Clinical Neurophysiology 169 (2025) 33–46

Contents lists available at ScienceDirect

Clinical Neurophysiology

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

Metrics for evaluation of automatic epileptogenic zone localization in intracranial electrophysiology

Valentina Hrtonova [a,b,c], Petr Nejedly [a,b], Vojtech Travnicek [b,d], Jan Cimbalnik [d], Barbora Matouskova [a,b,c], Martin Pail [e], Laure Peter-Derex [f,g], Christophe Grova [h], Jean Gotman [i], Josef Halamek [b], Pavel Jurak [b], Milan Brazdil [e,j], Petr Klimes [b,*], Birgit Frauscher [k,l,*]

[a] First Department of Neurology, Faculty of Medicine, Masaryk University, Pekarska 53, 602 00 Brno, Czech Republic
[b] Institute of Scientific Instruments of the CAS, v. v. i., Kralovopolska 147, 612 00 Brno, Czech Republic
[c] Department of Neurology, Duke University School of Medicine, 2424 Erwin Rd, Durham, NC 27705, the United States of America
[d] International Clinical Research Center, St. Anne's University Hospital, Pekarska 53, 602 00 Brno, Czech Republic
[e] Brno Epilepsy Center, Department of Neurology, St. Anne's University Hospital, member of ERN-EpiCARE, Faculty of Medicine, Masaryk University, Pekarska 53, 602 00 Brno, Czech Republic
[f] Center for Sleep Medicine, Lyon University Hospital, Lyon 1 University, 103 Grande Rue de la Croix-Rousse, 69004 Lyon, France
[g] Lyon Neuroscience Research Center, CH Le Vinatier - Batiment 462 - Neurocampus, 95 Bd Pinel, 69500 Lyon, France
[h] Multimodal Functional Imaging Lab, Department of Physics and Concordia School of Health, Concordia University and Biomedical Engineering Department, McGill University, Montreal Neurological Hospital, Concordia University, 7141 Sherbrooke Street West, Montreal, QC H4B 1R6
[i] Montreal Neurological Institute, McGill University, 3801 Rue University, Montreal, QC H3A 2B4, Quebec, Canada
[j] Behavioral and Social Neuroscience Research Group, CEITEC Central European Institute of Technology, Masaryk University, Zerotinovo nám 617/9, 601 77 Brno, Czech Republic
[k] Montreal Neurological Hospital, McGill University, 3801 Rue University, Montreal, QC H3A 2B4, Quebec, Canada
[l] Department of Neurology, Duke University Medical School and Department of Biomedical Engineering, Pratt School of Engineering, 2424 Erwin Road, Durham, NC 27705, the United States of America

ARTICLE INFO

Keywords:
Epilepsy
Epileptogenic zone
Seizure onset zone
Epileptogenic tissue localization
Intracranial electroencephalography
Machine learning
Binary classification
Evaluation metrics
Class imbalance

ABSTRACT

Introduction: Precise localization of the epileptogenic zone is critical for successful epilepsy surgery. However, imbalanced datasets in terms of epileptic vs. normal electrode contacts and a lack of standardized evaluation guidelines hinder the consistent evaluation of automatic machine learning localization models.
Methods: This study addresses these challenges by analyzing class imbalance in clinical datasets and evaluating common assessment metrics. Data from 139 drug-resistant epilepsy patients across two Institutions were analyzed. Metric behaviors were examined using clinical and simulated data.
Results: Complementary use of Area Under the Receiver Operating Characteristic (AUROC) and Area Under the Precision-Recall Curve (AUPRC) provides an optimal evaluation approach. This must be paired with an analysis of class imbalance and its impact due to significant variations found in clinical datasets.
Conclusions: The proposed framework offers a comprehensive and reliable method for evaluating machine learning models in epileptogenic zone localization, improving their precision and clinical relevance.
Significance: Adopting this framework will improve the comparability and multicenter testing of machine learning models in epileptogenic zone localization, enhancing their reliability and ultimately leading to better surgical outcomes for epilepsy patients.

1. Introduction

Epilepsy is one of the most common neurological disorders, affecting more than 70 million people worldwide (Thijs et al. 2019). Approximately 40 % of patients with epilepsy do not respond to anti-seizure medications (Chen et al. 2018). The most effective treatment for these patients is surgery (Jehi 2018). To optimize surgical outcomes in this patient group, precise localization of the epileptogenic zone (EZ)

* Corresponding authors at: Institute of Scientific Instruments of the CAS, v. v. i., Kralovopolska 147, 612 00 Brno, Czech Republic (P. Klimes) and Duke University Medical School, 2424 Erwin Road, Durham, NC 27705, the United States of America (B. Frauscher).
E-mail addresses: petr.klimes@isibrno.cz (P. Klimes), birgit.frauscher@duke.edu (B. Frauscher).

https://doi.org/10.1016/j.clinph.2024.11.007
Accepted 14 November 2024
Available online 19 November 2024

is necessary (Vakharia et al. 2018). Invasive intracranial electroencephalography (EEG) recordings through stereo-EEG (SEEG) are used to delineate the cortical areas where seizures start and rapidly propagate based on visual inspection of the SEEG signals (Frauscher et al. 2024). However, traditional methods relying on visual inspection of recordings by medical professionals are time-consuming and subjective. In recent years, machine learning approaches have emerged as promising tools to provide additional information for precise EZ localization from SEEG recordings, potentially improving accuracy and saving time (Cimbalnik et al. 2019; Bernabei et al. 2022; Gunnarsdottir et al. 2022; Grinenko et al. 2018). The increasing number of such studies creates a need for well-defined evaluation guidelines that consider the specific challenges encountered in this field.

One of the primary challenges encountered in EZ localization is the inherent imbalance within the analyzed data. Intracranial EEG recordings including SEEG often exhibit a scarcity of samples corresponding to the region of interest (EZ), leading to an underrepresentation of EZ electrodes compared to non-EZ electrodes. The class imbalance poses significant issues during both the training and evaluation of machine learning models. In training, the algorithm may become biased towards the majority class, affecting its ability to learn patterns from the minority class adequately. This imbalance also impacts the evaluation phase, where traditional metrics, such as accuracy, may not effectively reflect the model's performance due to the dominance of the majority class, necessitating the use of evaluation metrics suitable for imbalanced domains (He and Garcia 2009). Consequently, addressing the class imbalance becomes crucial to ensuring that machine learning models are trained and evaluated in a manner that accurately captures their performance in identifying the region of interest within SEEG recordings. Additionally, the degree of class imbalance varies among clinical datasets, introducing variability that inevitably impacts the models' outcomes. This variability, often overlooked in studies, challenges the interpretability and comparability of results. To tackle this issue effectively, adopting an evaluation framework becomes imperative. Such a framework should comprehensively assess model performance, prioritize the evaluation of the minority class, and demonstrate robustness to variations in class distribution. By adhering to these criteria, the chosen evaluation framework ensures a more clinically relevant and standardized assessment of model performance across diverse datasets.

Our study has three primary objectives. First, we will demonstrate the prevalence of class imbalance in clinical datasets. Second, we will rigorously analyze commonly used evaluation metrics in the field of EZ localization, with special attention to their sensitivity to class imbalance. Third, we will design a framework that proposes a standard evaluation approach.

To demonstrate the issue of class imbalance and its variations across clinical datasets, we analyzed the class distribution in a clinical dataset of 139 patients gathered from two Institutions, namely St. Anne's University Hospital in Brno (SAUH) and the Montreal Neurological Institute & Hospital (MNI). Our analysis aimed to specifically highlight variations: (i) within patients from the same Institution, where differences may emerge due to diverse pathologies and other factors influencing implantation and resection strategies, (ii) within each Institution across localization target definitions, as a result of a non-standardized approximation of the potential EZ region (Jehi 2018), (iii) within each Institution over time, reflecting how implantation strategies may evolve, and lastly (iv) across different Institutions, where variations may arise from distinct patient demographics, implantation protocols, or methodology for target definition.

Subsequently, to comprehensively investigate the effect of the inherent class imbalance on model results, we analyzed how changes in class distributions affect commonly used evaluation metrics, namely accuracy, Area Under the Receiver Operating Characteristic (AUROC), Area Under the Precision-Recall Curve (AUPRC), and F1-score. This analysis is performed on both outputs of a simulated binary classifier, as well as real outputs of logistic regression models trained on clinical data, serving as example models for EZ localization. In the analysis of clinical data, we highlight the main limitations and advantages associated with accuracy, AUROC, AUPRC, and F1-score through an examination of four patient cases, and we show group analysis results across Institutions and localization targets with different levels of class imbalance.

Based on our findings, we propose a combination of evaluation metrics that comprehensively assess binary classification models for EZ localization. By establishing a more robust and comprehensive evaluation framework, our study aims to standardize the model evaluation process. This standardization aims to ultimately enhance the overall classification performance and reliability of machine learning models in the critical task of automatic EZ localization for patients with drug-resistant epilepsy.

2. Methods

To promote transparency and reproducibility of the methodology, the codes for analysis of simulated and clinical data are available online (https://gitlab.com/bbeer_group/public_codes/metrics-for-evaluation-of-automatic-epileptogenic-zone-localization).

2.1. Patients

The patient cohort for the analysis of class imbalance consisted of all consecutive adult patients with drug-resistant focal epilepsy who underwent SEEG and subsequent resective surgery at SAUH between 2012 and 2022 and the MNI between 2009 and 2019. Patients without information on the seizure-onset zone (SOZ) or the resected contacts were not considered for the study, resulting in a patient cohort of 139 patients (59 from SAUH and 80 from the MNI). The study was approved by the Brno Epilepsy Center – SAUH Research Ethics Committee and MNI Ethics Review Board. All patients granted written informed consent in accordance with the Declaration of Helsinki.

2.2. Localization target definition

In EZ localization studies, class imbalance refers to the ratio of the target SEEG contacts, defined based on the approximation of the EZ, and the non-target SEEG contacts outside the region of interest. In our study, the localization target was identified according to the most common approximations of the EZ as

(i) **SOZ contacts** (marked by visual inspection of SEEG by board-certified epileptologists based on the earliest detectable changes at seizure onset irrespective of the fast activity content (Spanedda, Cendes, Gotman 1997)) (Saboo et al. 2021; Thomas et al. 2023; Conrad et al. 2023),

(ii) **resected contacts** (all contacts removed during surgery identified based on pre and postsurgical MRI) (Karunakaran et al. 2018; Zweiphenning et al. 2022; Shahabi, Nair, Leahy 2023), and

(iii) **resected SOZ contacts** (SOZ contacts removed during surgery) (Klimes et al. 2019; Bernabei et al. 2022).

2.3. Class imbalance analysis

To analyze the level of class imbalance in the data, we defined the term "relative target size" as the percentage of target SEEG contacts from the total number of all SEEG contacts used for analysis (all SEEG contacts with a confirmed location inside the brain). We investigated the distributions of relative target size values (i) within each Institution across individual patients, (ii) within each Institution across the three localization target definitions, (iii) within each Institution over time, and (iv) across the two Institutions. The parts of the class imbalance analysis are illustrated in Fig. 1, and the methodology is described in detail in the Supplementary Material.

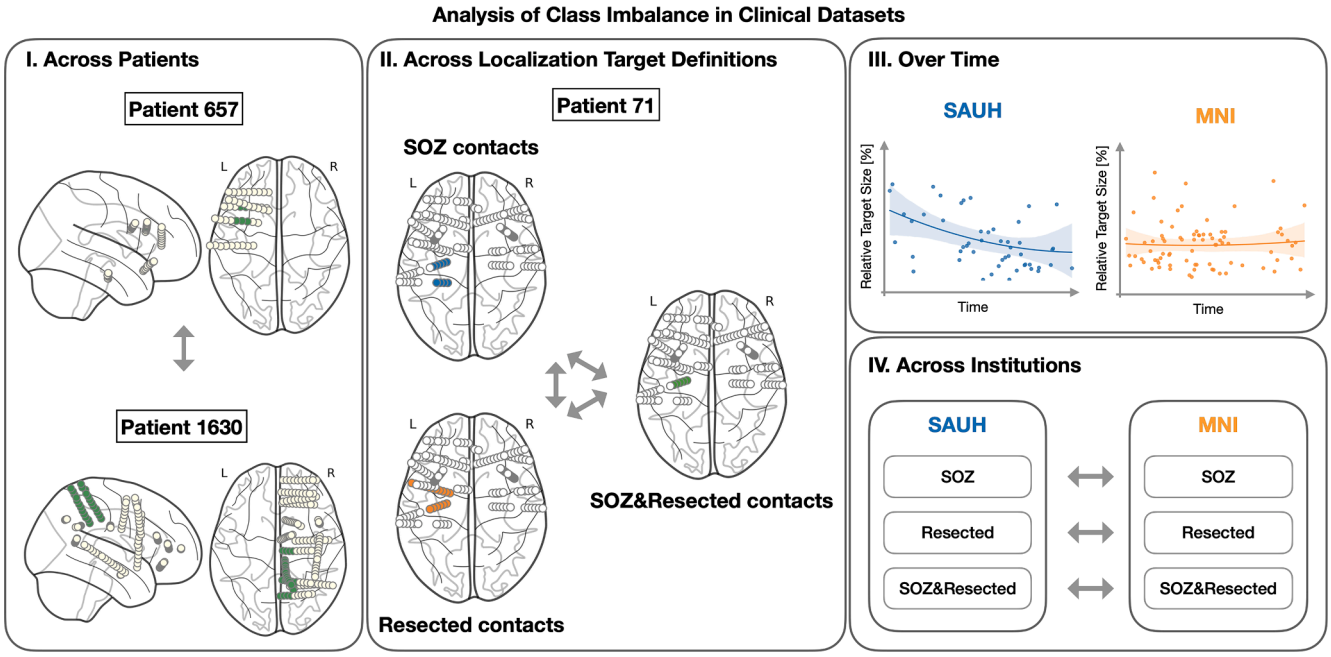

Fig. 1. Class imbalance in clinical datasets was analyzed in four different ways. (I.) Variability among patients: the brain diagrams show SEEG implantations for patients 657 and 1630 and resected contacts in green, demonstrating the differences in the number and distribution of implanted electrode contacts and target contacts. (II.) Variability across localization target definitions: the brain diagrams show the SOZ (blue), Resected (orange), and SOZ&Resected (green) contacts for patient 71, showing how target definitions can vary in size. (III.) Changes over time: the plots illustrate how changes in relative target size at SAUH and the MNI were analyzed over time (approximately 10 years). (IV.) Variability across Institutions: the panel shows how the class imbalance at SAUH and the MNI were compared for different target definitions.

2.4. Metrics for evaluation of epileptogenic zone localization

2.4.1. Metric selection

In the evaluation of models for localizing the EZ in intracranial electrophysiology studies, the following metrics are commonly used:

- **Sensitivity**, also known as "true positive rate" or "recall" (von Ellenrieder et al. 2016; Sumsky and Santaniello (2019); Cimbalnik et al. 2018)
- **Specificity**, also known as "true negative rate" or "selectivity" (Klimes et al. 2019; Lai et al. 2020; Lundstrom, Brinkmann, Worrell 2021)
- **Positive Predictive Value** (PPV), also known as "precision" (Elahian et al. 2017; Gunnarsdottir et al. 2022; Jiang et al. 2022)
- **Negative Predictive Value** (NPV) (Lundstrom, Brinkmann, Worrell 2021; Mooij et al. 2016; Wang et al. 2022)
- **Accuracy** (Gunnarsdottir et al. 2022; Gireesh et al. 2023; Jose et al. 2023)
- **AUROC** (Akter et al. 2020; Conrad et al. 2022; Wang et al. 2023)
- **AUPRC** (Chybowski et al. 2024; Bernabei et al. 2022)
- **F1-score**, also known as "F1-measure" (Gireesh et al. 2023; Modur and Miocinovic 2015; Varotto et al. 2021)

Their definitions are provided in the Supplementary Material. While each of the simple metrics (specificity, sensitivity, PPV, and NPV) plays a crucial role in model evaluation, individually they offer only a limited perspective on the model performance (Branco, Torgo, Ribeiro 2015). Given these limitations, our analysis centered primarily on metrics that evaluate the trade-off between the simple metrics, providing a more comprehensive evaluation. Specifically, we focused on **accuracy, AUROC**, **AUPRC**, and **F1-score. Accuracy**, despite being proven to be unsuitable for imbalanced data (Provost, Fawcett, Kohavi 1998 Branco, Torgo, Ribeiro 2015), is still one of the most frequently used metrics. **AUROC** is the most widely adopted metric in imbalanced SEEG studies, despite limitations caused by equal evaluation of both classes (Davis and Goadrich 2006; Webb and Ting 2005). On the other hand, **AUPRC** and **F1-score** are metrics that have been widely recommended for dealing with imbalanced data within the machine learning community (Davis and Goadrich 2006), yet these metrics are not commonly used in EZ localization studies, suggesting a gap between general machine learning practices and this specific field.

In our study, we highlighted the importance of accuracy, AUROC, AUPRC, and F1-score in offering a comprehensive view of model performance and their limitations in dealing with imbalanced data, which is a common occurrence in clinical datasets. While acknowledging their limitations, we have chosen them for their relevance and use in epilepsy diagnostics.

2.4.2. Metric chance levels

Comparing a model's performance to chance levels is crucial as it serves as a benchmark for assessing whether the model genuinely learns from data or merely guesses. The definition of chance level performance varies depending on the metric.

The chance level for accuracy, defined by the always negative classifier, depends on the class distribution. In a binary classification task, a random classifier predicting all samples as non-target achieves accuracy corresponding to the proportion of negative samples in the dataset.

The chance level for AUROC is 0.5 because it represents a scenario where the true positive rate equals the false positive rate at all thresholds, producing a diagonal line across the ROC space that covers half the unit square (Fawcett 2006).

For the precision-recall metrics, the chance level is usually defined by the always positive classifier, and it also depends on class distribution. Specifically, for AUPRC, the chance classifier's performance is equivalent to the precision (p) achieved when classifying all samples as positive. This precision is directly tied to the proportion of positive samples in the dataset, representing the relative target size. When calculating the chance level for the F1-score, recall is set to 1, indicating perfect recall, and the precision is set to the chance level precision (p). The chance level F1-score is then calculated using the F1-score formula

as $(2 \times p)/(p + 1)$.

2.5. Metric analysis using simulated data

The Monte Carlo simulation was used to analyze how various evaluation metrics depend on the level of class imbalance of data in a binary classification task, such as the classification of EZ vs. non-EZ contacts for EZ localization. Here, the class imbalance is given by the relative size of the localization target, which is defined based on one approximation of the EZ, for example, as either the "SOZ", the "Resected, or the "SOZ&Resected" contacts as defined in Section 2.2. The Monte Carlo method was used to approximate the results of a binary classification model under a specified level of class imbalance and fixed classification performance. By performing the simulation for a range of class imbalance levels and classifier parameters, we analyzed how the classification results change depending on those variables. The simulation process is illustrated in Fig. 2 and described in detail in the Supplementary Material.

2.6. Metric analysis using clinical data

To analyze metric properties on clinical data, three logistic regression EZ localization models were trained and evaluated, each corresponding to a specific target definition ("SOZ", "Resected", and "SOZ&Resected"). The models were used to classify electrode contacts as either target (pathologic) or non-target (normal) based on the rate of interictal epileptiform discharges (IEDs) detected in the SEEG signals. The patient selection process for the localization cohort, data preprocessing, and the logistic regression models are described in detail in the Supplementary Material.

Trained models were evaluated separately for each patient by metrics described in the Methods section. For group analysis, the results across patient groups were compared using the Hanley-McNeil and randomization tests, which are suitable for the comparison of results across groups with different levels of class imbalance.

To confirm the superiority of higher metric scores over lower scores across all statistical tests, we used a one-tailed approach designed specifically to assess whether the higher score was statistically greater than the lower score. A Hanley-McNeil test (Hanley and McNeil 1982) was used to compare AUROC values derived from independent ROC curves ($\alpha = 0.05$, one-tailed). Bonferroni correction was applied to maintain an overall significance level of $\alpha = 0.05$ across multiple comparisons, yielding a corrected significance level of 0.017.

To assess the statistical significance of differences in accuracy, AUPRC, and F1-score values, we employed a randomization test (Smucker, Allan, Carterette 2007) with Bonferroni correction ($\alpha = 0.05$, one-tailed). The test involves shuffling ground truth labels and recomputing metric values in numerous permutations. By comparing the observed differences in metric values between Model A and Model B with the distribution of differences from shuffled permutations, we obtained a p-value for each of the metrics. We performed 1000 permutations, ensuring both statistical robustness and computational efficiency.

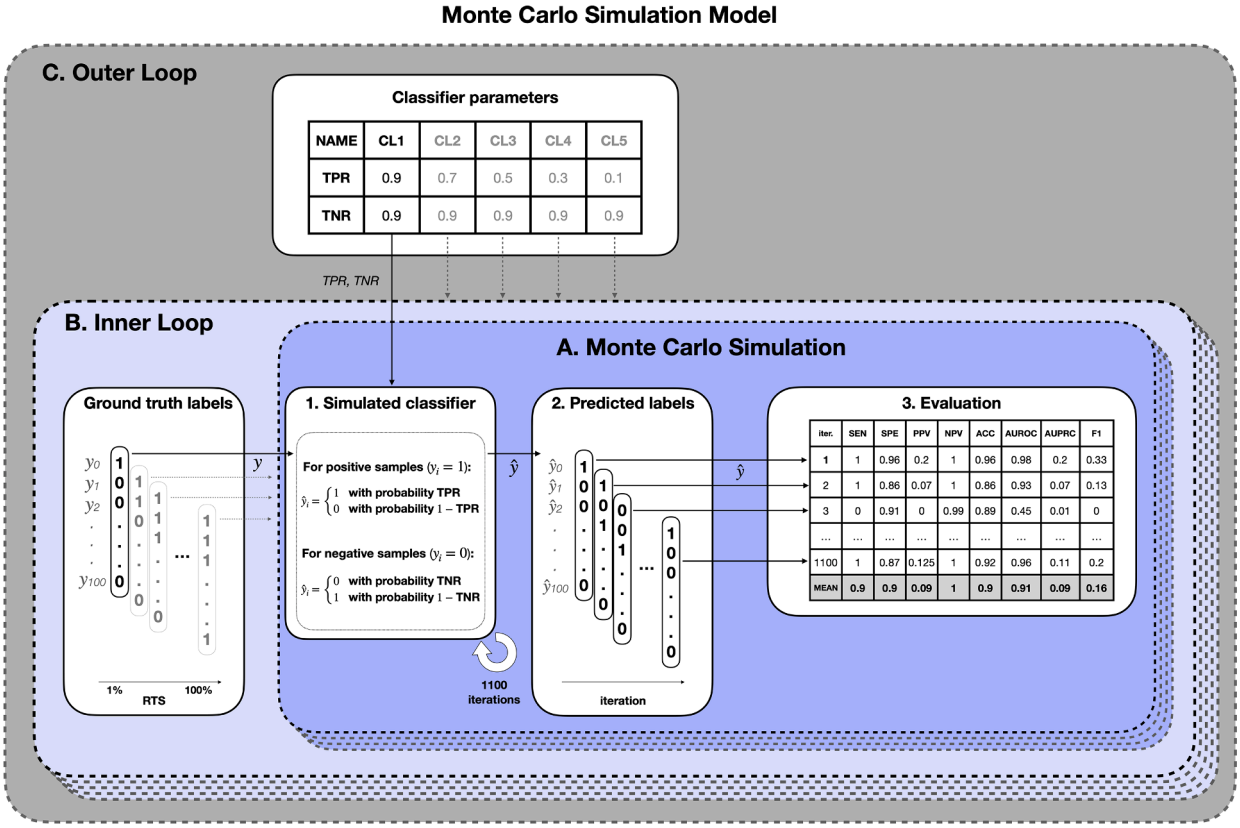

Fig. 2. A. Classification results were simulated using the Monte Carlo method: 1) In a single iteration of the Monte Carlo simulation, a vector of predicted labels $\hat{y}$ was generated based on the vector of ground truth labels y and the model's classification accuracy (TPR, TNR). Samples in the y and $\hat{y}$ vectors represent SEEG contacts, each with a ground truth label y_i (1 for target and 0 for non-target) and a label predicted by the classification model $\hat{y}_i$ (1 for target and 0 for non-target). Size of both vectors was 100 samples. 2) The simulation was repeated 1100 times to generate 1100 samples of the predicted label vector $\hat{y}$. 3) Classification results were evaluated for each iteration, illustrated by multiple dashed lines, and averaged to obtain the most probable performance of the classifier under given conditions. **B.** In an inner loop, the Monte Carlo simulation (block A) was performed 100 times, each time for a ground truth label vector y with a different relative target size (RTS) (1 to 100%). **C.** In an outer loop, the inner loop (block B) was repeated for 5 classifiers, illustrated by multiple dashed lines, with classification performance defined by true positive rate (TPR) and true negative rate (TNR).

3. Results

3.1. Patients

A cohort of 139 patients, with 59 patients from SAUH and 80 patients from the MNI, was included in the analysis of class imbalance. For additional patient information, please refer to the Supplementary Patient Table.

3.2. Class imbalance analysis

Across both Institutions and all localization target definitions, the localization target constituted a median (interquartile range, IQR) of 8.89 % (14.24 %) of all electrode contacts per patient, with 50 % of values between 3.88 % and 18.12 % of relative target size. The range of relative target size values between 4 and 18 % will further be used as a range of interest in Monte Carlo simulation results. The distribution of relative target size values across patients, localization targets, and Institutions, along with statistical test results, is visualized in Fig. 3.

In the SAUH dataset, 1.12 % to 13.46 % of contacts were marked as SOZ by clinicians, 0.54 % to 36.47 % of contacts were resected during surgery, and 0 % to 12.94 % were resected SOZ contacts. In the MNI dataset, between 2.22 % to as much as 85.92 % of contacts were marked as SOZ, 2.42 % to 66.71 % of contacts were resected during surgery, and 0 % to 42.25 % were resected SOZ contacts. None of the analyzed data distributions were normally distributed. Basic statistics and the normality test results for all distributions are summarized in Supplementary Table S1.

The class imbalance between the target definitions was significantly different for both Institutions, with p-values below 0.0001 (Kruskal-Wallis test). In the SAUH dataset, *post hoc* testing revealed significant variations in class imbalance for all targets. In the MNI dataset, while the "SOZ&Resected" target showed significant differences from the other targets, the variations in class imbalance between "SOZ" and "Resected" were not significant. The complete results of the analysis are shown in Supplementary Table S2.

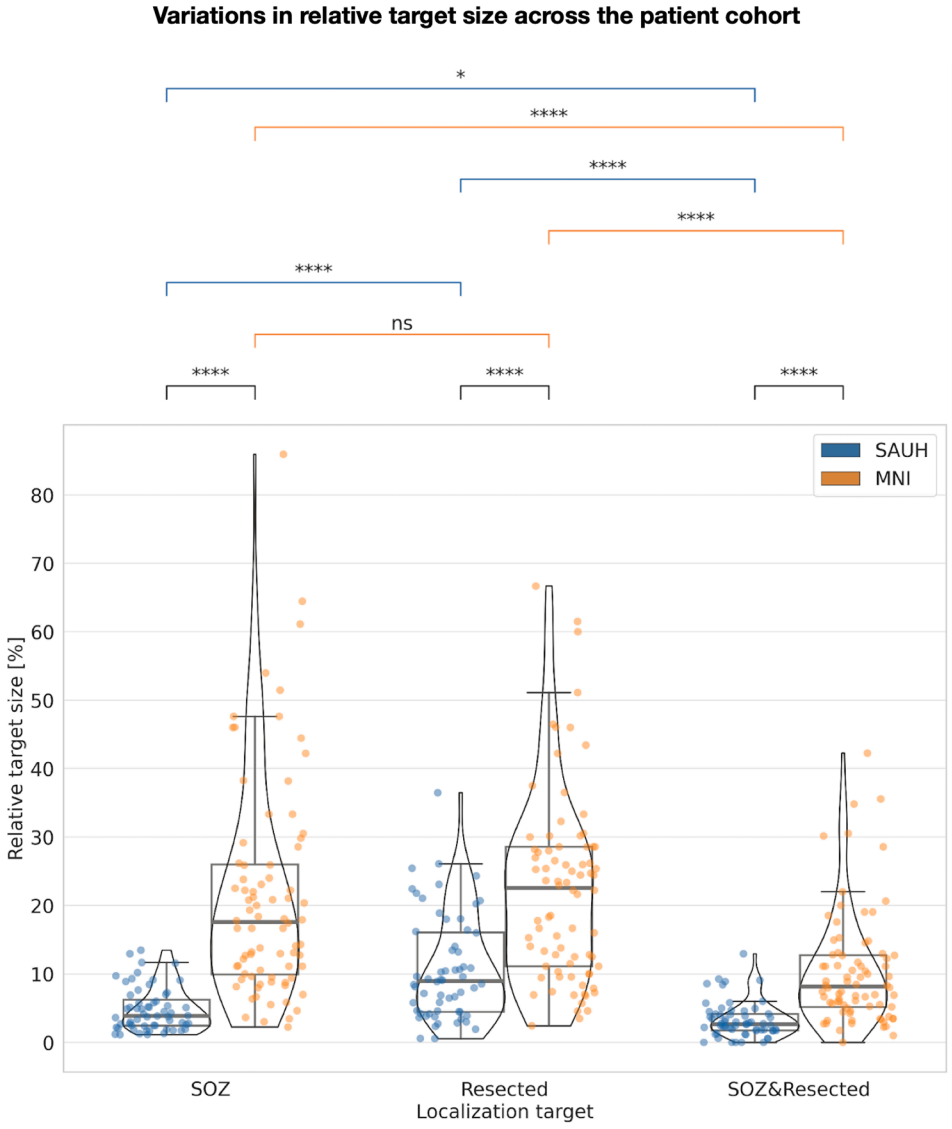

Fig. 3. Distribution of relative target size among patients shows significant differences across localization target definitions and Institutions. For each distribution, the box represents the interquartile range with the bold horizontal line inside the box corresponding to the median. Whiskers extend from the box to the minimum and maximum values within a range specified by 1.5 times the interquartile range. The violin represents the density of data points corresponding to individual patients. The results of the Mann–Whitney rank-sum test are visualized in black and results of *post hoc* testing with the Dunn's test for target pairs for SAUH and the MNI are visualized in blue and orange. P-value annotation legend: ns: $0.05 < p <= 1.00$, *: $0.01 < p <= 0.05$, **: $0.001 < p <= 0.01$, ***: $0.0001 < p = 0.001$, ****: $p <= 0.0001$.

The temporal variations in class imbalance were examined by analyzing the relationship between the relative target size and the date of SEEG implantation (used as a proxy for time). For the SAUH dataset, the percentage of resected contacts per patient decreased significantly between 2012 and 2022 (Spearman's correlation coefficient (SCC) = -0.42, p-value = 0.002). In contrast, the percentage of SOZ and resected SOZ contacts showed no significant change. For the MNI patients, the analysis revealed no significant correlations, suggesting that the relative size of none of the localization targets has changed over time. The dependency plots and analysis results are depicted in Fig. 4.

The explanation of the changes in relative target size occurring over time may lie in the significant increase in the total number of implanted electrode contacts per patient, which was found at both Institutions (SCC of 0.63 at SAUH and 0.43 at the MNI), as visualized in Supplementary Figure 1.

For all localization target definitions, the class imbalance between the two Institutions was different, with a significantly larger class imbalance at the MNI as compared to SAUH patients ($p < 0.001$), as shown in Fig. 3. The effect size was large for all comparisons with Cliff's delta of 0.84, 0.56, and 0.70 for the "SOZ", "Resected" and "SOZ&Resected" targets, respectively, showing the largest difference in the relative counts of contacts determined to be in the SOZ by clinicians across the Institutions. The SOZ contacts constituted a median (interquartile range, IQR) of 3.75 % (3.43 %) of all contacts implanted per patient at SAUH and, in contrast, 15.38 % (12.51 %) at the MNI. Similarly, the median (IQR) percentage of contacts resected during surgery was 9.26 % (11.71 %) at SAUH compared to 17.49 % (16.16 %) at the MNI and median of 2.76 % (2.24 %) of contacts were SOZ contacts resected during surgery at SAUH and 6.85 % (6.08 %) at the MNI. Subsequent analysis of the differences in absolute target size (number of target electrode contacts) across Institutions with the Mann-Whitney test also showed significant differences for all target definitions. The complete results are provided in Supplementary Table S3 and S4.

3.3. Metric analysis using simulated data

3.3.1. Monte Carlo simulation

During the Monte Carlo simulation, we fixed the model's classification accuracy for both the positive and negative classes (TPR, TNR) and varied the relative target size of input data between 1 % and 100 %. This simulation aimed to observe the effect of class imbalance on the evaluation metrics. The outcomes of this simulation are depicted in Fig. 5. Notably, a critical range between 4 % and 18 % relative target size is emphasized to reflect the range observed in clinical datasets.

The **sensitivity** (TPR) and **specificity** (TNR) plots depict the consistent behavior of simulated models in classifying positive and negative samples. In contrast, the **PPV** and **NPV** metrics showed a distinct dependency on class distribution. As the relative target size increased, PPV rose while NPV declined, despite the model's consistent performance, illustrating their lack of robustness to class imbalance variations.

The **accuracy** plot demonstrates the assignment of equal weight to both classes and why this metric is unacceptable in imbalanced domains. For small relative target sizes, the accuracy metric is predominantly influenced by the accuracy in classifying the non-target samples (quantified by TNR), resulting in values around 0.9 for all models in our simulation. Consequently, a specific accuracy value, such as 0.85, may be achieved by multiple classifier models. This property makes it impossible to distinguish the superior classifier based on accuracy value without the knowledge of the level of class imbalance.

The **AUROC** metric stands out among the analyzed metrics as it demonstrated independence on relative target size. Its consistency despite changes in relative target size makes AUROC a reliable measure for discriminating superior classifiers in terms of their overall performance. Nevertheless, the AUROC metric assigns equal weight to both classes, which can lead to an overly optimistic evaluation in datasets with a majority of non-target samples. This issue is demonstrated in an

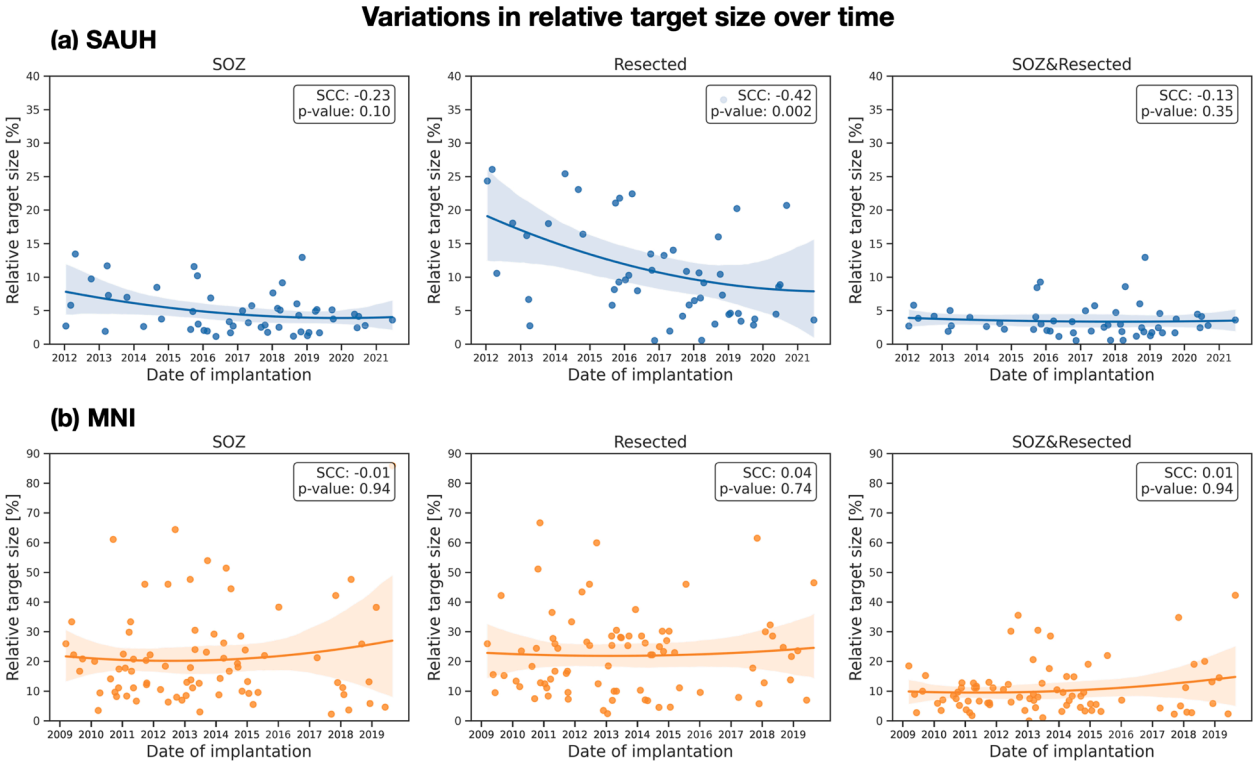

Fig. 4. The percentage of resected contacts at SAUH decreased over time. Plots of the dependency of relative target size on time for SAUH (A) and the MNI (B) datasets are visualized with the date of SEEG implantation used as a proxy for time. The Spearman's correlation coefficient (SCC) between relative target size and time as well as the p-value is reported for localization targets: "SOZ", "Resected", and "SOZ&Resected". Second-order polynomial regression was fitted to the data and visualized along with the 95% confidence interval.

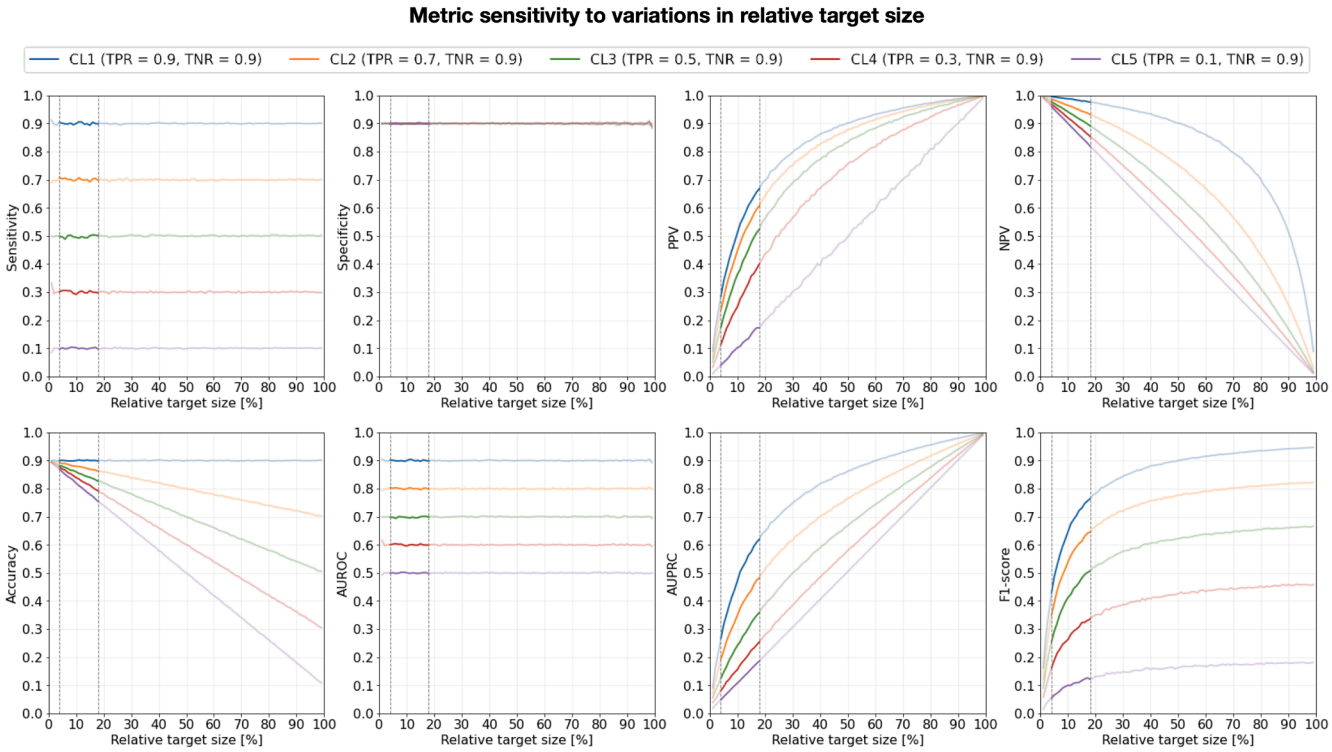

Fig. 5. PPV, NPV, accuracy, AUROC, and F1-score are sensitive to changes in relative target size. The dependency of sensitivity (also "true positive rate" or "recall"), specificity (also "true negative rate"), PPV (also "precision"), NPV, accuracy, AUROC, AUPRC, and F1-score on relative target size is visualized for 5 classification models (CL1-5). Models are defined by true positive rate (TPR) and true negative rate (TNR). The range between 4 and 18% relative target size, found in clinical data, is highlighted.

additional simulation in Supplementary Figure 2.

In contrast, the **AUPRC** metric exhibited a direct dependency on class imbalance levels, displaying variations in its values in response to changes in the relative target size. High metric values were observed in scenarios with higher relative target size values, signifying less imbalanced data. Conversely, lower AUPRC values were observed in instances of greater class imbalance. Consequently, relying solely on AUPRC proves insufficient for effectively distinguishing between strong and weak classifiers, as the class distribution heavily influences its outcomes in the dataset. To illustrate, a classification result of AUPRC = 0.3 could be produced by either of three classifiers, contingent on the specific class imbalance level.

The **F1-score** exhibited similar properties to the AUPRC, particularly when non-target samples dominated the dataset. Similar to AUPRC, the F1-score showed a dependency on the relative target size and favored less imbalanced data. Consequently, the F1-score faced similar challenges as AUPRC when distinguishing between classifiers based solely on the metric values. This was particularly evident in the lower ranges of relative target size values prevalent in EZ localization datasets.

As such, reporting these metrics in the context of imbalanced datasets requires careful consideration of the class distribution to avoid potentially biased conclusions about model performance.

3.4. Metric analysis using clinical data

The patient cohort was further refined according to the criteria described in the Supplementary Methods section, resulting in a localization patient cohort of 25 patients, with 8 gathered from SAUH and 17 from the MNI. Logistic regression models, each trained and tested on a specific localization target ("SOZ", "Resected", and "SOZ&Resected"), were evaluated for each patient through a leave-one-patient-out cross-validation approach for the localization of their respective targets. A median (IQR) of 9.2 % (13.3 %) of all contacts per patient was identified as the SOZ. In contrast, a median of 10.7 % (18.4 %) of contacts per patient was resected during the surgical intervention, while a median of 5.7 % (6.0 %) constituted SOZ contacts within the resected zone.

To analyze the evaluation metrics, we present four clinical cases from the localization cohort, highlighting specific limitations and advantages associated with selected metrics. Then, we show two examples of group analysis of results (across Institutions and across localization targets), including suitable statistical testing for datasets with different levels of class imbalance, and interpretation of the results.

3.4.1. Accuracy: Unmasking suboptimal performance

Solely relying on accuracy as an evaluation metric can be misleading, as exemplified by the case of patient number 89. Despite a high accuracy score of 0.953, a closer examination of the confusion matrix in Fig. 6 reveals the model as a no-skill classifier, misclassifying all electrodes as normal. The model achieved an F1-score of zero, with AUROC and AUPRC values only marginally exceeding chance levels, thus exposing the shortcomings of accuracy in reflecting true model performance within imbalanced datasets.

3.4.2. AUROC: An incomplete picture

The inadequacy of AUROC in addressing imbalanced data is highlighted by the case of patient 77, presented in Fig. 7. Despite a seemingly high AUROC of 0.970 for localizing "SOZ&Resected" contacts, a detailed analysis of AUPRC shows a less optimistic perspective. The AUPRC, which represents the model's average precision, indicates the overall ability of the model to localize the target with an average precision of 0.236 across all possible classification targets, although for the threshold chosen by the model, the F1-score was zero. This suggests that the model's predictions, while achieving a high AUROC, have limited clinical relevance as a substantial portion of the predicted positive instances does not correspond to actual positive cases. Additionally, the model yielded an F1-score of zero since no target contacts were correctly

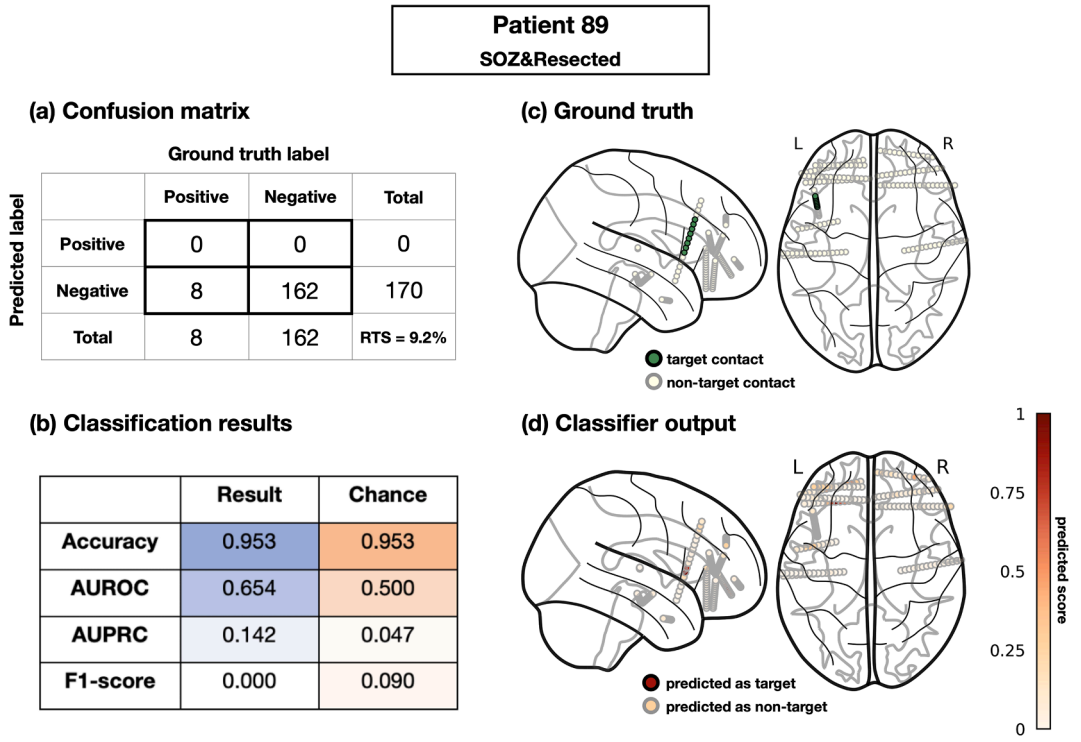

Fig. 6. Classification results for patient 89 and "SOZ&Resected" model, including the confusion matrix (A), model results with corresponding chance levels (B), and visualization of SEEG contacts projected on a standard MNI brain model with ground truth (C) and predicted labels (D). Contacts with black edges have a positive label, while contacts with gray edges have negative labels. The color gradient symbolizes the scores assigned to each contact by the classifier normalized to a range between 0 and 1.

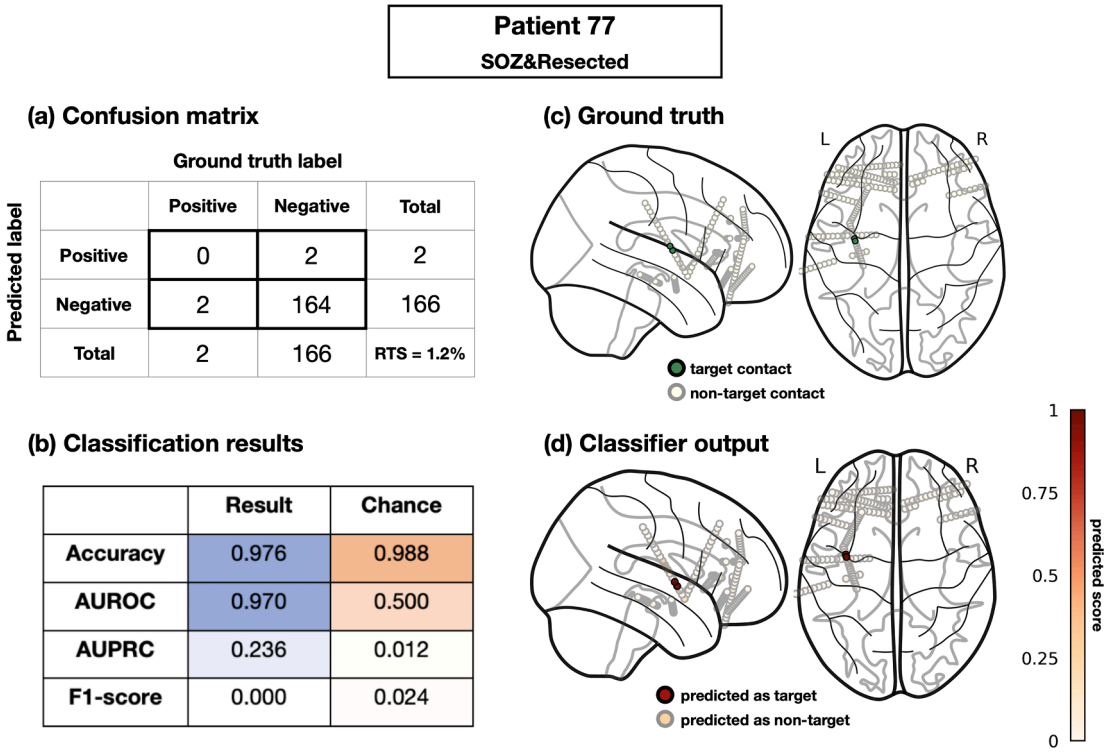

Fig. 7. Classification results for patient 77 and "SOZ&Resected" model, including the confusion matrix (A), model results with corresponding chance levels (B), and visualization of SEEG contacts projected on a standard MNI brain model with ground truth (C) and predicted labels (D). Contacts with black edges have a positive label, while contacts with gray edges have negative labels. The color gradient symbolizes the scores assigned to each contact by the classifier normalized to a range between 0 and 1.

identified with the proposed threshold. This case underscores the risk of overly optimistic conclusions based on high AUROC values, especially in EZ localization where the underrepresented positive class holds primary importance.

3.4.3. AUPRC: Addressing class imbalance

The limitations of AUPRC are exemplified in the case of patient 1153 and "SOZ" localization, illustrated in Fig. 8. Despite achieving a high AUPRC of 0.817, the high proportion (51.5 %) of target SOZ contacts calls the metric's validity into question. Even a naive strategy of labeling all contacts as positive would yield an AUPRC of 0.515, raising concerns about the clinical relevance of the model's performance. The inherent imbalance leads to an inflated AUPRC, emphasizing the need for a nuanced interpretation and consideration of alternative metrics, such as AUROC. In this case, AUROC showed an average performance of 0.786, aligning with the model's suboptimal accuracy of 0.495 (slightly above chance) and an F1-score of 0.038. These findings underscore the model's limitations, which could not be captured by the AUPRC, and emphasize the need for a comprehensive evaluation strategy that goes beyond individual metrics.

3.4.4. F1-score: Balancing precision and recall

The significance of the F1-score is underscored in the case of patient 965 and the "SOZ&Resected" model, shown in Fig. 9. Despite the model's success in assigning higher scores to target contacts, reflected in excellent AUROC and AUPRC values of 0.996 and 0.917, it struggled to identify an optimal classification threshold to distinguish between target and non-target contacts. The F1-score of 0.5 indicates a suboptimal trade-off between precision and recall, offering valuable insights into the model's performance and its clinical relevance..

3.5. Group analysis – Across Institutions

To perform a group analysis across Institutions, patient results were aggregated separately for each Institution and evaluated by comparing results between two Institutions, SAUH and the MNI, across four evaluation metrics: accuracy, AUROC, AUPRC, and F1-score. This evaluation example simulates the real-life scenario of cross-institutional testing, which is essential for model validation. The results for model "SOZ", depicted in Fig. 10, reveal important insights into the model's effectiveness at each Institution. Results for "Resected" and "SOZ&Resected" models are visualized in Supplementary Figure S3.

Firstly, the accuracy metric with median values of 0.953 for SAUH and 0.873 for the MNI shows no statistically significant difference between the two Institutions ($p = 0.996$). This suggests that the "SOZ" model performs similarly in terms of accuracy at both Institutions despite the difference in metric values.

In terms of AUROC, the median values were 0.791 for SAUH and 0.895 for the MNI. With a p-value of 0.379, this difference was also not statistically significant, suggesting that the ability of the "SOZ" model to distinguish between classes is comparable at both Institutions.

A notable difference was observed in the AUPRC metric, where SAUH had a median value of 0.277, significantly lower than the 0.817 observed for the MNI ($p < 0.001$). This implies that the MNI's model performance in terms of precision-recall trade-off is markedly better than that of SAUH.

Lastly, the F1-score showed median values of 0.292 for SAUH and a significantly better score of 0.400 for the MNI ($p = 0.015$). This indicates that the MNI achieves a better harmonic mean of precision and recall compared to SAUH.

In summary, while the model showed higher accuracy scores for SAUH, a more comprehensive analysis of model performance revealed a significant superiority of AUPRC and F1-score performance for the MNI. These findings suggest that while the model's ability to correctly classify

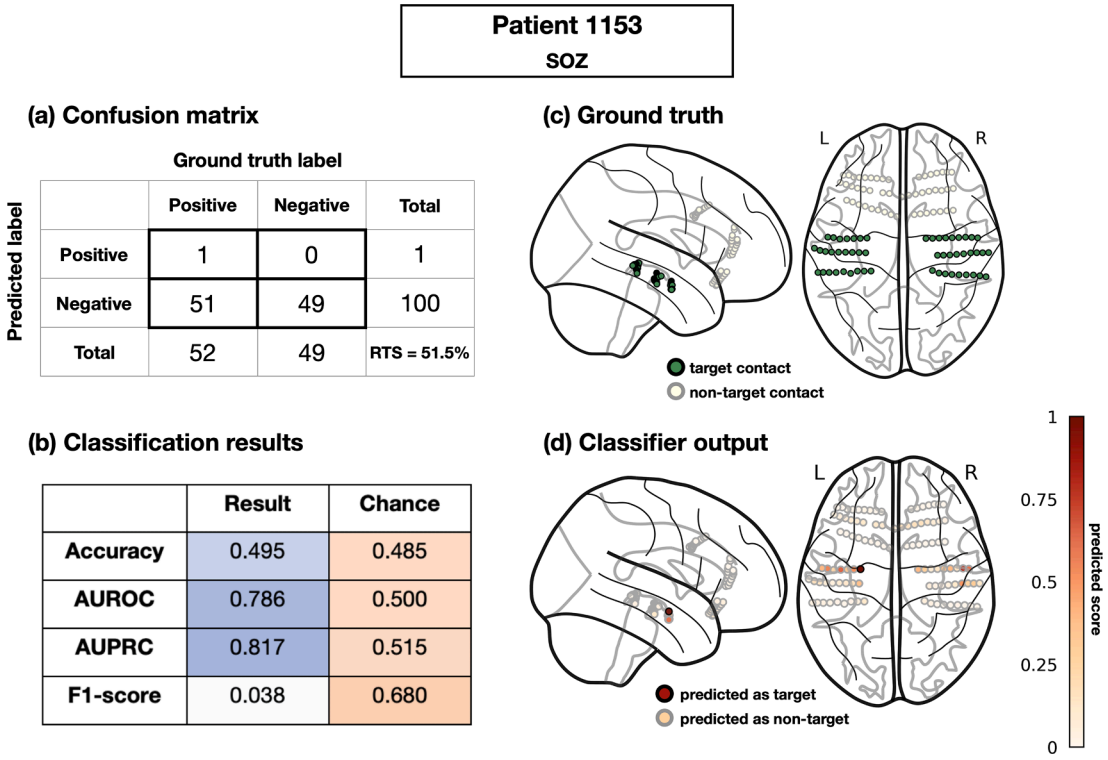

Fig. 8. Classification results for patient 1153 and "SOZ" model, including the confusion matrix (A), model results with corresponding chance levels (B), and visualization of SEEG contacts projected on a standard MNI brain model with ground truth (C) and predicted labels (D). Contacts with black edges have a positive label, while contacts with gray edges have negative labels. The color gradient symbolizes the scores assigned to each contact by the classifier normalized to a range between 0 and 1.

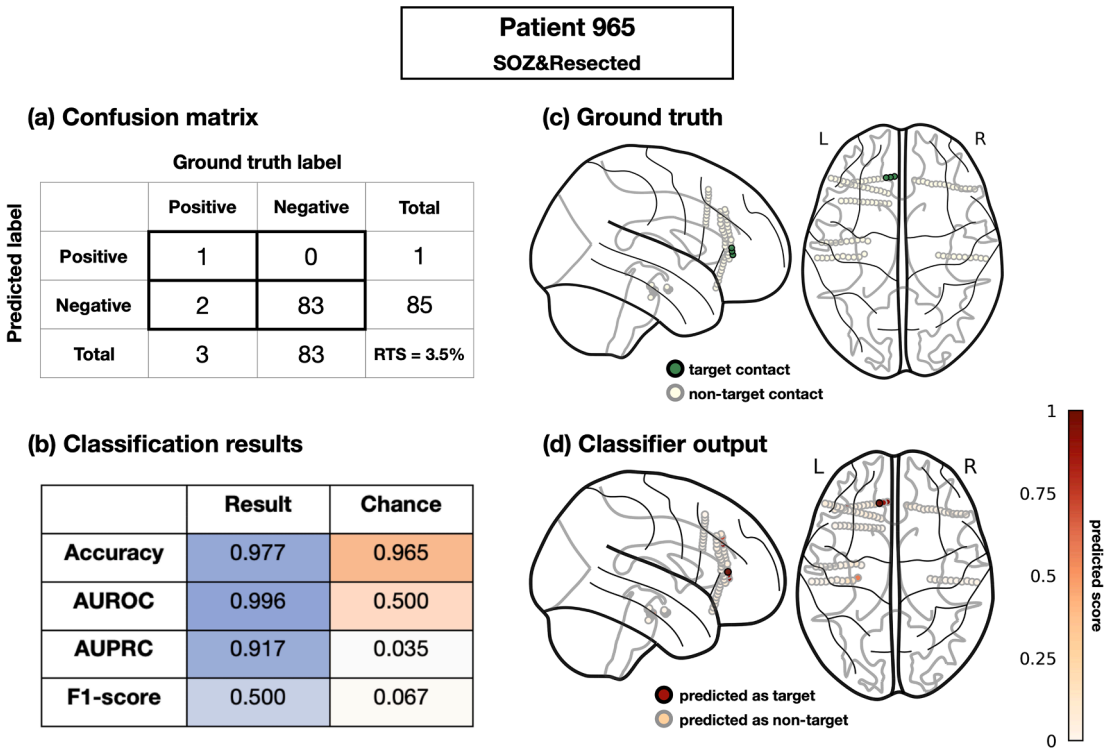

Fig. 9. Classification results for patient 965 and "SOZ&Resected" model, including the confusion matrix (A), model results with corresponding chance levels (B), and visualization of SEEG contacts projected on a standard MNI brain model with ground truth (C) and predicted labels (D). Contacts with black edges have a positive label, while contacts with gray edges have negative labels. The color gradient symbolizes the scores assigned to each contact by the classifier normalized to a range between 0 and 1.

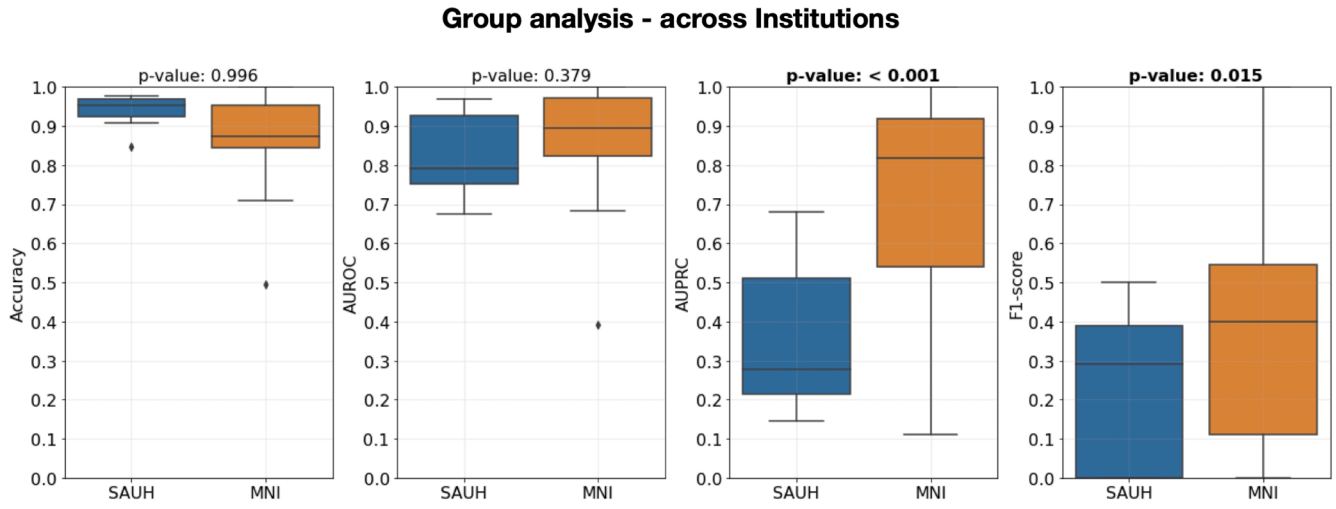

Fig. 10. Group analysis of the "SOZ" model's predictions reveals significantly better performance for the MNI. The distributions of classification metrics across patients from SAUH (N = 8) and the MNI (N = 17) are visualized, with a horizontal line as the median. Results of statistical testing (randomization test for accuracy, AUPRC and F1-score, and Hanley-McNeil test for AUROC) are reported with significant results in bold.

instances is similar at both Institutions, the MNI benefits from better precision-recall characteristics considering the class imbalance in respective datasets.

3.6. Group analysis – Across localization targets

For group analysis across localization targets, patient results were aggregated across all 25 patients in the localization cohort and evaluated by comparing results of the three models, ("SOZ", "Resected", and "SOZ&Resected"), across four evaluation metrics: accuracy, AUROC, AUPRC, and F1-score. A summary of the results is presented in Fig. 11 with complete results in Supplementary Table S6.

The model trained for localizing "SOZ&Resected" contacts demonstrated the highest accuracy (0.953), outperforming "SOZ" (0.907) and "Resected" (0.893) models in the localization of their respective targets. However, the differences in accuracy values lacked statistical significance when tested with the randomization test. In terms of AUROC, "SOZ&Resected" (0.895) outperformed both the "SOZ" (0.870) and "Resected" (0.787) models with statistical significance (p-values of 0.006 and < 0.001) measured by the Hanley-McNeil test. The model

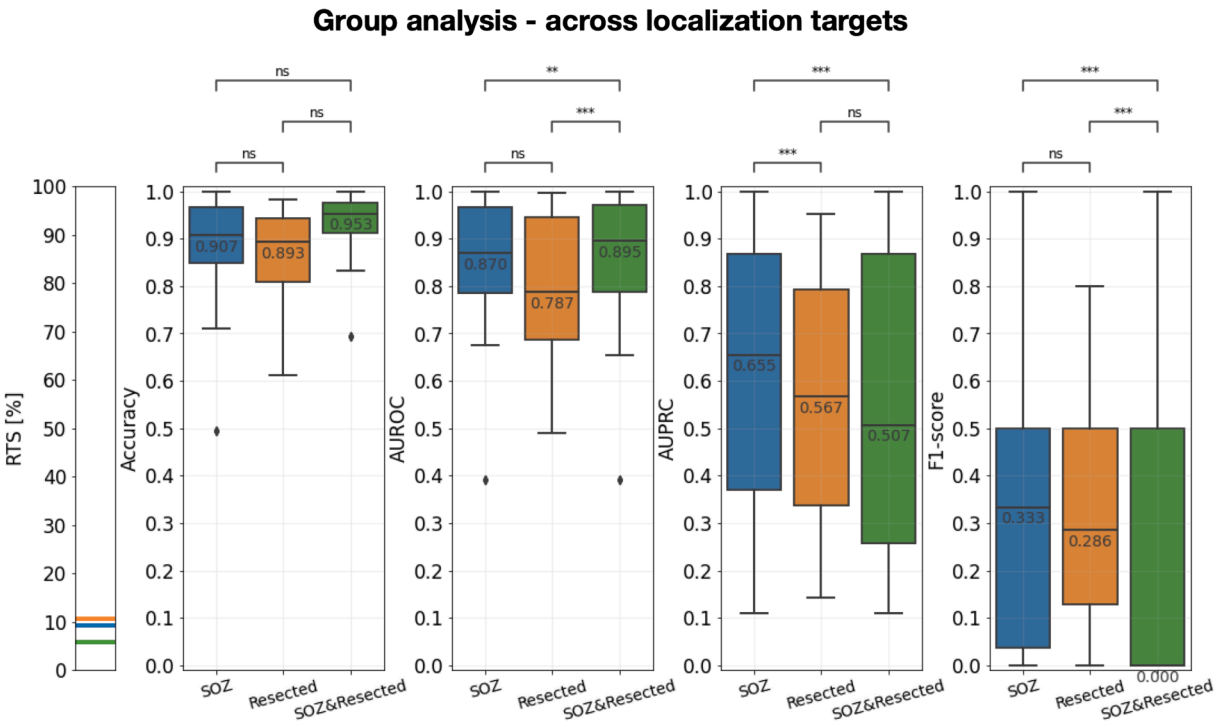

Fig. 11. On the left, the relative target size (RTS) for "SOZ" (blue), "Resected" (orange), and "SOZ&Resected" (green) targets are visualized. On the right, the distributions of classification metrics across cross-validation folds for the "SOZ", "Resected", and "SOZ&Resected" classification models are visualized, with a horizontal line as the median and the median values reported for each boxplot. Results of statistical testing (randomization test for accuracy, AUPRC and F1-score, and Hanley-McNeil test for AUROC) are reported with p-value annotation legend: ns: $0.017 < p <= 1.00$, *: $0.01 < p <= 0.017$, **: $0.001 < p <= 0.01$, ***: $0.0001 < p = 0.001$, ****: $p <= 0.0001$.

aimed at localizing "SOZ" contacts, with an AUPRC of 0.655, significantly outperformed the remaining models (p-value < 0.001) in a randomization test. The "Resected" contacts localization model followed with an AUPRC of 0.567 and the "SOZ&Resected" model with an AUPRC of 0.507 without a significant difference between the two results. The "SOZ" model achieved the best F1-score of 0.333, followed again by the "Resected" model with a score of 0.286 without a significant difference and the "SOZ&Resected" model with a score of 0.

Based on the results, we may conclude that the "SOZ&Resected" model has outperformed the remaining models in terms of their overall ability to assign higher scores to target contacts and lower scores to non-target contacts. However, in our dataset, the median relative size of the "SOZ&Resected" target (5.7 %) was significantly smaller than both the "SOZ" target (9.2 %) and the "Resected" target (10.7 %). The non-target contacts, which constitute 94.3 % of all contacts for the "SOZ&Resected" target, are easier to classify since they provide more training samples for the machine learning model. Consequently, the AUROC result may be driven by an excellent performance of the "SOZ&Resected" model in classifying the non-target contacts, potentially resulting in an overly optimistic assessment of the localization model's performance. In terms of performance focused on the target contacts, the "SOZ" model has shown the best results as it outperformed the remaining models in AUPRC with statistical significance. To interpret the achieved AUPRC value, on average over all possible classification thresholds, a median of 65.5 % of the contacts marked as SOZ by the "SOZ" model were actual SOZ contacts per patient.

Supplementary Figure S4 visualizes the ROC and PR curves to illustrate model performance over the range of classification thresholds.

4. Discussion

This study has systematically addressed the pivotal challenges in evaluating EZ localization models, presenting an in-depth analysis of commonly used evaluation metrics within this domain. Previous research, such as the work by Bernabei et al. (Bernabei et al. 2023) has highlighted several pitfalls in automatic EZ localization, including the significant variability in data related to implant types, therapeutic approaches, underlying pathologies, and outcome metrics. Furthermore, contributions from studies by Zhao et al. (Zhao et al. 2022) and Varotto et al. (Varotto et al. 2021) have advanced our understanding in areas of data augmentation and classification model design to mitigate some of these challenges. Despite these advancements, our study represents the first investigation into evaluating EZ localization models in intracranial electrophysiology, with a particular focus on the implications of class imbalance.

Our main findings underscore that relying solely on any single of the analyzed metrics provides an incomplete perspective on model performance, particularly when not accounting for class imbalance inherent in clinical datasets. Simple metrics like specificity, recall, precision, and negative predictive value, while straightforward, fail to assess model performance comprehensively. Similarly, widely used metrics such as accuracy, AUROC, AUPRC, and F1-score each fail to adequately address at least one of the unique challenges EZ localization poses. A combination of AUROC and AUPRC is therefore advised for robust evaluation.

4.1. Impact of class imbalance on model evaluation

The analysis of class imbalance in the clinical datasets clearly demonstrates the issue we face in the field of automatic EZ localization. Due to the different class distributions in the datasets, it is important to conduct appropriate statistical testing and acknowledge the impact of class imbalance on each metric when interpreting the results.

Patient-level evaluations, especially for datasets where the proportion of target to non-target contacts ranges widely (e.g., 3 % to 86 % SOZ contacts in the MNI dataset), must be approached with caution. The chance levels of precision-recall metrics, such as PPV, AUPRC, or F1-

score, span over a broad range of values, depending on the range of class imbalance in the data, emphasizing the importance of understanding class distributions for objective metric interpretation.

Moreover, the method of approximation of the EZ significantly impacts class distribution, as demonstrated in both Institutions. When comparing results for models trained on identical patients but under different target definitions, the differences in class distributions among the targets result in differences in metric chance levels. Given these circumstances, conventional paired tests become unsuitable, and alternative statistical methods must be considered. In our study, we propose the Hanley-McNeil test for the comparison of two independent AUROC values and the randomization test for the comparison of independent accuracy, AUPRC, and F1-score values.

Throughout our analysis, we also observed variations in the relative size of targets over time within one Institution. Notably, at SAUH, there was a significant decrease in the relative number of resected contacts from 2012 to 2022, attributed to a simultaneous increase in the total number of implanted contacts at the Institution. In contrast, we did not detect significant changes in relative target sizes at the MNI, although there was a noteworthy increase in the overall number of implanted contacts. This underscores the importance of exercising caution when splitting data into training and testing datasets in retrospective studies to avoid accentuating differences in class imbalance through the split.

Cross-institutional validation is crucial for assessing model generalization abilities (Jehi 2023). Our analysis revealed notable differences in class imbalance within clinical datasets from SAUH and the MNI across all target definitions. This underscores the importance of considering class distribution when interpreting the results of cross-institutional testing and the necessity to employ suitable statistical tests, such as the Hanley-McNeil and randomization tests, as demonstrated in the group analysis.

4.2. Critical analysis of common evaluation metrics

To summarize the main findings from the metric analysis, none of the commonly used evaluation metrics meets all the criteria defined for the evaluation framework. The criteria were that the evaluation framework must (i) comprehensively assess model performance, (ii) emphasize the evaluation of the minority class, and (iii) be robust to variations in class distribution.

Sensitivity, specificity, PPV, and NPV do not comprehensively evaluate the model performance since they focus only on a limited aspect of model performance. Accuracy weights each class according to its frequency in the dataset, leading to misleading conclusions in imbalanced datasets. AUROC, by assigning equal weight to both pathologic and normal contacts, tends to be influenced by the more prevalent negative class samples (i.e., normal contacts), which are of less clinical interest. This can potentially result in an overly optimistic evaluation since majority-class samples are typically easier to classify correctly compared to minority-class samples. As demonstrated in Supplementary Figure S2, AUROC cannot capture the difference between accuracy on the minority positive class (TPR) and accuracy on the majority negative class (TNR), although TPR is undeniably more relevant in EZ localization. In contrast, precision-recall metrics, including AUPRC and the F1-score, prioritize the minority class, representing pathologic contacts critical for accurate diagnosis and treatment in EZ localization. However, these metrics inherently favor less imbalanced data and are not robust to changes in class distribution.

4.2.1. Alternatives beyond traditional metrics

Alternative metrics that directly address the imbalance problem exist, such as balanced accuracy and localized ROC. Balanced accuracy, for example, averages the sensitivity and specificity, thus treating both classes equally regardless of their size in the dataset and, therefore, suffering the same limitation as AUROC. Localized ROC, or variations that focus on specific regions of the ROC curve, can provide insights into the performance of a model at clinically relevant decision thresholds. By focusing on the trade-offs between sensitivity and specificity in a balanced manner, these metrics can offer a more detailed view of model performance in contexts where certain errors are more costly than others. However, clinical datasets are inherently imbalanced, often significantly so, and metrics that require balanced conditions for optimal evaluation could, therefore, provide a distorted view of how a model performs in actual clinical settings.

Due to the sensitivity of AUPRC to class imbalance, several studies have proposed modifications to this metric based on its normalization. Boyd et al. (Kendrick, 2012) point out the existence of an “unachievable region” in the PR space, which limits the possible AUPRC values a model can achieve depending on the level of class imbalance in the data. As a solution, they propose the Area Under the Normalized PR Curve (AUCNPR), which is essentially the value of AUPRC normalized to a range of its achievable values. Flach and Kull (Flach and Kull 2015) rigorously analyzed the shortcomings of PR metrics and plots and defined a new metric, “AUPRC Gain”, as an alternative to AUPRC robust to changes in class distribution. Although both of the metrics show potential in addressing the sensitivity of AUPRC to class imbalance, neither of them proved to be completely robust to variations in relative target size in our analysis, as visualized in Supplementary Figure S5.

4.3. Recommendations for robust model evaluation

Based on our findings, we recommend reporting AUROC and AUPRC values as primary model results. AUROC, a widely accepted metric, assesses the overall model performance, maintaining robustness to class imbalance variations. Conversely, AUPRC provides insights into the model’s performance on target contacts, emphasizing its clinical utility, as the value of AUPRC represents the average precision at different thresholds of the model in localizing the EZ. To address AUPRC bias towards less imbalanced data, it is essential to report chance levels. Together, these metrics comprehensively capture key aspects of model performance.

To supplement these metrics, we suggest including the F1-score when a specific threshold’s performance is of interest. However, the F1-score should not be used as a substitute for AUPRC. Alternatively, the generalized F_β-score allows customization based on specific needs, emphasizing precision (with a lower beta, e.g., 0.5) or recall (with a higher beta, e.g., 2) depending on the application. This adaptability suits scenarios like defining the localization target as resected contacts or SOZ contacts resected during surgery, respectively, and provides further clinical insight.

Additionally, we suggest statistical tests (Hanley-McNeil and randomization tests) that enable effective model comparisons, accommodating different class distributions in training datasets.

5. Conclusions

In conclusion, we propose that the value of AUROC and AUPRC should be reported together for a comprehensive assessment of binary classification models for epileptogenic zone localization. Alongside metric values, it is crucial to report the class distribution and its impact on classification results should be discussed to draw valid conclusions about model performance. Furthermore, the inclusion of the F1-score is recommended when evaluating class assignments of samples. The adoption of this evaluation framework will not only enhance the comparability of study results but also contribute to the development of more reliable machine-learning models for epileptogenic zone localization in intracranial electrophysiology. By systematically addressing the challenges of class imbalance and providing a robust analytical framework, our study lays a foundation for more accurate and clinically relevant evaluations, ensuring better generalization of models across diverse clinical datasets.

Declaration of Generative AI and AI-assisted technologies in the writing process

During the preparation of this work the author(s) used ChatGPT by OpenAI in order to improve text readability. After using this tool, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the publication.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

This work was supported by the Canadian Institutes of Health Research (PJT-175056, B.F.), Duke University start-up funding (B.F.), the European Union's Next Generation EU (project nr. LX22NPO5107), the Czech Science Foundation (project 22-28784S), the Ministry of Health of the Czech Republic (project NU22-08-00278), the Ministry of Education, Youth and Sport of the Czech Republic (EATRIS-CZ, LM2023053), and The Czech Academy of Sciences (project RVO: 68081731).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.clinph.2024.11.007.

References

AKTER, Most. Sheuli,, et al., 2020. Multiband entropy-based feature-extraction method for automatic identification of epileptic focus based on high-frequency components in interictal iEEG. *Sci. Rep.*. 10 (1), 7044. https://doi.org/10.1038/s41598-020-62967-z.

Bernabei, j. m., et al., 2022. Normative intracranial EEG maps epileptogenic tissues in focal epilepsy. *Brain*. 145 (6), 1949–1961. https://doi.org/10.1093/brain/awab480.

Bernabei, j. m., et al., 2023. Quantitative approaches to guide epilepsy surgery from intracranial EEG. *Brain*. 146 (6), 2248–2258. https://doi.org/10.1093/brain/awad007.

BOYD, Kendrick et al., 2012. Unachievable Region in Precision-Recall Space and Its Effect on Empirical Evaluation. *Proceedings of the International Conference on Machine Learning. International Conference on Machine Learning*. Vol. 2012, p. 349.

BRANCO, Paula, TORGO, Luis and RIBEIRO, Rita, 2015. *A Survey of Predictive Modelling under Imbalanced Distributions*. arXiv. DOI: 10.48550/ARXIV.1505.01658.

CHEN, Zhibin,, et al., 2018. Treatment Outcomes in Patients With Newly Diagnosed Epilepsy Treated With Established and New Antiepileptic Drugs: A 30-Year Longitudinal Cohort Study. *JAMA Neurol.*. 75 (3), 279. https://doi.org/10.1001/jamaneurol.2017.3949.

CHYBOWSKI, Bartlomiej,, et al., 2024. Timing matters for accurate identification of the epileptogenic zone. *Clin. Neurophysiol.*. 161, 1–9. https://doi.org/10.1016/j.clinph.2024.01.007.

Cimbalnik, j., et al., 2018. Physiological and pathological high frequency oscillations in focal epilepsy. *Ann Clin Transl Neurol.* 5 (9), 1062–1076. https://doi.org/10.1002/acn3.618.

Cimbalnik, j., et al., 2019. Multi-feature localization of epileptic foci from interictal, intracranial EEG. *Clin Neurophysiol.* 130 (10), 1945–1953. https://doi.org/10.1016/j.clinph.2019.07.024.

Conrad, e. c., et al., 2023. Spike patterns surrounding sleep and seizures localize the seizure-onset zone in focal epilepsy. *Epilepsia*. 64 (3), 754–768. https://doi.org/10.1111/epi.17482.

Conrad, e. c.,, et al., 2022. Addressing spatial bias in intracranial EEG functional connectivity analyses for epilepsy surgical planning. *J Neural Eng*. Vol. 19, no. 5. https://doi.org/10.1088/1741-2552/ac90ed.

DAVIS, Jesse and GOADRICH, Mark, 2006. The relationship between Precision-Recall and ROC curves. In : *Proceedings of the 23rd international conference on Machine learning - ICML '06*, pp. 233–240. Pittsburgh, Pennsylvania : ACM Press. 2006. ISBN 978-1-59593-383-6. DOI: 10.1145/1143844.1143874.

Elahian, b.,, et al., 2017. Identifying seizure onset zone from electrocorticographic recordings: A machine learning approach based on phase locking value. *Seizure*. 51, 35–42. https://doi.org/10.1016/j.seizure.2017.07.010.

Ellenrieder, V.O.N., N., et al., 2016. Interaction with slow waves during sleep improves discrimination of physiologic and pathologic high-frequency oscillations (80-500 Hz). *Epilepsia*. 57 (6), 869–878. https://doi.org/10.1111/epi.13380.

FAWCETT, Tom,, 2006. An introduction to ROC analysis. *Pattern Recogn. Lett.*. 27 (8), 861–874. https://doi.org/10.1016/j.patrec.2005.10.010.

FLACH, Peter and KULL, Meelis, 2015. Precision-Recall-Gain Curves: PR Analysis Done Right. In : CORTES, C. et al. (eds.), *Advances in Neural Information Processing Systems* [online]. Curran Associates, Inc. 2015. Retrieved from : https://proceedings.neurips.cc/paper_files/paper/2015/file/33e8075e9970de0cfea955afd4644bb2-Paper.pdf.

Frauscher, b., et al., 2024. Learn how to interpret and use intracranial EEG findings. *Epileptic Disord.*. 26 (1), 1–59. https://doi.org/10.1002/epd2.20190.

GIREESH, Elakkat D., et al., 2023. Deep neural networks and gradient-weighted class activation mapping to classify and analyze EEG. *Intell. Decis. Technol.*. 17 (1), 43–53. https://doi.org/10.3233/IDT-228040.

Grinenko, o., et al., 2018. A fingerprint of the epileptogenic zone in human epilepsies. *Brain*. 141 (1), 117–131. https://doi.org/10.1093/brain/awx306.

Gunnarsdottir, k. m., et al., 2022. Source-sink connectivity: A novel interictal EEG marker for seizure localization. *Brain*. 145 (11), 3901–3915. https://doi.org/10.1093/brain/awac300.

HANLEY, J A and MCNEIL, B J,, 1982. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*. 143 (1), 29–36. https://doi.org/10.1148/radiology.143.1.7063747.

He, H., Garcia, E.A., 2009. Learning from Imbalanced Data. *IEEE Trans. Knowl. Data Eng.*. 21 (9), 1263–1284. https://doi.org/10.1109/TKDE.2008.239.

JEHI, Lara,, 2018. The Epileptogenic Zone: Concept and Definition. *Epilepsy Currents*. 18 (1), 12–16. https://doi.org/10.5698/1535-7597.18.1.12.

JEHI, Lara,, 2023. Machine Learning for Precision Epilepsy Surgery. *Epilepsy Currents*. 23 (2), 78–83. https://doi.org/10.1177/15357597221150055.

JIANG, H. et al., 2022. Interictal SEEG resting-state connectivity localizes seizure onset zone and predicts seizure outcome. . No. (Jiang H.; Ye S.; He B., bhe1@andrew.cmu.edu) Department of Biomedical Engineering, Carnegie Mellon University, Pittsburgh, PA, United States. DOI: 10.1101/2021.12.30.21268524.

Jose, b.,, et al., 2023. Improving the accuracy of epileptogenic zone localization in stereo EEG with machine learning algorithms. *Brain Res*. 1820, 148546. https://doi.org/10.1016/j.brainres.2023.148546.

Karunakaran, s., et al., 2018. The interictal mesial temporal lobe epilepsy network. *Epilepsia*. 59 (1), 244–258. https://doi.org/10.1111/epi.13959.

Klimes, p., et al., 2019. NREM sleep is the state of vigilance that best identifies the epileptogenic zone in the interictal electroencephalogram. *Epilepsia*. 60 (12), 2404–2415. https://doi.org/10.1111/epi.16377.

LAI, Dakun,, et al., 2020. Channel-Wise Characterization of High Frequency Oscillations for Automated Identification of the Seizure Onset Zone. *IEEE Access*. 8, 45531–45543. https://doi.org/10.1109/ACCESS.2020.2978290.

LUNDSTROM, Brian Nils, BRINKMANN, Benjamin H and WORRELL, Gregory A,, 2021. Low frequency novel interictal EEG biomarker for localizing seizures and predicting outcomes. *Brain. Communications*. Vol. 3, no. 4, p. fcab231. https://doi.org/10.1093/braincomms/fcab231.

Modur, P., Miocinovic, S., 2015. Interictal high-frequency oscillations (HFOs) as predictors of high frequency and conventional seizure onset zones. *Epileptic Disord.* 17 (4), 413–424. https://doi.org/10.1684/epd.2015.0774.

MOOIJ, Anne H., et al., 2016. Differentiating epileptic from non-epileptic high frequency intracerebral EEG signals with measures of wavelet entropy. *Clin. Neurophysiol.*. 127 (12), 3529–3536. https://doi.org/10.1016/j.clinph.2016.09.011.

PROVOST, F., FAWCETT, Tom and KOHAVI, R.,, 1998. In: The case against accuracy estimation for comparing induction algorithms. on Ma-chine Learning, pp. 445–453.

Saboo, k. v., et al., 2021. Leveraging electrophysiologic correlates of word encoding to map seizure onset zone in focal epilepsy: Task-dependent changes in epileptiform activity, spectral features, and functional connectivity. *Epilepsia*. 62 (11), 2627–2639. https://doi.org/10.1111/epi.17067.

Shahabi, H., Nair, D.R., Leahy, R.M., 2023. Multilayer brain networks can identify the epileptogenic zone and seizure dynamics. *Elife*. 12. https://doi.org/10.7554/elife.68531.

SMUCKER, Mark D., ALLAN, James and CARTERETTE, Ben, 2007. A comparison of statistical significance tests for information retrieval evaluation. In : *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, pp. 623–632. Lisbon Portugal : ACM. 6 November 2007. ISBN 978-1-59593-803-9. DOI: 10.1145/1321440.1321528.

Spanedda, F., Cendes, F., Gotman, J., 1997. Relations Between EEG Seizure Morphology, Interhemispheric Spread, and Mesial Temporal Atrophy in Bitemporal Epilepsy. *Epilepsia*. 38 (12), 1300–1314. https://doi.org/10.1111/j.1528-1157.1997.tb00068.x.

Sumsky, S.L., Santaniello, S., 2019. Decision Support System for Seizure Onset Zone Localization Based on Channel Ranking and High-Frequency EEG Activity. *IEEE J Biomed Health Inform.* 23 (4), 1535–1545. https://doi.org/10.1109/jbhi.2018.2867875.

THIJS, Roland D,, et al., 2019. Epilepsy in adults. *Lancet*. 393 (10172), 689–701. https://doi.org/10.1016/S0140-6736(18)32596-0.

Thomas, j., et al., 2023. A Subpopulation of Spikes Predicts Successful Epilepsy Surgery Outcome. *Ann. Neurol.*. 93 (3), 522–535. https://doi.org/10.1002/ana.26548.

VAKHARIA, Vejay N., et al., 2018. Getting the best outcomes from epilepsy surgery. *Ann. Neurol.*. 83 (4), 676–690. https://doi.org/10.1002/ana.25205.

VAROTTO, Giulia,, et al., 2021. Comparison of Resampling Techniques for Imbalanced Datasets in Machine Learning: Application to Epileptogenic Zone Localization From Interictal Intracranial EEG Recordings in Patients With Focal Epilepsy. *Frontiers in Neuroinformatics*. 15, 715421. https://doi.org/10.3389/fninf.2021.715421.

Wang, a.,, et al., 2023. Resting-state SEEG-based brain network analysis for the detection of epileptic area. *J. Neurosci. Methods*. 390. https://doi.org/10.1016/j.jneumeth.2023.109839.

WANG, Yiping,, et al., 2022. Automatic Localization of Seizure Onset Zone Based on Multi-Epileptogenic Biomarkers Analysis of Single-Contact from Interictal SEEG. *Bioengineering*. 9 (12), 769. https://doi.org/10.3390/bioengineering9120769.

WEBB, Geoffrey I. and TING, Kai Ming, 2005. On the Application of ROC Analysis to Predict Classification Performance Under Varying Class Distributions. *Machine Learning*. Vol. 58, no. 1, pp. 25–32. DOI: 10.1007/s10994-005-4257-7.

Zhao, x.,, et al., 2022. Seizure onset zone classification based on imbalanced iEEG with data augmentation. *J Neural Eng*. Vol. 19, no. 6. https://doi.org/10.1088/1741-2552/aca04f.

ZWEIPHENNING, Willemiek J. E. M., et al., 2022. Correcting for physiological ripples improves epileptic focus identification and outcome prediction. *Epilepsia*. 63 (2), 483–496. https://doi.org/10.1111/epi.17145.