Journal of the Neurological Sciences 440 (2022) 120335

Contents lists available at ScienceDirect

Journal of the Neurological Sciences

journal homepage: www.elsevier.com/locate/jns

Machine learning algorithms identify demographics, dietary features, and blood biomarkers associated with stroke records

Jundong Liu [a], Elizabeth L. Chou [b], Kui Kai Lau [c,d], Peter Y.M. Woo [e], Jun Li [f], Kei Hang Katie Chan [a,g,h,*]

[a] Department of Biomedical Sciences, Jockey Club College of Veterinary Medicine and Life Sciences, City University of Hong Kong, Hong Kong, China
[b] Massachusetts General Hospital, Boston, MA, USA
[c] Division of Neurology, Department of Medicine, The University of Hong Kong, Hong Kong, China
[d] The State Key Laboratory of Brain and Cognitive Sciences, The University of Hong Kong, Hong Kong, China
[e] Department of Neurosurgery, Kwong Wah Hospital, Hong Kong, China
[f] Department of Infectious Diseases and Public Health, Jockey Club College of Veterinary Medicine and Life Sciences, City University of Hong Kong, Kowloon, Hong Kong, China
[g] Department of Electrical Engineering, College of Engineering, City University of Hong Kong, Hong Kong, China
[h] Department of Epidemiology, Centre for Global Cardiometabolic Health, Brown University, RI, USA

ARTICLE INFO

Keywords:
Stroke record
Stroke prevalence
Machine learning
NHANES
Nomogram

ABSTRACT

Objective: We conducted a comprehensive evaluation of features associated with stroke records.
Methods: We screened the dietary nutrients, blood biomarkers, and clinical information from the National Health and Nutrition Examination Survey (NHANES) 2015–16 database to assess a self-reported history of all strokes (136 strokes, $n = 4381$). We computed feature importance, built machine learning (ML) models, developed a nomogram, and validated the nomogram on NHANES 2007–08, 2017–18, and the baseline UK Biobank. We calculated the odds ratios with/without adjusting sampling weights (OR/OR_w).
Results: The clinical features have the best predictive power compared to dietary nutrients and blood biomarkers, with 22.8% increased average area under the receiver operating characteristic curves (AUROC) in ML models. We further modeled with ten most important clinical features without compromising the predictive performance. The key features positively associated with stroke include age, cigarette smoking, tobacco smoking, Caucasian or African American race, hypertension, diabetes mellitus, asthma history; the negatively associated feature is the family income. The nomogram based on these key features achieved good performances (AUROC between 0.753 and 0.822) on the test set, the NHANES 2007–08, 2017–18, and the UK Biobank. Key features from the nomogram model include age ($OR = 1.05$, $OR_w = 1.06$), Caucasian/African American ($OR = 2.68$, $OR_w = 2.67$), diabetes mellitus ($OR = 2.30$, $OR_w = 1.99$), asthma ($OR = 2.10$, $OR_w = 2.41$), hypertension ($OR = 1.86$, $OR_w = 2.10$), and income ($OR = 0.83$, $OR_w = 0.81$).
Conclusions: We identified clinical key features and built predictive models for assessing stroke records with high performance. A nomogram consisting of questionnaire-based variables would help identify stroke survivors and evaluate the potential risk of stroke.

1. Introduction

Stroke is a leading cause of death and disability worldwide. Stroke survivors are at greater risk of suffering a second stroke. Stroke often results from the combined effects of genes and their complex interactions with environmental determinants [1]. Many epidemiological

Abbreviations: NHANES, National Health and Nutrition Examination Survey; ML, machine learning; UKB, UK Biobank; IF, isolation forest; H2O AI, H2O Driverless AI; DNN, Cost-Sensitive Neural Network; LR, logistic regression; AUROC, area under the receiver operating characteristic curves; AUCPR, area under the precision-recall curve; NRI, categorical net reclassification improvement; IDI, integrated discrimination improvement; SHAP, SHapley Additive exPlanations; HGB, histogram-based gradient boosting; FSRS, Framingham Stroke Risk Score.

* Corresponding author at: 1A-313 TYB, City University of Hong Kong, Hong Kong, China.
E-mail address: katie.kh.chan@cityu.edu.hk (K.H.K. Chan).

https://doi.org/10.1016/j.jns.2022.120335
Received 8 March 2022; Received in revised form 26 May 2022; Accepted 5 July 2022
Available online 9 July 2022

studies have offered valuable insight regarding the assessment of an individual's stroke risk. Traditional risk scores, including the Framingham Stroke Risk Score, the Cardiovascular Health Study (CHS), and the Atherosclerosis Risk in Communities (ARIC), have been extensively utilized for several years [2]. Machine learning (ML) has also been compared to such risk assessment metrics based on traditional regression models for stroke prediction. Studies have found that ML can integrate a vast amount of data, assist in identifying novel risk factors, achieve superior performance, and provide feature interpretations for prediction [3–5]. For example, T-wave abnormalities on electrocardiography and hematocrit levels were discovered to be novel risk factors by the ML approach. They improved predictive model performance, which depended on the existing revised Framingham stroke risk calculator [3].

However, fewer studies have adopted ML to explore the comprehensive features for those who have suffered a stroke event and stroke prevalence, possibly due to the cost or difficulty of collecting such data. Currently, the publicly available National Health and Nutrition Examination Survey (NHANES)[1] provides comprehensive cross-sectional information for United States residents. Cross-sectional studies are helpful in planning public health interventions [6]. Many studies [7–13] have taken advantage of the NHANES to explore cardiovascular diseases, diabetes mellitus, and all-cause mortality. In contrast, few have used a comprehensive comparison of features for stroke prevalence.

In this study, we utilized the NHANES database to study the features associated with stroke records. We hypothesized that diet, certain blood biomarkers, and other clinical data (including demographics, anthropometrics, and health status) could serve as independent features of stroke survivors. Our ML tool can be used to scrutinize and advance the existing support services for stroke survivors and their caregivers. Our study would provide a deeper insight into features associated with stroke, and offer clinicians a nomogram to detect stroke survivors and estimate stroke prevalence.

2. Materials and methods

2.1. Materials

2.1.1. Study population and outcome

Data were collected from NHANES, a research program gathering annual cross-sectional data from a nationally representative sample of approximately 5000 participants of all ages from the US population. In this study, the NHANES 2015–16 dataset was used for training and internal validation, while the NHANES 2007–08 and 2017–18 datasets and data at initial assessment for the UK Biobank (UKB) were utilized for external validation. UKB is a longitudinal large-scale and regularly augmented biomedical database, which collects genetic and health data from half a million UK participants [14]. The NHANES was approved by the National Center for Health Statistics Research Ethics Review Board, and all participants provided written informed consent. The UKB studies were approved by the National Health Service Research Ethics Service (11/N.W./0382). Access to the UKB was granted under application number 45788. Informed consent prior to enrollment in the study was obtained from all participants.

The correlated variables involved were based on stroke risk/protective factors and stroke symptoms and signs identified from the literature and covered by the NHANES 2015–16 database. Stroke was defined as a self-reported diagnosis by the participants. Self-reported stroke is defined as the participant who answered 'yes' to the NHANES questionnaire "Has a doctor or other health professional ever told you that you had a stroke?". Therefore, the positive samples are stroke survivors, while the negative samples are those who had not a stroke record before the survey. We did not distinguish between ischemic and hemorrhagic stroke or between diagnosis and recovery.

2.1.2. Features from established stroke risk factors and symptoms and signs

In this study, we involved the correlated variables from the documented factors present in NHANES. From the five NHANES 2015–2016 databases, namely Demographics Data, Dietary Data, Examination Data, Laboratory Data, and Questionnaire Data, we screened the dietary intake, blood biomarkers, demographics, anthropometrics, and health status as potential stroke risk/protective factors, according to the methodologies proposed by the American Heart Association[2] and elsewhere [15–21]. From our screenings, we defined four categories of features, dietary intakes (denoted by a diet set from 'Dietary Data'), blood biomarkers (blood set from 'Laboratory Data'), clinical features (clinical set from 'Demographics, Examination and Questionnaire Data'), and a combination of all aforementioned datasets (union set). In addition, we defined a second clinical set (clinical+ for short) by adding symptoms, complications, and the neurological deficits of stroke as well as possible confounders (e.g., medication information of insulin, anxiety, and cholesterol) to the clinical set in the sensitivity analysis to assess the robustness of the important clinical variables. These symptoms and signs of stroke are biologically associated with stroke survivors and prevalence. Therefore, we evaluated whether the essential risk factors could even be comparable to these post-stroke features in classification. The definitions of the above features were summarized and presented in Appendix A Data.

3. Methods

3.1. Data preprocessing and imbalanced classification

Each of the four datasets from the NHANES 2015–16 data was split into training (70%) and test sets (30%) by randomly stratified sampling of stroke. Missing values were filled by the Datawig algorithm [22], which was trained on the training set and applied to the training and test sets to avoid data leakage. However, variables with over 30% of their values missing were excluded. Subsequently, a feature selection algorithm, BoostARoota [23], was used to exclude the superfluous variables. Before applying BoostARoota, we processed the nominal variables (race and dichotomous variables) with one-hot encoding, which augmented the dimensions of variables, while ordinal features (lifestyle and stroke symptoms and signs) were regarded as integers. Continuous and ordinal variables were then processed with min-max scaling based on min-max values calculated from the training and applied to the training/test/external set. Imbalance classification models, which included an unsupervised algorithm, isolation forest (IF) [24], and three supervised algorithms, H2O Driverless AI (H2O) [25], Cost-Sensitive Neural Network [26] (DNN), and weighted logistic regression (LR) [27], were used to account for skewed stroke cases. Feature engineering, including imputation, feature selection, and standardization, was only used for training IF, DNN, and LR, since H2O could automate the whole process of end-to-end modeling. The hyperparameters for IF, DNN, and LR were optimized using cross-validation and Optuna [28]. Optuna can automate the search for various hyperparameters efficiently. It was set to maximize the area under the receiver operating characteristic curve of the training set with 100 trials; these hyperparameters included weight regularization, layers, units in each layer, and epochs for DNN; solver and penalty for LR; and the number of base estimators in the ensemble, samples, and features to draw from the training set to train each base estimator for IF. Fig. 1 summarizes the aforementioned modeling steps in a flow chart. The preprocessing of UKB data is presented in Fig. A1.

[1] https://www.cdc.gov/nchs/nhanes/index.htm

[2] https://www.stroke.org

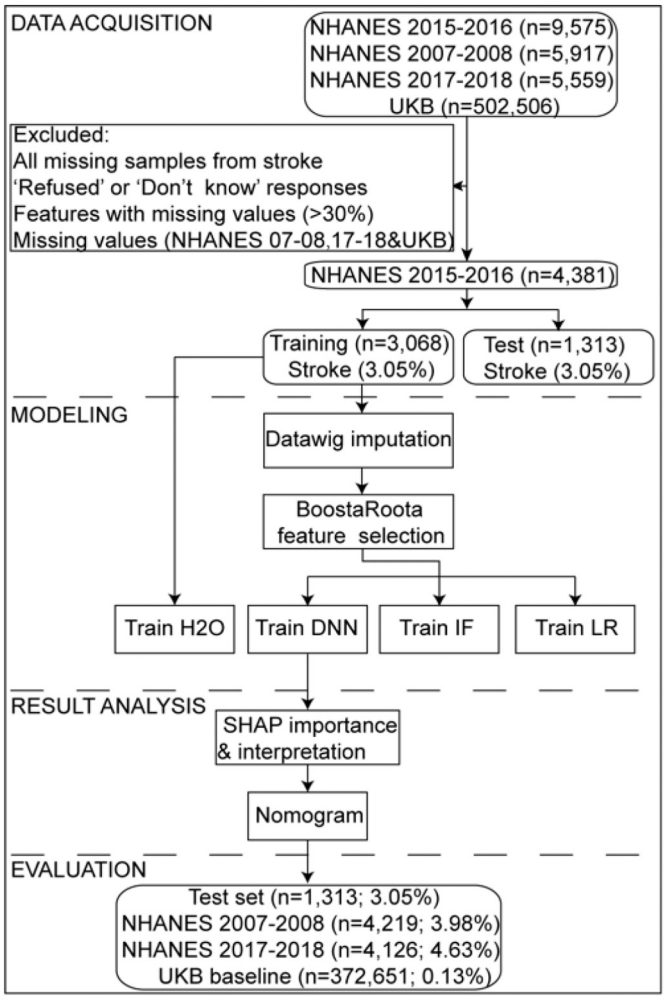

Fig. 1. Flow chart of modeling. The sample size (n) and stroke prevalence rate in percentage are shown inside the parentheses. NHANES, the National Health and Nutrition Examination Survey; H_2O, H_2O Driverless AI; DNN, deep neural network; IF, isolation forest; SHAP, SHapley Additive exPlanations; UKB, UK Biobank.

3.2. Metrics computation

Model evaluation was mainly based on the following metrics: area under the receiver operating characteristic curve (AUROC) and the precision-recall curve (AUCPR) [29]. The AUC is often considered a reliable performance metric for imbalanced binary classification problems [30]. For further comparison of important feature performance, we added categorical net reclassification improvement (NRI) and integrated discrimination improvement (IDI). The R package 'pROC' was used to calculate the optimal cutoff threshold from the Youden index and AUROC test statistics between models [31], while 'PredictABEL' [32] was used for IDI/NRI calculation. For AUPRC, the stratified bootstrap differences between models and the corresponding *p*-values (two-sided) were calculated by ourselves using the 'boot' package [33]. The threshold for significance is 0.05. For IF, LR, and DNN, we repeated the modeling 40 times to obtain the stable means for AUROC/AUCPR and score to calculate NRI/IDI. R Studio (v1.2.5001), H2O Driverless AI (v1.8.1), and Python (v3.7.3) were used for the above analyses. The positive and negative predictive values (PPV and NPV, respectively) were also provided.

3.2.1. Feature explanations and nomogram based on SHapley additive exPlanations

To apply the results from the modeling for practical use, we further limited the variables by selecting the top ten variables based on SHapley Additive exPlanations values (SHAP) [34]. The set of ten variables was denoted by the SHAP set in the analysis. As shown in the result described in Section 3, we selected these key contributors from the clinical set without compromising the performance. Then, we applied SHAP to interpret the variables for the individuals with/without a prior stroke. Furthermore, we created a nomogram [35] and validated its performance based on these variables on the test set, the NHANES 2007–08, 2017–18, and the UKB data. The interpretable variables and simple-to-use nomogram provide understandable characteristics for patients and facilitate discrimination of stroke survivors and estimates of stroke prevalence, which could be of clinical significance.

3.2.2. Comparison and bias assessment using SHAP set variables

To compare the performance and evaluate the robustness of our SHAP set features, we applied NRI and IDI to compare the SHAP set with all datasets. We even added the stroke symptoms and signs to the clinical set to assess whether the SHAP set variables could be comparable to well-known symptoms and signs in poststroke patients as sensitivity analysis. We also compare the SHAP variables with the set of risk factors used in the Framingham Stroke Risk Score (FSRS) [36] using logistic regression. We utilized the 'Caret' R package [37] to preprocess the data (k nearest neighbor missing value imputation and standardization) to avoid data leakage and assessed the performance in cross-validation with three folds and five repeats. We further explored the bias of the SHAP set variables by considering the adjustment of sampling weights and unlabeled data on the test set, the NHANES 2007–08 and 2017–18, and UKB databases. As NHANES uses a complex survey design, we involved multistage stratified weights in a generalized linear model to adjust the national population's effect size. Meanwhile, we explored the raw weighted SHAP set characteristics to assess their time differences across the NHANES 2015–16, 2007–08, and 2017–18, and the repeated measures in longitudinal UKB database. To take into account the impact of a large amount of excluded unlabeled data (with 'Don't know' or missing values for stroke status), we constructed a semi-supervised model based on histogram-based gradient boosting (HGB) [38]. We also built the supervised HGB for comparison. Moreover, to assess the bias from the use of different splits of data for training, we applied HGB to compute the out-of-fold predictions on the whole NHANES 2015–16 data and calculated the five-fold AUROC. Optuna and k-nearest neighbor algorithm [39] were applied to hyperparameters optimization and missing value imputation for the semi-supervised and HGB models.

Finally, considering NHANES is limited to all strokes, which is self-report, we utilized the UKB data to assess the bias of stroke subtypes and recall bias. Based on the self-report stroke status [40], we supplemented the subtype information by the hospital inpatient data, which provides the International Classification of Diseases, Ninth and Tenth Revision (ICD9 and 10) and Office of Population, Censuses and Survey (OPCS4) [41,42] for patients on hospital admissions. We selected the initial stroke records to avoid bias resulting from recurrent stroke. We dichotomized each subtype, including AS, and computed AUROC using the nomogram.

4. Results

4.1. Datasets and comparison of models

We retrieved 4381 observations (135 cases with a previous stroke) for the diet set (41 variables), the blood set (30 variables), the clinical set (34 variables), the union set (122 variables), and the clinical+ set (51 variables) from the NHANES 2015–16 data (Table A1). For the NHANES 2007–08 and 2017–18 data, we retrieved 4219 (168 cases) and 4126 (191 cases) samples for validation. The UKB baseline data have dozens

of times sample size than that of the NHANES data, consisting of 372,651 (467 cases) samples (Table 3).

For each category of feature applied to the NHANES 2015–16 data, models were trained on 3068 (95 cases; 8.10% mean missing rate) and tested on 1313 (40 cases; 7.08% mean missing rate) samples (Fig. 1). The results from Fig. 2 depict a pattern with increasing AUROC and AUCPR among the diet, blood, and clinical/union sets. Table 2 represents the confirmation of such a significant difference between the diet set and the clinical/union set in both statistics gained from the AUROC and AUCPR tests. However, between the blood and clinical/union sets, significant differences were only seen in the AUROC test, i.e., the clinical/union sets blood set significantly/non-significantly outperformed the blood set. For the clinical set and the union set, the results were not significantly different. Thus, the clinical variables are more powerful than or at least comparable to the other dataset variables in the classification.

We used the DNN model trained on the clinical set to calculate Deep SHAP [34] values, among which the top ten features (shown in Fig. 3A) were selected. These ten features (the SHAP set) can easily be obtained from patients via the questionnaire, and they are age, monthly family income, cigarette smoking, tobacco smoking, two race variables (Caucasian and African American), hypertension, diabetes mellitus, asthma history, and snore frequency. On the test set, the AUROCs resulting from these SHAP set variables were 0.704 (0.678–0.724), 0.817 (0.785–0.849), 0.797 (0.794–0.800), and 0.753 (0.731–0.774) using the IF, H2O, DNN, and LR models respectively; those for the clinical set are 0.731 (0.724–0.738), 0.788 (0.747–0.829), 0.779 (0.771–0.787), and 0.785 (0.749–0.797) respectively (Table 1). AUROC and AUCPR test statistics both showed a non-significant difference between the SHAP and the clinical sets (Table A3). Furthermore, we computed the IDI and NRI values to quantify the improved discrimination based on the SHAP set models compared to each of the diet, blood, clinical, and union sets models. Positive (negative) IDI/NRI, with a significant 95% confidence interval and *p*-value, indicates higher (lower) performance from the SHAP set. For NRI, the thresholds were derived from the training data using the SHAP set variables. As can be seen in Table 1 or Fig. A2, the results, calculated from SHAP set over the diet, blood, clinical, and union sets, respectively, were significantly positive or non-significantly negative, indicating SHAP set variables improved predictive performance, compared to other sets. Therefore, it seems that we can apply the ten SHAP set features from the clinical variables to assess the probability of individuals who reported a stroke without compromising the performance.

In the sensitivity analysis, the addition of the covariates in poststroke (symptoms and signs) to the clinical set. For example, the clinical+ set, resulted in the AUROC values of 0.772 (0.768–0.777), 0.811 (0.774–0.848), 0.804 (0.798–0.808), and 0.799 (0.757–0.814), and AUCPR values of 0.101 (0.098–0.106), 0.131 (0.087–0.174), 0.154 (0.135–0.199), and 0.148 (0.088–0.187) from IF, H2O, DNN, and LR models, respectively (Table A2). Nevertheless, the test statistics showed the above AUROC and AUCPR were all non-significant in comparison to those in the SHAP set (Table A4, all $p > 0.05$). We also compared the values garnered from the SHAP set with those from the clinical+ set using IDI and NRI. As shown in Table A2 and Fig. A2, out of the eight index results, there are three significant (IDI: 3.4% (H2O), 2.6% (DNN), $p < 0.01$; NRI: 20.8% (DNN), $p < 0.01$) and two non-significant (NRI: 0.085, $p = 0.329$ (IF); 0.053, $p = 0.152$ (H2O)) improvements favoring the SHAP set models. Only one improved discrimination (IDI: −3.4%, $p < 0.01$ (IF)) favored the clinical + set model. Therefore, in our supervised models, it could benefit more from the SHAP set variables. The above results suggest that we only need to apply these ten questionnaire-based SHAP set variables to provide comparable performances.

From the above results, we could infer that clinical features play an important role in classification. The top ten features derived from the clinical set, i.e., the SHAP set, should be comparable to other datasets in detecting the stroke survivors, thereby clinically informative. Besides the clinical signs and symptoms of stroke, which are informative for the medical practitioners, these ten features might also be worth the attention of medical practitioners/caregivers/researchers in the stroke risk studies.

4.2. Interpretations of key features, utility, and nomogram

Fig. 3B shows an overview of feature interpretations for the top 20 features over the population of training data, i.e., global explanations. The features are sorted by the sum of SHAP values. Dots show the distribution of the impacts that a feature has on the model output and indicate the density when piling up, with colors representing the feature value (red for high, blue for low). We could also visualize how these features affect the model arriving at a score for an individual via individual/local explanation. For example, Figs. 3C and 4 show a 'force plot' [43] from SHAP values and its 'waterfall plot' [44] to interpret the score for an individual and quantify each feature's contribution. Based on the derived threshold of 0.353, the higher score of 0.41 suggests a case of a stroke survivor. With the set of features that increase (red) and decrease (blue) the risk, we could observe that, although the patient was young, 27 years of age, without hypertension, which lowers the probability of a positive case, his/her relatively low family income, asthma history, tobacco smoking, and Caucasian race increased the likelihood of a post-stroke event.

The features in **(A)** and **(B)** are sorted by the absolute and the mean Deep SHAP magnitudes respectively for DNN trained on the clinical set. For each feature, dots show the distribution of the effects of the feature on the model output and indicate the density when piling up, with colors representing the feature value (red for high, blue for low). Features in **(C)** that increase the risk of stroke are in red, while those that decrease the risk are in blue. Under a cutoff at 0.353 derived from the Youden index derived, the outcome can be classified as positive because the relatively low family income, asthma, tobacco use, and Caucasian increase the probability.

Next, we further exploited the SHAP features to develop a nomogram. Fig. 5 presents the nomogram for practitioners to calculate the score manually. In addition, an online dynamic nomogram [45] tool was generated for online users (https://jdliu4-c.shinyapps.io/DynNomapp/). Because nomograms are based on logistic regression in our study, we deleted 'Cigarette use' and preserved the more generalized 'Tobacco use' to avoid the possible strong correlation. 'Snore frequency' and 'Race' were dichotomized by snoring over or less than five nights a week and being non-Hispanic white/Black or not. However, for UKB, which included twenty ethnic backgrounds, we defined all non-Caribbean White/Black as 'Caucasian/African American'. Table 3 shows the variables used in the nomogram. Using the nomogram, the AUROC on the test set, the NAHANES 2007–08 and 2017–18 data and the UKB dataset were 0.800 (0.739–0.855), 0.827 (0.801–0.852), 0.822 (0.797–0.846), and 0.753 (0.731–0.774) respectively (Table A5 or Fig. A3-A). AUROC between 0.7 and 0.8 is considered acceptable, 0.8 to 0.9 is considered excellent, and over 0.9 is considered outstanding [46]. Considering that the sample size of the UKB database (372,651) is dozens of times larger than that of the NHANES datasets and included many more distinct ethnic backgrounds, the performance was still reasonable and acceptable. These results show that our selected features have strong discriminative power and indicate that the nomogram could be applied in practice with decent performance compared with the machine learning models.

Our findings of these SHAP set variables could result in some aspects of utility. Our model based on the SHAP features would lower the time and financial cost of information retrieval and reduce the information noise without compromising the predictive power. The simplified and accurate nomogram would make the risk assessment and early warning of stroke feasible in a timely manner, especially in low-income countries. For stroke patients, these most stroke-relevant factors can lower the chance of a second/third stroke by recommending lifestyle modification

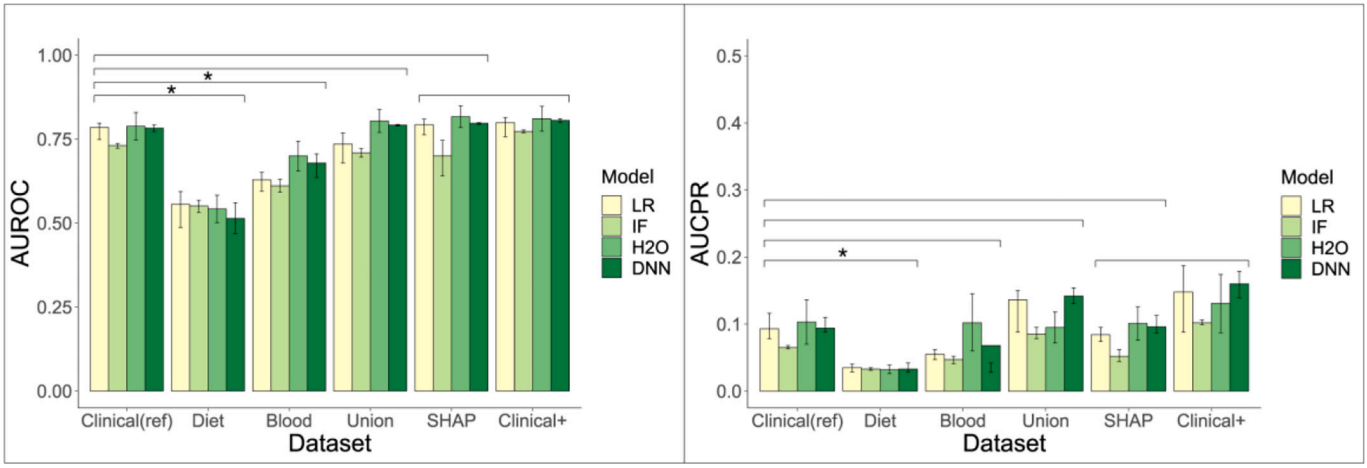

Fig. 2. AUROC and AUCPR of diet, blood, clinical, and union sets. **(A)** AUROC; **(B)** AUCPR. AUROC, area under the receiver operating characteristic curve; AUCPR, area under the precision-recall curve. *: significant test statistics for AUROC/AUCPR (p-value <0.05) between models.

or improving the post-stroke health management. For example, diabetes mellitus is associated with poorer cognitive function and higher mortality in poststroke patients. [47] After an acute stroke, patients with diabetes are more likely to suffer from hyperglycemia. [47] Health care providers should pay more attention to stroke survivors with diabetes and ensure their mental competency to fulfill the complex management of diabetes.

For the NHANES 2007–08, 2015–16, and 2017–18 datasets, income is the encoded monthly and annual family income, ranging from 1 to 12 (we recoded 12 for those over 12 for the NHANES 2017–18 data). For the UKB data, income is the encoded average total household income before tax ranging from 1 to 5, and we recoded them as 2.4, 4.8, 7.2, 9.6, and 12, respectively. The descriptive statistic was mean (standard deviation), median (interquartile range) or absolute frequency (relative frequency) in terms of the row-variable considered to be continuous normal-distributed, continuous non-normal distributed, or categorical by R 'compareGroups' package.[3]

Moreover, we also computed the odds ratio (OR) using LR and LR with sampling weights adjusted (OR_w) on the training set. The consistently significant ORs (Table 4) were from age (OR = 1.05 (1.03,1.07), OR_w = 1.06 (1.03,1.08)), Caucasian/African American (OR = 2.68 (1.58,4.53), OR_w = 2.67 (1.71,4.17)), diabetes mellitus (OR = 2.30 (1.44,3.67), OR_w = 1.99 (1.04,3.84)), asthma (OR = 2.10 (1.26,3.48), OR_w = 2.41 (1.35,4.32)), HBP (OR = 1.86 (1.11,3.13), OR_w = 2.10 (1.04,4.25)), and income (OR = 0.83 (0.77,0.90), OR_w = 0.81 (0.73,0.89)); tobacco (OR = 1.64 (1.01,2.68)) became non-significant involving sampling weights (OR_w = 1.64 (0.97,2.79)). Therefore, it seemed that the most robust features associated with stroke prevalence were age, income, diabetes, and asthma via linear modeling.

4.3. Bias assessment results of SHAP set features

4.3.1. Involving sampling weights to adjust population effect size

When sampling weights were adjusted, our SHAP set variables still achieved high AUROC on the test set, the NHANES 2007–08 and 2017–18 data (AUROC 0.824 (0.764–0.875), 0.825 (0.799–0.851), and 0.810 (0.783–0.835), respectively from Table A5 or Fig. A3-A), which meant these variables still play an important part in the US national population setting. Simultaneously, we assessed the selection bias: the test set performance showed that after adjusting the sampling weights, it might benefit more from the variables (AUROC 0.824 (0.764–0.875) as compared to LR 0.793 (0.763–0.809) without weights). On the other hand, on the whole population level, the mean (%) stroke records for NHANES 2007–08, 2015–16, and 2017–18 were 6,784,293.16 (3.2%), 6,369,193.13 (2.7%), and 7,890,966.34 (3.3%) (Table A7). The patients had average age (SD) of 64.66 (14.04), 63.89 (14.39), and 65.19 (12.44). Hypertension, diabetes, cigarette use, and asthma remained significant, while snoring, income, and tobacco use remained non-significant. We observed some features might change. Race was not found to be significantly different across 2007–08 and 2015–16; it was significant (p = 0.037) in 2017–18. Some variables' proportions changed by over 5%. Diabetes (%) was 31.2%, 29.9%, and 36.8%, and asthma 21.3%, 29.4%, and 25.8% for the three 2-year cycles. In our longitudinal UKB (N = 502,506) data, the mean age, age of stroke, and age of hypertension increased from 56.5 to 65.1, 51.6 to 61.1, and 45.5 to 50.2 in the four collections. Asthma age fluctuated between 27.6 and 29.4. Participants with snoring habits also fluctuated between 32.0% and 34.6%. Current tobacco smoking on most or all days decreased from 7.82% to 1.21%.

Involving 3856 missinAg labels and 5 'Don't know'/'Refuse to answer' in the semi-supervised HGB model resulted in an AUROC of 0.813 (0.749–0.870), while excluding them led to an AUROC of 0.793 (0.727–0.854) (0.02 AUROC improved) on the test set. Considering the sample size of unlabeled data is larger than that of the training set, it seems we did not lose much information from unlabeled data. Finally, with the same set of hyperparameters and model of HGB trained on the NHANES 2015–16 data, the out-of-fold score AUROC was 0.808 (0.779–0.836), which was near the test set AUROC, 0.793 (0.727–0.854). So, it seems our result could be stable compared to those using different splits of data (Table A5 or Fig. A3-A).

4.3.2. Comparison with FSRS variables

From NHANES 2015–16 (4379 samples with 134 cases), we screened and recoded age, systolic blood pressure, diabetes mellitus, cigarette smoking, cardiovascular disease history, atrial fibrillation, and use of hypertensive medication as required by FSRS (Table A6). However, we lacked the information on left ventricular hypertrophy. We also extracted the same participants (4379 samples with 134 cases) with SHAP set variables (Table A6). AUROC from cross-validation (3 folds and 5 repeats) showed SHAP set variables achieved 0.825 (0.790–0.858) AUROC, higher than 0.736 (0.700–0.778) of the FSRS. This demonstrated that our SHAP variables are more informative than the risk factors from FSRS in the cross-sectional design.

4.3.3. Nomogram on stroke subtypes of UKB

Out of 372,651 participants, 5796 (1.56%) participants reported no stroke while they had hospital inpatient stroke records. We updated

[3] https://cran.r-project.org/package=compareGroups

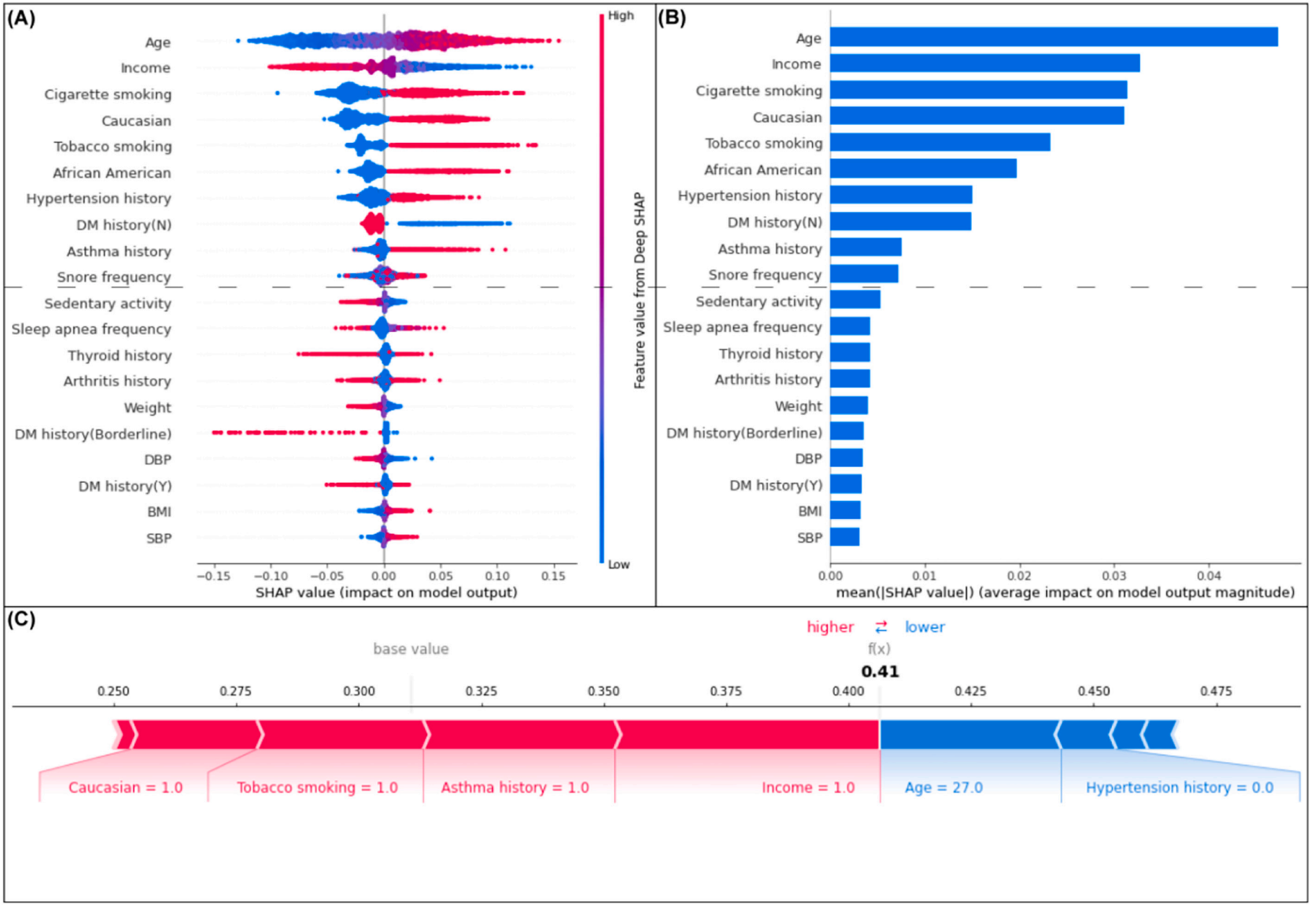

Fig. 3. Global explanation over training samples **(A-B)** and individual explanation for a test set example **(C)**. **(A)** SHAP bar plot; **(B)** SHAP plot. Age (year); Income: recoded as 1 to 12 to indicate 0$ to $8400 and over; Cigarette smoking (Y/N); Tobacco smoking (Y/N); Caucasian (Y/N); African American (Y/N); hypertension history ("Have you ever been told by a doctor or other health professional that you had hypertension, also called high blood pressure?") (Y/N); DM: diabetes mellitus ("Have you ever been told by a doctor or health professional that you have diabetes or sugar diabetes?") (Y/N/Borderline); Asthma history ("Has a doctor or other health professional ever told you that you had asthma?") (Y/N); Snore frequency ("How often do you snore?"); Sedentary activity ("How many minutes do you usually spend sitting on a typical day?); Sleep apnea frequency ("How often do you snort or stop breathing?") (Y/N); Thyroid history ("Has a doctor or other health professional ever told you that you had another thyroid problem?") (Y/N); Arthritis history ("Has a doctor or other health professional ever told you that you had arthritis?"); Weight (kg); DBP: diastolic blood pressure (mm Hg); BMI: Body Mass Index (kg/m^2); SBP: systolic blood pressure (mm Hg).

Table 1
Performances of datasets.

Dataset	Model	AUROC	AUCPR	IDI	Categorical NRI	NPV	PPV
Union	LR	0.735	0.136	0.0414(0.373)	0.176(0.017)	0.99	0.055
	CI	0.679–0.768	0.088–0.150	−0.050-0.133	0.032–0.320	0.981–0.996	0.051–0.139
Clinical	LR	0.785	0.093	0.007(0.744)	-3e-04(0.994)	0.991	0.077
	CI	0.749–0.797	0.078–0.116	−0.037-0.052	−0.070-0.069	0.986–0.997	0.058–0.109
Diet	LR	0.556	0.035	0.284(<0.001)	0.473(<0.001)	0.978	0.041
	CI	0.487–0.594	0.029–0.040	0.195–0.374	0.300–0.645	0.973–0.990	0.034–0.045
Blood	LR	0.629	0.055	0.215(<0.001)	0.365(<0.001)	0.981	0.063
	CI	0.595–0.651	0.047–0.062	0.152–0.278	0.219–0.510	0.976–0.989	0.042–0.103
SHAP	LR	0.793	0.084			0.995	0.068
	CI	0.763–0.809	0.074–0.095			0.988–0.999	0.062–0.117
Union	IF	0.711	0.091	0.006(0.452)	0.292(<0.001)	0.984	0.069
	CI	0.696–0.720	0.086–0.096	−0.010-0.022	0.148–0.437	0.980–0.992	0.047–0.073
Clinical	IF	0.731	0.065	−0.011(0.207)	0.240(0.006)	0.987	0.062
	CI	0.724–0.738	0.062–0.070	−0.028-0.006	0.070–0.411	0.983–0.998	0.046–0.090
Diet	IF	0.551	0.033	0.035(<0.001)	0.300(<0.001)	0.98	0.036
	CI	0.529–0.567	0.031–0.035	0.020–0.049	0.154–0.443	0.975–1	0.034–0.059
Blood	IF	0.610	0.047	0.029(<0.001)	0.292(<0.001)	0.978	0.055
	CI	0.591–0.633	0.041–0.056	0.019–0.038	0.147–0.436	0.976–1	0.035–0.082
SHAP	IF	0.704	0.051			0.995	0.054
	CI	0.678–0.724	0.046–0.055			0.989–1	0.049–0.067
Union	H2O	0.804	0.095	0.038(<0.001)	0.122(<0.001)	0.99	0.091
	CI	0.770–0.838	0.072–0.118	0.024–0.052	0.051–0.193	0.986–0.997	0.058–0.144
Clinical	H2O	0.788	0.103	0.032(<0.001)	0.084(0.020)	0.994	0.072
	CI	0.747–0.829	0.070–0.136	0.019–0.045	0.013–0.154	0.987–0.998	0.065–0.125
Diet	H2O	0.542	0.032	0.049(<0.001)	0.506(<0.001)	0.985	0.039
	CI	0.501–0.583	0.026–0.039	0.032–0.065	0.386–0.627	0.975–0.995	0.035–0.050
Blood	H2O	0.700	0.102	0.036(<0.001)	0.273(<0.001)	0.982	0.072
	CI	0.655–0.744	0.060–0.145	0.021–0.051	0.132–0.414	0.979–0.993	0.046–0.130
SHAP	H2O	0.817	0.101			0.99	0.102
	CI	0.785–0.849	0.076–0.126			0.987–0.999	0.059–0.133
Union	DNN	0.793	0.146	−0.021(0.091)	0.390(<0.001)	0.99	0.102
	CI	0.791–0.794	0.135–0.153	−0.045-0.003	0.243–0.537	0.987–0.999	0.059–0.133
Clinical	DNN	0.779	0.088	0.039(<0.001)	0.495(<0.001)	0.993	0.071
	CI	0.771–0.787	0.083–0.093	0.030–0.048	0.369–0.622	0.986–0.999	0.058–0.127
Diet	DNN	0.503	0.035	0.064(<0.001)	0.510(<0.001)	0.991	0.068
	CI	0.479–0.554	0.030–0.040	0.032–0.096	0.384–0.637	0.987–0.999	0.056–0.099
Blood	DNN	0.669	0.072	0.034(<0.001)	0.400(<0.001)	0.973	0.042
	CI	0.547–0.710	0.044–0.092	0.016–0.052	0.253–0.547	0.971–1	0.031–0.109
SHAP	DNN	0.797	0.096			0.991	0.077
	CI	0.794–0.800	0.089–0.112			0.988–0.999	0.059–0.119

Categorical NRI, categorical net reclassification improvement; IDI, integrated discrimination improvement; CI: 95% confidence interval. For IDI and NRI, improvements were shown from the SHAP set over other sets (*p* value in brackets).

these 5796 labels and achieved 6173 stroke records of all types (AS). By this adjustment, we could assess the reporting bias. Self-reporting stroke (AS report, $N = 467$) AUROC reduced from 0.753 to 0.7 after correction (AS, $N = 6173$) (Fig. A3-B). Our nomogram still achieved an AUROC of >0.7.

Stroke subtypes included ischemic stroke (IS, $N = 2975$), other strokes (OS, $N = 2808$), and hemorrhagic stroke (HS, $N = 570$). From Fig. A3-B, IS and OS had AUROC of ~0.7, inferior to AS report but similar to AS. Our nomogram presents a similar discrimination ability for ischemic stroke and other strokes. The predictive performance for hemorrhagic stroke was lower (AUROC: 0.623) than the other types ($AUROC_{IS} = 0.702$; $AUROC_{OS} = 0.711$), while it still presented certain diagnostic ability [48].

5. Discussion

This study compared the ability of NHANES-derived dietary nutrients, blood biomarkers, clinical features, and the combination of these three data domains to classify individuals with a prior stroke or not using three ML models. We discovered that dietary nutrient intake contributed the least to the performance, followed by blood biomarker data. Models based on clinical features showed no difference in performance compared to those based on a combination of all three data domains. We subsequently extracted the ten most powerful features. Our external validation study showed that the performance from these features could be generalized. Subsequently, we summarized the profiles of training samples, provided a specific risk stratification for an individual, and developed a nomogram to facilitate the manual classification of stroke records. One main objective of this study is to provide highly accurate predictive models based on much fewer and more informative features without compromising the performance compared to the ML model using all the features. These set variables, which are also relatively easy to collect, can be used for the physicians or health authorities to have a preliminary and quick risk assessment on potential stroke status of high-risk patients. Since some of the variables have not been thoroughly investigated as risk factors for stroke (e.g., asthma [49] and snoring [50]), these SHAP features would also inspire future large-scale longitudinal studies or experiments to validate the potential causality or reveal the underlying mechanism.

We referred to NHANES studies and compared the diet and blood features. However, features related to stroke records have been seldomly explored in NHANES. In an NHANES cardiovascular disease (CVD) and diabetes study [9], we still found that top ten important features were almost clinical features, instead of diet or blood features; top fine were all clinical features. Therefore, we still did not identify any diet and blood biomarkers that contributed much to stroke records while they could have high predictive power in mortality. On the other hand, different forms of nutrient representations, e.g., nutrition indices, might be helpful. In a study on NHANES data, dietary features were shown to have a strong influence on stroke-related mortality [51], which contradicts our findings where diet was observed to have a lower significance. However, that study also included broad cardiometabolic food

Table 2
AUROC and AUCPR test statistics for diet, blood, clinical, and union sets.

		Union				Clinical				Diet				Blood			
		LR	IF	H2O	DNN	LR	IF	H2O	DNN	LR	IF	H2O	DNN	LR	IF	H2O	DNN
Union	LR					0.582				0.027				0.098			
	IF						0.239				0.032				0.080*		
	H2O							0.814				0.002				0.968*	
	DNN								0.117				0.019				0.124
Clinical	LR	0.078								0.025				0.109			
	IF		0.556								0.036*				0.241*		
	H2O			0.358								0.007				0.945	
	DNN				0.608								0.020				0.385
Diet	LR	0.005				<0.001								0.171			
	IF		<0.001				0.001								0.287		
	H2O			<0.001				<0.001								0.045	
	DNN				<0.001				<0.001								0.052
Blood	LR	0.007				<0.001				0.290							
	IF		0.012				0.007				0.368						
	H2O			<0.001				0.012				0.010					
	DNN				0.028				0.041				0.003				

The underline indicates significance.
* p values below and above the diagonal are for AUROC and AUCPR tests.

categories such as fruits, vegetables, and unprocessed meats as part of its dietary intake analysis, while the current study only utilized dietary nutrient and supplement data. Another study examined if CVD mortality prediction could benefit from nutrition data through ML algorithms; it involved all nutrition variables, including micronutrients (e.g., sodium and selenium), macronutrients (e.g., fat, carbohydrates, and protein), and commonly utilized composite nutrition indices (e.g., Alternate Healthy Eating Index, Mediterranean Diet Score, and the Dietary Approaches to Stop Hypertension diet score) [10]. The investigators revealed that micronutrients and macronutrients, instead of nutrition indices, improved the predictive capacity of ML-based models. However, when adopting conventional Cox modeling, such nutrient information was found to have little contribution to stroke prediction. In light of these inconsistent findings, we postulate that incorporating macronutrients and composite nutrition indices in the dietary dataset could help reexamine the importance of diet in stroke prevalence. On the other hand, in our H2O model, protein and vitamin B6 were ranked fourth and ninth out of their top ten features. Therefore, different forms of nutrients and their transformations in different models warrant further investigation.

Agreements between ML model performances founded on laboratory- or non-laboratory-based information have been observed in previous studies that utilized NHANES data for CVD mortality predictions [52,53]. In those studies, the definition of non-laboratory mainly included clinical features plus dietary nutrients, while laboratory data included clinical features plus blood/urine biomarkers; therefore, they had an overlap of clinical features. Specifically, a recent study [9] on diabetes mellitus and CVDs using the NHANES compared the non-laboratory with the laboratory dataset. The investigators noted a significant influence on CVD prediction from their non-laboratory variable-based model, while the laboratory-based model did not enhance performance. This was similar to our findings in that the addition of blood biomarkers and diet data did not significantly improve performance when comparing the union- with the clinical-dataset-based models. We made an extra comparison of the modified diet and blood set. We added the clinical set to both the diet and the blood set, and then found the AUROC and AUCPR were non-differential (results not shown). It meant that the overlapped non-laboratory variables might account for the non-differential performances. Moreover, in that recent study [9], AUROC values were high (0.816–0.839). Two reasons could possibly explain it. The first is the data linkage from the dataset preprocessing step. The normalization was before the train-test split, which could result in unreliable better test performances because the training and the test set shared the same prior information (mean and standard deviation) from the original data for normalization. The second underlying cause is its moderately imbalanced CVDs data (17% vs. ours 3%). Classifiers could benefit more from moderately imbalanced data than severely imbalanced data [54]. On the other hand, that study both downsampled the training and the test sets, which modified the test structure and differed from ours.

H2O Driverless AI is rarely used for cardiovascular disease. To the best of our knowledge, this is the first time this framework has been applied to the stroke study. Compared with other preventive studies on CVDs using automatic ML frameworks with AutoPrognosis (AUROC = 0.78), Auto-sklearn (AUROC =0.76), Auto-weka (AUROC = 0.75), and TPOT (AUROC = 0.74) [4,55], our H2O model produced a higher AUROC (0.804 for the union, 0.817 for the SHAP set and 0.832 for the external validation), so it is an effective classifier.

As the data were cross-sectional, the explanation of the variables suggesting those who had strokes would be more challenging. In our global explanation (explanations for the sampled population), the ten most influential features (the SHAP set) were broadly consistent with what was reported in the literature, while other features, some (such as sedentary activity, diastolic blood pressure (DBP), and weight) seemed to need more explorations to understand how effective they are in assessing prior strokes for individuals. We reviewed more studies for

Table 3
Baseline table for nomogram variables.

	NHANES 2015–2016			NHANES 2007–2008			NHANES 2017–2018			UKB baseline		
	Control	Stroke	p	Control	Stroke	p	Control	Stroke	p	Control	Stroke	p
	4246	135		4051	168		3935	191		372,184	467	
Age	47.3 (17.2)	64.8 (14.1)	<0.001	49.2 (17.4)	66.5 (13.4)	<0.001	50.1 (17.4)	65.0 (12.8)	<0.001	55.9 (8.09)	60.9 (6.52)	<0.001
Income	7.11 (3.25)	5.18 (2.55)	<0.001	6.56 (3.18)	5.34 (2.57)	<0.001	8.07 (3.44)	6.52 (3.27)	<0.001	6.43 (2.85)	4.71 (2.70)	<0.001
Race			<0.001			<0.001			<0.001			1
Others	2033 (47.9%)	41 (30.4%)		1289 (31.8%)	26 (15.5%)		1636 (41.6%)	41 (21.5%)		360,686 (96.9%)	453 (97.0%)	
Caucasian/Africa American	2213 (52.1%)	94 (69.6%)		2762 (68.2%)	142 (84.5%)		2299 (58.4%)	150 (78.5%)		11,498 (3.09%)	14 (3.00%)	
Diabetes mellitus:			<0.001			<0.001			<0.001			<0.001
N	3623 (85.3%)	82 (60.7%)		3526 (87.0%)	107 (63.7%)		3254 (82.7%)	109 (57.1%)		355,266 (95.5%)	364 (77.9%)	
Y	623 (14.7%)	53 (39.3%)		525 (13.0%)	61 (36.3%)		681 (17.3%)	82 (42.9%)		16,918 (4.55%)	103 (22.1%)	
Asthma:			<0.001			0.247			0.001			0.003
N	585 (13.8%)	35 (25.9%)		3536 (87.3%)	141 (83.9%)		3344 (85.0%)	145 (75.9%)		333,045 (89.5%)	438 (93.8%)	
Y	3661 (86.2%)	100 (74.1%)		515 (12.7%)	27 (16.1%)		591 (15.0%)	46 (24.1%)		39,139 (10.5%)	29 (6.21%)	
Hypertension:			<0.001			<0.001			<0.001			<0.001
N	1388 (32.7%)	94 (69.6%)		2675 (66.0%)	36 (21.4%)		2539 (64.5%)	45 (23.6%)		283,081 (76.1%)	198 (42.4%)	
Y	2858 (67.3%)	41 (30.4%)		1376 (34.0%)	132 (78.6%)		1396 (35.5%)	146 (76.4%)		89,103 (23.9%)	269 (57.6%)	
Tobacco:			0.168			0.498			0.125			0.02
N	883 (24.0%)	37 (29.8%)		2929 (72.3%)	126 (75.0%)		3042 (77.3%)	138 (72.3%)		334,303 (89.8%)	399 (85.4%)	
Y	2792 (76.0%)	87 (70.2%)		1122 (27.7%)	42 (25.0%)		893 (22.7%)	53 (27.7%)		37,881 (10.2%)	68 (14.6%)	
Snore:			0.7			0.305			0.023			0.186
<5/week	3130 (73.7%)	97 (71.9%)		2699 (66.6%)	105 (62.5%)		2766 (70.3%)	119 (62.3%)		232,992 (62.6%)	278 (59.5%)	
≥5/week	1116 (26.3%)	38 (28.1%)		1352 (33.4%)	63 (37.5%)		1169 (29.7%)	72 (37.7%)		139,192 (37.4%)	189 (40.5%)	

Table 4
Odds ratio for nomogram features calculated by LR and weighted LR.

	LR	Weighted LR
Age	1.05 ***	1.06 **
	CI [1.03,1.07]	CI [1.03,1.08]
Income	0.83 ***	0.81 **
	CI [0.77,0.90]	CI [0.73,0.89]
Caucasian or African American	2.68 ***	2.67 **
	CI [1.58,4.53]	CI [1.71,4.17]
Diabetes	2.30 ***	1.99 *
	CI [1.44,3.67]	CI [1.04,3.84]
Asthma	2.10 **	2.41 *
	CI [1.26,3.48]	CI [1.35,4.32]
HBP	1.86 *	2.10 *
	CI [1.11,3.13]	CI [1.04,4.25]
Tobacco	1.64 *	1.64
	CI [1.01,2.68]	CI [0.97,2.79]
Snore	1.18	1.61
	CI [0.72,1.93]	CI [0.81,3.18]
N	3068	3068
AIC	686.24	579.18
BIC	740.50	
Pseudo R^2	0.23	0.25

Standard errors are heteroskedasticity robust. *** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$. AIC, Akaike information criterion; BIC, Bayesian information criterions.

other possible explanations for them. Take DBP as an example. DBP levels in the low-normal range after a stroke (<70) could be related to an increased risk of major vascular events and ischemic stroke among patients who are recently suffering from non-cardioembolic strokes [55]. On the other hand, several SHAP algorithms could be selected, and Deep SHAP was selected for two reasons. For one thing, we did not explore feature interactions in feature engineering manually. Feature explanation could involve interaction information from Deep SHAP because the DNN model is characterized by feature crossing in hidden layers. In addition, when we calculated the SHAP values using other models based on other algorithms, such as support vector machine (SVM) with Kernel SHAP algorithm and Catboost with Tree SHAP [34,56], the performances of these models were lower than that of our DNN. Moreover, we found that some variables with less SHAP importance than our SHAP set features had interpretations changed in different models (especially DNN); for example, sedentary activity was a positively correlated feature in Catboost and BMI a negative feature in another trained DNN model while these two features had the opposite interpretations in our DNN model. This was possibly due to fluctuations of DNN performance for variables with small SHAP values. However, we found we could avoid possible changes in interpretations by focusing on the most important features, i.e., focusing on the top ten feature explanations. Consequently, in this study, we did not explore more explanations other than the top ten features.

Due to the cross-sectional nature of the NHANES database, the scope of work of our study cannot be used for risk assessment. It also restricted the comparison between our study and traditional risk scores/Cox regression used in longitudinal studies to assess the years the analyzed people were at risk. However, in some NHANES studies, follow-up outcomes of CVDs were provided, but few solely focused on stroke [10,52,53]. In a cross-sectional NHANES study aiming at predicting CVD risk, which was most similar to our data structure, we observed that ML, rather than the Cox model, could benefit from the nutrition data and had higher performances [10]. Therefore, pertaining to our data structure, ML might still have the potential to capture the complexity of nonlinear relationships. It is speculated that ML could outperform the Cox models

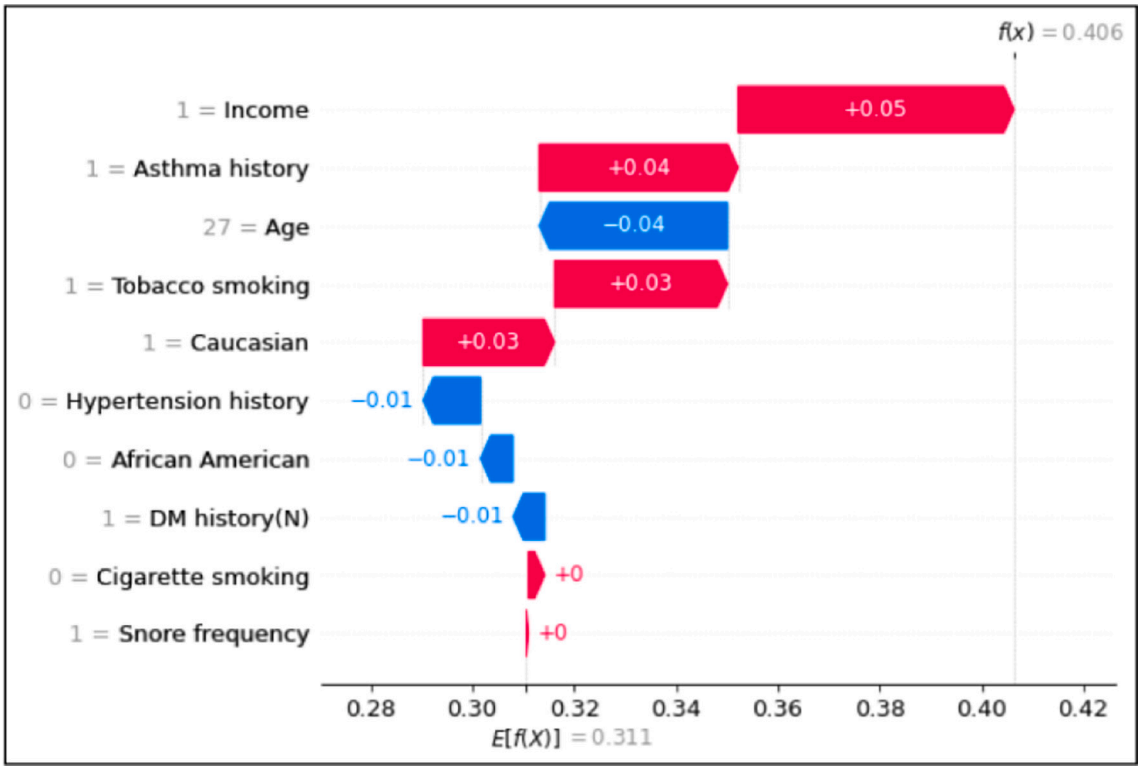

Fig. 4. Waterfall plot for individual explanation.

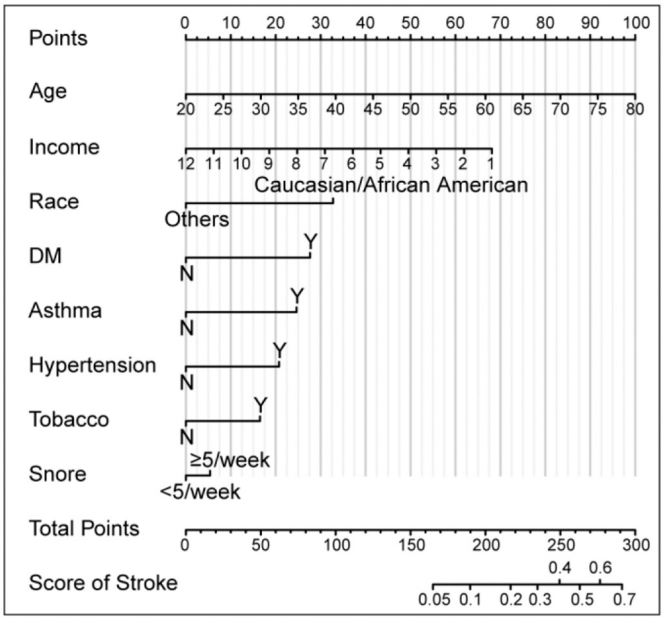

Fig. 5. Nomogram developed from SHAP set variables. For each feature, a point scale is provided to mark the points vertically in terms of the top scale of 'Points'. The points of all features are summed to calculate 'Total Points'. Finally, the 'Score of Stroke', at the bottom of the nomogram, is determined from 'Total Points'. Total Points <160 correspond to no scale of score. Hence, more than half the scale points could be classified as approximately no history of stroke.

or traditional risk scores if follow-up data are available for stroke outcomes.

There are limitations in our study that need to be addressed. First, this dataset is cross-sectional in nature, and therefore detected associations might be less robust than other study designs based on prospectively collected data. Although cross-sectional studies can provide information on the prevalence of a particular disease, which is helpful in planning public health interventions [6], our result cannot be used for inference of causality or treatment decision. However, as we did not infer the causality between the outcome and the exposures, there should not be temporal bias.

Second, our outcome was self-reported in the questionnaire, which might suffer from selection bias or reporting bias. For selection bias, NHANES utilized national representatives through complex sampling and manual quality control to mitigate the selection bias; we also involved the sampling weights in the modeling to adjust the effect size for the whole population and to evaluate the selection bias. For the reporting bias assessment, we adopted UKB hospital inpatient data to correct the labels.

Third, as a matter of concern, is the proportion of observation deletion and feature deletion. We excluded: 1) observations with non-valid ('Refused to answer' or 'Don't know') responses in covariates; 2) observations with unlabeled (missing values and non-valid responses in outcomes) data; 3) in NHANES 2015–16 for model development, features with over 30% missing values. We could infer the biases were small. First, the non-valid answers were out of the scope of our analysis. We want to explore features that are clear in definitions to the medical practitioners/caregivers; excluding non-valid answers might magnify the effect of valid answers. Additionally, the total proportions for these answers were small, ranging from 14%–18% of raw data for NHANES and 23% for UKB. So, the bias seemed to be small by deleting non-valid answers in covariates. Then, for the unlabeled data, the sensitivity analysis suggested low bias, with 0.02 AUROC reduction from over 200% sample size in NHANES 2015–16. Finally, for feature deletion, we deleted mainly the biomarkers with over 30% missing values (including selenium, cadmium, and LDL-cholesterol); they tended to be collected by subgroups and ended up with a small quantity. Still, we had large amounts of features in each dataset (30 to 122) in model development, and some of them could be the proxy variables for the deleted variables

(e.g., LDL cholesterol could be approximately derived from Total cholesterol and HDL cholesterol), thereby reducing the bias of lost information from the deleted features.

Fourth, related to feature deletion was that many important established blood biomarkers were not involved in our modeling. NHANES lacked some important stroke-associated variables, such as lipoprotein-associated phospholipase A2, D-dimer, and interleukin 6, etc., of which the involvement might further promote the performance. However, based on the results of blood biomarkers, we deduced that these lacking laboratory biomarkers might still be outperformed by the SHAP set in our modeling. Moreover, we prioritized using several features that are the informative and easily collected features to improve the simplicity and utility of our nomogram model and gain a higher generalization capability of the model.

Fifth, the NHANES data still lacks information on other features. For example, arthritis history had 5708 observations, while the type of arthritis had only 1433; thyroid history had 5806, while current thyroid status only had 593. Thus, the more informative features like arthritis type and current thyroid were with small sample size and therefore deleted because of too many missing values, although it also resulted in bias in our results.

Despite the limitations, our results, using a few questionnaire-based clinical variables, have high and robust performances and can potentially provide generalized and specialized characteristics for stroke survivors for a newer estimate of stroke prevalence.

6. Conclusion

This study compared diet, blood biomarkers, and other clinical features in assessing individuals who had a previous stroke. We selected the key contributors with comparable performances to the comprehensive set of features, provided individuals with interpretability of features using ML methods, and developed a nomogram to facilitate assessment. Our results could lead to a better understanding of the profile of stroke survivors and a newer estimate of stroke prevalence, which could shed light on reasonable estimates of stroke burden and public health intervention, international comparisons on stroke risk, recurrent stroke prevention, and clinical management strategies for stroke.

Authors' contribution

Conceptualization of project and analytic strategy instructor: KHKC. Data Extraction: JDL. Data Analysis and Interpretation: JDL. Drafting the article: JDL. Critical revision of the article: KHKC, JDL, ELC, KKL, PYMW, JL. Approval of the final version: KHKC, JDL, ELC, KKL, PYMW, JL.

Funding

This work is supported by the City University of Hong Kong New Research Initiatives/Infrastructure Support from Central (APRC; grant number 9610401), which had no involvement in the study design, data analysis, writing, or submission.

Ethics statement

Informed consent was offered by all participants from NHANES and UKB. The use of NHANES data was approved by the National Center for Health Statistics Research Ethics Review Board, while the use of UKB data was approved by the National Health Service (NHS) Research Ethics Service (11/N.W./0382). Access to the UKB data was granted under application number 45788.

Data availability statement

The NHANES data used in this work are publicly available at https://www.cdc.gov/nchs/nhanes/index.htm, while the UKB data are available on application at https://www.ukbiobank.ac.uk/.

Declaration of competing interest

All authors report no competing interests relevant to the manuscript.

Acknowledgments

We thank H2O.ai Academic Program for providing the license for the use of the automatic machine learning platform, H2O Driverless AI.

Appendix A. Data and methodologies

Data

Clinical and second clinical set variables from demographic, examination, and questionnaire data

The lists for the clinical set and the second clinical set included: 1) influential factors: high blood pressure, smoking, diabetes, physical inactivity, obesity, high blood cholesterol, heart diseases, sickle cell disease, age, race, gender, income, alcohol, drug abuse, sleep habits, oral health, gout, asthma, angina, thyroid, cancer, and hepatitis; 2) symptoms: face, limb weakness/numbness, speech slurred (confusion), trouble seeing, walking and severe headache; 3) complications: urinary tract infection and/or bladder control, pneumonia, swallowing problem, clinical depression, shoulder pain/anxiety, breathing problems, aspirin. Moreover, possible confounders, including diabetes risk, taking insulin, medication for depression, anxiety, and cholesterol, were also added. The clinical set consisted of the influential factors, while the second clinical set was a combination of the above variables.

Blood biomarkers from laboratory data

A list of blood biomarkers, including oxidative stress, metabolic and inflammatory ones, was compiled. It included: 1) metabolic: calcium, iron, cadmium, chloride, total cholesterol, triglyceride, percentage of segmented neutrophils, red cell distribution width, glycohemoglobin, potassium, sodium, high-density lipoprotein cholesterol (HDL—C), folic acid, and glucose; 2) inflammatory: white blood cell count, hematocrit and platelet count, aspartate and alanine aminotranferase, gamma glutamyl transferase, lactate dehydrogenase, creatinine, high-sensitivity C-reactive protein, Monocyte/HDL-C ratio, hemoglobin; 3) oxidative stress: Segmented neutrophils/Lymphocyte ratio, total bilirubin, uric acid; 4) neurohormone: cotinine.

Dietary nutrients from dietary data

The American Heart Association Diet and Lifestyle Recommendations encourages eating a variety of nutritious foods from all the food groups and eating less nutrient-poor foods to fight cardiovascular disease.[4] In our work, we involved all the nutrients that comprise foods or beverages in our dietary intake for stroke prediction; dietary supplement was also counted. These nutrients included: 1) nutrients that offered the most calories:

[4] https://www.stroke.org/en/about-stroke/stroke-risk-factors/stroke-risk-factors-you-can-control-treat-and-improve

carbohydrates, sugars, total fats, and HDL—C, protein, fiber, saturated and unsaturated (monounsaturated and polyunsaturated) fatty acids; 2) vitamins: vitamin A/B1/B2/B6/B12/C/D/E/K, alpha-carotene, beta-carotene, lycopene, lutein, riboflavin, niacin, folic acid, B-cryptoxanthin, theobromine, and folate; 3) minerals: sodium, phosphorus, zinc, intake of such as potassium, magnesium, iron, copper, and selenium; 4) other nutrients: water, alcohol, and caffeine.

Variable definitions and extraction

After the variables mentioned above were extracted from NHANES 2015–16 database in different files, they were merged according to the unique ID 'SEQN' to generate the four datasets. The occurrence of stroke was determined by the subject's answer to the questions 'Has a doctor or other health professional ever told {you/SP} that {you/s/he} . . . had a stroke?'. Out of 9575 individuals, 5714 answered with either 'Yes' (209) or 'No' (5505), 3856 had missing values, and five individuals refused to answer or wrote 'Don't know'. Based on this stroke proportion, other features were merged or removed if there were over 30% of values missing or replaced by similar and proxy variables if available. For example, smoking was defined based on the response to the question 'Smoked at least 100 cigarettes in life?' rather than 'Do you now smoke cigarettes?' because the latter produced over 30% missing values. Observations with 'Don't know' or 'Refused to answer' answers of stroke would also be deleted, but they would be counted in the semi-supervised model. Systolic blood pressure (SBP) and diastolic blood pressure (DBP) in the examination data were obtained from three consecutive readings after the subject had rested for five minutes in a seated position and after the determination of the maximum inflation level. The fourth read would be needed if a BP measurement is interrupted or incomplete. The values for SBP and DBP were then averaged by ourselves from the three readings. Urine biomarkers in the 'Laboratory Data' were excluded because of the small sample size and few variables after deleting variables with over 30% missing values, but the 'Dietary Data' dietary supplement was considered and added to the total nutrient intakes. Moreover, 'Added alpha-tocopherol (Vitamin E) (mg)' and 'Added vitamin B12 (mcg)' were also added to 'Vitamin E as alpha-tocopherol (mg)' and 'Vitamin B12 (mcg)'. The final values for diet were averaged by ourselves from the first and the second-day records.

Methodologies

Datawig imputation

DataWig imputation, which can be used for numerical, categorical, and unstructured text data [22],was adopted in this study. Inspired by established approaches [57], Datawig follows the process of multivariate imputation by chained equations (MICE) [58]. First, for strings and character sequence features, they are dealt with string and numeric representation and then further transformed into embeddings and hashing from (Long Short Term Memory) LSTM [59] or n-gram [60]; for numeric features, they are transformed into embeddings. Then all embeddings are concatenated and finally fitted with a regression or cross-entropy loss in terms of the type of missing value. DataWig compares favorably with other implementations (mean [61], k-nearest neighbor (KNN) [39], matrix factorization [62], MissForest [57], MICE [58]) for numeric and unstructured text imputation, even in the complex condition of missing-not-at-random [22].

Feature selection by BoostARoota

We used BoostARoota [23] to filter out redundant features and select important ones. BoostaARoota, as a modified version of Boruta algorithm [63], is a wrapper feature selection algorithm. Compared to Boruta, BoostARoota uses Xgboost [64] as the base model and modifies the feature elimination process, being computationally faster than Boruta. We first repeated BoostARoota for thirty times (by changing its random seed) to choose the overlaps as the robust features to be used for further analysis.

Imbalance classification analysis

In general, there are two strategies to handle class imbalance classification, i.e., data-level approach and algorithm-level approach. [65]

The data-level approach employs a preprocessing step to rebalance the class distribution. Samplings, as a preprocessing step, are very effective methods to address class imbalance [66] and have been proven to improve the predictive power of modeling in class-imbalanced datasets [67]. Our H2O model adopted a sampling method to adjust skewed stroke distribution to improve training. In addition to sampling methods, feature selection is another preprocessing step gaining popularity in class imbalance classification tasks. Feature selection removes irrelevant, redundant, or noisy data present in the problem of class overlapping in class imbalance [65,68]. We applied feature selection to reduce features to be used in our nomogram.

The algorithm-level approach, where the algorithms are fine-tuned to improve the learning of smaller classes, includes one-class learning and cost-sensitive learning (16). Our IF model is a one-class classification algorithm aiming at outlier or anomaly detection [24]. It can be effective for imbalanced classification datasets where stroke cases are both few in number and different in the feature space. Our DNN is a Cost-Sensitive Neural Network (25). It is trained with the Focal Loss function (29) to assign a larger error weight to stroke cases and reshape the standard cross-entropy loss to improve class-imbalance learning during the standard DNN training. Our LR adjusts observation weights inversely proportional to stroke frequencies in the training data to improve class imbalance training [27].

Thresholding is another cost-sensitive approach that is applied at the data level in a postprocessing step, aiming to identify the optimal decision thresholds for classification. [69] For binary classification, 0.5 is typically the threshold, while it may be biased with respect to the major class in imbalanced data [69,70], so using decision thresholds is an alternative technique that can deal with class imbalance [71]. The Youden index is a linear transformation of the mean sensitivity and specificity. It can define thresholds to avoid failure in evaluating the algorithm's ability and is applied in imbalance cases due to its invariance to imbalance ratios. [72–74] Therefore, we also used the Youden index to define thresholds for classification.

SHapley Additive exPlanations

SHapley Additive exPlanations (SHAP), as a tool for model interpretation, is based on a game-theoretic approach and can explain the output of any ML model [75]. Thus it could provide clinical value for interpreting the influential factors of stroke. SHAP was developed to solve the problem of inconsistency in many feature attribution methods. The so-called inconsistency means that the role of a certain feature in the model plays an important role, but the calculation methods of the importance of the feature, such as "Gain", "Split", and "Saabas", assign a lower importance value to it; SHAP guarantees this consistency in theory [76].

SHAP assigns each feature an importance value affecting a particular classification with that feature. For a given feature, $feature_i$, the Shapley value [77] is the weighted average of all possible marginal contributions of $feature_i$, which is used as the feature attribution. Eq. (1) presents a classic Shapley value estimation [77] for $feature_i$.

$$SHAP_{feature_i}(x) = \sum_{set:feature_i \in subset} \left[|subset| * \binom{F}{|subset|} \right]^{-1} \left[Prediction_{subset}(x) - Prediction_{subset \setminus feature_i}(x) \right] \quad (1)$$

where,

F: set of all features,

subset: subset of F,

|subset|: number of features in subset including $feature_i$,

$\left[|subset|*\binom{F}{|subset|}\right]^{-1}$: weight of marginal contribution of $feature_i$,

$Prediction_{subset}(x) - Prediction_{subset\setminus feature_i}(x)$: marginal contribution of $feature_i$ where $Prediction_{subset}(x)$ is the prediction from a model trained on subset and $Prediction_{subset\setminus feature_i}(x)$ the prediction from another model trained on subset with $feature_i$ withheld.

However, there are some issues and limitations as to the above formula. For example, the calculation is very time-consuming because of the numerous subsets, requiring much computation time to calculate the model's output under each subset accurately. Therefore, many algorithms have been developed and proposed, such as Kernel SHAP, Linear SHAP, Deep SHAP, and Tree SHAP [77,78].

Appendix B. Tables and figures

Table A1. All characteristics stratified by stroke for NHANES 2015–16

Table A1 can be found in the supplementary material.

Table A2
Performances of clinical+ set.

Model	AUROC	AUCPR	IDI	Categorical NRI	NPV	PPV
LR	0.799	0.148	−0.025(0.472)	−0.006(0.901)	0.99	0.087
CI	0.757–0.814	0.088–0.187	−0.092-0.042	−0.102-0.089	0.989–1	0.054–0.109
IF	0.772	0.101	−0.034(<0.001)	0.085(0.329)	0.986	0.089
CI	0.768–0.777	0.098–0.106	−0.055–0.013	−0.086-0.255	0.983–0.998	0.050–0.133
H2O	0.811	0.131	0.034(<0.001)	0.053(0.152)	0.992	0.078
CI	0.774–0.848	0.087–0.174	0.019–0.048	−0.019-0.125	0.986–0.999	0.058–0.144
DNN	0.804	0.154	0.026(<0.001)	0.208(0.001)	0.992	0.07
CI	0.798–0.808	0.135–0.199	0.015–0.037	0.081–0.335	0.986–0.999	0.057–0.123

Table A3
AUROC and AUCPR test statistics between SHAP set and clinical set.

	AUROC				AUCPR			
	LR	IF	H2O	DNN	LR	IF	H2O	DNN
LR	0.517				0.560			
IF		0.405				0.291		
H2O			0.067				0.857	
DNN				0.332				0.924

Table A4
AUROC and AUCPR test statistics between SHAP set and clinical+ set.

	AUROC				AUCPR			
	LR	IF	H2O	DNN	LR	IF	H2O	DNN
LR	0.777				0.204			
IF		0.059				0.063		
H2O			0.740				0.317	
DNN				0.722				0.140

Table A5
Performances of sensitivity analyses.

Model	AUROC	AUCPR	NPV	PPV
Nomogram 07–08	0.827	0.138	0.992	0.091
CI	0.801–0.852	0.115–0.168	0.989–0.995	0.085–0.097
Nomogram 17–18	0.822	0.153	0.988	0.116
CI	0.797–0.846	0.130–0.188	0.985–0.992	0.108–0.124
Nomogram test	0.800	0.089	0.99	0.075
CI	0.739–0.855	0.065–0.136	0.984–0.995	0.062–0.088
Weight 07–08	0.825	0.128	0.993	0.09
CI	0.799–0.851	0.109–0.155	0.99–0.996	0.085–0.095
Weight 17–18	0.81	0.145	0.986	0.109
CI	0.783–0.835	0.124–0.177	0.982–0.99	0.101–0.117
Weight test	0.824	0.1	0.994	0.073

(*continued on next page*)

Table A5 (*continued*)

Model	AUROC	AUCPR	NPV	PPV
CI	0.764–0.875	0.074–0.145	0.989–0.999	0.064–0.082
Semi-supervised test	0.813	0.104	0.993	0.069
CI	0.749–0.87	0.074–0.160	0.987–0.998	0.059–0.078
HGB test	0.793	0.096	0.992	0.066
CI	0.727–0.854	0.068–0.147	0.987–0.997	0.057–0.074
HGB 15–16	0.808	0.107	0.993	0.074
CI	0.779–0.836	0.089–0.132	0.99–0.996	0.069–0.079
UKB	0.753	0.004	0.999	0.002
CI	0.731–0.774	0.004–0.005	0.999–1.000	0.002–0.002

Table A6
Characteristics of predictors from NHANS 2015–16 based on Framingham Stroke Risk Score variables and raw SHAP set.

FSRS risk factors	Level	Control	Case	P
N		4245	134	
Gender (%)	F	2175 (51.2)	59 (44.0)	0.12
	M	2070 (48.8)	75 (56.0)	
Age (SD)		47.32 (17.25)	64.82 (14.13)	<0.001
Systolic pressure (SD)		124.23 (17.61)	134.93 (23.30)	<0.001
Diabetes (%)	Y	3623 (85.3)	81 (60.4)	<0.001
	N	622 (14.7)	53 (39.6)	
Cigarette use (%)	Y	1713 (40.4)	80 (59.7)	<0.001
	N	2532 (59.6)	54 (40.3)	
Hypertension medicine (%)	Y	168 (4.0)	7 (5.2)	<0.001
	N	1219 (28.7)	86 (64.2)	
CVD (%)	Y	278 (6.5)	53 (39.6)	<0.001
	N	3967 (93.5)	81 (60.4)	
Raw SHAP set	Level	Control	Case	P
N		4245	134	
Age (SD)		47.32 (17.25)	64.82 (14.13)	<0.001
Race (%)	Mexican American	724 (17.1)	15 (11.2)	0.002
	Other Hispanic	611 (14.4)	13 (9.7)	
	Non-Hispanic White	1398 (32.9)	61 (45.5)	
	Non-Hispanic Black	815 (19.2)	33 (24.6)	
	Other Race	697 (16.4)	12 (9.0)	
Hypertension (%)	Y	1387 (32.7)	93 (69.4)	<0.001
	N	2858 (67.3)	41 (30.6)	
Diabetes (%)	Y	534 (12.6)	48 (35.8)	<0.001
	N	3623 (85.3)	81 (60.4)	
	Borderline	88 (2.1)	5 (3.7)	
Cigarette use (%)	Y	1713 (40.4)	80 (59.7)	<0.001
	N	2532 (59.6)	54 (40.3)	
Snore (%)	Never	1202 (28.3)	48 (35.8)	0.019
	Rarely	1131 (26.6)	20 (14.9)	
	Occasionally	797 (18.8)	29 (21.6)	
	Frequently	1115 (26.3)	37 (27.6)	
Income (SD)		7.11 (3.25)	5.18 (2.56)	<0.001
Tobacco use (%)	Y	883 (20.8)	37 (27.6)	0.061
	N	2791 (65.7)	86 (64.2)	
Asthma (%)	Y	585 (13.8)	35 (26.1)	<0.001
	N	3660 (86.2)	99 (73.9)	

CVD, cardiovascular disease, is defined by 1) “Ever told had congestive heart failure”; 2) “Ever told you had coronary heart disease”; 3) “Ever told you had angina/angina pectoris”; 4)“Ever told you had heart attack”.

Table A7. Weighted raw characteristics of predictors from NHANS based on SHAP set variables.

Table A7 can be found in the supplementary material.

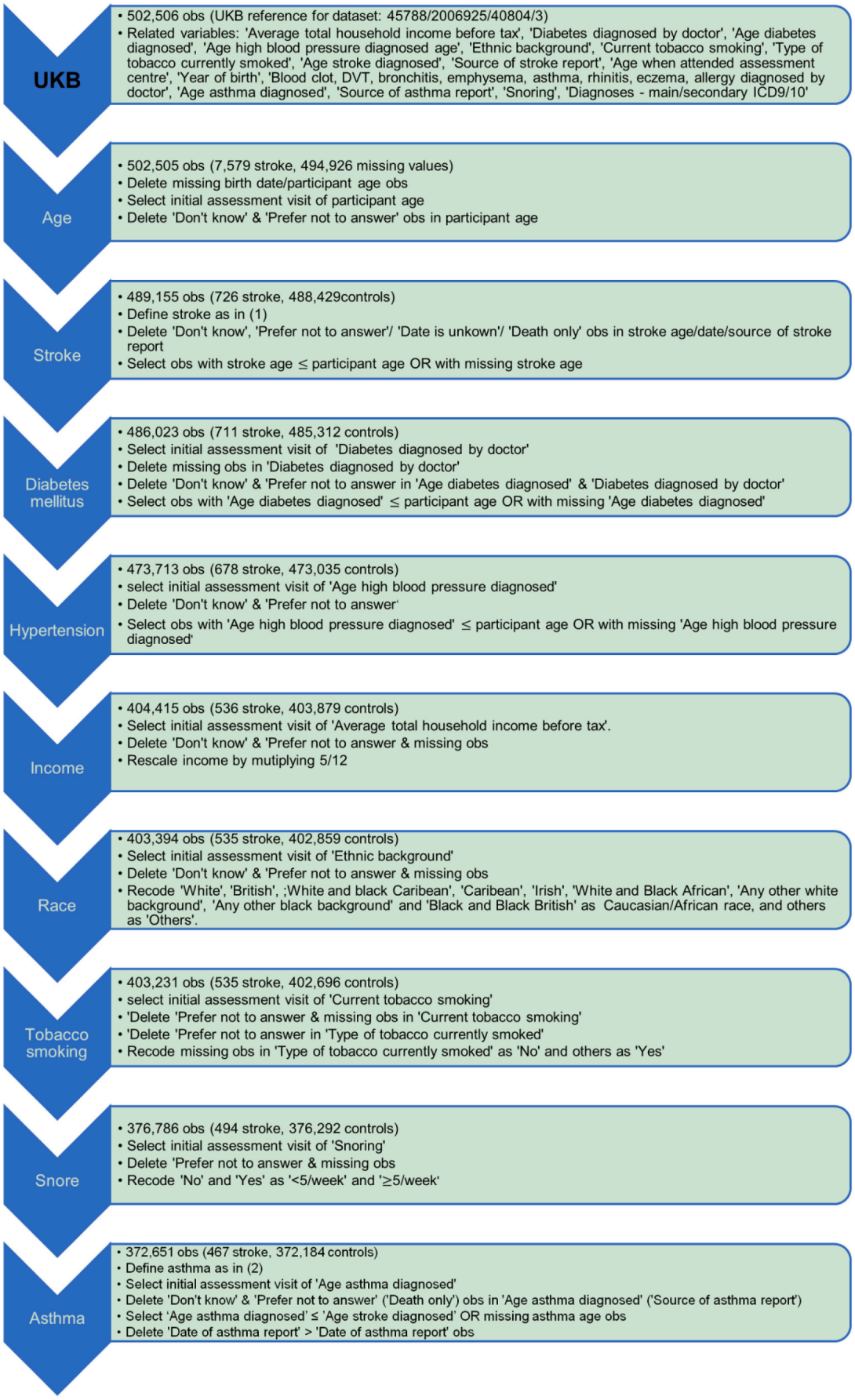
UKB
• 502,506 obs (UKB reference for dataset: 45788/2006925/40804/3)
• Related variables: 'Average total household income before tax', 'Diabetes diagnosed by doctor', 'Age diabetes diagnosed', 'Age high blood pressure diagnosed age', 'Ethnic background', 'Current tobacco smoking', 'Type of tobacco currently smoked', 'Age stroke diagnosed', 'Source of stroke report', 'Age when attended assessment centre', 'Year of birth', 'Blood clot, DVT, bronchitis, emphysema, asthma, rhinitis, eczema, allergy diagnosed by doctor', 'Age asthma diagnosed', 'Source of asthma report', 'Snoring', 'Diagnoses - main/secondary ICD9/10'
Age
• 502,505 obs (7,579 stroke, 494,926 missing values)
• Delete missing birth date/participant age obs
• Select initial assessment visit of participant age
• Delete 'Don't know' & 'Prefer not to answer' obs in participant age
Stroke
• 489,155 obs (726 stroke, 488,429controls)
• Define stroke as in (1)
• Delete 'Don't know', 'Prefer not to answer'/ 'Date is unkown'/ 'Death only' obs in stroke age/date/source of stroke report
• Select obs with stroke age ≤ participant age OR with missing stroke age
Diabetes mellitus
• 486,023 obs (711 stroke, 485,312 controls)
• Select initial assessment visit of 'Diabetes diagnosed by doctor'
• Delete missing obs in 'Diabetes diagnosed by doctor'
• Delete 'Don't know' & 'Prefer not to answer in 'Age diabetes diagnosed' & 'Diabetes diagnosed by doctor'
• Select obs with 'Age diabetes diagnosed' ≤ participant age OR with missing 'Age diabetes diagnosed'
Hypertension
• 473,713 obs (678 stroke, 473,035 controls)
• select initial assessment visit of 'Age high blood pressure diagnosed'
• Delete 'Don't know' & 'Prefer not to answer'
• Select obs with 'Age high blood pressure diagnosed' ≤ participant age OR with missing 'Age high blood pressure diagnosed'
Income
• 404,415 obs (536 stroke, 403,879 controls)
• Select initial assessment visit of 'Average total household income before tax'.
• Delete 'Don't know' & 'Prefer not to answer & missing obs
• Rescale income by mutiplying 5/12
Race
• 403,394 obs (535 stroke, 402,859 controls)
• Select initial assessment visit of 'Ethnic background'
• Delete 'Don't know' & 'Prefer not to answer & missing obs
• Recode 'White', 'British', ;White and black Caribean', 'Caribean', 'Irish', 'White and Black African', 'Any other white background', 'Any other black background' and 'Black and Black British' as Caucasian/African race, and others as 'Others'.
Tobacco smoking
• 403,231 obs (535 stroke, 402,696 controls)
• select initial assessment visit of 'Current tobacco smoking'
• 'Delete 'Prefer not to answer & missing obs in 'Current tobacco smoking'
• 'Delete 'Prefer not to answer in 'Type of tobacco currently smoked'
• Recode missing obs in 'Type of tobacco currently smoked' as 'No' and others as 'Yes'
Snore
• 376,786 obs (494 stroke, 376,292 controls)
• Select initial assessment visit of 'Snoring'
• Delete 'Prefer not to answer & missing obs
• Recode 'No' and 'Yes' as '<5/week' and '≥5/week'
Asthma
• 372,651 obs (467 stroke, 372,184 controls)
• Define asthma as in (2)
• Select initial assessment visit of 'Age asthma diagnosed'
• Delete 'Don't know' & 'Prefer not to answer' ('Death only') obs in 'Age asthma diagnosed' ('Source of asthma report')
• Select 'Age asthma diagnosed' ≤ 'Age stroke diagnosed' OR missing asthma age obs
• Delete 'Date of asthma report' > 'Date of asthma report' obs

Fig. A1. Flow chart of processing of UK biobank data.

Fig. 1 References.

Malik R, Rannikmäe K, Traylor M, Georgakis MK, Sargurupremraj M, Markus HS, Hopewell JC, Debette S, Sudlow CLM, Dichgans M. Genome-wide meta-analysis identifies 3 novel loci associated with stroke. Ann Neurol (2018) 84:934–939. doi:https://doi.org/10.1002/ana.25369

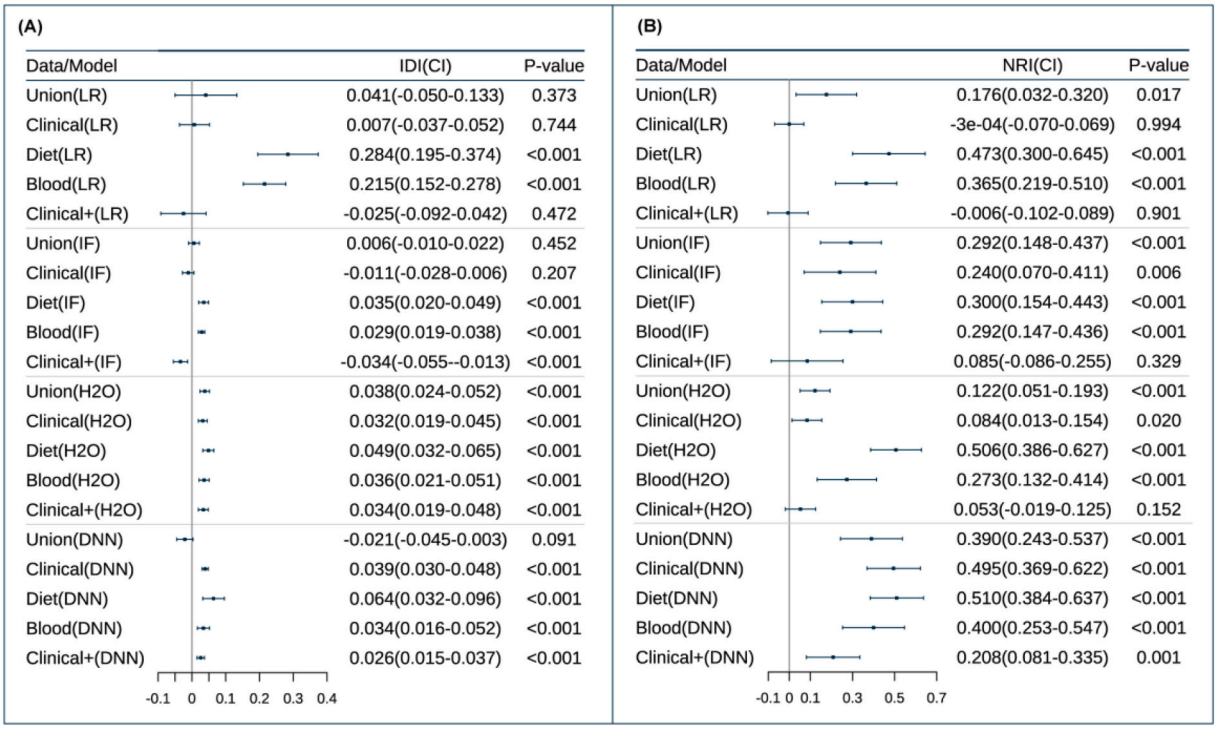

Fig. A2. IDI and NRI between SHAP and other datasets. **(A)** IDI; **(B)** NRI.

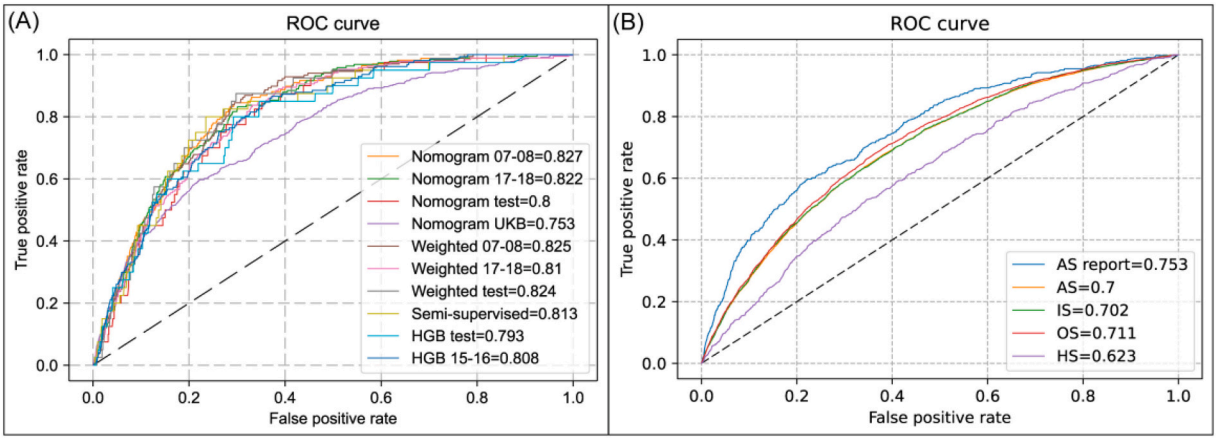

Fig. A3. SHAP set performance concerning Nomogram, sampling weightings, semi-supervised models and stroke subtypes. **(A)** SHAP set on Nomogram, weighted, semi-supervised algorithms. **(B)** Nomogram on stroke subtypes from UKB. HGB: Histogram-based Gradient Boosting; Semi-supervised semi-supervised model based on HGB; AS: all stroke; IS: ischemic stroke; OS: other stroke; SAH: subarachnoid hemorrhage; ICH: intracerebral hemorrhage; HS: hemorrhagic stroke.

Appendix C. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jns.2022.120335.

References

[1] S. Rubattu, R. Giliberti, M. Volpe, Etiology and pathophysiology of stroke as a complex trait, Am. J. Hypertens. 13 (2000) 1139–1148.
[2] P.A. Wolf, Stroke risk profiles, Stroke. 40 (2009) 2008–2011, https://doi.org/10.1161/STROKEAHA.108.530725.
[3] A. Orfanoudaki, A.M. Nouh, E. Chesley, C. Cadisch, B. Stein, M. Alberts, D. Bertsimas, Novel machine learning proves stroke risk is not linear, Stroke. 51 (2020) A153.
[4] A.M. Alaa, T. Bolton, E. Di Angelantonio, J.H.F.F. Rudd, M. van der Schaar, Cardiovascular disease risk prediction using automated machine learning: a prospective study of 423,604 UK Biobank participants, PLoS One 14 (2019) 1–17, https://doi.org/10.1371/journal.pone.0213653.
[5] A.M. Alaa, M. van der Schaar, AutoPrognosis: automated clinical prognostic modeling via Bayesian optimization with structured kernel learning, in: ArXiv Prepr. ArXiv1802.07207, 2018.
[6] L.P. Bignold, Principles of tumors: a translational approach to foundations, Princ. Tumors A Transl Approach Found. (2019) 1–675.

[7] J. Semerdjian, S. Frank, An ensemble classifier for predicting the onset of type II diabetes, in: ArXiv Prepr. ArXiv1708.07480, 2017.
[8] W. Yu, T. Liu, R. Valdez, M. Gwinn, M.J. Khoury, Application of support vector machine modeling for prediction of common diseases: the case of diabetes and pre-diabetes, BMC Med. Inform. Decis. Mak. 10 (2010) 1–7.
[9] A. Dinh, S. Miertschin, A. Young, S.D. Mohanty, A data-driven approach to predicting diabetes and cardiovascular disease with machine learning, BMC Med. Inform. Decis. Mak. 19 (2019) 211.
[10] J. Rigdon, S. Basu, Machine learning with sparse nutrition data to improve cardiovascular mortality risk prediction in the USA using nationally randomly sampled data, BMJ Open 9 (2019) 1–9, https://doi.org/10.1136/bmjopen-2019-032703.
[11] X. Mai, X. Liang, Risk factors for stroke based on the National Health and nutrition examination survey, J. Nutr. Health Aging 24 (2020) 791–795, https://doi.org/10.1007/s12603-020-1430-4.
[12] A.P. Abreo, S.R. Bailey, K. Abreo, Associations between calf, thigh, and arm circumference and cardiovascular and all-cause mortality in NHANES 1999-2004, Nutr. Metab. Cardiovasc. Dis. 31 (5) (2021) 1410–1415.
[13] N. Vangeepuram, B. Liu, P.H. Chiu, L. Wang, G. Pandey, Estimating youth diabetes risk using NHANES data and machine learning, MedRxiv. (2020) 19007872, https://doi.org/10.1101/19007872.
[14] U.K. Biobank, About UK Biobank, Available at, Https://Www.Ukbiobank.Ac.Uk/about-Biobank-Uk, 2014.
[15] N. Parakh, H.L. Gupta, A. Jain, Evaluation of enzymes in serum and cerebrospinal fluid in cases of stroke, Neurol. India 50 (2002) 518.
[16] R. Yang, A. Wang, L. Ma, Z. Su, S. Chen, Y. Wang, S. Wu, C. Wang, Hematocrit and the incidence of stroke: a prospective, population-based cohort study, Ther. Clin. Risk Manag. 14 (2018) 2081–2088, https://doi.org/10.2147/TCRM.S174961.
[17] H.Y. Wang, W.R. Shi, X. Yi, Y.P. Zhou, Z.Q. Wang, Y.X. Sun, Assessing the performance of monocyte to high-density lipoprotein ratio for predicting ischemic stroke: insights from a population-based Chinese cohort, Lipids Health Dis. 18 (2019) 1–11, https://doi.org/10.1186/s12944-019-1076-6.
[18] T.S. Perlstein, R.L. Pande, M.A. Creager, J. Weuve, J.A. Beckman, Serum total bilirubin level, prevalent stroke, and stroke outcomes: NHANES 1999-2004, Am. J. Med. 121 (2008) 781–788, https://doi.org/10.1016/j.amjmed.2008.03.045.
[19] M. Söderholm, Y. Borné, B. Hedblad, M. Persson, G. Engström, Red cell distribution width in relation to incidence of stroke and carotid atherosclerosis: a population-based cohort study, PLoS One 10 (2015) 1–14, https://doi.org/10.1371/journal.pone.0124957.
[20] H.G. Oh, E.J. Rhee, T.W. Kim, K.B. Lee, J.H. Park, K.I. Yang, D. Jeong, H.K. Park, Higher glycated hemoglobin level is associated with increased risk for ischemic stroke in non-diabetic Korean male adults, Diabetes Metab. J. 35 (2011) 551–557, https://doi.org/10.4093/dmj.2011.35.5.551.
[21] M. Emdin, C. Passino, L. Donato, A. Paolicchi, A. Pompella, Serum gamma-glutamyltransferase as a risk factor of ischemic stroke might be independent of alcohol consumption, Stroke. 33 (2002) 1163–1164, https://doi.org/10.1161/01.STR.0000012344.35312.13.
[22] F. Bießmann, T. Rukat, P. Schmidt, P. Naidu, S. Schelter, A. Taptunov, D. Lange, D. Salinas, DataWig: missing value imputation for tables, J. Mach. Learn. Res. 20 (2019) 1–6.
[23] C. DeHan, BoostARoota, 2017.
[24] W.-R. Chen, Y.-H. Yun, M. Wen, H.-M. Lu, Z.-M. Zhang, Y.-Z. Liang, Representative subset selection and outlier detection via isolation forest, Anal. Methods 8 (2016) 7225–7231.
[25] P. Hall, N. Gill, M. Kurka, W. Phan, A. Bartz, Machine Learning Interpretability with H2O Driverless AI: First Edition Machine Learning Interpretability with H2O Driverless AI, H_2O. Ai, 2019, pp. 1–40. http://docs.h2o.ai.
[26] Z.-H. Zhou, X.-Y. Liu, Training cost-sensitive neural networks with methods addressing the class imbalance problem, IEEE Trans. Knowl. Data Eng. 18 (2006) 63–77, https://doi.org/10.1109/TKDE.2006.17.
[27] W.S. Lee, B. Liu, Learning with positive and unlabeled examples using weighted logistic regression, in: ICML, 2003, pp. 448–455.
[28] T. Akiba, S. Sano, T. Yanase, T. Ohta, M. Koyama, Optuna: a next-generation hyperparameter optimization framework, in: Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min, 2019, pp. 2623–2631, https://doi.org/10.1145/3292500.3330701.
[29] A. Verma, Evaluation of classification algorithms with solutions to class imbalance problem on bank marketing dataset using WEKA, IJACSA, Int. J. Adv. Comput. Sci. Appl. 5 (13) (2019) 54–60.
[30] Q. Zou, S. Xie, Z. Lin, M. Wu, Y. Ju, Finding the best classification threshold in imbalanced classification, Big Data Res. 5 (2016) 2–8, https://doi.org/10.1016/j.bdr.2015.12.001.
[31] X. Robin, N. Turck, A. Hainard, N. Tiberti, F. Lisacek, J.-C. Sanchez, M. Müller, pROC: an open-source package for R and S+ to analyze and compare ROC curves, BMC Bioinformat. 12 (2011) 77.
[32] S. Kundu, Y.S. Aulchenko, A.C.J.W. Janssens, M.S. Kundu, S. GenABEL, Package 'PredictABEL', 2020.
[33] A.J. Canty, Resampling methods in R: the boot package, Newsl. R Proj. 2 (2002) 3.
[34] S.M. Lundberg, S.I. Lee, A unified approach to interpreting model predictions, Adv. Neural Inf, Process. Syst. (2017) 4766–4775. Decem (2017).
[35] Z. Zhang, M.W. Kattan, Drawing nomograms with R: applications to categorical outcome and survival data, Ann. Transl. Med. 5 (2017), https://doi.org/10.21037/atm.2017.04.01.
[36] R.B. D'Agostino, P.A. Wolf, A.J. Belanger, W.B. Kannel, Stroke risk profile: adjustment for antihypertensive medication the Framingham study, Stroke. 25 (1994) 40–43, https://doi.org/10.1161/01.STR.25.1.40.
[37] M. Kuhn, Caret: classification and regression training, Astrophys. Source Code Libr. (2015) ascl: 1505.003.
[38] Y.J. Ong, Y. Zhou, N. Baracaldo, H. Ludwig, Adaptive histogram-based gradient boosted trees for federated learning, in: ArXiv Prepr. ArXiv2012.06670, 2020.
[39] S. Zhang, Nearest neighbor selection for iteratively kNN imputation, J. Syst. Softw. 85 (2012) 2541–2552.
[40] R. Malik, K. Rannikmäe, M. Traylor, M.K. Georgakis, M. Sargurupremraj, H. S. Markus, J.C. Hopewell, S. Debette, C.L.M. Sudlow, M. Dichgans, Genome-wide meta-analysis identifies 3 novel loci associated with stroke, Ann. Neurol. 84 (2018) 934–939, https://doi.org/10.1002/ana.25369.
[41] M. Abdullah Said, R.N. Eppinga, E. Lipsic, N. Verweij, P. Vvan der Harst, Relationship of arterial stiffness index and pulse pressure with cardiovascular disease and mortality, J. Am. Heart Assoc. 7 (2018), https://doi.org/10.1161/JAHA.117.007621.
[42] C. Schnier, K. Bush, J. Nolan, C. Sudlow, Definitions of asthma for UK Biobank phase 1 outcomes adjudication documentation prepared by: definitions of asthma, in: UK Biobank Phase 1 Outcomes Adjudication, 2017.
[43] S.M. Lundberg, B. Nair, M.S. Vavilala, M. Horibe, M.J. Eisses, T. Adams, D. E. Liston, D.K.W. Low, S.F. Newman, J. Kim, S.I. Lee, Explainable machine-learning predictions for the prevention of hypoxaemia during surgery, Nat. Biomed. Eng. 2 (2018) 749–760, https://doi.org/10.1038/s41551-018-0304-0.
[44] T.W. Gillespie, Understanding waterfall plots, J. Adv. Pract. Oncol. 3 (2012) 106.
[45] A. Jalali, A. Alvarez-Iglesias, D. Roshan, J. Newell, Visualising statistical models using dynamic nomograms, PLoS One 14 (2019), e0225253.
[46] J.N. Mandrekar, Receiver operating characteristic curve in diagnostic test assessment, J. Thorac. Oncol. 5 (2010) 1315–1316.
[47] J.W. Lo, J.D. Crawford, K. Samaras, D.W. Desmond, S. Köhler, J. Staals, F.R. J. Verhey, H.J. Bae, K.J. Lee, B.J. Kim, R. Bordet, C. Cordonnier, T. Dondaine, A. M. Mendyk, B.C. Lee, K.H. Yu, J.S. Lim, N. Kandiah, R.J. Chander, C. Yatawara, D. M. Lipnicki, P.S. Sachdev, Association of prediabetes and Type 2 diabetes with cognitive function after stroke: a STROKOG collaboration study, Stroke. (2020) 1640–1646, https://doi.org/10.1161/STROKEAHA.119.028428.
[48] A.J. Bowers, X. Zhou, Receiver operating characteristic (ROC) area under the curve (AUC): a diagnostic measure for evaluating the accuracy of predictors of education outcomes, J. Educ. Stud. Placed Risk 24 (2019) 20–46.
[49] A. Corlateanu, I. Stratan, S. Covantev, V. Botnaru, O. Corlateanu, N. Siafakas, Asthma and stroke: a narrative review, Asthma Res. Pract. 7 (2021) 1–17, https://doi.org/10.1186/s40733-021-00069-x.
[50] J. Li, R.D. McEvoy, D. Zheng, K.A. Loffler, X. Wang, S. Redline, R.J. Woodman, C. S. Anderson, Self-reported snoring patterns predict stroke events in high-risk patients with obstructive sleep apnea: post-hoc analyses of the SAVE study, Chest. 158 (5) (2020) 2146–2154.
[51] R. Micha, J.L.L. Peñalvo, F. Cudhea, F. Imamura, C.D.D. Rehm, D. Mozaffarian, Association between dietary factors and mortality from heart disease, stroke, and type 2 diabetes in the United States, JAMA - J. Am. Med. Assoc. 317 (2017) 912–924, https://doi.org/10.1001/jama.2017.0947.
[52] A. Pandya, M.C. Weinstein, T.A. Gaziano, A comparative assessment of non-laboratory-based versus commonly used laboratory-based cardiovascular disease risk scores in the NHANES III population, PLoS One 6 (2011), https://doi.org/10.1371/journal.pone.0020416.
[53] T.A. Gaziano, C.R. Young, G. Fitzmaurice, S. Atwood, J.M. Gaziano, Laboratory-based versus non-laboratory-based method for assessment of cardiovascular disease risk: the NHANES I follow-up study cohort, Lancet. 371 (2008) 923–931, https://doi.org/10.1016/S0140-6736(08)60418-3.
[54] D. Veganzones, E. Séverin, An investigation of bankruptcy prediction in imbalanced datasets, Decis. Support. Syst. 112 (2018) 111–124, https://doi.org/10.1016/j.dss.2018.06.011.
[55] J.-H. Park, B. Ovbiagele, Post-stroke diastolic blood pressure and risk of recurrent vascular events, Eur. J. Neurol. 24 (2017) 1416–1423.
[56] G. Huang, L. Wu, X. Ma, W. Zhang, J. Fan, X. Yu, W. Zeng, H. Zhou, Evaluation of CatBoost method for prediction of reference evapotranspiration in humid regions, J. Hydrol. 574 (2019) 1029–1041.
[57] S. Van Buuren, Flexible Imputation of Missing Data, CRC press, 2018.
[58] S. van Buuren, K. Groothuis-Oudshoorn, Mice: multivariate imputation by chained equations in R, J. Stat. Softw. 45 (2011) 1–67, https://doi.org/10.18637/jss.v045.i03.
[59] M. Sundermeyer, R. Schlüter, H. Ney, LSTM neural networks for language modeling, in: Thirteen. Annu. Conf. Int. Speech Commun. Assoc, 2012.
[60] W. Cheng, C. Greaves, M. Warren, From n-gram to skipgram to concgram, Int. J. Corpus Linguist. 11 (2006) 411–433.
[61] W. Young, G. Weckman, W. Holland, A survey of methodologies for the treatment of missing values within datasets: limitations and benefits, Theor. Issues Ergon. Sci. 12 (2011) 15–43.
[62] Y. Koren, R. Bell, C. Volinsky, Matrix factorization techniques for recommender systems, Computer (Long Beach. Calif). 42 (2009) 30–37.
[63] M.B. Kursa, W.R. Rudnicki, Feature selection with the Boruta package, J. Stat. Softw. 36 (2015) 1–13, https://doi.org/10.18637/jss.v036.i11.
[64] T. Chen, C. Guestrin, XGBoost: a scalable tree boosting system, in: Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min. 13-17-Augu, 2016, pp. 785–794, https://doi.org/10.1145/2939672.2939785.
[65] A. Ali, S.M. Shamsuddin, A.L. Ralescu, Classification with class imbalance problem: a review, Int. J. Adv. Soft Comput. Its Appl. 7 (2015) 176–204.
[66] N. Japkowicz, The class imbalance problem: Significance and strategies, in: Proc. Int'l Conf. Artif. Intell, 2000.

[67] P.H. Lee, Resampling methods improve the predictive power of modeling in class-imbalanced datasets, Int. J. Environ. Res. Public Health 11 (2014) 9776–9789, https://doi.org/10.3390/ijerph110909776.
[68] G. Cuaya, A. Munoz-Meléndez, E.F. Morales, A minority class feature selection method, in: Iberoam. Congr. Pattern Recognit, 2011, pp. 417–424.
[69] C. Esposito, G.A. Landrum, N. Schneider, N. Stiefl, S. Riniker, GHOST: adjusting the decision threshold to handle imbalanced data in machine learning, J. Chem. Inf. Model. 61 (2021) 2623–2640, https://doi.org/10.1021/acs.jcim.1c00160.
[70] X. Zhang, H. Gweon, S. Provost, Threshold moving approaches for addressing the class imbalance problem and their application to multi-label classification, in: ACM Int. Conf. Proceeding Ser. PartF16925, 2020, pp. 72–77, https://doi.org/10.1145/3441250.3441274.
[71] G. Collell, D. Prelec, K. Patil, Reviving Threshold-Moving: a Simple Plug-in Bagging Ensemble for Binary and Multiclass Imbalanced Data. http://arxiv.org/abs/1606.08698, 2016.
[72] V.V. Starovoitov, Y.I. Golub, Comparative study of quality estimation of binary classification, Informatics. 17 (2020) 87–101, https://doi.org/10.37661/1816-0301-2020-17-1-87-101.
[73] F.A.G. Pena, P.D.M. Fernandez, P.T. Tarr, T.I. Ren, E.M. Meyerowitz, A. Cunha, J regularization improves imbalanced multiclass segmentation, in: 2020 IEEE 17th Int. Symp. Biomed. Imaging, 2020, pp. 1–5.
[74] M. Usman, S. Khan, J.A. Lee, AFP-LSE: antifreeze proteins prediction using latent space encoding of composition of k-spaced amino acid pairs, Sci. Rep. 10 (2020) 1–13, https://doi.org/10.1038/s41598-020-63259-2.
[75] K. Zhang, Y. Zhang, M. Wang, A unified approach to interpreting model predictions Scott, in: Nips, 2012, pp. 426–430.
[76] S.M. Lundberg, G.G. Erion, S.-I. Lee, Consistent Individualized Feature Attribution for Tree Ensembles. http://arxiv.org/abs/1802.03888, 2018.
[77] S.M. Lundberg, S.I. Lee, A unified approach to interpreting model predictions, in: Adv. Neural Inf. Process. Syst, 2017, pp. 4766–4775. https://github.com/slundberg/shap (accessed March 23, 2020).
[78] S.M. Lundberg, G. Erion, H. Chen, A. DeGrave, J.M. Prutkin, B. Nair, R. Katz, J. Himmelfarb, N. Bansal, S.-I. Lee, From local explanations to global understanding with explainable AI for trees, Nat. Mach. Intell. 2 (2020) 56–67, https://doi.org/10.1038/s42256-019-0138-9.