A SYSTEM-LEVEL FRAMEWORK FOR EEG-BASED SCHIZOPHRENIA ASSESSMENT: METHODOLOGICAL RIGOR, UNCERTAINTY QUANTIFICATION, AND HARDWARE FEASIBILITY

Samiksha B. Chandrasekaran[1] and Eric Raymond[1]

[1]Department of Computer Science, Indiana University South Bend, South Bend, IN 46615, USA
[1]Purdue University Indianapolis, Indianapolis, IN 46202, USA

Abstract

Schizophrenia diagnosis remains predominantly subjective, relying on clinical interviews and behavioral observation, while machine learning approaches using electroencephalography (EEG) have shown promise for objective biomarker discovery. However, many published studies suffer from methodological flaws—particularly "subject leakage," where recordings from the same individual contaminate both training and testing sets, artificially inflating reported accuracies. This work presents a rigorously validated EEG classification pipeline that prioritizes methodological integrity over inflated performance metrics, alongside a low-cost hardware prototype suitable for resource-limited clinical settings. Using the ASZED-153 dataset (N=153 subjects; 77 healthy controls, 76 schizophrenia patients; 1,931 recordings), we implemented strict subject-level cross-validation ensuring no identity leakage. Feature extraction yielded 264 features spanning spectral power, coherence, phase-lag index, and nonlinear complexity measures, with Random Forest pre-specified as the primary classifier to avoid post-hoc selection bias. Subject-level classification achieved 83.7% accuracy (95% CI: 77.8–89.5%) with ROC-AUC of 0.869, representing an approximate 7-point reduction from recording-level accuracy (90.9%)—quantifying the inflation caused by identity leakage in naive evaluation schemes. Feature importance analysis revealed frontal channels (Fp1, Fp2) as top predictors, biologically validating our single-channel ESP32-based hardware design. By transparently reporting honest metrics obtained through rigorous methodology, this work establishes a reproducible baseline for EEG-based schizophrenia screening and offers a pathway toward accessible psychiatric assessment tools for underserved populations.

1 Introduction

Schizophrenia affects approximately 1% of the global population, with diagnosis typically occurring years after symptom onset due to the disorder's heterogeneous presentation and the absence of objective biomarkers [1]. Current diagnostic practice relies heavily on clinical interviews and behavioral observation, introducing subjectivity that can delay treatment initiation and contribute to poor long-term outcomes [2].

The application of machine learning (ML) to electroencephalography (EEG) has emerged as a promising avenue for developing objective, quantitative markers of schizophrenia [3, 4]. EEG offers several practical advantages: it is non-invasive, relatively inexpensive, and captures the neural oscillatory dynamics known to be disrupted in psychotic disorders [5]. Published studies have reported classification accuracies exceeding 90%, suggesting that EEG-based diagnostic tools may soon augment clinical practice.

However, the ML-for-health literature faces a well-documented replication crisis [6, 7]. A substantial proportion of studies suffer from methodological flaws that artificially inflate reported performance. Chief among these is *identity leakage*—the contamination of test sets with recordings from subjects present in the training set. When multiple recordings per subject exist (as is common in EEG datasets), naive random splitting allows the model to exploit subject-specific patterns (voice, electrode impedance, head shape) rather than learning disorder-relevant biomarkers. The resulting accuracy estimates are overly optimistic and fail to generalize to new individuals [8, 9].

Table 1: Demographic Characteristics of the ASZED-153 Dataset

Characteristic	**Controls (HC)**	**Patients (SZ)**
N (subjects)	77	76
N (recordings)	990	941
Recordings/subject	12.9	12.4
Recording sites	2 (Nigeria)	
Sampling rate	256 Hz	
Channels	16 (10–20 system)	

1.1 Contributions of This Work

This paper makes four primary contributions:

1. **Rigorous Evaluation Protocol:** We implement strict subject-level stratified cross-validation, ensuring that all recordings from a given individual remain entirely within either the training or testing fold—never split between them. This eliminates identity leakage and provides honest generalization estimates.
2. **Transparent Quantification of Methodological Impact:** We explicitly compare recording-level accuracy (the inflated metric) with subject-level accuracy (the honest metric), quantifying the "cost of rigor" at approximately 7 percentage points.
3. **Pre-Specified Analysis Plan:** The primary classifier (Random Forest) was designated before evaluation to prevent post-hoc model selection bias, adhering to principles of registered reporting.
4. **Hardware Feasibility Demonstration:** We developed a low-cost ($50) single-channel EEG prototype based on the ESP32 microcontroller and BioAmp EXG Pill, with feature importance analysis validating that frontal channels carry sufficient discriminative information.

We emphasize that this work prioritizes *methodological integrity* over state-of-the-art accuracy claims. The resulting 83.7% subject-level accuracy, while modest compared to inflated literature benchmarks, represents an honest, reproducible baseline upon which future work can build.

2 Materials and Methods

2.1 Dataset: ASZED-153

We utilized the African Schizophrenia EEG Dataset (ASZED), version 1.1, publicly available through Zenodo (DOI: 10.5281/zenodo.14178398) [10]. This dataset comprises EEG recordings from 153 participants: 77 healthy controls (HC) and 76 patients with schizophrenia (SZ), recruited from two clinical sites in Nigeria. Demographic characteristics are summarized in Table 1.

Recordings were acquired using a 16-channel montage following the international 10–20 system (Fp1, Fp2, F3, F4, C3, C4, P3, P4, O1, O2, F7, F8, T3, T4, T5, T6) at 256 Hz sampling rate. Multiple paradigms were recorded per subject, including resting-state (eyes open/closed), cognitive tasks, mismatch negativity (MMN), and auditory steady-state response (ASSR), yielding a total of 1,931 usable recordings (mean: 12.6 recordings per subject).

2.2 Preprocessing Pipeline

All preprocessing was implemented in Python 3.10 using MNE-Python [11]. The pipeline comprised:

1. **Channel Standardization:** Raw channel names (e.g., "Fp1[1]") were canonicalized to standard 10–20 nomenclature. Missing channels were zero-padded in-place to maintain consistent feature indexing.
2. **Filtering:** Fourth-order Butterworth bandpass filter (0.5–45 Hz) followed by a 50 Hz notch filter (Nigeria power grid frequency).
3. **Quality Control:** Files with fewer than 10 matched channels or fewer than 500 samples were rejected. Rejection rates were monitored for selection bias using Fisher's exact test.
4. **Resampling:** All recordings were resampled to 250 Hz for computational consistency.

Quality control analysis confirmed no differential rejection between diagnostic groups (rejection rate: HC = 0.1%, SZ = 0.0%; Fisher exact $p = 1.0$), ensuring that preprocessing did not introduce selection bias.

2.3 Feature Extraction

We extracted 264 features per recording, organized into six categories:

1. **Spectral Power (80 features):** Band power computed via Welch's method for delta (0.5–4 Hz), theta (4–8 Hz), alpha (8–13 Hz), beta (13–30 Hz), and gamma (30–45 Hz) across all 16 channels.
2. **ERP-like Components (20 features):** Temporal dynamics in windows corresponding to N100, P200, MMN, and P300 latencies, including peak amplitude, latency, mean, standard deviation, and area under curve.
3. **Inter-channel Coherence (30 features):** Magnitude-squared coherence between six electrode pairs (Fp1-O1, Fp2-O2, Fp1-Fp2, O1-O2, C3-T3, C4-T4) across five frequency bands.
4. **Phase-Lag Index (6 features):** PLI computed for the same electrode pairs, quantifying phase synchronization while minimizing volume conduction artifacts [12].
5. **Statistical Moments (96 features):** Mean, standard deviation, skewness, kurtosis, RMS, and peak-to-peak amplitude for each channel.
6. **Nonlinear Complexity (32 features):** Sample entropy and Higuchi fractal dimension for each channel, capturing signal complexity reductions associated with schizophrenia [13].

2.4 Strict Subject-Level Cross-Validation

To prevent identity leakage, we implemented a strict subject-level evaluation protocol (Figure **??**):

1. **Subject-Level Stratification:** Five-fold stratified splitting was performed on the 153 *subjects* (not recordings), ensuring approximately equal class proportions in each fold.
2. **Fold Expansion:** Subject-level folds were then expanded to include all recordings belonging to each subject, guaranteeing that no subject appeared in both training and testing partitions.
3. **Within-Fold Normalization:** Feature standardization (z-scoring via StandardScaler) was applied *inside* each fold, fitting only on training data to prevent information leakage.
4. **Subject-Level Aggregation:** Test predictions were aggregated by subject using mean probability voting, with final classification determined at threshold 0.5.

This protocol ensures that reported metrics reflect true generalization to *new individuals*, not memorization of subject-specific artifacts.

2.5 Classification Models

Four classifiers were evaluated:

- **Random Forest (RF):** 300 trees, max depth 20, min samples split 5. Designated as the *primary model* before analysis.
- **Logistic Regression (LR):** L2 regularization, max 1000 iterations.
- **Gradient Boosting (GB):** 100 estimators, default hyperparameters.
- **Support Vector Machine (SVM):** RBF kernel, probability calibration enabled.

All models were wrapped in scikit-learn Pipelines to ensure proper scaler fitting within each CV fold [14].

2.6 Hardware Prototype

To demonstrate feasibility for low-resource settings, we developed a single-channel EEG acquisition system comprising:

- **Microcontroller:** ESP32 (dual-core, Wi-Fi/Bluetooth enabled, $5)
- **Analog Front-End:** BioAmp EXG Pill (instrumentation amplifier with ×1000 gain, $25)
- **Electrodes:** Dry Ag/AgCl electrodes at Fp1 position
- **Sampling:** 256 Hz, 12-bit ADC resolution

Total hardware cost was approximately $50 USD, compared to $5,000–50,000 for clinical-grade EEG systems.

Table 2: Classification Performance (Subject-Level, N=153)

Model	Accuracy (%)	AUC	F1
Random Forest*	**83.7**	**0.869**	0.837
Gradient Boosting	83.7	0.871	0.837
SVM (RBF)	81.7	0.852	0.823
Logistic Regression	76.5	0.811	0.768

*Pre-specified primary model.

Table 3: Confusion Matrix (Subject-Level, Random Forest)

	Predicted HC	Predicted SZ
Actual HC (n=77)	57	20
Actual SZ (n=76)	5	71

Sensitivity (SZ recall) = 93.4%, Specificity (HC recall) = 74.0%

3 Results

3.1 Quality Control and Selection Bias Analysis

Of 1,932 EEG files, 1,931 (99.95%) passed quality control. The single rejection was due to insufficient recording length (<500 samples). Crucially, rejection rates did not differ between diagnostic groups (HC: 0.1%, SZ: 0.0%; Fisher exact test $p = 1.0$), confirming that preprocessing did not introduce selection bias.

All 153 subjects were retained for analysis, with matched channel counts (mean = 16.0, range: 16–16) indicating successful channel canonicalization.

3.2 Classification Performance

Table 2 presents classification performance across all models. Random Forest, the pre-specified primary model, achieved 83.7% subject-level accuracy (95% CI: 77.8–89.5%) with ROC-AUC of 0.869.

The subject-level confusion matrix for Random Forest is presented in Table 3. Note that all cells sum to N=153 subjects, not recordings.

The model demonstrated high sensitivity (93.4%) for detecting schizophrenia cases, at the cost of moderate specificity (74.0%). This trade-off is appropriate for a screening tool where missing true cases carries greater clinical cost than false positives.

3.3 Quantifying Identity Leakage

To quantify the inflation caused by identity leakage, we compared recording-level and subject-level accuracy (Table 4). Recording-level accuracy (90.9%) exceeded subject-level accuracy (83.7%) by 7.2 percentage points, representing the "hidden cost" of naive evaluation protocols.

3.4 Cross-Validation Stability

Per-fold subject-level accuracy ranged from 74.2% to 90.0% across the five folds, with balanced class distributions in each fold confirming successful stratification (Table 5).

3.5 Feature Importance and Hardware Validation

Random Forest feature importance analysis revealed that frontal channels contributed disproportionately to classification (Figure **??**). The top 10 features included:

- Fp1 theta power (rank 1)

Table 4: Impact of Evaluation Methodology on Reported Accuracy

Evaluation Level	**N**	**Accuracy (%)**
Recording-level (pooled OOF)	1,931	90.9
Subject-level (aggregated)	153	83.7
Inflation due to leakage	—	**+7.2**

Table 5: Per-Fold Performance (Random Forest)

Fold	**Train (C/S)**	**Test (C/S)**	**Acc**	**AUC**
1	62/60	15/16	0.894	0.916
2	61/61	16/15	0.802	0.808
3	61/61	16/15	0.743	0.773
4	62/61	15/15	0.871	0.919
5	62/61	15/15	0.908	0.947

C = Controls, S = Schizophrenia patients

- Fp2 alpha power (rank 2)
- Fp1 sample entropy (rank 4)
- F3 beta power (rank 6)
- Fp1-Fp2 coherence (rank 8)

The prominence of Fp1 and Fp2 channels validates our single-channel hardware design targeting the frontal region. Frontal abnormalities are well-documented in schizophrenia, including reduced frontal alpha power and increased theta activity associated with hypofrontality [15].

4 Discussion

4.1 Quantifying the Cost of Methodological Rigor

The central finding of this work is not the 83.7% accuracy itself, but rather the *7.2 percentage point gap* between recording-level and subject-level evaluation. This gap quantifies the inflation introduced by identity leakage—a pervasive but often unreported methodological flaw in the EEG-ML literature.

When multiple recordings exist per subject, models can exploit subject-specific artifacts (electrode impedance patterns, head geometry, environmental noise signatures) rather than learning disorder-relevant biomarkers. The resulting accuracy estimates, while numerically impressive, fail to generalize to new individuals—the actual clinical use case.

We argue that the apparent “drop” in accuracy from 90.9% to 83.7% should be reframed as a *scientific correction* rather than a limitation. The subject-level estimate reflects honest generalization performance; the recording-level estimate is an artifact of evaluation methodology. By explicitly reporting both, we enable the field to calibrate expectations and identify studies that may have inadvertently inflated their results.

4.2 Clinical Interpretation of the Confusion Matrix

The confusion matrix (Table 3) reveals an asymmetric error pattern: the model exhibits high sensitivity (93.4%) but moderate specificity (74.0%). This means:

- 71 of 76 patients (93.4%) are correctly identified
- 5 patients are missed (false negatives)
- 20 of 77 controls are incorrectly flagged (false positives)

For a *screening* tool, this trade-off is clinically appropriate. Missing a schizophrenia case (false negative) carries severe consequences: delayed treatment, disease progression, and increased risk of adverse outcomes. False positives, while burdensome, can be resolved through subsequent clinical evaluation.

The 74% specificity suggests that approximately one in four healthy individuals would be incorrectly flagged for follow-up. In a two-stage screening paradigm—EEG screening followed by clinical interview for positive cases—this false positive rate may be acceptable, particularly in high-risk populations where base rates of schizophrenia are elevated.

4.3 The Case for Uncertainty Quantification

Beyond point predictions, we advocate for uncertainty quantification in clinical ML systems. Models that report only binary classifications provide false confidence; those that report calibrated probabilities enable clinicians to triage cases appropriately.

In our analysis, prediction probabilities clustered near 0.5 for many subjects, indicating low model confidence. We propose that predictions with probability < 0.6 or > 0.4 (i.e., within 10 points of the decision boundary) should be flagged as "uncertain" and prioritized for clinical review. This approach embodies the principle of "knowing what you don't know"—a critical requirement for safe deployment of ML in healthcare.

4.4 Hardware Implications for Global Health

Feature importance analysis revealed that frontal channels (Fp1, Fp2) were among the top predictors, validating our single-channel hardware concept. This finding has significant implications for accessibility:

1. **Cost Reduction:** A single-channel system costs approximately $50, compared to $5,000–50,000 for clinical-grade 16+ channel systems.
2. **Ease of Use:** Single-channel acquisition requires minimal training, enabling deployment by community health workers.
3. **Biological Plausibility:** Frontal abnormalities are well-established in schizophrenia, including hypofrontality, reduced frontal gamma synchrony, and aberrant prefrontal connectivity [5].

Our ESP32-based prototype demonstrates technical feasibility. Future work will validate classification performance using hardware-acquired signals, addressing the domain shift between research-grade and low-cost acquisition systems.

4.5 Comparison with Literature

Published EEG-based schizophrenia classification studies have reported accuracies ranging from 75% to 99% [4, 3]. However, critical examination reveals that many high-accuracy claims stem from:

- Recording-level evaluation (identity leakage)
- Small sample sizes (N < 50 subjects)
- Post-hoc model selection
- Absence of confidence intervals

Our 83.7% accuracy, while lower than some reported values, is derived from a rigorously validated protocol on a relatively large dataset (N=153). We contend that this estimate is more likely to generalize to real-world clinical deployment than inflated metrics from methodologically flawed studies.

5 Limitations

Several limitations warrant acknowledgment:

1. **Single Dataset:** All analyses were conducted on the ASZED-153 dataset. External validation on independent cohorts is essential to confirm generalizability.
2. **Geographic Specificity:** The dataset was collected in Nigeria. Performance may differ in populations with different genetic backgrounds, medication regimens, or comorbidity profiles.
3. **Hardware Prototype Stage:** The low-cost hardware system was developed as a proof-of-concept. The main classification results were obtained from research-grade EEG; validation with hardware-acquired signals remains future work.
4. **Cross-Sectional Design:** This study evaluated diagnostic classification (HC vs. SZ) at a single timepoint. Longitudinal prediction of disease onset, progression, or treatment response was not addressed.

5. **Binary Classification:** Schizophrenia is a heterogeneous disorder with multiple subtypes. Binary classification may obscure clinically meaningful subgroup differences.
6. **Medication Effects:** Patient recordings were collected during naturalistic treatment. Antipsychotic medications are known to alter EEG patterns, potentially confounding diagnostic biomarkers.

6 Conclusion

This work demonstrates that rigorous methodology—including strict subject-level cross-validation, pre-specified primary analysis, and transparent reporting—yields honest but modest classification performance for EEG-based schizophrenia detection. The 83.7% subject-level accuracy represents approximately 7 percentage points less than recording-level metrics, quantifying the inflation attributable to identity leakage.

We frame this “accuracy drop” as a feature, not a bug. By explicitly correcting for methodological artifacts, we establish a reproducible baseline upon which the field can build. The integration of low-cost hardware feasibility analysis further advances the goal of accessible psychiatric screening tools for underserved populations.

Future work should prioritize external validation, prospective clinical trials, and longitudinal prediction of disease trajectories. We advocate for registered analysis protocols and mandatory reporting of subject-level metrics in EEG-ML research.

In an era of inflated claims and replication failures, this work offers a template for honest, transparent, and clinically grounded AI in psychiatry.

Data Availability

The ASZED-153 dataset is publicly available through Zenodo (DOI: 10.5281/zenodo.14178398). Analysis code will be made available upon reasonable request.

Acknowledgments

The authors thank the original ASZED dataset creators for making their data publicly available. We acknowledge the use of computational resources at Indiana University South Bend.

References

[1] World Health Organization. Schizophrenia Fact Sheet. WHO, 2022.

[2] Tandon R, Gaebel W, Barch DM, et al. Definition and description of schizophrenia in the DSM-5. *Schizophr Res*. 2013;150(1):3–10.

[3] Murphy M, Whitton AE, Deccy S, et al. Abnormalities in EEG during sleep and wakefulness in schizophrenia. *JAMA Psychiatry*. 2021;78(9):986–994.

[4] Phang CR, Noman F, Hussain H, Ting CM, Ombao H. A multi-domain connectome convolutional neural network for identifying schizophrenia from EEG connectivity patterns. *IEEE J Biomed Health Inform*. 2020;24(5):1333–1343.

[5] Uhlhaas PJ, Singer W. Abnormal neural oscillations and synchrony in schizophrenia. *Nat Rev Neurosci*. 2010;11(2):100–113.

[6] Roberts M, Driggs D, Thorpe M, et al. Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans. *Nat Mach Intell*. 2021;3(3):199–217.

[7] Varoquaux G, Cheplygina V. Machine learning for medical imaging: methodological failures and recommendations for the future. *NPJ Digit Med*. 2022;5(1):48.

[8] Little MA, Varoquaux G, Saeb S, et al. Using and understanding cross-validation strategies. *GigaScience*. 2017;6(5):gix020.

[9] Saeb S, Lonini L, Jayaraman A, Mohr DC, Kording KP. The need to approximate the use-case in clinical machine learning. *GigaScience*. 2017;6(5):gix019.

[10] African Schizophrenia EEG Dataset (ASZED). Zenodo. 2024. DOI: 10.5281/zenodo.14178398.

[11] Gramfort A, Luessi M, Larson E, et al. MEG and EEG data analysis with MNE-Python. *Front Neurosci*. 2013;7:267.

[12] Stam CJ, Nolte G, Daffertshofer A. Phase lag index: assessment of functional connectivity from multi-channel EEG and MEG with diminished bias from common sources. *Hum Brain Mapp*. 2007;28(11):1178–1193.

[13] Kim DJ, Jeong J, Chae JH, et al. An estimation of the first positive Lyapunov exponent of the EEG in patients with schizophrenia. *Psychiatry Res Neuroimaging*. 2000;98(3):177–189.

[14] Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: Machine learning in Python. *J Mach Learn Res*. 2011;12:2825–2830.

[15] Boutros NN, Arfken C, Galderisi S, Warrick J, Pratt G, Iacono W. The status of spectral EEG abnormality as a diagnostic test for schizophrenia. *Schizophr Res*. 2008;99(1-3):225–237.