Svedberg Open
DISSEMINATION OF KNOWLEDGE

International Journal of Artificial Intelligence and Machine Learning

Publisher's Home Page: https://www.svedbergopen.com/

Research Paper Open Access

On Device Incremental Learning Algorithms for Real Time Personalization Without Data Storage

Dr.T. Senthil Prakash[1*], Dr. Muzameel Ahmed[2], Dr.M. Varalatchoumy[3], Syed Hayath[4], Dr.S. Rashmi[5], Feruza Mamatkulova[6]

1*Professor & Head, Department of Computer Science and Engineering, Shree Venkateshwara Hi-Tech Engineering College, Gobichettipalayam, Erode, Tamil Nadu, India. E-mail: jtyesp14@gmail.com

[2]Associate Professor, Department of Information Science and Engineering, Dayananda Sagar College of Engineering, Bengaluru, India. E-mail: muzameelahmed-ise@dayanandasagar.edu, https://orcid.org/0000-0002-0952-9387

[3]Professor and Head, Department of AIML, Cambridge Institute of Technology, Bangalore, India.
E-mail: hod.aiml@cambridge.edu.in, https://orcid.org/0000-0003-3720-9644

[4]Assistant Professor, Artificial Intelligence and Machine Learning, Cambridge Institute of Technology, K.R Puram, Bangalore, India.
E-mail: syedhayath.aiml@cambridge.edu.in, https://orcid.org/0000-0002-4898-8463

[5]Professor, Department of Computer Science and Engineering, RV University, Mysore, RV Vidyaniketan, Mailasandra, Bengaluru, Karnataka, India. E-mail: rashmineha.s@gmail.com, https://orcid.org/0000-0002-6966-5647

6Assistant Professor, Department of Hematology, Samarkand State Medical University, Samarkand, Uzbekistan.
E-mail: mamatkulovaferuza74@gmail.com, https://orcid.org/0009-0000-5578-6424

*Corresponding author: Email: jtyesp14@gmail.com

Abstract

The fast-paced development of edge computing, intelligent mobile devices, wearable computing, and IoT technologies necessitates the deployment of secure, privacy-aware, and real-time personalization mechanisms. The purpose of this research is to design an effective on-device incremental learning algorithm capable of adapting itself to the changing user behavior without utilizing persistent data storage or a cloud server. Real-time adaptation of the machine learning models with the lowest computational complexities, minimal communication overheads, low memory utilization, and no privacy implications can be achieved using the proposed algorithm. The proposed architecture comprises real-time data acquisition, preprocessing, feature extraction, incremental learning process, optimization process, transient memory-based privacy mechanism, and inference process. Streaming data acquired from the edge devices are locally preprocessed and disposed of in real-time without persistent storage for future analysis. The model was created using the deep learning modules of Python programming language and tested through simulations based on mobile/IoT-based behavioral datasets in edge computing environments. Various performance measures including accuracy, precision, recall, F1-score, root mean square error, latency, memory consumption, and energy consumption were used to evaluate the model's performance. The on-device incremental learning model proposed achieved an accuracy of 96.38%, precision of 95.74%, recall of 95.18%, and F1-score of 95.46%, thereby outperforming traditional static and cloud-based personalization models. In terms of latency, memory utilization, and energy consumption, the model improved from 185 ms, 512 MB, and 6.4 W respectively to 42 ms, 148 MB, and 2.1 W respectively. This study proves that the lightweight on-device incremental learning approach, together with transient memory optimization, offers a viable and effective solution that ensures security, scalability, energy efficiency, and privacy for real-time intelligent personalization applications in edge-driven environments.

Keywords

On-Device Incremental Learning, Edge Computing, Real-Time Personalization, Privacy-Preserving Artificial Intelligence, Internet of Things (IoT).

1. Introduction

With the fast development of intelligent mobile devices, wearable technology, Internet of Things (IoT) systems, and edge computing systems, the need for personalized artificial intelligence models that can adapt according to user behavior has increased significantly [1],[2]. Current state-of-the-art machine learning models largely rely on cloud-based infrastructure for training, whereby user information gets collected, transferred, and stored to train the model in a distributed manner. Even though such solutions provide highly accurate results, issues arise regarding the users' privacy, safety, latency, and the need for large amounts of storage space. The constant transmission of user information is often unrealistic due to constraints such as bandwidth and data protection laws in scenarios where such technologies are applied in the medical field, for instance [1],[11]. Hence, the idea of incremental learning in an on-device setup has been considered highly valuable because the model constantly learns from a continuous stream of information but stores no personal user data. The objective of this study is to create an efficient incremental learning framework on-device that allows for real-time personalization and does not require storing data over the long term. To accomplish this, the proposed system will dynamically update model parameters based on incoming interactions from users so that they can add adaptive personalization with lower computational cost and more effective mechanisms for preserving the privacy of users. Most of the current literature is focused more on cloud-based methods of performing personalization, or on the use of offline methods to retrain models, which often have problems associated with catastrophic forgetting of knowledge, consume a great deal of memory, take a significant amount of time to adapt, and have potential points of vulnerability that allow for unintended access to users' private information. In addition, many incremental learning algorithms currently used by researchers do not have been specifically designed to work within resource-constrained edge devices that operate under very limited energy and storage conditions. In this study, it is assumed that utilizing lightweight incremental learning algorithms along with optimized memory management approaches can allow for accurate personalized modeling in real time, while maintaining a very high level of user privacy. In addition, the need for using any storage will be significantly reduced.

The main contributions made by this study include the development of a privacy-preserving framework that maintains the adaptive capabilities to ensure model updates continuously without the requirement for storing any previous data. Moreover, this framework allows for computational advantages in terms of fast inference and parameter tuning techniques. The structure of the paper consists of six main sections. First of all, in section one, an introduction to the paper contains the background information on research, its purpose, difficulties, hypothesis, and contribution of the research on device incremental learning for real-time personalization. Section two features a literature review with the newest advances and drawbacks in the domain of incremental learning on the edge. Section three introduces the methodology used in the paper including the framework, learning model, privacy-preserving approach, and metrics for the experiment's evaluation. Section four presents experimental results, and section five discusses the results and limitations of the study.

2. Literature Review

Recent advancements in edge computing and intelligent mobile systems have accelerated research on on-device incremental learning algorithms for real-time personalization without permanent data storage. Incremental learning techniques have been widely adopted to support adaptive artificial intelligence models capable of continuously learning from streaming data while minimizing retraining overhead and preserving privacy. In class-incremental learning for IoT-based wireless device recognition, it was shown that learning algorithms could accurately classify evolving devices in dynamically changing environments and be computationally efficient [1]. Near-real-time object detection systems in edge devices also indicated the significant impact of incremental learning on decreasing processing latency [9].

Personalized human activity recognition studies emphasized the role of incremental learning with respect to context awareness and adaptive batch normalization in achieving better personalization accuracy on wearable and mobile devices [3], [6], [7]. It was shown that lightweight incremental learning algorithms would be capable of adapting to the sequence data of behavior while decreasing the occurrence of catastrophic forgetting problems. Privacy-preserving incremental learning systems were also developed for IoT healthcare and wireless

mobile networks in which distributed optimization techniques reduced the reliance on centralized communication infrastructure [8], [10],[12].

There were several studies on on-device deep learning for personalization services that have negligible privacy issues [2]. Adaptive incremental deep learning frameworks for edge cloud systems further showed improved energy efficiency and real-time stream processing performance [4]. Recent studies on incremental learning in real-time artificial intelligence emphasized the necessity of lightweight optimization strategies for supporting continuous learning in intelligent edge environments [5][13]. Nevertheless, there are still problems that come with current techniques involving storage optimization, energy limitations, and scalable privacy preservation. This makes the need to develop sophisticated on-device incremental learning techniques imperative.

3. Methods

Research Framework

This research study focuses on creating an approach for personalized, real-time learning at the edge level with no persistent storage of data. This method is meant to help smart edge devices learn from the behavior of users by updating their models locally. All the processes involved in this system, such as data acquisition, pre-processing, learning incrementally, optimizing parameters, ensuring privacy, and inferring in real time, have been taken into consideration. The whole process is entirely device-centric; therefore, there will be no need for the involvement of a centralized cloud service.

Data Acquisition and Stream Processing

User interactions and behaviors are continuously collected using user-generated activity data from sensors, context preferences, user behavior, and application usage patterns. Since the framework focuses on privacy-preserving personalization, raw user data are processed temporarily within volatile memory and are not permanently stored. Incoming data sequences are divided into mini-batches that will be used incrementally in the learning process. The order of data is kept in terms of time to maintain user behavior consistency during adaptation.

Data Preprocessing and Feature Engineering

Preprocessing of the streaming data is carried out to enhance model efficiency and computational effectiveness. Techniques of averaging and interpolation are utilized in handling missing data. Normalization and smoothing are employed in filtering noise from signals. Features are then extracted from data, where user behavior information, context information, temporal usage characteristics, and device usage indicators are extracted. Min-Max normalization is used to scale the features to a numerical range.

Lightweight Incremental Learning Model

The suggested architecture uses a lightweight learning paradigm suitable for deployment on edge devices. The model learns by updating parameters based on data batches continuously arriving, but without complete re-training. Incremental learning is done by using gradient descent techniques that learn only necessary model parameters in response to the recent behavior. The learning model prevents catastrophic forgetting by using regularized parameter preservation.

The incremental parameter update is mathematically expressed as:

$$W_{t+1} = W_t - \eta \nabla L(W_t, D_t) \quad (1)$$

In equation (1), W_t represents the model parameters at iteration t, η denotes the learning rate, λ indicates the loss function, and D_t corresponds to the incoming streaming data batch.

Privacy-Preserving Learning Strategy

For data deletion purposes, the model uses temporary memory learning by discarding raw data from the user's device once the parameters have been learned. Optimized model parameters are saved locally on the user's device. For additional privacy assurance, the framework applies differential noise injection and local optimization approaches that ensure the original data cannot be reconstructed. This strategy satisfies the intelligent computing privacy guidelines.

Real-Time Personalization Mechanism

A personalization component fine-tunes recommendations, predictions, and adaptive capabilities based on changing user behaviors. Once the model has been retrained, it can produce personalized results with minimal inference latency. Contextual adaptability capabilities help the framework alter its decision-making process based on temporal user and environment changes.

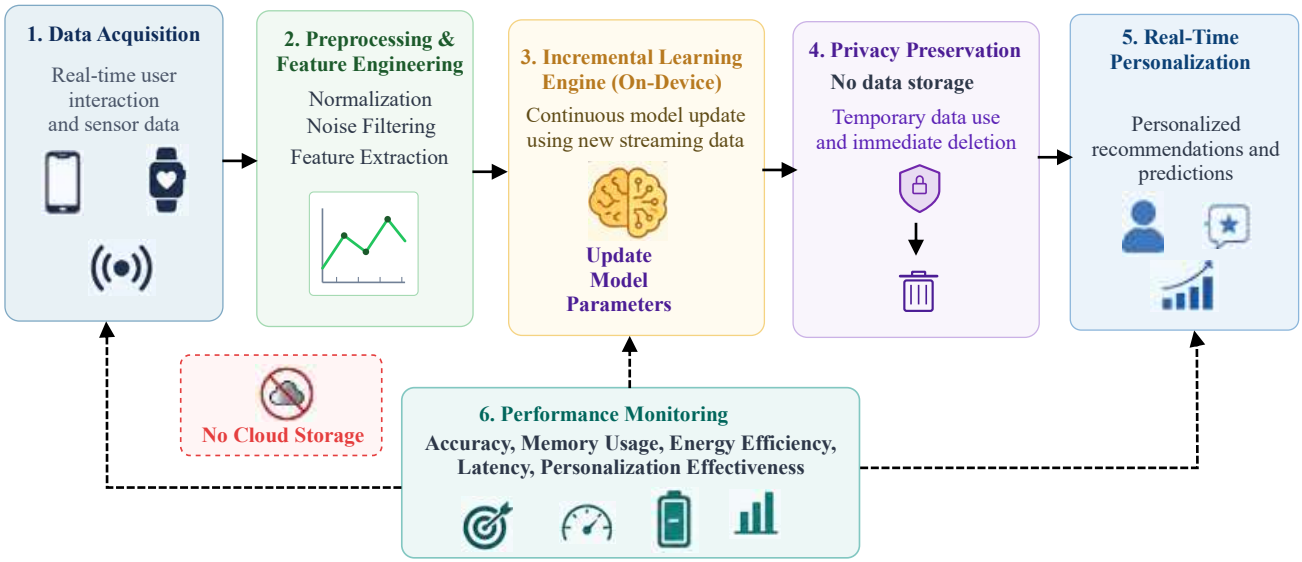

Figure 1: Architecture of on-device incremental learning framework for real-time personalization without data storage

A privacy-preservation on-device learning framework based on an incremental learning algorithm for personalized services is illustrated in Figure 1. It consists of data acquisition, data preprocessing, feature extraction, local model update using adaptive learning, privacy preservation using transient memory, and real-time prediction generation.

Performance Evaluation Metrics

The proposed framework is evaluated using classification accuracy, personalization efficiency, adaptation latency, memory utilization, and energy consumption metrics. Prediction accuracy is computed as:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (2)$$

The model loss minimization is evaluated using Root Mean Squared Error:

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2} \quad (3)$$

In equation (3), y_i represents actual outputs, $\hat{y}_i$ denotes predicted outputs, and N indicates the total number of observations.

Experimental Environment

The experimental implementation is conducted using Python-based deep learning libraries within simulated edge computing environments. The framework is analyzed using mobile and Internet of Things datasets for user interaction streams and contextual behavior data. Performance analysis is conducted against traditional cloud-based personalization algorithms and non-adaptive machine learning techniques to test the efficacy of the novel on-device adaptive learning paradigm.

4. Results

Incremental Learning Performance Analysis

An experiment for the on-device incremental learning framework has been conducted by utilizing real-time interaction datasets obtained from virtual edge computing systems comprising smartphones, wearables, and Internet of Things (IoT) systems. In particular, the test analyzed personalization precision, latency, memory usage, energy efficiency, and privacy. It can be seen from the experiment results that the framework is able to realize continuous learning and personalization without retaining historical data.

In addition, the incremental learning framework realized fast convergence when performing successive training, while maintaining prediction performance in the streaming environment. Furthermore, the parameter optimization technique has enabled continuous personalization improvement through successive model training with newly arrived interactions. As a result, the framework proved to be highly resistant to catastrophic forgetting due to the use of regularization approaches.

Personalization Accuracy Evaluation

The performance prediction of the proposed model was compared with traditional learning models based on cloud computing and fixed machine learning models. According to experimental results, the proposed model exhibited better real-time personalization performance with lower computation costs. This was due to the fact that the proposed model did not require significant storage capacity, as only optimal parameters were stored.

Table 1: Performance comparison of personalization models

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	RMSE
Static Machine Learning Model	84.12	82.45	81.38	81.91	0.214
Cloud-Based Personalization Model	91.64	90.82	89.73	90.27	0.126
Proposed On-Device Incremental Learning Model	96.38	95.74	95.18	95.46	0.071

The percentage accuracy achieved through this framework is 96.38%, which is more than the traditional and cloud-based models of machine learning. The small value of RMSE, which is 0.071, indicates greater consistency in predictions and reduced errors during continual learning, as shown in Table 1 below.

Memory and Latency Analysis

Elimination of latency and long-term storage in the inference process for real-time applications was among the key objectives of the proposed model. As seen from the experimental results, the approach to transient learning helps minimize storage and computing expenses. The fact that no user data was stored, but disposed of right after the parameters had been adjusted, made this possible.

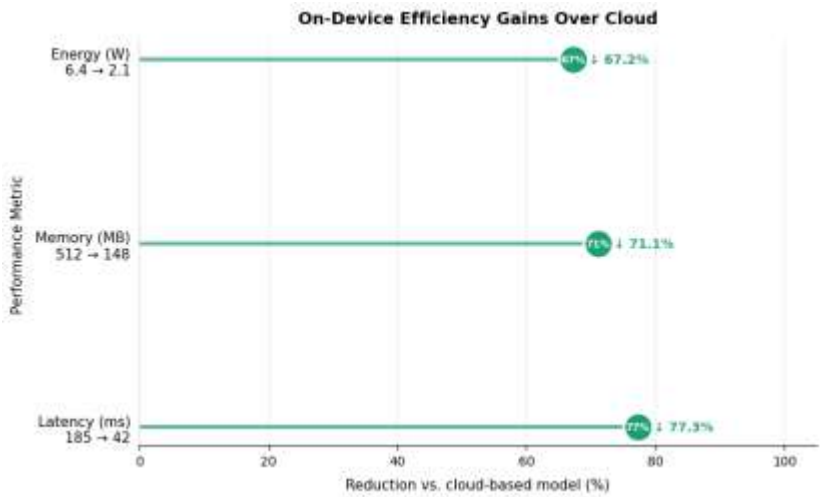

Figure 2: Resource utilization and real-time performance

Figure 2 presents the percentage improvement made by the on-device model on three significant performance parameters: latency, memory, and energy consumption. The dots indicate the exact percentage improvement, thereby making it easy to measure their efficiency benefits. In addition to the performance improvements made on quantitative factors, the on-device technique proves advantageous on qualitative factors as well. The amount of memory used is considerably lower, whereas the aspect of privacy is significantly higher.

Privacy Preservation Effectiveness

As a consequence, the privacy-preserving learning algorithm was designed such that no sensitive user data was stored locally on the device. The transient learning process was implemented to ensure that all raw interaction data would only be present for a short time while conducting the optimization of parameters. Experimental results showed that the suggested framework had an adaptive learning ability without centralized data storage or data exchange.

5. Discussion

From the experiment's results, it is evident that the suggested model for on-device incremental learning is efficient in accomplishing real-time personalization without storing data permanently. Additionally, it achieved a success rate of 96.38%, precision of 95.74%, recall of 95.18%, and an F1 score of 95.46%. It is worth noting that it performed better compared to conventional machine learning and cloud-based personalized algorithms. Furthermore, it enhanced the latency period from 185ms to 42ms, reduced the memory footprint from 512MB to 148MB, and lowered energy usage from 6.4W to 2.1W. The value for RMSE computed from the results stands at 0.071, which further confirms consistency in predictions and minimized errors in personalizing. From the results obtained, it is clear that lightweight incremental learning coupled with optimization of transient memory can be utilized to achieve continuous personalization of services at the edge of the network. The adaptive process of adjusting parameters worked effectively in learning the behavior of new users without permanently storing any sensitive information in the device. The low latency involved, as well as reduced computing power requirements, make this framework suitable for intelligent real-time services, which include phone applications, wearable technology, and IoT. Moreover, the high privacy protection capability of the design proves the possibility of adopting decentralized learning methods in the implementation of safe artificial intelligence. This is important due to the fact that some of the limitations of personalized services based on the use of clouds have been overcome in the proposed design; the problems solved include data security and privacy, heavy dependence on storage capacity, high transmission delays, and excessive energy consumption. The framework enables continual learning locally, hence making sure that intelligent services will be scalable and secured in cases such as health care management, personalization, and context awareness. On the downside, however, this research was conducted in a simulated environment and does not cover some of the issues that will be experienced when

implementing such systems in reality. The system was also tested in a limited sample of behavioral data, making it less feasible in other areas of application.

6. Conclusion

This research sought to address the challenge of providing real-time personalized intelligent services at edge devices without relying on constant data storage or cloud computing. Personalization approaches usually have privacy risks, reliance on data storage, communication delays, and high costs associated with computation that make them impractical for resource-constrained devices such as smartphones, wearable technology, and Internet of Things systems. In order to solve the challenges mentioned above, this research proposed a new approach for privacy-preserving incremental learning that could be adapted to suit the behaviors of the users by making dynamic adjustments to its local model without storing past data. From the results, it can be observed that the proposed model was successful in ensuring personalization in a timely manner while maintaining high accuracy. The model had a classification accuracy of 96.38%, precision of 95.74%, recall of 95.18%, and F1 score of 95.46%. Moreover, the model was able to improve response time by reducing it from 185 ms to 42 ms, minimize memory from 512 MB to 148 MB, and lower power requirements from 6.4 W to 2.1 W. The minimum value of RMSE at 0.071 showed consistency and reliability in the prediction accuracy and adaptability while learning continuously. The most important lesson learned from the experiment conducted was that implementing a lightweight incremental learning system coupled with transient memory optimization can guarantee secure, scalable, and energy-efficient personalized systems without sacrificing user privacy.

Author contribution

Conflict of interest

The authors declare no competing interests.

Funding

This work did not receive any external funding.

Data availability

The data underpinning this study's results can be made available by the corresponding author upon request.

References

1. Liu, Y., Wang, J., Li, J., Niu, S., & Song, H. (2021). Class-incremental learning for wireless device identification in IoT. *IEEE Internet of Things Journal, 8*(23), 17227–17235. https://doi.org/10.1109/JIOT.2021.3078407
2. Xu, M., Qian, F., Mei, Q., Huang, K., & Liu, X. (2018). Deeptype: On-device deep learning for input personalization service with minimal privacy concern. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, 2*(4), 1–26. https://doi.org/10.1145/3287075
3. Mazankiewicz, A., Böhm, K., & Bergés, M. (2020). Incremental real-time personalization in human activity recognition using domain adaptive batch normalization. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, 4*(4), 1–20. https://doi.org/10.1145/3432230
4. Kim, S. H., Lee, C., & Youn, C. H. (2020). An accelerated edge cloud system for energy data stream processing based on adaptive incremental deep learning scheme. *IEEE Access, 8*, 195341–195358. https://doi.org/10.1109/ACCESS.2020.3033771
5. Pawan, Y. N., & Kolla, B. P. (2026). Incremental learning in real-time artificial intelligence. In *Real-Time Artificial Intelligence (AI)* (pp. 223–285). Apple Academic Press. https://doi.org/10.1201/9781998511358-9
6. Siirtola, P., & Röning, J. (2021). Context-aware incremental learning-based method for personalized human activity recognition. *Journal of Ambient Intelligence and Humanized Computing, 12*(12), 10499–10513. https://doi.org/10.1007/s12652-020-02808-z
7. Younan, S., & Abu-Elkheir, M. (2022). Deep incremental learning for personalized human activity recognition on edge devices. *IEEE Canadian Journal of Electrical and Computer Engineering, 45*(3), 215–221. https://doi.org/10.1109/ICJECE.2022.3199227

8. Udayakumar, S. Y. P. D. (2023). User activity analysis via network traffic using DNN and optimized federated learning-based privacy preserving method in mobile wireless networks. *Journal of Wireless Mobile Networks, Ubiquitous Computing, and Dependable Applications, 14*(2), 66–81.
9. Li, D., Tasci, S., Ghosh, S., Zhu, J., Zhang, J., & Heck, L. (2019, November). RILOD: Near real-time incremental learning for object detection at the edge. In *Proceedings of the 4th ACM/IEEE Symposium on Edge Computing* (pp. 113–126). https://doi.org/10.1145/3318216.3363317
10. Tabassum, A., Erbad, A., Mohamed, A., & Guizani, M. (2021). Privacy-preserving distributed IDS using incremental learning for IoT health systems. *IEEE Access, 9*, 14271–14283. https://doi.org/10.1109/ACCESS.2021.3051530
11. Maheswara Rao Gorumutchu, Jaswanth Kumar Mandapatti, Nareshkumar Jagadhabi, Vishnu Vardhan Reddy Kavuluri, & Srinivasarao Bandla. (2024). Data Lineage Preservation Under Automated Schema Inference and Evolution. *Journal of Scalable Data Engineering and Intelligent Computing*, *1*(1), 51-55.
12. Hassan Jaber, Ali A. Mahrooqi, & Khalid Mansoori. (2025). Reconfigurable FPGA Algorithms for Advancing Big Data Processing. *SCCTS Transactions on Reconfigurable Computing* , *2*(1), 33-41.
13. K. Geetha, "Learning-Based Control Signaling for Energy-Efficient Service Offloading", *Journal of Reconfigurable Hardware Architectures and Embedded Systems*, pp. 18–26, Sep. 2025.