Plan Overview

A Data Management Plan created using DMP Tool

Title: Closing the Gap Between Lab and Clinic: Validating Lightweight AI for ECG Arrhythmia Classification

Creator: Makabongwe Sibisi

Affiliation: Iie Varsity College

Funder: Digital Curation Centre (dcc.ac.uk)

Template: Digital Curation Centre

Project abstract:

This study will develop and validate a lightweight Recursive Depthwise Separable Convolutional Neural Network (RDSCNN) to bridge the gap between laboratory and real-world clinical performance in ECG arrhythmia classification. The model will be rigorously evaluated using the inter-patient paradigm and AAMI EC57 standards on diverse, publicly available datasets (MIT-BIH, PTB Diagnostic) to ensure generalizability. Performance will be assessed through accuracy, F1-score, and sensitivity metrics, alongside computational efficiency (parameters, FLOPs, inference time). By prioritizing clinically aligned validation and resource efficiency, this research will deliver a robust assistive tool to enhance diagnostic workflows in resource-constrained settings.

Start date: 02-28-2025

End date: 11-28-2025

Last modified: 06-13-2025

Closing the Gap Between Lab and Clinic: Validating Lightweight AI for ECG Arrhythmia Classification

Data Collection

The study will utilize raw ECG time-series signals obtained from standard leads, sourced from the MIT-BIH Arrhythmia Database (2-channel, 360 Hz) and the PTB Diagnostic ECG Database (15-lead, 1000 Hz). Alongside these signals, expert-provided annotations. Including R-peak locations, beat-level arrhythmia classifications, rhythm abnormalities, and relevant patient metadata (e.g., age, sex, clinical history) that will be collected as part of the original datasets. During preprocessing, these annotations will be programmatically remapped to align with the AAMI EC57 standard, categorizing heartbeats into five clinically recognized classes: Normal (N), Supraventricular Ectopic (S), Ventricular Ectopic (V), Fusion (F), and Unknown (Q). No new data will be generated; all outputs will be derived from the structured transformation of existing records.

This study will utilize existing, publicly available ECG datasets, specifically the MIT-BIH Arrhythmia Database and the PTB Diagnostic ECG Database, accessed primarily via PhysioNet and mirrored sources such as Kaggle. Raw ECG signals and expert-provided annotations, including R-peak locations and beat labels, will be retrieved in their native formats (e.g., .dat, .hea WFDB files). Using Python-based preprocessing scripts, these annotations will be mapped to standardized AAMI EC57 arrhythmia classes (N, S, V, F, Q), creating structured data without generating new ECG recordings. The resulting filtered signals and segmented beats will be organized into arrays and tensors using libraries like NumPy and Pandas, forming the input for deep learning model training and evaluation.

Documentation and Metadata

Ethics and Legal Compliance

Ethical issues will be managed through formal ethics clearance from the university committee, strict adherence to data protection regulations (e.g., POPIA/GDPR), and exclusive use of de-identified public datasets. The AI tool will be explicitly framed as a clinician decision-support (not diagnostic) system to ensure human oversight. Transparency will be maintained via open methodology documentation, and potential algorithmic biases will be evaluated and disclosed.

This study will fully comply with the licensing terms of the publicly available ECG datasets used, specifically the MIT-BIH Arrhythmia Database (OSL 1.0) and the PTB Diagnostic ECG Database (CC-BY-4.0), by properly attributing all sources in both publications and code. The novel implementation of the RDSCNN model will be released under an OSI-approved open-source license, such as the MIT License, to encourage reuse while preserving authorship recognition. All work will be carried out in accordance with The Independent Institute of Education’s policies on student-generated intellectual property. Derivative outputs, including AAMI-mapped data and preprocessing scripts, will be shared exclusively for non-commercial research and reproducibility purposes. All manuscripts and supplementary materials will clearly cite the original dataset creators, following academic and licensing standards.

Storage and Backup

All data used in this research will be securely stored on an encrypted, access-controlled local drive, with regular backups maintained on a secure cloud storage service such as Google Drive or OneDrive linked to an institutional account. The original ECG datasets, along with processed files, model checkpoints, and analysis scripts, will be organized using a version-controlled folder structure to ensure traceability and reproducibility. Data integrity will be preserved through routine checksum validation, and access will be restricted to the primary researcher and supervisor to maintain confidentiality. Additionally, all code and essential data artifacts will be backed up using a private GitHub repository to facilitate version tracking and disaster recovery.

Access to all research data, scripts, and models will be strictly limited to the primary researcher and academic supervisor. Data will be stored on password-protected devices with full-disk encryption enabled to prevent unauthorized access. Cloud-based backups will be secured using institutional accounts with two-factor authentication. Publicly available datasets will be handled in compliance with their licensing terms, and any derivative files used for analysis will be anonymized and stored in restricted-access folders. Version control systems like GitHub will be used in private mode to maintain secure, trackable code development, with access granted only to approved collaborators. These measures ensure data confidentiality, integrity, and compliance with institutional and legal data protection standards.

Selection and Preservation

This research will preserve all essential outputs needed for transparency, reproducibility, and future reuse. Key assets include structured indices and metadata files (in CSV/JSON format with accompanying METADATA.yml) that detail patient IDs, beat indices, AAMI class labels, and dataset split assignments—enabling exact replication of experiments without redistributing raw data. All preprocessing and validation scripts, covering signal filtering, R-peak detection, AAMI label mapping, and inter-patient splitting, will be preserved as fully documented Python Jupyter notebooks and hosted in a public GitHub repository under the MIT License. The trained RDSCNN model, including its weights (.h5), architecture (TensorFlow SavedModel), and hyperparameter configurations, will be stored in standardized AI model formats such as ONNX or TFLite to support deployment on edge devices. Performance benchmarking results, including accuracy, F1-scores per AAMI class, and computational efficiency metrics (e.g., inference time, FLOPs), will be saved as structured CSV tables and visual figures for publication and future comparison. Raw ECG signals and annotations will not be preserved, as they are already maintained by PhysioNet and Kaggle under restrictive licenses, and intermediate files such as filtered signals will be excluded due to size and reproducibility via shared scripts.

The datasets will not be preserved, as they are already publicly available.

Data Sharing

Responsibilities and Resources