAUE and produced similar generalization to RCD, DDD and DP. Independent of the factors of our analysis, OSNN is consistently among the highest ranked approaches. OSNN's ability to adapt and to exploit unlabeled data could compensate for the use of ensembles in existing methods when very few labels are available.

B. Nonuniformly distributed labels

The Scott-Knott test was performed for each group of streams with nonuniform labeling distribution. The rankings of these groups for F-score, precision and recall are shown in Tables XVI, XVII and XVIII, respectively. Algorithms with significantly superior predictive performance are highlighted in green.

Since the F-score metric is the harmonic mean between precision and recall and most of the streams have balanced classes, the results for F-score in Table XVI follow the rankings of prequential accuracy for nonuniformly distributed labels in the experiments in main manuscript. OSNN was able to consistently deliver the highest F-score in most groups. For 10% and 20% of labeled data, OSNN was superior to state-of-the-art ensemble methods (RCD, DDD and DP). For abrupt drifts, it delivered higher F-score than HTNB, OAUE, RCD, DDD and DP. It was the superior approach for gradual drifts. This analyses for types of drifts demonstrate OSNN's ability to exploit unlabeled data to adapt its centers and weights when a sudden or gradual drift occurs. For artificial data streams, OSNN was the superior method for Sine and STAGGER, and superior to most algorithms for Agrawal. For real-world data streams, OSNN was superior to DDD and DP for all streams, except Elec, for which it was a tie. OSNN was the superior approach for NOAA. This fact indicate the presence of useful underlying manifold structures in the data. When all streams and all factors are analyzed, OSNN produced significantly superior predictive performance than all other approaches.

TABLE XVI: Statistical ranking of prequential F-score on streams with nonuniformly distributed labels.

Groups	HTNB	OZABAG	OAUE	RCD	DDD	DP	OSNN
Grouped by amount of labels							
5%	**1**	**1**	2	**1**	**1**	**1**	**1**
10%	**1**	**1**	**1**	**1**	**1**	**1**	**1**
20%	**1**	**1**	**1**	**1**	**1**	**1**	**1**
Grouped by type of concept drift							
Abrupt	**1**	**1**	**1**	**1**	**1**	**1**	**1**
Gradual	2	2	2	2	2	2	**1**
Real-world	**1**	**1**	2	2	2	2	**1**
Grouped by stream							
Sine	2	2	2	2	**1**	**1**	**1**
Agrawal	2	**1**	2	2	2	2	**1**
SEA	**1**	**1**	2	**1**	**1**	**1**	**1**
STAGGER	3	2	3	3	2	3	**1**
Elec	**1**	**1**	**1**	**1**	**1**	**1**	**1**
NOAA	2	2	2	2	2	2	**1**
Power S.	**1**	**1**	2	**1**	2	2	2
Sensor	**1**	**1**	2	2	2	2	2
All streams	2	2	3	2	2	2	**1**

Highlighted ranks denote significant superiority.

For prequential precision (Table XVII), OSNN was consistently between the highest scoring algorithms, especially for Sine, Agrawal and NOAA streams. However, DDD and DP also delivered low false positives in most groups. There is typically a trade-off between precision and recall, and the results for precision were contrasted by the recall metric as shown in Table XVIII. OSNN produced the statistically lowest number of false negative in most groups, especially for gradual drifts, STAGGER and NOAA. In fact, when all streams were considered, OSNN was the method with the lowest number of false negatives.

TABLE XVII: Statistical ranking of prequential precision on streams with nonuniformly distributed labels.

Groups	HTNB	OZABAG	OAUE	RCD	DDD	DP	OSNN
Grouped by amount of labels							
5%	**1**	**1**	2	**1**	**1**	**1**	**1**
10%	**1**	**1**	**1**	**1**	**1**	**1**	**1**
20%	**1**	**1**	**1**	**1**	**1**	**1**	**1**
Grouped by type of concept drift							
Abrupt	2	2	2	2	**1**	**1**	**1**
Gradual	**1**	**1**	2	**1**	**1**	**1**	**1**
Real-world	**1**	**1**	**1**	**1**	2	2	**1**
Grouped by stream							
Sine	3	3	3	3	2	2	**1**
Agrawal	2	2	3	2	2	2	**1**
SEA	**1**	**1**	2	**1**	**1**	**1**	**1**
STAGGER	2	2	2	2	**1**	**1**	2
Elec	**1**	**1**	**1**	**1**	**1**	**1**	**1**
NOAA	2	2	2	2	2	2	**1**
Power S.	**1**	**1**	**1**	**1**	**1**	**1**	**1**
Sensor	2	2	**1**	**1**	**1**	**1**	**1**
All streams	**1**	**1**	2	**1**	**1**	**1**	**1**

Highlighted ranks denote significant superiority.

TABLE XVIII: Statistical ranking of prequential recall on streams with nonuniformly distributed labels.

Groups	HTNB	OZABAG	OAUE	RCD	DDD	DP	OSNN
Grouped by amount of labels							
5%	1	1	1	1	1	1	1
10%	2	1	2	2	2	2	1
20%	2	1	2	2	2	2	1
Grouped by type of concept drift							
Abrupt	2	1	2	2	2	2	1
Gradual	2	2	2	2	2	2	1
Real-world	1	1	1	1	1	1	1
Grouped by stream							
Sine	2	2	2	2	1	1	1
Agrawal	2	1	2	2	2	2	1
SEA	1	1	2	1	1	1	1
STAGGER	3	2	3	3	2	3	1
Elec	1	1	1	1	1	1	1
NOAA	2	2	2	2	2	2	1
Power S.	1	1	2	1	2	2	2
Sensor	1	1	2	2	2	2	1
All streams	3	2	3	3	3	3	1

Highlighted ranks denote significant superiority.

For nonuniformly distributed labels, OSNN is the highest ranking method in precision in most cases, however, when considering all streams, there is no significant difference between five of these algorithms. On the other hand, OSNN is able to significantly reduce the number of false negatives in comparison to the other approaches. In fact, when considering all streams, OSNN delivers the statistically highest recall. The F-score metric (which combines precision and recall) shows consistent results to the ones found for prequential accuracy in the main manuscript. When label arrival depends on the region of input space instead of time, the advantages of the data representation and regularization mechanisms in OSNN over single and ensemble learners become more evident. OSNN was the approach with highest performance in the vast majority of cases. In fact, when we grouped all streams, OSNN produced superior generalization compared to all other algorithms. Such a result demonstrates that OSNN is the most robust classifier to scarce labels, different types of concept drift and diverse data from different environments, with uniformly and nonuniformly distributed labels.

VI. Experiments with visual data

Despite our approach not being proposed for the specific problem of image classification, we have also run experiments with four image datasets to evaluate its predictive performance on such a problem. In this Section, we analyze the results for the accuracy, F-score, precision and recall metrics evaluated in a prequential manner in our experiments. We group 4 visual streams according to amount of labels and data stream in two analyses: with uniform and nonuniform labeling probabilities, shown in Subsections VI-A and VI-B, respectively. We used the Scott-Knott multiple comparison procedure to evaluate statistical differences in prequential accuracy, F-score, precision and recall. Best-performing methods are successively assigned ranks $1, 2, \dots, 7$.

We used 4 data streams: Outdoor [8], Rialto [8], CIFAR [7] and Rotated MNIST [9], where the instances in Outdoor and Rialto are naturally ordered forming a true data stream. CIFAR and Rotated MNIST are datasets typically used for offline learning, as they have images in randomized orders. However, they were used to simulate data streams by presenting such images sequentially to the machine learning approaches. The summary of these streams is as follows:

- **Outdoor** [8] consists of color images recorded by a smartphone camera in a garden environment of 40 different objects, such as balls, shoes, pliers, cans, among others. We selected objects 0 and 19 as the classes for our classification problem. This stream has 200 instances and 21 features (dimensions).
- **Rialto** [8] contains color images extracted from time-lapse videos recorded by a webcam in a fixed position. The recordings cover 20 consecutive days from May to June 2016, capturing various colorful buildings next to the famous Rialto bridge in Venice. We employed the buildings number 0 and 4 were considered as the classes for our classification problem. It has 16,450 instances and 27 features.
- **CIFAR** contains resized (16x16 pixels) gray-scale images from the original CIFAR10 dataset [7]. We selected instances from classes “automobile” and “dog”. This stream has 12,000 instances and 256 features.
- **Rotated MNIST** [9] consists of rotated resized (16x16 pixels) gray-scale images of handwritten “0” and “1” digits. It has 12,670 instances and 256 features.

A. *Uniformly distributed labels*

The Scott-Knott test was performed for each group of streams with uniform labeling distribution. The rankings of these groups for accuracy, F-score, precision and recall are shown in Tables XIX, XX, XXI, and XXII, respectively. Algorithms with significantly superior predictive performance are highlighted in green.

Table XIX demonstrates the significantly superior performance of OSNN for most groups according to prequential accuracy for streams with uniformly distributed labels. When OSNN was not the highest scoring method, it was among the top performing approaches. Although none of these methods was specially designed for visual data, OSNN was able to exploit unlabeled data and learn useful structures in the data when compared to the other approaches. In fact, when grouping by streams, OSNN obtained the highest accuracy for CIFAR and Rotated MNIST. When considering all visual streams, OSNN was the approach with significantly best accuracy among all methods.

TABLE XIX: Statistical ranking of prequential accuracy on visual streams grouped by factors with uniformly distributed labels.

Groups	HTNB	OZABAG	OAUE	RCD	DDD	DP	OSNN
Grouped by amount of labeled data							
5%	1	1	1	1	1	1	1
10%	2	2	2	2	2	2	1
20%	2	2	2	2	2	2	1
Grouped by streams							
Outdoor	1	1	1	1	1	1	1
Rialto	1	1	1	1	1	1	1
CIFAR	2	2	2	2	2	2	1
Rotated MNIST	2	2	2	2	2	2	1
All streams	2	2	2	2	2	2	1

Highlighted ranks denote significant superior performance.

In Table XX, we show the results for the prequential F-score metric with uniformly distributed labels. F-score is the harmonic mean of precision and recall. Such a metric shows the trade-off in each method between the amounts of false positives and false negatives. This Table supports the results of Table XIX, as it also indicates that OSNN delivered significantly higher predictive performance for most groups, delivering the best compromises between false positives and false negatives among all approaches. In fact, when grouping by streams, OSNN outperformed all other methods in terms of accuracy for CIFAR and Rotated MNIST, while obtaining competitive results for Outdoor and Rialto. Therefore, when considering the behaviour across all visual streams, OSNN was the approach with significantly best accuracy among all methods.

TABLE XX: Statistical ranking of prequential F-score on visual streams grouped by factors with uniformly distributed labels.

Groups	HTNB	OZABAG	OAUE	RCD	DDD	DP	OSNN
Grouped by amount of labeled data							
5%	1	1	2	1	1	1	1
10%	2	2	2	2	2	2	1
20%	2	2	2	2	2	2	1
Grouped by streams							
Outdoor	1	1	1	1	1	1	1
Rialto	1	1	1	1	1	1	1
CIFAR	2	2	2	2	2	2	1
Rotated MNIST	2	2	2	2	2	2	1
All streams	2	2	2	2	2	2	1

Highlighted ranks denote significant superior performance.

Table XXI present the results for prequential precision with uniformly distributed labels. For most cases, all algorithms obtained similar amounts of false positives. This might denote the challenge of learning a good decision boundary for visual data when the algorithms do not take advantage of particular features of the images (e.g. notion of neighborhood among pixels).

TABLE XXI: Statistical ranking of prequential precision on visual streams grouped by factors with uniformly distributed labels.

Groups	HTNB	OZABAG	OAUE	RCD	DDD	DP	OSNN
Grouped by amount of labeled data							
5%	1	1	2	1	1	1	1
10%	1	1	1	1	1	1	1
20%	1	1	1	1	1	1	1
Grouped by streams							
Outdoor	1	1	1	1	1	1	1
Rialto	2	1	3	2	1	2	3
CIFAR	2	2	2	2	2	2	1
Rotated MNIST	1	1	1	1	1	1	1
All streams	1	1	1	1	1	1	1

Highlighted ranks denote significant superior performance.

Table XXII shows the prequential recall with uniformly distributed labels. OSNN obtained the significantly lower number of false negatives than the other algorithms for most groups. For the other groups, OSNN was still able to be among the highest performing approaches. OSNN delivered the highest recall for CIFAR and Rotated MNIST. When considering the performance across all streams, OSNN was the best approach with significantly better recall than the other methods.

These results indicate that OSNN is able to learn from both labeled and unlabeled data to construct meaningful models for visual data with uniformly distributed labels when compared to the other approaches. Such outcome is important since none of these algorithms are specially designed to learn from visual data.

TABLE XXII: Statistical ranking of prequential recall on visual streams grouped by factors with uniformly distributed labels.

Groups	HTNB	OZABAG	OAUE	RCD	DDD	DP	OSNN
Grouped by amount of labeled data							
5%	1	1	1	1	1	1	1
10%	2	2	2	2	2	2	1
20%	2	2	2	2	2	2	1
Grouped by streams							
Outdoor	1	1	1	1	1	1	1
Rialto	1	1	1	1	1	1	1
CIFAR	2	2	2	2	2	2	1
Rotated MNIST	2	2	2	2	2	2	1
All streams	2	2	2	2	2	2	1

Highlighted ranks denote significant superior performance.

B. *Nonuniformly distributed labels*

The Scott-Knott test was performed for each group of streams with nonuniform labeling distribution. The rankings of these groups for accuracy, F-score, precision and recall are shown in Tables XXIII, XXIV, XXV, and XXVI, respectively. Algorithms with significantly superior predictive performance are highlighted in green.

For prequential accuracy (Table XXIII), OSNN was significantly better than all other approaches for CIFAR and Rotated MNIST. The results for F-score, in Table XXIV, show that OSNN delivered the best trade-off between false positives and false negatives for 20% of labeled data and for CIFAR. For the other groups, there was single superior algorithm. When considering the performance across all data streams, most approaches performed similarly in terms of accuracy and F-score. A similar outcome is observed for prequential precision, as shown in Table XXV. In terms of prequential recall, as shown in Table XXVI, OSNN performed significantly better than all other methods in terms of performance across all data streams.

TABLE XXIII: Statistical ranking of prequential accuracy on visual streams grouped by factors with nonuniformly distributed labels.

Groups	HTNB	OZABAG	OAUE	RCD	DDD	DP	OSNN
Grouped by amount of labeled data							
5%	1	1	1	1	1	1	1
10%	1	1	1	1	1	1	1
20%	1	1	1	1	1	1	1
Grouped by streams							
Outdoor	1	1	2	1	1	1	2
Rialto	1	1	2	1	1	1	2
CIFAR	2	2	2	2	2	2	1
Rotated MNIST	2	2	2	2	2	2	1
All streams	1	1	2	1	1	1	1

Highlighted ranks denote significant superior performance.

TABLE XXIV: Statistical ranking of prequential F-score on visual streams grouped by factors with nonuniformly distributed labels.

Groups	HTNB	OZABAG	OAUE	RCD	DDD	DP	OSNN
Grouped by amount of labeled data							
5%	1	1	2	1	1	1	1
10%	1	1	1	1	1	1	1
20%	2	2	2	2	2	2	1
Grouped by streams							
Outdoor	1	1	2	1	1	1	2
Rialto	1	1	1	1	1	1	1
CIFAR	2	2	2	2	2	2	1
Rotated MNIST	2	2	2	2	2	1	1
All streams	1	1	2	1	1	1	1

Highlighted ranks denote significant superior performance.

TABLE XXV: Statistical ranking of prequential precision on visual streams grouped by factors with nonuniformly distributed labels.

Groups	HTNB	OZABAG	OAUE	RCD	DDD	DP	OSNN
Grouped by amount of labeled data							
5%	1	1	2	1	1	1	1
10%	1	1	1	1	1	1	1
20%	1	1	1	1	1	1	1
Grouped by streams							
Outdoor	1	1	1	1	1	1	1
Rialto	1	1	2	1	1	1	2
CIFAR	2	2	2	2	2	2	1
Rotated MNIST	1	1	2	1	1	1	2
All streams	1	1	2	1	1	1	1

Highlighted ranks denote significant superior performance.

TABLE XXVI: Statistical ranking of prequential recall on visual streams grouped by factors with nonuniformly distributed labels.

Groups	HTNB	OZABAG	OAUE	RCD	DDD	DP	OSNN
Grouped by amount of labeled data							
5%	1	1	2	1	1	1	1
10%	1	1	1	1	1	1	1
20%	2	2	2	2	2	2	1
Grouped by streams							
Outdoor	1	1	2	1	1	1	2
Rialto	1	1	1	1	1	1	1
CIFAR	2	2	2	2	2	2	1
Rotated MNIST	2	2	2	2	2	2	1
All streams	2	2	3	2	2	2	1

Highlighted ranks denote significant superior performance.

The results for nonuniformly distributed labels might be explained by the fact that these algorithms are not designed to exploit specific features in visual data (e.g. neighborhood among pixels) that help in the learning of a predictive models for images. Without the ability to fully access many useful features in images, these algorithms suffer with the nonuniform distribution of labels. The scarce and nonuniform labels hinders the learning of good decision boundaries by not revealing and distorting important structures in the data. Without the ability to use image-specific structures, none of these approaches was able to significantly overcome the lack and misleading of label information for the majority of groups. Only OSNN was able to be the exception for prequential accuracy in CIFAR and Rotated MNIST; for F-score in 20% of labels and CIFAR; for precision in CIFAR; and for prequential recall in 20% of labels, CIFAR, Rotated MNIST and prequential recall across all streams. These results indicate that the SLVQ may be able to extract useful information from scarce and misleading labels even without the use of image-specific features.

VII. Severity of concept drifts

Our experiments evaluate the proposed approach on synthetic data streams containing different types of drift with different severities, including recurrent drifts. Table XXVII summarizes the data streams. In particular, the streams Sine1, Sine2, Agrawal3, SEA1, SEA2, STAGGER1 and STAGGER2 have recurrent concepts, as shown in the column containing the concept sequences. Table XXVIII shows the severities.

TABLE XXVII: Summary of data streams.

Stream	Concept sequences	Number of inst.	Dim.
	Artificial data streams		
Sine1	$r_3 \rightarrow r_4 \rightarrow r_3$	12000	4
Sine2	$r_1 \rightarrow r_2 \rightarrow r_3 \rightarrow r_4 \rightarrow r_1$	20000	4
Agrawal1	$r_1 \rightarrow r_3 \rightarrow r_4 \rightarrow r_7 \rightarrow r_{10}$	20000	36
Agrawal2	$r_7 \rightarrow r_4 \rightarrow r_6 \rightarrow r_5 \rightarrow r_2 \rightarrow r_9$	24000	36
Agrawal3	$r_4 \rightarrow r_2 \rightarrow r_1 \rightarrow r_3 \rightarrow r_4$	20000	36
Agrawal4	$r_1 \rightarrow r_3 \rightarrow r_6 \rightarrow r_5 \rightarrow r_4$	20000	36
SEA1	$r_4 \rightarrow r_3 \rightarrow r_1 \rightarrow r_2 \rightarrow r_4$	20000	3
SEA2	$r_4 \rightarrow r_1 \rightarrow r_4 \rightarrow r_3 \rightarrow r_2$	20000	3
STAGGER1	$r_1 \rightarrow r_2 \rightarrow r_3 \rightarrow r_2$	16000	7
STAGGER2	$r_2 \rightarrow r_3 \rightarrow r_1 \rightarrow r_2$	16000	7
	Real-world data streams		
Elec	–	27549	7
NOAA	–	18159	8
Power S.	–	29928	2
Sensor	–	130073	5

TABLE XXVIII: Severity of Drifts as the Percentage Difference Between the Old and New Concepts. Adjusted from: [5]

	Sine			SEA				STAGGER	
	r1	r2	r3	r1	r2	r3	r4	r1	r2
r2	100.0%	-	-	8.5%	-	-	-	59.3%	-
r3	26.8%	73.2%	-	7.4%	16.0%	-	-	77.8%	48.1%
r4	73.2%	26.8%	100.0%	13.1%	4.6%	20.6%	-	-	-
r5	-	-	-	23.9%	32.5%	16.5%	37.1%	-	-

	Agrawal								
	r1	r2	r3	r4	r5	r6	r7	r8	r9
r2	53.9%	-	-	-	-	-	-	-	-
r3	53.1%	50.8%	-	-	-	-	-	-	-
r4	53.9%	20.5%	50.8%	-	-	-	-	-	-
r5	53.4%	47.6%	50.7%	47.7%	-	-	-	-	-
r6	69.9%	28.9%	51.2%	35.5%	48.1%	-	-	-	-
r7	50.5%	53.3%	50.1%	53.5%	60.1%	57.2%	-	-	-
r8	33.5%	60.4%	46.5%	59.6%	59.6%	59.8%	49.8%	-	-
r9	50.4%	53.3%	50.2%	53.5%	59.9%	57.3%	6.0%	49.5%	-
r10	32.9%	61.3%	46.5%	61.3%	60.0%	59.9%	51.1%	1.8%	51.1%

All percentage differences were calculated using Eq. 1 based on one million random generated examples.

The severity of a drift is calculated as the percentage difference between the old concept and the new concept, calculated as follows [5]:

$$diff(r_a, r_b) = \frac{\sum_{i=1}^{n} |y^i_{r_a} - y^i_{r_b}|}{n} \tag{1}$$

where $y^i_{f_a}$ and $y^i_{f_b}$ are the class labels determined by the a-th and b-th functions of a generator, respectively, and n is the total number of examples generated uniformly at random to calculate the severity. According to [5], a concept drift could be considered severe when the concepts before and after the drift have at least around 50% difference, and mild if the difference is around 25%.

VIII. Adaptive learning rate

Our approach does not need to explicitly detect the type of drift in order to decide how much to increase (decrease) the learning rate. The amount by which the learning rate increases (decreases) is determined by backtracking line search [1], [3] (Algorithm 1). This procedure starts with a large learning rate of $\eta = 1$ and iteratively reduces the learning rate until a learning rate that results in a decrease of the loss is found (or the minimum allowed learning rate of tol is reached).

Algorithm 1 Backtracking line search

```
1:  **Input:** $B^{(t)}, w^{(t)}, H^{-1}$
2:  **Output:** $w^{(t+1)}$
3:  $\eta \leftarrow 1$
4:  $tol \leftarrow 10^{-8}$
5:  **while** $\eta > tol$ **do**
6:      $\Delta w \leftarrow -\eta \mathbf{H}^{-1} \nabla_w \mathcal{L}$
7:      **if** $\mathcal{L}(B^{(t)}, w^{(t)}) < \mathcal{L}(B^{(t)}, w^{(t)} + \Delta w)$ **then**
8:          $\eta \leftarrow \eta/2$
9:      **else**
10:         $w^{(t+1)} \leftarrow w^{(t)} + \Delta w,$
11:         $\eta \leftarrow 0$
12:     **end if**
13: **end while**
14: **if** $\eta > 0$ **then**
15:     Armijo condition not fulfilled.
16: **end if**
```

Algorithm 1 works in conjunction with the Newton-Raphson method (or other optimizers, such as gradient descent). One of the advantages of using the Newton-Raphson method is that it calculates the curvature information of the loss function [3] as it uses Hessian matrix (i.e. second-order properties of the error surface, which are controlled by the Hessian matrix) [3]. Such information is very useful for identifying changes in the slope of the loss function. When a concept drift occurs, the current loss function slope changes. Such a change is affected by, among other factors, the type of concept drift [6], [4]. Abrupt drifts present potentially sudden changes to the loss surface, whilst gradual drifts reveal more parsimonious changes on the current curvatures.

Therefore, Algorithm 1 attempts to fit Δw into the current weights $w^{(t)}$ with a decreasing η. When an abrupt drift happens, OSNN learns a substantial correction in the weights Δw. In particular, the loop from line 6 will iterate few times, leading to the adoption of a larger learning rate η, as a fairly large learning rate will result in a decrease in the loss. When a gradual drift happens, OSNN learns more cautious correction in the weights Δw. In particular, the loop from line 6 will iterate several times, because large learning rates will not result in a reduction in the loss.

In Figures 23 and 24, we show the training of C and η throughout the Sine1 stream with abrupt and gradual drifts, respectively. We plot η (orange plot) and the function $\Delta C = \sum_i ||c_i^{(t)} - c_i^{(t-1)}||$ (blue plot) as a measure of the adaptation of C at each time step[1].

[1] The codebook starts to vary (blue line) after the first H instances are received to form this set, that is, $t > H$.

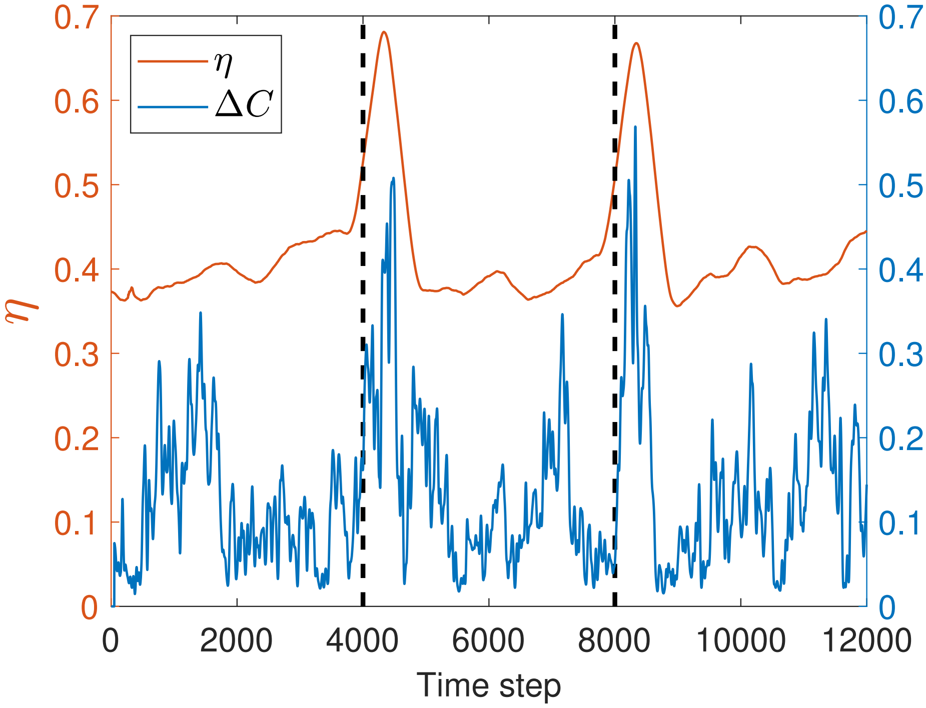

Fig. 23: Adaptive C and η for Sine1 with abrupt concept drifts (dashed lines) and uniform labeling distribution.

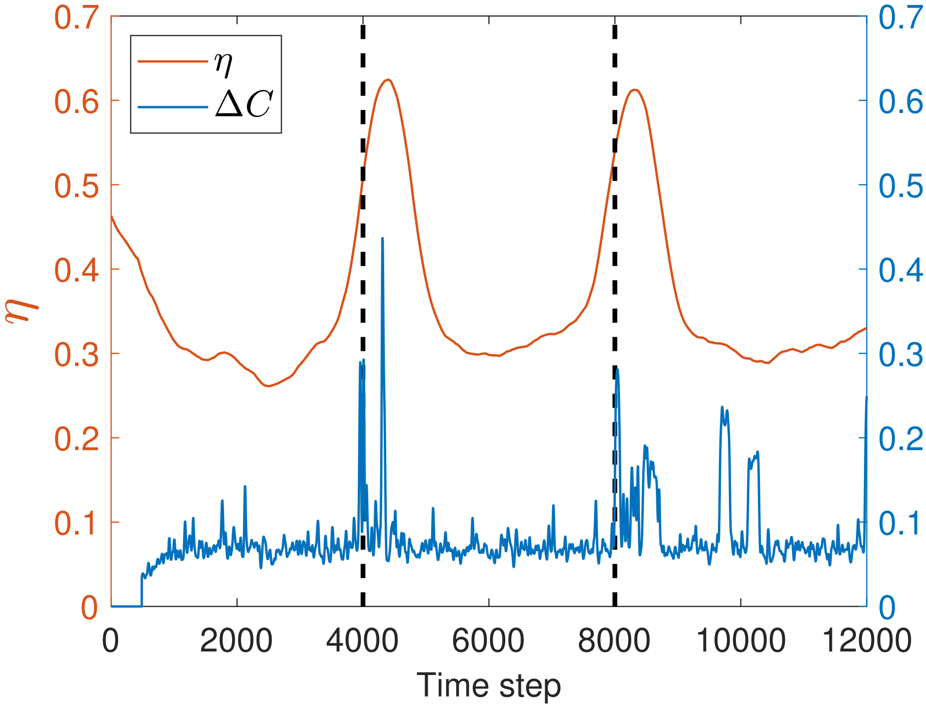

Fig. 24: Adaptive C and η for Sine1 with gradual concept drifts (dashed lines) and uniform labeling distribution.

In Figure 23, after each abrupt drift (with time steps denoted by dashed lines), the learning rate increases because OSNN produces a weight update Δw that will potentially cause a large decrease in the loss function and Algorithm 1 will identify (line 7) and take advantage of that fact by delivering a high η (the while loop will have few steps). At the other time steps, the weight update should be smaller because there is no change in the current concept (the impact of update Δw should be reduced), therefore Algorithm 1 will produce smaller η (the while loop will have several steps). In contrast, in Figure 24, OSNN produces a more parsimonious response to gradual drifts (dashed lines). The learning rate increases and decreases more smoothly after drifts because our algorithm is able to detect that a certain amount of the knowledge of the previous concept should be kept while learning the weights for a new one. In line 7, the algorithm would detect that a large weight update towards the new concept would cause a higher loss (due to the presence of the previous concept in the stream), then it would iterate and produce a smaller η (line 8) to fit Δw so that both concepts can be learned gradually. It is important to highlight that the peak learning rate produced by OSNN for abrupt drifts (Figure 23) is higher than the peak for gradual drifts (Figure 24). This result is expected since OSNN is able to adapt faster to abrupt drifts, which demands larger step sizes (learning rates).

The codebook variations (highlighted in blue) in Figures 23 and 24 show that OSNN, via SLVQ, is able to train the centers differently in presence of abrupt and gradual drifts, respectively. For abrupt drifts, SLVQ relocates the centers more quickly towards the region of the new concept due to the sudden arrival of instances in new regions and sudden absence of instances in the regions of the previous concept. In gradual drifts, instances from two different concepts arrive at the same time window. SVLQ partially adapts its centers to the new concept, while maintaining knowledge from the previous one. This is the reason why the variation of the positions of the codebook in Figure 24 is smoother than in Figure 23, as the latter required quicker adaptation.

Overall, our training method is able to learn appropriate directions and curvatures for the optimization of $w^{(t)}$ for abrupt and gradual drifts. The line search can adjust the learning step size (i.e. the amount of correction) according to the type of changes that are occurring.

IX. Table of symbols

In Table XXIX, we present the table of symbols of this work.

TABLE XXIX: Table of symbols

Symbol	Description
x	Input instance
y	True instance label
t	Time step
D	Input dimension
$B^{(t)}$	Minibatch at time step t
$B_l^{(t)}$	Set of labeled instances in $B^{(t)}$
$B_u^{(t)}$	Set of unlabeled instances in $B^{(t)}$
L	Size of $B_l^{(t)}$
U	Size of $B_u^{(t)}$
N	Size of $B^{(t)}$
$C^{(t)}$	Set of network centers (codebook) at time step t
$V^{(t)}$	Set of graph vertices at time step t
H	Number of network hidden neurons
$S^{(t)}$	Similarity matrix
f_i	Learner output as posterior class probabilities for instance $x_i \in B^{(t)}$
u_i	Pseudo-label for instance $x_i \in B^{(t)}$
w	Weight vector
$\mathcal{L}$	Loss function
ϕ_{ij}	Basis function output for instance x_i and neuron j
z_i	Net input for neuron i
R_i	Region of influence of neuron i in $\mathbb{R}^D$
σ_i	Width of basis function of neuron i
CL	Number of classes
$q(x)$	Vector quantization of x
$\mathcal{I}_1$	Functional for the minimization of the average quantization error
g_i	Density function for basis function of neuron i
K_i	Scaling factor for g_i
$\mathcal{I}_2$	Functional for the minimization of the average quantization error
$\mathcal{I}_{emp}$	Approximation of $\mathcal{I}_2$ via empirical risk minimization
$B'^{(t)}$, $B''^{(t)}$	Sets containing labeled instances of the majority and minority classes of R_i, respectively
β	Scalar that controls the radius of influence of each basis function
H	Hessian matrix
	Acronyms
1NN	1-Nearest Neighbor
DDD	Diversity for Dealing with Drift

TABLE XXIX: Table of symbols

Symbol	Description
HTNB	Hoeffding Tree with Naive Bayesian Learning
MR	Manifold Regularization
OAUE	Online Accuracy Update Ensemble
OSNN	Online Semisupervised Radial Basis Function Neural Network
OzaBag	Online Bagging
RCD	Recurring Concept Drift
SLVQ	Semisupervised Learning Vector Quantization
VQ	Vector quantization

X. Table of baseline hyperparameters

Table XXX depicts the hyperparameter probability distributions used in randomized search for tuning the baseline algorithms in our study.

TABLE XXX: Table of hyperparameter ranges for randomized search.

Hyperparameter	Probability distributions and ranges [2] used by the random search used for hyperparameter tuning
	Ozabag
ensemble size	uniform in [1,20]
	OAUE
ensemble size	uniform in $[1, 20]$
window size d	uniform in $[1, 1000]$
base learner	HTNB
	RCD
ensemble size	uniform in $[1, 20]$
p-value s	uniform in $[0.01, 0.05]$
rate the tests t	uniform in $[1, 1000]$
k	fixed at 1
batch size b	uniform in $[1, 1000]$
base learner	HTNB
drift detection	uniform in {DDM, EDDM, ADWIN}
	DDD
ensemble size	uniform in $[1, 20]$
weight W	uniform in $[5 * 10^{-4}, 5 * 10^{-1}]$
p_l	uniform in $[0, 1]$
p_h	uniform in $[0, 1]$
base learner	HTNB
drift detection	uniform in {DDM, EDDM, ADWIN}
	DP
ensemble size	uniform in $[1, 20]$
batch size	uniform in $[1, 1000]$
statistical test	uniform in {Entropy, Q-Statistics}
base learner	HTNB
drift detection	uniform in {DDM, EDDM, ADWIN}

XI. Runtime

To show the runtime of OSNN on specific hardware, we employed machines with Intel(R) Xeon(R) CPU E5-2690 v3 at 2.60GHz and 16Gb of RAM. In the Table XXXI below, we depict the speed at which this machine could process several data streams. It is important to highlight that this runtime includes not only the learning time, but also the time for the machine to read the stream, predict a new instance, obtain and save the output and calculate the accuracy measures.

TABLE XXXI: Speed (rate of instances per second) at which OSNN could process instances in our experiments run in an Intel(R) Xeon(R) CPU E5-2690 v3 at 2.60GHz and 16Gb of RAM.

Data stream	speed (instances processed per second)
Sine1	16.48
Sine2	16.64
Agrawal1	3.76
Agrawal2	3.74
SEA1	14.25
SEA2	14.33
STAGGER1	1.36
STAGGER2	2.20

Acknowledgments

This work was funded by EPSRC Grant No. EP/R006660/2.

References

[1] Larry Armijo. Minimization of functions having lipschitz continuous first partial derivatives. *Pacific J. Math.*, 16(1):1–3, 1966.

[2] James Bergstra and Yoshua Bengio. Random search for hyper-parameter optimization. *J. Mach. Learn. Res.*, 13:281—-305, 2012.

[3] Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

[4] Dariusz Brzezinski and Jerzy Stefanowski. Combining block-based and online methods in learning ensembles from concept drifting data streams. *Inf. Sci.*, 265:50–67, 2014.

[5] C. W. Chiu and L. L. Minku. A diversity framework for dealing with multiple types of concept drift based on clustering in the model space. *IEEE TNNLS*, pages 1–11, 2020.

[6] G. Ditzler, M. Roveri, C. Alippi, and R. Polikar. Learning in nonstationary environments: A survey. *IEEE Comput. Intell. Mag.*, 10(4):12–25, 2015.

[7] Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

[8] Vinicius M. A. Souza, Denis M. dos Reis, André G. Maletzke, and Gustavo E. A. P. A. Batista. Challenges in benchmarking stream learning algorithms with real-world data. *Data Mining and Knowledge Discovery*, 34(6):1805–1858, 2020.

[9] Rupesh K Srivastava, Jonathan Masci, Sohrob Kazerounian, Faustino Gomez, and Jürgen Schmidhuber. Compete to compute. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013.