Federated Learning for Network Intrusion Detection in Ambient Assisted Living Environments

Ana Cholakoska, Hristijan Gjoreski, Valentin Rakovic, Daniel Denkovski and Marija Kalendar Faculty of Electrical Engineering and Information Technologies, Ss. Cyril and Methodius University in Skopje, North Macedonia

Bjarne Pfitzner and Bert Arnrich

Digital Health - Connected Healthcare, Hasso Plattner Institute, Potsdam, Germany

Abstract—

Given the Internet of Things rapid expansion and widespread adoption, it is of great concern to establish secure interaction between devices without worsening the quality of their performance. Using machine learning techniques has been shown to improve detecting anomalous behavior in these types of networks, but their implementation leads to poor performance and compromised privacy. To better address these shortcomings, federated learning is being introduced. It enables devices to collaboratively train and evaluate a shared model while keeping personal data on-site (e.g., smart homes, intensive care units, hospitals, etc.), thus minimizing the possibility of an attack and fostering real-time distribution of models and learning. The paper investigates the performance of federated learning in comparison to deep learning, with respect to network intrusion detection in ambient assisted living environments. The results demonstrate comparable performances of federated learning with deep learning, while achieving improved data privacy and security.

THE INCREASE in sensors, cloud and big data analytics, as well as the need to automate and ease processes, is contributing to the fast-paced development of Internet of things (IoT) networks. IoT facilitates many diverse functions that are provided for the household, industry, infrastructure, transportation, by a massive number of unique devices from diverse manufacturers. It has also received attention from the medical community as a promising way for early diagnosis, prevention, treatment, and administration of drugs while the patients remain in the comfort of their own homes

[1].

Nevertheless, the diversity of sensors and devices creates a number of issues, primarily related to security and privacy. As health and personal information become remotely obtainable, the risk of violating the privacy of patients' data and their electronic health records increases substantially. Also, these devices can be reconfigured or turned off [2], which may contribute to severe consequences for the patient's health. As most network intrusion detection systems (NIDS) cannot fully provide the necessary protection for IoT networks because of the ever increasing pace of new attack types and methods [3], novel ways to detect potential anomalies must be sought.

Recently, machine learning (ML) algorithms have been applied for NIDS, and have shown good results in detecting such anomalies. Because IoT devices are limited in storage, power, and cannot apply complex artificial neural networks (NNs), it is necessary to have a central server which will process the data. The centralized approach introduces several limitations, most notably sharing patients' data and compromising their privacy. Federated learning (FL) is showing great potential as a new distributed variant that can speed up the detection and handling of network anomalies without compromising patient data and maintaining privacy intact. However, the FL-based solutions lack the accuracy, robustness and ubiquity compared to their centralized learning counterparts. Moreover, the number of works focusing on FL-based network intrusion in ambient assisted living (AAL) environments, is not prevalent in the literature.

This paper analyzes the aspect of anomaly detection in AAL environments regarding network intrusion by utilizing Federated Learning. It also performs parameter characterization and attack grouping in order to improve the models accuracy and robustness. The structure of the paper is as follows. Section II discusses the state of the art in FL and Deep Learning (DL) approaches for anomaly detection in IoT networks. Section III presents the used dataset, the experimental and evaluation setup, as well as the performance metrics of interest. Section IV gives an overview of the experiments and obtained results and compares the FL and DL models. It also discusses how grouping attacks and parameter characterization can improve the overall accuracy of the models. Section V summarizes the paper and presents possible future directions and improvements regarding the given problem.

RELATED WORK

Conventional signature-based techniques focus on detecting already known and established patterns, while network intrusion detection techniques can detect both known and unknown attacks. This implies that network intrusion detection demands more computational power and achieves lower overall accuracy in the process. Recent research has shown that leveraging different ML and DL algorithms for network anomaly detection purposes is highly beneficial for building more adaptable and accurate intrusion detection systems.

Saheed et al. [4] suggest an ML-supervised algorithm-based IDS for IoT networks. After performing normalization and dimensionality reduction on the UNSW-NB15 dataset, six different ML models were trained. All models present an accuracy in detection of 99%. The authors in [5] also evaluated the possibility of using different ML algorithms to detect security attacks in medical devices. The results show that the decision tree-based algorithms achieve the highest detection accuracy (~90%). Intelligent and dynamic ransomware spread detection in medical cyberphysical systems was the topic of interest in [6]. In this research, two different ML models have proved to be successful in detecting and classifying these types of attacks, with Naive Bayes (NB) obtaining an accuracy of 99.99%. Otoum et al. in [7] as well as the paper in [8] present DL-based solutions which tackle IDS systems for IoT networks. The first use a Spider Monkey Optimization algorithm (SMO) and Stacked-Deep Polynomial Network (SDPN), achieving an overall accuracy of 99.02%, while the second use deep Q-learning-based neural network with privacy preservation method (DQ-NNPP) and achieve an accuracy of 93.74%.

However, these algorithms also have their drawbacks, mainly because of the centralized approach. Having the entire dataset on one server can be computationally expensive and timeconsuming. In addition, data signatures can be very large in size, so it can be very difficult to collect the data in an efficient and real-time manner. In most network intrusion and detection scenarios, swift detection is of utmost importance. Moreover, this can compromise security and privacy when transferring data from IoT nodes to the server and vice versa. As such, FL [9], which enables distributed training of models, has emerged as a potential and adaptable strategy that can address these drawbacks.

The authors in [10] designed LocKedge, an FL-based IDS for IoT networks, which detects anomalies at the edge layer. Nevertheless, when evaluated on the BoT-IoT dataset, the FL model achieved lower performances than a DL model. Rahman et al. [11] tried to keep data privacy intact, while suggesting a new FL-based system for IoT intrusion detection. However, the evaluation process on the NSL-KDD dataset shows an oscillating accuracy of around 83.09%, which is not substantial for real-time IDS purposes. The authors in [12] utilize homomorphic encryption, as well as a convolutional neural network for the development of a distributed IDS system based on FL. The model is tailored to analyze and block only DDoS traffic on satellite-terrestrial networks. The authors in [13] also develop an FL model to deliver real-time anomaly detection of DDoS attacks in IoT networks, which is based on the gated recurrent unit (GRU) concept. Both models in [12], [13] exhibit high accuracies, but are tailored only for a specific type of attack, and lack ubiquitous applicability.

The related FL works primarily focus on a limited number of attacks, such as DDoS, which significantly limits their applicability to real-world scenarios. This is highly detrimental for classification purposes, since anomaly detection systems require diverse and updated data in order to foster high accuracy and robustness. This work presents a novel FL solution based on anomaly detection that builds upon the weakness of the state of the art works. It achieves satisfactory performances for a large plethora of IoT-based attacks, and it is comparable to the results achieved by DL. Also, to the best of the authors' knowledge, this work is the first to focus on anomaly detection-based NIDS in AAL. Additionally, the paper presents a novel idea for grouping attacks based on their similarity, which can significantly improve the performance of the FL-based network intrusion detection (above 98%), while preserving the detection capabilities for different types of attacks.

DATASET AND METHODOLOGY

This section provides insight related to the dataset of interest. It also gives a thorough explanation of the system architecture, as well as the design of the DL model. Moreover, it introduces the specific performance metrics of interest.

Used dataset

For the purpose of this research, we used a publicly available dataset called "IoTID20" [14]. The testbed is a typical AAL environment, which includes a camera, a smartphone, a home speaker (AI speaker) and several computers. By simulating network traffic and monitoring at different time periods, researchers were able to create the dataset and then extract 83 network features from the pcap files using Wireshark [15].

The distribution between the normal network traffic and the different types of anomalies is as follows. The Normal category is about 28%, compared to the rest of the dataset. The rest of the dataset contains attacks normally found in an IoT environment, such as: Mirai UDP flooding (contributing with 183,554 instances or 22.5%), Mirai Hostbruteforce (121,181 instances or 14.8%), DoS SYN flooding (59,391 instances or 7.3%), Mirai HTTP flooding (55,818 instances or 6.8%), Mirai ACK flooding (55,124 instances or 6.8%), Scan Port OS (53,073 instances or 6.5%), Man in the middle (MITM) ARP spoofing (35.377 instances or 4.3%), Scan Hostport (22,192 instances or 2.8%) and Bot (1,966 instances or 0.2% of the whole dataset).

System architecture and NN model

The proposed FL system architecture is given in **Figure 1a**. It is constructed of two main components; i) FL clients (AAL environments) and ii) central server. The FL clients train a local model on-site using their local data. After specific number of local epochs the FL clients send the trained models (i.e., model weights) to the central server. The central server aggregates the received models by averaging each model weight across all clients, known as Federated Averaging (FedAvg) strategy. The updated model is then sent back

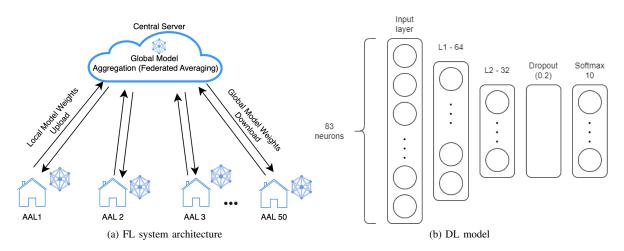


Figure 1. FL system architecture and data flow (left), and DL model architecture (right)

to the clients, completing one FL round. The process is repeated for a number of rounds, until achieving the required performances or model convergence. As such, the FL approach does not share nor exposes any information from the AAL dataset and environment, fostering high level of privacy.

It is assumed that all clients have the same NN model (**Figure 1b**) and use the same number of local epochs. The model consists of a feed-forward neural network (FFNN) with two fully-connected layers with 64 and 32 neurons, respectively. Both layers utilize the ReLU activation function. The two layers are followed by a dropout layer with 0.2 rate. The output layer is a softmax layer consisting of 10 neurons, which represent the classes of attacks in the dataset. In the experiments where attack grouping is performed, the number of neurons in the output layer is reduced to 7.

Evaluation setup and metrics of interest

The dataset consists of one normal traffic data class and 9 different types of attacks, resulting in 10 classes in total. The data is further split into a training and test subset. The training subset contains 80% of the data, while the remaining 20% are present in the test subset. The evaluation does not incorporate any tuning of the neural network parameters, so no validation dataset is necessary.

For performance comparison, we use a DL baseline. From an information theoretical per-

spective no FL model can achieve higher accuracy than a centralized DL model, when the FL is using the same underlying neural network. The reason is related to the manipulation with the dataset. Specifically, the DL model is trained on the whole dataset, while the FL trains the local models on portions on the dataset and then aggregates them into a global model. Hence, loosing valuable information due to the partitioning and averaging. The DL model is based on the same FFNN from **Figure 1b**. In the DL experiments, the training dataset is used for training and the test dataset is used for evaluation. We used a maximum of 35 epochs to train the DL model.

For the FL, the experiments are executed with a different dataset distribution because of the nature of the FL itself - no data leaves the device. Therefore, the complete data is split among 50 clients (in our case, each client refers to an AAL environment), where every client holds a different portion of the test and training dataset (see **Figure 1a**). The training subset of each client is used to train the local models. The global model is evaluated (in each round) using the combined test subsets from all the clients.

In each round of the FL, a subset of random clients is selected for local training, controlled by the fraction fit parameter. Each of the clients uses only 5 epochs for the training of the local DL models. As mentioned, the FedAvg optimizer is used for the aggregation of the local DL models into the global FL model, which comes as a sim-

ple, yet effective solution. After each round, the aggregated global model weights are distributed to the clients and used as starting point for the local model training in the next round. In the experiments, we use a maximum of 35 rounds for training of the global FL model.

By careful analysis of the dataset, it can be concluded that many attacks are highly related by their type and inherent family features. For example, there are several distinct Mirai attacks that exhibit very similar network intrusion behavior. As the primary goal of network intrusion detection systems is to accurately and timely detect attacks, it can be highly beneficial if the system can group the attacks and improve its detection capabilities. Since the grouping will be done over the same family of attacks, the system will still be able to identify the type of attack, however, its granularity will be coarser.

The performance metrics of interest in this study is the models' accuracy as a function of the number of epochs/rounds required to finish the training. Specifically, the evaluation focuses on the FL's accuracy in dependence of the number of FL rounds as well as the fraction fit parameter (i.e. percentage of FL clients used in each round).

RESULTS AND DISCUSSION

In this study, we conducted three experiments in order to investigate the capabilities and limitations of the FL model for anomaly detection in AAL environments. The first experiment focuses on training, testing and comparing of the FL model with the baseline DL model. In this experiment we focus on classification performances for all 10 available classes in the dataset. The second experiment serves to examine the benefits of attack grouping with the aim to improve the detection performances of the FL model. Finally, the last experiment is concentrated on the parameter characterization of the FL models.

In the head-to-head comparisons between the DL and the FL models concerning the convergence, we associate training epochs for the DL model with training rounds for the FL model. Even though this may appear to be unfair, since the FL additionally uses 5 epochs for the training of the local models, the local models are trained on a significantly smaller dataset portions (1/50). Models accuracy

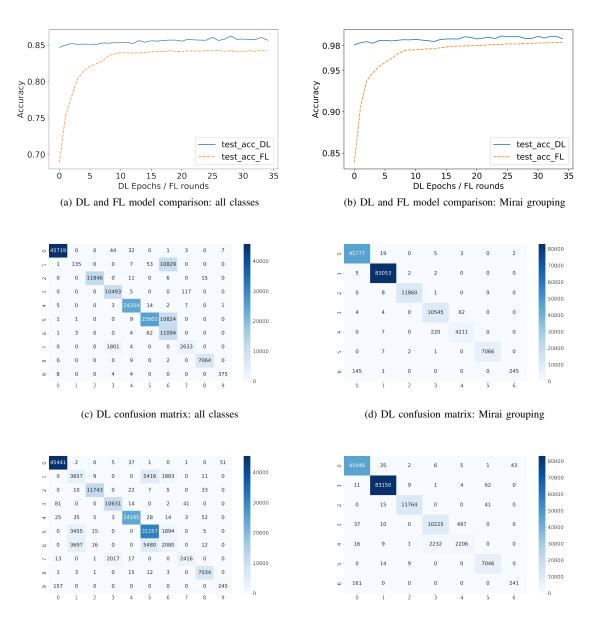
The first experiment is used to compare the accuracy and convergence performances of the FL model with the baseline DL model. In this experiment, the fraction fit parameter is fixed to 1.0, meaning that all clients are participating in each round of the FL training process.

Figure 2a depicts the models' accuracy in dependance of the number of epochs (for DL) and rounds (for FL). It can be seen that the FL model achieves a slightly lower accuracy ($\sim 84\%$) compared to the DL model ($\sim 86\%$) for the test dataset. Furthermore, it can be noted that after the 20th round, the FL model seems to achieve its convergence. On the contrary, the DL model still tends to improve its accuracy as the number of epochs increases, but it encounters slow convergence and a higher performance variability (model instability). This result clearly shows the benefits of using FL for anomaly detection in AAL scenarios. At the price of slight classification performance decrease, one can preserve the user privacy in these scenarios, as the FL model does not share and expose the AAL dataset, only the model weights. Furthermore, the FL provides better stability (mostly due to the FFNN weights averaging) and faster convergence.

Figures 2c and **2e** show the confusion matrices for the DL and the FL models, respectively. The results show that most of the misclassifications of both models occur between classes 1, 4, 5 and 6, which correspond to similar types of attacks, i.e., the Mirai attacks. Intuitively, this indicates that grouping the Mirai types of attacks into one class would improve the anomaly classification performances.

Attack grouping

The second experiment groups the four Mirai classes into one class, leaving the dataset with seven distinct classes. This also decreases the number of nodes in the final layer of the FFNN. Same as in the first experiment, the fraction fit parameter for the FL model is set to 1.0. Figure **2b** shows the head-to-head comparison of the DL and the FL model in terms of accuracy and convergence when applying the Mirai attacks grouping, while Figures **2d** and **2f** show the DL and FL confusion matrices for this case, respectively.



(e) FL confusion matrix: all classes

(f) FL confusion matrix: Mirai grouping

Figure 2. DL and FL models experiments. For confusion matrices c) and d) the classes go as follows: class 0 represents normal traffic; classes 1, 4, 5 and 6 represent the distinct Mirai attacks; class 2 - DoS; class 3 - Scan Port OS; class 7 - Scan Hostport; class 8 - MITM; class 9 - Bot. For confusion matrices d) and f) classes 1, 4, 5 and 6 have been grouped to class 1 (Mirai).

The results clearly show that the attack grouping substantially improves the accuracy of both DL and FL models, and it is also noticeable that the grouping has more benefits for the FL model. Specifically, the performance difference between the DL and FL models is smaller, compared to the case when there is no attack grouping. Furthermore, the FL model tends to improve its performance even after the 20th round. The reason behind this behavior can be found reconsidering the confusion matrices in Figures 2c and 2e (without the Mirai grouping). It is evident that the FL model is more affected by misclassifications between the Mirai types of attacks. In specific, due to the similarity between these attacks and the substantially smaller datasets (1/50), the local models fail to learn the differences between the Mirai classes. Therefore, the grouping of the multiple Mirai classes into one, results in more substantial performance gain for the FL model.

FL parameter characterization

The final experiment focuses on the parameter characterization of the two FL models, i.e., the FL model using all classes and the FL model using the Mirai grouping. Besides the number of rounds, this experiment also investigates the fraction fit parameter and its contribution to the accuracy and the convergence of the models. The fraction fit is an important parameter in federated learning, since it controls the client selection and the stochasticity of the learning process. Randomly choosing a subset of clients in each training round reduces the computation and communication overhead and can reduce the overfitting in the resulting global FL model. The results are obtained for three fraction fit parameter values: 0.2, 0.6 and 1.0. In particular, a fraction fit of 0.2 means that in each round of the federated learning, only 20% randomly chosen clients participate in the FL training (i.e., 10 randomly selected clients out of 50 in our case).

Figures 3a and 3b show the convergence and accuracy results for both FL models, with respect to the fraction fit parameter. Similar behavior can be observed for both FL models, i.e., the fraction fit does not significantly impact the accuracy and convergence performances for the chosen problem of anomaly classification. Only minor differences (<0.3%) can be seen between the

different choices of fraction fit. However, there are few important considerations that should be noted. A smaller fraction fit provides a slightly better convergence rate for smaller number of FL rounds (<10). A fraction fit of 0.6 provides the best accuracy when the number of FL training rounds is above 15, e.g., the FL model with Mirai grouping achieves accuracy of 98.3% at round 25. The fraction fit of 1.0 seems to experience some minor model overfitting. In conclusion, the results clearly show that there is an optimal fraction fit in the trade-off between accuracy, convergence and FL overhead.

CONCLUSION

This paper discusses the applicability of FL for network intrusion detection for AAL environments. The paper also introduces the concept of attack grouping in order to improve the overall detection performance of the FL models. The analysis show that FL achieves very similar performances to its DL counterpart, without sharing any personal and patient's data. Additionally, the results show that the attack grouping significantly improves the detection accuracy of both DL and FL, with FL having a larger benefit from the grouping process.

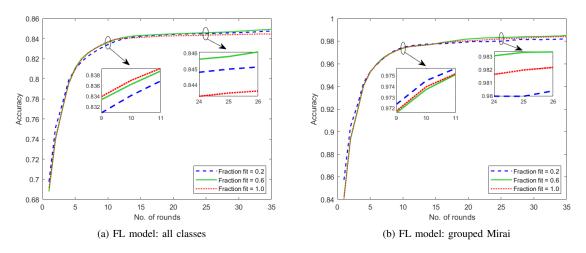
Further work will include implementing some security mechanisms (e.g. differential privacy) to the FL models and evaluating the trade-off between privacy and accuracy. New federated learning optimizers can also be tested and evaluated on the same and new AAL datasets. Another potential venue for future exploration is the system level specifics of FL, with respect to bandwidth efficiency, noisy data and computational overhead.

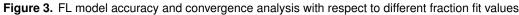
ACKNOWLEDGMENT

This work has been supported by the Wide-Health project - European Union's Horizon 2020 research and innovation programme under grant agreement No. 952279.

REFERENCES

 E. Adeniyi, R. Ogundokun, A.J. Bamidele, "IoMT-Based Wearable Body Sensors Network Healthcare Monitoring System," *IoT in Healthcare and Ambient Assisted Living*, 2021.





- M.M.U. Rehman, H.Z.U. Rehman, and Z.H. Khan. "Cyber-Attacks on Medical Implants: A Case Study of Cardiac Pacemaker Vulnerability," *IJCDS Journal*, vol. 10, Jul. 2020.
- A. Cholakoska, M. Karanfilovska, D. Efnusheva, "Survey of Security Issues, Requirements, Challenges and Attacks in Internet of Things", *Informatics and Cybernetics in Intelligent Systems*, vol. 228, 2021.
- Y. Saheed, et al., "A machine learning-based intrusion detection for detecting internet of things network attacks", *Alexandria Engineering Journal*, vol. 61, 2022.
- S. Gao, G. Thamilarasu, "Machine-Learning Classifiers for Security in Connected Medical Devices", *ICCCN* 2017, Sep. 2017.
- L. Fernández Maimó, et al., "Intelligent and Dynamic Ransomware Spread Detection and Mitigation in Integrated Clinical Environments", *Sensors*, vol. 19, no. 5, Mar. 2019.
- Y. Otoum, D. Liu, A. Nayak, "DL-IDS: a deep learning-based intrusion detection framework for securing IoT", *Transactions on ETT*, vol. 33, no.3, Mar. 2022.
- N. D. Kathamuthu, et al., "Deep Q-Learning-Based Neural Network with Privacy Preservation Method for Secure Data Transmission in Internet of Things (IoT) Healthcare Application," *Electronics*, vol. 11, no. 1, Jan. 2022.
- H. B. McMahan, et al., "Federated learning of deep networks using model averaging,", *CoRR*, 2016.
- T. T. Huong, et al., "LocKedge: Low-complexity cyberattack detection in IoT edge computing", *IEEE Access*, vol. 9, 2021.
- S. A. Rahman, et al., "Internet of Things intrusion detection: Centralized, on-device, or federated learning?",

IEEE Network, vol. 34, no. 6, Nov. 2020.

- K. Li, et al., "Distributed network intrusion detection system in satellite-terrestrial integrated networks using federated learning", *IEEE Access*, vol. 8, 2020.
- V. Mothukuri, et al., "Federated learning-based anomaly detection for IoT security attacks", *IEEE IoT Journal*, vol.9, no.4, May, 2021.
- I.Ullah and Q.H.Mahmoud, "A Scheme for Generating a Dataset for Anomalous Activity Detection in IoT Networks," *Advances in Artificial Intelligence*, May 2020.
- U. Banerjee, A. Vashishtha, and M. Saxena, "Evaluation of the Capabilities of WireShark as a tool for Intrusion Detection," *IJCA Journal*, vol. 6, no. 7, Sep. 2010.

Ana Cholakoska is a research and teaching assistant at the Faculty of Electrical Engineering and Information Technologies, Ss. Cyril and Methodius University in Skopje. Currently, she is part of the FEEIT team working on the WideHealth project, developing FL methods for anomaly detection in AAL environments, as well as improving privacy in IoT environments. Her interests include cybersecurity and network analysis. Contact her at acholak@feit.ukim.edu.mk.

Bjarne Pfitzner is a research assistant and doctoral student at the Digital Health - Connected Healthcare group at the Hasso Plattner Institute (HPI) in Germany. He did his M.Eng. degree in Computing at Imperial College London. Over the last four years, he worked in the area of federated machine learning with a focus on privacy-preserving algorithms using differential privacy and healthcare applications such as medical imaging and risk stratification for the intensive

care unit. Contact him at bjarne.pfitzner@hpi.de.

Prof. Dr. Hristijan Gjoreski received his Ph.D. degree in ICT from the Jozef Stefan Postgraduate School in Slovenia in 2015. From 2010 to 2016, he was a researcher at the Jozef Stefan Institute in Slovenia. In 2017 he was a Postdoc at the University of Sussex, United Kingdom. Currently he is an Associate Professor at the Ss. Cyril and Methodius University in Skopje, N. Macedonia. His research interests include Artificial Intelligence, Machine Learning, Wearable Computing. Contact him at hristijang@feit.ukim.edu.mk.

Prof. Dr. Valentin Rakovic currently holds the position of associate professor and the head of the Laboratory for Wireless and Mobile Networks at the Faculty of Electrical Engineering and Information Technologies (FEEIT), Ss Cyril and Methodius University in Skopje. He received his Dipl.-Ing., M.Sc. and Ph.D. degree in Telecommunications at the Faculty of Electrical Engineering and Information Technologies, Ss Cyril and Methodius University (UKIM) in Skopje, in 2008, 2010 and 2016 respectively. His research work focuses on the areas of wireless networks, signal processing, optimization theory, machine learning as well as prototyping of wireless networking solutions. Contact him at valentin@feit.ukim.edu.mk.

Prof. Dr. Daniel Denkovski is an Associate Professor at the Faculty of Electrical Engineering and Information Technologies, Ss. Cyril and Methodius University in Skopje. His major research interest is concentrated on signal processing, information theory, wireless communications, cloud computing, and recently machine learning and federated learning and their application in different domains. He has notable research experience having worked on 12 internationally funded research projects (FP7, H2020 and NATO SpS), as well as several domestic projects in his research areas. Contact him at denield@feit.ukim.edu.mk.

Prof. Dr.-Ing. Bert Arnrich is head of the Chair Digital Health - Connected Healthcare at the joint Digital-Engineering Faculty of Hasso Plattner Institute (HPI) and the University of Potsdam. He has been a PI in several European and national research projects. At ETH Zurich he established and headed the research group Pervasive Healthcare in the Wearable Computing Laboratory. He received a Marie Curie Cofound Fellowship from the European Union and was appointed to tenure track professorship at the Computer Engineering Department at Bosporus University. He worked as a Science Manager for Emerging Technologies at Accenture Technology Solutions. Contact him at bert.arnrich@hpi.de.

Prof. Dr. Marija Kalendar is a full professor at Faculty of Electrical Engineering and Information Technologies, University Ss Cyril and Methodius in Skopje, N. Macedonia, where she received her B.Sc. (2002), M.Sc. (2007) and Ph.D. (2011) degrees. She is currently the President of the ETAI Macedonia society and member of the Steering Committee of the Engineering Institution of Macedonia. Her research interests include communication networks and protocol design, IoT devices, systems and protocols design, (RT)OS, Cloud and HPC systems. Contact her at marijaka@feit.ukim.edu.mk.