Fuzzy Soft Sets for Stroke Risk Feature Selection: Age-Adjusted Ranking, Clinical Validity and Robustness to Severe Class Imbalance

Sisir Kumar Rajbongshi

Assistant Professor, Department of Computer Science, Pandit Deendayal Upadhyaya Adarsha Mahavidyalaya (affiliated to Gauhati University), Amjonga, Goalpara, Assam-783124, India

E-mail: sisirrajbongshi@gmail.com; ORCID: 0000-0003-1874-6882

Abstract

Feature relevance for stroke risk is usually judged with crisp statistics or post-hoc model explanations, neither of which describes a risk factor in the graded, linguistic terms clinicians use. This study ranks the candidate features of a public, de-identified stroke dataset (5,109 patients, 4.87% with stroke) with a fuzzy soft set discrimination score (FSDS), which sums the absolute differences between the mean Low, Medium and High memberships of stroke and non-stroke patients. The score is shown to be twice the total variation distance between the two classes' term distributions, which puts continuous and binary features on one scale; for a binary feature it is twice the difference in prevalence. Because crude scores mix risk factors with proxies for age, an age-adjusted score is introduced, computed within age strata and tested by permutation within strata. Without adjustment, "ever married" ranked second and "work type: children" sixth; after adjustment both lost significance ($q = 0.16$ and 0.84), and only average glucose level, hypertension, BMI and heart disease remained significant beside age ($q \leq 0.048$). In ten repetitions of five-fold cross-validation, with ranking repeated inside every training fold, the top eight FSDS features matched the full feature set and ANOVA, chi-square and mutual-information selection (logistic regression ROC-AUC 0.840 against 0.838–0.841; corrected resampled t-test, all $p > 0.4$). With class weighting, balanced random forests or SMOTE, recall was 42%–80%. FSDS selections were less stable than ANOVA selections in small samples. The contribution is interpretability and confound control, not higher accuracy.

Keywords

Fuzzy soft sets, feature selection, stroke risk, confounding, class imbalance, interpretable machine learning

1. Introduction

Stroke is the second leading cause of death and a leading cause of disability worldwide [1,2]. Its main modifiable risk factors (raised blood pressure, raised blood glucose, high body-mass index and heart disease) are well established, and age is its strongest non-modifiable risk factor [1]. Machine-learning studies of stroke risk on public electronic health record (EHR) extracts report high accuracy with boosting or stacking ensembles, often explained after training with SHAP or LIME [3–6]. In these studies the relevance of a feature is a crisp statistic or an attribution to a fitted model. Neither says, in the terms a clinician would use,

how the distribution of a risk factor differs between patients who had a stroke and those who did not.

Fuzzy soft sets [7–9] can express such statements directly: each attribute is described by linguistic terms such as Low, Medium and High, and every patient belongs to each term to a degree between 0 and 1. Fuzzy soft sets have a long record in medical decision-making [10,11], but a recent review [12] lists no use for feature selection and none for stroke. Fuzzy-rough methods are widely used for feature selection [13,14] but give no per-term description of the classes, and fuzzy scoring for stroke has so far served as a classifier input rather than as a ranking tool [15].

This study uses a simple fuzzy soft set discrimination score (FSDS) for this purpose and examines it critically on a public stroke dataset [16]. It addresses three problems that a naive application would ignore. First, most features in tabular EHR extracts are binary, so the score must be comparable across feature types. Second, univariate scores rank proxies of age highly, because unmarried patients and children are young and rarely have strokes. Third, with 4.87% positives accuracy is uninformative, and selecting features on the full dataset before cross-validation biases the evaluation [17,18].

The contributions are: (i) a proof (Proposition 1) that FSDS is twice a total variation distance, which explains what it measures and why continuous and binary features are comparable; (ii) an age-adjusted FSDS with a stratified permutation test, which on the stroke data recovers the four established modifiable risk factors and removes the age proxies; and (iii) a leakage-free comparison with three standard filters and four classifiers, with selection stability, a sweep over the number of selected features and sensitivity analyses. The comparison shows that FSDS matches, but does not exceed, the predictive value of standard filters and is less stable in small samples; these limitations are reported because they define where the method is useful.

2. Related work

Machine learning for stroke risk. The public stroke dataset used here [16] and close variants have been analysed many times. Dubey et al. [3] trained gradient boosting, AdaBoost and XGBoost on a 4,981-record version of a public Kaggle stroke dataset, explained the models with SHAP and LIME, and reported a test accuracy of 92.13% for XGBoost. Dritsas and Trigka [4] restricted the data to adults, balanced the classes with SMOTE [19] and reported an AUC of 0.989 for a stacking ensemble. Both aim at maximal predictive performance; age, hypertension, heart disease, glucose and BMI are repeatedly reported as the most influential inputs. Ben Kahla et al. [15] passed a fuzzy risk score computed from clinical data to standard classifiers and found that it improved accuracy. In their work the fuzzy model is part of the predictor; here it is used to rank and describe features before any classifier is trained.

Feature selection, leakage and stability. Feature selection methods are grouped into filter, wrapper and embedded methods [20,21]. Filters such as the ANOVA F-test, the chi-square statistic and mutual information [22] score each feature independently of the classifier and are the natural comparison for a univariate fuzzy score. Selecting features on all the data and then cross-validating the selected set gives optimistically biased estimates [17,18], and a good

selector should give similar subsets on similar samples, which is usually measured with the Kuncheva index [23,24].

Fuzzy soft sets and fuzzy-rough sets. Molodtsov's soft sets [8] parameterise a universe by attributes, and fuzzy soft sets [9] attach fuzzy subsets to those parameters. The algebraic structure of collections of fuzzy subsets has also been studied, including families of fuzzy subsets that form Boolean algebras [25–27]. In medicine, fuzzy soft sets have mainly been used for patient-level diagnosis by relation matching [10,11]; the review of Singh and Bhardwaj [12] covers these applications. Fuzzy-rough feature selection [13,14] measures the dependency between features and the class through lower and upper approximations; it is effective but does not give a class-wise linguistic description.

3. Methods

3.1. Preliminaries

Definition 1 (fuzzy set) [7]. A fuzzy set A on a universe U is a mapping $\mu_A: U \to [0, 1]$; $\mu_A(x)$ is the degree to which x belongs to A.

Definition 2 (soft set) [8]. Let E be a set of parameters and $\mathrm{P}(U)$ the power set of U. A pair (F, E) with $F: E \to \mathrm{P}(U)$ is a soft set over U.

Definition 3 (fuzzy soft set) [9]. Let $\mathrm{I}(U)$ be the set of fuzzy subsets of U. A pair (F, E) with $F: E \to \mathrm{I}(U)$ is a fuzzy soft set over U; for each parameter e, $F(e)$ is a fuzzy subset of U.

3.2. Linguistic partitions of the features

Let $U = \{u_1, \ldots, u_n\}$ be the patients, y_i the stroke label (1 for stroke, 0 otherwise), and U_1 and U_0 the stroke and non-stroke patients, with sizes n_1 and n_0. For a feature c with values x_i, a set of linguistic terms T_c is chosen, and the fuzzy soft set (F, T_c) assigns to each patient a membership $\mu_t(x_i)$ in each term t of T_c. The memberships form a partition of unity:

$$\Sigma_t \, \mu_t(x) = 1 \text{ for every value } x, \text{ the sum running over all terms } t \text{ of } T_c. \qquad (1)$$

Continuous features. For age, average glucose level and BMI, $T_c = \{\text{Low, Medium, High}\}$. With three breakpoints $p_1 < p_2 < p_3$ the memberships are

$$\mu_{\text{Low}}(x) = \min\{1, \max[0, (p_2 - x) / (p_2 - p_1)]\},$$
$$\mu_{\text{High}}(x) = \min\{1, \max[0, (x - p_2) / (p_3 - p_2)]\}, \qquad (2)$$

and $\mu_{\text{Medium}}(x) = 1 - \mu_{\text{Low}}(x) - \mu_{\text{High}}(x)$. Medium therefore rises from 0 at p_1 to 1 at p_2 and falls to 0 at p_3, and (1) holds by construction. Clinical breakpoints were used where they exist: 18.5, 25 and 30 kg/m² for BMI (WHO [28]), and 100 and 126 mg/dL for glucose (the fasting thresholds of the American Diabetes Association [29]) with their midpoint of 113 mg/dL. For age the quartiles of the training data were used (25, 45 and 61 years on the full data). Figure 1 shows the partitions.

Binary features. For a feature with values in $\{0, 1\}$, $T_c = \{\text{No, Yes}\}$ with $\mu_{\text{No}}(x) = 1 - x$ and $\mu_{\text{Yes}}(x) = x$. Each category of a categorical feature is a separate binary indicator. This is the crisp special case of (1).

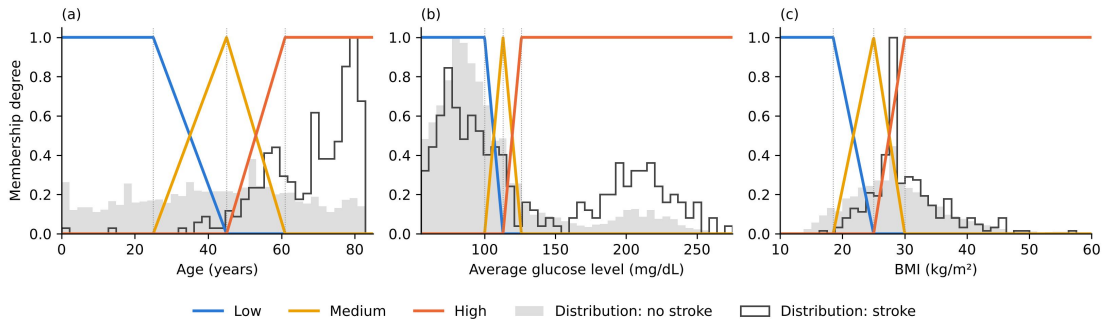


Figure 1. Linguistic partitions of (a) age, using training-data quartiles; (b) average glucose level, using the clinical thresholds 100 and 126 mg/dL and their midpoint; (c) BMI, using the WHO thresholds. Grey bars and black outlines show the distributions of non-stroke and stroke patients.

3.3. The fuzzy soft set discrimination score

For each term the class-conditional mean memberships are

$$M_1(t) = (1/n_1)\,\Sigma_{U1}\,\mu_t(x_i), \quad M_0(t) = (1/n_0)\,\Sigma_{U0}\,\mu_t(x_i), \qquad (3)$$

where Σ_{U1} and Σ_{U0} denote sums over the stroke and non-stroke patients. $M_1(t)$ is the degree to which stroke patients, on average, are described by term t; for example, M_1(High) = 0.86 for age means that stroke patients are, on average, "old" to degree 0.86. The fuzzy soft set discrimination score of feature c is

$$\Phi(c) = \Sigma_t\,|M_1(t) - M_0(t)|, \text{ the sum running over all terms } t \text{ of } T_c. \qquad (4)$$

Features are ranked by decreasing Φ. The following property explains what the score measures.

Proposition 1. For any feature whose memberships satisfy (1):

(a) M_1 and M_0 are probability distributions on T_c, and $0 \le \Phi(c) \le 2$;

(b) $\Phi(c) = 2\,\max_A\,|M_1(A) - M_0(A)|$, where the maximum is over all subsets A of T_c and $M_y(A)$ is the sum of $M_y(t)$ over t in A; that is, $\Phi(c)$ is twice the total variation distance between M_1 and M_0;

(c) For a binary feature with π_1 and π_0 the proportions of stroke and non-stroke patients having the attribute, $\Phi(c) = 2|\pi_1 - \pi_0|$.

Proof. (a) Memberships are non-negative and, by (1), the $M_1(t)$ sum to $(1/n_1)\,\Sigma_{U1}\,\Sigma_t\,\mu_t(x_i) = 1$; the same holds for M_0. Since $|a - b| \le a + b$ for non-negative a and b, $\Phi(c) \le 1 + 1 = 2$. (b) Let B be the set of terms with $M_1(t) > M_0(t)$ and $d = M_1(B) - M_0(B)$. Because both distributions sum to 1, the terms outside B contribute $M_0 - M_1 = d$ as well, so $\Phi(c) = 2d$. For any other set A, $M_1(A) - M_0(A)$ lies between $-d$ and d, so d is the maximum in (b). (c) For a binary feature, M_y(Yes) = π_y and M_y(No) = $1 - \pi_y$, so $\Phi(c) = |\pi_1 - \pi_0| + |(1 - \pi_1) - (1 - \pi_0)| = 2|\pi_1 - \pi_0|$. This completes the proof.

Proposition 1 has three consequences. Scores of continuous and binary features lie on the same scale and can be ranked together. For binary features FSDS is simply twice the difference in the prevalence of the attribute between the classes, so any advantage of the fuzzy representation can come only from continuous features. Finally, the terms of (4) form a

readable profile: the sign and size of $M_1(t) - M_0(t)$ say which linguistic description is over-represented among stroke patients.

3.4. Age-adjusted score and permutation test

A univariate score cannot distinguish a risk factor from a feature that is merely associated with a strong risk factor. In stroke data age dominates, and marital status, being coded as a child and employment type are strongly age-dependent. Following stratified analysis in epidemiology [30], the score is computed within age strata and averaged.

Definition 4 (age-adjusted FSDS). Divide the patients into age strata $s = 1, \ldots, S$, let $\Phi_s(c)$ be the score (4) computed from the patients of stratum s, and let n_{1s} be the number of stroke patients in stratum s. The age-adjusted score is

$$\Phi_{\text{adj}}(c) = \Sigma_s \, w_s \, \Phi_s(c), \quad \text{with weights } w_s = n_{1s} / n_1, \qquad (5)$$

where strata without stroke or without non-stroke patients are skipped and the weights are renormalised. Weighting by the number of strokes gives most weight to the strata in which strokes occur, so that the young strata, with very few strokes, cannot dominate through noisy estimates. Age keeps its unadjusted score. The main analysis used three strata (under 40, 40–59 and 60 years or older), containing 8, 60 and 181 strokes.

To decide which scores exceed chance, stroke labels were permuted 2,000 times and the scores recomputed. For the unadjusted score labels were permuted over all patients; for the adjusted score they were permuted within each age stratum, which keeps the association between age and stroke but removes any association between the feature and stroke beyond age stratum. The p-value is $(1 + B_c)/2001$, where B_c is the number of permutations with a score at least as large as the observed one. P-values were adjusted by the Benjamini–Hochberg procedure [31], and $q < 0.05$ was taken as significant. Computing all 17 scores for 5,109 patients takes about 5 ms.

4. Experimental setup

4.1. Data and candidate features

The study used the public Stroke Prediction Dataset [16]: 5,110 de-identified records with gender, age, hypertension, heart disease, ever married, work type, residence type, average glucose level, BMI, smoking status and a binary stroke outcome. The provider does not disclose the original source. Because the data are public and de-identified, no ethical approval was required. One record with gender “Other” was removed, leaving 5,109 patients, of whom 249 (4.87%) had a stroke. BMI was missing for 201 patients (3.9%) and was replaced by the median of the training fold. Gender, marital status and residence were coded as single indicators, work type as five indicators and smoking status as four (never, former, current, unknown). With hypertension and heart disease this gives 14 binary and 3 continuous candidate features. The partition of unity (1) held exactly for all 5,109 × 17 membership vectors.

4.2. Feature selection methods and classifiers

FSDS was compared with three filters on the same 17 features: the ANOVA F-statistic, the chi-square statistic (on features scaled to [0, 1]) and mutual information (nearest-neighbour

estimator for continuous and discrete estimator for binary features [22]). Two variants of FSDS were also evaluated: *crisp FSDS*, in which the triangular memberships are replaced by crisp intervals with boundaries at $(p_1 + p_2)/2$ and $(p_2 + p_3)/2$, to isolate the effect of fuzziness; and *age-adjusted FSDS* (5). Each method selected its top $k = 8$ features unless stated otherwise.

Four classifiers with standard settings were each combined with an imbalance strategy suited to it [32]: logistic regression and an RBF-kernel support vector machine (SVM), both with class weights inversely proportional to class frequency; a balanced random forest of 200 trees, each grown on a class-balanced bootstrap sample [33]; and k-nearest neighbours ($k = 7$) trained after SMOTE oversampling (SMOTE-NC when binary features were present) [19]. Scaling and oversampling were fitted on training folds only. The implementation used scikit-learn [34] and imbalanced-learn [35].

4.3. Evaluation protocol

Performance was estimated by ten repetitions of stratified five-fold cross-validation (50 splits). In every split, imputation, breakpoints, all feature scores and the selection of the top k features were computed from the training fold alone, and the classifier was evaluated on the untouched test fold. The primary metric was the area under the ROC curve (ROC-AUC), which does not depend on a threshold; balanced accuracy, recall and precision at the default decision rule are also reported [36]. Plain accuracy is not used, because a classifier that never predicts stroke already reaches 95.1%. FSDS was compared with each other feature set by the corrected resampled t-test [37,38], in which the variance of the 50 paired differences is inflated by the factor $1/50 + n_{\text{test}}/n_{\text{train}}$ to account for overlapping training sets (49 degrees of freedom).

Selection stability was measured on 50 stratified subsamples of 250, 500, 1,000 and 2,000 patients as the mean pairwise Kuncheva index of the selected subsets at $k = 5$ and $k = 8$ [23]; the index is 1 for identical subsets and about 0 for random agreement. Sensitivity of the rankings was checked against alternative age strata and against quartile breakpoints for glucose and BMI.

5. Results

5.1. Feature ranking and significance

Table 1 and Figure 2 give the scores on the full dataset. Age had by far the largest score ($\Phi = 1.022$). Ten features had significant unadjusted scores, including four with no plausible causal link with stroke: ever married (second), work type “children” (sixth), smoking status “unknown” (eighth) and self-employment (tenth). The unadjusted ranking agreed closely with the baselines (Spearman correlation 0.93 with ANOVA, 0.77 with chi-square and 0.91 with mutual information, all $p < 0.001$), and all four methods placed “ever married” among their top five features and “children” among their top six.

After age adjustment only four features remained significant beside age: average glucose level ($q = 0.003$), hypertension ($q = 0.003$), heart disease ($q = 0.008$) and BMI ($q = 0.048$), the four established modifiable risk factors [1]. “Ever married” fell from 0.478 to 0.091 ($q = 0.157$) and “children” from 0.266 to 0.004 ($q = 0.842$). Overall, 0.3% of patients coded as

children had a stroke, against 5.6% of the others, but all of them are young; within age strata the difference disappears. The risk factors' scores also fell, because part of their crude association is shared with age (hypertension from 0.352 to 0.184), but they stayed clearly above the permutation null.

Table 1. Fuzzy soft set discrimination scores on the full dataset (n = 5,109). The q-values are Benjamini–Hochberg-adjusted permutation p-values (2,000 permutations; within age strata for the adjusted score). "Adj." denotes the age-adjusted score; age is the stratification variable and is not tested after adjustment. The last three columns give each feature's rank under the ANOVA, chi-square and mutual-information (MI) filters.

Feature	Φ	q	Rank	Φ adj.	q adj.	Rank adj.	ANOVA	Chi-sq.	MI
Age	1.0221	< 0.001	1	1.0221	–	1	1	3	1
Ever married	0.4778	< 0.001	2	0.0910	0.157	8	5	5	3
Average glucose level	0.4038	< 0.001	3	0.2926	0.003	2	3	8	7
Hypertension	0.3523	< 0.001	4	0.1844	0.003	3	4	2	4
Heart disease	0.2833	< 0.001	5	0.1519	0.008	5	2	1	5
Work: children	0.2658	< 0.001	6	0.0043	0.842	16	6	4	6
BMI	0.2467	< 0.001	7	0.1543	0.048	4	10	15	2
Smoking: unknown	0.2385	< 0.001	8	0.0628	0.765	12	9	9	9
Smoking: former	0.2273	< 0.001	9	0.0541	0.765	13	7	6	8
Work: self-employed	0.2118	0.002	10	0.0760	0.655	9	8	7	10
Residence: urban	0.0716	0.478	11	0.0690	0.765	10	11	11	12
Work: private	0.0548	0.569	12	0.1042	0.651	6	13	13	13
Gender: male	0.0415	0.652	13	0.0952	0.655	7	14	14	14
Smoking: current	0.0299	0.640	14	0.0306	0.881	15	15	12	15
Smoking: never	0.0187	0.844	15	0.0666	0.765	11	16	16	16
Work: never worked	0.0091	0.569	16	0.0007	1.000	17	12	10	11
Work: government	0.0083	0.918	17	0.0425	0.818	14	17	17	17

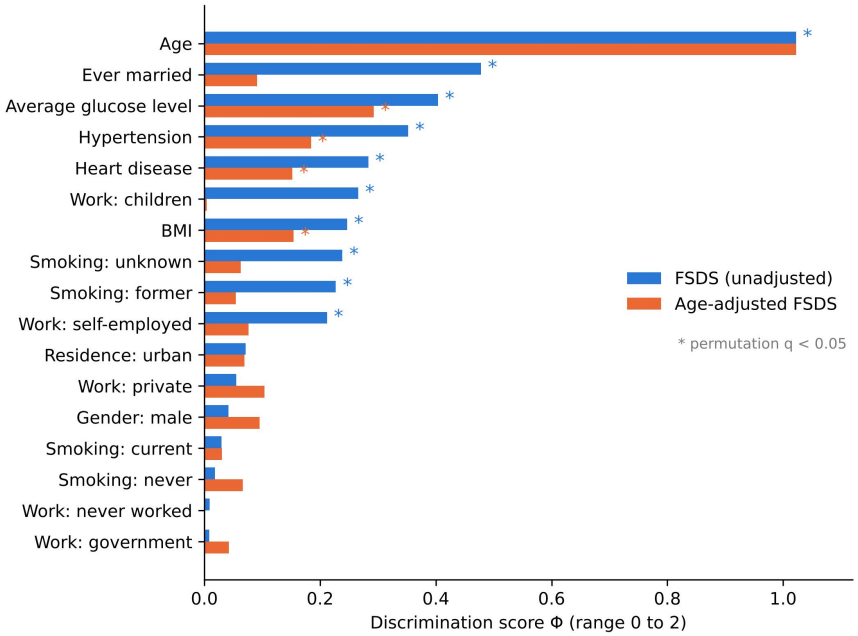


Figure 2. Unadjusted (blue) and age-adjusted (orange) FSDS of the 17 candidate features. Asterisks mark permutation q < 0.05. The adjusted bar for age repeats its unadjusted score.

5.2. Linguistic profiles

Figure 3 shows the class-conditional mean memberships behind the scores of the continuous features. Stroke patients belonged to "High age" (61 years or older) to degree 0.86 on average, against 0.35 for other patients, and almost never to "Low age" (0.02 against 0.39). Their mean membership in "High glucose" was 0.41, twice that of other patients (0.21). For BMI

the difference lay mainly in the "Low" term (0.06 against 0.18), while membership in "Medium" was the same in both classes. Among the binary features, hypertension was present in 26.5% of stroke patients against 8.9% of others, and heart disease in 18.9% against 4.7%. These profiles are the main output of the method for a clinical reader: each score in Table 1 decomposes into statements such as "stroke patients are more often high-glucose", with a measured size.

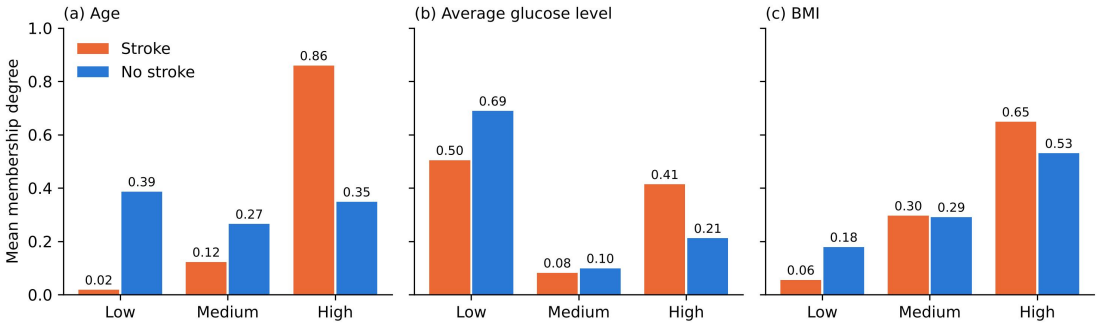


Figure 3. Linguistic profiles of the continuous features: mean membership of stroke (orange) and non-stroke (blue) patients in each term.

5.3. Classification performance

Table 2 reports the cross-validated performance. With logistic regression all selections of eight features performed alike (ROC-AUC 0.839–0.841, balanced accuracy 76.4%–76.6%, recall 79.2%–79.6%), as did the full set of 17 features (0.838, 76.4% and 79.2%). With selected features, recall was 70%–80% for logistic regression, the SVM and the balanced random forest, and 42%–51% for SMOTE-KNN, against 30% for SMOTE-KNN with all features. Precision was low for every method (9%–15%): with 4.87% prevalence, a threshold that catches most strokes produces many false alarms, which is why threshold-free AUC is the primary metric [36]. Crisp FSDS gave results identical to FSDS to within 0.2 points and is omitted from the table.

Table 2. Cross-validated performance (mean ± standard deviation over 10 × 5 folds, in %), with feature selection (k = 8) repeated inside every training fold.

Classifier	Feature set	ROC-AUC	Balanced accuracy	Recall	Precision
Logistic regression	All 17 features	83.8 ± 1.8	76.4 ± 2.2	79.2 ± 4.2	13.4 ± 0.9
	FSDS	84.0 ± 1.8	76.6 ± 2.4	79.6 ± 4.6	13.4 ± 0.9
	Age-adjusted FSDS	83.9 ± 1.8	76.4 ± 2.2	79.4 ± 4.4	13.3 ± 0.8
	ANOVA	84.1 ± 1.9	76.4 ± 2.2	79.2 ± 4.0	13.4 ± 0.9
	Chi-square	84.0 ± 1.8	76.5 ± 2.2	79.4 ± 4.2	13.4 ± 0.9
	Mutual information	84.0 ± 1.8	76.5 ± 2.3	79.4 ± 4.4	13.4 ± 0.9
SVM (RBF)	All 17 features	77.6 ± 2.7	69.7 ± 3.3	62.0 ± 6.6	12.3 ± 1.3
	FSDS	81.6 ± 2.1	74.2 ± 2.5	75.2 ± 5.1	12.6 ± 0.9
	Age-adjusted FSDS	80.0 ± 2.3	72.3 ± 2.5	69.8 ± 5.4	12.4 ± 0.9
	ANOVA	81.7 ± 2.3	74.7 ± 2.7	75.4 ± 5.3	13.0 ± 1.0
	Chi-square	81.3 ± 2.2	74.0 ± 2.8	73.7 ± 5.7	12.9 ± 1.1
	Mutual information	81.3 ± 2.0	73.9 ± 2.1	74.8 ± 4.2	12.5 ± 0.9
Balanced random forest	All 17 features	83.2 ± 2.0	76.2 ± 2.2	73.2 ± 4.6	15.4 ± 1.0
	FSDS	83.2 ± 2.2	75.8 ± 2.7	72.8 ± 5.4	15.0 ± 1.2
	Age-adjusted FSDS	83.2 ± 2.0	76.0 ± 2.7	72.6 ± 5.4	15.3 ± 1.2
	ANOVA	81.8 ± 1.9	74.0 ± 2.6	70.4 ± 5.2	13.9 ± 1.0

Classifier	Feature set	ROC-AUC	Balanced accuracy	Recall	Precision
	Chi-square	81.9 ± 1.8	74.3 ± 2.7	70.9 ± 5.4	14.1 ± 1.1
	Mutual information	83.1 ± 2.0	75.5 ± 2.4	72.0 ± 5.0	15.0 ± 1.1
SMOTE + KNN	All 17 features	65.9 ± 3.2	57.4 ± 3.3	29.6 ± 6.8	9.3 ± 1.9
	FSDS	71.9 ± 2.8	64.2 ± 3.3	47.6 ± 7.6	11.3 ± 1.2
	Age-adjusted FSDS	69.7 ± 3.2	62.1 ± 3.1	42.4 ± 6.4	10.6 ± 1.4
	ANOVA	71.7 ± 3.0	65.9 ± 2.7	51.3 ± 5.4	11.9 ± 1.3
	Chi-square	71.8 ± 2.9	65.5 ± 2.8	50.3 ± 6.1	11.8 ± 1.3
	Mutual information	71.5 ± 2.8	63.6 ± 3.1	46.1 ± 6.7	11.1 ± 1.3

Feature selection mattered for the kernel- and distance-based classifiers: FSDS selection raised the SVM AUC from 0.776 to 0.816 (p = 0.002) and the SMOTE-KNN AUC from 0.659 to 0.719 (p = 0.005). Table 3 compares FSDS with every other feature set. None of the nine comparisons with the baselines was significant for logistic regression, the SVM or KNN. For the balanced random forest FSDS was 1.4 AUC points better than ANOVA (p = 0.055) and chi-square (p = 0.047), because these filters ranked BMI tenth and fifteenth and excluded it; given the number of comparisons, this is not treated as evidence of superiority. Fuzzy and crisp FSDS were indistinguishable ($p \geq 0.58$). The age-adjusted selection matched FSDS for logistic regression and the random forest but was weaker for the SVM (recall −5.3 points, p = 0.036) and KNN, because it drops age proxies such as “ever married” that help prediction although they are not risk factors.

Table 3. Difference in ROC-AUC (percentage points) between FSDS (k = 8) and each other feature set, with the p-value of the corrected resampled t-test in parentheses. Positive values favour FSDS.

FSDS compared with	Logistic regression	SVM (RBF)	Balanced random forest	SMOTE + KNN
All 17 features	+0.21 (0.426)	+3.97 (0.002)	+0.05 (0.921)	+6.03 (0.005)
Crisp FSDS	+0.01 (0.826)	+0.10 (0.796)	−0.21 (0.584)	+0.21 (0.748)
Age-adjusted FSDS	+0.10 (0.663)	+1.50 (0.162)	+0.06 (0.913)	+2.25 (0.268)
ANOVA	−0.08 (0.423)	−0.10 (0.893)	+1.44 (0.055)	+0.21 (0.912)
Chi-square	−0.05 (0.713)	+0.25 (0.755)	+1.36 (0.047)	+0.18 (0.913)
Mutual information	+0.03 (0.766)	+0.28 (0.647)	+0.16 (0.720)	+0.42 (0.726)

5.4. Number of selected features and stability

Figure 4 shows the logistic-regression AUC against the number of selected features. Age alone gave 0.834 and no feature set exceeded 0.842, so the other 16 features add less than one AUC point. The age-adjusted ranking reached its plateau with two features (age and glucose, AUC 0.840); unadjusted FSDS and ANOVA needed four, because they placed “ever married” second. Chi-square was poor for $k \leq 2$ (AUC 0.566 and 0.641), because it ranks heart disease and hypertension above age. Beyond eight features all methods converged, so the choice k = 8 is not critical.

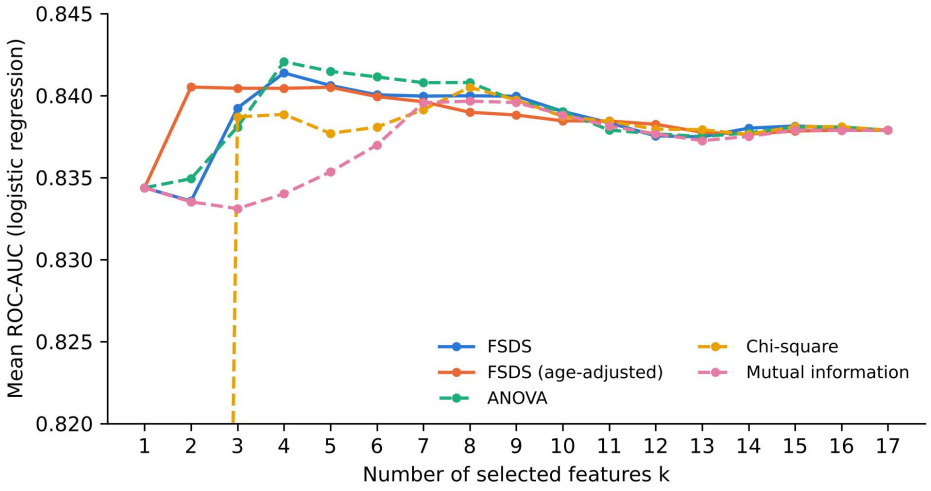


Figure 4. Logistic-regression ROC-AUC (mean over 50 folds) against the number of selected features. Chi-square at k = 1 and 2 (AUC 0.566 and 0.641) lies below the plotted range.

Across the 50 training folds, unadjusted FSDS always selected age, ever married, glucose and hypertension. The age-adjusted top five was exactly {age, glucose, hypertension, heart disease, BMI} in 33 of the 50 folds; no other method selected this set in any fold. On small subsamples (Table 4), however, FSDS was less stable than ANOVA at every sample size; at $k = 5$ and $n = 500$, for example, the Kuncheva index was 0.459 for FSDS against 0.562 for ANOVA, 0.526 for chi-square and 0.478 for mutual information. The reason is visible in Table 1: below the first four features, six FSDS scores lie between 0.21 and 0.28, so small sampling changes reorder them. Fuzzy and crisp FSDS differed by at most 0.04. The age-adjusted score was the least stable (0.255 at $k = 5$, $n = 500$), because a subsample of 500 contains about 24 strokes, of which only a handful fall in the younger strata.

Table 4. Selection stability (mean pairwise Kuncheva index over 50 stratified subsamples of size n) for k = 5 and k = 8.

Method	k = 5, n = 250	500	1000	2000	k = 8, n = 250	500	1000	2000
FSDS	0.277	0.459	0.578	0.664	0.205	0.474	0.593	0.710
Crisp FSDS	0.306	0.459	0.564	0.646	0.242	0.475	0.604	0.725
Age-adjusted FSDS	0.229	0.255	0.340	0.434	0.197	0.238	0.306	0.390
ANOVA	0.321	0.562	0.679	0.924	0.248	0.582	0.688	0.782
Chi-square	0.264	0.526	0.629	0.771	0.303	0.539	0.684	0.802
Mutual information	0.330	0.478	0.618	0.657	0.249	0.551	0.691	0.760

5.5. Sensitivity analyses

Under four age stratifications (the three strata above; under 45, 45–64 and 65 or older; quintiles; ten-year bands), glucose and hypertension were always in the age-adjusted top five and heart disease in the top eight. BMI was in the top eight under three of the four; it dropped out with ten-year bands, where several strata contain very few strokes. “Children” never entered the adjusted top eight, and “ever married” did so once, in eighth place. With quartile breakpoints for glucose and BMI instead of the clinical ones, the unadjusted ranking was almost unchanged (Spearman 0.99) and the top eight contained the same features; clinical breakpoints were kept so that the terms have their usual clinical meaning.

6. Discussion

What the representation adds. On this dataset FSDS did not select better features for prediction than ANOVA, chi-square or mutual information; all reached the same AUC, and age alone provided most of the signal. Its value lies elsewhere. Proposition 1 shows that the score is a total variation distance between class-wise term distributions, which gives continuous and binary features a common, bounded scale and makes every score decompose into term-level statements. The profile "stroke patients are high-glucose to degree 0.41 against 0.21" is a quantitative version of the way clinicians describe risk, and it is obtained without training a model. SHAP and LIME [5,6] explain a particular fitted model, whereas the fuzzy profile describes the data; the two are complementary.

Confounding by age. The most important result is the contrast between the crude and the adjusted rankings. All four standard methods placed "ever married" in their top five, and a reader could conclude that marital status is a stroke risk factor. Stratifying by age and permuting within strata removed it and the other age proxies and left exactly the four established modifiable risk factors. The stratified test is not specific to the fuzzy representation and could be applied to any filter score, but it combines naturally with FSDS because the within-stratum scores keep their linguistic meaning. The results also separate prediction from explanation: age proxies help prediction (removing them weakened the SVM and KNN), so crude scores are adequate for building a classifier, while the adjusted score should be used to identify and describe risk factors.

Evaluation under imbalance. Earlier analyses of this dataset reported accuracies above 90% [3,4]; here the best AUC was 0.84 and the best balanced accuracy 76.6%. Because protocols, data versions and metrics differ, this does not show that either set of results is wrong, but it shows how strongly results on this dataset depend on the protocol. With 95.1% negatives, accuracy rewards ignoring strokes, and oversampling or selecting features before splitting lets test information leak into training [17,18]. Under a leakage-free protocol, the relative ranking of methods, not the absolute performance, is the meaningful result.

When does fuzziness matter? Fuzzy and crisp FSDS gave the same rankings (Spearman 0.99), performance and stability. Proposition 1(c) explains why: 14 of the 17 features are binary, for which the two scores coincide, and the three continuous features differ so strongly between classes that smoothing the term boundaries changes little at this sample size. Datasets with many continuous measurements near clinical thresholds are where graded memberships should make a measurable difference.

7. Limitations

The data come from one public dataset whose original source is not disclosed [16], and no external cohort was available. The glucose variable is an average glucose level rather than a fasting measurement, so the fasting thresholds give the terms an approximate clinical meaning. Age adjustment by strata is coarse: residual confounding within strata is possible, other confounders such as sex were not adjusted for, and the adjusted score is unstable in small samples. The scores are univariate and do not model interactions or redundancy. FSDS showed no predictive advantage over standard filters and was less stable than ANOVA. Classifiers were not tuned, because the comparison concerned feature selection.

8. Conclusion

A fuzzy soft set discrimination score for ranking stroke risk features was examined on a severely imbalanced public dataset. The score equals twice the total variation distance between the stroke and non-stroke distributions over linguistic terms, which justifies ranking continuous and binary features together and gives each score a readable profile. Without adjustment the score, like the ANOVA, chi-square and mutual-information filters, ranked age proxies such as marital status among the top features; with age stratification and a stratified permutation test it recovered exactly the four established modifiable risk factors (glucose, hypertension, heart disease and BMI) beside age. In a leakage-free evaluation its selections predicted stroke as well as all 17 features and as well as the standard filters, and they improved the SVM and KNN substantially, but they were not more accurate than the filters and were less stable in small samples. Future work should test the adjusted score on independent cohorts with more continuous clinical measurements and adjust for several confounders at once.

Declarations

Data availability. The data are publicly available from Kaggle [16]. The analysis code is available from the author on reasonable request.

Funding. This research did not receive any specific grant from funding agencies in the public, commercial or not-for-profit sectors.

Competing interests. The author declares no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Ethics statement. The study used only a public, de-identified dataset; no ethical approval or informed consent was required.

Author contribution. Sisir Kumar Rajbongshi: conceptualization, methodology, software, formal analysis, writing – original draft, writing – review and editing.

Use of generative AI. During the preparation of this work the author used Claude (Anthropic) to assist with code, data analysis and language editing. After using this tool, the author reviewed and edited the content as needed and takes full responsibility for the content of the publication.

References

[1] GBD 2019 Stroke Collaborators. (2021). Global, regional, and national burden of stroke and its risk factors, 1990–2019: A systematic analysis for the Global Burden of Disease Study 2019. The Lancet Neurology, 20(10), 795–820. https://doi.org/10.1016/S1474-4422(21)00252-0

[2] Feigin, V. L., Brainin, M., Norrving, B., Martins, S., Sacco, R. L., Hacke, W., et al. (2022). World Stroke Organization (WSO): Global stroke fact sheet 2022. International Journal of Stroke, 17(1), 18–29. https://doi.org/10.1177/17474930211065917

[3] Dubey, Y., Tarte, Y., Talatule, N., Damahe, K., Palsodkar, P., & Fulzele, P. (2024). Explainable and interpretable model for the early detection of brain stroke using optimized boosting algorithms. Diagnostics, 14(22), 2514. https://doi.org/10.3390/diagnostics14222514

[4] Dritsas, E., & Trigka, M. (2022). Stroke risk prediction with machine learning techniques. Sensors, 22(13), 4670. https://doi.org/10.3390/s22134670

[5] Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In Advances in Neural Information Processing Systems 30 (pp. 4765–4774).

[6] Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?": Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 1135–1144). https://doi.org/10.1145/2939672.2939778

[7] Zadeh, L. A. (1965). Fuzzy sets. Information and Control, 8(3), 338–353. https://doi.org/10.1016/S0019-9958(65)90241-X

[8] Molodtsov, D. (1999). Soft set theory—First results. Computers & Mathematics with Applications, 37(4–5), 19–31. https://doi.org/10.1016/S0898-1221(99)00056-5

[9] Maji, P. K., Biswas, R., & Roy, A. R. (2001). Fuzzy soft sets. Journal of Fuzzy Mathematics, 9(3), 589–602.

[10] Sanchez, E. (1979). Inverse of fuzzy relations. Application to possibility distributions and medical diagnosis. Fuzzy Sets and Systems, 2(1), 75–86. https://doi.org/10.1016/0165-0114(79)90017-4

[11] Çelik, Y., & Yamak, S. (2013). Fuzzy soft set theory applied to medical diagnosis using fuzzy arithmetic operations. Journal of Inequalities and Applications, 2013, 82. https://doi.org/10.1186/1029-242X-2013-82

[12] Singh, R., & Bhardwaj, N. (2025). Fuzzy soft set theory applications in medical diagnosis: A comprehensive review and the roadmap for future studies. New Mathematics and Natural Computation, 21(2), 597–619. https://doi.org/10.1142/S1793005725500279

[13] Jensen, R., & Shen, Q. (2004). Semantics-preserving dimensionality reduction: Rough and fuzzy-rough-based approaches. IEEE Transactions on Knowledge and Data Engineering, 16(12), 1457–1471. https://doi.org/10.1109/TKDE.2004.96

[14] Zhang, X., Mei, C., Li, J., Yang, Y., & Qian, T. (2023). Instance and feature selection using fuzzy rough sets: A bi-selection approach for data reduction. IEEE Transactions on Fuzzy Systems, 31(6), 1981–1994.

[15] Ben Kahla, M., Kanzari, D., Ben Amor, S., Ayachi Ghannouchi, S., & Martinho, R. (2025). Enhanced fuzzy score-based decision support system for early stroke prediction. ACM Transactions on Computing for Healthcare, 6(1), 1–23. https://doi.org/10.1145/3703461

[16] fedesoriano. (2021). Stroke prediction dataset [Data set]. Kaggle. https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset

[17] Ambroise, C., & McLachlan, G. J. (2002). Selection bias in gene extraction on the basis of microarray gene-expression data. Proceedings of the National Academy of Sciences, 99(10), 6562–6566. https://doi.org/10.1073/pnas.102102699

[18] Kaufman, S., Rosset, S., Perlich, C., & Stitelman, O. (2012). Leakage in data mining: Formulation, detection, and avoidance. ACM Transactions on Knowledge Discovery from Data, 6(4), 15. https://doi.org/10.1145/2382577.2382579

[19] Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. Journal of Artificial Intelligence Research, 16, 321–357. https://doi.org/10.1613/jair.953

[20] Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. Journal of Machine Learning Research, 3, 1157–1182.

[21] Chandrashekar, G., & Sahin, F. (2014). A survey on feature selection methods. Computers & Electrical Engineering, 40(1), 16–28. https://doi.org/10.1016/j.compeleceng.2013.11.024

[22] Kraskov, A., Stögbauer, H., & Grassberger, P. (2004). Estimating mutual information. Physical Review E, 69(6), 066138. https://doi.org/10.1103/PhysRevE.69.066138

[23] Kuncheva, L. I. (2007). A stability index for feature selection. In Proceedings of the 25th IASTED International Multi-Conference: Artificial Intelligence and Applications (pp. 390–395).

[24] Nogueira, S., Sechidis, K., & Brown, G. (2018). On the stability of feature selection algorithms. Journal of Machine Learning Research, 18(174), 1–54.

[25] Rajbongshi, S. K., & Talukdar, D. (2013). An introduction to a family of fuzzy subsets forming Boolean algebra. International Journal of Computer Applications, 68(24).

[26] Rajbongshi, S. K., & Talukdar, D. (2013). Some aspects of fuzzy Boolean algebra formed by fuzzy subsets. International Journal of Advanced Research in Computer Science and Software Engineering, 3(7).

[27] Rajbongshi, S. K., & Gogoi, P. (2023). On some theoretic aspects of fuzzy subsets. International Journal for Research in Applied Science & Engineering Technology, 11(2).

[28] World Health Organization. (2000). Obesity: Preventing and managing the global epidemic. Report of a WHO consultation (WHO Technical Report Series 894). World Health Organization.

[29] American Diabetes Association Professional Practice Committee. (2024). 2. Diagnosis and classification of diabetes: Standards of care in diabetes—2024. Diabetes Care, 47(Suppl. 1), S20–S42. https://doi.org/10.2337/dc24-S002

[30] Mantel, N., & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. Journal of the National Cancer Institute, 22(4), 719–748.

[31] Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. Journal of the Royal Statistical Society: Series B, 57(1), 289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

[32] He, H., & Garcia, E. A. (2009). Learning from imbalanced data. IEEE Transactions on Knowledge and Data Engineering, 21(9), 1263–1284. https://doi.org/10.1109/TKDE.2008.239

[33] Chen, C., Liaw, A., & Breiman, L. (2004). Using random forest to learn imbalanced data (Technical Report 666). Department of Statistics, University of California, Berkeley.

[34] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12, 2825–2830.

[35] Lemaître, G., Nogueira, F., & Aridas, C. K. (2017). Imbalanced-learn: A Python toolbox to tackle the curse of imbalanced datasets in machine learning. Journal of Machine Learning Research, 18(17), 1–5.

[36] Saito, T., & Rehmsmeier, M. (2015). The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PLOS ONE, 10(3), e0118432. https://doi.org/10.1371/journal.pone.0118432

[37] Nadeau, C., & Bengio, Y. (2003). Inference for the generalization error. Machine Learning, 52(3), 239–281. https://doi.org/10.1023/A:1024068626366

[38] Bouckaert, R. R., & Frank, E. (2004). Evaluating the replicability of significance tests for comparing learning algorithms. In Advances in Knowledge Discovery and Data Mining (PAKDD 2004), Lecture Notes in Computer Science, 3056 (pp. 3–12). Springer. https://doi.org/10.1007/978-3-540-24775-3_3