Supplementary Table S1A. Lexical consistency of negation cues following NLP preprocessing

Cue	Extracted clinical text	Final text
sem	121	121
não / nao	141	141
ausência / ausencia	20	20
Total	282	282

Note. Counts were identical within each of the 161 reports, not merely in aggregate. Some occurrences correspond to non-pathological or technical expressions; therefore, these counts should not be interpreted as 282 clinically validated negations.

Supplementary Table S1B. Illustrative preservation of clinically relevant negated expressions

Group	Extracted clinical text	Final representation
BD	Não há desvio das estruturas da linha média, nem sinais de hidrocefalia.	nao ha desvio estruturas linha media nem sinais hidrocefalia
BD	Sem sinais de hidrocefalia a montante.	sem sinais hidrocefalia montante
BD	Não existe herniação das amígdalas cerebelosas.	nao existe herniacao amigdalas cerebelosas
BD	Não há sinais de conflito espaço no plano do buraco magno.	nao ha sinais conflito espaco plano buraco magno
BD	Ausência de herniações encefálicas.	ausencia herniacoes encefalicas
NBD	Sem hidrocefalia.	sem hidrocefalia
NBD	Não existem sinais de hidrocefalia ativa.	nao existem sinais hidrocefalia ativa
NBD	Não se observam desvios das estruturas da linha mediana.	nao se observam desvios estruturas linha mediana
NBD	Sem sinais de transformação hemorrágica.	sem sinais transformacao hemorragica
NBD	Sem evidência de lesão hemorrágica aguda, parenquimatosa ou extra-axial.	sem evidencia lesao hemorragica aguda parenquimatosa ou extra axial

Note. Ten illustrative expressions (five BD, five NBD) were checked by comparing extracted clinical text with the finalized normalized representation. This targeted qualitative comparison is not an independent clinician-adjudicated assessment of semantic accuracy across all reports. Original excerpts are anonymized.

Supplementary Table S2. NLP, model configuration, internal validation, and software environment

Parameter	Final analytical configuration
A. Automated NLP and TF-IDF feature extraction	**A. Automated NLP and TF-IDF feature extraction**
Source and preprocessing	161 index brain CT reports (94 BD, 67 NBD); deterministic final extraction of clinical report sections, removal of administrative material, text normalization, and preservation of natural Portuguese negation cues.
Feature source	Final processed text (nlp_text); no manually curated radiological vocabulary/whitelist.
Vectorizer and analyzer	scikit-learn TfidfVectorizer; analyzer=word; ngram_range=(1, 3).
Tokenization	token_pattern=r"(?u)\b\w+\b" (single-character tokens allowed); tokenizer=None; preprocessor=None.
Case and accent settings	lowercase=False (final input already normalized); strip_accents=unicode.
Vocabulary and frequency limits	vocabulary=None; min_df=2; max_df=1.0; max_features=5000; stop_words=None.
TF-IDF weighting	sublinear_tf=True; use_idf=True; smooth_idf=True; norm=l2; binary=False.
Other TF-IDF settings	dtype=float64; input=content; encoding=utf-8; decode_error=strict.
Observed feature counts	Median 3,267 per outer fold (range 3,077–3,441); 4,000 in a separate full-cohort refit used only for descriptive reporting.
B. Machine-learning models (fixed hyperparameters)	**B. Machine-learning models (fixed hyperparameters)**
Logistic regression (LR)	C=1.0; penalty=l2 (default); solver=liblinear; max_iter=1000; class_weight=balanced; random_state=42.
Random forest (RF)	n_estimators=500; min_samples_leaf=3; class_weight=balanced_subsample; n_jobs=1; random_state=42; other parameters at package defaults.
Support vector machine (SVM)	C=1.0; kernel=rbf; gamma=scale; probability=False; class_weight=balanced; random_state=42; probabilities obtained by separate sigmoid calibration.
XGBoost	n_estimators=150; max_depth=2; learning_rate=0.05; subsample=0.8; colsample_bytree=0.8; reg_lambda=2.0; reg_alpha=0.0; objective=binary:logistic; eval_metric=logloss; n_jobs=1; random_state=42.
Model selection and imbalance	No data-driven hyperparameter search. Class weighting: balanced (LR/SVM), balanced_subsample (RF); no additional explicit class weighting in XGBoost.
C. Internal validation and threshold selection	**C. Internal validation and threshold selection**
Outer validation	RepeatedStratifiedKFold: five folds × 20 repetitions (100 dependent validation folds); random_state=42; one held-out prediction per patient per repetition.
Training-only fitting	TF-IDF vocabulary/IDF estimation and model training restricted to the appropriate outer training partitions; held-out patients excluded from fitting.
Probability calibration	CalibratedClassifierCV; method=sigmoid; cv=3; ensemble=False; calibration fitted within each outer training partition.
Threshold selection	Separate inner three-fold stratified cross-validation on the outer training patients; calibrated out-of-fold probabilities used to select the specificity-maximizing threshold subject to training sensitivity ≥0.95.
Held-out assessment	Selected threshold applied unchanged to outer test-fold probabilities; median fold threshold reported descriptively only, not applied globally to primary predictions.
Uncertainty estimation	2,000 patient-cluster bootstrap resamples preserving all 20 repeated out-of-fold predictions per patient; conditional on fitted CV models.
Sensitivity analyses	Alternative training sensitivity target 1.00; documentary feature ablation (final_b); temporally restricted/stratified analyses and cross-period transfer; metadata-based negative controls.
D. Verified computational environment	**D. Verified computational environment**
Python and core ML	Python 3.12.8; scikit-learn 1.9.1; XGBoost 3.0.2; SHAP 0.48.0.
Data and numerical libraries	NumPy 2.4.6; pandas 3.1.0; SciPy 1.18.1; openpyxl 3.1.5.
Visualization and PDF library	Matplotlib 3.11.2; PyMuPDF 1.26.3 (version recorded in the final ML environment; original final extraction environment should be verified separately).

Note. BD, brain death; NBD, non-brain death; IDF, inverse document frequency; OOF, out-of-fold; TF-IDF, term frequency–inverse document frequency. Hyperparameters not explicitly specified used package defaults. The 4,000-feature full-data refit was descriptive and did not enter held-out evaluation. Software versions refer to the final modelling environment; the original PDF-extraction environment should be independently verified.

Supplementary Table S3. Internal calibration of machine-learning models

Model	Calibration intercept[1]	Calibration slope[2]	Brier score[3]
Logistic regression	0.026	0.951	0.163
Random forest	0.062	1.148	0.171
Support vector machine	0.041	0.983	0.167
XGBoost	0.021	1.475	0.180

Calibration estimates used patient-level mean out-of-fold probabilities (n = 161; 94 BD and 67 NBD). Each patient contributed one average of 20 held-out predicted probabilities, one from each cross-validation repetition.

Notes. For each patient, 20 held-out predictions were obtained across repeated stratified five-fold cross-validation; the probabilities were averaged before calculating these calibration summaries. Each patient therefore contributed one averaged prediction to the analysis.

[1] Calibration-in-the-large: intercept from logistic recalibration with the logit of the predicted probability as an offset (slope fixed at 1).

[2] Calibration slope: coefficient of the predicted log-odds from logistic recalibration with a freely estimated intercept. The reported intercept and slope are estimated using different calibration specifications and should not be interpreted as coefficients from the same fitted regression.

[3] Brier scores were calculated from patient-level mean out-of-fold probabilities; these are distinct from Brier scores averaged over individual cross-validation repetitions.

The study cohort was retrospectively selected and is not representative of the prevalence of brain death in consecutive clinical admissions. Accordingly, these calibration estimates should not be extrapolated to the intended screening population without external validation.

Supplementary Table S4. Exploratory temporal analyses, documentary controls, and additional sensitivity analyses

A. Distribution of analytic cases by CT-report period

Period	BD, n	NBD, n	Total, n
2011–2012	27	16	43
2013–2021	49	0	49
2022–2024	18	51	69
Total	94	67	161

B. Exploratory temporal performance of logistic regression (ROC-AUC)

Analysis	Training period	Evaluation period	ROC-AUC
Period-specific CV evaluation	2011–2012 and 2022–2024	2011–2012	0.642
Period-specific CV evaluation	2011–2012 and 2022–2024	2022–2024	0.713
Cross-period transfer	2011–2012	2022–2024	0.630
Cross-period transfer	2022–2024	2011–2012	0.704

Note. All analyses in Panel B are exploratory. For period-specific cross-validation, the model was trained on both periods (2011–2012 and 2022–2024), using repeated five-fold cross-validation (20 repetitions) stratified jointly by period and outcome; out-of-fold ROC-AUC was then calculated separately within each period. In the cross-period transfer analyses, the model was trained exclusively in one period and evaluated in the other. These estimators are not directly equivalent to the primary full-cohort repeated cross-validation estimate. Small sample sizes and differences in case selection across periods limit conclusions about temporal generalizability.

C. Documentary and temporal metadata negative-control analyses

Model / inputs	Mean fold ROC-AUC	Validation folds
Original NLP (LR), comparator	0.820	100
Year only	0.837	100
Report template only	0.770	100
Year + report template	0.872	100
Year + template + report length	0.873	100

Note. Panel C reports mean ROC-AUC across 100 dependent folds from exploratory metadata negative-control analyses. In particular, the 0.820 NLP comparator here is a mean fold-level ROC-AUC and must not be confused with the primary patient-level repeated-CV ROC-AUC of approximately 0.812. Metadata discrimination may reflect temporal selection and report-format differences; these analyses do not establish independent temporal transportability or causal explanations.

D. Additional sensitivity analyses (logistic regression; full analytic cohort)

Analysis	Training sensitivity target	Patients, n	ROC-AUC	Sensitivity	Specificity
Primary analysis	95%	161	0.812	94.1%	42.6%
Higher-sensitivity threshold	100%	161	0.812	98.5%	19.3%
Selected documentary feature ablation	95%	161	0.804	94.6%	42.8%

Note. All analyses used the same 161 patients (94 BD and 67 NBD) and repeated stratified five-fold cross-validation (20 repetitions). The 100% target was applied only to training-stage threshold selection; held-out sensitivity remained below 100%. The documentary ablation excluded selected generic/report-related terms and TF-IDF n-grams containing them while preserving negation cues. Performance estimates are mean values across the 20 held-out cross-validation repetitions. These exploratory analyses do not eliminate the possibility of temporal or documentary confounding.