An Adaptive Random Forest for Data Stream Sentiment Classification under Concept Drift

Brian Farrel Arkana[1, *], Sudianto Sudianto[2], Nenen Isnaeni[3]

[1,2,3]*Informatics Engineering Study Program, Telkom University, Purwokerto, Indonesia*

(Received: September 20, 2025; Revised: November 18, 2025; Accepted: February 18, 2026; Available online: March 17, 2026)

Abstract

Data labeling plays a crucial role in determining the performance of machine learning models, especially in data stream environments where concept drift frequently occurs. The primary objective of this study is to analyze the effectiveness of adaptive learning models in managing dynamic data distribution changes and to evaluate the influence of different labeling strategies on sentiment classification performance using user reviews from the OVO mobile application. The research contributes to understanding how labeling approaches interact with adaptive modeling under real-time data stream conditions. Two labeling methods were employed: score-based labeling derived from user ratings and content-based labeling generated automatically using the IndoRoBERTa language model. These labeled data streams were evaluated using two classifiers: a conventional Random Forest model and an Adaptive Random Forest model designed to handle evolving data distributions. The evaluation was conducted through streaming experiments that continuously fed new review data to simulate real-world drift scenarios. The results reveal that in the score-based labeling scenario, the conventional Random Forest model's accuracy gradually declined, reaching a final accuracy of 31%, while the Adaptive Random Forest achieved 80%, reflecting a 49% performance gap. In the content-based labeling scenario, both models improved over time, with final accuracies of 57% for Random Forest and 76% for the adaptive model, resulting in a 19% difference. These findings indicate that Adaptive Random Forest is more robust in adapting to distributional and temporal changes in data streams regardless of the labeling strategy used. This study implies that combining adaptive learning with semantically rich labeling approaches can substantially enhance model reliability in real-time sentiment analysis tasks. Future research may further explore hybrid adaptive mechanisms to improve the resilience of data stream classification models across various domains.

Keywords: Adaptive Random Forest, Concept Drift, Data Stream, Labeling Strategy, Sentiment Analysis

1. Introduction

Machine learning plays a crucial role in analyzing continuously generated data, where challenges arise due to concept drift—changes in the statistical relationship between input features and output labels that cause model degradation over time [1]. Traditional models such as Random Forest struggle in these environments because they require full retraining to remain effective, making them unsuitable for real-time scenarios [2]. This retraining process introduces significant latency, as the model must be rebuilt each time new data are received, and it demands substantial computational and storage resources, which limit scalability in continuous data stream applications. The complexity increases when dealing with heterogeneous, large-scale textual data collected through methods like web scraping, which demand robust preprocessing and adaptive feature extraction approaches such as TF-IDF with Indonesian stopword filtering to maintain contextual relevance in evolving data streams [3], [4], [5], [6].

To address these issues, recent research emphasizes adaptive and ensemble-based learning methods capable of responding dynamically to distributional shifts. Advances in knowledge discovery have introduced frameworks for real-time drift detection and adaptation [7], alongside insights into handling recurring concepts efficiently [8]. Emerging deep learning approaches further enhance model responsiveness by enabling architectures to adjust automatically to shifting data distributions [9], while systematic reviews confirm that accurate drift detection and adaptation are essential for maintaining performance stability in non-stationary environments [10]. In addition to algorithmic adaptability, labeling strategies significantly influence data stream classification performance, particularly

***Corresponding author: Brian (brianfarrel@student.telkomuniversity.ac.id)**
DOI: https://doi.org/10.47738/jads.v5i2.1153
This is an open access article under the CC-BY license (https://creativecommons.org/licenses/by/4.0/).
© Authors retain all copyrights

in sentiment analysis where star ratings may oversimplify nuanced linguistic expression. This study examines these challenges using OVO e-wallet reviews from the Google Play Store, comparing traditional star-based labeling with IndoRoBERTa-based model labeling. The OVO platform was selected due to its extensive review volume (over one million entries) and moderately balanced average rating of 3.6, and the ease of systematic data collection through the google-play-scraper library, which minimizes polarity bias while providing a rich, continuously updated data stream representative of real-world user sentiment. By evaluating Random Forest and Adaptive Random Forest under simulated concept drift, the study demonstrates that ARF offers superior adaptability, and that labeling choices substantially affect long-term model performance. The findings highlight the critical interplay between adaptive algorithms and labeling strategies in real-world streaming data contexts.

2. Literature Review

Adaptive machine learning has become essential in non-stationary environments where data distributions evolve continuously, such as IoT ecosystems, financial technologies, cybersecurity systems, and social media platforms. Prior work establishes that model usability depends on robustness to concept drift, scalability, and the extent to which systems minimize manual intervention during deployment and maintenance [1], [11], [12]. Without automation and adaptability, traditional learning pipelines become impractical as they require continual retraining in response to shifting data conditions. A substantial portion of recent research therefore focuses on automated drift detection and adaptation strategies. Methods such as Adaptive Deep Forest [13], dynamic-detector Random Forests [14], and online boosting across multistream environments [12] demonstrate how classifiers can autonomously replace outdated components and adjust to new distributions. Complementary work in meta-learning further enhances automation by selecting optimal adaptation strategies with minimal human oversight [15], while automated repair mechanisms improve model stability during drift events [16].

Efficiency and lightweight computation also play a critical role in adaptive system design. Studies targeting IoT and mobile environments highlight the need for resource-aware drift detection frameworks [17], scalable neural-forest hybrids [18], and unsupervised approaches that reduce labeling cost such as robust random cut forests [19]. Fog–cloud collaborative architectures further illustrate how distributed adaptive learning can sustain low-latency performance [20], while Adaptive Random Forest has been validated as both scalable and stable for large-scale data streams [21].

Domain-specific research reinforces these principles across cybersecurity, distributed systems, and sentiment analysis. Evolving malware and intrusion detection contexts underscore the superiority of adaptive classifiers in reducing maintenance overhead [22], [23], while federated drift-aware frameworks enable adaptation across decentralized environments without heavy coordination [11]. In sentiment analysis and financial applications, adaptive ensembles such as SOKNL, Forgetful Forests, and optimized ARF variants show that automated drift handling supports accuracy and usability in evolving opinion streams [24], [25], [26], [27], [28]. Ensemble-based methods also manage linguistic variability and class imbalance in social media environments, improving resilience to recurrent concepts [29], [30], [31], [32]. Despite progress, several studies identify persistent gaps. Many adaptive systems still prioritize accuracy over interpretability, deployment simplicity, or real-world generalizability [29], [33], [34]. Evaluations often rely on constrained or synthetic datasets, raising concerns about scalability to complex, high-dimensional, or multilingual data [35], [36], [37]. These limitations highlight the need for approaches that integrate automation, transparency, and computational efficiency to make adaptive learning frameworks both accurate and operationally feasible in real-world streaming scenarios.

3. Methodology

Figure 1 presents the research methodology employed in this study, outlining the sequential stages from literature review to model evaluation. The process includes literature review, data collection, preprocessing data, implementation of Adaptive Random Forest, and model evaluation. Meanwhile, figure 2 provides a detailed view of the data collection process. This structured workflow ensures a systematic approach in addressing sentiment classification under concept drift conditions.

 ISSN 2723-6471

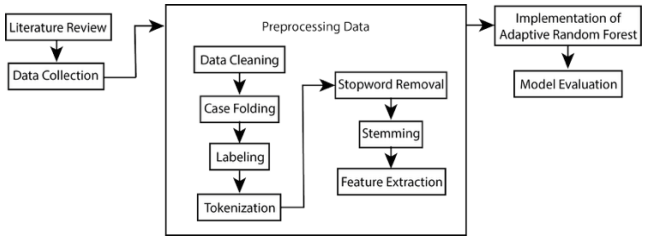

Figure 1. Research Flow

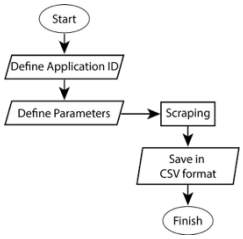

Figure 2. Data Collection Process

3.1.Literature Review

As summarized in Chapter 2, existing studies indicate the need for adaptive models in dynamic sentiment streams but reveal a gap in understanding how different labeling strategies influence their performance.

3.2. Data Collection

The dataset was collected through web scraping of user reviews of the OVO application from the Google Play Store using the google_play_scraper Python library. The reviews were stored in .csv format for further processing and include textual reviews, star ratings, and metadata such as review date and application version.

3.3. Data Preprocessing

To ensure data quality and consistency, a series of preprocessing steps were performed. First, text cleaning was applied by removing non-alphabetic characters, extra spaces, and unnecessary symbols. This was followed by case folding, where all characters were converted to lowercase to maintain uniformity. Next, two types of labeling were implemented: rating-based labeling, in which reviews with 1–2 stars were categorized as negative, 3 stars as neutral, and 4–5 stars as positive; and content-based labeling, which utilized a fully automated approach based on the pre-trained IndoRoBERTa language model. After labeling, tokenization was conducted to split sentences into individual tokens. Subsequently, stopword removal was performed to eliminate common words such as “di” (in/at), “yang” (which/that), and “dari” (from) that carry minimal semantic meaning. Stemming was then applied to reduce words to their base forms, for example, “berguna” (useful) to “guna” (use). Finally, feature extraction was carried out using Term Frequency–Inverse Document Frequency (TF-IDF) to convert textual data into numerical feature vectors suitable for machine learning models.

The final preprocessed dataset consisted of structured, labeled tokens with TF-IDF values, ready for model training and evaluation. No class balancing technique (e.g., undersampling or SMOTE) was applied after labeling. This decision was intentional to preserve the natural class distribution of the data stream, reflecting real-world conditions where sentiment classes are inherently imbalanced and continuously evolving.

3.4. Implementation of Adaptive Random Forest (ARF)

Since the dataset originates from historical reviews (November 2016 – May 2025), a data stream simulation was required. Google Colab was employed to replicate real-time data streaming, enabling sequential processing of reviews. The Adaptive Random Forest model was then implemented to classify sentiment while handling potential concept drift as illustrated in figure 3. ARF maintains an ensemble of decision trees and leverages the Adaptive Windowing (ADWIN) mechanism to detect changes in data distribution. When drift is detected, underperforming trees are replaced with new ones, ensuring adaptability.

 ISSN 2723-6471

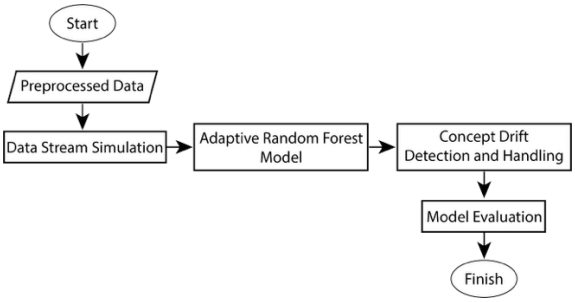

Figure 3. Adaptive Random Forest Implementation Flow

3.5. Model Evaluation

Model performance was evaluated using the prequential evaluation method, where each incoming instance is tested before being used for training, thus simulating a realistic streaming scenario. This approach was chosen over holdout or sliding window validation because it continuously measures model performance under evolving data distributions, allowing a more accurate assessment of real-time adaptability and concept drift handling. The evaluation employed standard metrics including Accuracy, Precision, Recall, and F1-Score. Parameter tuning for both Random Forest and Adaptive Random Forest was conducted through preliminary manual testing to balance performance stability and computational efficiency. To ensure a fair and unbiased comparison, both models were configured with the same number of trees (*n_estimators* = 10), and most other parameters were kept at their default values. This design allows the evaluation to reflect each model's inherent learning capability rather than the effects of extensive hyperparameter optimization. The Random Forest used the default "sqrt" feature selection strategy (*max_features*='sqrt') to optimize training speed without significantly compromising accuracy, while the Adaptive Random Forest relied on its default adaptive mechanisms to simulate realistic, resource-constrained streaming conditions. This configuration provided consistent performance while maintaining computational feasibility on limited hardware. The assessment also examined how the model adapts to concept drift and maintains stability under both labeling strategies, ensuring sustained performance in dynamic sentiment streams.

4. Results and Discussion

4.1. Data Collection

The dataset used in this study was obtained through web scraping of user reviews of the OVO application from the Google Play Store using the google_play_scraper library in Python. The collected attributes include review id, username, review content, rating score, timestamp, and application version. Initial inspection revealed missing values across several attributes, with the majority occurring in non-essential fields such as username and appVersion. Table 1 presents the distribution of missing values across all attributes. Despite these missing values, the content and score attributes—which are central to sentiment analysis and labeling—remained sufficiently populated to allow for further preprocessing and analysis.

Table 1. Breakdown of Missing Values per Attribute

Column Name	Total Missing Value
reviewID	0
username	99023
content	99176
score	99227
timestamp	99262
appVersion	99269

ISSN 2723-6471

4.2. Data Preprocessing

Before conducting the analysis, incomplete records containing missing values were removed. The preprocessing stage was then carried out to prepare the textual data for modeling. This process consisted of several steps, namely data cleaning to remove HTML tags, symbols, numbers, and redundant spaces; case folding to normalize all text into lowercase; tokenization to segment reviews into individual words; stopword removal to eliminate common but less informative words; and feature extraction using TF-IDF to quantify the importance of each term within the corpus.

A crucial step in this study was the labeling process, which was conducted using two different approaches. The first approach assigned sentiment labels based on the rating score provided by users, where scores 1–2 were categorized as negative, 3 as neutral, and 4–5 as positive. The second approach utilized a semantic-based method by applying the IndoRoBERTa model to classify sentiment directly from the content of reviews. This dual-labeling design allowed for a comparative analysis of sentiment distribution and improved the robustness of concept drift detection.

Figures 4 and 5 illustrate the sentiment distribution and the ten most frequent words for each sentiment category under both labeling approaches. The comparison shows that score-based labeling tends to produce a dataset dominated by positive reviews, while content-based labeling with IndoRoBERTa yields a more balanced distribution across positive, neutral, and negative classes, as summarized in table 2.

Table 2. Sentiment Distribution for Both Labeling Methods

Labeling Method	Negative	Neutral	Positive	Total Reviews
Score-based labeling	71631	9655	149834	231120
Content-based labeling	76491	23202	131427	231120

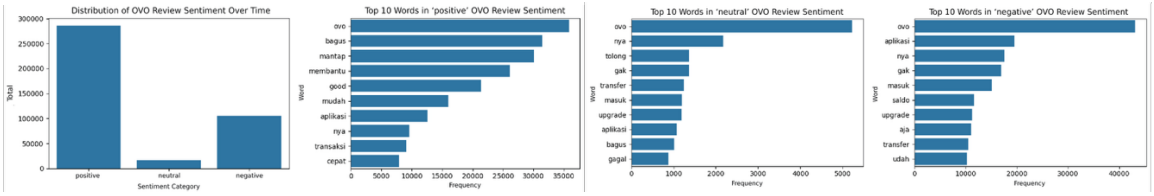

Figure 4. Sentiment Distribution and Top 10 Words (Score-Based Labeling)

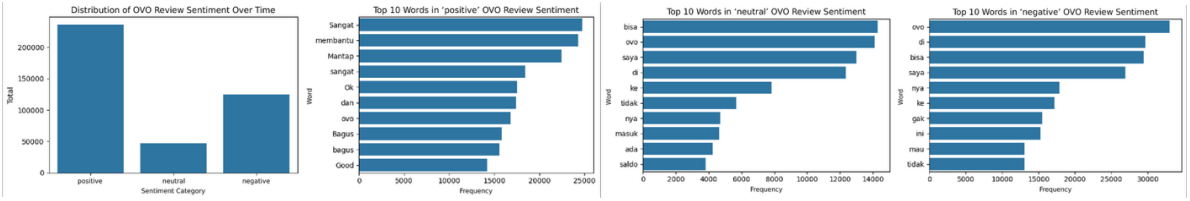

Figure 5. Sentiment Distribution and Top 10 Words (Content-Based Labeling with IndoRoBERTa)

To further capture temporal dynamics, the proportions of sentiment over time were visualized for both approaches, as shown in figures 6 and 7. These visualizations reveal subtle differences in sentiment trends depending on the labeling method, which is essential for subsequent drift detection analysis.

ISSN 2723-6471

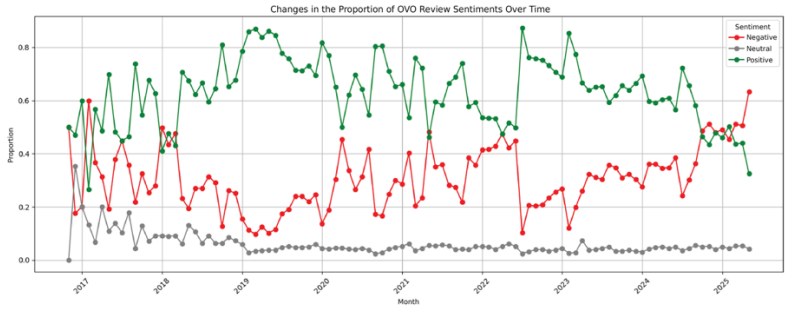

Figure 6. Temporal Changes in Sentiment Proportions (Score-Based Labeling)

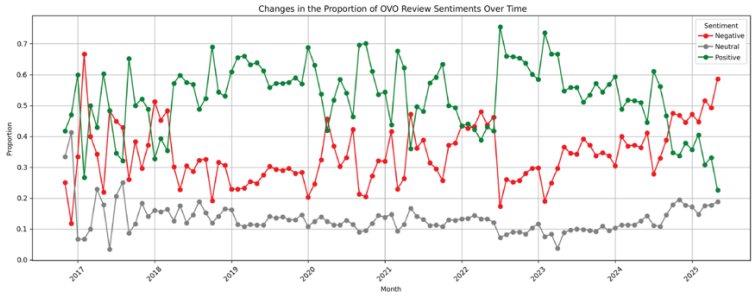

Figure 7. Temporal Changes In Sentiment Proportions (Content-Based Labeling)

4.3. Concept Drift Detection

To operationalize the detection, a threshold of 10% change in sentiment distribution was established. This fixed threshold was chosen as a model-agnostic and data-level heuristic to detect observable distributional shifts without relying on classifier-specific mechanisms. Established drift detection methods such as DDM, EDDM, and ADWIN primarily monitor error rate fluctuations in adaptive or incremental models and therefore cannot be directly applied to static learners like the conventional Random Forest used in this study. The 10% threshold was thus adopted to provide a consistent and comparable measure of data drift across both the adaptive and non-adaptive models. Any deviation exceeding this threshold was considered an indication of concept drift. Examples of detected drifts are summarized in tables 3 and 4, which report sentiment changes greater than 10% across different periods. The results reveal that both labeling approaches consistently detected instances of drift, though the magnitude and direction of changes varied. Score-based labeling tended to capture sharper fluctuations in negative and positive sentiments, whereas IndoRoBERTa-based labeling highlighted a more balanced interplay between negative, neutral, and positive shifts.

Table 3. Sentiment Changes >10% Indicating Concept Drift (Score-Based Labeling)

Sentiment	Period	Negative	Neutral	Positive	Information
0	2016-12	-32.4%	+35.3%	-2.9%	Negative decline
1	2017-01	+2.4%	-15.3%	+12.9%	Positivity increased sharply
2	2017-02	+40.0%	-6.7%	-33.3%	Negative spikes drastically, positive declines
3	2017-03	-23.3%	-6.7%	+30.0%	Negative decline, positive increase sharply
4	2017-04	-5.2%	+13.3%	-8.1%	-

Table 4. Sentiment Changes >10% Indicating Concept Drift (Content-Based Labeling)

Sentiment	Period	Negative	Neutral	Positive	Information
0	2016-12	-13.2%	+7.8%	+5.4%	Negative decline
1	2017-01	+21.6%	-34.5%	+12.9%	Negatives spike drastically, positives rise sharply
2	2017-02	+33.3%	+0.0%	-33.3%	Negative spikes drastically, positive declines
3	2017-03	-26.7%	+3.3%	+23.3%	Negative decline, positive increase sharply

4	2017-04	-5.7%	+12.9%	-7.1%	-

While this approach effectively identifies observable distributional changes, it assumes that variations in label proportions directly reflect underlying drift in the data. This assumption is more reliable under score-based labeling, where star ratings provide explicit ground truth, but may be less stable for IndoRoBERTa-based labeling, where inferred sentiments depend on model predictions that can introduce additional variability.

Subsequent visualizations integrated sentiment proportion trends with drift events. Figures 8 and 9 show the temporal evolution of sentiment proportions alongside detected drift for both labeling methods, revealing frequent shifts between 2017–2019 and 2020–2022. For the modeling experiments, the latter period (2020–2022) was selected because it contained a substantially larger volume of reviews (231,120 entries compared to only 94,383 from 2017–2019), providing a richer and more stable basis for evaluating model adaptability. Moreover, this period reflects more mature user behavior and system stability after OVO's initial launch phase, making drift patterns more representative of real-world sentiment evolution. This selection prioritizes recency and data sufficiency rather than arbitrary filtering, minimizing potential bias in drift interpretation.

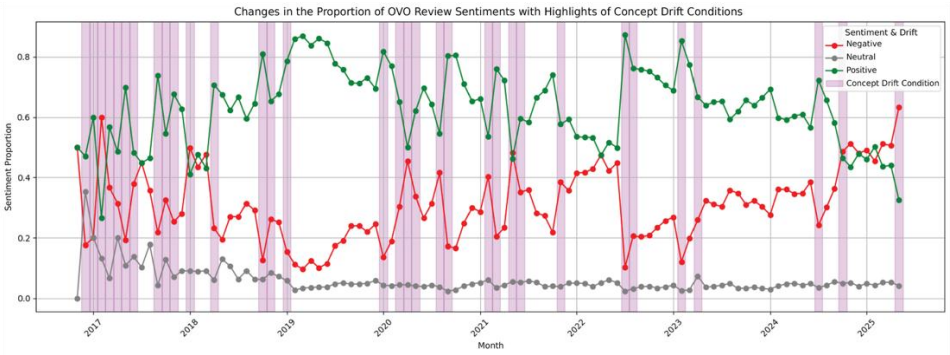

Figure 8. Sentiment Proportion Trends with Highlighted Drift Events (Score-Based Labeling)

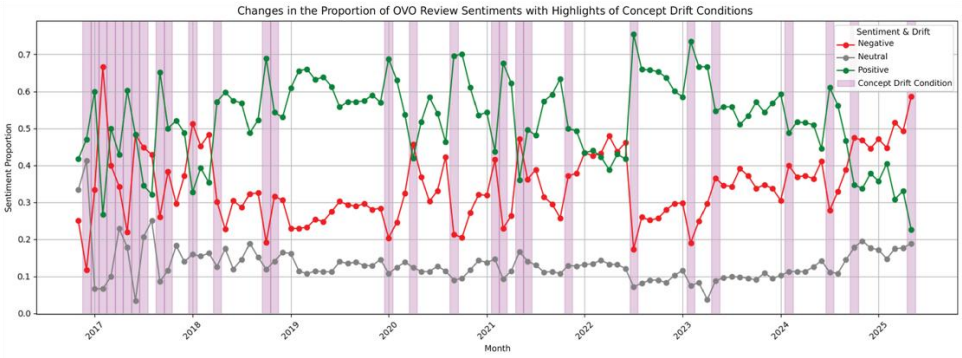

Figure 9. Sentiment Proportion Trends with Highlighted Drift Events (Content-Based Labeling)

4.4. Effectiveness of Random Forest and Adaptive Random Forest in Handling Concept Drift

To assess model performance under concept drift, OVO review data from 2020–2022 were streamed over 240 minutes, averaging 963 instances per minute for Adaptive Random Forest. Random Forest was trained on the first 10 instances and tested on equal-sized batches to ensure comparable conditions. Both models used aligned hyperparameters—10 trees and the maximum number of features set to sqrt—and were evaluated on two labeling schemes: score-based labeling using star ratings and content-based labeling using IndoRoBERTa classifications.

4.4.1. Random Forest and Adaptive Random Forest with Score-Based Labeling

The first evaluation employed score-based labeling, with results presented in tables 5–6 and figures 10–11. The Random Forest model achieved an overall accuracy of only 31% (table 5), with a strong bias toward predicting negative sentiment. The confusion matrix (figure 10) indicates that the model frequently misclassified neutral and positive sentiments, highlighting its inability to generalize effectively in the presence of concept drift.

Table 5. Classification Report of Random Forest (Score-Based Labeling)

	precision	recall	f1-score	support
negative	0.31	1.00	0.47	71631
neutral	0.20	0.00	0.00	9655
positive	0.33	0.00	0.00	149834
accuracy			0.31	231120
macro avg	0.28	0.33	0.16	231120
weighted avg	0.32	0.31	0.15	231120

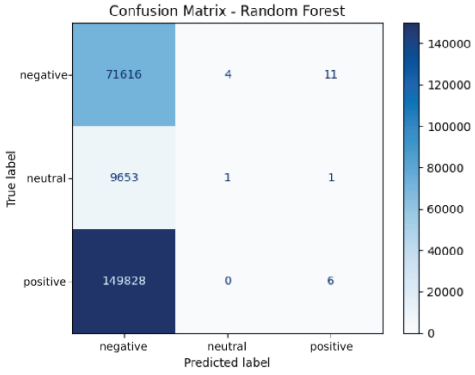

Figure 10. Confusion Matrix of Random Forest (Score-Based Labeling)

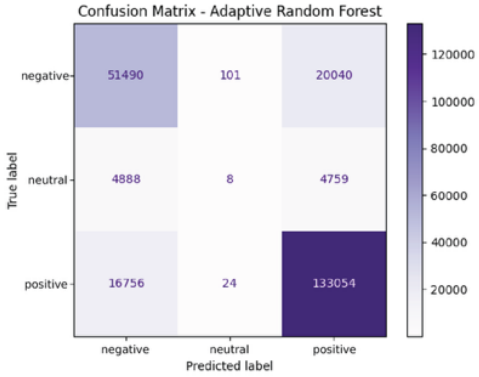

Figure 11. Confusion Matrix of Adaptive Random Forest (Score-Based Labeling)

The extreme misclassification observed in figure 10, where Random Forest predominantly predicts the negative class, indicates an early model collapse caused by limited and imbalanced initial training data. The model was initialized with only ten samples—six negative, two neutral, and two positive reviews—resulting in a strong bias toward the dominant class. Because Random Forest operates as a static learner without incremental updates, this initial imbalance persisted throughout the data stream, preventing the model from adapting to subsequent distributional changes. The phenomenon highlights how small and biased initialization windows can severely impair non-adaptive models in streaming environments, particularly under dynamic sentiment shifts.

In contrast, the Adaptive Random Forest achieved significantly higher performance, with an overall accuracy of 80% (table 6). The confusion matrix (figure 11) shows that the model successfully identified both negative and positive sentiments with balanced precision and recall. Accuracy trends in figure 12 (daily accuracy) and figure 13 (cumulative accuracy) further confirm that the Adaptive Random Forest adapts well to evolving data, as its accuracy improves steadily over time. In this context, "daily accuracy" refers to accuracy computed based on the original review timestamps grouped by calendar date, reflecting the model's performance progression across actual temporal periods rather than simulated time steps. Meanwhile, Random Forest performance declines as more data are introduced, underscoring its lack of adaptability.

Table 6. Classification Report of Adaptive Random Forest (Score-Based Labeling)

	precision	recall	f1-score	support
negative	0.70	0.72	0.71	71631
neutral	0.06	0.00	0.00	9655
positive	0.84	0.89	0.86	149834
accuracy			0.80	231120
macro avg	0.54	0.54	0.53	231120

weighted avg	0.77	0.80	0.78	231120

Although the Adaptive Random Forest demonstrates balanced performance between negative and positive sentiments, the neutral class remains notably underrepresented. This underperformance stems from the naturally skewed sentiment distribution in OVO reviews, where neutral ratings account for fewer than 10% of all instances. Consequently, the adaptive ensemble receives insufficient exposure to neutral examples, limiting its ability to learn stable decision boundaries for this class. This issue underscores the need for future work to explore class rebalancing or weighted adaptation strategies to improve minority-class recognition in real-time data streams.

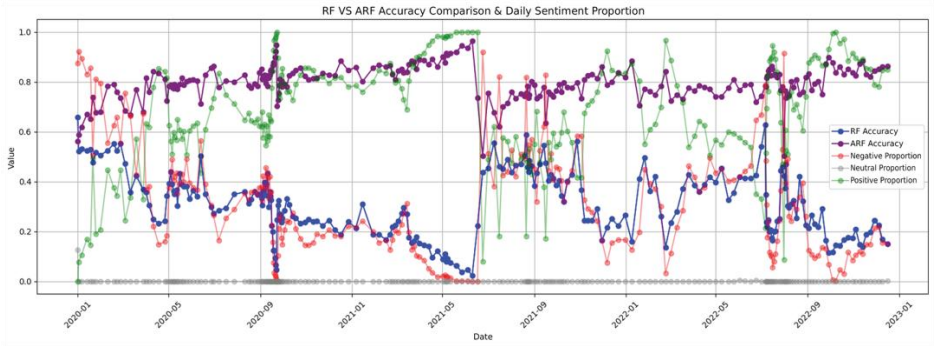

Figure 12. Daily Accuracy Comparison (Score-Based Labeling)

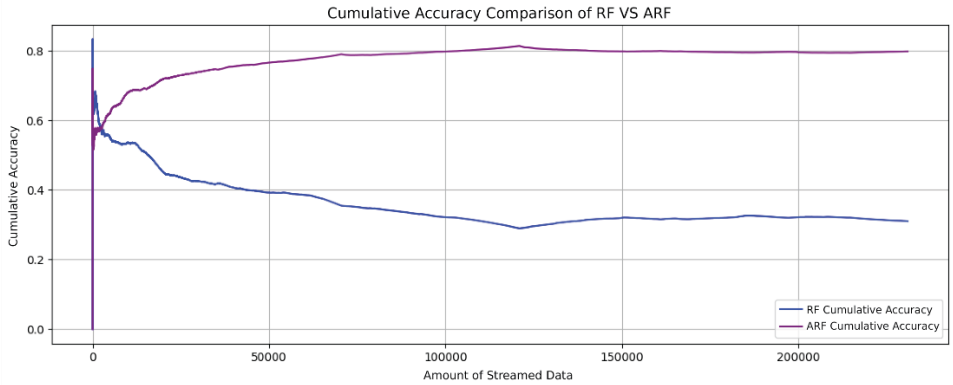

Figure 13. Cumulative Accuracy Comparison (Score-Based Labeling)

4.4.2. Random Forest and Adaptive Random Forest with Content-Based Labeling

The second evaluation used content-based labeling, where sentiment categories were assigned based on IndoRoBERTa's classification of review text. Results are reported in tables 7–8 and figures 14–15. The Random Forest achieved an accuracy of 57% (table 7), higher than in the previous setup, but remained heavily skewed toward predicting positive sentiment (figure 14). Its performance on negative and neutral classes was particularly weak.

Table 7. Classification Report of Random Forest (Content-Based Labeling)

	precision	**recall**	**f1-score**	**support**
negative	0.69	0.01	0.01	76491
neutral	0.34	0.02	0.03	23202
positive	0.57	1.00	0.73	131427
accuracy			0.57	231120
macro avg	0.53	0.34	0.26	231120
weighted avg	0.59	0.57	0.42	231120

ISSN 2723-6471

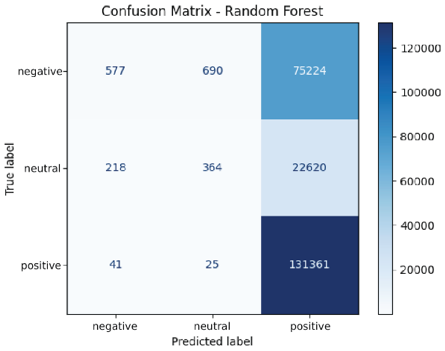

Figure 14. Confusion Matrix of Random Forest (Content-Based Labeling)

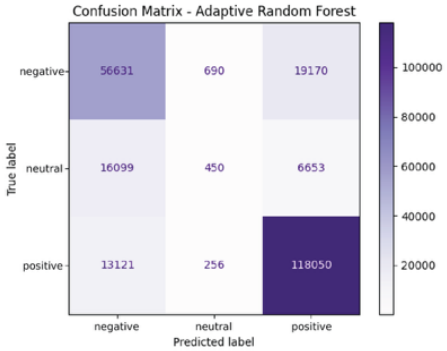

Figure 15. Confusion Matrix of Adaptive Random Forest (Content-Based Labeling)

Similar to the previous experiment, the Random Forest model under content-based labeling exhibits a skewed classification pattern, but in this case biased toward the positive class. The initial batch used for training consisted of five negative, two neutral, and three positive reviews, indicating a mild dominance of the negative class. However, because the model was trained only once and could not update incrementally, early tendencies in IndoRoBERTa's labeling likely influenced its subsequent behavior. IndoRoBERTa tends to assign relatively higher confidence scores to reviews containing positive or appreciative language cues, even when such texts include mixed sentiments. This subtle bias, combined with the overall dominance of positive reviews in the full dataset, amplified the model's inclination to overpredict the positive sentiment throughout the stream.

Adaptive Random Forest again outperformed the baseline, achieving 76% accuracy (table 8). The confusion matrix (figure 15) demonstrates its strength in identifying both negative and positive classes, though neutral sentiment remained difficult to classify. Figures 16 and 17, which present daily and cumulative accuracy comparisons, reveal that both models improve as more data are processed. As in the previous experiment, "daily accuracy" corresponds to performance calculated over reviews grouped by their actual submission dates in the dataset, preserving the temporal structure of real user activity rather than artificial simulation intervals. However, Adaptive Random Forest consistently maintains a margin of roughly 20% higher accuracy, showing greater resilience to evolving data distributions.

Table 8. Classification Report of Adaptive Random Forest (Content-Based Labeling)

	precision	recall	f1-score	support
negative	0.66	0.74	0.70	76491
neutral	0.32	0.02	0.04	23202
positive	0.82	0.90	0.86	131427
accuracy			0.76	231120
macro avg	0.60	0.55	0.53	231120
weighted avg	0.72	0.76	0.72	231120

Adaptive Random Forest once again demonstrated superior stability and adaptability under content-based labeling, yet its classification performance remained weakest for the neutral category. This limitation is consistent with the underlying class imbalance of the dataset and the ambiguous linguistic nature of neutral expressions, which are often context-dependent and harder to detect even for pre-trained models like IndoRoBERTa.

 ISSN 2723-6471

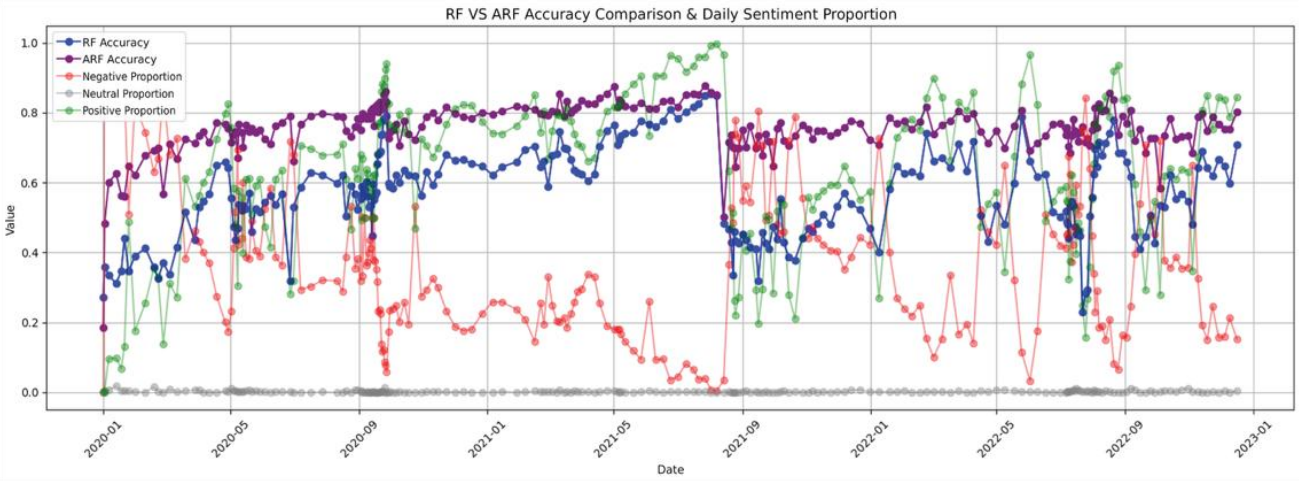

Figure 16. Daily Accuracy Comparison (Content-Based Labeling)

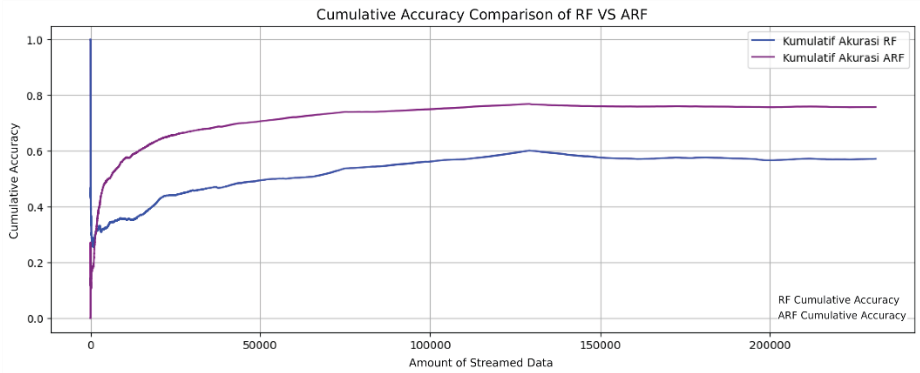

Figure 17. Cumulative Accuracy Comparison (Content-Based Labeling)

The comparative results highlight that the performance gap between Random Forest and Adaptive Random Forest is more pronounced under score-based labeling than under content-based labeling. This can be explained by the deterministic nature of score-based labeling, where star ratings provide explicit ground truth, making the impact of concept drift clearer. Conversely, content-based labeling introduces additional uncertainty, since IndoRoBERTa infers sentiment from textual features, which may blur the distinction between classes.

To provide a clearer comparison across all experimental settings, table 9 summarizes the final accuracy and macro F1-scores of both models under the two labeling strategies, along with their notable performance characteristics.

Table 9. Summary of Model Performance Across Labeling Methods

Model Type	Labeling Method	Accuracy	Macro F1-Score	Characteristics
Random Forest	Score-based	31%	0.16	High negative bias
Adaptive Random Forest	Score-based	80%	0.53	Stable accuracy growth
Random Forest	Content-based	57%	0.26	High positive bias
Adaptive Random Forest	Content-based	76%	0.53	Consistent improvement

Overall, Adaptive Random Forest demonstrates clear superiority in handling concept drift. Its continuous learning mechanism allows it to adjust to new patterns in the data stream, thereby sustaining high accuracy over time. In contrast, the conventional Random Forest is restricted to its initial training data and cannot adapt to changes, resulting in declining performance.

5. Conclusion

This study investigated the effectiveness of Random Forest (RF) and Adaptive Random Forest (ARF) in handling concept drift within sentiment analysis of OVO application reviews collected from the Google Play Store. Two labeling strategies were employed: one based on user ratings (score) and another on textual interpretation using IndoRoBERTa. The results consistently demonstrated that ARF substantially outperformed RF across both labeling approaches, particularly under conditions of dynamic data streams. When labeling was derived from scores, ARF achieved 80% accuracy compared to only 31% for RF, highlighting its ability to adapt to abrupt distributional changes. Under content-based labeling, ARF also exhibited superior performance with an accuracy of 76%, maintaining a notable 20% margin over RF.

The findings further emphasize that score-based labeling provided more deterministic ground truth, enabling clearer distinctions in model performance, while content-based labeling introduced greater variability due to linguistic ambiguity. Overall, the study confirms that adaptive ensemble methods such as ARF are significantly more robust than conventional static models in addressing concept drift, making them suitable for real-time sentiment analysis in evolving digital platforms. These results underline the critical role of adaptability in ensuring model reliability over time. Building on these findings, future research should extend beyond the limitations of this study by incorporating more diverse data sources, such as multi-platform reviews or cross-domain user feedback, to ensure broader generalizability. In addition, future work could explore the comparative effectiveness of the fixed 10% drift detection threshold used in this study against algorithmic approaches such as ADWIN, which monitor error-rate fluctuations to detect adaptive changes. This comparison would clarify the trade-offs between heuristic and statistical drift detection methods, particularly in non-stationary data streams. Furthermore, although the Adaptive Random Forest demonstrated superior adaptability in this study, future work should benchmark its performance against other adaptive ensemble or streaming models, such as Online Bagging, Leveraging Bagging, or neural-based adaptive architectures, to better understand the trade-offs between accuracy, computational efficiency, and drift responsiveness. Finally, integrating temporal and contextual information, such as major application updates or policy changes, may provide deeper insights into the underlying causes of drift and further enhance the robustness of adaptive sentiment analysis models.

6. Declarations

6.1. Author Contributions

Conceptualization: B.F.A., S., and N.I.; Methodology: S.; Software: B.F.A.; Validation: B.F.A., S., and N.I.; Formal Analysis: B.F.A., S., and N.I.; Investigation: B.F.A.; Resources: S.; Data Curation: S.; Writing Original Draft Preparation: B.F.A., S., and N.I.; Writing Review and Editing: S., B.F.A., and N.I.; Visualization: B.F.A.; All authors have read and agreed to the published version of the manuscript.

6.2. Data Availability Statement

The data presented in this study are available on request from the corresponding author.

6.3. Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

6.4. Institutional Review Board Statement

Not applicable.

6.5. Informed Consent Statement

Not applicable.

6.6. Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] G. J. Aguiar and A. Cano, "Enhancing Concept Drift Detection in Drifting and Imbalanced Data Streams through Meta-Learning," in *Proceedings - 2023 IEEE International Conference on Big Data, BigData 2023, Institute of Electrical and Electronics Engineers Inc*., vol. 2023, no. Dec., pp. 2648–2657, 2023. doi: 10.1109/BigData59044.2023.10386364.

[2] M. A. Shyaa, N. F. Ibrahim, Z. Zainol, R. Abdullah, M. Anbar, and L. Alzubaidi, "Evolving cybersecurity frontiers: A comprehensive survey on concept drift and feature dynamics aware machine and deep learning in intrusion detection systems," *Eng. Appl. Artif. Intell.*, vol. 2024, no. Nov., pp. 1–15, 2024. doi: 10.1016/j.engappai.2024.109143.

[3] B. Halstead , "Analyzing and repairing concept drift adaptation in data stream classification," *Mach Learn*, vol. 111, no. 10, pp. 3489–3523, Oct. 2022, doi: 10.1007/s10994-021-05993-w.

[4] X. Jin and Y. Zhang, "Adaptive Random Forest with Dynamic Detectors for Evolving Data Stream Classification," in *ACM International Conference Proceeding Series, Association for Computing Machinery*, vol. 2023, no. Mar., pp. 678–684, 2023. doi: 10.1145/3594315.3594390.

[5] N. Abdulla, M. Demirci, and S. Ozdemir, "Adaptive Learning on Fog-Cloud Collaborative Architecture for Stream Data Processing," in *2021 International Symposium on Networks, Computers and Communications, ISNCC 2021, Institute of Electrical and Electronics Engineers Inc*., vol. 2021, no. Jul., pp. 1–6, 2021. doi: 10.1109/ISNCC52172.2021.9615824.

[6] A. O. Alqabbany and A. M. Azmi, "Measuring the effectiveness of adaptive random forest for handling concept drift in big data streams," *Entropy*, vol. 23, no. 7, pp.1-12, Jul. 2021, doi: 10.3390/e23070859.

[7] K. Parnow, Z. Li, and H. Zhao, "Grammatical error correction: More data with more context," *IEEE Access,* vol. 2020, no. Dec., pp. 1–10, 2020, doi: 10.1109/IALP51396.2020.9310498.

[8] Ł. Korycki and B. Krawczyk, "Adaptive Deep Forest for Online Learning from Drifting Data Streams," *arXiv*, vol. 2020, no. Oct., pp. 1–12, 2020. Available: http://arxiv.org/abs/2010.07340

[9] M. G. Rahman and M. Z. Islam, "Adaptive Decision Forest: An Incremental Machine Learning Framework," *arXiv*, vol. 2021, no. Jan., pp. 1–12, 2021, Available: http://arxiv.org/abs/2101.11828

[10] L. Yang and A. Shami, "A Lightweight Concept Drift Detection and Adaptation Framework for IoT Data Streams," *IEEE Internet Things Mag.*, vol. 2021, no. Apr., pp. 1–10, 2021, doi: 10.1109/IOTM.0001.2100012

[11] F. Ceschin, M. Botacin, H. M. Gomes, F. Pinagé, L. S. Oliveira, and A. Grégio, "Fast & Furious: Modelling Malware Detection as Evolving Data Streams," *Expert Syst. Appl.*, vol. 2022, no. May, pp. 1–12, 2022, doi: 10.1016/j.eswa.2022.118590.

[12] M. Badar, W. Nejdl, and M. Fisichella, "FAC-fed: Federated adaptation for fairness and concept drift aware stream classification," *Mach Learn*, vol. 112, no. 8, pp. 2761–2786, Aug. 2023, doi: 10.1007/s10994-023-06360-7.

[13] E. Yu, J. Lu, B. Zhang, and G. Zhang, "Online Boosting Adaptive Learning under Concept Drift for Multistream Classification," *arXiv*, vol. 2024, no. Jan., pp. 1–12, 2024. Available: http://arxiv.org/abs/2312.10841

[14] Z. Han , "A Survey on Event Tracking in Social Media Data Streams," *Big Data Mining and Analytics*, vol. 7, no. 1, pp. 217–243, Mar. 2024, doi: 10.26599/BDMA.2023.9020021.

[15] E. Yu, Y. Song, G. Zhang, and J. Lu, "Learn-to-adapt: Concept drift adaptation for hybrid multiple streams," *Neurocomputing,* vol. 496, no.1, pp. 121–130, Jul. 2022, doi: 10.1016/j.neucom.2022.05.025.

[16] Y. Vivek, "ROSFD: Robust Online Streaming Fraud Detection with Resilience to Concept Drift in Data Streams," *AI Res. J.*, vol. 2025, no. Jan., pp. 1–12, 2025.

[17] M. A. Khder, "Web scraping or web crawling: State of art, techniques, approaches and application," *International Journal of Advances in Soft Computing and its Applications*, vol. 13, no. 3, pp. 144–168, 2021, doi: 10.15849/ijasca.211128.11.

[18] Z. Jiang, B. Gao, Y. He, Y. Han, P. Doyle, and Q. Zhu, "Text Classification Using Novel Term Weighting Scheme-Based Improved TF-IDF for Internet Media Reports," *Math. Probl. Eng.,* vol. 2021, no. Jan., pp. 1–12, 2021, doi: 10.1155/2021/6619088.

[19] S. J. S. K. V. S. Jain, "An Effective TF-IDF Model to Improve the Text - Classification Performance," *Comput. Sci. Rev.*, vol. 2024, no. Jan., pp. 1–10, 2024.

[20] S. G. Liu, R. Liu, and S. Y. Rao, "Secure and efficient two-party collaborative SM9 signature scheme suitable for smart home," *Journal of King Saud University - Computer and Information Sciences*, vol. 34, no. 7, pp. 4022–4030, Jul. 2022, doi: 10.1016/j.jksuci.2022.05.008.

[21] Y. Zhong, H. Yang, Y. Zhang, P. Li, and C. Ren, "Long short-term memory self-adapting online random forests for evolving data stream regression," *Neurocomputing*, vol. 457, no. Oct., pp. 265–276, Oct. 2021, doi: 10.1016/j.neucom.2021.05.026.

[22] Z. Pang, J. Cen, and M. Yi, "Unsupervised concept drift detection method based on robust random cut forest," *International Journal of Machine Learning and Cybernetics*, vol. 14, no. 12, pp. 4207–4222, Dec. 2023, doi: 10.1007/s13042-023-01890-x.

[23] F. Ridder, K. H. Chen, and N. Alachiotis, "Accelerated Real-Time Classification of Evolving Data Streams using Adaptive Random Forests," in *Proceedings - International Conference on Field-Programmable Technology, ICFPT, Institute of Electrical and Electronics Engineers Inc*., vol. 14, no. Dec., pp. 4207–4222, 2023. doi: 10.1109/ICFPT59805.2023.00031.

[24] M. M. Yacoub, A. Rezk, and M. B. Senousy, "Adaptive classification in data stream mining," *J. Theor. Appl. Inf. Technol.*, vol. 2020, no. Jan., pp. 1–10, 2020

[25] S. si Zhang, J. wei Liu, and X. Zuo, "Adaptive online incremental learning for evolving data streams," *Appl Soft Comput,* vol. 105, no. Jul., pp. 1–12, 2021, doi: 10.1016/j.asoc.2021.107255.

[26] J. Haug, A. Braun, S. Zürn, and G. Kasneci, "Change Detection for Local Explainability in Evolving Data Streams," in *International Conference on Information and Knowledge Management, Proceedings, Association for Computing Machinery*, vol. 2022, no. Oct., pp. 706–716, 2022. doi: 10.1145/3511808.3557257.

[27] G. Goos, E. Bertino, W. Gao, B. Steffen, and M. Yung, "Artificial Neural Networks and Machine Learning – ICANN 2022," *Neural Netw. J.*, vol. 2022, no. Jan., pp. 1–10, 2022. Available: https://link.springer.com/bookseries/558

[28] E. B. Gulcan and F. Can, "Implicit Concept Drift Detection for Multi-label Data Streams," *arXiv*, vol. 2022, no. Jan., pp. 1–12, 2022. Available: http://arxiv.org/abs/2202.00070

[29] W. Ren, P. Wang, X. Li, C. E. Hughes, and Y. Fu, "Semi-supervised Drifted Stream Learning with Short Lookback," in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Association for Computing Machinery*, vol. 2022, no. Aug., pp. 1504–1513, 2022. doi: 10.1145/3534678.3539297.

[30] A. L. Suárez-Cetrulo, D. Quintana, and A. Cervantes, "A survey on machine learning for recurring concept drifting data streams," *Expert Syst. Appl.*, vol. 2023, no. Mar., pp. 1–15, 2023. doi: 10.1016/j.eswa.2022.118934.

[31] S. Kumar, R. Singh, M. Z. Khan, and A. Noorwali, "Design of adaptive ensemble classifier for online sentiment analysis and opinion mining," *PeerJ Comput Sci*, vol. 7, no. 1, pp. 1–24, 2021, doi: 10.7717/peerj-cs.660.

[32] Y. Sun, B. Pfahringer, H. M. Gomes, and A. Bifet, "SOKNL: A novel way of integrating K-nearest neighbours with adaptive random forest regression for data streams," *Data Min Knowl Discov*, vol. 36, no. 5, pp. 2006–2032, Sep. 2022, doi: 10.1007/s10618-022-00858-9.

[33] Z. Yuan, Y. Sun, and D. Shasha, "Forgetful Forests: Data Structures for Machine Learning on Streaming Data under Concept Drift," *Algorithms,* vol. 16, no. 6, pp. 1-12, Jun. 2023, doi: 10.3390/a16060278.

[34] Q. Xiang, L. Zi, X. Cong, and Y. Wang, "Concept Drift Adaptation Methods under the Deep Learning Framework: A Literature Review," *Algorithms*, vol. 16, no. Jun., pp. 1–12, 2023. doi: 10.3390/app13116515.

[35] T. Mahmood and T. Fatima, "Concept Drift in Streaming Data: A Systematic Literature Review," *KJCIS*, vol. 2021, no. Jan., pp. 1–12, 2021. doi: 10.51153/KJCIS.V4I1.43.

[36] D. Mulimani, P. Patil, S. Totad, and R. Benni, "Online Detection and Adaptation of Concept Drift in Streaming Data Classification," in *Procedia Comput. Sci.,* vol. 2024, no. Apr., pp. 2803–2811, 2024. doi: 10.1016/j.procs.2024.04.265

[37] K. Goel and S. Batra, "Adaptive Online Learning for Classification under Concept Drift." *AI Res. J.*, vol. 2025, no. Jan., pp. 1–10, 2025