Distilling knowledge from high quality biobank data towards the discovery of risk factors for patients with cardiovascular diseases and depression

Vasileios C. Pezoulas, *Student Member IEEE*, Georg Ehret, Jos Bosch, Dimitrios I. Fotiadis, *Fellow, IEEE*, Antonis Sakellarios

Abstract— **Cardiovascular disease (CVD) is the leading cause of death worldwide. Patients with CVD may also suffer from mental disorders, such as, depression which is a common comorbid condition. However, the risk factors for depression in CVD patients have not been extensively investigated in the literature. In this work, we utilized a hybrid and explainable AI-empowered workflow to identify underlying factors for CVD and depression. Towards this direction, we acquired a subset of the UK Biobank (UKB), including 157,302 patients with depression assessment and CVD. At the first step, 701 features were selected from the UKB, upon clinical guidance, including demographics, blood tests, mental examinations, and clinical assessments. An automated biobank data curation pipeline was applied to transform the UKB subset into a high-quality dataset by removing outliers, and genes with increased variability. A hybrid version of the XGBoost classifier was used to classify patients with CVD and depression, where a scalable loss function was utilized to overcome overfitting effects. Our results demonstrate that we can diagnose patients with comorbid conditions of CVD and depression with 0.80, 0.82, accuracy, and sensitivity, respectively, where the mood swings, BMI, and age, were identified as biomarkers, among others. To our knowledge, this is the first case study aiming to distil knowledge from the UKB to identify cost effective risk factors for patients with CVD and depression.**

Clinical relevance— Using a hybrid and explainable AI model, as the one presented in our work, we can effectively identify patients with both diseases in a cost-effective way since the identified and used biomarkers can be easily collected in everyday clinical practice.

I. Introduction

Cardiovascular disease (CVD) is the leading cause of death worldwide. In parallel, depression is the third leading cause of non-fatal health loss globally [1]. It has been proposed that depression and CVD present a bidirectional relationship in which a CVD patient is more likely to be depressive and vice versa [2], [3]. Depression and CVD share common risk factors, such as age, inflammation and oxidative stress [4]. Regardless of the shared risk factors between CVD and depression, the link between the two diseases is still unclear.

The recent years, machine learning based methodologies have been developed to predict depression. The increase of data availability contributed to this research area. For example, depression was predicted with 86.20% accuracy employing the Random Forest (RF) classifier and using data from 6,588 patients including hundreds of features [5]. RF presented the highest accuracy also in another study focused on an elderly population. In that case, the predictive model has 91% accuracy applied to an external validation dataset [6]. In a similar concept, many other studies have been presented with the general aim of diagnosis or prediction of depression under different populations or pathologic conditions [7], [8]. None of these studies, however, has focused on the classification of patients with both CVD and depression across large scale data from biobanks.

To address this need, we propose a hybrid, explainable AI-empowered pipeline to identify patients with comorbid conditions by harnessing knowledge from the well-known UK Biobank database which provides clinical, lifestyle, and omics related information regarding the CVD and depression. An automated biobank data curation workflow was applied on each batch to remove outliers, and genes with increased variability. A hybrid version of the XGBoost algorithm was trained on the aggregated, curated UKB subset to classify patients with CVD and mental disorders by utilizing a hybrid loss function which is resilient against overfitting effects. Explainability analysis was finally applied to identify risk factors for CVD and mental disorders, such as, mood swings, fed-up feelings, BMI and age. Our results highlight the favorable performance of the hybrid XGBoost which achieved similar results (sensitivity 0.86, specificity 0.73) with the conventional XGBoost (sensitivity 0.83, specificity 0.76) and the Random Forests (sensitivity 0.87, specificity

* The project is co-financed by Greece and the European Union—European Regional Development Fund (ERDF) under the Operational Program “Competitiveness Entrepreneurship Innovation” (EPAnEK), NSRF 2014-2020 (Project code: 82863, MIS: 5047133). The data used in the analysis were acquired by the TO_AITION project (GA: 848146).

V. C. Pezoulas, D. I. Fotiadis, and A. I. Sakellarios are with the Unit of Medical Technology and Intelligent Information Systems, Dept. of Materials Science and Engineering, University of Ioannina, GR 45110, Ioannina, Greece, and with the Biomedical Research Institute, Foundation for Research and Technology-Hellas (FORTH) (phone: +30 26510 09006; fax: +302651005588; emails: fotiadis@uoi.gr , bpezoulas@gmail.com, ansakel13@gmail.com).

G. Ehret is with the University of Geneva. J. Bosch is with the University of Amsterdam.

0.71) algorithm. Since the identified risk factors can be easily obtained in clinical practice, we can efficiently detect patients with both diseases using an explainable AI model, such as the one presented in our work.

II. Materials and Methods

A. An overview of the workflow

According to Fig. 1, the workflow for the analysis of biobank data consists of two primary stages, namely: (i) the incremental data curation stage, and (ii) the hybrid machine learning and explainability analysis stage. The workflow uses as input an anonymized subset from the UKB biobank (Section B) with M patients and N features. The dataset is divided into Q batches, with J patients, where $Q > 2$. Each batch is introduced into an efficient data curation pipeline (Section C) to automatically remove outliers, duplicated features, and genes with increased variability, as well as, incompatible and inconsistent fields across the data. The pipeline produces a data quality evaluation report and a curated dataset. Each curated batch is then collected and aggregated to formulate the final curated dataset with M patients and N' records, where $N' \leq N$.

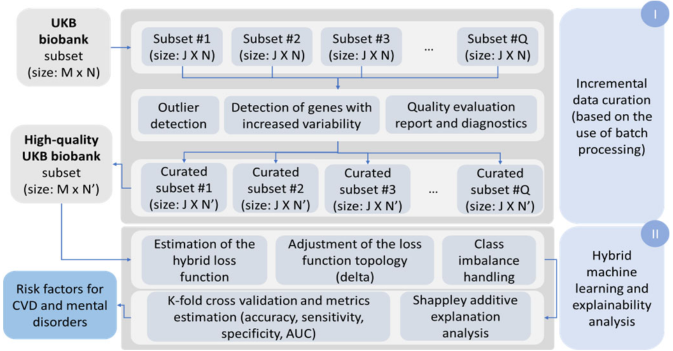
UKB biobank subset (size: M x N)
Subset #1 (size: J X N)
Subset #2 (size: J X N)
Subset #3 (size: J X N)
...
Subset #Q (size: J X N)
I
Incremental data curation (based on the use of batch processing)
Outlier detection
Detection of genes with increased variability
Quality evaluation report and diagnostics
High-quality UKB biobank subset (size: M x N')
Curated subset #1 (size: J X N')
Curated subset #2 (size: J X N')
Curated subset #3 (size: J X N')
...
Curated subset #Q (size: J X N')
II
Hybrid machine learning and explainability analysis
Estimation of the hybrid loss function
Adjustment of the loss function topology (delta)
Class imbalance handling
Risk factors for CVD and mental disorders
K-fold cross validation and metrics estimation (accuracy, sensitivity, specificity, AUC)
Shappley additive explanation analysis

Figure 1. An illustration of the proposed workflow.

A hybrid machine learning workflow based on boosting ensembles is applied on the high-quality biobank data to shed light into risk factors for CVD and mental disorders through the application of a hybrid loss function which was explicitly designed to deal with overfitting effects across complex data. Class imbalance handling methods are also used to deal with the significant difference among the number of patients in the control group and in the target group (those having both CVD and mental disorders). The performance of the trained model is evaluated based on a stratified cross validation schema by estimating the accuracy, sensitivity, specificity, and area under the ROC curve (AUC). Shappley additive explanation analysis is finally applied to identify risk factors for CVD and MD.

The workflows were executed under the high-performance computing infrastructure (HCI) which has been explicitly designed for data intensive tasks as part of the PRECIOUS system (MIS 5047133). The HCI currently includes 576 Intel(R) Xeon(R) Gold 5220R physical cores, 86000 CUDA cores, 4.6 TB RAM, and 0.5 PB raw storage.

B. Data origins

Anonymized data were obtained from 157,302 patients and 701 features related to cardiovascular diseases and mental health assessment (e.g., demographics, laboratory tests, blood biomarkers) from the UK Biobank (UKBB), upon approval. UK Biobank is a large-scale biomedical database and research resource, containing in-depth genetic and health information from half a million UK participants. The database is regularly augmented with additional data and is globally accessible to approved researchers undertaking vital research into the most common and life-threatening diseases [9]. In this work, we selected a sub-set population of 157,302 individuals from whom mental assessment for depression was performed using specialized questionnaires. In this subset, 37,294 patients were diagnosed with at least one CVD condition, 31,856 patients were diagnosed with depression and 7,644 with both the comorbid conditions.

C. Incremental data curation based on batch processing

Considering the large size of the UKBB database and its underlying complexity, the available anonymized dataset was split into 15 batches, where each batch included 10,000 patients and 701 features. Then, a fully automated medical data curation pipeline which was developed in Python and was presented in a previous study [10] was incrementally applied on each batch to generate data quality evaluation reports along with the curated datasets. Additional memory buffer functionalities were added as part of the workflow to improve the execution time. Data incompatibilities, such as, mixed data types and additional data inconsistencies, such as, data type representations were automatically resolved. Genes with increased variability were identified by estimating the covariance matrix from the input space and eliminating pairs of genes with increased variability. The isolation forests (IF) algorithm was trained on the available data to detect and remove outliers. The intermediate curated batches from each round were aggregated to extract metadata information, including the total number of missing values, the number of discrete and continuous features, the number of features with good (no missing values), bad (more than 30% missing values, and fair (less than 30% missing values) quality. The bad features were removed from the analysis. The final high-quality UKB subset included 157,302 patients and 335 features (those with good and fair quality).

D. Hybrid machine learning and explainability analysis

1) Hybrid machine learning

The extreme gradient boosting (XGBoost) technique has excelled in numerous Kaggle contests, and boosting ensembles have been frequently employed to handle classification tasks with improved performance [11]. The objective is to find an estimated function , $\tilde{G}(\boldsymbol{x})$, mapping $\boldsymbol{x}$ to y that minimizes the expected value of a loss function, suppose $L(y, G(\boldsymbol{x}))$, given a collection of N-observations, say $\{(\boldsymbol{x_1},y_1), (\boldsymbol{x_2},y_2), \ldots, (\boldsymbol{x_N},y_N)\}$. At stage $q \in Q$, the XGBoost aims to derive mapper estimations, say $G_q(x)$, in a sequential manner, as in [11], [12]:

$$G_q(x) = G_{q-1}(x) - \gamma_q \sum_{i=1}^{N} \nabla_{G_{q-1}} L(y_n, G_{q-1}(x_n)), \quad (1)$$

where the objective can be approximated by Taylor's theorem:

$$B(q) \approx \sum_{i=1}^{N} \left[l(y_i, \tilde{y}_{i,q-1}) + p_i g_q(x_i) + \frac{1}{2} t_i g_q{}^2(x_i) \right] + r, \qquad (2)$$

where p_i and t_i are the first and second order gradients of the loss function, respectively. However, a significant and rather crucial issue in the XGBoost schema lies on the fact that the early additions of regression trees to the ensemble tend to have a greater influence on the decision-making process than those trees that are added later in the ensemble. Dropout rates [13] can overcome this problem by incorporating a portion of discarded trees into the decision-making process. In fact, trivial trees may be avoided since the over-fitting problem can be solved using drop trees. However, the dropout rate is predetermined and frequently arbitrarily chosen, which has a significant negative impact on how well the XGBoost performs. To overcome this problem, we used a hybrid loss function presented in previous studies [14] that combines the modified Huber loss and the logcosh loss, where the delta value (i.e., the scale factor in the modified Huber loss) is employed to account for the hybrid topology's shape. In the "dart" booster, we also link the delta value to the dropout rate so that higher dropout rates can result in steeper loss topologies that prevent overfitting. To this end, the objective function in Eq. (2) can be re-written as::

$$B(q) \approx \sum_{i=1}^{N} \left[l(y_i, \tilde{y}_{i,q-1}) + t_i g_q(x_i) + \frac{1}{2} p_i g_q{}^2(x_i) \right] =$$

$$= \sum_{i=1}^{N} \left[l(y_i, \tilde{y}_{i,q-1}) + tanh(i) d/\sqrt{s} + \frac{1}{2} 1/cosh^2(i)\sqrt{s}/s \right], \qquad (3)$$

where d is the scale of the topology, and s in an approximation factor. The conventional XGBoost algorithm with the "gbtree" booster and the XGBoost with the "dart" booster from Python's xgboost package were used for comparison purposes. Random downsampling with replacement was applied to deal with the class imbalance to ensure a 1:1 ratio among the control and the target groups. The whole process was applied ten times to avoid biases in the training procedure. Additional classifiers like the AdaBoost and the Random Forests were also used in the analysis for comparison purposes.

2) *Explainability analysis*

An innovative approach from coalition game theory called the Shapley Additive explanation analysis (SHAP) can provide insight into the decision-making process of an AI model [15]. To achieve this, SHAP makes use of explanation models that produce interpretable and explicable categorization results. The SHAP value of a feature $d_j \in D$, say S_j, is defined as the total contribution of this feature to the outcome, given a subset of input features, say $P \complement \{d_1, d_2, \dots, d_Z\}$, from a larger set of K-features $\{d_1, d_2, \dots, d_K\}$, where $Z \leq K$, as in [15]:

$$S_j = \sum \frac{|D|!\,(P - |D| - 1)!}{P!} (f_d(D \cup \{d\}) - f_d(D)), \qquad (4)$$

where, $f_d(D)$ is the expected value of the function conditioned on P, K is the set of all input features, and $|D|$ is the number of features in D. The number of observations that are connected to a specific feature was also counted using the cover metric, by estimating the number of splits that each feature participated in the ensemble.

III. Results

A. *Batch based data curation*

According to Fig. 2, the initial UKB subset (701 features and 152,307 instances) was split into 16 batches; 15 batches with 701 features and 10,000 instances, and 1 batch with 701 features and 7,302 instances. The incremental data curation workflow presented in Section C was applied to each batch and the results were aggregated. In total, 111 features had "good" quality, 294 features with "fair quality", in average. The 296 features with "bad" quality were discarded. The total percentage of detected outliers was 0.5% and the percentage of features with unknown value types was 0.044%. The final dataset included 157,302 instances and 400 features.

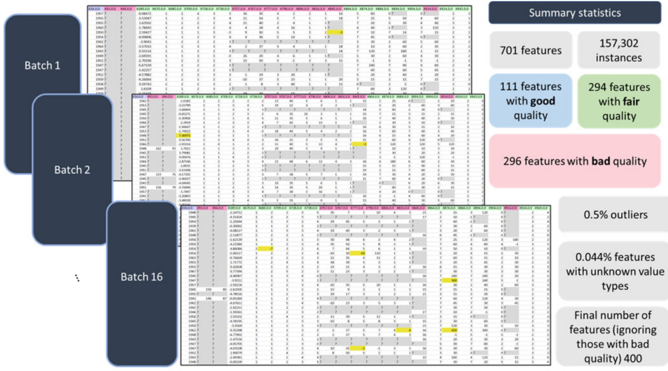
Batch 1
Batch 2
Batch 16
Summary statistics
701 features
157,302 instances
111 features with good quality
294 features with fair quality
296 features with bad quality
0.5% outliers
0.044% features with unknown value types
Final number of features (ignoring those with bad quality) 400

Figure 2. Indicative instances of the curated batches along with summary statistics (averaged across the total number of pre-defined batches).

B. *Hybrid AI model performance*

According to the classification results, the three algorithms perform equally good (Table II). In particular, the same accuracy is observed (0.79) for all models. The highest sensitivity is found using the RF classifier, while the hybrid XGBoost presents a better combination of sensitivity and specificity having in mind that the highest the sensitivity the better for a diagnostic model applied in the clinical practice. The highest AUC is found for the hybrid XGBoost and the RF (Table II, Fig. 1).

TABLE I. Classification Results (Number of Controls: 7,644, Number of Targets: 7,644).

	XGBoost	**Random Forests**	**Hybrid XGBoost**
Accuracy	0.79	0.79	0.79
Sensitivity	0.83	0.87	0.86
Specificity	0.76	0.71	0.73
AUC	0.87	0.88	0.88

C. *Risk factors for CVD and mental disorders*

Fig. 3 provides the ROC curve of the hybrid XGBoost classifier, whereas Fig. 4 provides the information-dense summary of how the top features in the dataset impact the model's output, where in each instance the given explanation is represented by a single dot on each feature row. According to Fig. 3 and Fig. 4, it is clear that for the comorbid condition of CVD and depression, the features of the model are distributed into two main categories: one mostly related to the patient mood, and one mostly related to cardiovascular disease and blood biomarkers. Also, we can easily conclude that all the biomarkers can be easily measured in clinical practice, meaning that a cost-effective solution of the diagnosis of CVD/depression is possible. More specifically, the feelings of the individual (unenthusiasm, tenseness, tiredness) are highly associated with the comorbid condition. An interesting finding is that although the LDL direct and cholesterol are highly ranked, they have the opposite direction than the expected (high cholesterol – high probability of CVD). This may be the effect of getting lipid lowering treatment such as using statins.

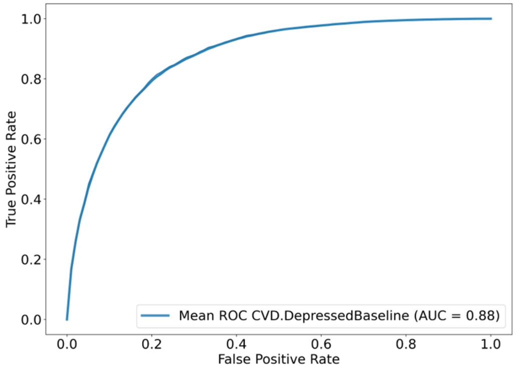

Figure 3. ROC curve of the hybrid XGBoost classifier.

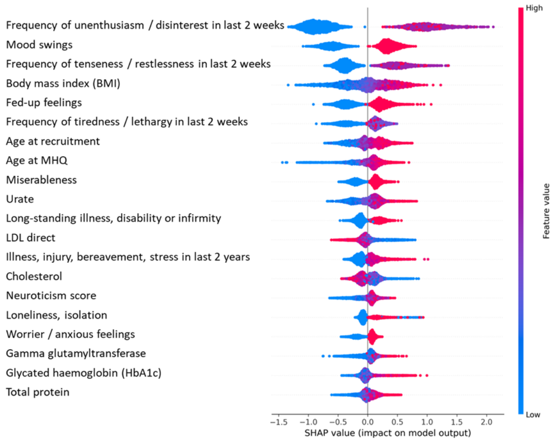

Figure 4. Information-dense summary per feature impact in the classification.

IV. Discussion And Conclusions

In this work, for the first time, we present an AI-empowered pipeline to detect patients who are most likely diagnosed with CVD and depression. For this purpose, we utilized the well-known cohort of UKB, which provides information about the CVD and depression. A hybrid representation of the XGBoost classifier has been employed to identify the patients with comorbid conditions. This classifier performed equally well with other traditional classifiers. Depression is associated with complications for optimal CVD management, including low adherence to healthy lifestyles and taking medications in accordance with medical recommendations. Also, it may increase mortality, disability, healthcare expenditures and reduced quality of life among patients with CVD.

Unfortunately, strategies for screening and treating depression are poorly implemented in patients with CVD. Using a hybrid and explainable AI model, as the one presented in our work, we can effectively identify patients with both diseases in a cost-effective way since the identified and used biomarkers can be easily collected in everyday clinical practice.

References

[1] S. L. James *et al.*, “Global, regional, and national incidence, prevalence, and years lived with disability for 354 diseases and injuries for 195 countries and territories, 1990–2017: a systematic analysis for the Global Burden of Disease Study 2017,” *The Lancet*, vol. 392, no. 10159, pp. 1789–1858, Nov. 2018, doi: 10.1016/S0140-6736(18)32279-7.

[2] Y. Xue, G. Liu, and Q. Geng, “Associations of cardiovascular disease and depression with memory related disease: A Chinese national prospective cohort study,” *J. Affect. Disord.*, vol. 260, pp. 11–17, Jan. 2020, doi: 10.1016/j.jad.2019.08.081.

[3] N. Graham *et al.*, “Impact of major depression on cardiovascular outcomes for individuals with hypertension: prospective survival analysis in UK Biobank,” *BMJ Open*, vol. 9, no. 9, p. e024433, Sep. 2019, doi: 10.1136/bmjopen-2018-024433.

[4] M. Shao *et al.*, “Depression and cardiovascular disease: Shared molecular mechanisms and clinical implications,” *Psychiatry Res.*, vol. 285, p. 112802, Mar. 2020, doi: 10.1016/j.psychres.2020.112802.

[5] K.-S. Na, S.-E. Cho, Z. W. Geem, and Y.-K. Kim, “Predicting future onset of depression among community dwelling adults in the Republic of Korea using a machine learning algorithm,” *Neurosci. Lett.*, vol. 721, p. 134804, Mar. 2020, doi: 10.1016/j.neulet.2020.134804.

[6] A. Sau and I. Bhakta, “Predicting anxiety and depression in elderly patients using machine learning technology,” *Healthc. Technol. Lett.*, vol. 4, no. 6, pp. 238–243, 2017, doi: 10.1049/htl.2016.0096.

[7] Md. S. Zulfiker, N. Kabir, A. A. Biswas, T. Nazneen, and M. S. Uddin, “An in-depth analysis of machine learning approaches to predict depression,” *Curr. Res. Behav. Sci.*, vol. 2, p. 100044, Nov. 2021, doi: 10.1016/j.crbeha.2021.100044.

[8] A. Priya, S. Garg, and N. P. Tigga, “Predicting Anxiety, Depression and Stress in Modern Life using Machine Learning Algorithms,” *Procedia Comput. Sci.*, vol. 167, pp. 1258–1267, Jan. 2020, doi: 10.1016/j.procs.2020.03.442.

[9] C. Sudlow *et al.*, “UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age,” *PLoS Med.*, vol. 12, no. 3, p. e1001779, Mar. 2015, doi: 10.1371/journal.pmed.1001779.

[10] V. C. Pezoulas *et al.*, “Medical data quality assessment: On the development of an automated framework for medical data curation,” *Comput. Biol. Med.*, vol. 107, pp. 270–283, Apr. 2019, doi: 10.1016/j.compbiomed.2019.03.001.

[11] T. Chen and C. Guestrin, “XGBoost: A Scalable Tree Boosting System,” *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.*, pp. 785–794, Aug. 2016, doi: 10.1145/2939672.2939785.

[12] T. Chen and T. He, “xgboost: eXtreme Gradient Boosting,” p. 4.

[13] K. V. Rashmi and R. Gilad-Bachrach, “DART: Dropouts meet Multiple Additive Regression Trees,” arXiv, arXiv:1505.01866, May 2015. doi: 10.48550/arXiv.1505.01866.

[14] V. Pezoulas *et al.*, “Metabolomics in the prediction of prodromal stages of carotid artery disease using a hybrid ML algorithm,” in *2022 IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI)*, Sep. 2022, pp. 1–4. doi: 10.1109/BHI56158.2022.9926774.

[15] S. M. Lundberg and S.-I. Lee, “A Unified Approach to Interpreting Model Predictions,” in *Advances in Neural Information Processing Systems*, Curran Associates, Inc., 2017. Accessed: May 18, 2023. [Online]. Available: https://papers.nips.cc/paper_files/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html