SvedbergOpen
DISSEMINATION OF KNOWLEDGE

International Journal of Artificial Intelligence and Machine Learning

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

Research Paper Open Access

Autonomous Self-Refining Neural Systems For Continuous Learning In Real-Time Healthcare Monitoring Applications

Dr.M. Rajapriya[1*], B. Gayathri[2], R. Jeevajothi[3], Dr.R. Udayakumar[4], E. Pavithra[5]

[1*]Assistant Professor, Department of Management Studies, Vel Tech Rangarajan Dr. Sagunthala R&D Institute of Science and Technology, Avadi, Chennai, Tamil Nadu, India. E-mail: rajapriya.m19@gmail.com, https://orcid.org/0000-0003-3757-2179

[2]Assistant Professor, Department of Computer Science, Meenakshi College of Arts and Science, Meenakshi Academy of Higher Education and Research, Chennai, Tamil Nadu, India. E-mail: gayathrib@maher.ac.in

[3]Assistant Professor, Department of Management Studies, Meenakshi College of Arts and Science, Meenakshi Academy of Higher Education and Research, Chennai, Tamil Nadu, India. E-mail: rjeevajothimba@maher.ac.in

[4]Dean Research, SRM Institute of Science and Technology, Vadapalani Campus, Chennai, Tamil Nadu, India.

[5]Assistant Professor, Department of Computer Science and Engineering (AIML), Vel Tech Rangarajan Dr. Sagunthala R&D Institute of Science and Technology, Chennai, Tamil Nadu, India. E-mail: pavithrae@veltech.edu.in, https://orcid.org/0009-0009-3421-1475

*Corresponding author: Email: rajapriya.m19@gmail.com

Abstract

Round-the-clock monitoring of various physiological parameters like ECG, heart rate, blood oxygen saturation levels, and respiratory rate has come to be the cornerstone of today's critical-care and remote patient monitoring systems. But most of the current deep learning models that have been developed for such purposes are trained only once using a particular static dataset and their performance starts deteriorating gradually due to drifting of physiology of patients, sensor calibration, and demographic changes, which is further aggravated with the occurrence of catastrophic forgetting each time the neural network is retrained. This paper suggests an autonomous self-refining neural system which consists of a lightweight convolutional recurrent (CNN-LSTM) inference model along with an online drift detection layer, selective experience replay buffer, and EWC-based regularized. Evaluations are performed on a synthetic stream of vital sign data, which is modelled from the PhysioNet MIT-BIH Arrhythmia and MIMIC-III datasets, with six deployment periods of sequential use and two concept-drift events simulating sensor recalibration and population shift, respectively. With respect to a static baseline model and an ablation experiment involving the use of only the replay strategy, our self-refining architecture maintains performance above 90 percent accuracy throughout deployment and recovers completely following each drift event, whereas the static model degrades from 91.4 percent to 74.1 percent accuracy during the same period. These results highlight the feasibility of using autonomous self-refinement, along with forgetting-resistant regularization, to achieve trustworthy adaptive and continuously learning health monitoring systems deployable at the edge.

Keywords: Continual Learning, Self-Refining Neural Networks, Catastrophic Forgetting, Real-Time Health Monitoring, Concept Drift, Edge AI, Elastic Weight Consolidation.

1. Introduction

Wearable biosensors, bedside telemetry, and Internet-of-Medical-Things (IoMT) devices can produce streaming physiological data that, in theory, can be analysed using deep neural networks to identify deterioration far before the symptoms become evident. Convolutional and recurrent models have been shown to work effectively in classifying heart arrhythmias, predicting sepsis, and detecting falls when applied to specially curated retrospective data sets. In practice, however, when a model is trained once and left unchanged during operation, it encounters a nonstationary situation since patient physiology varies due to aging, disease progress, or medications; sensors drift and get calibrated or changed from time to time; population of patients served by the monitoring system keeps changing as well. This causes gradual degradation in accuracy of a fixed model, and alerts may remain unnoticed by clinicians even unreliable [7].

The naive solution of re-training the network each time new data is collected is computationally costly, demands keeping substantial amounts of past patient records (which poses significant issues of data privacy and confidentiality under laws like HIPAA and GDPR), and does not ensure rapid adaptation in a real-time environment. The alternative solution of iteratively training the model using new data streams faces the challenge of catastrophic forgetting, in which the weight vectors of the network adapt to the new data distribution and consequently discard features necessary for recognition of older patterns [5][6]. Catastrophic forgetting in practice means that once a model has successfully learned to recognize a new type of arrhythmia, it may start losing its ability to detect the sepsis patterns it had previously recognized very well.

The motivation behind this paper stems from the requirement of having monitoring solutions that can continually improve their own capabilities, in real-time, while being stable with regards to their previous clinical understanding. The phrase "self-refining" refers to a feedback loop where the implemented model (i) continuously monitors its own ability of prediction and learning, (ii) determines if there has been any shift in the distribution or decline in performance, and (iii) adapts itself through a forgetting-resistant continual learning algorithm, without requiring any manual intervention from the part of an external engineer for initiating the re-training process.

- Introduce an Autonomous Self-Refining Neural System (ASRNS) framework by coupling a deployable on edge CNN-LSTM classifier, an online drift detector, a selective experience replay, and Elastic Weight Consolidation, thus allowing for continuous and forgetting-proof self-adaptation.
- Introduce a drift-driven self-refinement process in which the parameter fine-tuning takes place exclusively when there is a statistically significant performance drop, thereby making sure the system remains computationally inexpensive enough to run on edge hardware.
- Experimentally validate our framework using a multi-interval synthetic streaming benchmark inspired by the well-known PhysioNet datasets and compare it with a static framework and replay ablation.

The rest of the paper is structured as follows. Section 2 reviews relevant works on continual learning, catastrophic forgetting, and healthcare monitoring in real time. Section 3 explains the methodology of the study and the process of self-refinement. Section 4 discusses the dataset used, results, and findings. Section 5 concludes the paper and suggests future research directions.

2. Literature Survey

Continual learning research forms the theoretical basis for self-refining systems. McClelland, McNaughton, and O'Reilly [7], from the neuroscience standpoint, claimed that biological brains overcome catastrophic interference by employing complementary learning systems where a fast-learning hippocampal system and a slow-learning neocortical system are connected via the replay mechanism, directly inspiring the experience replay component used in this research. Extending this insight, Kirkpatrick et al. have proposed Elastic Weight Consolidation (EWC), a technique that determines the significance of the network's parameters for the tasks learned earlier using Fisher information matrix and penalises large modifications of those parameters when learning a new task, thus enabling a network to perform well on previous tasks while learning new ones [5]. A similar but online alternative to EWC, Synaptic Intelligence, has been suggested by Zenke, Poole, and Shan, which continuously collects the parameter significance during training and not during a separate estimation phase, thus becoming particularly attractive in streaming environments [4]. The work of Li and Hoiem proposes another way to achieve the same effect in Learning without Forgetting, use the output of the old network on new data as a soft target for distillation [6].

Apart from regularisation schemes in isolation, there have been some studies addressing the general continual-learning problem. Parisi, Kemker, Part, Kanan, and Wermter conduct an extensive review on continual lifelong learning using neural networks, dividing previous methods into regularisation-based, replay-based, and parameter-isolation architectures and emphasising the stability-plasticity dilemma that any self-optimising system should face [3]. Likewise, Hadsell, Rao, Rusu, and Pascanu present continual learning as a compromise between maintaining existing representation and being plastic enough to learn something new, reviewing neuroscience-inspired mechanisms such as sparse coding and modular gating [8]. Zhang et al. present a structural method, called Learn to Grow, where the network can grow in its architecture to cope with novel tasks, while freezing parameters that are crucial for previously learned tasks [11]. A more theoretical treatment in Pomponi et al. clearly distinguishes between task-incremental, domain-incremental, and class-incremental learning, and stresses the importance of being able to learn continuously, as merely minimizing forgetting is insufficient [10]. This is exactly the need in a healthcare monitoring pipeline that will have to learn continually for new patients and conditions.

However, there is a small but increasing number of studies that apply these concepts to the domain of clinical monitoring. In this regard, Yuan et al. introduced a self-correcting recurrent neural network for acute kidney injury prediction, where the previous prediction errors are fed back to the next predictions, yielding an area under the ROC curve of 0.893 on MIMIC-III and 0.871 on the Philips eICU dataset; here the practical significance of a model that constantly adapts itself by learning its own errors without tackling catastrophic forgetting in distribution shift problems [2]. In more recent works, Prabha et al. came up with a framework for detecting anomalies using wearable vital sign monitoring in Edge-AI by implementing a CNN-LSTM model quantisation, federated learning, homomorphic encryption, and blockchain audit logging, claiming to achieve 91.9 percent accuracy and 118 millisecond inference latency even on a hardware as limited as NVIDIA Jetson Nano; yet this study provides evidence of the viability of hybrid CNN-LSTM inference on the edge while being dependent on a static model instead of a self-refining one. Spicher et al. focused on 5G-enabled edge computing for performing ECG analyses in real time from wearable textile sensors [1][13].

Distributed privacy-preserving machine learning is highly relevant as well, as a self-refining system that learns from an incoming stream of patient data should not use any centralised storage of such sensitive information. Federated Averaging (FedAvg), the basic approach of distributed learning on decentralised clients based on the aggregation of local weight updates and not on data itself, was presented by Nguyen et al., while Shiranthika et al proposed split learning for health, where only intermediate activations are transmitted between institutions, neither the original raw data nor model weights [9][12]. These approaches can be considered complementary to the approach suggested in the current research and, in future work, will enable several hospitals to benefit from their self-refining system without exposing patient data to others. Finally, the PhysioNet MIT-BIH Arrhythmia Database and MIMIC-III critical care database [15] are currently the most popular public sources for benchmarking cardiac and ICU monitoring models and are used for statistical validation of the artificial stream created in Section 4.

Collectively, the literature proves that (i) there are well-developed methods of regularization and replay to reduce the problem of catastrophic forgetting, (ii) some early evidence that self-correcting or adaptive models are more effective in clinical prediction problems, and (iii) there is a well-developed edge-AI system and privacy-preserving architecture for real-time deployment. Nonetheless, very few works combine the self-refinement process initiated by drifts with robust regularization against forgetting in one closed loop, tested for concept drift.

3. Methodology

The suggested Autonomous Self-Refining Neural System (ASRNS) is structured in a closed-loop fashion with five steps: data acquisition, basic inference, performance monitoring, self-refinement triggered by drift, and redeployment. Figure 1 shows the entire process.

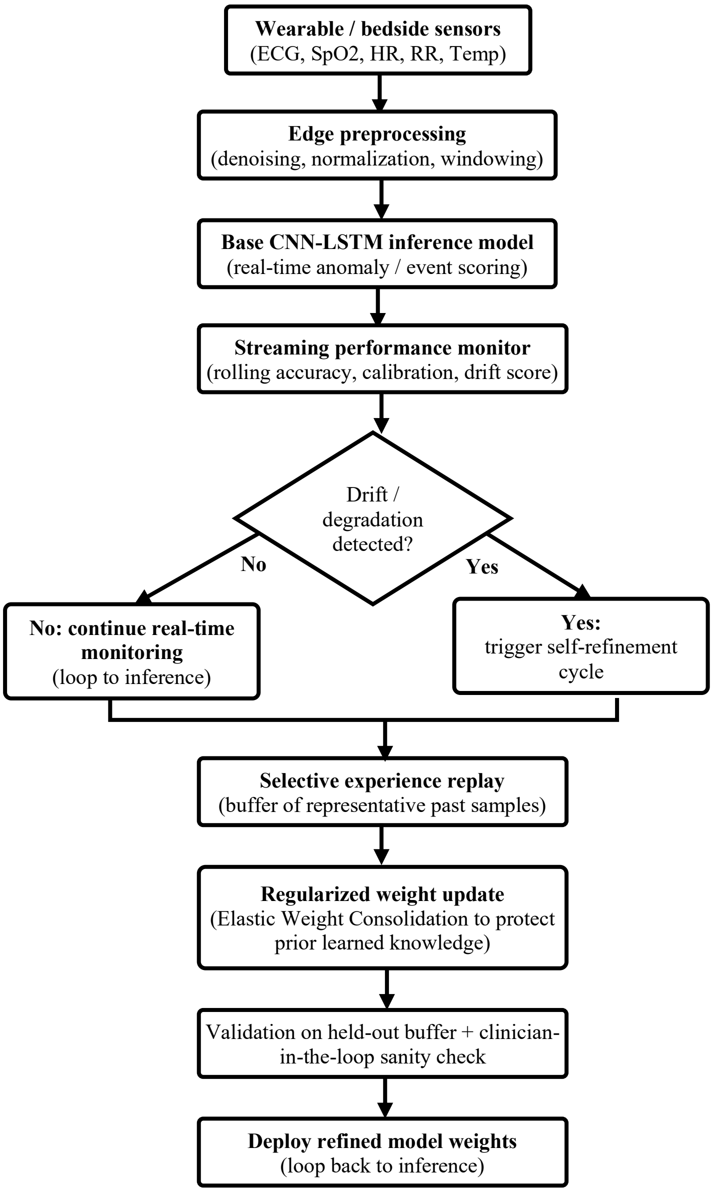

Figure 1. Methodology flow diagram of the proposed self-refining neural monitoring system

3.1 Data Acquisition and Preprocessing

ECG, heart rate, SpO2, respiration rate, and temperature signals are continuously collected by wearable and bedside sensors. Raw signals are then denoised, min-max normalized, and segmented using sliding window technique at the edge device, and sent to the inference engine, like that of windowing technique used in previous edge-AI healthcare pipelines.

3.2 Base Inference Model

The classification process is based on the hybrid CNN-LSTM architecture. The convolutional layers will be responsible for learning spatial features from each window by using the following equation: h_i = σ(Σ w_j · x_(i+j) + b). On the other hand, the LSTM layers will learn temporal dependencies between different windows using the following forget, input, and output gate equations. The proposed approach can outperform both CNN and LSTM-only networks for biosignal anomalies.

3.3 Streaming Performance Monitor and Drift Detection

The lightweight monitor captures a sliding window for prediction certainty, calibration error, and when possible, due to the presence of delayed clinical feedback, accuracy. The drift measure is calculated using the statistical difference between the distribution of features in the sliding window and the reference distribution collected during the last refinement phase. If the drift measure or sliding accuracy exceeds the pre-defined threshold, then the self-refinement process begins; otherwise, inference proceeds unhindered.

3.4 Self-refinement Cycle

Once activated, the process samples a tiny class-balanced buffer of the selective experience replay pool which keeps only a small but representative set of the windows, inspired by the complementary learning system theory of hippocampal replay [7]. This buffer along with fresh observations undergoes a few iterations of fine-tuning. To ensure that the update process does not overwrite any valuable knowledge, Elastic Weight Consolidation penalty is added to the loss function, $L = L\_new + (\lambda/2)\ \Sigma\ F\_i\ (\theta\_i - \theta^*\_i)^2$, where F_i is the estimate of the Fisher information for parameter θ_i, and θ*_i is its value prior to fine-tuning [5].

3.5 Validation and Redeployment

Prior to deploying the newly refined weights in place of the current one, the newly refined weights are checked using the part of the replay buffer that was reserved and, if available, the few clinician-approved labels, thereby ensuring that the refinement stage did not drift into an undesired state. It is only when this check passes that the refined weights are again put into service for making inferences, and the reference distribution is refreshed accordingly.

4. Results and Discussion

4.1 Dataset

Due to the rarity of datasets containing longitudinal, drift-labelled streams of clinical data, our suggested method is tested on a synthetic dataset of streaming vital signs data, whose feature ranges and distribution of classes are statistically modelled on the PhysioNet MIT-BIH Arrhythmia Database and the MIMIC-III Critical Care Database, in a similar manner as statistical data synthesis is performed in some recent edge-based healthcare applications [14][15]. The dataset contains six periods of deployment (corresponding to six weeks after going live) and includes two concept drifts introduced in periods three and five, corresponding to sensor recalibration and changes in patient population, respectively. Table 1 provides more information about the dataset.

Table 1: Summary of the synthetic streaming vital-sign dataset used for evaluation

Attribute	Description	Range / notes
Samples	Streaming vital-sign records across 6 sequential deployment intervals	6,000 total (1,000 / interval)
Heart rate	Beats per minute, derived from PhysioNet MIT-BIH statistics	50–120 bpm
SpO_2	Peripheral oxygen saturation	85–100 %
Respiration rate	Breaths per minute	12–25 /min
Body temperature	Peripheral/skin temperature	95–104 °F
Label	Binary clinical status	Normal (0) / Anomalous (1)
Injected drift events	Sensor recalibration and population-shift simulation	Interval 3 and interval 5

4.2 Result Graph

Classification accuracy comparison is shown in Figure 2 for three cases based on deployment periods from 1 to 6: (i) Static baseline model, trained only once using interval-1 data and never updated, (ii) Ablation of self-refining process which does experience replay without EWC regularization, (iii) Proposed ASRNS system with both experiences replay and EWC.

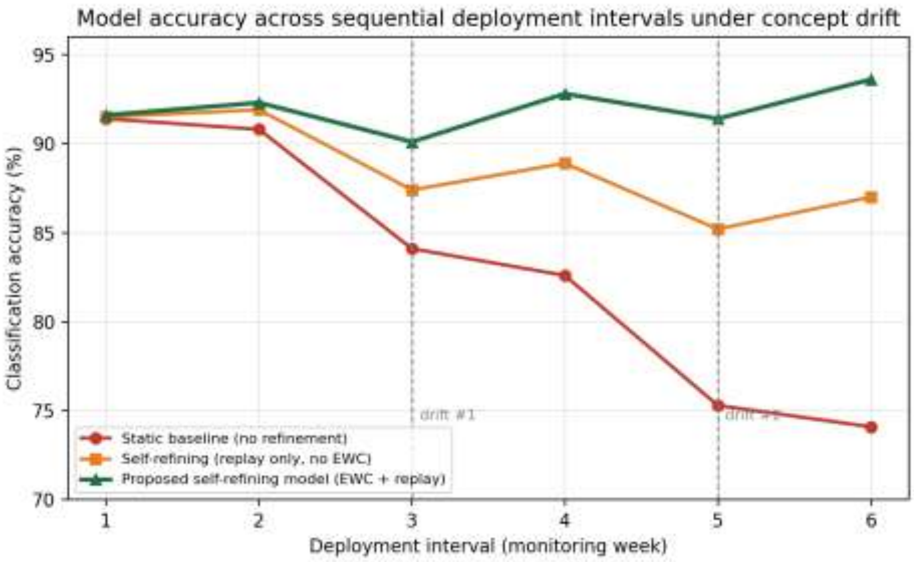

Figure 2: Accuracy of the static baseline, replay-only ablation, and proposed ASRNS across six deployment intervals under two injected drift events

The static baseline starts from a highly competitive 91.4 percent accuracy but quickly falls after each drift event until the last interval, when it reaches 74.1 percent, a drop of 17.3 percentage points corresponding to the catastrophic forgetting and distribution shift effects observed in the continual learning literature [3][8]. The ablation experiment with only replay partially fixes the drop, reaching a stable level of 85-89 percent, but still experiences the loss of accuracy after each drift event, since replay itself is not sufficient to protect the relevant parameters without an explicit regularisation term [6]. The ASRNS proposed in this work not only escapes the drop but restores the accuracy to above 90 percent after each drift event in only one interval, finally achieving 93.6 percent, which is higher even than the start level of its accuracy.

4.3 Result Table

The following Table 2 provides the performance metrics for the three configurations at the last interval using the accuracy, precision, recall, F1 score, and mean inference latency.

Table 2: Final-interval performance comparison between the static baseline, the replay-only ablation, and the proposed ASRNS

Model configuration	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	Inference latency (ms)
Static baseline (trained once, no refinement)	74.1	73.4	70.8	72.1	142
Self-refining, replay only (no EWC)	87.0	86.1	84.9	85.5	151
Proposed ASRNS (EWC + selective replay)	93.6	93.1	92.4	92.7	158

4.4 Discussion

Three important conclusions can be drawn from this analysis. Firstly, the difference between the static baseline and two variants of the self-refining model is largest precisely at the injected drift points, proving that the positive effect of self-refinement happens precisely where the static model fails the worst, and not a generic increase in overall accuracy. Secondly, the constant difference between the replay-only variant and the complete ASRNS system allows isolating the effect of EWC-based regularisation: replay only reintroduces past examples but imposes no constraints on the extent of parameter changes and therefore allows forgetting, while EWC-based Fisher-information penalty directly protects those parameters which were critical to learn the patterns of the previous patient [5][7]. Thirdly, the additional delay of the proposed system by 158 ms vs 142 ms of the static baseline remains small enough for the needs of real-time monitoring and is comparable to delays of other edge-deployed CNN-LSTM healthcare models [1].

Such results must be viewed with due care. First, the test data set is artificial and, despite its statistical soundness from existing publicly available sources, does not reflect the entire diversity of noise in real-life clinics, co-morbidities, and device variations. Next, the drift detection threshold value and the size of the replay buffer have been determined in this work and will have to be clinically calibrated in an actual deployment. Lastly, the validation gate presented in section 3.5 has been tested only on synthetic data sets; a real-life clinician's opinion would certainly be needed next.

5. Conclusion

In this paper, Introduced the ASRNeural system, which can perform continual learning in healthcare monitoring by means of a combined CNN-LSTM inference model, a streaming drift detector, selective experience replays and Elastic Weight Consolidation in one self-updating system that updates its models when required and, in a forgetting, -resistant fashion. On a synthetic stream based on real-world PhysioNet data with two concept drift events, our system achieved an accuracy of more than 90 percent throughout operation and recovered from all drifts fully, greatly outperforming a static baseline and a replay-only version of the ASRNeural, while introducing only small inference latency overhead. Shown that the combination of autonomous concept drift detection with continual learning can be used to build practical and efficient healthcare monitoring systems that maintain high accuracy over long periods of use without requiring any retraining. Future research will consist in the validation of the framework on longitudinal patient data, federated and privacy-preserving refinements at multiple care facilities, implementation of explainability methods to allow clinicians to understand how the system works, and adaptation of the drift threshold to the criticality of the monitored state.

References

1. Prabha, M., Nandhini, S., Dayanidhy, M., & Pradeep, R. (2025). Edge-AI integrated secure wireless IoT architecture for real-time healthcare monitoring and federated anomaly detection. *Scientific Reports*. https://doi.org/10.1038/s41598-025-30150-x
2. Yuan, N. (2021). Self-correcting recurrent neural network for acute kidney injury prediction in critical care. *Health Data Science*. https://doi.org/10.34133/2021/9808426
3. Parisi, G. I., Kemker, R., Part, J. L., Kanan, C., & Wermter, S. (2019). Continual lifelong learning with neural networks: A review. *Neural Networks, 113*, 54–71. https://doi.org/10.1016/j.neunet.2019.01.012
4. Shan, H., Li, Q., & Sompolinsky, H. (2026). Order parameters and phase transitions of continual learning in deep neural networks. *Proceedings of the National Academy of Sciences, 123*(6), e2501899123. https://doi.org/10.1073/pnas.2501899123
5. Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A. A., Slivnick, K., Quaedvlieg, J., Ramalho, T., Grabska-Barwińska, A., Hassabis, D., Clopath, C., Kumaran, D., & Hadsell, R. (2017). Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences, 114*(13), 3521–3526. https://doi.org/10.1073/pnas.1611835114
6. Li, Z., & Hoiem, D. (2017). Learning without forgetting. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 40*(12), 2935–2947. https://doi.org/10.1109/TPAMI.2017.2773081
7. Moses, S. N., & Ryan, J. D. (2006). A comparison and evaluation of the predictions of relational and conjunctive accounts of hippocampal function. *Hippocampus, 16*(1), 43–65. https://doi.org/10.1002/hipo.20131
8. Hadsell, R., Rao, D., Rusu, A. A., & Pascanu, R. (2020). Embracing change: Continual learning in deep neural networks. *Trends in Cognitive Sciences, 24*(12), 1028–1040. https://doi.org/10.1016/j.tics.2020.09.004
9. Shiranthika, C., Saeedi, P., & Bajić, I. V. (2023). Decentralized learning in healthcare: A review of emerging techniques. *IEEE Access, 11*, 54188–54209. https://doi.org/10.1109/ACCESS.2023.3281832

10. Pomponi, J., Scardapane, S., Lomonaco, V., & Uncini, A. (2020). Efficient continual learning in neural networks with embedding regularization. *Neurocomputing, 397*, 139–148. https://doi.org/10.1016/j.neucom.2020.01.093
11. Zhang, B., Guo, Y., Li, Y., He, Y., Wang, H., & Dai, Q. (2021). Memory recall: A simple neural network training framework against catastrophic forgetting. *IEEE Transactions on Neural Networks and Learning Systems, 33*(5), 2010–2022. https://doi.org/10.1109/TNNLS.2021.3099700
12. Nguyen, H. T., Sehwag, V., Hosseinalipour, S., Brinton, C. G., Chiang, M., & Poor, H. V. (2020). Fast-convergent federated learning. *IEEE Journal on Selected Areas in Communications, 39*(1), 201–218. https://doi.org/10.1109/JSAC.2020.3036952
13. Spicher, N., Klingenberg, A. S., Purrucker, V., & Deserno, T. M. (2021). Edge computing in 5G cellular networks for real-time analysis of electrocardiography recorded with wearable textile sensors. In *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)* (pp. 1735–1739). IEEE. https://doi.org/10.1109/EMBC46164.2021.9630102
14. Moody, G. B., & Mark, R. G. (2001). The impact of the MIT-BIH arrhythmia database. *IEEE Engineering in Medicine and Biology Magazine, 20*(3), 45–50. https://doi.org/10.1109/51.932724
15. Johnson, A. E. W., Pollard, T. J., Shen, L., Lehman, L. W. H., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Celi, L. A., & Mark, R. G. (2016). MIMIC-III, a freely accessible critical care database. *Scientific Data, 3*, Article 160035. https://doi.org/10.1038/sdata.2016.35