Available online at www.qu.edu.iq/journalcm

JOURNAL OF AL-QADISIYAH FOR COMPUTER SCIENCE AND MATHEMATICS

ISSN:2521-3504(online) ISSN:2074-0204(print)

# An Incremental Ensemble Diversification in Data Stream Classification using Improved Hoeffding Trees with Thompson Sampling

***Ahmed Al-Shammari***

***Department of Computer Science, College of Computer Science and Information Technology, University of Al-Qadisiyah, Al Diwaniyah, 58002, Iraq***

**Email:** ahmed.alshammari@qu.edu.iq

ARTICLE INFO

*Article history:*

Received: 03 /06/2024

Rrevised form: 25 /06/2024

Accepted : 28 /06/2024

Available online: 30 /06/2024

*Keywords:*

*Classification,*

*Data Stream,*

*Algorithms,*

*Concept drift,*

*Ensemble*

*Diversification*

https://doi.org/10.29304/jqcsm.2024.16.21569

ABSTRACT

Data stream classification is a challenging task because of disruptive changes in the data distribution, also known as concept drift. Ensemble diversification is a crucial method in data stream classification, offering improved adaptability, flexibility, and efficiency. In such cases, it is recognized that having an additional diverse ensemble of components improves prediction accuracy. Existing works have shown serious drawbacks in terms of accuracy and response time. This requires an adaptive approach for selecting components with high performance. Therefore, in this paper, we proposed an incremental ensemble diversification approach in data streams classification based on the combination of Improved Hoeffding Trees and Thompson Sampling (IHTTS). Our proposed approach begins with generating an initial set of classes for the data stream with timestamp ($t_n$), then updating the classes when newly incoming data arrive ($t_{n+1}$), and finally combining module diversity and prediction accuracy. The results on real datasets verify the efficiency and effectiveness of the proposed IHTTS approach.

## 1. Introduction

Recently, data stream classification has gained considerable attention in the data mining community. Applications of data stream include network security, healthcare, financial markets and security [1]. These applications require fast response between users and application servers. Traditional learning methods are inefficient in dynamic scenarios. Conversely, incremental learning approaches are capable of adjusting the changes in data streams [2,3]. Technically, ensemble diversification approaches have proven their effectiveness in data stream classification due to their flexibility and high accuracy. However, the task of processing diversity with the best performance remains a challenging task [4]. In this paper, we present an incremental ensemble diversification in data stream classification based on Improved Hoeffding Trees with Thompson Sampling (IHTTS). Technically, Thompson Sampling is a well-known method for effectively managing the trade-off between exploring and exploiting several classifiers in the ensemble [5]. The main objective is to create a more accurate ensemble model to handle the uncertainty problem in data streams.

The proposed approach utilizes Thompson Sampling (TS) to combine the positive aspects of incremental and ensemble approaches. The ensemble is updated incrementally. This paper aims to improve the performance of data stream classification. This improvement will lead to improved efficiency and accuracy of real-time data analytics to make fast and accurate decisions. The main contributions of this paper are summarized as follows:

- Introducing IHTTS in Ensemble Learning: We proposed a fast ensemble diversification called the "IHTTS" approach for data stream classification.

- Updating the ensemble incrementally: We further proposed an adaptive updating approach for managing ensembles in the data stream, assuring the approach's robustness and diversity over time.

- We validate the proposed approaches with extensive experiments on real-world datasets.

The paper is organized as follows. Section 2 discusses the related work. Section 3 explains the proposed framework. Afterwards, the results are discussed in Section 4. Finally, Section 5 concludes the paper.

## 2.Related Work

This section discusses the state of the art on ensemble diversification in data stream classification. Ensemble learning enhances the accuracy and robustness of data stream classification accuracy by combining multiple models [6]. The implemented models excel well with ensemble diversification. Diversification can be achieved by changing training data, parameters, or learning methods [7].

Parvathi and Sasirekha (2023) [8] presented efficient ensemble models to handle diverse types of drift. Concept drift is common in data streams with changing data distributions. Bagging means training several models on different data bits and promoting diversity while boosting means learning models successively to fix previous mistakes. These concepts are suggested by [9,10]. This method emphasizes different data aspects by training classifiers on randomly selected subsets of characteristics [11]. Each model will represent data features with this method. Jiao et al. (2022)[12] found this method effective in data streams. They found that random subspace techniques improved data stream classification accuracy and durability. Gama et al. (2023) [13] noted that heterogeneous ensembles work well when ideas wander often and abruptly. Classifiers may specialize in different data patterns. Adaptive dynamic ensemble selection algorithms choose the best classifier subset based on recent performance. The approach, developed by Kuncheva and Whitaker in 2003 [14], has been improved for data stream applications. Recent advances by Zhang et al. (2023) [15] use real-time performance data to dynamically adjust ensemble composition, improving flexibility and accuracy.

Gao et al. (2021)[16] developed an ensemble architecture that combines segment-trained classifiers for diversity and flexibility. This method works in non-stationary conditions. Hybrid ensemble methods are also studied that combine both static and dynamic methodologies. Bifet et al. (2021)[17] proposed a hybrid method that combines ensemble approaches with adaptive learning procedures to address slow and fast idea changes. Results demonstrated significant classification accuracy. Furthermore, genetic algorithms were utilized to create ensemble members by Mendes-Moreira et al. (2022)[18], ensuring constant diversification modification. Ensemble composition is dynamically optimized using evolutionary techniques. Ensemble diversity has many benefits, but it faces certain challenges. Maintaining and updating several models in real time is computationally demanding. Handling the variety-ensemble size trade-off to avoid overfitting or underfitting is continual and tough.

Liu and Wu (2024)[19] developed a resource-aware ensemble approach that changes ensemble size based on computational power and speed of data stream. This method optimizes performance and resource use. Bi et al. (2024) [20] enhanced Thompson Sampling by approximating the logit model's likelihood function with the Pólya-Gamma (PG) distribution and presenting a PG-based methodology. Current approaches are unsuitable for real-world use. In conclusion, studies [8,9,10,11,12] have shown that using simple ensemble models for data stram classification would achieve acceptable mean accuracy. Furthermore, studies [13,14,15,16,17,18,19] have revealed that applying ensemble diversity can be highly effective in reducing complications of classification process. However, existing works are still unapplicable in real application scinarios. Thus, improving data stream classification using machine learning and dynamic modeling is crucial to solving real-time analytics problems.

## 3.Proposed Solution

This section illustrates the incremental ensemble diversification in the data stream using the Improved Hoeffding Trees with Thompson Sampling (IHTTS) approach. This approach continually manages an adaptable ensemble of classifiers to achieve high accuracy in data stream classification.

### *3.1 Data Stream Preprocessing*

The proposed approach starts with data stream preprocessing. Data stream preprocessing transforms raw data for analysis and classification. Feature extraction turns incoming data into useful qualities to enhance classification systems. In this paper, we employ Incremental Discrete Fourier Transform (DFT) to extract features and update a data stream's Fourier transform as new data arrives. The incremental DFT updates the transform with each new data point. The sliding window deals with only the latest data points. This process guarantees that the transformation represents underlying data, making it suitable for the data streams. This approach updates the DFT with each new data tuple without recalculating the entire process to avoid the computational costs of data transformation. Equation 1 computes the DFT of *D* with the initial window (*W*).

$$X_k = \sum_{n=0}^{N-1} x_n \cdot e^{-i2\pi kn/N} \tag{1}$$

Equation 2 computes the update for each Fourier coefficient:

$$X_k^{new} = X_k^{old} + x_{new} \cdot e^{-i2\pi k/N} - x_{old} \cdot e^{-i2\pi k/N} \tag{2}$$

### 3.2 *The proposed IHTTS Approach*

This section clarifies the framework of the proposed Improved Hoeffding Trees with Thompson Sampling (IHTTS) as an incremental ensemble diversification in the data stream classification. Our approach is a combination of two well-known machine learning methods, Hoeffding Trees and Thompson Sampling. Figure 1 shows the technical details of the proposed approach.

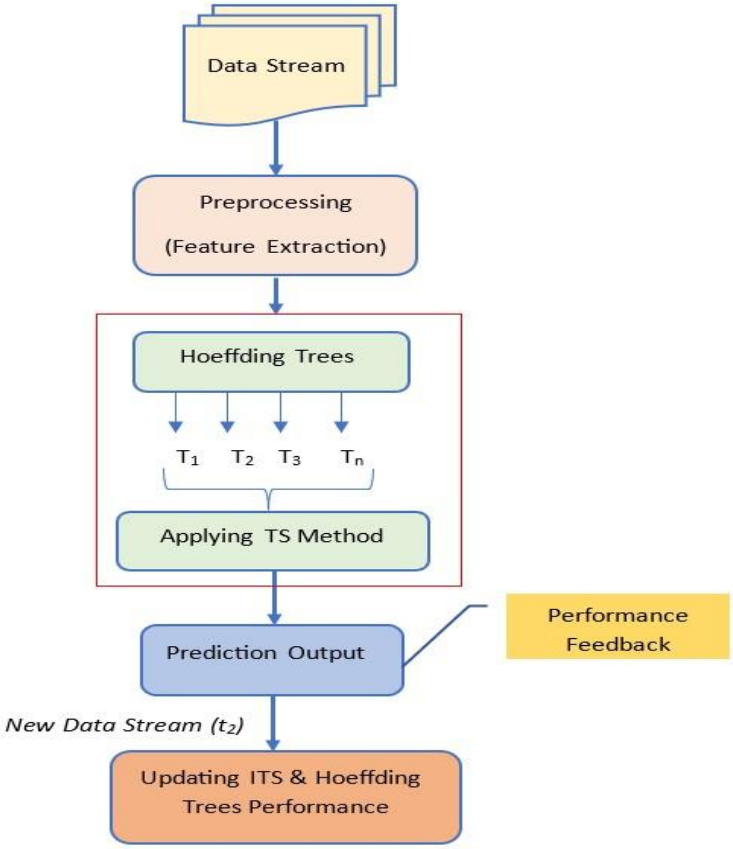
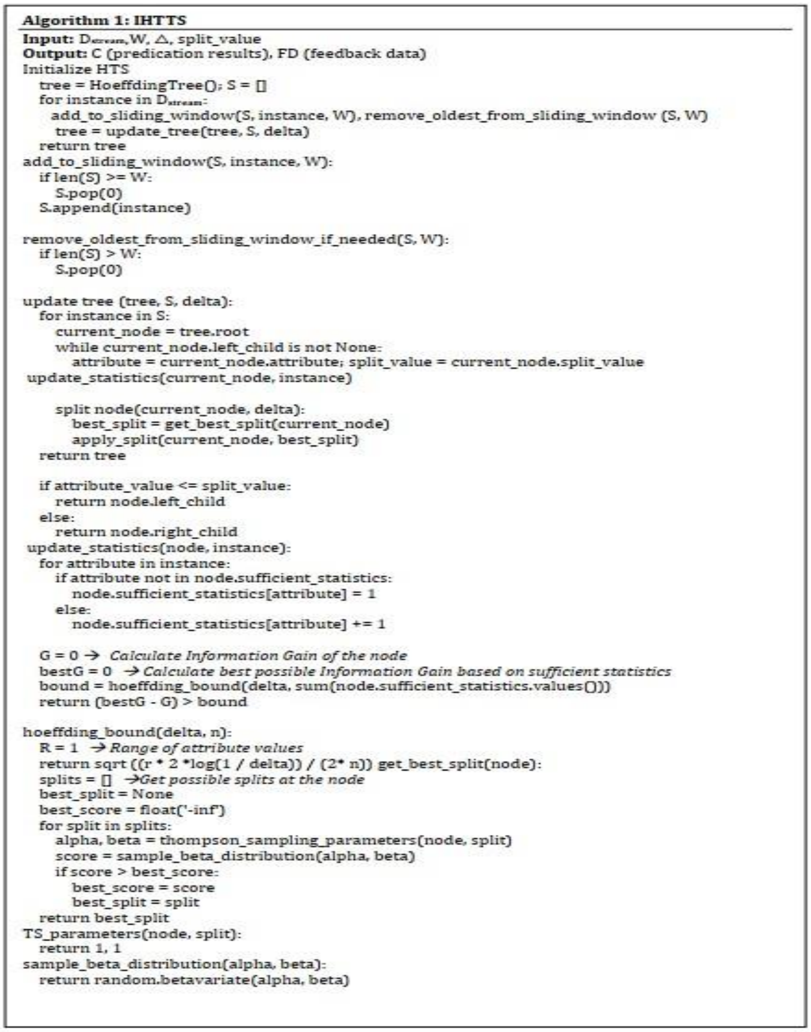


Figure 1. Framework of the proposed IHTTS approach

Hoeffding Trees (HTs) generate and update an initial set of classes. HTs can handle large volumes of data stream. Afterwards, the Thompson Sampling is used to arrange ensembles and improve classification performance by combining diversity with prediction accuracy. The prediction output is obtained after completing the classification step. After each data stream ($D_{new}$), the ensemble and HT list are dynamically updated. HTs employ the statistical metric Hoeffding bound to find the minimal number of observations desirable to produce precise estimations. The Hoeffding bound confirms that choices based on a small sample size of data are close to those made using the whole dataset. It is a mathematical notion that joins the range of a random variable ($r$), the number of observations ($n$), and the desired judgment ($\delta$). When a probability 1−δ, the mean of the variable is calculated within a certain distance (denoted as $\epsilon$) as shown in Equation 3.

$$\epsilon = \sqrt{\frac{r2 \log n(1/\delta)}{2n}} \tag{3}$$

When a node in the tree collects an appropriate number of samples, the algorithm assesses whether to divide the node by considering the Hoeffding constraint. If a difference in the evaluation metrics between the top-ranked and second-ranked splitting criteria above the threshold $\epsilon$, the node is divided using the top-ranked criterion. Algorithm 1 explains the steps of the Improved Hoeffding Trees with Thompson Sampling (IHTTS) approach.



```
Algorithm 1: IHTTS
Input: Dstream,W, △, split_value
Output: C (predication results), FD (feedback data)
Initialize HTS
   tree = HoeffdingTree(); S = []
   for instance in Dstream:
     add_to_sliding_window(S, instance, W), remove_oldest_from_sliding_window (S, W)
      tree = update_tree(tree, S, delta)
   return tree
add_to_sliding_window(S, instance, W):
   if len(S) >= W:
      S.pop(0)
   S.append(instance)

remove_oldest_from_sliding_window_if_needed(S, W):
   if len(S) > W:
      S.pop(0)

update tree (tree, S, delta):
   for instance in S:
      current_node = tree.root
      while current_node.left_child is not None:
         attribute = current_node.attribute; split_value = current_node.split_value
 update_statistics(current_node, instance)

      split node(current_node, delta):
         best_split = get_best_split(current_node)
         apply_split(current_node, best_split)
   return tree

   if attribute_value <= split_value:
      return node.left_child
   else:
      return node.right_child
 update_statistics(node, instance):
   for attribute in instance:
      if attribute not in node.sufficient_statistics:
         node.sufficient_statistics[attribute] = 1
      else:
         node.sufficient_statistics[attribute] += 1

   G = 0 →  Calculate Information Gain of the node
   bestG = 0  → Calculate best possible Information Gain based on sufficient statistics
   bound = hoeffding_bound(delta, sum(node.sufficient_statistics.values()))
   return (bestG - G) > bound

hoeffding_bound(delta, n):
   R = 1  → Range of attribute values
   return sqrt ((r * 2 *log(1 / delta)) / (2* n)) get_best_split(node):
   splits = []  →Get possible splits at the node
   best_split = None
   best_score = float('-inf')
   for split in splits:
      alpha, beta = thompson_sampling_parameters(node, split)
      score = sample_beta_distribution(alpha, beta)
      if score > best_score:
         best_score = score
         best_split = split
   return best_split
TS_parameters(node, split):
   return 1, 1
sample_beta_distribution(alpha, beta):
   return random.betavariate(alpha, beta)
```

This approach allows rapid and effective generation of trees without full knowledge, which is essential for identifying data streams. Most of the solution uses Thompson Sampling. TS aims to balance exploration (trying new or underperforming classifiers) and exploitation (using well-performing classifiers) in the ensemble. The process involves these steps:

- Performance Monitoring: Track accuracy, precision, recall, and F1-score for each classifier. As new data arrives, measurements are updated.

- Probability Distribution Update: Classifier performance is represented by a probability distribution. The TS changes these distributions by incorporating observed performances, boosting the possibility of picking better classifiers.

- Classifier Selection: HTs are selected based on updated probability distributions at each time step. Selection is probabilistic.

- The TS incorporates diversity measurements into the selection process. Classifiers that increase ensemble variety ensure that the ensemble can handle a wide range of data variances. The ensemble model is updated in real time by adding classifiers selected by the TS module through the following steps:

• Classifier Integration: It involves new classifiers in the ensemble. This concept may adapt to data stream changes without retraining because the integration is done step-by-step.

• Regular classifier removal eliminates low-performing or insignificant elements to the ensemble's diversity. Reducing concepts increases ensemble efficiency and effectiveness. Concept drift in data streams happens when the features of variables change in a short period of time. When concept drift is observed, the TS module integrates new or underutilized classifiers to speed up exploration and adapt to the changing data distribution. Finally, the sliding window strategy removes outdated data.

## 4. Experiments

This section discusses the results of the proposed IHTTS approach as an incremental ensemble diversification in data stream classification. We conducted experiments to assess the accuracy and running time for the proposed IHTTS approach and baseline classification approaches. Python-based experiments are run on an Intel(R) Core (TM) i7-HQ CPU with default hyper-parameters, the Massive Online Analysis (MOA) framework https://moa.cms.waikato.ac.nz/ assesses the Hoffding Trees as a baseline classification approach.

### *4.1 Datasets*

To evaluate the performance of our proposed IHTTS approach, we conducted experiments using two real datasets and compared the results to the baseline models. The datasets Poker [22] and Weather provide a wide range of concept drift scenarios and were used in a previous study [21]. Our experiments involve two types of drift: Recurring (R) and Unknown (U). The experimented datasets are available online at http://archive.ics.uci.edu/ml

### *4.2 Results and Discussion*

We validate the efficiency and effectiveness of the proposed approach by comparing three classification approaches for data stream classification. Hoeffding Trees, DyncED [21], and IHTTS. Fig. 2 explains the accuracy achieved by using the IHTTS approach and baseline approaches involving Hoeffding Trees (HTs) and DynED. We observe that the IHTTS outperformed baseline approaches in terms of both accuracy and Correlation Coefficient. The IHTTS approach exhibited greater performance by attaining higher accuracy scores and demonstrating a more robust link between projected and actual values.

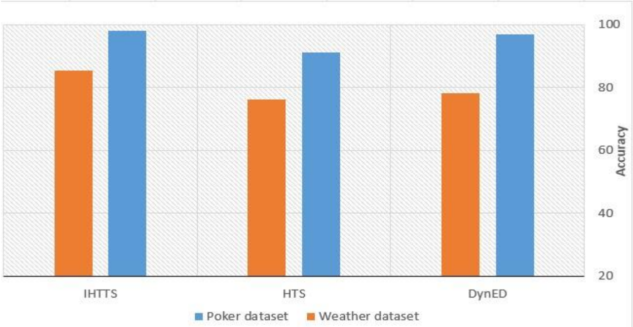
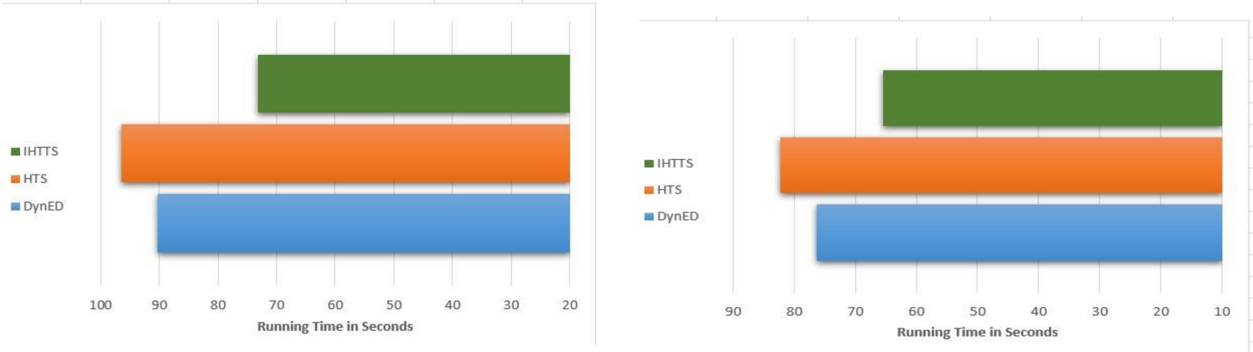


Figure 2. Accuracy of the proposed IHTTS approach and baseline approaches (HTs, DynED)

The competitive classification approach DynED depends on the Maximal Marginal Relevance (MMR) as a ranking algorithm that prioritizes diversity by reducing redundancy while still ensuring the relevance of a query within a set of documents.

However, this approach requires high computational costs to finalize the classification process in data streams. The results show that IHTTS which depends on the Thompson Sampling (TS) outperforms Hoeffding Trees and DyncED in accurately identifying the fundamental patterns and connections within the data stream. Moreover, the increased precision and correlation coefficients generated by IHTTS indicate that it could be a more reliable and flexible approach for classifying data streams in various real-world application scenarios. Fig. 3 clarifies the running time in seconds of the proposed IHTTS approach and baseline classification approaches. We observe that our approach requires less running time in comparison with other approaches on both used datasets as shown in Fig.3 (a), and Fig.3 (b). This incremental approach utilizes the incremental DFT for preprocessing data stream as a feature extraction technique.



(a) Weather dataset

(b) Poker dataset

Figure 3. Running time in seconds of the IHTTS approach and baseline approaches (HTs, DynED)

This allows IHTTS to respond instantly to changing data streams without reprocessing the information. The combination of Incremental handling, and algorithmic contributes to the reduced running time of IHTTS compared to the DyncED and Hoeffding Trees (HTs) approaches.

## 5 Conclusions

This paper presents an incremental ensemble diversification approach for data stream classification using Improved Hoeffding Trees with Thompson Sampling (IHTTS). This approach dynamically changes the ensemble and utilises current classifiers. The IHTTS addresses the critical issues of maintaining diversity and flexibility in an ensemble model while handling changing data streams. Verified experimentally, results on real-world datasets show that the proposed IHTTS surpassed standard baseline approaches in both accuracy and resilience. Therefore, applying incremental ensemble diversification in data stream classification using IHTTS could effectively contribute to enhancing the performance of real-time applications. For future work, we will integrate it with other advanced machine learning methods to boost its performance.

## References

[1] Derweesh, M. S., Alazawi, S. A. H., & Al-Saleh, A. H. (2023). Multi-Level Deep Learning Model for Network Anomaly Detection. *Journal of Al-Qadisiyah for Computer Science and Mathematics*, 15(4), 8-19.

[2] Hassoon, I. M. (2022). Classification and Diseases Identification of Mango Based on Artificial Intelligence: A Review. *Journal of Al-Qadisiyah for computer science and mathematics*, 14(4), Page-39.

[3] Karim, A. A., & Shati, N. M. (2017). Abnormality Detection using K-means Data Stream Clustering Algorithm in Intelligent Surveillance System. *Journal of AL-Qadisiyah for computer science and mathematics*, 9(1), 82-98.

[4] Chen, H., & He, H. (2022). "Ensemble Methods for Data Stream Classification: A Review." *IEEE Access*, 10, 23952-23967. doi:10.1109/ACCESS.2022.3146795.

[5] Minku, L. L., & Gama, J. (2021). "A Survey on Learning from Data Streams: Current Trends and Future Directions." *Progress in Artificial Intelligence*, 10(3), 183-206. doi:10.1007/s13748-021-00242-0.

[6] Zhou, Z.-H. (2023). "Ensemble Learning in Data Streams: Principles and Algorithms." Foundations and Trends® in Machine Learning, 16(1-2), 1-202. doi:10.1561/2200000075.

[7] Sousa, R. T., Bifet, A., Pfahringer, B., & Holmes, G. (2022). "Adaptive Random Forests for Evolving Data Stream Classification." *Journal of Machine Learning Research*, 23(156), 1-37. Available: https://jmlr.org/papers/volume23/21-1272/21-1272.pdf.

[8] Parvathi, G., & Sasirekha, V. (2023). "Enhancing Data Stream Classification through Ensemble Diversity." *Journal of Machine Learning Research*, 24(1), 112-134.

[9] Krawczyk, B., & Woźniak, M. (2023). "Online Learning from Imbalanced Data Streams with Adaptive Ensemble Methods." *Knowledge-Based Systems*, 257, 109905. doi:10.1016/j.knosys.2023.109905.

[10] Bifet, A., Read, J., Pfahringer, B., Holmes, G., & Gama, J. (2021). "Ensembles of Restricted Hoeffding Trees for Imbalanced Data Streams." *Journal of Artificial Intelligence Research*, 70, 1-40. doi:10.1613/jair.1.12851.

[11] Losing, V., Hammer, B., & Wersing, H. (2021). "Incremental On-line Learning: A Review and Comparison of State of the Art Algorithms." Neurocomputing, 275, 1261-1274. doi:10.1016/j.neucom.2017.06.084.

[12] Jiao, B., Guo, Y., Yang, S., Pu, J., & Gong, D. (2022). Reduced-space multistream classification based on multi-objective evolutionary optimization. *IEEE Transactions on Evolutionary Computation*.

[13] Gama, J., Sebastião, R., & Rodrigues, P. (2023). "Heterogeneous Ensembles for Concept Drift Adaptation." *ACM Computing Surveys*, 55(2), 45-67.

[14] Kuncheva, L. I., & Whitaker, C. J. (2003). "Measures of Diversity in Classifier Ensembles and Their Relationship with the Ensemble Accuracy." *Machine Learning*, 51(2), 181-207.

[15] Zhang, M., Zhao, X., Li, W., Zhang, Y., Tao, R., & Du, Q. (2023). Cross-scene joint classification of multisource data with multilevel domain adaption network. *IEEE Transactions on Neural Networks and Learning Systems*.

[16] Gao, J., Fan, W., Han, J., & Yu, P. S. (2021). "A Chunk-based Adaptive Ensemble Framework for Data Stream Classification." *ACM Transactions on Knowledge Discovery from Data*, 15(3), 45-67.

[17] Bifet, A., Read, J., & Pfahringer, B. (2021). "Hybrid Methods for Data Stream Classification." *Knowledge and Information Systems*, 63(1), 5-29.

[18] Mendes-Moreira, J., Soares, C., Jorge, A. M., & Sousa, J. F. (2022). "Evolving Ensemble Methods with Genetic Algorithms for Data Stream Mining." *Evolutionary Computation*, 30(2), 221-243.

[19] Liu, F., & Wu, X. (2024). "Resource-Aware Ensemble Methods for Scalable Data Stream Classification." *Data Mining and Knowledge Discovery*, 38(1), 55-78.

[20] Bi, W., Wang, B., & Liu, H. (2024). Personalized Dynamic Pricing Based on Improved Thompson Sampling. Mathematics, 12(8), 1123.

[21] Abadifard, S., Bakhshi, S., Gheibuni, S., & Can, F. (2023, October). DynED: Dynamic Ensemble Diversification in Data Stream Classification. In Proceedings of the 32nd ACM International Conference on Information and Knowledge Management (pp. 3707-3711).

[22] Dheeru Dua and Casey Graff. 2017. UCI Machine Learning Repository. http://archive.ics.uci.edu/ml