Executive Summary

Second reporting period (1st May 2020 – 30th June 2021)

The digital revolution, in particular big data and artificial intelligence (AI), offer new opportunities to transform healthcare. However, it also harbors risks to the safety of sensitive clinical data stored in critical healthcare ICT infrastructure. In particular data exchange over the internet is perceived as insurmountable, posing a roadblock hampering big data-based medical innovations. FeatureCloud's transformative security-by-design concept will minimize the cyber-crime potential and enable secure cross-border collaborative data mining endeavors. FeatureCloud will be implemented into a community-extendible software platform for substantially reducing cyber risks to healthcare infrastructure by employing the world-wide first privacy-by-architecture approach, which has two key characteristics: (1) no sensitive data is communicated through any communication channels, and (2) data is not stored in one central point of attack. Federated machine learning (for privacy-preserving data mining) integrated with relevant privacy-enhancing technology, like secure multiparty computation or differential privacy, will safely apply next generation AI technology for medical purposes. Our ground-breaking new cloud-AI infrastructure only exchanges learned model representations which are anonymous by default. Collectively, our highly interdisciplinary consortium, from IT to medicine covers all aspects of the value chain: assessment of cyber risks, legal considerations and international policies, development of federated AI technology, extendible app store and user interface design, implementation as prognostic medical devices, evaluation and translation into clinical practice, commercial exploitation, as well as dissemination and patient trust maximization. FeatureCloud's goals are bold, necessary, achievable, and paving the way for a socially agreeable big data era of the Medicine 4.0 age.

Work performed from the beginning of the project to the end of the period covered by the report and main results achieved so far

We implemented a first version of the FeatureCloud app store including software development packages for corresponding computer-computer interfaces running as web server and fostering an advanced and user-extendible app store functionality. It was developed using documented bi-weekly platform developer online conferences with all source code stored in Git repositories. Software development was organized into sprints and extended by two app development hackathons to test and improve the platform. More than 20 apps for federated computations have been developed, from the computation of standard statistics like t-tests or regression analyses to advanced apps on federated principle component analysis (PCA), artificial neural networks, random forest classifiers and survival time predictors. An app development and testing environment was deployed as well as a full documentation to aid app developers. In total, we also prepared five live demos to illustrate FeatureCloud's capabilities developed in the first two periods of FeatureCloud. Using the first FeatureCloud apps, we have worked on demonstrating the power of federated machine learning coupled to relevant related privacy-enhancing technologies. Specifically, we have worked on typical medical application scenarios. We began with a federated genome-wide association study (GWAS) tool: sPLINK, which mimics the non-federated standard GWAS tool PLINK. We demonstrate that currently available distributed GWAS software (so-called meta-analysis tools) massively loses accuracy when the data suffers heterogeneously distributed outcomes or confounders. In contrast, sPLINK gives the exact same results as PLINK, and thus has the potential to become the new standard tool for genotyping in the future as it does not require any exchange of raw data between the participating institutions/hospitals and on top is not suffering any accuracy loss compared to the state-of-the-art centralized tools. sPLINK implements federated Chi-squared tests, as well as federated multimodal linear and logistic regression models coupled to secure multiparty computation (SMPC). Likewise, we have developed the first prototype software for federated survival analysis: PARTEA. It combines federated statistical modelling and differential privacy approaches based on Laplacian noise to generate privacy-preserving Kaplan-Meier plots. A corresponding paper is under review. In addition, we developed, evaluated and published a federated gene expression data analysis tool called "Flimma". The FeatureCloud AI store itself was also evaluated and tested in real-world scenarios, and a corresponding publication is under review (see preprint).

Progress beyond the state of the art and expected potential impact (including the socio-economic impact and the wider societal implications of the action so far)

FeatureCloud contributes significantly to all three expected impacts mentioned in the work programme:

- **Improved security of Health and Care services, data and infrastructures**
  - By addressing the evident roadblock in medical data mining – centralized data mining but distributed clinical data – we improve the cyber security of computational health care services, patient data and communication infrastructure. FeatureCloud's federated machine learning and SMPC engines erase the necessity to share sensitive data with a cloud.
- **Less risk of data privacy breaches caused by cyberattacks**
  - FeatureCloud significantly reduces the risk of data privacy breaches caused by cyberattacks on health cloud services or on the communication channels between hospital and cloud. Instead of bringing the data to the AI, we bring the AI to the data.
- **Increased patient trust and safety**
  - Based on trusted authority technology, like blockchains, we work on ensuring full control over the access rights to own sensitive data combined with the guarantee that no sensitive data is exchanged to learn the federated AI which could be traced back to individual patients. This strategy will increase patient trust and safety significantly. Our FeatureCloud platform is designed to be in accordance with E.U. GDPR and NISD policies, and it is developed with respect to the criteria for software-supported medical devices of the FDA and EMA, respectively.

FeatureCloud furthermore contributes to the following most significant impacts not mentioned in the work programme:

- **The novel FeatureCloud technology will create new market opportunities.**
  - The FeatureCloud's app store for client-side machine learning tools will have an enormous impact worldwide and foster pan-European business, e.g. with spin-offs and start-ups because of a huge emerging market in privacy-aware machine learning.
- **The European society will benefit from new levels of personalized medicine, new possibilities for research of complex diseases like e.g., cancer, and lower costs of medical research.**
  - FeatureCloud enables open science without boundaries, cross-domain and pan-European, which will particularly allow new levels of cancer research because FeatureCloud apps address current privacy, ethical, security, and safety restrictions at the core, and will thus reduce increasing health costs in Europe by rising medical quality at the same time.

FeatureCloud Acknowledgement

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 826078. This reflects only the author's view and the European Commission is not responsible for any use that may be made of the information it contains.

Address (URL) of the action's public website

https://featurecloud.eu/