REACH

NEXT GENERATION DATA INCUBATOR

This project has received funding from the European Union's H2020 research and innovation programme under Grant Agreement no 951981

CORE PARTNERS

DATA PROVIDERS

1 ANNEX I. Technical Specification Double-side Page

1. **TECHNICAL SCOPE:** The mock-up solution is suitable and correctly addresses the challenge/theme selected over the REACH dataset/s. The Big Data solution architecture proposed is adequate to tackle the data management issues associated to the solution in mind. "To what extent does the applications handle the data provided?"

Spirometry is a common clinical test performed in clinical environments in order to assess one's respiratory condition. In case of asthma, spirometry also serves as a diagnostic plus severity evaluation tool. Results from such tests have been labelled with **audio recordings** of patients performing respiratory tasks in front of a smartphone at different levels of respiratory impairments. This allows the usage of a machine learning solution in order to **predict spirometry** results from respiratory sounds, projecting pulmonary function values to a **vocal biomarker** domain, under the assumption that airflow dynamics impairments lead to **relevant acoustic changes**.

In this regard, a solution comprising a **respiratory sound detection tool** and a **spirometry-correlated estimator** could boost the use of such solution for telemonitoring of asthmatic patients, enabling a constant and periodic production of respiratory data from the user to provide also **trend analysis**, thus allowing more intervention time in case of an imminent worsening detection.

Finally, the user can periodically produce and download a "performance report" that summarises the results of the last period. This item can be further sent and shared with the reference doctor.

The Big Data architecture is based on a flexible and secure storage system that can store an unlimited amount of voice data thanks to the cloud infrastructure object storage solution. In addition, the application backend consists of multiple containerized modules that can be deployed using a container orchestrator, thus taking advantage of advanced load balancing and auto scaling mechanisms which enable the application to adapt to the available resources (in terms of bandwidth and computation), according to the number of parallel connections and concurrent users.

The application will handle the data provided to develop a model for predicting spirometry results and asthma worsening.

2. **SELECTION OF ALGORITHMS AND TOOLS:** The indicated Data Science approach, i.e. algorithms chosen, and Big Data architecture approach, i.e. tools chosen may successfully accomplish the required data governance, processing and analysis. A clear understanding of the used REACH dataset/s is demonstrated.

As the analysis pipeline can be splitted into two main components: respiratory sound detection and clinical acoustic analysis, datasets are going through the following steps:

1. *Labelling* (**Python**): a digital platform providing **data listening** and **data visualisation** to create metadata files regarding sound event occurrences, sound artefacts occurrences, noises and so on;
2. *Features extraction* (**Python**): a digital platform providing several techniques for respiratory sounds analysis to give the best physical/mathematical description of the data, following the respiratory-mechanics theory (i.e. **formant analysis**, **cepstral analysis** etc.);
3. *Feature selection* (**Python and Orange**): a digital platform providing several services as feature ranking, redundant features filtering, **correlation analysis**, **mutual information analysis** etc. in order to streamline and optimise the dataset's content for the best model building practice;
4. *Model investigation and production* (**Python, Orange**): a digital platform where to use many different machine learning algorithms (**neural network regression**, **support vector machine regression** etc.) in different combinations of their hyper-parameters to assess the best type/combination for the final model production.

Audio data from the CERTH health dataset from the REACH catalogue will be exploited for **data augmentation** purposes in order to build more **generalising models** and to **stress-test** the final solution.

Dockerized Big Data architecture can be deployed thanks to an orchestrator such as Kubernetes to offer capabilities useful for managing large datasets and fast and accurate analisys.

TECHNICAL SCALABILITY AND FLEXIBILITY OF THE SOLUTION: The solution can truly cope with humongous and increasing datasets, potentially from diverse data providers, and is flexible it to adapt to other related domains.

Every illustrated analysis-step exploits machine learning techniques that usually gains from **data heterogeneity** and **dataset size**. As both quantitative and heterogeneity increase, models can be updated to improve their prediction power. Specifically, in the case of respiratory diseases, the solution can be adapted to other cases besides Asthma: Chronic Obstructive Pulmonary Disease (COPD), sleeping apnea and generic dyspnea are a few examples.
Is worth noting that an increase in personal audio recordings would also enable state of the art prediction systems to accurately predict future days rather than just trends.
For instance, in the future, the solution could also use predictive modelling to guess the vocal biomarker value for the following "N" days using models such as **XGBoost**.

3. **DATA GOVERNANCE AND LEGAL COMPLIANCE:** Data sharing challenges, data governance and legal compliance, must be observed. The proposed solution is compliant with the current data legislations concerning security and privacy (e.g. GDPR).

With REACH, the solution is created based on an anonymous proprietary dataset from the Data Provider.

This solution will be integrated in an existing GDPR-compliant app and stored on EU based GDPR compliant cloud backend. All the data are stored in a secured cloud environment and are exchanged via encrypted connections (https, PGP, authenticated REST APIs with JWT token).

The user creates and manages their own account in the application. Any sharing of data with the doctors or other third-parties is initiated and controlled by the user directly (for example, the user shares a report from the app with their doctor). The user retains full control over their data at all times.

4. **QUALITY ASSURANCE AND RISK MANAGEMENT:** Feasible and credible quality process followed for the final product generation. The potential risks in all the phases of the project (design of the solution, development, testing, deployment…) are identified and convincing mitigation plans put in place.

A potential risk is to come up with **poorly-correlated** features due to the systemic distance between the lower airways and the device's microphone. The acoustical production is indeed a result of several filtering steps applied by the human cavities. A **backward inverse filtering technique** will be implemented in order to automatically model the human internal acoustic and fetch those lower components that are relevant for the biological problem.

Another risk is to find averagely less **standardised data** due to the user's performance when dealing with new incoming data (i.e. **wrong sound production**, uncompleted task, **too noisy environment** etc.). In this regard, the issue will be addressed by providing a **quality measurement tool** that checks the user's performance right before sending the data to further analysis and eventually corrects him to improve his acquisition.

Moreover, from an analysis perspective, regarding the final operative product, the final user could miss some data acquisition appointments leading to **missing data points** in his personal dataset. As a periodic acquisition is requested to enable a reliable trend analysis, several interpolation techniques will be introduced to fill the gap and still keep a satisfying performance (i.e. spline curves).

Finally, in order to ensure immutability, auditability and tamper-resilience for each user performance report generated by the system, the CERTH **Audit Messages Storage Platform** will be employed to store the SHA-3 hash of each report in the Blockchain network while the report itself will be stored off-chain in a managed database, so whenever a document is added to the system a transaction is recorded on the Blockchain.

REACH

NEXT GENERATION DATA INCUBATOR

This project has received funding from the European Union's H2020 research and innovation programme under Grant Agreement no 951981

CORE PARTNERS

DATA PROVIDERS