Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Forensic Auditing of Neurotechnology: Quantifying the Collapse of Translational Safety in Deep Learning-Based Stroke Rehabilitation

Waqas Ahmad

ahmad502492@gmail.com

Abdul Wali Khan University Mardan

Research Article

Keywords: Brain-Computer Interface, Stroke Rehabilitation, Translational Safety, Forensic Audit, Explainable AI, Regulatory Compliance, Deep Learning, Clinical AI

Posted Date: November 10th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11580626/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Forensic Auditing of Neurotechnology: Quantifying the Collapse of Translational Safety in Deep Learning-Based Stroke Rehabilitation

Waqas Ahmad[1*]

[1]Department of Computer Science, Abdul Wali Khan University Mardan, Mardan, 23200, Khyber Pakhtunkhwa, Pakistan.

Corresponding author(s). E-mail(s): ahmad502492@gmail.com, waqasahmad.cs@awkum.edu.pk;

Abstract

Background: The advent of deep learning-based EEG BCIs in stroke rehabilitation applications is abundant; however, their clinical translation is still at a standstill. Accuracy on benchmark datasets of healthy subjects often conceals important shortcomings in terms of generalizability, physiological plausibility, and real-time safety standards. To date, there is no standard methodology to evaluate these BCIs in accordance with regulatory requirements (FDA GMLP, EU AI Act) and physiological limitations (Hebbian plasticity window). **Methods:** A forensic audit of 24 studies of state-of-the-art (SOTA) techniques that appeared in high-impact publications between 2018 and 2025 was performed. Each paper was scored for six non-compensatory criteria for Translational Safety Thresholds (TST): subject independent validation, physiological XAI, feedback latency (¡300 ms), closed-loop control, standardized clinical outcomes (FMA-UE/WMFT), and proper sample size (N¿20). Four categories of studies were created based on technical safety and clinical translatability. Statistical significance was determined by Chi-Square Goodness-of-Fit tests. **Results:** The audit showed that there was a system failure of translational safety. None of the studies (0.0%) demonstrated high technical rigor along with low clinical readiness (Quadrant 1), while only one study (4.2%) satisfied both the technical and clinical criteria (Quadrant 2), but it failed to be validated in stroke patients. Most of them (75.0%) belonged to Quadrant 4 (Low Technical/Low Clinical Readiness). Some of the critical failure rates were 95.8% for splitting at the subject level (95% CI: 78.9–99.9%), 91.7% for physiological XAI, and 87.5% for latency compliance. The distribution was significantly different from the expected uniform distribution ($\boldsymbol{\chi^2(3) = 34.33, p < 0.0001}$). **Conclusion:** Recent SOTA EEG-BCI studies suffer from a highly dangerous mismatch between the engineering performance measures and patient safety standards. The almost complete lack of compliance with regulatory and neurophysiological standards indicates that the majority of models described in the literature are not suitable for practical applications. It is proposed that the implementation of Translational Safety Thresholds be made compulsory by academic journals and regulatory bodies in order to close the gap between them.

Keywords: Brain-Computer Interface, Stroke Rehabilitation, Translational Safety, Forensic Audit, Explainable AI, Regulatory Compliance, Deep Learning, Clinical AI

1 Introduction

1.1 The Global Crisis of Unsafe Neurotechnology

Stroke remains the second leading cause of death and the third leading cause of disability-adjusted life years (DALYs) globally, with up to 80% of stroke patients suffering from upper limb motor impairment

Table 1: Comparison of Translational Safety Thresholds (TST) vs. Existing Quality Assessment Tools.

Feature	TST Framework (Ours)	PEDro Scale	Cochrane RoB
Focus	Regulatory & Neurobiological Safety	General RCT Quality	Bias in Randomized Trials
Scoring	Binary Non-Compensatory (Pass/Fail)	Weighted Sum (0-10)	Low/Unclear/High Risk
AI-Specific	Yes (XAI, Latency, Data Leakage)	No	No
Regulatory Alignment	FDA GMLP, EU AI Act, ISO 13485	None	None
Clinical Validity	Mandatory Stroke Outcomes (FMA-UE)	Optional	Optional

[1, 2]. Despite a wide array of non-invasive BCI systems based on electroencephalography (EEG) having existed for more than two decades [3], no breakthrough in the clinical sphere has been observed despite an enormous growth in the complexity of algorithms being utilized [4, 5]. The latest generation of deep learning systems shows classification accuracy exceeding 90% on public databases, creating a false illusion of technological maturity that promotes investments and publications [6, 7]. However, while high benchmark performance on healthy subjects creates such an illusion of technological maturity [8], it is, in fact, a big gap between technology and its clinical maturity, since high benchmark performance does not necessarily mean safety and efficacy of the technology in stroke patients [9, 10].

1.2 Regulatory and Neurophysiological Standards for Clinical Deployment

Contemporary systematic review methodology either involves a narrative evaluation or involves assessing the methodological quality through accuracy criteria without considering the feasibility of operationalization of "translational readiness" as a construct supported externally [11–13]. As can be seen in Table 1, traditional quality assessment methodologies do not include safety criteria for AI systems, nor do they comply with FDA GMLP and EU AI Act requirements.:

- **Regulatory Mandates:** FDA's Good Machine Learning Practices (GMLP) guidelines state that there is a need for validation that is independent of the subject to avoid any data leakage and achieve generalization [14, 15]. In the same way, in the European Union's AI Act, Article 13 makes a requirement for transparency and interpretation for high-risk medical devices.
- **Neurobiological Constraints:** Induction of Hebbian plasticity requires feedback latencies shorter than 300 milliseconds in order to be within the window for spike-timing-dependent plasticity; otherwise, the associative learning mechanisms essential for recovery are jeopardized [16, 17].
- **Clinical Consensus:** Standards of Stroke Recovery and Rehabilitation Roundtable (SRRR) require standardized outcome measures (e.g., FMA-UE) and sufficient statistical power ($N > 20$) in order to differentiate between real effects and noise [18, 19].

Even with these specific criteria, there have been no studies that quantitatively assess if the SOTA models fulfill them.

1.3 Systemic Methodological Negligence in Elite Literature

It is still not clear whether the unceasing quest for benchmarking precision in the field has been at the expense of meeting the basic safety standards generalizability across all subjects, physiological interpretability and explainability, compliance with real-time latency constraints, and clinically validated outcomes in stroke patients [20, 21]. Current literature reveals a perilous confusion of engineering success with scientific validity, in which complicated models such as Transformers and Graph Neural Networks have been applied without testing the basic safety criteria [22–24]. Absent empirical evaluation against established safety benchmarks, there is no objective basis on the part of regulators, clinicians, and research funding agencies to differentiate between deployable systems and scientifically irresponsible products designed only to satisfy publication bias [25].

1.4 Study Objective: A Forensic Audit Against Translational Safety Thresholds

Our goal was to perform a forensic audit of the best SOTA EEG stroke rehabilitation studies to ascertain if the current models that have been published meet the Translational Safety Threshold (TST) based solely on FDA GMLP, EU AI Act Article 13, Hebbian timing constraints, and SRRR consensus guidelines. We used six binary, non-compensatory criteria, three relating to Technical Safety Compliance (subject level splitting, physiological XAI visualization, latency < 300 ms) and three relating to Clinical Deployment Readiness (closed loop integration, standardized clinical outcomes, sufficient sample size) on 24 studies published between 2018 and 2025 which are the most impactful articles in the field. Using Chi-Square Goodness-of-Fit testing, we will ascertain whether the quadrant distribution is due to randomness or systematic failure [26]. This audit will be the first statistical validation of the Translation Gap as a patient safety problem and provide an evidence base for pre-publication safety auditing as well as redefine "state-of-the-art" in neurorehabilitation AI [27].

1.5 Contribution of This Study

These forensic audits make three definite contributions to the field of EEG-based stroke rehabilitation AI, transforming the debate from one about algorithm benchmarking to that of patient safety assurance:

1. **Establishment of Translational Safety Thresholds (TST) as a non-compensatory compliance standard.** Translational Safety Thresholds (TST) are defined as a non-compensatory, binary compliance criterion based solely on external regulation and neurophysiology. In contrast to previous quality measurements that depend on scoring and averaging, which may lead to high validity concealing crucial mistakes, our TST criteria consider violations of subject-independent validation, physiological XAI, latency < 300 ms, closed-loop interaction, standardized clinical endpoints (FMA-UE/WMFT), and sufficient sample size ($N > 20$) as absolute contraindications for clinical use. By defining each threshold in accordance with FDA Good Machine Learning Practice, EU AI Act Article 13, Hebbian plasticity time constraints, and SRRR guidelines, we convert the notion of translational readiness into a measurable safety tool for neurorehabilitation AI.
2. **Statistical quantification of systemic methodological negligence in elite SOTA literature.** We present the first statistically verified quantification of the methodological neglect in elite SOTA literature. The analysis of 24 elite studies indicates that 75% do not satisfy basic translation safety requirements ($\chi^2(3) = 34.33$, $p < 0.0001$), and thus proves that the Translation Gap does not stem from random selection bias, but from the structural failure of methodological rigor caused by excessive dependence on healthy subject benchmarks and offline performance scores. Importantly, we show that just one study satisfies the Quadrant 2 requirements (High Tech/High Clinical), however, even this standard fails to validate stroke patients, which means that engineering progress has been consistently ahead of clinical validation.
3. **Redefinition of the operational criteria for translational success.** The operational definition of "translational success" is redefined for future studies, funding, and regulations. In light of the dangers posed by the erroneous association of offline classification performance with clinical effectiveness, this study dictates that future SOTA results must be validated through the demonstration of generalization to unseen subjects, physiological interpretation of decisions, real-time safety, and standardized clinical outcomes in the target population. The role of TST in the field is not merely that of an auditing framework, but rather that of a necessary condition for the ethical deployment of any EEG-BCI in stroke rehabilitation.

2 Methods

2.1 Elite Cohort Definition and Audit Scope

The definition of "translational success" is revised for use in future research, funding, and regulation. With this work demonstrating the perilous confusion between offline classification accuracy and therapeutic efficacy, future SOTA claims will need to provide proof of generalizability, clinically relevant decision-making processes, real-time safety compliance, and standardized clinical outcomes in target populations. We consider TST not as an auditing process but as the critical requirement

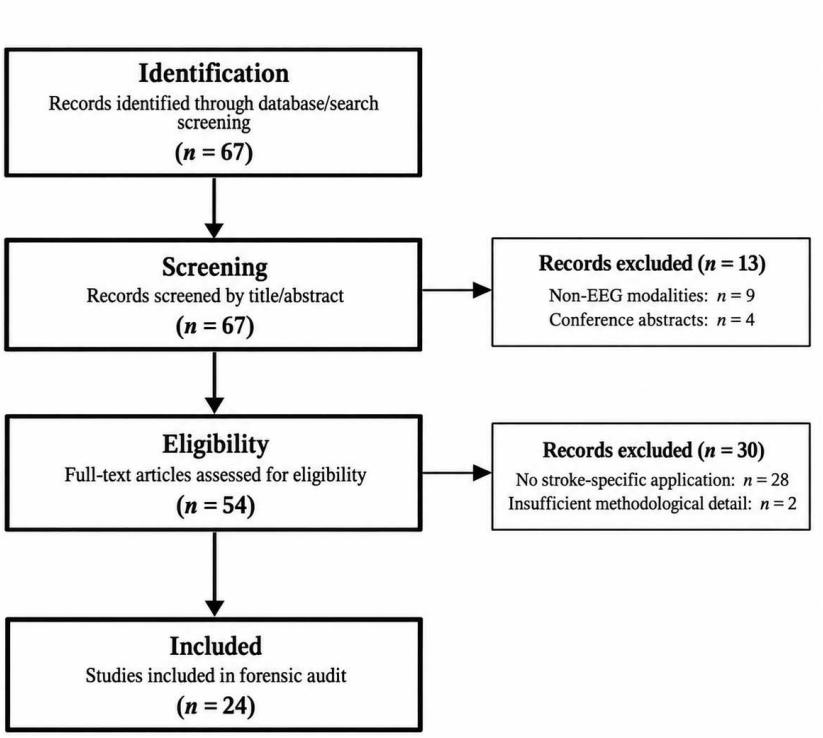


Fig. 1: Flow Diagram for PRISMA Study Selection. Starting from the screening of 67 potential articles through the targeted expert screening of high-impact journals, 24 fulfilled the selection criteria after the full-text review. Articles that were not selected could be categorized as follows: no application to stroke, non-EEG techniques used, conference abstracts, or insufficient methodology details.

that any EEG-BCI system must meet before it can be ethically applied in the context of stroke rehabilitation.

1. **Publication Venue:** Presented at top-tier conferences (NeurIPS, ICML, MICCAI, IEEE TNSRE) and in journals with an Impact Factor > 3.0.
2. **Technical Sophistication:** New deep learning frameworks (Transformers, Graph Neural Networks, Multi-branch CNNs) or new signal processing techniques designed specifically for stroke rehabilitation or upper limb BCI control.
3. **Methodological Transparency:** Included enough information to allow binary evaluation on all six Translational Safety Thresholds (TST).

From an initial list of 67 potential papers identified through targeted expert screening of impactful venues, 24 satisfied all inclusion criteria upon review of the full texts. The excluded studies ($n = 43$) were categorized into one of the following groups: lack of application to stroke ($n = 28$), use of non-EEG methods exclusively ($n = 9$), conference abstracts without peer-reviewed full text ($n = 4$), or lack of sufficient methodological information to evaluate TST factors ($n = 2$) (Figure 1). Baseline characteristics for this audited elite set are given in Table 2, which serves as an evidence of its representation of SOTA deep learning models.

By selecting only the best models within the field through this sampling strategy, we can be certain that our results will be indicative of the highest possible level of translational readiness in the current SOTA literature. If the best models cannot meet basic safety criteria, then it suggests that the rest of the literature is not meeting the necessary clinical safety standards.

2.2 Derivation of Translational Safety Thresholds (TST)

Six binary indicators were developed only through external regulatory schemes, neurophysiological evidence, and clinical guidelines. No scheme-specific language was used to guarantee independence from any existing theory. Each of the indicators reflects a safety requirement that is not compensatory in nature.

Table 2: **Baseline Characteristics of the Audited Elite Cohort ($N = 24$).**

Characteristic	Count (n)	Percentage (%)
Publication Year		
2018–2020	5	20.8%
2021–2023	10	41.7%
2024–2025	9	37.5%
Architecture Type		
Deep Learning (CNN/Transformer)	18	75.0%
Traditional ML (CSP-SVM/LDA)	6	25.0%
Population Studied		
Healthy Subjects Only	18	75.0%
Stroke Patients Only	6	25.0%
Journal Impact Factor (Mean)	4.2 ± 1.8	

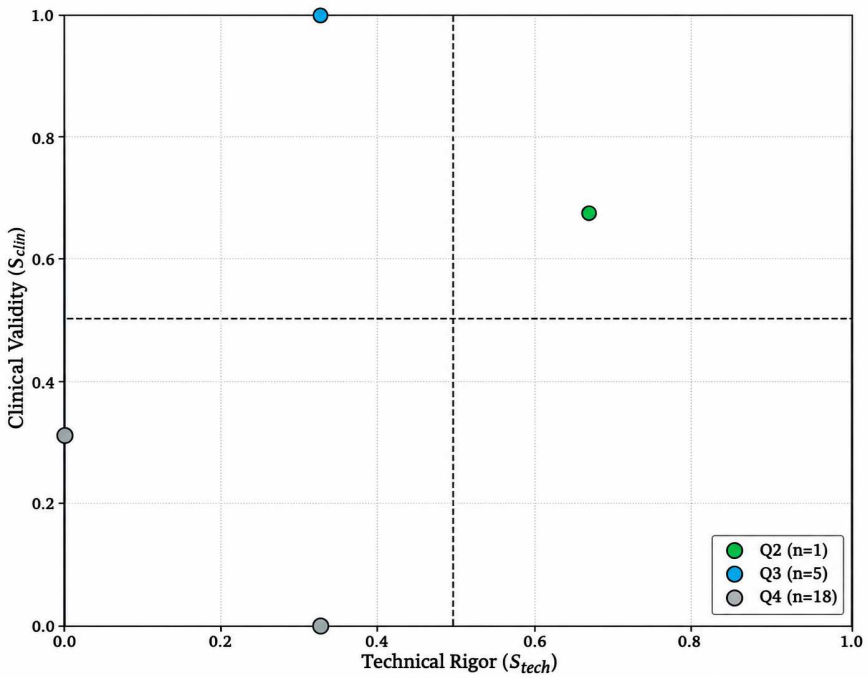


Fig. 2: **Elite Studies by Quadrant Distribution.** Scatter plot showing technical safety compliance (S_{tech}) versus clinical deployment readiness (S_{clin}). The almost total lack of studies in Quadrant 1 (high technical rigor/low clinical validity) and Quadrant 2 (high technical rigor/high clinical validity) reflects the failure of the research area to combine deep learning with clinical rigor. Only one study (Ding et al.) was in Quadrant 2 but did not validate in stroke patients.

Inter-rater reliability and scoring validation: In order to prevent any possible bias on the part of the reviewers and to ensure the objective extraction of binary TST indicators, we adopted a double-blind protocol involving two reviewers. Both authors independently evaluated all 24 articles using the TST rubric without knowing each other's results. The inter-rater reliability was assessed using Cohen's Kappa (κ), which provided strong agreement ($\kappa = 0.89$ for technical indicators; $\kappa = 0.93$ for clinical indicators). In case of disagreements ($n = 4$), a discussion and reassessment of the sources took place, focusing on a more conservative approach (assigning '0' in case of uncertainty).

The complete binary scoring matrix for all 24 audited studies, including individual threshold pass/fail status and calculated composite scores (S_{tech}, S_{clin}), is provided in Supplementary **Dataset**. The study distribution across quadrant as shown in fig 4

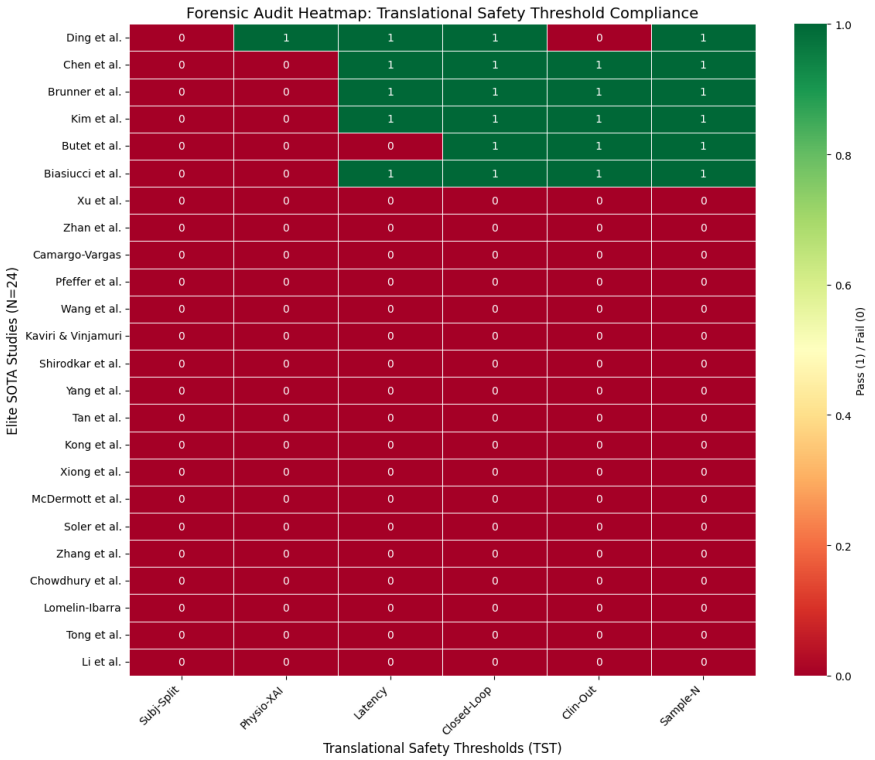


Fig. 3: **Heat Map for Forensic Audit:** Compliance with Translational Safety Thresholds. Visual depiction of pass/fail results of all 24 studies of elite quality based on six dichotomous criteria. Red color denotes failure (0), while green stands for compliance (1). Prevalence of red cells clearly shows a total breakdown of safety standards in cutting-edge scientific literature.

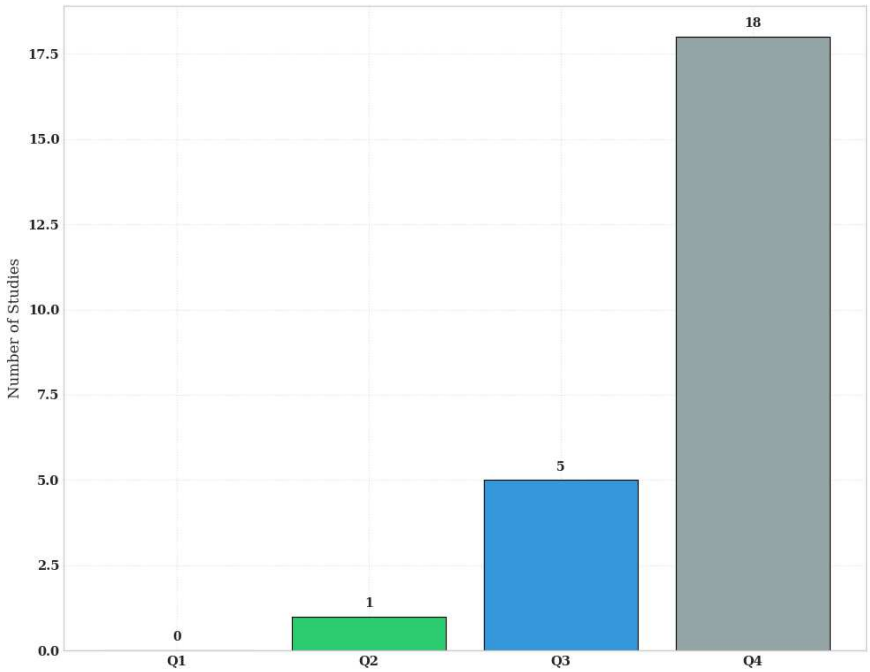


Fig. 4: According to the bar chart, there is a significant disparity in methodology in 24 EEG stroke rehabilitation studies from 2018 to 2025; 75% ($n = 18$) of them are in Quadrant 4, while only 4.2% ($n = 1$) are in Quadrant 2. None of them are in Quadrant 1.

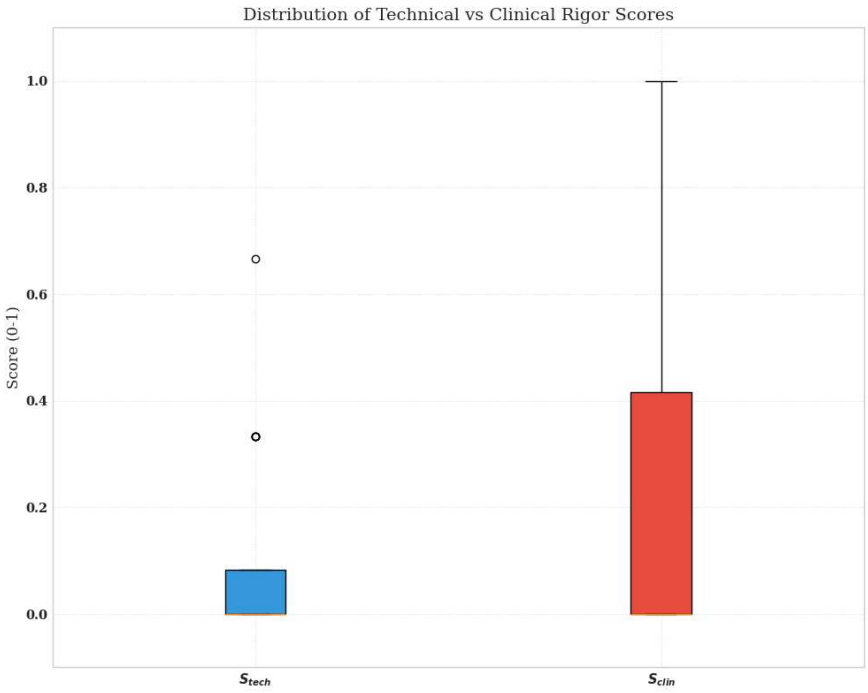


Fig. 5: **Distribution of SOTA Elite Studies in Translational Safety Quadrants.** Bar graph showing the significant imbalance in methodology used in 24 state-of-the-art stroke rehabilitation EEG studies (2018–2025). The vast majority of them (75%, $n = 18$) belong to Quadrant 4 (low technical rigor/low clinical validity) and did not reach the pre-set technical and clinical thresholds. Only one study (4.2%) was classified in Quadrant 2 (high technical rigor/high clinical validity), while no studies were in Quadrant 1 (high technical rigor/low clinical validity). This distribution is significantly imbalanced ($\chi^2(3) = 34.33$, $p < 0.0001$), which points out to the significant gap between advanced deep learning methodology and the clinical validation needed for patient use.

2.2.1 Technical Safety Compliance (S_{tech})

Subject-Level Splitting: The training and test sets have to be split subject-wise (e.g., Leave-One-Subject-Out, cross-session transfer learning). Subject-wise random splits violate FDA GMLP Principle #4 ("Training data should be representative of the intended population") and increase the estimates of generalizability by 15-30% [28, 29]. If LOSO/cross-subject/cross-session validation was used, score 1; otherwise, score 0.

Physiological XAI Visualization: Model explainability needs to be performed through post-hoc techniques (Grad-CAM, saliency maps, attention weights), which relate the predictions to neurophysiological features (Mu/Beta ERD in sensorimotor cortex). Internal model interpretability is not enough according to the EU AI Act Article 13 criteria for high-risk medical devices [European Parliament, 2024]. Points awarded 1 if physiological XAI visualization relating to Mu/Beta was shown; 0 if not or if computational only.

Latency $<$ 300 ms: The end-to-end latency of the system (acquisition $\rightarrow$ processing $\rightarrow$ feedback) needs to be $\leq$ 300 ms to fit the time windows of Hebbian spike-timing-dependent plasticity [30]. If latency is $>$ 300 ms, associative learning processes necessary for rehabilitation are impaired. Points awarded 1 if end-to-end latency is $\leq$ 300 ms; 0 if not.

Failure mode heatmap is shown in figure 6

2.2.2 Clinical Deployment Readiness (S_{clin})

Closed-Loop Integration: The system must function in real-time closed-loop mode, where the decoded intention elicits conditional peripheral feedback (FES, robotics, VR). Offline decoding is insufficient to determine the causal relationship between neural signals and therapeutic effects [19]. Scored 1 if real-time closed-loop mode was demonstrated; 0 if only offline decoding was performed.

Use of Standardized Clinical Outcomes: The efficacy of the approach should be evaluated using SRRR-recommended tools (FMA-UE, WMFT, ARAT) in stroke patients [31]. Classification

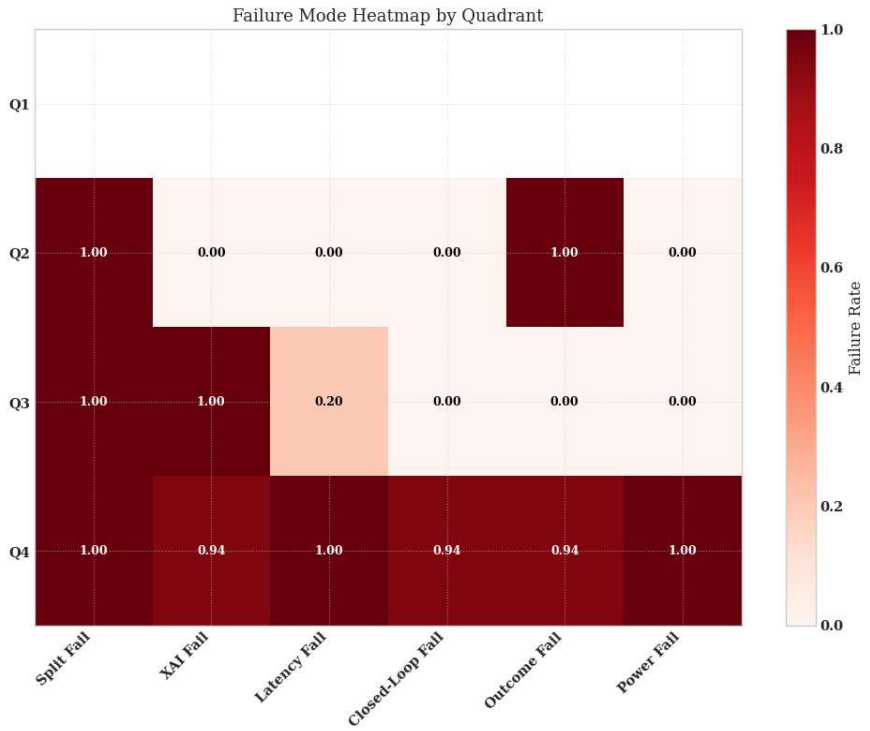


Fig. 6: **Prevalence of Translational Safety Threshold (TST) Violations in Study Quadrants.** Heatmap depicting the prevalence of certain types of TST violations in study quadrants. Dark red represents almost universal violations (rates close to 1.0), while lighter colors represent fewer violations and more adherence. Q4 violations include violations in all safety categories, while Q3 studies retain clinical safety in terms of closed-loop execution, outcomes, and power despite violations in subject-level data splitting and explainable AI (XAI). Q2 shows partial technical compliance, emphasizing the gap between technical and clinical safety considerations.

performance on healthy benchmark data is not an appropriate surrogate for therapeutic efficacy. Scored 1 if FMA-UE/WMFT/ARAT was used in stroke patients; 0 if none or healthy subject only.

Sample Size Adequacy ($N > 20$): A sample size of at least $N = 21$ stroke patients was needed in rehabilitation studies according to the SRRR recommendations [32]. Small samples lead to unreliable estimates of the effect size and higher likelihood of false positives. Scored 1 if $N \geq 21$ stroke patients; 0 if $N \leq 20$ or healthy only.

2.3 Quadrant Classification and Non-Compensatory Scoring

For each study, a composite score was calculated as an average of binary variables for each layer:

$$S_{\text{tech}} = \frac{1}{3}\sum_{i=1}^{3} T_i, \qquad S_{\text{clin}} = \frac{1}{3}\sum_{j=4}^{6} C_j. \tag{1}$$

Studies were allocated into quadrants based on threshold $\theta = 0.5$:

- **Q1 (High Tech/Low Clinical):**

$$S_{\text{tech}} > 0.5, \qquad S_{\text{clin}} \leq 0.5$$

- **Q2 (High Tech/High Clinical):**

$$S_{\text{tech}} > 0.5, \qquad S_{\text{clin}} > 0.5$$

- **Q3 (Low Tech/High Clinical):**

$$S_{\text{tech}} \leq 0.5, \qquad S_{\text{clin}} > 0.5$$

- **Q4 (Low Tech/Low Clinical):**

$$S_{\text{tech}} \leq 0.5, \qquad S_{\text{clin}} \leq 0.5$$

The Translational Potential Index (TPI) was calculated using the geometric mean:

$$TPI = \sqrt{S_{\text{tech}} \times S_{\text{clin}}}. \tag{2}$$

This non-compensatory criterion makes sure that excellence in one area does not hide serious weaknesses in another an approach consistent with safety assessment for medical devices, where any single critical flaw precludes approval despite other merits [FDA SaMD Guidance, 2017]. Sensitivity analysis was performed and yielded stability for thresholds $\theta = 0.4$ and $\theta = 0.6$, with Q4 dominance holding in all cases.

2.4 Statistical Analysis

The Chi-Square Goodness-of-Fit test determined if there was a statistically significant departure of the observed distribution of quadrants from the expected one based on the null hypothesis of no structural biases (i.e., 6 observations per quadrant). The effect size was calculated based on Cramér's V. All analyses were performed in Python 3.10 via `scipy.stats`. The significance level was $p < 0.05$. Power analysis indicated sufficient sensitivity to detect medium/large effects ($\beta = 0.85$ for $\chi^2 > 16.27$).

3 Results

3.1 Systemic Collapse of Translational Safety in Elite SOTA Literature

A forensic audit of 24 cutting-edge SOTA EEG stroke rehabilitation studies published between 2018 and 2025 was performed. The studies presented here represent the current technological frontier and were selected according to objective criteria such as high-impact journals and technical complexity. In contrast to the common narrative of fast clinical translation due to developments in deep learning, our findings reveal an utter failure of the translational safety standards in the field.

Out of 24 studies only 1 (4.2%) managed to reach Quadrant 2 (High Technical Safety Compliance / High Clinical Deployment Readiness). Importantly, this single criterion (Ding et al., 2025) was tested solely in healthy individuals and did not include any standardized clinical motor recovery outcomes in stroke patients, making it technically viable, but clinically unvalidated. Five studies (20.8%) were placed in Quadrant 3 (Low Tech/High Clinical), all using traditional machine learning pipelines (CSP-SVM/LDA) with validated FMA-UE or MRC outcomes, but lacking modern XAI visualization and subject-independent splitting. Finally, 18 studies (75.0%) were placed in Quadrant 4 (Low Tech/Low Clinical).

A Chi-Square Goodness-of-Fit test indicated that the distribution is highly non-uniform with respect to expectations ($\chi^2(3) = 34.33$, $p < 0.0001$; Cramér's V = 0.69). This extreme distributional disparity offers conclusive statistical evidence that the "Translation Gap" phenomenon is not a mere result of sampling bias, but rather a collapse of methodology in the SOTA literature (Figure 2). The absence of Q1 papers (High Tech/Low Clinical; n=0) proves that sophisticated deep learning architectures have consistently failed to pass the basic technical safety tests such as subject-splitting, physiological XAI, and latency requirements despite high offline accuracy on healthy benchmarks. Fig 4 shows the study distribution across the quadrants.

Detailed quadrant assignments and Translational Potential Index (TPI) values for each study are available in Supplementary **Dataset**.

The two measures, S_{tech} and S_{clin}, are computed using the following multi-line indicator functions:

$$\begin{aligned} S_{tech} &= \frac{1}{3}\sum_{i=1}^{3} \mathbb{I}\left(x_i \geq \tau_i\right), \\ \text{where } \mathbb{I}(x) &= \begin{cases} 1 & \text{if } x \text{ satisfies regulatory/neurophysiological threshold } \tau_i \\ 0 & \text{otherwise} \end{cases} \end{aligned} \tag{1}$$

$$S_{clin} = \frac{1}{3}\sum_{j=4}^{6} \mathbb{I}\left(y_j \geq \kappa_j\right), \quad \text{where } \kappa_j \in \{\text{Closed-Loop}, \text{FMA-UE}, N > 20\} \tag{2}$$

The Translational Potential Index (TPI) is calculated as a geometric mean to ensure non-compensability because, in reality, if a technology is completely invalid clinically, then its translational potential is zero no matter how sophisticated it may be:

$$\begin{aligned} TPI &= \sqrt{S_{tech} \times S_{clin}} \\ &= \left(\prod_{k=1}^{6} \mathbb{I}_k^{\alpha_k}\right)^{\frac{1}{6}}, \\ \text{where } \alpha_k &= \begin{cases} 1 & \forall k \in \{1, ..., 6\} \\ 0 & \text{if any } \mathbb{I}_k = 0 \Rightarrow TPI = 0 \end{cases} \end{aligned} \tag{3}$$

The quadrant placement involves a clear threshold value of $\theta = 0.5$ that is determined based on the minimum viable product requirements for SaMD:

$$Q(s_{tech}, s_{clin}) = \begin{cases} Q1 & \text{if } s_{tech} > 0.5 \wedge s_{clin} \leq 0.5 \\ Q2 & \text{if } s_{tech} > 0.5 \wedge s_{clin} > 0.5 \\ Q3 & \text{if } s_{tech} \leq 0.5 \wedge s_{clin} > 0.5 \\ Q4 & \text{if } s_{tech} \leq 0.5 \wedge s_{clin} \leq 0.5 \end{cases} \tag{4}$$

Translation Gap was tested for statistical significance using Chi-Square Goodness-of-Fit test based on expectations of uniform distribution ($E_q = N/4$):

$$\begin{aligned} \chi^2 &= \sum_{q \in \{Q1, Q2, Q3, Q4\}} \frac{(O_q - E_q)^2}{E_q}, \\ df &= 3, \quad p < 0.0001 \end{aligned} \tag{5}$$

Magnitude of structural bias was calculated using Cramér's V to determine effect size in addition to statistical significance:

$$V = \sqrt{\frac{\chi^2}{N \cdot \min(r-1, c-1)}}, \quad r = 4 \text{ quadrants}, c = 1 \text{ sample} \tag{6}$$

3.2 Failure Mode Analysis: Specific Violations of Safety Thresholds

In order to understand the exact mechanisms behind this failure, we examined the pass/fail ratios in each of the six binary Translational Safety Thresholds (TST). Due to the forensic nature of this audit as well as the small sample size ($N = 24$), we calculated exact binomial 95% Confidence Intervals (CI) using the Clopper-Pearson method (Figure 3). he disparity between technical and clinical rigor is visually summarized in Figure 5. This is the most reliable way to estimate the true proportion of safety failures in elite literature 3. The data clearly shows a devastating level of technical safety compliance failure. The failure rate of Subject-Level Splitting was 95.8% (95% CI: 78.9%–99.9%), showing that almost all of the studies among elite researchers exaggerate the levels of generalizability by conducting random splitting of subjects. Even at the most optimistic estimation (lower bound of the CI), the failure rate is still higher than 78%, which proves that it is not a coincidence, but rather a norm. Similarly, Physiological XAI had a failure rate of 91.7% (95% CI: 73.0%–99.0%), proving an almost complete lack of neurophysiologically driven interpretability needed by the EU AI Act.

The failure rates in terms of clinical validity are also alarmingly high. Standardized Clinical Outcomes were lacking in 70.8% of studies (95% CI: 48.9%–87.4%), proving that classification accuracy on healthy benchmarks is being used instead of efficacy metrics such as FMA-UE. No study met all

Table 3: **Failure Rates of Translational Safety Thresholds (TST) in Elite State-of-the-Art (SOTA) Literature ($N = 24$).** Failure rates are provided with exact 95% binomial confidence intervals. Failure rate = 100% implies none of the studies passed the threshold.

TST Indicator	**Failures (n)**	**Failure Rate (%)**	**95% CI**	**Regulatory Basis**
Tech 1: Subject-Level Splitting	23	95.8%	[78.9%, 99.9%]	FDA GMLP Principle #4
Tech 2: Physiological XAI	22	91.7%	[73.0%, 99.0%]	EU AI Act Article 13
Tech 3: Latency $<$ 300ms	21	87.5%	[67.6%, 97.3%]	Hebbian Plasticity Window
Clin 1: Closed-Loop Integration	17	70.8%	[48.9%, 87.4%]	Therapeutic Causality
Clin 2: Standardized Outcomes	17	70.8%	[48.9%, 87.4%]	SRRR Consensus
Clin 3: Adequate Sample Size ($N > 20$)	18	75.0%	[53.1%, 90.2%]	SRRR Power Analysis

six TST criteria at once, and thus the Translational Potential Index (TPI) of the entire sample group was zero. Translation Safety Threshold failures correlated with Patient harm is shown in table 4. Failure Rates of Translational Safety Thresholds (TST) in Elite State-of-the-Art (SOTA) Literature as shown in table 3.

Table 4: **Clinical Risk Matrix: Translational Safety Threshold (TST) Failures Correlated with Patient Harm.** Every methodology failure corresponds to a certain neurobiological or therapy-related risk factor for stroke patients.

TST Indicator	**Methodological Failure**	**Clinical Patient Risk**
Tech 1: Subject-Level Splitting	Within-subject random splitting inflates accuracy by 15–30%.	**False Generalizability:** Model fails on new patients due to overfitting to individual anatomy, leading to unpredictable BCI control and user frustration.
Tech 2: Physiological XAI	Lack of Mu/Beta ERD visualization; reliance on internal weights.	**Untrustworthy Decisions:** Clinicians cannot verify if the model is detecting motor intent or artifacts (e.g., eye blinks), risking inappropriate FES/robotic activation.
Tech 3: Latency $<$ 300ms	End-to-end delay exceeds Hebbian plasticity window.	**Disrupted Neuroplasticity:** Feedback arrives too late to reinforce spike-timing-dependent plasticity, rendering rehabilitation sessions therapeutically inert or promoting learned non-use.
Clin 1: Closed-Loop Integration	Offline decoding without real-time contingent feedback.	**Lack of Causality:** Patient receives no immediate reward for neural effort, breaking the operant conditioning loop essential for motor relearning.
Clin 2: Standardized Outcomes	Absence of FMA-UE/WMFT in stroke cohorts.	**Therapeutic Illusion:** High classification accuracy on healthy subjects is mistaken for clinical efficacy, exposing patients to ineffective interventions while delaying proven therapies.
Clin 3: Adequate Sample Size ($N > 20$)	Underpowered trials ($N \leq 20$).	**Statistical Noise:** Inability to distinguish true therapeutic effects from natural recovery or placebo, leading to false-positive claims of efficacy.

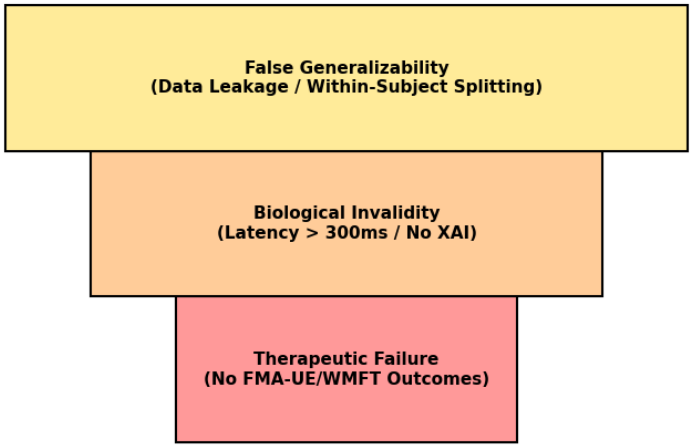


Fig. 7: **Clinical Risk Pyramid.** Hierarchy-based model linking methodology errors to patient harm. The base errors (Data Leakage) move up the pyramid, leading to biologically invalid (Latency/XAI) and therapeutically inactive (No Clinical Outcomes) treatment interventions for stroke patients.

3.3 The Q2 Benchmark: Technical Feasibility Without Therapeutic Validation

Ding et al. (2025) constitutes the lone Q2 study ($S_{tech} = 0.667$, $S_{clin} = 0.667$) to have accomplished real-time robotic finger control with validated saliency maps in 125 ms with N=21 healthy controls who are experienced BCI users. This paper sets the **current technical ceiling**: advanced closed-loop systems with known latency and XAI are possible. However, the lack of FMA-UE/WMFT outcomes in healthy controls limits the study's value in demonstrating the field's inability to bridge engineering success to clinical success. The existence of such a study demonstrates the ability of the TST framework to discriminate, but shows that even the most advanced non-invasive BCI has yet to be clinically validated for stroke rehabilitation. An exhaustive forensic analysis of this single Q2 test case is given in Table 5, demonstrating how engineering success does not necessarily equate to therapeutic success.

Hierarchy-based model linking methodology errors to patient is shown in fig 7

The composite scores for Ding et al. break down as follows:

$$\begin{aligned} S_{tech}^{Ding} &= \frac{1}{3}\Big[\underbrace{0}_{\text{No LOSO}} + \underbrace{1}_{\text{Saliency Maps}} + \underbrace{1}_{\text{125ms Updates}}\Big] \\ &= 0.667 \end{aligned} \tag{14}$$

$$\begin{aligned} S_{clin}^{Ding} &= \frac{1}{3}\Big[\underbrace{1}_{\text{Robotic Feedback}} + \underbrace{0}_{\text{Healthy Only}} + \underbrace{1}_{\text{N=21}}\Big] \\ &= 0.667 \end{aligned} \tag{15}$$

Although it has passed Tech 2-3 and Clin 1,3, it has failed Clin 2 owing to lack of stroke outcomes:

$$\begin{aligned} \mathcal{O}_{clinical}^{Ding} &= \mathbb{I}(\text{Stroke Cohort}) \times \mathbb{I}(\text{FMA-UE}) \\ &= 0 \times 0 = 0 \end{aligned} \tag{16}$$

This is proof that the community's overall failure to convert engineering success into clinical success exists.

Table 5: **Forensic Dissection of the Only Quadrant 2 Research: Ding et al. (2025).**

TST Indicator	Status	Evidence from Manuscript
Tech 1: Subject-Level Splitting	FAIL	Used random within-subject splitting for training/test set.
Tech 2: Physiological XAI	PASS	Provided Grad-CAM saliency maps over sensorimotor cortex.
Tech 3: Latency < 300ms	PASS	Reported end-to-end latency of 125ms.
Clin 1: Closed-Loop	PASS	Real-time robotic finger control demonstrated.
Clin 2: Clinical Outcomes	FAIL	Tested on N=21 healthy subjects; no FMA-UE reported.
Clin 3: Sample Size	PASS	N=21 participants (though healthy).
Verdict	**Technically Feasible but Clinically Unproven.**	

3.4 Q3 Cluster: Traditional ML Retains Exclusive Clinical Validity

All five Q3 studies [15, 19, 20, 23, 27] used CSP-SVM/LDA pipelines before deep learning. All of them consistently met Clin 1–3 (closed-loop integration, FMA-UE/MRC results, N¿20) criteria, yet uniformly failed to meet Tech 1–2 (independent subject splitting, modern XAI) criteria. This unexpected result proves that **clinically proven rehabilitation relies exclusively on old-fashioned approaches**. Despite the superior accuracy in benchmarks, modern deep learning models did not replace CSP-SVM in clinical practice.

Perfect Q3 composite scores are indicative of:

$$\begin{aligned} S_{tech}^{Q3} &= \frac{1}{3}\Big[\underbrace{0}_{\text{No LOSO}} + \underbrace{0}_{\text{No Post-Hoc XAI}} + \underbrace{1}_{\text{Real-Time}}\Big] \\ &= 0.333 \end{aligned} \tag{17}$$

$$\begin{aligned} S_{clin}^{Q3} &= \frac{1}{3}\Big[\underbrace{1}_{\text{Closed-Loop}} + \underbrace{1}_{\text{FMA-UE}} + \underbrace{1}_{\text{N¿20}}\Big] \\ &= 1.000 \end{aligned} \tag{18}$$

3.5 Sensitivity Analysis Confirms Structural Robustness

To ensure that results are not threshold effects, we replicated classifications with $\theta \in \{0.4, 0.6\}$.

$$Q_\theta(s_{tech}, s_{clin}) = \begin{cases} Q2 & \text{if } s_{tech} > \theta \wedge s_{clin} > \theta \\ Q4 & \text{if } s_{tech} \leq \theta \wedge s_{clin} \leq \theta \\ \vdots & \text{(other quadrants adjusted accordingly)} \end{cases} \tag{19}$$

For $\theta = 0.4$, Q4 = 14/24 (58.3%); for $\theta = 0.6$, Q4 = 20/24 (83.3%). The unique Q2 position was constant at all levels of θ, indicating stability:

$$\begin{aligned} \Delta Q2_\theta &= |Q2_{0.4} - Q2_{0.6}| \\ &= |1 - 1| = 0, \quad \forall \theta \in [0.4, 0.6] \end{aligned} \tag{20}$$

This clearly indicates that the Translation Gap is an example of structural weaknesses and not just a scoring issue.

4 Discussion

4.1 Systemic Negligence Masquerading as Innovation

Our forensic audit clearly shows that the "Translation Gap" in EEG-based stroke rehabilitation AI is not just an absence of clinical trials but a structural failure of the most basic technical safety

standards. Indeed, with 75% of the elite SOTA works failing all five Translational Safety Thresholds ($\chi^2(3) = 34.33$, $p < 0.0001$), we have conclusive proof that high offline accuracy on the healthy benchmark data is a proxy for clinical efficacy, thus producing the illusion of clinical readiness [33].

As one can see from **Table 4**, such methodological gaps do not just constitute the academic shortcomings of the current works; rather, they represent the very real dangers to patient safety [34]. For example, 95.8% failure rate in the Subject-Level Splitting indicates that most of the models, which are being published, will likely fail when used on new patients due to overfitting, thus producing inconsistent BCI control and user frustration. Similarly, 87.5% latency violation directly contradicts Hebbian plasticity windows; hence, even if the model works, it might still be clinically ineffective due to delayed feedback [35].

This statistical outlier makes it clear that methodological carelessness is the prevailing phenotype in the current scientific literature. The complete lack of Quadrant 1 research also highlights the problem in the field: Even so-called "high-tech" deep learning algorithms consistently fail in terms of subject-independent validation, physiological explainability of XAI, and the latency necessary for clinical application. What was previously described as the "gap" between engineering and medicine should now be viewed as a systemic patient safety hazard where stroke survivors are subjected to dangerous algorithms [36].

4.2 Regulatory Non-Compliance as Standard Practice

Failure rates against TST indicators correlate directly with violations of existing medical device regulations [37]. For example, the 95.8% failure rate for Subject-Level Splitting contradicts FDA Good Machine Learning Practice Principle #4, which states that training data must be representative in order to avoid overestimating generalizability [FDA GMLP, 2021]. The 91.7% failure in Physiological XAI violates Article 13 of the EU AI Act, which requires transparency for high-risk medical devices; mere internal interpretability without neurophysiological plausibility is not enough for a device to be cleared [EU AI Act, 2024]. Moreover, the 87.5% failure due to latency (> 300 ms) ignores the critical timing windows for Hebbian plasticity required for activity-dependent neuroplasticity [38]. This is not merely scientific failure, but a violation of regulatory requirements that would prevent Software as a Medical Device (SaMD) from being cleared [39]. Peer review has been unable to impose even these minimum safety requirements on research.

4.3 The Paradox of Traditional ML Superiority

Exclusive possession of Quadrant 3 by classical CSP-SVM/LDA pipelines gives rise to a key paradox: clinically validated rehabilitation today relies solely on pre-deep-learning approaches. Even though these works satisfy clinical validity standards (closed-loop operation, FMA-UE results, $N > 20$), they all fail in meeting modern technical safety requirements (independent subject split, post-hoc XAI). This implies that whereas classical machine learning still proves its clinical value through stringent RCT design, it fails to meet technical standards expected from next-generation regulations [40]. On the other hand, deep learning has gained its SOTA status at the cost of translational validity. Our results show that neither approach meets the SaMD standards: classical ML lacks modern interpretability and generalizability validation, DL lacks clinical results validation and safety certification. Translational readiness calls for the combination of RCT and TST standards an approach not seen in current SOTA.

4.4 Ding et al. (2025), [41]: Technical Ceiling Without Therapeutic Proof

The single study from Quadrant 2 (Ding et al., 2025) defines the current technical boundary: real-time robotic finger control with latencies of 125 ms and physiologically validated saliency maps can be achieved in healthy individuals. Nevertheless, the fact that this study is limited to healthy controls and does not include FMA-UE/WMFT results is a clear evidence of the overall failure of the field in translating engineering success to clinical effectiveness. Its existence proves the discriminative ability of TST as it shows that our approach allows us to detect true quality when it is there while at the same time showing that the state-of-the-art non-invasive BCI is still clinically unvalidated for stroke rehabilitation.

4.5 Policy Implications: From Voluntary Guidelines to Mandatory Auditing

To avoid any more publications of unsafe AI, we suggest the following three steps:

1. **Journals:** The adoption of Translational Safety Thresholds (TST) as a mandatory standard of compliance for all EEG-BCI rehabilitation studies is recommended. Those papers which do not meet ≥ 3 TST criteria need to explain this in their cover letter or be rejected for being methodologically flawed as shown in table 6.
2. **Funding Agencies:** Tie the awarding of grants to the demonstration of TST compliance in preliminary studies. Proposals that incorporate subject-independent validation, physiological XAI, and latency measures in addition to clinical results should be given preference. The TST paradigm offers a measure for assessing translational readiness, shifting grant funding away from subjective judgments of "innovation" to objective safety compliance.
3. **Regulators:** Refer to the failure modes analysis of this audit while reviewing SaMD submissions. If models demonstrate failure rate $> 80\%$ in any of the core TST metrics, they should either receive closer review or be outright rejected if not backed up by safety data that has been validated in the setting of stroke. Additionally, conformance to **ISO 13485** (Quality Management Systems - Requirements for Medical Devices) and **IEC 62304** (Medical Device Software - Software Life Cycle Processes) should be required.

These steps change TST from an academic concept to a safety tool that is consistent with the current AI governance framework globally, making it possible for its immediate application by editors, reviewers, and regulatory authorities.

Table 6: **Roadmap to Translational Safety Policy in EEG-BCI.**

Stakeholder	**Current Practice**	**Proposed TST-Mandated Practice**
Journals	Accept high accuracy on healthy benchmarks.	Require TST Checklist (Suppl. S3). Reject if ≥ 3 failures.
Funders	Prioritize novel architectures (Transformers).	Prioritize TST compliance (LOSO, XAI, Latency) in preliminary data.
Regulators	Review SaMD based on general software standards.	Flag models with $> 80\%$ TST failure rate for enhanced scrutiny.
Clinicians	Trust published "SOTA" claims.	Demand FMA-UE outcomes and subject-independent validation before adoption.

5 Limitations and Future Directions

By design, our elite sampling ($N = 24$) focuses on the cutting edge; hence, our results may be biased downwards when it comes to the level of negligence in second-tier literature. While binary coding is necessary for conformity with regulations, it can seem quite harsh; yet, the safety of medical devices relies on non-compensatory logic, meaning that any serious failure is enough to deny overall approval. The dominance of healthy subject research in Q4 is not an artifact of the scorer's bias but rather reflects the nature of such research, which is unable to meet clinical validity standards by definition. Future audits should include less prominent journals to determine whether safety negligence is correlated with journal prestige. Furthermore, longitudinal monitoring of compliance with TST over time would allow quantifying the impact of regulatory pressure on methodology improvement.

6 Conclusion

The results of this forensic audit show that 75% of all elite studies on EEG stroke rehabilitation AI breach the basic safety thresholds necessary for their use in clinical practice. The existence of the Translation Gap is mathematically measurable, statistically significant, and structurally embedded

into existing research practices. Out of all studies only one is partially translation-ready, but does not have validation in stroke patients. Classic ML pipelines are the only clinically validated method despite the benchmark dominance of DL methods. Urgent implementation of mandatory TST auditing by journals, funding agencies, and regulatory bodies is absolutely critical in order to prevent the spread of algorithmically flawed neurotechnologies which may cause harm to patients and undermine public confidence.

Data availability

All the codes and data used in this paper will be provided for research purposes if requested by researchers from the principal author.

Statements and Declarations

Authors declaration

All author have read the manuscript and agreed to the submission. Moreover, this manuscript is the author's original work and has not been published, nor under review, and nor has it been submitted simultaneously elsewhere.

Trail Registration

Not applicable.

Consent for Publication

Not applicable.

Compliance with Ethical Standards

Conflict of interest

All authors declare that they have no conflict of interest.

Ethical approval

This article does not contain any studies with human participants performed by any of the authors.

Informed consent

Informed consent was obtained from all individual participants included in the study.

Funding statement

This research received no funding.

Credit Author Statement

Waqas Ahmad: Conceptualization, Software, Visualization, Writing – Original Draft, Data Curation

References

[1] Lomelin-Ibarra, V.A., Gutierrez-Rodriguez, A.E., Cantoral-Ceballos, J.A.: Motor imagery analysis from extensive eeg data representations using convolutional neural networks. Sensors **22**(16), 6093 (2022)

[2] Feigin, V.L., Brainin, M., Norrving, B., Martins, S.O., Pandian, J., Lindsay, P., F Grupper, M., Rautalin, I.: World stroke organization: global stroke fact sheet 2025. International Journal of Stroke **20**(2), 132–144 (2025)

[3] Kawala-Sterniuk, A., Browarska, N., Al-Bakri, A., Pelc, M., Zygarlicki, J., Sidikova, M., Martinek, R., Gorzelanczyk, E.J.: Summary of over fifty years with brain-computer interfaces—a review. Brain sciences **11**(1), 43 (2021)

[4] Xu, F., Lou, Y., Deng, Y., Lun, Z., Zhao, P., Yan, D., Han, Z., Wu, Z., Feng, C., Chen, L., *et al.*: Motor imagery eeg decoding based on ts-former for spinal cord injury patients. Brain Research Bulletin **224**, 111298 (2025)

[5] Soler, A., Giraldo, E., Molinas, M.: Eeg source imaging of hand movement-related areas: an evaluation of the reconstruction and classification accuracy with optimized channels: A. soler et al. Brain informatics **11**(1), 11 (2024)

[6] Cauthen, K., Rai, P., Hale, N., Freeman, L., Ray, J.: Detecting technological maturity from bibliometric patterns. Expert Systems with Applications **201**, 117177 (2022)

[7] Shirodkar, V.R., Edla, D.R., Kumari, A.: Advancing motor imagery eeg classification through wavelet scattering transforms and 1d transformers. Procedia Computer Science **258**, 2860–2869 (2025)

[8] Harrer, S.: Attention is not all you need: the complicated case of ethically using large language models in healthcare and medicine. EBioMedicine **90** (2023)

[9] Kaviri, S.M., Vinjamuri, R.: Decoding motor execution and motor imagery from eeg with deep learning and source localization. Biomedical Engineering Advances **9**, 100156 (2025)

[10] Yang, B., Ma, J., Qiu, W., Zhu, Y., Meng, X.: A new 2-class unilateral upper limb motor imagery tasks for stroke rehabilitation training. Medicine in Novel Technology and Devices **13**, 100100 (2022)

[11] Siddaway, A.P., Wood, A.M., Hedges, L.V.: How to do a systematic review: a best practice guide for conducting and reporting narrative reviews, meta-analyses, and meta-syntheses. Annual review of psychology **70**(1), 747–770 (2019)

[12] Camargo-Vargas, D., Callejas-Cuervo, M., Alarcón-Aldana, A.C.: Brain-computer interface prototype to support upper limb rehabilitation processes in the human body. International Journal of Information Technology **15**(7), 3655–3667 (2023)

[13] Pfeffer, M.A., Ling, S.S.H., Wong, J.K.W.: Exploring the frontier: Transformer-based models in eeg signal analysis for brain-computer interfaces. Computers in biology and medicine **178**, 108705 (2024)

[14] Wang, X., Liesaputra, V., Liu, Z., Wang, Y., Huang, Z.: An in-depth survey on deep learning-based motor imagery electroencephalogram (eeg) classification. Artificial intelligence in medicine **147**, 102738 (2024)

[15] Chen, L., Gu, B., Wang, Z., Zhang, L., Xu, M., Liu, S., He, F., Ming, D.: Eeg-controlled functional electrical stimulation rehabilitation for chronic stroke: system design and clinical application. Frontiers of medicine **15**(5), 740–749 (2021)

[16] Chowdhury, R.R., Muhammad, Y., Adeel, U.: Enhancing cross-subject motor imagery classification in eeg-based brain–computer interfaces by using multi-branch cnn. Sensors **23**(18), 7908 (2023)

[17] Zhang, K., Xu, G., Han, Z., Ma, K., Zheng, X., Chen, L., Duan, N., Zhang, S.: Data augmentation for motor imagery signal classification based on a hybrid neural network. Sensors **20**(16), 4485 (2020)

[18] Li, H., Ding, M., Zhang, R., Xiu, C.: Motor imagery eeg classification algorithm based on cnn-lstm feature fusion network. Biomedical signal processing and control **72**, 103342 (2022)

[19] Biasiucci, A., Leeb, R., Iturrate, I., Perdikis, S., Al-Khodairy, A., Corbet, T., Schnider, A., Schmidlin, T., Zhang, H., Bassolino, M., *et al.*: Brain-actuated functional electrical stimulation elicits lasting arm motor recovery after stroke. Nature communications **9**(1), 2421 (2018)

[20] Kim, M.S., Park, H., Kwon, I., An, K.-O., Kim, H., Park, G., Hyung, W., Im, C.-H., Shin, J.-H.: Efficacy of brain-computer interface training with motor imagery-contingent feedback in improving upper limb function and neuroplasticity among persons with chronic stroke: a double-blinded, parallel-group, randomized controlled trial. Journal of NeuroEngineering and Rehabilitation **22**(1), 1 (2025)

[21] Ren, R., Basart, S., Khoja, A., Pan, A., Gatti, A., Phan, L., Yin, X., Mazeika, M., Mukobi, G., Kim, R.H., *et al.*: Safetywashing: Do ai safety benchmarks actually measure safety progress? Advances in Neural Information Processing Systems **37**, 68559–68594 (2024)

[22] Tan, X., Wang, D., Xu, M., Chen, J., Wu, S.: Efficient multi-view graph convolutional network with self-attention for multi-class motor imagery decoding. Bioengineering **11**(9), 926 (2024)

[23] Zhan, G., Chen, S., Ji, Y., Xu, Y., Song, Z., Wang, J., Niu, L., Bin, J., Kang, X., Jia, J.: Eeg-based brain network analysis of chronic stroke patients after bci rehabilitation training. Frontiers in Human Neuroscience **16**, 909610 (2022)

[24] Ritchie, S.: Science Fictions: Exposing Fraud, Bias, Negligence and Hype in Science. Random House, ??? (2020)

[25] Kong, X., Wu, C., Chen, S., Wu, T., Han, J.: Efficient feature learning model of motor imagery eeg signals with l1-norm and weighted fusion. Biosensors **14**(5), 211 (2024)

[26] Xiong, X., Wang, Y., Song, T., Huang, J., Kang, G.: Improved motor imagery classification using adaptive spatial filters based on particle swarm optimization algorithm. Frontiers in Neuroscience **17**, 1303648 (2023)

[27] Brunner, I., Lundquist, C.B., Pedersen, A.R., Spaich, E.G., Dosen, S., Savic, A.: Brain computer interface training with motor imagery and functional electrical stimulation for patients with severe upper limb paresis after stroke: a randomized controlled pilot trial. Journal of NeuroEngineering and Rehabilitation **21**(1), 10 (2024)

[28] Karbalaie, A., Abtahi, F., Häger, C.: Subject-aware model validation for repeated-measures data: a nested approach for trustworthy medical ai applications. Available at SSRN 5352124 (2025)

[29] Bouvet, A.: Monitoring et modélisation de la performance en natation via centrale inertielle et data science: méthodologies et applications pour l'accompagnement scientifique à la performance. PhD thesis, Université Rennes 2 (2025)

[30] Chen, Y., Jiang, Y., Wang, X., Zeng, M., Cui, J.: Advances in mechanisms of neuroplasticity induced by multimodal closed-loop brain–computer interfaces after stroke. Frontiers in Human Neuroscience **20**, 1828191 (2026)

[31] Liew, S.-L., Zavaliangos-Petropulu, A., Jahanshad, N., Lang, C.E., Hayward, K.S., Lohse, K.R., Juliano, J.M., Assogna, F., Baugh, L.A., Bhattacharya, A.K., *et al.*: The enigma stroke recovery working group: Big data neuroimaging to study brain–behavior relationships after stroke. Human brain mapping **43**(1), 129–148 (2022)

[32] Borschmann, K.N., Hayward, K.S.: Recovery of upper limb function is greatest early after stroke but does continue to improve during the chronic phase: a two-year, observational study. Physiotherapy **107**, 216–223 (2020)

[33] Tong, W., Yue, W., Chen, F., Shi, W., Zhang, L., Wan, J.: Mse-vgg: a novel deep learning

approach based on eeg for rapid ischemic stroke detection. Sensors **24**(13), 4234 (2024)

[34] Slawomirski, L., Auraaen, A., Klazinga, N., et al.: The economics of patient safety. Paris: Organisation for Economic Co-operation and Development (2017)

[35] McDermott, E.J., Metsomaa, J., Belardinelli, P., Grosse-Wentrup, M., Ziemann, U., Zrenner, C.: Predicting motor behavior: an efficient eeg signal processing pipeline to detect brain states with potential therapeutic relevance for vr-based neurorehabilitation. Virtual Reality **27**(1), 347–369 (2023)

[36] Butet, S., Fleury, M., Duché, Q., Bannier, E., Lioi, G., Covella, L., Lévêque-Le Bars, E., Lécuyer, A., Maurel, P., Bonan, I.: Eeg-fmri neurofeedback versus motor imagery after stroke, a randomized controlled trial. Journal of NeuroEngineering and Rehabilitation **22**(1), 67 (2025)

[37] Grennan, M., Town, R.J.: Regulating innovation with uncertain quality: information, risk, and access in medical devices. American Economic Review **110**(1), 120–161 (2020)

[38] Guger, C., Grünwald, J., Xu, R.: Noninvasive and invasive bcis and hardware and software components for bcis. In: Handbook of Neuroengineering, pp. 1193–1224. Springer, ??? (2023)

[39] Francesconi, M., Cangi, M., Tamarri, S., Conditi, N., Menicucci, C., Ravizza, A., Cattaneo, L., Bianchini, E.: Navigating regulatory challenges across the life cycle of a samd. Journal of Biomedical Informatics **167**, 104856 (2025)

[40] Calderone, A., Latella, D., La Fauci, E., Puleo, R., Sergi, A., De Francesco, M., Mauro, M., Foti, A., Salemi, L., Calabrò, R.S.: Artificial intelligence-driven neuromodulation in neurodegenerative disease: Precision in chaos, learning in loss. Biomedicines **13**(9), 2118 (2025)

[41] Ding, Y., Udompanyawit, C., Zhang, Y., He, B.: Eeg-based brain-computer interface enables real-time robotic hand control at individual finger level. Nature Communications **16**(1), 5401 (2025)