Article

SMMR
STATISTICAL METHODS IN MEDICAL RESEARCH

Statistical Methods in Medical Research
0(0) 1–10

DOI: 10.1177/0962280220980484
journals.sagepub.com/home/smm
SAGE

Class imbalance in gradient boosting classification algorithms: Application to experimental stroke data

Olga Lyashevska[1,2], Fiona Malone[1], Eugene MacCarthy[1], Jens Fiehler[3], Jan-Hendrik Buhk[3] and Liam Morris[1]

Abstract

Imbalance between positive and negative outcomes, a so-called class imbalance, is a problem generally found in medical data. Imbalanced data hinder the performance of conventional classification methods which aim to improve the overall accuracy of the model without accounting for uneven distribution of the classes. To rectify this, the data can be resampled by oversampling the positive (minority) class until the classes are approximately equally represented. After that, a prediction model such as gradient boosting algorithm can be fitted with greater confidence. This classification method allows for non-linear relationships and deep interactive effects while focusing on difficult areas by iterative shifting towards problematic observations. In this study, we demonstrate application of these methods to medical data and develop a practical framework for evaluation of features contributing into the probability of stroke.

Keywords

Imbalanced data, gradient boosting, classification algorithm, trees, stroke, oversampling

1 Introduction

Medical datasets are often imbalanced, i.e. the number of positive and negative cases is not equal.[1,2] Ironically, we might be interested in occurrence of some rare disease or high risk patients which tend to be in the minority. When positive cases are rare, for example, 90% of patients without disease (class 0) and 10% with disease (class 1), conventional classification methods will typically have accuracy up to 90%. In this case, even if all data points are predicted as 0's, results still will be correct 90% of the times. This due to the fact that most classification algorithms implicitly assume an equal occurrence of classes and aim to improve the overall accuracy of the model without accounting for uneven distribution of the majority and minority classes.[3] Therefore, the impact of class imbalance on classification performance can be detrimental. It results in failure to predict positive cases and in inability of classifier to predict high-impact cases when measures need to be taken. It is clear that the classifier performance cannot be simply expressed in terms of the accuracy, but other criteria reflecting the ability of the classifier to find all positive outcomes need to be taken into consideration.

Stroke is one of the leading causes of death worldwide. Over 80% of all strokes are acute ischemic strokes (AIS), from which 15–20% are due to cardiogenic emboli. Atrial Fibrillation (AF) is the most significant contributor to thrombus formation within the heart and is responsible for 45% of all cardio-embolic strokes.

[1]Enterprise Ireland Medical and Engineering Technologies Gateway, GMIT, Galway, Ireland
[2]Marine and Freshwater Research Centre, GMIT, Galway, Ireland
[3]Department of Diagnostic and Interventional Neuroradiology, University Medical Center Hamburg-Eppendorf, Hamburg, Germany

Corresponding author:
Olga Lyashevska, Enterprise Ireland Medical and Engineering Technologies Gateway, GMIT, Dublin Road, Galway, Ireland
Email: olga.lyashevska@gmit.ie

Embolisation is one of the most important issues facing clinicians today, especially due to the fatal consequences of stroke. It is not known why some patients with AF tend to embolise and others do not.[4]

The chance of such emboli, of varying sizes and composition, causing a stroke under various flow types has been previously evaluated in experimental settings.[4,5] However, a rigid statistical approach to establish relationship between the trajectory paths of blood clots, AF flowtypes and clot dimensions is required. An interesting hypothesis requiring investigation is the relationship between clot diameter and length and that relationship's influence in clot trajectory and stroke occurrence. It is difficult to record the original dimension and geometry of the stroke causing clot upon clot removal from a stroke patient. The medical device used to retrieve the clot from the cerebral vasculature has altered the clot and thus the clot removed was not exactly the way it was when it travelled to the brain. This study aims to investigate the influence that geometrical characteristics of cardio emboli and aortic arch may play in stroke occurrence.

However, identifying factors influencing the probability of strokes and quantifying their effect is difficult. Imbalance in data sets is a common problem that has been comprehensively studied in classical machine learning literature[6]; however, there is little uptake in medical literature.[7,8] A well-balanced dataset is very important, since imbalanced class distribution hampers the detection of minority class. Incorrect prediction of the minority, i.e. a blood clot travelling to the head, can affect the overall prediction of stroke among patients. There are two outcomes if this occurs: (1) a prediction of a blood clot travelling to the head, when this did not a occur, i.e. predicting stroke when a clot did not travel to the head, and (2) a prediction of no blood clot travelling to the head when this did occur, i.e. a stroke did occur but not predicted. Both cases are not helpful in determining stroke etiology, but perhaps case 2 is more dangerous in that a prediction of stroke has been missed which could be fatal to the patient. In terms of an experimental point of view, it hinders the prediction that clot dimension and flow type and stroke occurrence are related.[4]

A common strategy to overcome the class imbalance problem is to resample the original training dataset to decrease the overall level of class imbalance. Resampling is done either by oversampling the minority (positive) class and/or undersampling the majority (negative) class until the classes are approximately equally represented. Both strategies have limitations such as loss of data with undersampling or overfitting with oversampling, and the effect of resampling has rarely been evaluated.[9] Synthetic Minority Oversampling Technique (SMOTE) is one way of achieving class balance. It is designed to generate new synthetic data that is coherent with the minority class distribution while minimising overfitting.[10]

To evaluate features contributing into the increased probability of stroke, gradient boosting classification trees are particularly useful. Gradient boosting is an algorithmic model that does not require any assumptions on data distribution or relationships among features. It was intensively studied by the authors in literature.[11,12] It is a powerful algorithm that combines two algorithms: regression trees and boosting to predict binary outcomes and to estimate interpretable models.[13] It is successful because of its ability for automatic feature selection. Boosting is popular in medical literature.[13–15]

The goal of classification is to make valid predictions for unlabelled future or unobserved data. That is in the case of stroke, prediction probability of event (stroke = 1) using a set of features. When data is imbalanced, most classification methods perform poorly on minority class. In such cases, gradient boosting is particularly useful because it focuses on difficult areas by iterative shifting towards problematic observations. These problematic observations are identified by the large residual. Gradient boosting allows for non-linear relationships and deep interactive effects. It produces a prediction model in the form of an ensemble of weak prediction models, typically decision trees. Gradient boosting is a sequential process and thus every time it makes an incorrect prediction, it focuses more on that incorrectly predicted data point. So, if the first iteration gave you an accuracy of 90%, the second iteration would focus on the remaining 10%.

Gradient boosting is therefore not a single model, but rather a collection of simple additive models predictions that are averaged to give a more robust estimate of the response. Such methods are called ensemble methods.[12] Within ensemble methods, gradient boosting trees (GBT) form a supervised machine learning algorithm, which naturally allows for complex nonlinear interactions between features.[16] They do not require any assumptions on data distribution, but use an algorithmic model to learn the relationship between the response variable and the features and to find patterns. The objective of the algorithmic model is to minimise mean squared error (or other metric), by training each successive tree on the errors left over by the collection of earlier trees.

Despite the fact that gradient boosting is rather suitable for imbalanced data, it is very likely to be biased towards one of the classes at the cost of other. This situation can be rectified by oversampling of the minority class. Here, we combine gradient boosting approach which is particularly suitable for imbalanced data with a minority oversampling technique SMOTE, to analyse experimental stroke data. We investigate the effect of

oversampling on the performance of the classifier and contrast the results with another frequently applied classifier – random forest.

2 Materials and methods

2.1 Data

A total of 1084 bovine blood clot mimics were fabricated with varying concentrations of thrombin (0–20 NIHU/ml blood).[17] Clot length was maintained below 10 mm in length as this length is the most commonly seen among clots retrieved in stroke procedure.[18] A physiological simulation system was designed to experimentally analyse the trajectory patterns of blood clots of varying size and thrombin concentration through patient-specific aortic arch models, under different flow conditions: steady flow, normal pulsatile flow and AF pulsatile flows.

Four aortic arch image data sets in 2 D DICOM format were obtained to fabricate 3 D aortic arch models. These datasets were of AF patients who had suffered a stroke. The four aortic arches displayed all three major aortic arch geometries as described by Uflacker[19]: Romanesque ($n = 2$), Crenel ($n = 1$), and Gothic ($n = 1$). The Romanesque arch shape is the most common (80%) and resembles a normal rounded aortic arch compared to the less common Crenal and Gothic shapes.

The commercially available, reconstruction software, Mimics (Materialise, Leuven, Belgium) was used to generate the 3 D geometries from the 2 D medical image data sets. The models were 3 D printed in rigid Watershed stereolithography material (LPE 3 D printing, Belfast). One Romanesque model was replicated in silicone which was more compliant that the Watershed material.

The clots were released at the beginning of each cardiac cycle through a primed connector. The clots were injected into the system via a 3-arm connector that allowed for the clot to enter the system from the left, right and centre positions. Trajectory patterns were recorded through the left and right subclavian arteries (l-sub and r-sub), left and right common carotid arteries (L-CCA and R-CCA) and descending aorta. Two cameras (50 Hz frame rate, 12 MPixels) were used to monitor EA trajectories.[4,5]

Trajectory patterns were combined to produce the outcome variable Y with 2 classes coded as 0, 1. Blood clots that travelled through vessels that lead to the cerebral vasculature, i.e. the LCCA and RCCA, were combined together as clots with the ability to cause a stroke and referred to as Class 0. Class 1 referred to trajectories through all other vessels that do not lead to the cerebral vasculature, i.e. lsub, rsub and descending aorta, and thus no ability to cause a stroke. Other parameters, clot size, arch type, entry point, flowrate and thrombin were used as features.

2.2 Oversampling

There were 141 samples in the minority class and 943 in the majority class originally. The minority class was oversampled by creating synthetic examples, which were introduced using the k-nearest neighbour algorithm. The value of k depends on the amount of oversampling required. The amount of oversampling needed was 668% with five nearest neighbours. The amount of oversampling was defined by dataset size and class proportions.

The effect of oversampling on the performance of the Gradient boosting classifier was analysed using varying degrees of oversampling in the range of 0.25 and 1, with a step of 0.1. This was achieved through splitting data into training and validation folds and using stratified k-fold cross validation, then on each fold the minority class was oversampled for which the classifier was trained. The remaining fold was used to validate the classifier.

2.3 Gradient boosting

Formal treatment of the algorithm is presented in Friedman.[16]

Gradient boosting trees (GBT) considers additive models of the following form

$$F_m(x) = \sum_{m=1}^{M} \gamma_m h_m(x) \tag{1}$$

where γ_m is a learning rate and $h_m(x)$ are weak learners.

GBT uses decision trees of fixed size as weak learners. Decision trees have a number of abilities that make them valuable for boosting, namely the ability to handle data of mixed type and the ability to model complex functions. GBT builds the additive model in a forward stepwise fashion

$$F_m(x) = F_{m-1}(x) + \gamma_m h_m(x) \tag{2}$$

At each stage the weak learner $h_m(x)$ is chosen to minimise the loss function L given the current model F_{m-1} and its fit $F_{m-1}(x_i)$

$$F_m(x) = F_{m-1}(x) + \arg\min_h \sum_{i=1}^{n} L(y_i, F_{m-1}(x_i) - h_m(x)) \tag{3}$$

GBT attempts to solve this minimisation problem numerically via steepest descent: The steepest descent direction is the negative gradient of the loss function evaluated in the current model F_{m-1} which can be calculated for any differentiable loss function

$$F_m(x) = F_{m-1}(x) + \gamma_m \sum_{i=1}^{n} \nabla_F L(y_i, F_{m-1}(x_i)) \tag{4}$$

where the step length γ_m is chosen using line search

$$\gamma_m = \arg\min_\gamma \sum_{i=1}^{n} L\left(y_i, F_{m-1}(x_i) - \gamma \frac{\partial L(y_i, F_{m-1}(x_i))}{\partial F_{m-1}(x_i)}\right) \tag{5}$$

This algorithm is similar for regression and classification and only differs in the loss function used.

The accuracy of gradient boosting can be improved by introducing randomisation into the procedure through taking randomly selected subsets of training data at each iteration (hence stochastic gradient boosting).

2.4 Performance criteria

To estimate the performance of any classification algorithm, a set of performance criteria can be derived from confusion matrix (Table 1).

Accuracy is calculated as proportion of correctly classified cases to the overall number, i.e. $\frac{a+d}{a+b+c+d}$. Other performance criteria which are more suitable for imbalanced data are precision $\frac{d}{b+d}$, and recall $\frac{d}{c+d}$.

Another performance measure that is used in classification algorithms is ROC curve (Receiver Operating Characteristics), which suggest how much model is able to distinguish between classes. The higher the curve, the better model at predicting zeros as zeros and ones as ones.

3 Results

The overall mean length of a blood clot was 6.33 mm with the diameter of 3.69 mm. Blood clots of class 1 on average were of greater length than those of class 0; the diameter, however, was slightly smaller (Figure 1). For both parameters, length and diameter uncertainty around the mean value was higher for blood clots of class 0.

Thirteen percent of clots went into one of the branches leading to head (class 1), and from the remaining (class 0) 87%, the majority went through the descending branch.

Table 1. Confusion matrix.

	Predicted negative	Predicted positive
True negative	a	b
True positive	c	d

3.1 Gradient boosting

Gradient boosting model achieved R^2 of 0.67 on the test dataset. A total of 100 boosting iterations with learning rate of 0.35 and maximum depth 8 was required by the model. Hyperparameters learning rate and maximum depth were tuned using exhaustive grid search which generated the best candidates from a grid of provided parameter value. Subsampling was set at 0.75.

3.2 Relative feature importance

Relative importance of features in predicting stroke expressed as class 1 (Figure 2) suggests that blood clot length is the most important feature contributing to the stroke. The diameter of the blood clot is among other relevant features.

3.3 Partial dependence plots

Feature interaction is summarised using one-way and two-way partial dependence plots (Figure 3). Partial dependence plots show the marginalised effect one or two features have on the predicted outcome of a model and indicate whether the relationship between the target and a feature is linear, monotonous or more complex.[16]

The probability of class 1 increases with the increase in blood clot dimensions expressed as clot length and diameter. However, at 3.75 mm diameter, there seem to be a shift towards negative contribution of blood clot

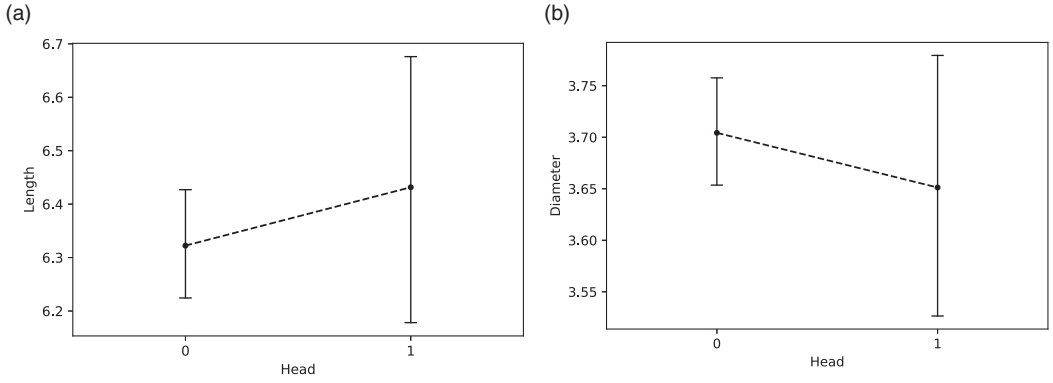

Figure 1. An estimate of central tendency for blood clot length (a) and blood clot diameter (b) based on the raw data (before oversampling). Error bars are calculated using the standard deviation of the observations. Uncertainty around the value is larger for class 1. The slope of the line that joins two outcomes allows changes in interaction to be judged.

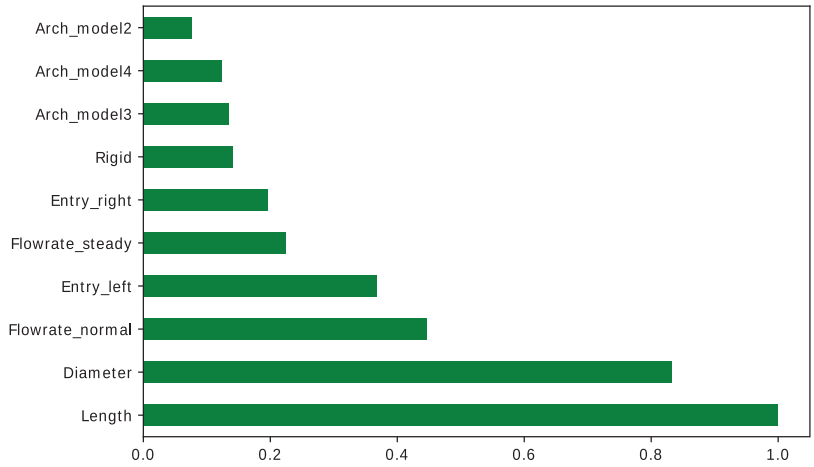

Figure 2. Relative feature importance ranked from low to high. The first level of each variable was removed to ensure that categorical variables are functionally independent linear combinations.

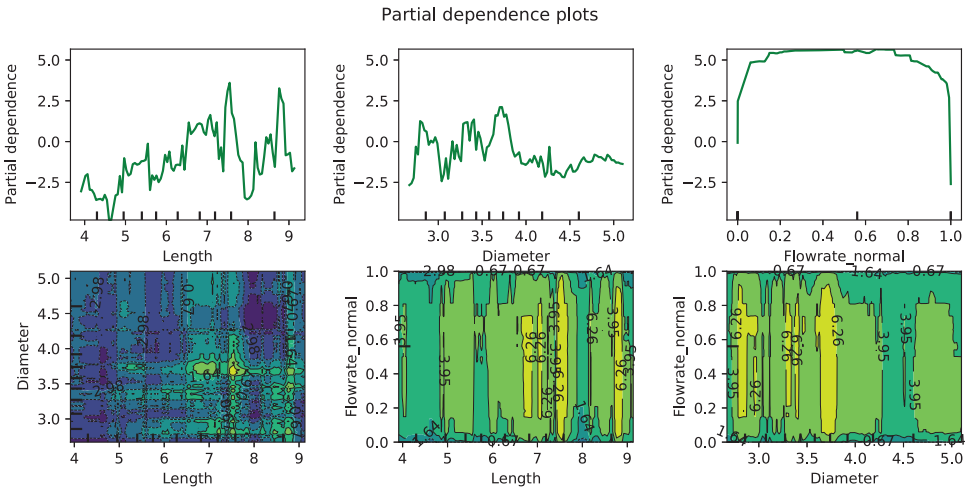

Figure 3. One-way and two-way partial dependence plots.

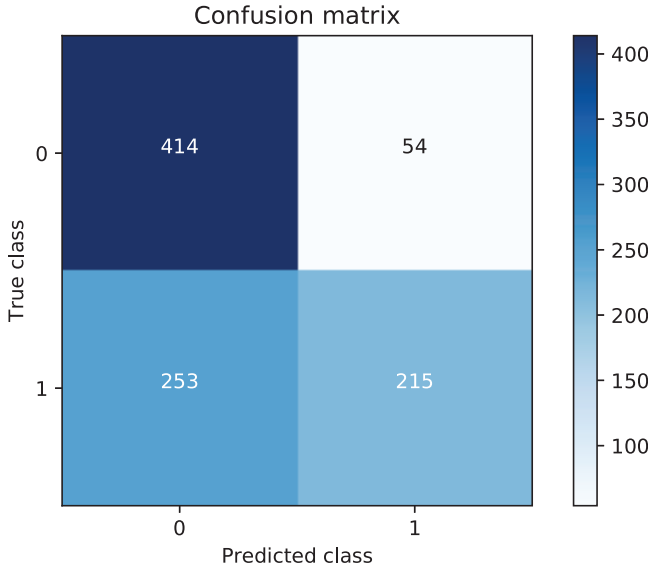

Figure 4. Confusion matrix.

diameter. The feature length shows a constant positive effect. The interaction between length and diameter seems to be complex.

3.4 Model performance

To describe performance of classification model, we use confusion matrix (Figure 4). Non-diagonal elements are inaccurate predictions. Precision or the ability of the classifier not to label a sample as positive if it is negative of prediction was higher for class 1 than for class 0, they were 0.78 and 0.61, respectively. Recall was the opposite. Recall of the positive class ('sensitivity') was 0.44, whereas recall of the negative class ('specificity') was 0.88.

The Receiver Operating Characteristic curve (Figure 5) shows that the model is able to distinguish between the majority and the minority classes.

3.5 The effect of oversampling

The effect of oversampling on the performance of gradient boosting classifier is shown in Figure 6(a). Recall that after the initial drop, followed by a steady increase between 0.4 and 0.7, there is a sharp increase between 0.8 and

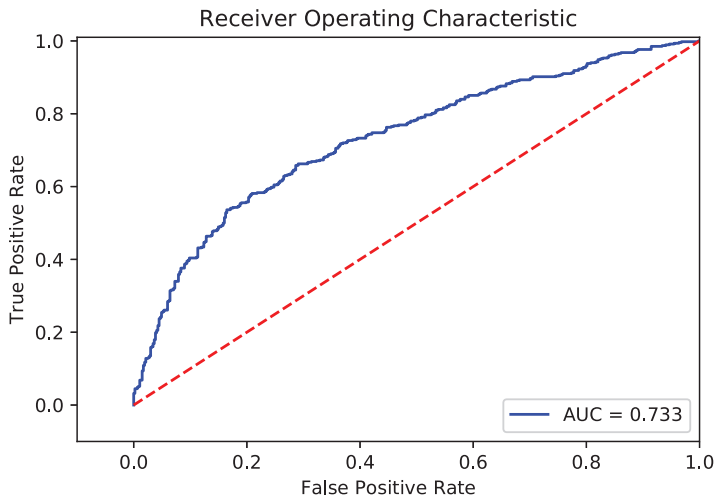

Figure 5. Receiver Operating Characteristic curve. The line at the top suggests that model is able to distinguish between the two classes.

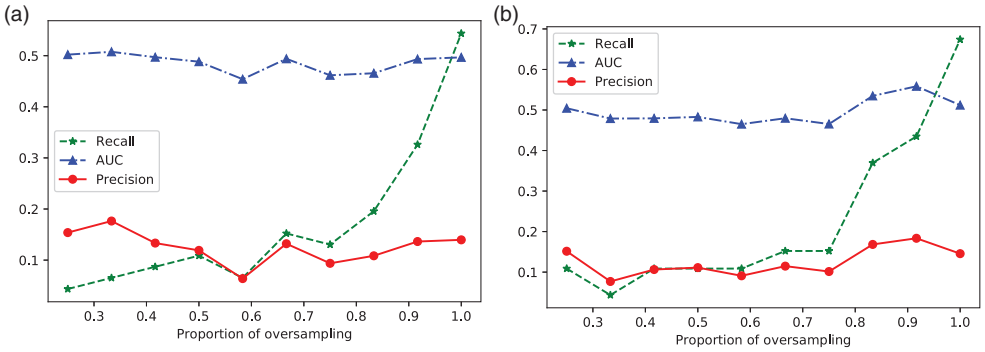

Figure 6. The effect of oversampling on precision, recall and AUC in gradient boosting classification (a) and random forest (b). Values are represented as average over all stratified k-folds.

1, reaching the best score of 0.71. This indicates that the ability of classifier to identify all positive outcomes (class 1) considerably improves when the proportion of oversampling is high. Precision or the ability not to label an outcome as positive that is negative, stays low with a slight increase after 0.8, decreasing again at 1. AUC is similar to precision, except for the higher values. Note that these values are slightly different from those reported in the main gradient boosting model, which is due to the fact that the stratified k-fold was implemented and the results were averaged over each fold.

The best score of random forest (Figure 6(b)) in contrast with gradient boosting was only 0.52, thus indicating the superiority of gradient boosting in terms of performance. It took 34 min to run gradient boosting compared to 108 min for random forest on laptop Intel(R) Core(TM) i5-5300U CPU @ 2.30 GHz.

4 Discussion and conclusions

Here we presented a study on class imbalance classification using simulated stroke data. We argued that rare cases are often of interest and correct prediction of them is of paramount importance in a medical diagnosis. Furthermore, we demonstrated that choosing accuracy as the performance criterion in class imbalance classification may be inaccurate and misleading. We indicated a range of other performance criteria that were more appropriate, in particular, the recall which is the ability of the classifier to find all positive outcomes (strokes). We also demonstrated the effect of oversampling and the importance of closing the gap between the majority and the minority class by combining gradient boosting with stratified k-fold while manipulating the proportion of oversampling. All metrics showed an improvement; however, the strongest effect was demonstrated by the recall.

This has important implications in medical diagnosis, since we are mostly interested in rare events such as diseases and their correct classification. Finally, gradient boosting algorithm compared with random forest algorithm confirmed its robustness and superiority. This is because gradient boosting is an additive (ensemble) model which works in a forward stage-wise manner building upon and improving the shortcomings of existing weak learners, whereas random forest is a collection of unrelated decision trees. However, care needs to be taken when data are noisy, and in such cases gradient boosting may result in overfitting.

Our results suggested that blood clots dimensions such as length and diameter were more important than a patient's aortic arch geometry. The probability of blood clot going to the head increased when the length was above average and diameter was below average. However, the uncertainty around the mean was large.

One way partial dependence plots showed that relationship between the feature length and the predicted outcome was approximately linear and monotonous. Similar effect was observed for feature diameter, although the relationship seemed to be less monotonous for the blood clots of large diameter. Furthermore, the two way partial dependence plots indicated a more complex relationship between the length and diameter in relation to the probability of blood clot entering the head.

4.1 Relationship between stroke and clot length

Figure 3 displays the one-way partial dependence of clot length. There is a linear relationship until a clot length of 7 mm. This shows that as the clot gets longer, there is an increased probability that the clot will travel toward the cerebral vasculature and cause a stroke. This was evident in experiments as longer clots indeed have the ability to change geometry under physiological flow conditions to fit through narrower blood vessels. During experimentation, it was often observed that longer clots tended to split resulting in either all of the broken clot travelling towards the brain or part of the clot travelling along the branching vessels and the rest through the descending aorta. However, the longer the clot was entering into the physiological simulation system, the heavier the specimen would become. The percentage flow rate that travelled through the branching vessels may be too low to transport the heavier clots which could account for the drop in dependency after 7 mm in length. Very short clots tended to travel through the descending aorta. This is most likely due again to percentage flow rate. Approximately 70% of a patient's blood flow travels through the descending aorta, which would have the power to transport the majority of clots.[4,5]

4.2 Relationship between stroke and clot diameter

Up until 3.5–4.0 mm, there seems to be a positive contribution of clot diameter on stroke (Figure 3). This is possibly due to the fact that the diameter of each branching vessel falls within this range. Therefore, the clot can physically fit through these vessels. After that, there is shift as the clots tend to be too large to travel through these narrow vessels and instead can fit through the much larger descending aorta, which averages between 2 and 3 cm.[19]

4.3 Other features

It is well known that aortic arch hemodynamics display large variability as a result of arch and branching variation and thus would influence clot embolisation. Numeric simulations suggest that aortic arch curvature is an important risk factor for embolic stroke.[20] The swirling helical flow associated with vascular curvature has been shown to play a substantial role in embolus transport within the carotid artery.[21]

However, our data show that, clot geometry is more important than arch geometry when analysing the hemodynamics that influence stroke occurrence. This study only used four cases of aortic arch of the most common type in the global population. These arches displayed a three-branch branching pattern, which is again the most commonly recorded in patients. However difficult to obtain, 1, 2 and 4 branch patterns are found within patient datasets. A greater number of datasets with aortic arch variation could provide more statistical significance with regard to the effect arch geometry has on emboli trajectories.

Another interesting result from this study is the limited difference shown between the rigid and flexible aortic arch models (Figure 2). This is important for future biomedical biosimulators as flexible model production is quite expensive and time consuming where rigid models can produce similar and quicker results.

4.4 Fabrication of emboli's size

The motivation related to the fabrication of the emboli's size was inspired by patient-specific cases reported in literature. Middle cerebral artery (MCA) occlusion is the most common site for cardioembolic strokes,[22] which measures a diameter of 3–4 mm and so would indicate a lodging embolus of similar diameter size. Thambidorai et al.[23] describe the size of atrial and atrial appendage emboli of 0.2–4.2 cm and 1.0–3.9 cm in width with an area of 0.1–8.0 cm^2 and 0.9–7.0 cm^2, respectively. Menke et al.[18] also describe clots associated with the left atrium ranging in size from a few millimeters to 4 cm. As mentioned, the original dimension and geometry of the stroke causing clot cannot be recorded. These values in literature are based on clot removed from a stroke patient.

Although the normal and AF flow profiles are considered representative of biological flow conditions in this study, it must be mentioned that the patient-specific outlet flow splits were unknown. The pressure at the outlet boundaries was also not measured and the rigidity of the models was also a limitation to the study. Future work should include more flexible model cases and patient physiological data such as outlet flow and blood pressure data, if possible, to create a true patient specific investigation. Also, patients who suffer a stroke risk usually present with other risk factors such as heart disease, age, duration of arrhythmia, chronic versus intermittent fibrillation, and atrial size.[24] Therefore, the presence of a possible cardiac source of embolism does not necessarily mean that the stroke was caused by a cardiogenic embolism.

For further research, we suggest including a greater sample size of clots/arches. Arches with varying branching patterns could investigate the role geometry plays in aortic hemodynamics and clot trajectories in much more detail. Two-branch and four branch patterns are also potential branching patterns found in patients. It is well known that aortic hemodynamics display large variability as a result of arch and branching variation. Larger clots need to be included. This study is important for the future work in stroke research. The data presented in this study can be used in future algorithms for stroke prediction and treatment. Employing a similar study using a variation in clot formation and aortic arch geometry would give a complete and satisfactory prediction mechanism for future work in stroke.

The platforms used in the analysis are given below:

```
Platform version:
Linux version 4.15.0-46-generic
(gcc version 7.3.0 (Ubuntu 7.3.0-16ubuntu3))
Python 3.6.5:: Anaconda, Inc.
```

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

ORCID iD

Olga Lyashevska https://orcid.org/0000-0002-8686-8550

Supplemental material

Data and source code supplementary to the paper are available on the Github of the main author.

References

1. Zhao Y, Wong ZSY and Tsui KL. A framework of rebalancing imbalanced healthcare data for rare events' classification: a case of look-alike sound-alike mix-up incident detection. *J Healthcare Eng* 2018; **2018**: 1–11.
2. Belarouci S and Chikh MA. Medical imbalanced data classification. *Adv Sci Technol Eng Syst J* 2017; **2**: 116–124.
3. Ali A, Shamsuddin SMH and Ralescu AL. Classification with class imbalance problem: a review. *Int J Adv Soft Comput Appl* 2015; **7**: 166–204.
4. Malone F, McCarthy E, Delassus P, et al. Investigation of the hemodynamics influencing emboli trajectories through a patient-specific aortic arch model. *Stroke* 2019; **50**: 1531–1538.
5. Malone F, McCarthy E, Delassus P, et al. Embolus analog trajectory paths under physiological flowrates through patient-specific aortic arch models. *J Biomech Eng* 2019; **141**: e101007.

6. Batista GEAPA, Prati RC and Monard MC. A study of the behavior of several methods for balancing machine learning training data. *SIGKDD Explor Newsl* 2004; **6**: 20–29, http://doi.acm.org/10.1145/1007730.1007735
7. Rahman MM and Davis DN. Addressing the class imbalance problem in medical datasets. *Int J Mach Learn Comput* 2013; **3**: 224–228.
8. Colak C, Karaaslan E, Colak C, et al. Handling imbalanced class problem for the prediction of atrial fibrillation in obese patient. *Biomed Res* 2017; **28**: 3293–3299.
9. Emanet N, Öz HR, Bayram N, et al. A comparative analysis of machine learning methods for classification type decision problems in healthcare. *Decis Analytics* 2014; **1**: 1–20.
10. Chawla NV, Bowyer KW, Hall LO, et al. Smote: Synthetic minority over-sampling technique. *J Artificial Intell Res* 2002; **16**: 321–357.
11. Freund Y and Schapire R. A decision-theoretic generalization of on-line learning and an application to boosting. *J Comput Syst Sci* 1997; **55**: 119–139.
12. Hastie T, Tibshirani R and Friedman J. *Elements of statistical learning*. Berlin, Germany: Springer, 2009.
13. Mayr A, Binder H, Gefeller O, et al. The evolution of boosting algorithms. from machine learning to statistical modelling. *Meth Inform Med* 2014; **53**: 419–427.
14. Mayr A, Hofner B, Waldmann E, et al. An update on statistical boosting in biomedicine. *Computat Math Meth Med* 2017; 2017: 1–12.
15. Zhang Z, Zhao Y, Canes A, et al. Predictive analytics with gradient boosting in clinical medicine. *Ann Translat Med* 2019; **7**, http://atm.amegroups.com/article/view/24543
16. Friedman JH. Greedy function approximation: a gradient boosting machine. *Ann Stat* 2000; **29**: 1189–1232.
17. Malone F, McCarthy E, Delassus P, et al. The mechanical characterisation of bovine embolus analogues under various loading conditions. *Cardiovasc Eng Technol* 2018; **9**: 489–502.
18. Menke J, Luthje L, Kastrup A, et al. Thromboembolism in atrial fibrillation. *Am J Cardiol* 2010; **105**: 502–510.
19. Uflacker R. Thoracic aorta and arteries of the trunk. In: *Atlas of vascularanatomy*. 2nd ed. Philadelphia, PA: Lippincott Williams & Wilkins, 2009.
20. Choi H, Luo T, Navia J, et al. Role of aortic geometry on stroke propensity based on simulations of patient-specific models. *Scientific Rep* 2017; **7**. DOI: 10.1038/s41598-017-06681-3.
21. Mukherjee D, Padilla J and Shadden SC. Numerical investigation of fluid–particle interactions for embolic stroke. *Theoretical Computat Fluid Dynam* 2016; **30**: 23–39.
22. Chung J, Park S, Kim N et al. Trial of ORG 10172 in Acute Stroke Treatment (TOAST) classification and vascular territory of ischemic stroke lesions diagnosed by diffusion-weighted imaging. *J Am Heart Assoc* 2014; **3**: e001119. DOI: 10.1161/JAHA.114.001119.
23. Thambidorai S, Murray R, Parakh K, et al. Utility of transesophageal echocardiography in identification of thrombogenic milieu in patients with atrial fibrillation (an ACUTE ancillary study). *Am J Cardiol* 2005; **96**: 935–941.
24. Leary MC and Caplan LR. Cardioembolic stroke: an update on etiology, diagnosis and management. *Indian Acad Neurol* 2008; **11**: 52–63.