Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon

Research article

Prediction of atrial fibrillation and stroke using machine learning models in UK Biobank

Areti Papadopoulou[a], Daniel Harding[a], Greg Slabaugh[b,c], Eirini Marouli[a,c,*,1], Panos Deloukas[a,d,**,1]

[a] William Harvey Research Institute, Barts and the London School of Medicine and Dentistry, Queen Mary University of London, London, UK
[b] School of Electronic Engineering and Computer Science, Queen Mary University of London, London, UK
[c] Digital Environment Research Institute, Queen Mary University of London, London, UK
[d] Princess Al-Jawhara Al-Brahim Centre of Excellence in Research of Hereditary Disorders (PACER-HD), King Abdulaziz University, Jeddah, Saudi Arabia

ABSTRACT

Objective: Atrial fibrillation (AF) is the most common cardiac arrythmia, and it is associated with increased risk for ischemic stroke, which is underestimated, as AF can be asymptomatic. The aim of this study was to develop optimal ML models for prediction of AF in the population, and secondly for ischemic stroke in AF patients.

Methods: To develop ML models for prediction of 1) AF in the general population and 2) ischemic stroke in patients with AF we constructed XGBoost, LightGBM, Random Forest, Deep Neural Network, Support Vector Machine and Lasso penalised logistic regression models using UK-Biobank's extensive real-world clinical data, questionnaires, as well as biochemical and genetic data, and their predictive performances were compared. Ranking and contribution of the different features was assessed by SHapley Additive exPlanations (SHAP) analysis. The clinical tool CHA_2DS_2-VASc for prediction of ischemic stroke among AF patients, was used for comparison to the best performing ML model.

Findings: The best performing model for AF prediction was LightGBM, with an area-under-the-roc-curve (AUROC) of 0.729 (95% confidence intervals (CI): 0.719, 0.738). The best performing model for ischemic stroke prediction in AF patients was XGBoost with AUROC of 0.631 (95% CI: 0.604, 0.657). The improved AUROC in the XGBoost model compared to CHA_2DS_2-VASc was statistically significant based on DeLong's test (p-value = 2.20E-06). In addition, the SHAP analysis showed that several peripheral blood biomarkers (e.g. creatinine, glycated haemoglobin, monocytes) were associated with ischemic stroke, which are not considered by CHA_2DS_2-VASc.

Implications: The best performing ML models presented have the potential for clinical use, but further validation in independent studies is required. Our results endorse the incorporation of some routinely measured blood biomarkers for ischemic stroke prediction in AF patients.

1. Introduction

Atrial fibrillation (AF) is the most common cardiac arrythmia, which is characterised by a rapid and irregular heartbeat [1,2]. The incidence of AF is increasing rapidly with 12.1 million people expected to be affected by 2030. This is mainly attributed to the ageing of the population, along with changes in lifestyle. AF, besides doubling the risk of cardiovascular mortality, is associated with increased

[*] Corresponding author. William Harvey Research Institute, Barts and The London School of Medicine and Dentistry, Queen Mary University of London, London, UK.

[**] Corresponding author. William Harvey Research Institute, Barts and The London School of Medicine and Dentistry, Queen Mary University of London, London, UK.

E-mail addresses: e.marouli@qmul.ac.uk (E. Marouli), p.deloukas@qmul.ac.uk (P. Deloukas).

[1] These authors contributed equally to this work.

https://doi.org/10.1016/j.heliyon.2024.e28034
Received 31 August 2023; Received in revised form 7 March 2024; Accepted 11 March 2024
Available online 17 March 2024

risk of stroke, ischemic heart disease, heart failure and cognitive dysfunction. More specifically, AF quintuple the risk for ischemic stroke, independent of age. However, AF is sometimes asymptomatic, and thus remains undetected, and subsequently the ischemic stroke risk attributed to AF is under-estimated [1,2].

Machine learning (ML) algorithms are promising to revolutionise disease prediction, classification of medical images and diagnosis revealing new features, which would have not been discovered using traditional statistical models [3]. ML models use a hypothesis-free approach with no prior assumptions either among the input features or between the features and the outcome. ML methods with varying degree of accuracy have been reported for the prediction of circulatory diseases. However, they have been limited from access to large-scale cohorts with integrated clinical, biochemical and genetic data [3,4].

There have been several studies that employed ML methods for prediction of circulatory diseases. A recent study in Geisinger's clinical MUSE database with no history of AF, within 1-year of an ECG, employed deep neural networks and reported an area under the receiver operating characteristic (AUROC) of 0.85 for AF prediction [3]. They also reported that 62% of patients who had a stroke caused by AF within 3 years of an ECG, with no prior AF diagnosis, would have been identified by their prediction tool before the stroke occurred [3]. Another study employed four ML models to predict modified Rankin Scale (mRS) at hospital discharge and in-hospital deterioration for acute ischemic stroke patients enrolled on the Stroke Registry in Chang Gung Healthcare System (SRICHS) [4]. Random forest performed well in both outcomes; the AUROC was 0.83 for discharge mRS and 0.71 for in-hospital deterioration [4]. There have also been several studies using ML methods for the prediction of ischemic stroke in AF-patients. In the Korean National Health Insurance (KNHIS) dataset, the authors aimed to predict ischemic stroke occurrence in AF patients using ML models such as DNN, XGBoost and RF, for more than 150,000 AF patients. The best performing model was DNN with an AUROC of 0.727, outperforming CHA_2DS_2-VASc with AUROC of 0.651 [5]. Another study using the Fushimi AF registry, showed that CatBoost ML method outperformed CHA_2DS_2-VASc, having AUROC 0.72 (95%CI, 0.66–0.79) and 0.62 (95%CI, 0.54–0.70) respectively [6]. Using the Korean Atrial Fibrillation Evaluation Registry in Ischemic Stroke Patients (K-ATTENTION), the authors showed that LightGBM performed the best, with AUROC of 0.772 (95% CI 0.715–0.829), for the prediction of early neurological deterioration (END) among AF-related stroke patients [7]. The studies mentioned above underlined the importance of ML methods, since besides the improved prediction performance that they display in contrast to current clinical tools, they exhibit the potential to unravel new and diverse risk factors associated with the disease.

The aim of this study was to develop optimal ML models for prediction of: 1) AF in the population and 2) ischemic stroke in AF patients. We constructed ML models with six different algorithms in UK-Biobank (500,000 participants with extensive questionnaires, clinical, biochemical and genetic data – Tables S1–S3) and assessed their predictive performances. For ranking of feature importance and contribution to the prediction outcome we used SHapley Additive exPlanations (SHAP) [8].

2. Methods

2.1. Overview of the research framework

We included clinical data, phenotypes, lifestyle, and medications from UK-Biobank. We imputed the missing values and employed a feature selection process, described in more detail at *Data pre-processing*, to reduce the number of features employed to the ones relative to the outcome. Six ML models were used to create predictive models as described at the *ML methods* below. Each model's hyperparameters were optimised using 10-fold cross validation at the training dataset. The ML models were validated on the test dataset and their performances were compared. Lastly, we employed the SHAP explanations to reveal the features' contributions to the prediction.

2.2. Phenotype and participant selection

2.2.1. Data pre-processing

We examined the UK-Biobank, a prospective cohort of 502,492 participants, aged 37–73 years old, recruited between 2006 and 2010. The dataset includes blood measurements, clinical assessments, anthropometry, cognitive function, hearing, arterial stiffness, hand grip strength, sociodemographic factors, lifestyle, family history, psychosocial factors and dietary intake [9]. Related individuals were removed, and the remaining dataset for analysis included 454,118 participants. Furthermore, we incorporated medications as features, derived from field 20003 (https://biobank.ndph.ox.ac.uk/showcase/field.cgi?id=20003). Additionally, clinical data were employed, coded in ICD10, derived from the Hospital Episodes Statistics (HES), which are linked to the UK-Biobank. From these, we constructed phenotype codes or "phecodes", using a phecode map [10], which are aggregated ICD10 codes defining specific diseases or traits. We employed only the umbrella phecode categories. Detailed list of all the features that we examined can be found at Tables S1–S3. Moreover, we created two polygenic scores (PGS) which were included as features for the prediction of ischemic stroke in people with AF. The first one is the AF score, based on 94 genome-wide variants derived from the Roselli et al. [11] genome-wide association study (GWAS) for AF. The second is the Ischemic STROKE score, based on 28 genome-wide variants derived from the Malik et al. [12] GWAS for ischemic stroke. The AF SCORE was also employed as a feature both for the prediction of AF and for the ischemic stroke in AF patients.

The investigator phenotypes dataset from UK-Biobank includes 2,199 fields for 454,118 participants. We set answers "Do not know" and "Prefer not to answer" as NA and removed features that had more than 25% missingness, resulting in 390 investigator phenotypes. Afterwards, we imputed the missing values using a multivariate imputer that estimates each feature from all the others, using *IterativeImputer* from Python [13]. Then, we added 419 phecodes, available for 278,177 participants, derived from HES in UK-Biobank. Lastly, we added the medications from field 20003 (https://biobank.ndph.ox.ac.uk/showcase/field.cgi?id=20003), after applying one-hot-encoding, resulting in 1,289 medications for 294,698 participants (Fig. 1).

Next, we decided to balance the outcome sample size, since imbalanced data has a negative impact on ML procedures [14]. The classification algorithms have the tendency to get biased estimates towards the majority class, ignoring the minority class. This happens because most of the classifying methods aim to maximize the accuracy rate, meaning the number of correctly classified observations [15,16]. Therefore, we employed the *fixed under-sampling* technique from Python [17], which is a process for reducing the number of samples in the majority class; the control group in this case. The algorithm randomly selects samples from the control group, in order to have equal representation of both classes. After balancing the outcome, we used *VarianceThreshold* from Python [13], which eliminates all features whose variance does not meet a threshold of 90%. Additionally, we removed the continuous correlated fields using Pearson correlation, at a 0.8 threshold; features strongly correlated with the outcome were maintained. Next, we performed feature selection in order to reduce the computational cost via dimensionality reduction, achieve higher classification accuracy by eliminating the noise, and include the most relevant features for the disease prediction [18]. A recent paper by Ismael et al. [19], suggests that the best practice is for the fixed under-sampling technique to precede the feature selection. Therefore, we filtered all the remaining features using recursive feature elimination with cross-validation from Python [13] in order to find the optimal number of features to include in the ML models.

2.2.2. Create the AF outcome

We removed participants from the UK-Biobank that had cardiac dysrhythmias before the time of enrolment, with one or more of the following codes: non-cancer illness code, self-reported (1471, 1483); operation code (1524); diagnoses – main/secondary ICD10 (I44, I44.1–I44.7, I45, I45.0–I45.6, I45.8–I45.9, I46, I46.2, I46.8–I46.9, I47, I47.0–I47.2, I47.9, I48, I48.0–4, I48.9, I49, I49.0–I49.5, I49.8–I49.9, R00.0, R00.1, R00.2, R94.3, Z86.7, Z95.0, Z95.8-Z95.9); underlying (primary/secondary) cause of death: ICD10 (I44, I44.1–I44.7, I45, I45.0–I45.6, I45.8–I45.9, I46, I46.2, I46.8–I46.9, I47, I47.0–I47.2, I47.9, I48, I48.0–4, I48.9, I49, I49.0–I49.5, I49.8–I49.9, I60–I61, I63–I64 (NOT I63.6), R00.0, R00.1, R00.2, R94.3, Z86.7, Z95.0, Z95.8-Z95.9); diagnoses – main/secondary ICD9 (4273, 430, 431, 4339, 4340, 4341, 4349, 436); operative procedures – main/secondary OPCS (K57.1, K62.1–4). In total, 20,584

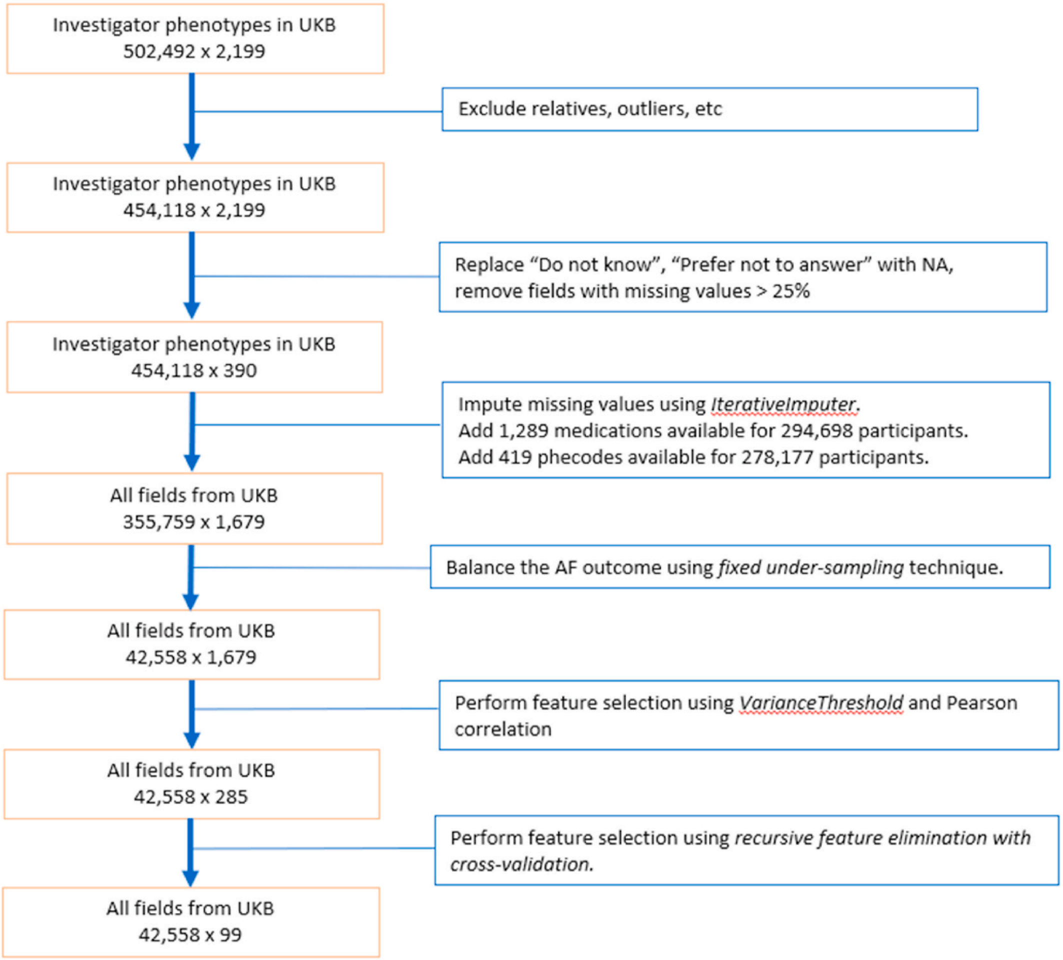
Investigator phenotypes in UKB
502,492 x 2,199
Exclude relatives, outliers, etc
Investigator phenotypes in UKB
454,118 x 2,199
Replace "Do not know", "Prefer not to answer" with NA, remove fields with missing values > 25%
Investigator phenotypes in UKB
454,118 x 390
Impute missing values using IterativeImputer.
Add 1,289 medications available for 294,698 participants.
Add 419 phecodes available for 278,177 participants.
All fields from UKB
355,759 x 1,679
Balance the AF outcome using fixed under-sampling technique.
All fields from UKB
42,558 x 1,679
Perform feature selection using VarianceThreshold and Pearson correlation
All fields from UKB
42,558 x 285
Perform feature selection using recursive feature elimination with cross-validation.
All fields from UKB
42,558 x 99

Fig. 1. Diagram depicting the data curation and feature selection process for the prediction of atrial fibrillation.

participants were excluded, having at least one of the above conditions, before enrolment in the UK-Biobank.

AF cases were defined when having one or more of the following codes: non-cancer illness code, self-reported (1471, 1483); operation code (1524); diagnoses – main/secondary ICD10 (I48, I48.0–4, I48.9); underlying (primary/secondary) cause of death: ICD10 (I48, I48.0–4, I48.9); operative procedures – main/secondary OPCS (K57.1, K62.1–4). In total, 21,279 people developed one of the conditions described above, after enrolment in UK-Biobank (Fig. 1).

2.2.3. Create the AF & stroke outcome

Cases were defined as participants who developed ischemic stroke after AF diagnosis in UK-Biobank with one or more of the following codes: diagnoses – main/secondary ICD10 (I63, I63.0–9, I64); diagnoses – main/secondary ICD9 (434, 436); underlying (primary/secondary) cause of death: ICD10 (I63, I63.0–9, I64). Thus, 3,150 people developed ischemic stroke after AF diagnosis and were included as cases, and the controls were people diagnosed with AF and did not develop stroke, as far as the data allow us to know. Based on the selection criteria for AF patients with and without ischemic stroke (Supplementary Fig. 1), 3,150 prospective cases who developed ischemic stroke after AF diagnosis and equal number of controls, along with 129 features, were included in the ML models (Table S8).

2.3. ML models

2.3.1. XGBoost

In more detail, XGBoost uses regression trees in a sequential learning process as weak learners into a single strong model, where each tree attempts to correct the residuals in the predictions made by previous trees. Regression trees include a continuous score on each leaf, which is the last node once the tree has grown. For a specific observation, the algorithm uses decision rules in the trees to classify it into the leaves. The sum of the scores on each leaf is the final prediction [20].

2.3.2. LightGBM

Machine learning methods relying on Gradient Boosting Decision Tree (GBDT) scan all the data instances, for all the features, to calculate the information gain for each possible split. As a result, the computational time and complexity will increase as the features accumulate. To this end, there are two techniques incorporated at LightGBM algorithm that contribute towards a faster implementation. Firstly, in the Gradient-based One-Side Sampling (GOSS) technique, instances that have larger gradients contribute more to the information gain, and the instances with smaller gradients are randomly sampled to provide accurate and fast estimation. Secondly, the Exclusive Feature Bundling (EFB) technique reduces the number of effective features. For datasets that are sparse, many features are mutually exclusive; they will rarely take nonzero values at the same time, therefore such features are tied into one [21].

2.3.3. Deep neural networks (DNN)

Deep learning is a subdomain of ML attempting to learn many levels of representation using multiple layers. These layers transform the data in a non-linear way, and as a result, more complex structure and relationships can be discovered. This method is inspired by the human brain, using a series of connected layers of neurons that constitute a whole network, including at least three layers: input, hidden and output. The input layer consists of multiple neurons, which use as input the original features. The hidden layers can be more than one, depending on the complexity of the dataset. Each layer includes multiple nodes, and each node from the previous layer is connected to each one from the next layer, constituting a fully connected network. Lastly, the output layer, using a sigmoid activation function, concludes in a number between 0 and 1, which represents the probability belonging to one of the two classes [22].

2.3.4. Support vector machine (SVM)

SVM is a high accuracy ML model, which can deal with non-linear spaces. It maps the input data into a higher dimension feature space, using a kernel function. Then, a linear decision surface (hyperplane), is created to classify the outcome, with properties that satisfy the generalisation of the algorithm. The optimal hyperplane classifies the data by using its maximal margin, employing a small percentage of the training data, which are named support vectors. The authors support that if the optimal hyperplane is created from a few support vectors, then the algorithm can be generalised, even in a space with infinite dimensions [23].

2.3.5. Cross-validation

The ML models aim to optimise the general model performance on datasets different from the ones used to train them. Therefore, evaluating the generalisation of ML methods requires the data to be split in three non-overlapping sets of training/validation/test, combined with stratified 10-fold cross-validation (CV), maintaining the same proportion of cases and controls in each fold. Grid search is performed using 9 sets for the parameter tuning, and the 1 remaining set is used for validation. This process is repeated 10 times, until every set is used once for training and once for validation. The best parameters for the model correspond to the highest score, which is calculated by averaging the results from all repetitions. The test dataset is used to check for overfitting and unbiased evaluation of the final model [13].

2.3.6. SHAP

ML models, although accurate and capable of capturing the non-linear relationships, are complex to interpret. A more widespread method for interpretation is SHAP, employed to understand each feature's contribution to the prediction, using cooperative game theoretic tools. The SHAP values are in theory the best solution up to now, however time-consuming, since all possible combinations need to be calculated. TreeExplainer is an expansion of SHAP, employing tree nodes instead of linear models for the estimation of Shapley values. The Shapley values of a tree-based algorithm are calculated as the weighted average of the Shapley values corresponding to individual trees. Thus, it is commonly used to explain tree-based machine learning models, reducing tremendously the computation time. In parallel, SHAP values seem to overcome the interpretability issue by employing both global and local interpretation. Global explanation relies on the effect of input features on the whole model, and local interpretation depicts the effect of input features on single predictions [8].

For the methods described above, Python computer language was employed [24]. The code and libraries that were employed are described in Table S5.

3. Results

Machine learning models can enhance prediction accuracy by utilising extensive datasets and incorporating potential predictors. In our present study, we demonstrated the improvement in prediction accuracy for ischemic stroke among AF patients, compared to current approaches, by employing machine learning modelling. The findings suggest inclusion of commonly measured blood biomarkers for prediction, while advocating for the incorporation of a genetic score for AF prediction. The approaches and modelling introduced in this study hold promise for clinical implementations.

3.1. AF

We examined 21,279 prospective AF cases and an equal number of controls in UK-Biobank. Baseline characteristics, along with comorbidities and medication, both overall and according to AF cases versus controls, are provided in Table 1.

In total, 99 features (Table S4) were employed, using five ML models to predict AF. The results presented in this section correspond to the optimal hyperparameters, derived after 10-fold cross-validation from the examined values included in Table S6. SVM did not converge after running 10 days and utilising 16 cores in Queen Mary's Apocrita HPC facility.[2]

The best AUROC value was achieved with LightGBM (Table 2) albeit De-Long's test (Table 3) showed that there is no evidence for significant difference in the AUROCs between LightGBM and XGBoost, DNN, or RF. In contrast, DeLong's test showed that there was statistically significant difference in the AUROCs between LightGBM and penalised LR (p-value = 1.38E-02), after considering multiple correction. The AUROC of penalised LR differed from the AUROC of all other examined ML models based on DeLong's test and this was statistically significant. The AUROC curves for the five models in the test dataset are shown in Fig. 2.

To estimate the contribution of each feature in each of the five models assessed for prediction of AF, we employed SHAP analysis, which is accurate, fast and stable. Fig. 3 displays the top 20 features, ranked according to their SHAP value, for the LightGBM model; features are listed in descending order, starting with the most significant for AF prediction. SHAP values depict the distribution of the effect of each feature on the model output.

Based on Fig. 3, SHAP analysis reveals that the top 3 most important variables contributing to the model were "Records in HES inpatient diagnoses dataset" which is the number of times an individual has been hospitalised (fieldID 41234), "Age at recruitment" (fieldID 21022) and "AF SCORE", using the unweighted sum of increasing alleles from Roselli et al. [11]. All the features' contributions, based on SHAP analysis, can be found in Table S7.

3.2. AF & stroke

We examined 3,150 prospective cases who developed ischemic stroke after being diagnosed with AF, and an equal number of controls in UK-Biobank including 129 features (Table S8) and using six models to predict ischemic stroke in AF cases. As indicated previously, results correspond to the optimal hyperparameters (Table S9).

The best AUROC value was achieved for XGBoost (Table 4). DeLong's test (Table 5) showed that there is no evidence for significant difference in the AUROCs between XGBoost and all other examined ML models but the penalised LR model (p-value = 2.00E-02) (Fig. 4).

As shown in Fig. 5, SHAP analysis revealed that the 3 most important variables contributing to prediction of ischemic stroke in AF cases in the model were "Records in HES inpatient diagnoses dataset" which is the number of times an individual has been hospitalised (fieldID 41234), "Age at recruitment" (fieldID 21022), and "Glycated haemoglobin (HbA1c)" which is a blood biochemistry measurement (fieldID 30750). Table S10 lists the contribution of each of the 129 features in the model based on SHAP analysis.

3.3. Comparison with CHA_2DS_2-VASc

The current tool used for prediction of ischemic stroke occurrence among AF patients is CHA_2DS_2-VASc which considers multiple risk factors; age, sex, heart failure, hypertension, stroke, vascular disease, diabetes [25]. Thus, we decided to compare the performance of the best ML model, XGBoost (Table 4), with CHA_2DS_2-VASc in UK-Biobank. To construct the CHA_2DS_2-VASc we employed the codes described in Table S11. The AUROC and 95% CI for CHA_2DS_2-VASc and XGBoost was 0.611 (0.585–0.638) and 0.631 (0.604–0.657) in

[2] This research utilised Queen Mary's Apocrita HPC facility, supported by QMUL Research-IT. http://doi.org/10.5281/zenodo.438045.

Table 1
Baseline characteristics for the 21,279 prospective AF cases and equal number of controls.

	Total	AF cases	AF controls	Pvalue*
Sex				
Females	20,231 (47.5%)	8122 (38.2%)	12,109 (56.9%)	<2.2E-16
Males	22,327 (52.5%)	13,157 (61.8%)	9170 (43.1%)	
Age (mean, sd)	59 (8)	62 (6)	57 (8)	<2.2E-16
Ethnicity				
EUR	41,042 (96.9%)	20,791 (97.7%)	20,251 (95.0%)	5E-03
AFR	535 (1.2%)	154 (0.7%)	381 (1.8%)	
EAS	127 (0.3%)	31 (0.2%)	96 (0.5%)	
SAS	854 (1.6%)	303 (1.4%)	551 (2.7%)	
Comorbidities				
Diabetes	6434 (15.1%)	4423 (20.8%)	2011 (9.5%)	<2.2E-16
Hypertension	22,019 (51.7%)	14,810 (69.6%)	7209 (33.9%)	<2.2E-16
Smoking				
Never	23,273 (54.7%)	11,627 (54.6%)	11,646 (54.7%)	0.8804
Previous	14791 (34.8%)	7389 (34.7%)	7402 (34.8%)	
Current	4494 (10.6%)	2263 (10.6%)	2231 (10.5%)	
Cholesterol lowering medication	7459 (17.5%)	3712 (17.4%)	3747 (17.6%)	0.4799
History of heart diseases	21,102 (49.6%)	11,233 (52.8%)	9869 (46.4%)	<2.2E-16
History of stroke	12,317 (28.9%)	6581 (30.9%)	5736 (26.9%)	<2.2E-16

Note. * *P*-values refer to chi-square test for dichotomous variables and to Mann-Whitney test for continuous data with non-parametric distribution.

Table 2
Performance of the ML models for AF prediction, on the test dataset, under various metrics.

Models	AUROC (95% CI)	Accuracy	Precision	Recall	F1 score
LightGBM	0.729 (0.719–0.738)	0.73	0.72	0.74	0.73
XGBoost	0.728 (0.718–0.737)	0.73	0.74	0.73	0.73
DNN	0.716 (0.706–0.725)	0.72	0.71	0.73	0.72
RF	0.715 (0.706–0.725)	0.72	0.71	0.74	0.72
LR (L1 penalty)	0.622 (0.612–0.633)	0.62	0.63	0.60	0.61

AUROC, the area under a receiver operating characteristic curve; Accuracy = (TP + TN)/(TP + TN + FP + FN); Precision = TP/(TP + FP), Recall = TP/(TP + FN) where TP stands for true positive, TN for true negative, FP for false positive, and FN for false negative; F1 score = 2 (precision × recall)/(precision + recall).

Table 3
DeLong's test for the ML model comparisons for AF prediction.

Models	LightGBM	XGBoost	DNN	RF
LightGBM	–			
XGBoost	8.28E-01	–		
DNN	3.67E-02	5.78E-02	–	
RF	1.17E-02	2.44E-02	9.91E-01	–
LR (L1 penalty)	1.38E-32	8.82E-32	2.41E-24	5.73E-27

the test set, respectively. The improved AUROC in the XGBoost model compared to CHA_2DS_2-VASc was statistically significant based on DeLong's test (p-value = 2.20E-06). Furthermore, the SHAP analysis for the XGBoost model (Fig. 5), shows that there is a significant number of peripheral blood markers associated with ischemic stroke, which are overlooked from CHA_2DS_2-VASc.

4. Discussion

4.1. Comparison of the performance of ML models for prediction of AF or ischemic stroke in patients with AF

We assessed six ML models in total for prediction of AF (XGBoost, LightGBM, RF, DNN, LR) or ischemic stroke in AF patients (XGBoost, LightGBM, RF, DNN, SVM, LR) and employed SHAP analysis to rank features for predictive importance. SHAP analysis was successful in the visualisation of non-linear relationships between the features used for prediction and the outcome. Additionally, the direction of the SHAP values for the top 20 features agrees with what has been reported so far in the literature. We found that the ensemble learning models LightGBM (best for AF prediction) and XGBoost (best for prediction of ischemic stroke in patients with AF) achieved higher AUROCs compared to the other examined models, suggesting that these models have better generalisation. DeLong's test showed that penalised LR model had a lower AUROC compared to all other models and these differences were statistically significant (Table 3), indicating that ML models capture useful information by modeling non-linear associations, leading to the discovery

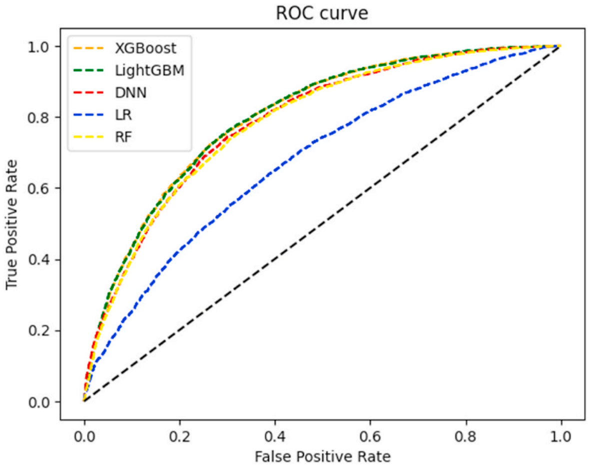

Fig. 2. AUROC for each ML model for AF prediction in the test dataset.

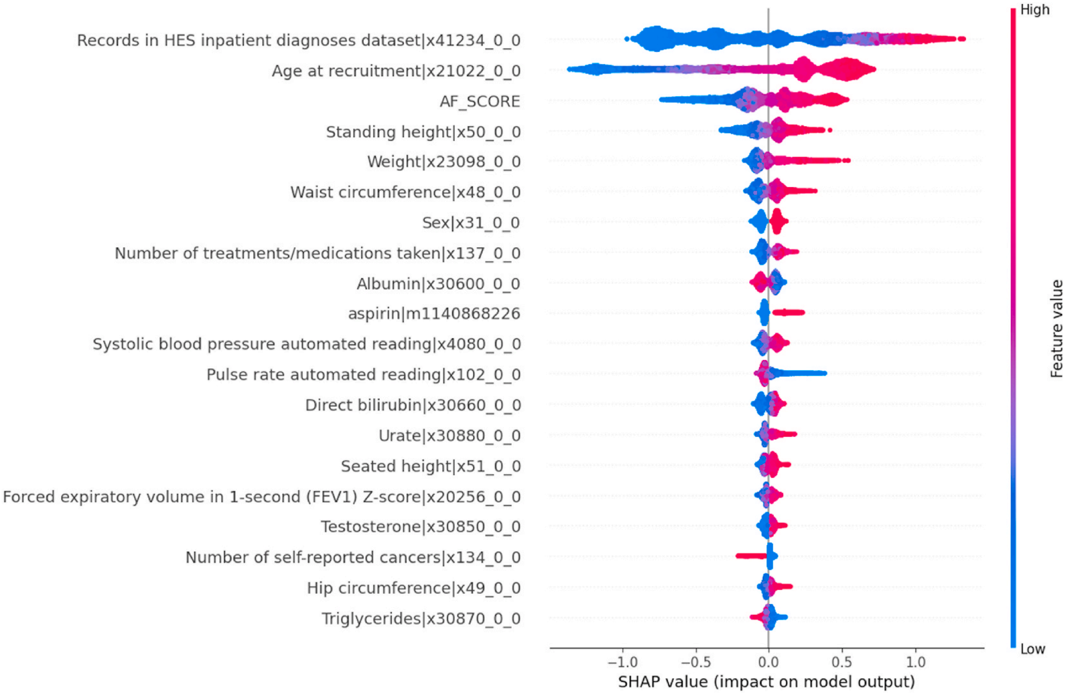

Fig. 3. Summary plot of the SHAP values (x-axis) for the top 20 features (y-axis), in descending order, showing the distribution of the impact that each feature has for the AF prediction on the test dataset, employing LightGBM model. Each dot represents a participant. The red dots represent a high feature value and blue dots represent a low feature value for each participant. For example, the AF SCORE had a positive impact on the model output, i.e., a higher AF SCORE increased AF risk. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

of new features.

4.2. AF results

Advancing age has been shown to be one of the most important risk factors for AF [26], which is corroborated by the present study and ranked as the second most important feature. The third most important feature in the model was the AF SCORE, a set of 94 genome-wide variants associated with AF and explaining 42% of the heritability in Europeans [11], which as expected had a positive impact on the model output, i.e. the higher the AF score the higher the risk of developing AF. Thus, the present results endorse the likely clinical utility of an AF score in disease prediction. However, an optimised AF score for prediction in multi-ethnic populations

Table 4
Performance of the ML models for the prediction of ischemic stroke in AF patients, on the test dataset, under various metrics.

Models	AUROC (95% CI)	Accuracy	Precision	Recall	F1 score
XGBoost	0.631 (0.604–0.657)	0.63	0.63	0.63	0.63
LightGBM	0.620 (0.593–0.647)	0.62	0.62	0.61	0.62
RF	0.599 (0.573–0.625)	0.60	0.61	0.56	0.58
SVM	0.599 (0.572–0.624)	0.60	0.63	0.50	0.55
DNN	0.589 (0.562–0.615)	0.59	0.59	0.60	0.59
LR (L1 penalty)	0.563 (0.536–0.591)	0.56	0.56	0.56	0.56

AUROC, the area under a receiver operating characteristic curve; Accuracy = (TP + TN)/(TP + TN + FP + FN); Precision = TP/(TP + FP), Recall = TP/(TP + FN) where TP stands for true positive, TN for true negative, FP for false positive, and FN for false negative; F1 score = 2 (precision × recall)/(precision + recall).

Table 5
DeLong's test for the ML model comparisons for ischemic stroke prediction in AF patients.

Models	XGBoost	LightGBM	RF	SVM	DNN
XGBoost	–				
LightGBM	5.65E-01	–			
RF	1.33E-01	3.45E-01	–		
SVM	1.71E-01	3.75E-01	9.80E-01	–	
DNN	1.34E-01	2.89E-01	7.54E-01	7.45E-01	–
LR (L1 penalty)	2.00E-02	5.70E-02	2.56E-01	4.50E-01	2.54E-01

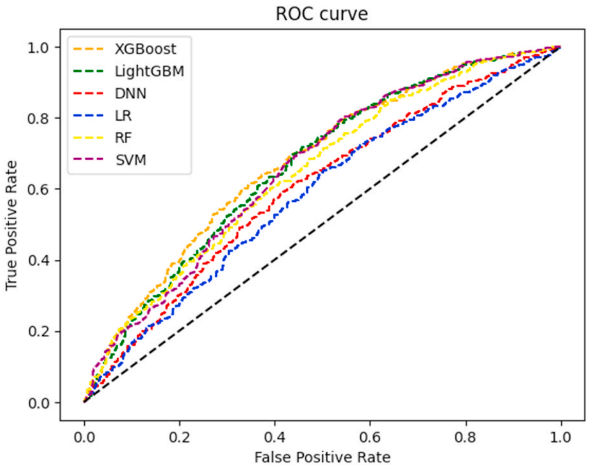

Fig. 4. AUROC for each ML model for predicting the development of ischemic stroke in AF patients, on the test dataset.

such as the UK population will be required prior to considering clinical use. Interestingly, standing height was ranked as the fourth most significant feature in LightGBM, which was the best performing model for AF prediction. Greater height has been identified as a risk factor for AF in several studies and in both males and females [27], and it is in agreement with the present analysis. Some studies report that taller people have greater heart chamber size [27], meaning a larger left atrial size, which may be potential explanation albeit not a very robust one as AF is driven by left atrial stretch and fibrosis. Two other anthropometric traits, weight and waist circumference, ranked just below standing height. Obesity is associated with increased risk of left atrial enlargement, atrial fibrosis, electrical derangements of the atria, impaired diastolic function, inflammation and accumulation of pericardial fat, which are all key mechanisms in the pathogenesis of AF [28], and it is supported by the present analysis. The ranking of sex as the seventh most significant feature in the model is also in agreement with epidemiological studies reporting sex differences in AF; males are at higher risk which is in agreement with the results, along with the electrophysiologic properties of the atria and structural remodelling [29]. The analysis presented here also found that participants with lower albumin levels had an increased risk of AF. This is in agreement with a meta-analysis revealing that an increase in albumin level decreased the risk of AF [30]. However, low albumin levels are associated with poor health overall and therefore we cannot exclude confounding. Among the remaining 20 most significant features in the model it is worth noting that (i) direct bilirubin has been reported as an important independent risk factor for AF development in both thyrotoxic patients [31] and a study in postoperative cardiac surgery [32], (ii) urate has been reported to increase the risk of AF and be causally associated to AF through MR analysis in Koreans [33], and (iii) the positive effect of increased testosterone on risk of AF has been reported in males but not in females in the ARIC study [34]; the present study corroborates these results. Finally, only two of the

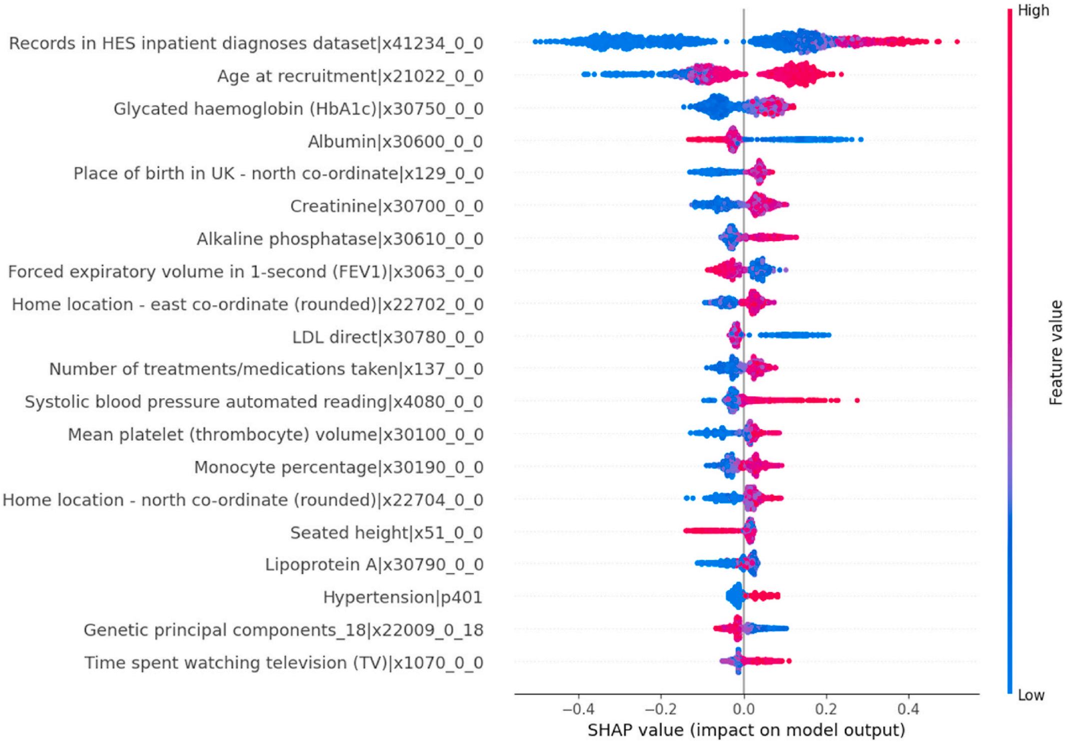

Fig. 5. Summary plot of the SHAP values (x-axis) for the top 20 features (y-axis), in descending order, showing the distribution of the impact that each feature has for the development of ischemic stroke in AF patients, on the test dataset, employing XGBoost model. Each dot represents a participant. The red dots represent a high feature value and blue dots represent a low feature value for each participant. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

20 top features have some conflicting data in the literature. FEV-1 levels have an increased risk of AF as shown in other studies [35], and it is corroborated by the present analysis, but the Korean National Health and Nutritional Examination Survey reported an adverse association between FEV-1 and AF development [36]. Decreased levels of triglycerides contribute to increased risk of AF, but a study in Chinese participants contradicts the present analysis, showing no evidence of association between triglycerides and incidence of AF [37].

4.3. *AF & ischemic stroke results*

In the present study, XGBoost model was the best in predicting ischemic stroke in AF patients and showed that it performs better than CHA$_2$DS$_2$-VASc, albeit marginal this result was statistically significant. Consistent with a recent French study for prediction of incident AF in a post-stroke population [38], the best performing ML model was DNN with a C index of 0.77 (95% CI 0.76–0.78) on the test set, performed better than CHA$_2$DS$_2$-VASc. In this study, XGBoost was identified as the best ML model for prediction of ischemic stroke in AF patients, with AUROC 0.631 (95% CI 0.604–0.657), in contrast to another two US studies that use more than 3.4 [39] and 6.4 [40] million participants, and reported c-index above 0.8. The lower performance of the ML model could be attributed to the fact that we used 6,300 participants in contrast to the million that were used in the US studies [39,40], thus leading to less power.

Unexpectedly, the genetic risk score for ischemic stroke, based on 28 genome-wide variants, was not among the top 20 features of the model, although ischemic stroke is highly heritable [41]. In the top 20 most significant features, medium to high feature values of HbA1c ranked third after sex and was associated with increased risk of stroke in AF patients. This agrees with the Clalit Health Services electronic medical records Israelian database, where participants with diabetes and AF were found to have an increased risk of stroke when their HbA1C levels were ranging from medium to high [42]. The fourth most significant feature was albumin which ranked ninth in the AF prediction model, suggesting a stronger relationship with ischemic stroke in AF patients than AF per se. This is corroborated by a Japanese study, which reported that lower albumin levels were associated with an increased risk of ischemic stroke in both sexes independently of AF status [43]. Four other blood biomarkers, creatinine, alkaline phosphatase, LDL cholesterol, and Lipoprotein A (Lp(a)) ranked among the top 20 features. These results are in agreement with the China National Stroke Registry reporting an association between high levels of alkaline phosphatase with recurrent stroke [44] and the Copenhagen General Population Study showing that high levels of Lp(a) were associated with increased risk of ischemic stroke [45]. It is worth noting that the latter although true for all examined ancestries it varies in strength e.g. higher in African than European Americans [46]. Interestingly, the use of creatinine as marker for increased risk of ischemic stroke in AF patients has not been previously reported and will merit further investigation. Lastly, the twentieth feature identified from the SHAP analysis – time spent watching television – could be considered as a surrogate marker for luck of sleep and physical inactivity; a recent study showed that physical inactivity increases the risk of stroke risk [47].

5. Conclusion

To conclude, there is a plethora of studies using ML methodology to predict circulatory diseases such as AF [3], cardiovascular disease [48], stroke [4,5], however none of them has the breadth and richness of electronic health record data that UK Biobank offers, including disease diagnosis, medications and laboratory tests. The strength of the present study is that makes use of the UK Biobank dataset, including up to 2,199 variables. The present study supports the incorporation of a few routinely measured blood biomarkers, whereas the results endorse the inclusion of a genetic score only in the model for AF prediction. The standardization of big data, along with the wide application of machine and deep learning methodologies, enables the identification of previously unknown risk factors for disease prediction. In the current study, the use of creatinine as marker for increased risk of ischemic stroke in AF patients has not been previously reported, however it requires further investigation. Machine learning models that employ large datasets, including potential predictors, can improve prediction accuracy, as presented in the current study, for the prediction ischemic stroke in AF patients using ML models in comparison to CHA_2DS_2-VASc, and provide graphical interpretation of the results using SHAP analysis. The models presented here have the potential for clinical use, but validation in further independent studies is required, since the models were developed and assessed in the UK Biobank and might not reflect other datasets with respect to age, sex, socio-economic status [49]. The models would need to be validated across all ancestries as some features vary by ethnicity e.g., Lp(a) and AF genetic score.

Funding source

AP is supported by National Institute for Health and Care Research Barts Biomedical Research Centre (NIHR203330).

Data availability

Individual level data could be accessed upon request and approval from UK Biobank. All the results discussed in this manuscript are available in the Supplementary Material.

CRediT authorship contribution statement

Areti Papadopoulou: Writing – review & editing, Writing – original draft, Visualization, Methodology, Formal analysis, Data curation. **Daniel Harding:** Writing – review & editing, Writing – original draft, Resources, Methodology. **Greg Slabaugh:** Writing – review & editing, Writing – original draft, Resources, Methodology, Investigation. **Eirini Marouli:** Writing – review & editing, Writing – original draft, Supervision, Methodology, Investigation, Conceptualization. **Panos Deloukas:** Writing – review & editing, Writing – original draft, Supervision, Methodology, Investigation, Funding acquisition, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.heliyon.2024.e28034.

References

[1] E.J. Benjamin, et al., Heart disease and stroke statistics-2019 update: a report from the American Heart Association, Circulation 139 (10) (2019) e56–e528.
[2] S. Khurshid, et al., Performance of atrial fibrillation risk prediction models in over 4 million individuals, Circ. Arrhythm Electrophysiol. 14 (1) (2021) e008997.
[3] S. Raghunath, et al., Deep neural networks can predict new-onset atrial fibrillation from the 12-lead ECG and help identify those at risk of atrial fibrillation-related stroke, Circulation 143 (13) (2021) 1287–1298.
[4] P.Y. Su, et al., Machine learning models for predicting influential factors of early outcomes in acute ischemic stroke: registry-based study, JMIR. Med. Inform. 10 (3) (2022) e32508.
[5] S. Jung, et al., Predicting ischemic stroke in patients with atrial fibrillation using machine learning, Front. Biosci. 27 (3) (2022) 80.
[6] H. Nishi, et al., Predicting cerebral infarction in patients with atrial fibrillation using machine learning: the Fushimi AF registry, J. Cerebr. Blood Flow Metabol.. 42 (5) (2022) 746–756.
[7] S.H. Kim, et al., Interpretable machine learning for early neurological deterioration prediction in atrial fibrillation-related stroke, Sci. Rep. 11 (1) (2021) 20610.
[8] S.M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, in: Proceedings of the 31st International Conference on Neural Information Processing Systems, 2017.
[9] L.A.C. Millard, et al., Searching for the causal effects of body mass index in over 300 000 participants in UK Biobank, using Mendelian randomization, PLoS Genet. 15 (2) (2019) e1007951.

[10] P. Wu, et al., Mapping ICD-10 and ICD-10-CM codes to phecodes: workflow development and initial evaluation, JMIR Med. Inform. 7 (4) (2019) e14325.
[11] C. Roselli, et al., Multi-ethnic genome-wide association study for atrial fibrillation, Nat. Genet. 50 (9) (2018) 1225–1233.
[12] R. Malik, et al., Multiancestry genome-wide association study of 520,000 subjects identifies 32 loci associated with stroke and stroke subtypes, Nat. Genet. 50 (4) (2018) 524–537.
[13] F. Pedregosa, et al., Scikit-learn: machine learning in Python, J. Mach. Learn. Res. 12 (2011) 2825–2830.
[14] G. Lemaître, F. Nogueira, C.K. Aridas, Imbalanced-learn: a python toolbox to tackle the curse of imbalanced datasets in machine learning, J. Mach. Learn. Res. 18 (1) (2017) 559–563.
[15] B. Krawczyk, et al., Evolutionary undersampling boosting for imbalanced classification of breast cancer malignancy, Appl. Soft Comput. 38 (2016) 714–726.
[16] M. AlJame, et al., Ensemble learning model for diagnosing COVID-19 from routine blood tests, Inform. Med. Unlocked 21 (2020) 100449.
[17] G. Lemaître, F. Nogueira, C.K. Aridas, Imbalanced-learn: a Python toolbox to tackle the curse of imbalanced datasets in machine learning, J. Mach. Learn. Res. 18 (17) (2017) 1–5.
[18] V. Berisha, et al., Digital medicine and the curse of dimensionality, NPJ Digit. Med. 4 (1) (2021) 153.
[19] R.-P. Ismael, et al., When is resampling beneficial for feature selection with imbalanced wide data? Expert Syst. Appl. 188 (2022) 116015.
[20] T. Chen, C. Guestrin, XGBoost: a scalable tree boosting System, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Association for Computing Machinery, San Francisco, California, USA, 2016, pp. 785–794.
[21] G. Ke, et al., LightGBM: A Highly Efficient Gradient Boosting Decision Tree, 2017.
[22] Y. LeCun, Y. Bengio, G. Hinton, Deep learning, Nature 521 (7553) (2015) 436–444.
[23] C. Cortes, V. Vapnik, Support-vector networks, Mach. Learn. 20 (3) (1995) 273–297.
[24] G. Van Rossum, F.L. Drake, The python Language Reference Manual, Network Theory Ltd, 2011.
[25] G.Y. Lip, et al., Refining clinical risk stratification for predicting stroke and thromboembolism in atrial fibrillation using a novel risk factor-based approach: the euro heart survey on atrial fibrillation, Chest 137 (2) (2010) 263–272.
[26] M.K. Chung, et al., Lifestyle and risk factor modification for reduction of atrial fibrillation: a scientific statement from the American Heart Association, Circulation 141 (16) (2020) e750–e772.
[27] C. Johansson, et al., Weight, height, weight change, and risk of incident atrial fibrillation in middle-aged men and women, J. Arrhythm. 36 (6) (2020) 974–981.
[28] T. Feng, et al., Weight and weight change and risk of atrial fibrillation: the HUNT study, Eur. Heart J. 40 (34) (2019) 2859–2866.
[29] S. Westerman, N. Wenger, Gender differences in atrial fibrillation: a review of epidemiology, management, and outcomes, Curr. Cardiol. Rev. 15 (2) (2019) 136–144.
[30] Y. Wang, et al., Relationship between serum albumin and risk of atrial fibrillation: a dose-response meta-analysis, Front. Nutr. 8 (2021) 728353.
[31] D. Sun, et al., Direct bilirubin level is an independent risk factor for atrial fibrillation in thyrotoxic patients receiving radioactive iodine therapy, Nucl. Med. Commun. 40 (12) (2019) 1289–1294.
[32] S.T. Turkkolu, E. Selcuk, C. Koksal, Biochemical predictors of postoperative atrial fibrillation following cardiac surgery, BMC Cardiovasc. Disord. 21 (1) (2021) 167.
[33] M. Hong, et al., A mendelian randomization analysis: the causal association between serum uric acid and atrial fibrillation, Eur. J. Clin. Invest. 50 (10) (2020) e13300.
[34] D. Berger, et al., Plasma total testosterone and risk of incident atrial fibrillation: the Atherosclerosis Risk in Communities (ARIC) study, Maturitas 125 (2019) 5–10.
[35] S.L. Au Yeung, et al., Impact of lung function on cardiovascular diseases and cardiovascular risk factors: a two sample bidirectional Mendelian randomisation study, Thorax 77 (2) (2022) 164–171.
[36] S.N. Lee, et al., Association between lung function and the risk of atrial fibrillation in a nationwide population cohort study, Sci. Rep. 12 (1) (2022) 4007.
[37] X. Li, et al., Lipid profile and incidence of atrial fibrillation: a prospective cohort study in China, Clin. Cardiol. 41 (3) (2018) 314–320.
[38] A. Bisson, et al., Prediction of incident atrial fibrillation in post-stroke patients using machine learning: a French nationwide study, Clin. Res. Cardiol. 112 (6) (2023) 815–823.
[39] G.Y.H. Lip, et al., Improving stroke risk prediction in the general population: a comparative assessment of common clinical rules, a new multimorbid index, and machine-learning-based algorithms, Thromb. Haemostasis. 122 (1) (2022) 142–150.
[40] G.Y.H. Lip, et al., Improving dynamic stroke risk prediction in non-anticoagulated patients with and without atrial fibrillation: comparing common clinical risk scores and machine learning algorithms, Eur. Heart J. Qual. Care Clin. Outcomes 8 (5) (2022) 548–556.
[41] J.W. O`Sullivan, et al., Combining clinical and polygenic risk improves stroke prediction among individuals with atrial fibrillation, Circ. Genom. Precis. Med. 14 (3) (2021) e003168.
[42] L. Kezerle, et al., Relation of hemoglobin A1C levels to risk of ischemic stroke and mortality in patients with diabetes mellitus and atrial fibrillation, Am. J. Cardiol. 172 (2022) 48–53.
[43] J. Li, et al., Serum albumin and risks of stroke and its subtypes- the circulatory risk in communities study (CIRCS), Circ. J. 85 (4) (2021) 385–392.
[44] L. Zong, et al., Alkaline phosphatase and outcomes in patients with preserved renal function: results from China national stroke registry, Stroke 49 (5) (2018) 1176–1182.
[45] P.R. Kamstrup, Lipoprotein(a) and cardiovascular disease, Clin. Chem. 67 (1) (2021) 154–166.
[46] P. Kumar, et al., Lipoprotein (a) level as a risk factor for stroke and its subtype: a systematic review and meta-analysis, Sci. Rep. 11 (1) (2021) 15660.
[47] P.T. Katzmarzyk, et al., Physical inactivity and non-communicable disease burden in low-income, middle-income and high-income countries, Br. J. Sports Med. 56 (2) (2022) 101–106.
[48] G. Joo, et al., Clinical implication of machine learning in predicting the occurrence of cardiovascular disease using big data (nationwide cohort data in Korea), IEEE Access 8 (2020) 157643–157653.
[49] A. Fry, et al., Comparison of sociodemographic and health-related characteristics of UK Biobank participants with those of the general population, Am. J. Epidemiol. 186 (9) (2017) 1026–1034.