Layman’s summary

Strokes resulting from the rupture of an aneurysm (bulge in a brain blood vessel) are a quite rare occurrence. Survival rates of these type of strokes are low compared to other types of strokes. Correct prediction of the rupture risk of an aneurysm could be a helpful way to help in the treatment of patients with an aneurysm. This prediction is done using risk factors for rupture of an aneurysm. Current known risk factors include female sex, smoking, excessive alcohol intake, high blood pressure and a family history of these type of strokes. These predictions using the currently known risk factors often wrongfully calculate the risk of rupture. In this study we tried to identify new risk factors for this type of stroke using an database from the UK containing both medical information and life style information of participants. Participants that got diagnosed with a stroke after they were assessed for the this database were selected and compared to the other participants that did not get a stroke. By comparison between these two groups, relevant information was selected from the database. This information was used to train machine learning algorithms to select the most relevant information. In the end the results from these machine learning algorithms was compared resulting in the proposal of three potential new risk factors: low sodium concentration in urine, low creatinine concentration in urine and low average 24 hours level of noise pollution. Low levels of sodium and creatinine are associated with several diseases related to the circulatory system and kidney’s. Currently there is no literature available that links noise pollution to strokes. One possibility is that noise pollution is related to something else, which in turn is a risk factor for strokes. During this study it became apparent that the low amount of strokes that occurred in comparison to the amount if participants in the database was very low. This imbalance in data causes some hardship for machine learning algorithms or their outcome metrics, and without accounting for them biased conclusions might be made.

Ways to deal with imbalanced data sets for machine-learning using the identification of potential new risk factors for aneurysmal subarachnoid hemorrhage from the UK Biobank as an example.

L. Edwards, supervisor: dr. Y.M. Ruigrok, daily supervisor: MSc. J. Kanning

Abstract

Imbalanced data which is the occurrence of one a minority class in a data set, often causes hardship for machine learning algorithms. A pipeline was built to preprocess the data and apply machine learning algorithms specifically built for imbalanced data sets. Different resulting metrics for model performance were considered (AUROC, AUPCR, precision, recall, accuracy, F-1 and F-beta). The pipeline was applied to the UK Biobank, a large-scale prospective cohort study that allowed to identify hypothesis free, new risk factors for aneurysmal subarachnoid hemorrhage (aSAH).

Introduction

Rupture of an intracranial aneurysm, resulting in a bleeding into the subarachnoid space, is the main cause of non-traumatic subarachnoid hemorrhage (aSAH) and accounts for 80% of SAH cases (1). Generally aSAH is characterized by high mortality where one third of the patients die within several days to weeks after occurrence. After incidence of aSAH most survivors tend to suffer from lasting disability or cognitive impairment (2).

Over time multiple metadata studies were performed determining the incidence rate of aSAH. Worldwide the incidence of aSAH is estimated to be ranging between 6 per 100.000 persons. This figure however varies significantly amongst regions (3). Although aSAH is an relatively uncommon type of stroke it accounts for a relatively high amount of lost productive life-years due to its high mortality and relatively young incidence compared to ischemic stroke which is the most common type of stroke (4).

Known risk factors for aSAH include sex, age, smoking, alcohol consumption, hypertension and a family history of SAH (5–8). These risk factors however do not fully account for the variation in aSAH. The exact mechanisms through which these risk factors cause aSAH are not yet fully known but are thought to include damage to the vessel wall, the loss of ability to repair the vessel wall, hemodynamic stress and the interaction of these effects (9). The risk of rupture of an aneurysm is correlated to the size of an aneurysm however due to the rarity of large aneurysms most patients tend to have smaller aneurysms (10).

Currently management of small aneurysms can be quite difficult as risk prediction models such as PHASES often fail to asses the risk of rupture for these kind of aneurysms (11). PHASES assesses risk of rupture according to a set of known risk factors (12). Knowing what risk factors play an important role in rupture of aneurysms could improve these models and could therefore be a key in prevention of aSAH cases and thus potentially lowering lost life years.

Machine learning algorithms are trained to classify data and are used in a range of applications inside medical science such as: the detection of chronic diseases (13), discovery of risk factors for diseases (14) or drug discovery and development (15). These algorithms generally work best when the number of data points per class is roughly the same. These imbalanced data sets are known for causing challenges (16). The event of rupture of an aneurysm is small, therefore it is expected that in prospective cohort studies such as the UK Biobank only a fraction of the participants are affected, resulting in an imbalanced data set.

Many different approaches to improve classification on imbalanced data are available but generally good results are booked by applying one or a combination of the following techniques. Sampling the classes so that these become more balanced, iteratively train a model on bootstrapped data (bagging) or the use of an ensemble of algorithms (boosting) (17–19).

This study will aim to provide better understanding of the effects of imbalanced data sets to scoring metrics of several machine learning algorithms. The UK Biobank was used as an real world application of imbalanced data to find new risk factors of aSAH.

Method

Study design and participants

Baseline data as well as data from follow-up assessments coming from the UK Biobank were analyzed (20). The UK Biobank is an population-based cohort consisting of 500.000 participants aged between 40 and 69 years old in the UK. Between 2006 and 2010 the participants were recruited and assessed on a voluntary basis. This assessment consisted of questionnaires such as mental health, cognitive function, diet or occupational history. Blood samples, urine samples and saliva samples were also taken (21,22). Additional data such as an dietary assessment (23) on large subsets of the participants were added later on (20). Follow up data on medical conditions as well as deaths is acquired by linkage to databases from the national health service (NHS) (20).

Cases of non-traumatic SAH were selected using the International Classification of Diseases, edition 10 (ICD-10) codes I600-I609. Participants diagnosed with SAH before the baseline assessment of the UK Biobank (field 53) were excluded. In case a person had multiple diagnoses of SAH the date of the first diagnosis was taken as reference.

Statistical analyses

Preprocessing

A pipeline was built to process data from the UK Biobank and select relevant features which then were used to train several machine learning algorithms (fig. 1). Using metadata downloaded from the website of the UK Biobank a selection of features was made consisting of data with continuous, integer, categorical value type. No features were filtered out a priori out in order to keep this analysis as much as possible free of hypotheses.

Further subselection of features was made by doing univariate statistical comparison between the selected SAH cases and a control group (24,25). A univariate approach was chosen due to resource constraints. To calculate p-values for the features different tests were used depending on the datatype of the feature. A Welch's t-test was performed on normally distributed continuous and integer features. Normality was checked by performing a D'Agostino-Pearson test. The non-parametric Mann-Whitney U test was used on the non-normally distributed continuous and integer features. A chi-squared test was used on the nominal categorical data and the non-parametric Kruskall-Wallis test was used on ordinal categorical features. A sub-selection of significant features was made using the results from the statistical tests ($p < 0.05$).

To remove non-sensical or duplicate data, features were removed manually from the data. This includes features like count freeze thaw cycles, acquisition route or percentages where counts are available for blood count, features that explain the acquisition methodology (e.g. sodium in urine device ID and weight method). Female specific features (e.g. had menopause and age at first live birth) were removed to prevent multicollinearity between these features as they would all be an indicator for female sex.

Nominal categorical features were processed using dummy variables, creating binary features as a result. Ordinal categorical features were transformed to binary features to improve efficiency by limiting the total number of features e.g. alcohol intake frequency consisting of six categories ranging from daily or almost daily to never was transformed so that the first three categories represent high alcohol intake and the last three no alcohol intake.

Model training

A sample of 100.000 controls were randomly selected from the total UK Biobank population due to performance constrictions. Both controls and SAH cases were used in a pipeline to train the model iterative for 10 times (fig. 1).

The data was split into a training set consisting of 80% of the data and a test set of consisting of the remaining 20% of the data. The data was split in a stratified fashion to preserve the ratio of SAH cases in the data.

Missing data was accounted for in both the train and test groups separately using means imputation from the scikit-learn simple imputer. Standardized scaling was applied to decrease the effect of the high variation in magnitude of the data.

The prevalence of SAH in the UK Biobank is ~1:1.000 resulting in a ~5:1.000 SAH to control ratio after selecting 100.000 controls. Imbalanced data sets can weaken the discriminative ability of models for minority classes (19). To correct for this imbalanced data synthetic minority over sampling technique (SMOTE) (26) was applied to the training subset followed by under sampling using the imbalanced-learn library for python (27). This resulted in a ratio of 1:10 after applying SMOTE to the SAH cases and a ratio of 1:2 after under sampling the control cases.

Both train and test groups were bootstrapped with replacement to create a new data set for each iteration.

Five different supervised learning methods covering different classes of algorithms were considered. These classifiers consisted of a penalized logistic regression (LogisticRegression) (28), a parametric support vector machine (parametric SVM) (29), a random forest (RandomForest) (30), a multilayer perceptron (MLP) (31) and a gradient booster (AdaBoost) (32). Hyperparameters for each model were tuned using a halving grid search (table S2). All models were implemented using scikit-learn (33). For each iteration of each model AUROC, AUPCR, precision, recall, accuracy, F-1 and F-beta metrics were calculated.

Performance analysis

The predictive performance of each model was assessed using the calculated scoring metrics as well as the receiver operating curve and precision recall curve. The resulting feature importance was assessed using the mean coefficients (logistic regression and SVM) or mean feature importance (random forest). Additionally SHAP values were generated using the python SHAP library to get a unified comparison between all models (34). Mean feature importance was calculated for the nonlinear models (AdaBoost, RandomForest and MLP) as these performed best and thus resulted the most explaining features. To get the variance in feature importance between the models, the absolute average SHAP feature importances for each of the nonlinear models were ranked.

Final model classification results were fine-tuned in postprocessing by optimizing the F-1 score through moving the decision threshold (35).

Software

All preprocessing and machine learning steps were performed using python version 3.6.9, scikit-learn version 0.24.2 (33), imbalanced-learn version 0.8.0 (27) and SHAP version 0.39.0 (34). The source code of this project can be found on Github (36,37).

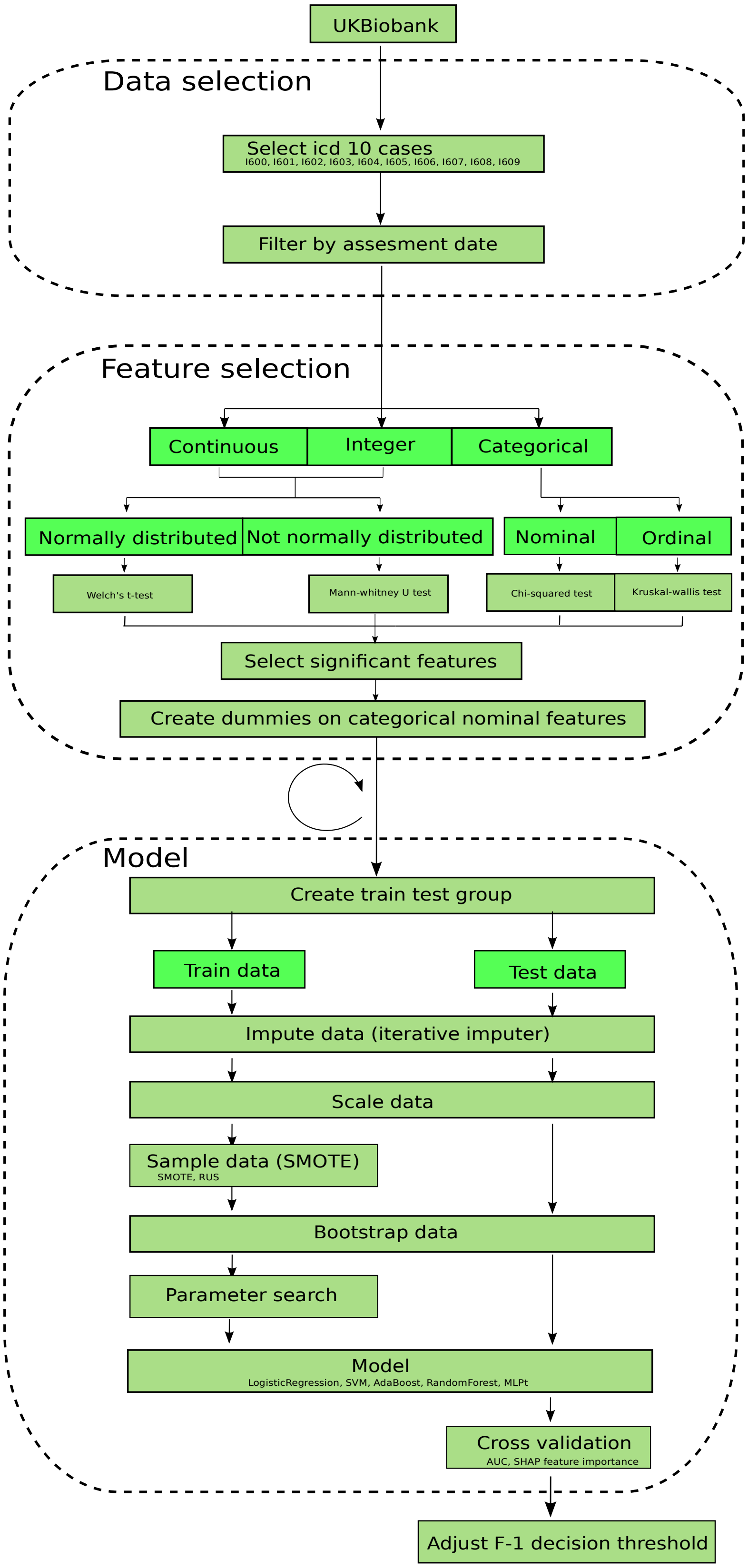


Figure 1) Pipeline used in this study. The pipeline consists of fours steps. 1) Selection of participants that are diagnosed with SAH after assessment. 2) Selection of significant different features between the SAH group and the control group. 3) Bagging of the machine learning models. 4) Evaluation of the outcome metrics.

Results

539 participants of the UK Biobank were diagnosed with SAH after recruitment. The average age at recruitment for participants in the UK Biobank was 57 year and 58 year for the SAH group. Overall 55% of the participants in the UK Biobank were female and 67% of the SAH group. The 5-year risk of SAH in the UK Biobank was 0.06% after recruitment giving SAH an incidence of 12 per 100.000 participants per year.

Statistical analysis and filtering of features with more than 90% of missing data resulted in 79 features that differ significantly between controls and SAHs (table S3).

Both linear classifiers (LogisticRegression and SVC) performed significantly worse on classifying SAHs in comparison to the nonlinear classifiers (AdaBoost, RandomForest, MLP) resulting in lower AUPCR and AUROC scores (table 1, fig. 2). Between the nonlinear classifiers AdaBoost performed best in terms of AUPCR (0.49 ± 0.14) followed by RandomForest (0.38 ± 0.10) and MLP (0.35 ± 0.11). Both AdaBoost and RandomForest show high recall scores of 0.99 ± 0.01 and 0.98 ± 0.02 respectively while MLP has a lower recall score of 0.39 ± 0.10. The precision score for both AdaBoost (0.01 ± 0.01 * 10^{-1}) and RandomForest (0.01 ± 0.03 * 10 $^{-1}$) is lower than the recall score of MLP (0.25 ± 0.10).

All models were trained to maximize AUPCR which was then used to inspect model performance. The precision recall plots show a clear difference in model performance between nonlinear models (fig. 3a) and linear (fig. 3b). For linear models the precision recall curve is slightly below or above baseline which was defined as the ratio of SAH in the test data.

For each model the iteration with AUPCR closest to the mean was chosen by moving the decision threshold to maximize F-1 score. This resulted in higher F-1 scores for all models (table 2) and in particularly AdaBoost, which suffered from a high number of false positives. The precision of the AdaBoost classifier went up from 0.005 to 0.72 after adjusting decision thresholds and recall went down from 0.99 to 0.45. This was also true for RandomForest for which the precision went up from 0.01 to 0.55 and the recall decreased from 0.98 to 0.52. For MLP precision increased from 0.25 to 0.84 and recall also increased from 0.39 to 0.58 after adjusting the threshold. The difference in performance between linear and nonlinear models still remained after adjusting decision thresholds.

Looking at the ranked feature importances models tend to agree on the five highest ranking features after which the variance in ranking between models increases (fig. 4b). Looking at the resulting average SHAP feature importances models predict a no association between SAH occurrence and the features: high sodium in urine, high creatinine in urine, high noise pollution, older age at recruitment, high neuroticism score and male sex. A association is found between SAH and the features are current smoking status and alcohol intake (fig. 4a).

Table 1) Average outcome metrics of the bagging steps for each model.

	Accuracy	F-1	Recall	AUPCR	Precision	AUROC	F-beta 0.5
AdaBoost	0.03 ± 0.03	0.01 ± 0.00	1.00 ± 0.01	0.49 ± 0.14	0.01 ± 0.00	0.93 ± 0.03	0.01 ± 0.00
RandomForest	0.46 ± 0.16	0.02 ± 0.01	0.98 ± 0.02	0.38 ± 0.10	0.01 ± 0.00	0.97 ± 0.01	0.01 ± 0.00
MLP	0.99 ± 0.01	0	0.39 ± 0.10	0.35 ± 0.11	0.25 ± 0.10	0.83 ± 0.05	0.27 ± 0.10
LogisticRegression	0.61 ± 0.05	0.01 ± 0.00	0.38 ± 0.06	0.01 ± 0.00	0.01 ± 0.00	0.58 ± 0.02	0.01 ± 0.00
SVC	0.54 ± 0.06	0.01 ± 0.00	0.31 ± 0.05	0.01 ± 0.00	0.00 ± 0.00	0.49 ± 0.03	0.00 ± 0.00

Table 2) Outcome metrics after adjusting the threshold of the bagging step that was closest to the average AUPCR score to optimize F-1 score for each model.

	Accuracy	F-1	Recall	Precision	F-beta 0.5
AdaBoost	1	0.55	0.45	0.72	0.64
RandomForest	0.99	0.52	0.5	0.54	0.54
MLP	1	0.58	0.45	0.84	0.71
LogisticRegression	0.51	0.02	0.75	0.01	0.01
SVC	0.99	0.02	0.02	0.02	0.02

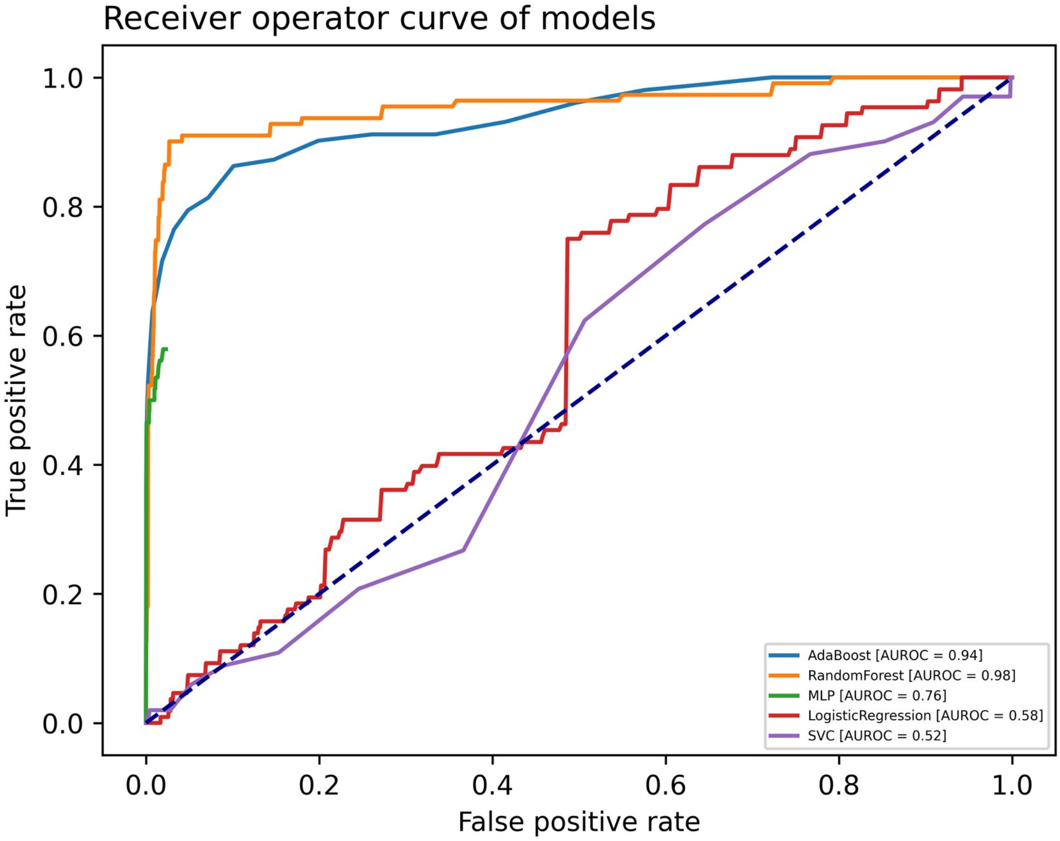


Figure 2) Receiver operator curves of the used models. The ROC of the bagging step that was closest to the mean AUPCR score was plotted. AUROC values of the corrsponding bagging step are given for each model: AdaBoost = 0.94, RandomForest = 0.98, MLP = 0.76, LogisticRegression = 0.58 and SVC = 0.52.

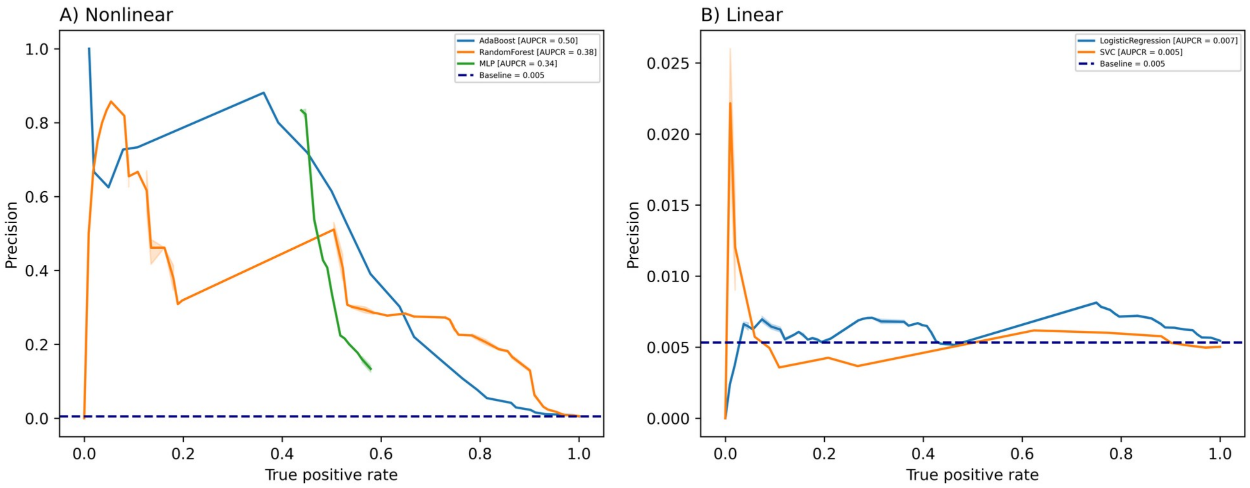


Figure 3) precision recall curves of the used algorithms. The precision recall curve of the bagging step that was closest to the mean AUPCR score was plotted. A) Precision recall curves of the nonlinear models. AUPCR values: AdaBoost = 0.50, RandomForest = 0.38 and MLP = 0.34. B) Precision recall curves of the linear models. AUPCR values: LogisticRegression = 0.007 and SVC = 0.005.

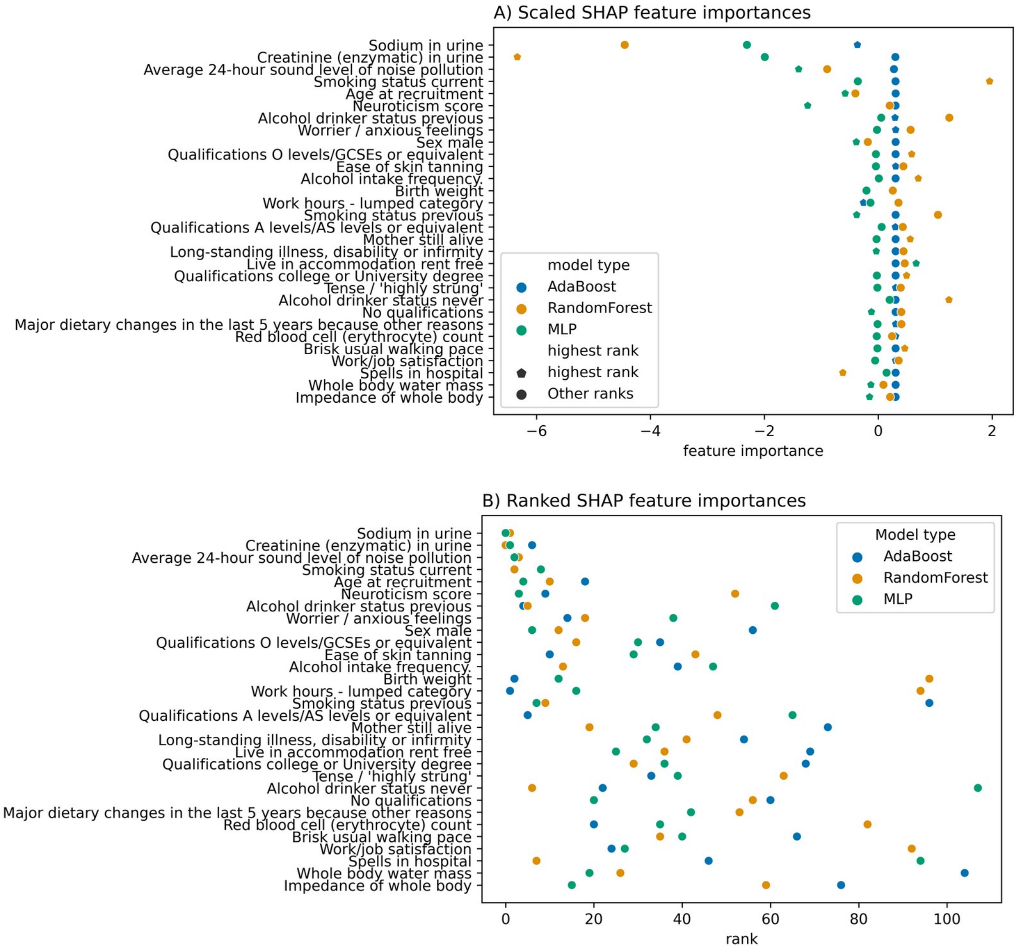


Figure 4) Resulting SHAP feature importances and ranks for the nonlinear models. On the y-axis the top 30 highest ranking features are given when looking at the absolute mean SHAP values. A) Scaled SHAP feature importances for the nonlinear models. For each feature the model with that ranks this feature highest was marked. B) For each nonlinear model ranks for the highest ranking features are given.

Discussion

In this study different machine learning algorithms were compared using a range of outcome metrics and applied to UKBiobank data to find new potential risk factors of SAH. Linear classifiers (LogisticRegression and SVM) performed worse than nonlinear classifiers (AdaBoost, RandomForest and MLP). The AUPCR score is a better measurement of model performance than the AUROC score in imbalanced data. Several potential new risk or protective factors for SAH were identified.

Data of each participant in the UK Bioabnk has been sampled systematically, generating a high number of features that can be used. This is as strength of using the Uk biobank for hypothesis free detection of new risk factors in comparison to other studies that also like the UK Biobank, include a high number of participants but only measure certain features (38). A difference compared to this study is that the UK Biobank (2006-present) has not been established as long as the FINRISK study (1972-2007) which found 437 cases of SAH out of ~64.000 participants compared to 539 cases this study found out of ~500.000 patients in the UK Biobank. Another study performed in Norway found 132 cases of SAH in ~75.000 participants and started in 1984 until 2005 (39). It is expected however that the number of cases of SAH in the UK Biobank will increase as the study continues keeping it relevant for the detection of new risk factors in the future.

The UK Biobank consists of participants that are that have on average higher economic status and better health in comparison to the general population (40). This could indicate that the found potential new risk factors might be specific risk factors for this group and not for the general population. However known risk factors such as smoking, alcohol intake, female sex occurred in this study as risk factors indicating the power of the trained models to identify important features. The UK Biobank consists of participants that were in the age range of 40-69 at recruitment (20) which gives these participants a higher risk for SAH when compared to the general population (41). This study found a 2:1 female to male ratio of SAH occurrence which was also found in the UK Biobank.

Other studies that searched for risk factors for SAH using the UK Biobank could not be found. There are however studies that searched for risk factors of other diseases (14,42). Both studies used primarily the GradientBooster algorithm as this algorithm performs well on high numbers of features. In this study the GradientBooster algorithm was not used due to resource constrictions. A difference between the studies is that in this study the outcomes of different algorithms are compared showing that there is a relative high variance in the ranking of risk factors between the models. In this study multiple outcome metrics were evaluated due to the imbalancedness of the data, in the other studies AUROC and F-1 was used (42) or only AUROC despite reporting high class differences (14).

Both linear classifiers (LogisticRegression and SVM) perform substantially worse than the nonlinear classifiers (AdaBoost, RandomForest and MLP) when looking at the AUROC and AUPCR scores. Worse performance of these classifiers on imbalanced data has been reported on different data sets (43–45). Generally however there tends to be a ~10% performance drop when compared to non-linear classifiers which is a significant difference when compared to the ~40% difference in performance found in this study. Variation in performance of each model was accounted for by bagging the dataset resulting in average model performance scores. A clear reason for this difference in performance could not be found as all data preprocessing steps between the models were kept equal and regularization was applied to the models. Overfitting of the nonlinear models could also not explain this difference since training and test AUROC for the nonlinear models do not display high variation. In order to get better performance, logistic or SVM algorithms that are better optimized for imbalance data could be used. One such algorithm is Rare Event Weighted Logistic Regression (46). This algorithm was not tested in this research due to its unavailability in the scikit-learn package.

Comparing the resulting AUROC scores between the models shows a clear difference between the linear and nonlinear models. Both the AdaBoost and RandomForest classifier show a high AUROC score of over 0.9. Looking at just AUROC score can however give a skewed view of the performance. This tends to be the case because of the imbalanced nature of the data in which the high number of true negatives can either mask the relatively high number of false positive compared to the true positives or the low number of true positives (47). For this reason it is better to use alternative metrics such as AUPCR, f1 or fbeta score to define model performance in imbalanced data sets. Looking at the result the models show a clear trade of between the number of false positives and the number of true positives which is reflected in the trade of between precision and recall (fig. S1). Both AdaBoost and RandomForest classifier score high in recall and relatively low on precision while for MLP classifier this is the other way around. However for MLP the difference between precision and recall was lower than the difference for AdaBoost and RandomForest. In short AdaBoost and RandomForest predict more true positives at the cost of more false positives while MLP is more balanced. This result is reflected in the F-1 and F-beta scores for which MLP scores highest.

Both AdaBoost and SVM classifiers suffered from high numbers of false positives. Moving the threshold so that the f1 score was optimized, decreased the number of false positives which was reflected in the increased precision for both models. As F-1 score is defined as a balance between precision and recall, moving thresholds comes at the cost of predicting true positives. In cases where it is important to classify all true positives or minimize the number of false positives alternative metrics for optimization could be used such as F-beta which allows to put emphasis on either precision or recall.

Adjusting decision thresholds in postprocessing could be a necessary step to finetune the balance between false positives and true positives. The balance that is required will be specific for the requirements of the project as in some cases it will be important to get all true positives and false positives are less important or other cases where the number of false positives is relevant. To get an idea of what the best scoring model is in the context of this research a combination of F1-score and recall would be best to look at. This is the case because in the search of risk factors it is important to get an high amount of correctly predicted SAH to associate risk factors with and thus accept somewhat higher number of false positives. For other cases such as the prediction of patients with high risk of SAH in patients with an aneurysm, a high number of false positives could mean that the workload for doctors would increase and also the added stress of operation for patients that fall in the category of high risk of SAH. In this case it could be better to look at a combination of F1-score and precision. For the case in this research the RandomForest classifier would score best after adjusting the thresholds for F1-score (table 2). For the other case the MLP classifier would score better (table 2).

Current smoking status and previous alcohol intake were predictors for SAH in the models which are already established risk factors for occurrence of SAH (5,6).

High levels of average 24 hour noise pollution have a negative association with SAH. Indicating that the inverse, low levels of 24 hour noise pollution are predictive for SAH. Currently no literature is available that associates noise pollution to SAH. The level of noise pollution could be an proxy for the real predictor which then was not included into the data. One hypothesis could be that that the level of noise polution is a proxy for the living environment/area of a participant.

Average creatinine and sodium levels in urine were lower in the SAH group compared to the levels in the control group. Measurements of features in the UK BioBank were conducted in a similar fashion for all participants so that this difference in means can not be explained by the principle of confounding by indication. AdaBoost, RandomForest and MLP ranked these features the highest on average. The average SHAP feature importance for both features was negative, implying that high creatinine or sodium levels in urine have a negative association with SAH. The other way around a low level of creatinine and sodium is predictive for SAH. Low levels of excretion of creatinine are generally associated with poor health, increased risk of cardiovascular disease such as coronary artery disease, chronic kidney disease and general increased risk of mortality (48–50). Out of these, chronic kidney disease is associated with SAH and other types of strokes (51).

Whether these potential new risk factors are real risk factors or a proxy for a different risk factor still remains unknown. Further research to determine wether low urinal creatinine, low urinal sodium levels and low noise polution are associated with SAH and if so, what undelying mechanisms are responsible. and whether low creatinine and sodium levels could be risk factors for SAH is needed. The UK Biobank has a variety of female specific features such as ever had hysterectomy, had menopause and ever used hormone replacement therapy. Since female sex is regarded as a risk factor for SAH, a model could be trained that include these features and only consists of female SAH cases to find female specific risk factors.

Conclusion

Outcome metrics of machine learning algorithms can give a overoptimistic result, especially when considering imbalanced data. Other metrics such as AUPCR or F-1 score give a more balanced result. A hypothesis free search for potential new risk factors for aneurysmal subarachnoid hemorrhage (aSAH) was performed using a set of machine learning algorithms. Resulting from the models high levels of sodium, creatinine and 24 hour noise level were associated with aSAH besides the traditional risk factors such as smoking, alcohol intake and female sex.

References

1. van Gijn J, Rinkel G. Subarachnoid haemorrhage: diagnosis, causes and management. Brain. 2001;124(2):249–278.

2. Nieuwkamp DJ, Setz LE, Algra A, Linn FH, de Rooij NK, Rinkel GJ. Changes in case fatality of aneurysmal subarachnoid haemorrhage over time, according to age, sex, and region: a meta-analysis. Lancet Neurol. 2009;8(7):635–642.

3. Etminan N, Chang H-S, Hackenberg K, De Rooij NK, Vergouwen MD, Rinkel GJ, et al. Worldwide incidence of aneurysmal subarachnoid hemorrhage according to region, time period, blood pressure, and smoking prevalence in the population: a systematic review and meta-analysis. JAMA Neurol. 2019;76(5):588–597.

4. Johnston SC, Selvin S, Gress DR. The burden, trends, and demographics of mortality from subarachnoid hemorrhage. Neurology. 1998;50(5):1413–1418.

5. Woo D, Khoury J, Haverbusch M, Sekar P, Flaherty M, Kleindorfer D, et al. Smoking and family history and risk of aneurysmal subarachnoid hemorrhage. Neurology. 2009;72(1):69–72.

6. Juvela S, Hillbom M, Numminen H, Koskinen P. Cigarette smoking and alcohol consumption as risk factors for aneurysmal subarachnoid hemorrhage. Stroke. 1993;24(5):639–646.

7. Korja M, Lehto H, Juvela S, Kaprio J. Incidence of subarachnoid hemorrhage is decreasing together with decreasing smoking rates. Neurology. 2016;87(11):1118–1123.

8. Lindekleiv H, Sandvei M, Njølstad I, Løchen M-L, Romundstad P, Vatten L, et al. Sex differences in risk factors for aneurysmal subarachnoid hemorrhage: a cohort study. Neurology. 2011;76(7):637–643.

9. Andreasen TH, Bartek Jr J, Andresen M, Springborg JB, Romner B. Modifiable risk factors for aneurysmal subarachnoid hemorrhage. Stroke. 2013;44(12):3607–3612.

10. Van Gijn J, Kerr RS, Rinkel GJ. Subarachnoid haemorrhage. The Lancet. 2007;369(9558):306–318.

11. Rutledge C, Jonzzon S, Winkler EA, Raper D, Lawton MT, Abla AA. Small aneurysms with low phases scores account for most subarachnoid hemorrhage cases. World Neurosurg. 2020;139:e580–e584.

12. Pagiola I, Mihalea C, Caroff J, Ikka L, Chalumeau V, Iacobucci M, et al. The

PHASES score: to treat or not to treat? Retrospective evaluation of the risk of rupture of intracranial aneurysms in patients with aneurysmal subarachnoid hemorrhage. J Neuroradiol. 2020;47(5):349–352.

13. Battineni G, Sagaro GG, Chinatalapudi N, Amenta F. Applications of machine learning predictive models in the chronic disease diagnosis. J Pers Med. 2020;10(2):21.

14. Madakkatel I, Zhou A, McDonnell MD, Hyppönen E. Combining machine learning and conventional statistical approaches for risk factor discovery in a large cohort study. Sci Rep. 2021;11(1):1–11.

15. Vamathevan J, Clark D, Czodrowski P, Dunham I, Ferran E, Lee G, et al. Applications of machine learning in drug discovery and development. Nat Rev Drug Discov. 2019;18(6):463–477.

16. Thabtah F, Hammoud S, Kamalov F, Gonsalves A. Data imbalance in classification: Experimental evaluation. Inf Sci. 2020;513:429–441.

17. Galar M, Fernandez A, Barrenechea E, Bustince H, Herrera F. A review on ensembles for the class imbalance problem: bagging-, boosting-, and hybrid-based approaches. IEEE Trans Syst Man Cybern Part C Appl Rev. 2011;42(4):463–484.

18. He H, Garcia EA. Learning from imbalanced data. IEEE Trans Knowl Data Eng. 2009;21(9):1263–1284.

19. Kaur H, Pannu HS, Malhi AK. A systematic review on imbalanced data challenges in machine learning: Applications and solutions. ACM Comput Surv CSUR. 2019;52(4):1–36.

20. Sudlow C, Gallacher J, Allen N, Beral V, Burton P, Danesh J, et al. UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age. PLOS Med. 2015 Mar 31;12(3):e1001779.

21. Elliott P, Peakman TC, on behalf of UK Biobank. The UK Biobank sample handling and storage protocol for the collection, processing and archiving of human blood and urine. Int J Epidemiol. 2008 Apr 1;37(2):234–44.

22. Ollier W, Sprosen T, Peakman T. UK Biobank: from concept to reality. 2005;

23. Liu B, Young H, Crowe FL, Benson VS, Spencer EA, Key TJ, et al. Development and evaluation of the Oxford WebQ, a low-cost, web-based method for assessment of previous 24 h dietary intakes in large-scale prospective studies. Public Health Nutr. 2011 Nov;14(11):1998–2005.

24. Cai J, Luo J, Wang S, Yang S. Feature selection in machine learning: A new perspective. Neurocomputing. 2018 Jul;300:70–9.

25. Khoshgoftaar T, Dittman D, Wald R, Fazelpour A. First Order Statistics Based Feature Selection: A Diverse and Powerful Family of Feature Seleciton Techniques. In: 2012 11th International Conference on Machine Learning and Applications [Internet]. Boca Raton, FL, USA: IEEE; 2012 [cited 2021 Nov 5]. p. 151–7. Available from: http://ieeexplore.ieee.org/document/6406743/

26. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: Synthetic Minority Over-sampling Technique. J Artif Intell Res. 2002 Jun 1;16:321–57.

27. Lemaıtre G, Nogueira F. Imbalanced-learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning. JMLR Org. 2017;18(1):559–63.

28. Hosmer Jr DW, Lemeshow S, Sturdivant RX. Applied logistic regression. Vol. 398. John Wiley & Sons; 2013.

29. Wang L. Support vector machines: theory and applications. Vol. 177. Springer Science & Business Media; 2005.

30. Breiman L. Random forests. Mach Learn. 2001;45(1):5–32.

31. Gardner MW, Dorling S. Artificial neural networks (the multilayer perceptron)—a review of applications in the atmospheric sciences. Atmos Environ. 1998;32(14–15):2627–2636.

32. Schapire RE. The boosting approach to machine learning: An overview. Nonlinear Estim Classif. 2003;149–171.

33. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine Learning in Python. JMLR Org. 2015;12:2825–30.

34. Lundberg SM, Lee S-I. A Unified Approach to Interpreting Model Predictions. In: Proceedings of the 31st international conference on neural information processing systems. 2017. p. 4768–4777.

35. Tosun A, Bener A. Reducing false alarms in software defect prediction by decision threshold optimization. In: 2009 3rd International Symposium on Empirical Software Engineering and Measurement. IEEE; 2009. p. 477–480.

36. lwaw. lwaw/UK-Biobank-risk-factors-ML: v1.1 [Internet]. Zenodo; 2022. Available from: https://doi.org/10.5281/zenodo.5866838

37. lwaw. lwaw/UK-Biobank-scraper: 1.1 [Internet]. Zenodo; 2022. Available from: https://doi.org/10.5281/zenodo.5914396

38. Korja M, Silventoinen K, Laatikainen T, Jousilahti P, Salomaa V, Hernesniemi J, et al. Risk factors and their combined effects on the incidence rate of subarachnoid hemorrhage–a population-based cohort study. PloS One. 2013;8(9):e73760.

39. Sandvei MS, Romundstad PR, Müller TB, Vatten L, Vik A. Risk factors for aneurysmal subarachnoid hemorrhage in a prospective population study: the HUNT study in Norway. Stroke. 2009;40(6):1958–1962.

40. Fry A, Littlejohns TJ, Sudlow C, Doherty N, Adamska L, Sprosen T, et al. Comparison of sociodemographic and health-related characteristics of UK Biobank participants with those of the general population. Am J Epidemiol. 2017;186(9):1026–1034.

41. Kongable GL, Lanzino G, Germanson TP, Truskowski LL, Alves WM, Torner JC, et al. Gender-related differences in aneurysmal subarachnoid hemorrhage. J Neurosurg. 1996;84(1):43–48.

42. Wong KC-Y, Xiang Y, Yin L, So H-C. Uncovering Clinical Risk Factors and Predicting Severe COVID-19 Cases Using UK Biobank Data: Machine Learning Approach. JMIR Public Health Surveill. 2021;7(9):e29544.

43. Muchlinski D, Siroky D, He J, Kocher M. Comparing random forest with logistic regression for predicting class-imbalanced civil war onset data. Polit Anal. 2016;24(1):87–103.

44. Ruiz A, Villa N. Storms prediction: Logistic regression vs random forest for unbalanced data. ArXiv Prepr ArXiv08040650. 2008;

45. Lin W, Wu Z, Lin L, Wen A, Li J. An ensemble random forest algorithm for insurance big data analysis. Ieee Access. 2017;5:16568–16575.

46. Maalouf M, Siddiqi M. Weighted logistic regression for large-scale imbalanced and rare events data. Knowl-Based Syst. 2014;59:142–148.

47. Jeni LA, Cohn JF, De La Torre F. Facing imbalanced data–recommendations for the use of performance metrics. In: 2013 Humaine association conference on affective computing and intelligent interaction. IEEE; 2013. p. 245–251.

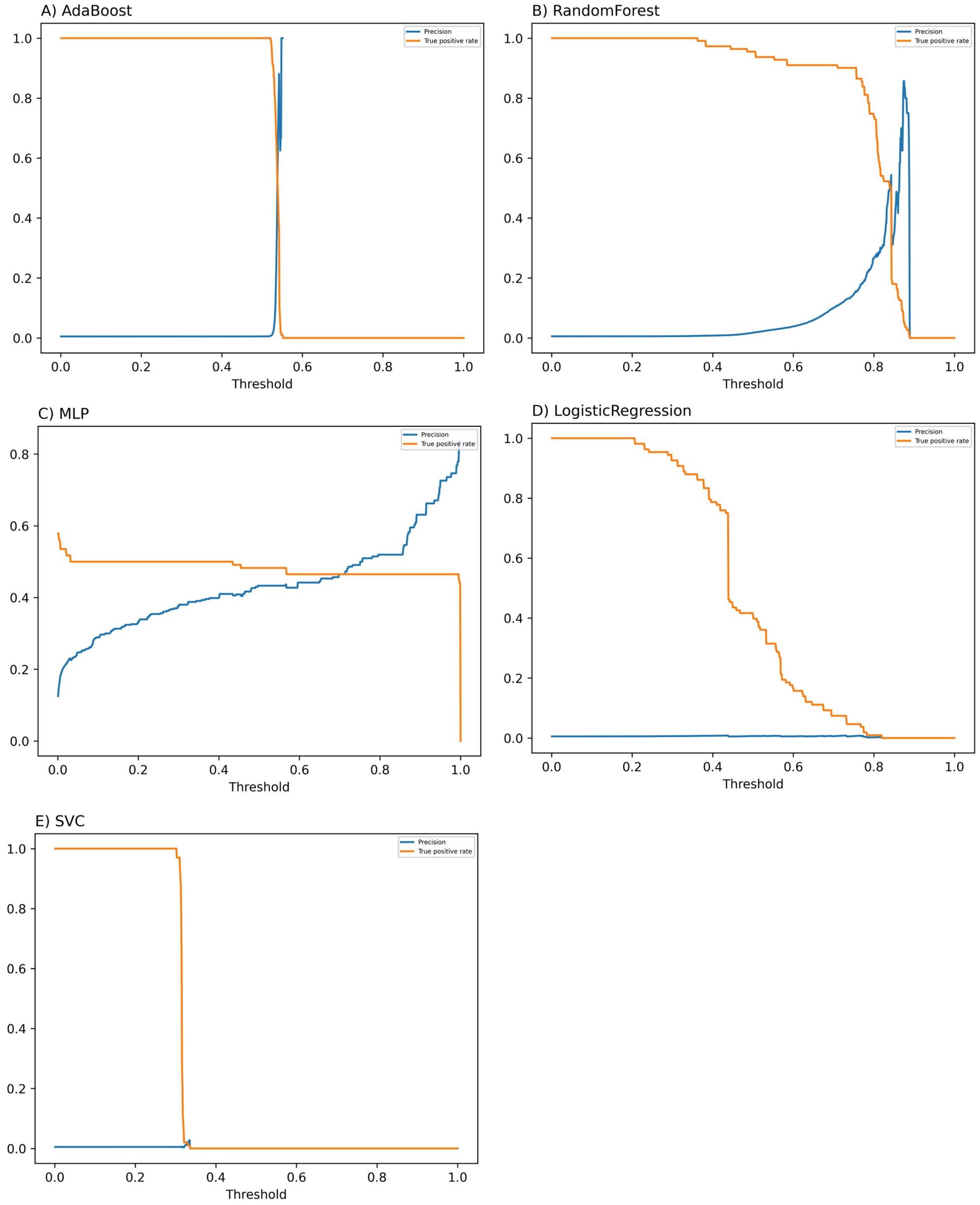
48. Diago CAA, Señaris JAA. Should we pay more attention to low creatinine levels? Endocrinol Diabetes Nutr Engl Ed. 2020;67(7):486–492.

49. Polinder-Bos HA, Nacak H, Dekker FW, Bakker SJ, Gaillard CA, Gansevoort RT. Low urinary creatinine excretion is associated with self-reported frailty in patients with advanced chronic kidney disease. Kidney Int Rep. 2017;2(4):676–685.

50. Ix JH, de Boer IH, Wassel CL, Criqui MH, Shlipak MG, Whooley MA. Urinary creatinine excretion rate and mortality in persons with coronary artery disease: the Heart and Soul Study. Circulation. 2010;121(11):1295–1303.

51. Ovbiagele B, Wing JJ, Menon RS, Burgess RE, Gibbons MC, Sobotka I, et al. Association of chronic kidney disease with cerebral microbleeds in patients with primary intracerebral hemorrhage. Stroke. 2013;44(9):2409–2413.

Supplements



A) AdaBoost

B) RandomForest

C) MLP

D) LogisticRegression

E) SVC

fig. S1) Trade of between precision and recall when moving the decision threshold for the different models. A) AdaBoost. B) RandomForest. C) MLP. D) LogisticRegression. E) SVC.

Table S2) Search grid used to search parameters for each model and the found optimal parameters for the bagging step that was the closest to the mean AUPCR.

	Search grid	Optimal parameters
AdaBoost	{'n_estimators': [5, 10, 50, 100, 200, 500, 700, 1000, 1500, 2000]}	n_estimators: 200
RandomForest	{'n_estimators': [10, 50, 100, 200, 500, 1000], 'max_depth': [2, 5, 10, 50, 100, 200, 500, 1000, None], 'min_samples_split': [2, 5, 10, 50, 100], 'min_samples_leaf': [2, 5, 10, 50, 100], 'class_weight': ['balanced', 'balanced_subsample']}	class_weight: balanced, max_depth: 500, min_samples_leaf: 2, min_samples_split: 2, n_estimators: 100
MLP	param_grid = {'hidden_layer_sizes': [x for x in itertools.product((10, 20, 30, 40, 50, 100), repeat = 3)], 'alpha': [0.0001, 0.001, 0.1]}	Alpha: 0.0001, hidden_layer_sizes: (100, 100, 50)
LogisticRegression	{'solver': ['saga'], 'penalty': ['elasticnet'], 'l1_ratio': [x*0.01 for x in range(1, 101, 1)], 'class_weight': ['balanced']}	class_weight: balanced, l1_ratio: 0.34, penalty: elasticnet, solver: saga
SVC	{'C': [x*0.01 for x in range(1, 101, 1)], 'kernel': ['linear'], 'class_weight': ['balanced']}	C: 0.35000000000000003, class_weight: balanced, kernel: linear

Table S3) Selected features and corresponding p-value resulting from the statistical tests.

Description	P-value
Doctor diagnosed silicosis	0
Doctor diagnosed cystic fibrosis	0
Doctor diagnosed mesothelioma of the lung	0
Doctor diagnosed fibrosing alveolitis/unspecified alveolitis	0
Doctor diagnosed alpha-1 antitrypsin deficiency	0
Basal metabolic rate	0
Whole body water mass	0
Peak expiratory flow (PEF)	0
Red blood cell (erythrocyte) count	0
Smoking status	0
Sex	0
Forced vital capacity (FVC)	0
Sitting height	0
Impedance of whole body	0
Doctor diagnosed asbestosis	0
Standing height	0
Type milk consumed	0
Drive faster than motorway speed limit	0
How are people in household related to participant	0
Medication for pain relief, constipation, heartburn	0
Mean sphered cell volume	0
Doctor diagnosed idiopathic pulmonary fibrosis	0
Haemoglobin concentration	0
Time spent watching television (TV)	0
Eye problems/disorders	0
Able to walk or cycle unaided for 10 minutes	0
Number of operations, self-reported	0
Maximum workload during fitness test	0
Body mass index (BMI)	0
Doctor diagnosed lung cancer (not mesothelioma)	0
Average total household income before tax	0
Contra-indications for spirometry	0
Alcohol drinker status	0.01
Headaches for 3+ months	0.01
Sodium in urine	0.01
Fluid intelligence score	0.01
Age at recruitment	0.01
Number of treatments/medications taken	0.01
Spells in hospital	0.01
Frequency of depressed days during worst episode of depression	0.01
Current employment status	0.01
Doctor restricts physical activity due to heart condition	0.01
Qualifications	0.01
Englyst dietary fibre	0.01
Alcohol intake frequency.	0.01
Attendance/disability/mobility allowance	0.01
Reticulocyte count	0.01
Age completed full time education	0.01

Vascular/heart problems diagnosed by doctor	0.01
Frequency of friend/family visits	0.01
Number in household	0.01
Position of the pulse wave peak	0.01
Work/job satisfaction	0.02
Worrier / anxious feelings	0.02
Neuroticism score	0.02
Long-standing illness, disability or infirmity	0.02
Heel bone mineral density (BMD)	0.02
Food weight	0.02
Work hours - lumped category	0.02
Time spent using computer	0.02
Creatinine (enzymatic) in urine	0.02
Usual walking pace	0.03
Sexual interference by partner or ex-partner without consent as an adult	0.03
Length of time at current address	0.03
Average 24-hour sound level of noise pollution	0.03
Ever depressed for a whole week	0.04
White blood cell (leukocyte) count	0.04
Bread intake	0.04
Drinking water intake	0.04
Own or rent accommodation lived in	0.04
Lifetime number of sexual partners	0.04
Tense / highly strung	0.04
Sleeplessness / insomnia	0.04
Taking other prescription medications	0.04
Ease of skin tanning	0.04
Mother still alive	0.05
Major dietary changes in the last 5 years	0.05
Birth weight	0.05
Father still alive	0.05