Project proposal: Early & accessible CVD risk prediction

Group 15:

eake191
awns132
qwan621
thil124
zden973
zfan700

Background & Motivation

Cardiovascular disease (CVD) is the cause of a **quarter** of all deaths every year (McNamara et al., 2019).

CVD is incurable, however intervention at early stages (such as medication) can slow its progression and extend life.

Identification of those at risk **before** symptoms become severe allows the best chance at a successful intervention

Many ML studies have been done in this area, however most risk models rely on clinical metrics (e.g., blood pressure), reducing practicality and accessibility.

Mc Namara, K., Alzubaidi, H., & Jackson, J. K. (2019). Cardiovascular disease as a leading cause of death: how are pharmacists getting involved? *Integr Pharm Res Pract, 8*. 1–11. doi: 10.2147/IPRP.S133088.

Lit review

There are over 300 articles published in online journals that use ML to predict CVD risk (Krittanawong et al., 2020). However, most of them use **prevalent** data and rely on **clinical factors**.

Dritsas et al. (2022) performed the most similar analysis, however included clinical factors like cholesterol and blood pressure. They also used prevalent data, limiting clinical applicability.

Our dataset was sourced from Weng et al. (2017), which we will be using to construct our project.

Improvements:

1. Removing clinical factors
2. Using a wider range of ML models

Dritsas, E., Alexiou, S., & Moustakas, K. (2022). Cardiovascular disease risk prediction with supervised machine learning techniques. In *Proceedings of the 8th International Conference on Information and Communication Technologies for Ageing Well and e-Health (ICT4AWE 2022)* (pp. 315–321). DOI: 10.5220/0011088300003188

Krittanawong C, Rogers AJ, Johnson KW, Wang Z, Turakhia MP, Halperin JL, Narayan SM. Integration of novel monitoring devices with machine learning technology for scalable cardiovascular management. *Nat Rev Cardiol*,*18*. 75–91 doi: 10.1038/s41569-020-00445-9

Weng, S. F., Reps, J., Kai, J., Garibaldi, J. M. & Qureshi, N. (2017). Can machine-learning improve cardiovascular risk prediction using routine clinical data? Stephen. *PLoS ONE 12*, 1–14. doi.org/10.1371/journal.pone.0174944

Objectives of research

Objective: Develop a model that uses readily available characteristics (e.g., diet, lifestyle, medical history) to predict whether a user is at risk for CVD without requiring specialised medical tests (e.g., blood pressure, LDL levels).

Approach:

1. Train and evaluate multiple ML models to identify the most accurate predictor.
2. Compare feature profiles of CVD cases vs. controls to highlight key risk indicators.

Outcome: If your feature profile aligns more closely with (eventual) CVD cases, consider consulting a primary care provider for further evaluation.

Dataset

308,854 UK-based participants

Cases: 24,971 10-year incident cases

Controls: 283,883 controls

17 baseline features:

1. **Lifestyle**: general health (1-5), exercise (y/n), weight, BMI, smoking (y/n), alcohol (0-30/week), fruit consumption (0-120/week), vegetable consumption (0-128), fried potato consumption (0-128)
2. **Demographic**: sex (m/f), age, height
3. **Medical**: skin cancer (y/n), other cancer (y/n), depression (y/n), diabetes (y/n), arthritis (y/n)

Methodology

1. **Preprocessing**: Remove irrelevant features, create participant IDs, etc.
2. **Data Splitting**: Split the data into 90% training and 10% test sets, stratified by case/controls.
3. **Class Imbalance**: Address imbalance using SMOTE on the training set, most likely via undersampling controls.
4. **Model Tuning**: Perform hyperparameter optimization via 5-fold cross-validation for all models: LightGBM, LR, RF, KNN, SVM, & MLP (if possible)
5. **Evaluation**: Assess the best model on the test set using:
   - ○ AUROC
   - ○ Precision, recall, sensitivity, specificity
   - ○ Feature importance

Risks & limitations

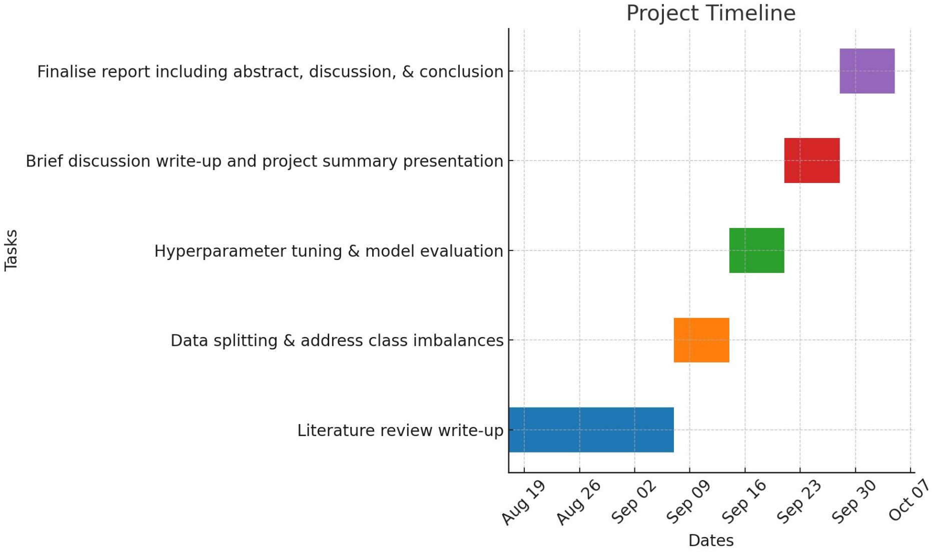
1. **Different CVD types**: CVD includes many diseases, and they might not all have the same risk factors.

2. **Missing data**: Trade-off between accuracy and accessibility - the removal of clinical features will most likely degrade our accuracy metrics.

3. **UK-only data**: Since the dataset is from the UK-based, the results may be as applicable to other countries with other ethnic makeups.

4. **Complex models**: Since we plan to use multiple complex models, available time may become an issue.

Timeline



Weeks 1-3:	Literature review write-up
Weeks 3-4:	Data splitting & address class imbalances
Weeks 5-6:	Hyperparameter tuning & model evaluation
Week 7:	Brief discussion write-up and project summary presentation
Weeks 8-9:	Finalise report including abstract, discussion, & conclusion.

Group contribution

Emma:	Write up, hyperparameter tuning, model evaluation.
Tom:	Write up, literature review, model evaluation
Qian:	Class imbalance investigation, write up
Siqi:	Preprocessing of data, splitting and preparing for evaluation.
Zishi:	Literature review, model evaluation
Ziheng:	Literature review, preprocessing of data

Q&A