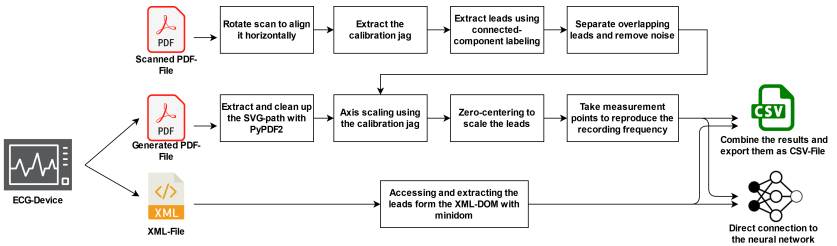
A Data Pipeline for Extraction and Processing of Electrocardiogram Recordings

Joshua Prim, Tim Uhlemann, Nils Gumpfer, Dimitri Grün, Sebastian Wegener, Sabrina Krug, Jennifer Hannig, Till Keller, Michael Guckert

Cognitive Information Systems, Technische Hochschule Mittelhessen, 61169 Friedberg, Germany

Aims: ECG devices used in clinical practice mostly generate PDF files or even paper printouts which cannot be directly processed by information systems or diagnostic algorithms. However, with the advent of deep learning (DL) methods yielding excellent results in detecting cardiac pathologies using electrocardiogram (ECG) data as input, availability of ECG recordings as training data gains in importance. This study aims to address this problem by developing a data-pipeline which generates machine-readable ECG data irrespective of the initial data format.
Methods: The pipeline can not only process data from modern digital ECG devices, e.g. in XML file format, but is also capable of extracting all necessary information from PDF files (both scanned hard copies and digitally generated PDFs). Different techniques were used to achieve this, including the adaption of open source libraries for computer vision technologies. The processed files from various sources are saved in CSV file format or can be processed directly with DL models.
Results: For validation the data-pipeline was applied to a set of 113 12-lead ECGs in PDF format from which CSV data was reconstructed. This was used for training a DL model that achieved comparable results with the same architecture trained on the original XML data.
Conclusion: Our ECG data-pipeline can accelerate ECG-based AI research and application of AI algorithms by providing access to ECG data irrespective of the format of available ECG recordings.



Overview of the Processing Pipeline.