A System-Level Framework for EEG-Based Schizophrenia Assessment: Methodological Rigor, Uncertainty Quantification, and Hardware Feasibility

Samiksha BC[1], Eric Raymond[1], and Divyashree Santhosh[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. Machine learning approaches using electroencephalography (EEG) have shown promise for objective biomarker discovery, but many published studies suffer from *identity leakage*—where recordings from the same individual contaminate both training and testing sets, artificially inflating reported accuracies. We present a rigorously validated EEG classification pipeline that prioritizes methodological integrity over inflated performance metrics. 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 prespecified 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%) and quantifying the inflation caused by identity leakage in naive evaluation schemes. Feature importance analysis revealed frontal channels (Fp1, Fp2) as top predictors, providing biological rationale for targeting frontal sites in future low-cost hardware designs. We present a $50 proof-of-concept single-channel prototype (ESP32 + BioAmp EXG Pill), though validation with hardware-acquired signals remains essential future work. By transparently reporting honest metrics obtained through rigorous methodology, this work establishes a reproducible baseline for EEG-based schizophrenia screening and proposes a pathway—contingent on external validation and prospective trials—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.

The ML-for-health literature, however, faces a well-documented replication crisis [6, 7]. A substantial proportion of studies suffer from methodological flaws that artificially inflate reported performance. Most problematic 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	200/256 Hz (resampled to 250 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. **Feature Importance Analysis and Proof-of-Concept Hardware Design:** Feature importance analysis identified frontal channels as top predictors, providing biological rationale for a low-cost ($50) single-channel prototype design (ESP32 + BioAmp EXG Pill). This serves as a proof-of-concept for future accessibility research; functional validation with hardware-acquired data remains essential next-step work.

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. All classification results reported here were obtained from research-grade EEG equipment; the hardware prototype serves as a proof-of-concept for future accessibility-focused validation studies.

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]. Participants were 153 adults recruited from two clinical sites in Nigeria: 77 healthy controls (HC) and 76 patients meeting DSM-5 criteria for schizophrenia (SZ). Clinical diagnoses were established by board-certified psychiatrists using structured clinical interviews; specific diagnostic instruments (e.g., SCID-5) were not reported in the original dataset documentation. Patient medication status, illness duration, and symptom severity scores were not available in the public release. Controls were screened for absence of psychiatric history via self-report; formal diagnostic exclusion criteria were not documented. Demographic matching between groups (age, sex, education) could not be verified from available metadata. These limitations constrain our ability to control for potential confounds and are addressed in Section 2.2.

Recordings were acquired using a 16-channel montage following the international 10-20 system (Fp1, Fp2, F3, F4, F7, F8, C3, C4, Cz, T3, T4, T5, T6, P3, P4, Pz) using two EEG systems: Contec-KT2400 (200 Hz) and Brain-Master Discovery24-E (256 Hz). Multiple paradigms were recorded per subject, including resting-state (eyes closed), arithmetic working-memory task, mismatch negativity (MMN), and 40 Hz auditory steady-state response (ASSR), yielding a total of 1,931 usable recordings (mean: 12.6 recordings per subject). Channel placement and paradigm specifications are per Mosaku et al. [**?**].

2.2 Clinical Characterization and Unmeasured Confounds

Several clinical variables that may influence EEG patterns were unavailable in the ASZED public release, limiting our ability to control for confounds:

- **Diagnostic Criteria:** The dataset documentation states that schizophrenia diagnoses were established by board-certified psychiatrists, but specific diagnostic instruments (e.g., SCID-5) or DSM-5/ICD-10 criteria confirmation were not reported. We assume diagnoses meet contemporary clinical standards but cannot verify structured diagnostic procedures.
- **Medication Status:** Antipsychotic medication types, doses, and treatment duration were not documented. Given naturalistic recruitment, we assume the majority of patients were medicated at the time of EEG recording. Antipsychotics (particularly dopamine D2 antagonists) are known to alter EEG spectral power, especially in beta and gamma bands [15]. Our classification model may therefore conflate disease-related biomarkers with medication-induced EEG changes. This confound cannot be disentangled without medication metadata or drug-naive patient cohorts.
- **Illness Characteristics:** Duration since diagnosis, number of psychotic episodes, current symptom severity (e.g., PANSS total scores), and illness subtype were not available. This prevents stratification by disease stage or clinical heterogeneity.
- **Comorbidities:** Substance use disorders (particularly cannabis), affective symptoms (depression, anxiety), and neurological conditions were not reported. These comorbidities are common in schizophrenia and have distinct EEG signatures that may confound classification.
- **Demographic Matching:** Group-level age, sex, and education distributions were not documented. Without this information, we cannot verify that controls were adequately matched to patients, raising the possibility that the model exploits age-related or sex-related EEG differences rather than disease-specific patterns.

These unmeasured variables represent threats to internal validity. The high sensitivity (93.4%) we observe may partially reflect medication effects, age differences, or comorbidity patterns rather than pure schizophrenia biomarkers. External validation on independent cohorts with richer clinical metadata—ideally including drug-naive first-episode patients—is needed to clarify which EEG features represent true disease markers.

2.3 Preprocessing Pipeline

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

1. **Referencing:** The ASZED dataset does not document the original acquisition reference. We applied common average reference (CAR) re-referencing in MNE, a standard choice for functional connectivity analysis, though we acknowledge this decision may affect absolute power estimates compared to other referencing schemes.
2. **Channel Standardization:** Raw channel names (e.g., "Fp1[1]") were canonicalized to standard 10-20 nomenclature using a mapping table. Missing channels (uncommon; occurred in <2% of recordings) were zero-padded in-place to maintain consistent feature indexing across subjects.
3. **Filtering:** Fourth-order Butterworth bandpass filter (0.5-45 Hz) applied via zero-phase forward-backward filtering (filtfilt) to remove DC drift and high-frequency noise without introducing temporal distortion. A 50 Hz notch filter (3 Hz width) removed Nigeria mains interference.
4. **Artifact Rejection:** We deliberately did *not* perform automated or manual artifact rejection (e.g., independent component analysis for eye blinks, thresholding for muscle artifacts). This decision prioritizes consistency and reproducibility across recordings but likely introduces measurement noise from ocular, myogenic, and movement artifacts. Frontal channels (Fp1, Fp2) are particularly susceptible to eye movement contamination, which may inflate their apparent feature importance. Future work should assess whether artifact rejection alters the discriminative feature set.
5. **Segmentation:** Task-based recordings (MMN, ASSR, cognitive tasks) were analyzed as continuous segments without epoching around stimulus events, as event markers were not available in the dataset. Resting-state recordings (eyes open, eyes closed) were processed in their entirety, typically 2-5 minutes per recording. We did not apply baseline correction.
6. **Quality Control:** Files with fewer than 10 matched channels or fewer than 500 samples (2 seconds at 250 Hz) were rejected as insufficient for spectral estimation via Welch's method. Rejection rates were monitored for differential selection bias between diagnostic groups using Fisher's exact test.
7. **Resampling:** All recordings were resampled to 250 Hz using MNE's Fourier-based antialiasing resampling to

ensure uniform sampling rate for subsequent feature extraction.

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.

2.4 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):** ERP-like temporal dynamics computed on the Global Field Power (GFP; spatial RMS across all 16 channels) in windows corresponding to N100 (80-120ms), P200 (150-250ms), MMN (150-200ms), and P300 (250-400ms) latencies. Features include peak amplitude, peak latency, window mean, standard deviation, and area under curve.
3. **Inter-channel Coherence (30 features):** Magnitude-squared coherence between six electrode pairs across five frequency bands. Pairs were chosen to capture interhemispheric connectivity commonly disrupted in schizophrenia: prefrontal (Fp1-Fp2), frontal (F3-F4), central (C3-C4), temporal (T3-T4), posterior temporal (T5-T6), and parietal (P3-P4) regions [5].
4. **Phase-Lag Index (6 features):** PLI computed for the same six electrode pairs in the alpha band (8-13 Hz), chosen because alpha-band synchronization abnormalities are well-documented in schizophrenia. PLI quantifies 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].

Methodological Clarifications. Three aspects of the feature extraction procedure require elaboration:

- **ERP-like components on continuous data:** Although traditional event-related potentials require stimulus-locked averaging, we computed "ERP-like" temporal features by applying ERP latency windows (N100: 80-120ms, P200: 150-250ms, etc.) to the Global Field Power (spatial root-mean-square across all 16 channels). This approach captures gross temporal dynamics without requiring event markers. Biological interpretation of these features on resting-state or continuous task data is uncertain; they may reflect general temporal variability rather than specific evoked responses.
- **Coherence and PLI electrode pairs:** The six electrode pairs for coherence and phase-lag index were chosen to capture interhemispheric dysconnectivity patterns implicated in schizophrenia: prefrontal (Fp1-Fp2), frontal (F3-F4), central (C3-C4), temporal (T3-T4), posterior temporal (T5-T6), and parietal (P3-P4). These pairs were selected to match the ASZED-153 electrode montage (which does not include occipital electrodes O1/O2). PLI was computed specifically in the alpha band (8-13 Hz), where synchronization abnormalities are well-documented in psychosis [5].
- **Nonlinear complexity sample length:** Sample entropy and Higuchi fractal dimension were computed on the first 250 samples (1 second at 250 Hz) per channel due to computational constraints. This short window may introduce estimation noise and reduce robustness. Future work should assess sensitivity to window length.

Justification for Feature Dimensionality. The 264-dimensional feature space with 153 subjects ($p/n \approx 1.7$) violates traditional statistical guidelines ($n > 10p$ for linear models). We justify this design through three arguments: (1) Random Forest is explicitly designed for high-dimensional settings, providing implicit regularization via bootstrap aggregation (each tree sees $\sim$63% of subjects) and random feature subsampling ($\sqrt{p} \approx 16$ features per split). The constraint further limits tree complexity. (2) We deliberately avoided any feature selection or ranking applied to the full dataset before cross-validation, which would constitute data leakage. Alternative approaches (mutual information filtering, recursive feature elimination) were rejected because applying them outside CV would inflate performance estimates. (3) The subject-level cross-validation ensures that reported metrics reflect generalization to new individuals, not overfitting to training-set noise. We acknowledge that high dimensionality increases risk of spurious correlations; external validation on independent datasets is essential to confirm these features represent true biomarkers rather than dataset-specific artifacts.

2.5 Strict Subject-Level Cross-Validation

To prevent identity leakage, we implemented a strict subject-level evaluation protocol:

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.

Methodological Design Choices. Two design decisions require justification: (1) *Decision threshold*: the 0.5 probability threshold was used as the scikit-learn default; threshold optimization was deliberately avoided to prevent overfitting to the validation data. In deployment, threshold adjustment based on clinical priorities (e.g., maximizing sensitivity) would be appropriate but should be performed on held-out data. (2) *Number of folds*: five-fold CV was pre-specified to balance training set size ($\sim$122 subjects, sufficient for Random Forest training) with number of independent test evaluations. Fewer folds would reduce test variance but limit training data; more folds would increase computational cost with diminishing returns.

2.6 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].

2.7 Hardware Prototype (Proof-of-Concept)

To demonstrate *design feasibility* for low-resource settings, we developed a proof-of-concept single-channel EEG acquisition system comprising:

- **Microcontroller:** ESP32 (dual-core, Wi-Fi/Bluetooth enabled, \$5)
- **Analog Front-End:** BioAmp EXG Pill (instrumentation amplifier with $\times$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. This prototype represents a proof-of-concept design motivated by feature importance analysis. All classification results reported in this paper were obtained from research-grade 16-channel EEG equipment; validation of classification performance using hardware-acquired signals remains an essential direction for future work.

3 Results

3.1 Quality Control and Selection Bias Analysis

Of 1,932 raw EEG files in the dataset, 1,931 (99.95%) passed quality control and were retained for analysis. 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.

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

Model	Accuracy (%)	AUC (95% CI)	F1 (95% CI)
Majority-class baseline†	50.3	0.500	—
Logistic Regression	76.5 (69.4–83.6)	0.811 (0.74–0.88)	0.768 (0.69–0.85)
SVM (RBF)	81.7 (75.2–88.2)	0.852 (0.79–0.91)	0.823 (0.75–0.89)
Gradient Boosting	83.7 (77.8–89.5)	0.871 (0.81–0.93)	0.837 (0.77–0.90)
Random Forest*	**83.7 (77.8–89.5)**	**0.869 (0.81–0.93)**	**0.837 (0.77–0.90)**

*Pre-specified primary model. †Always predicts HC (majority class: 77/153 = 50.3%).

Note: No pairwise model comparisons performed; overlapping CIs suggest comparable performance among RF, GB, and SVM.

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%

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.

Clinical Utility at Population Prevalence. Sensitivity and specificity alone are insufficient for clinical deployment decisions; predictive values at realistic base rates are essential. At the general population prevalence of $\sim$1%, applying Bayes' theorem yields:

- **Positive Predictive Value (PPV):** $\frac{0.934 \times 0.01}{0.934 \times 0.01 + 0.26 \times 0.99} = 3.5\%$
- **Negative Predictive Value (NPV):** $\frac{0.74 \times 0.99}{0.74 \times 0.99 + 0.066 \times 0.01} = 99.9\%$

The 3.5% PPV at general population prevalence is notably low, meaning approximately 97% of positive screens would be false positives. This underscores that EEG-based screening is not suitable for general population deployment. The 99.9% NPV, though, suggests excellent rule-out capability: a negative screen provides strong reassurance. Clinical utility improves substantially in enriched settings: at 10% prevalence (e.g., first-degree relatives of patients, or individuals presenting with prodromal symptoms), PPV rises to 28.5%, and at 30% prevalence (e.g., psychiatric outpatient clinics), PPV reaches 60.6%. These calculations demonstrate that the tool is most appropriate as a screening aid in high-risk populations or as a triage mechanism to prioritize clinical evaluation, not as a standalone diagnostic.

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.

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

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).

Interpreting Cross-Fold Variance. The 16-point accuracy swing across folds (74.2%-90.0%) requires explanation. Three factors contribute to this variance: (1) *Small test set sizes*: each fold contains only 15-16 subjects, where a single misclassification changes accuracy by $\sim$6 percentage points; (2) *Paradigm heterogeneity*: recordings within each subject span multiple EEG paradigms (resting-state, cognitive tasks, MMN, ASSR), and the distribution of paradigms may differ across subjects assigned to different folds; (3) *Expected statistical variance*: with only 30-31 test subjects per fold, binomial sampling variance yields an expected standard deviation of $\sim$7 percentage points at 85% accuracy. This variance does not indicate model instability but rather reflects the fundamental uncertainty inherent in small-sample clinical classification tasks.

3.5 Feature Importance Analysis

Random Forest feature importance analysis (computed using mean decrease in impurity across 300 trees) revealed that frontal channels contributed disproportionately to classification. The top 10 features included:

- Fp1 theta power (rank 1)
- 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 provides biological rationale for our proof-of-concept single-channel hardware design targeting the frontal region, though validation with hardware-acquired signals remains future work. 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 perva-

sive 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.

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.

Nuanced Interpretation of the Gap. We acknowledge that the 7.2 percentage point difference between recording-level and subject-level accuracy cannot be attributed *entirely* to identity leakage. A portion of this gap reflects the legitimate increased difficulty of cross-subject generalization: subjects vary in baseline EEG characteristics, medication regimens, disease severity, and recording conditions. Even without any information leakage, we would expect some performance reduction when evaluating on held-out subjects versus held-out recordings from known subjects. Nevertheless, the magnitude of this gap and its near-universal omission in published studies underscore the importance of honest reporting. Our contribution lies not in perfectly decomposing this gap, but in demonstrating that subject-level evaluation is both feasible and necessary.

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, including 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 (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 Why This Model Is Not Ready for Clinical Deployment

The 83.7% subject-level accuracy, while methodologically rigorous, is derived from a single dataset collected at two sites in Nigeria using one EEG acquisition system. Three critical generalizability threats preclude clinical deployment without further validation:

1. **Site-Specific Artifact Exploitation:** Machine learning models can inadvertently exploit site-specific patterns—electrical noise signatures, technician protocols, electrode impedance conventions—that masquerade as disease biomarkers. Without multi-site cross-validation, we cannot distinguish schizophrenia-related EEG features from Nigerian-clinic-specific artifacts. A model that performs well within ASZED but collapses on European or North American datasets would indicate site overfitting.

2. **Population and Treatment Heterogeneity:** Genetic background, medication regimens, and comorbidity profiles vary across populations. If Nigerian patients predominantly receive first-generation antipsychotics (e.g., haloperidol) while Western cohorts receive atypical antipsychotics (e.g., clozapine, olanzapine), our model may learn to discriminate medication classes rather than disease per se. Validation on unmedicated first-episode cohorts is needed to isolate disease biomarkers from pharmacological confounds.
3. **Hardware Domain Shift:** The model was trained on research-grade 16-channel wet-electrode EEG systems with high signal-to-noise ratio. Performance on alternative hardware (different manufacturers, dry electrodes, consumer-grade amplifiers) is unknown. The proposed $50 single-channel prototype introduces substantial domain shift: reduced spatial resolution, inferior electrode contact, lower ADC precision, increased susceptibility to motion artifacts. Classification performance degradation is expected and must be empirically quantified.

The path to responsible clinical translation requires: (1) validation on at least two geographically and demographically independent external cohorts, (2) prospective testing on treatment-naive first-episode patients to assess medication-free performance, (3) head-to-head comparison against clinical diagnostic interview (the current standard) to quantify incremental value, and (4) hardware validation demonstrating that classification performance holds when using low-cost acquisition systems. Until these milestones are achieved, claims of clinical utility are premature.

4.5 Hardware Implications for Global Health

Feature importance analysis revealed that frontal channels (Fp1, Fp2) were among the top predictors in our research-grade EEG analysis, providing biological rationale for a single-channel hardware design. 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 potential 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 *design* feasibility as a proof-of-concept. All classification results reported in this paper were obtained from research-grade 16-channel EEG equipment. Future work must validate classification performance using hardware-acquired signals, addressing the substantial domain shift between research-grade and low-cost acquisition systems, a non-trivial challenge involving differences in electrode quality, ADC precision, artifact susceptibility, and signal-to-noise ratio.

4.6 Comparison with Literature

Published EEG-based schizophrenia classification studies have reported accuracies ranging from 75% to 99% [4, 3]. 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). This estimate is more likely to generalize to real-world clinical deployment than inflated metrics from methodologically flawed studies.

5 Limitations

This work has several limitations:

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 Future Directions

6.1 Web Application Deployment and Model Testing

The trained Random Forest model (serialized as a file) enables deployment via web-based applications for real-world screening scenarios. A critical consideration for deployment is the testing strategy: how should incoming EEG recordings be processed to generate predictions?

The ASZED dataset structure—with multiple sessions and paradigms per subject—motivates two distinct testing approaches, each with different trade-offs for clinical deployment.

Strategy 1: Individual Phase/Session Testing. The most straightforward approach treats each EEG recording (phase or session) as an independent sample:

- **Advantages:** Increased data samples (individual phases can be processed independently), realistic clinical scenario (practitioners typically have a single recording from a patient), faster inference time, and ability to evaluate which paradigms (resting-state, cognitive tasks, MMN, ASSR) have highest discriminative power.
- **Implementation:** Load the trained model, extract 264 features from the uploaded EEG file, and generate a prediction with confidence score. This mirrors the current web application architecture.
- **Limitation:** Single-recording predictions may be more susceptible to noise and artifacts, potentially increasing false positive/negative rates compared to multi-recording ensemble approaches.

Strategy 2: Multi-Session Ensemble Testing. For scenarios where multiple recordings are available from the same subject, an ensemble approach can improve diagnostic reliability:

- **Advantages:** Subject-level diagnosis aggregating multiple recordings, increased robustness by averaging out recording-specific noise, and higher confidence through multiple independent predictions.
- **Implementation:** Process all available recordings from a subject, generate individual predictions, and aggregate via majority voting or probability averaging. Final classification determined by aggregated confidence.
- **Limitation:** Requires multiple recordings (may not be available in typical clinical screening), increased processing time, and need for session management in the web interface.

Recommended Deployment Strategy. For initial web application deployment, we recommend a hybrid approach:

1. **Primary Interface:** Support single-recording upload with individual phase prediction (Strategy 1), providing immediate screening results with appropriate uncertainty communication.
2. **Extended Interface:** Allow optional multi-recording upload for users who have multiple EEG sessions, with automatic ensemble aggregation (Strategy 2) and improved confidence estimates.
3. **Paradigm-Specific Guidance:** Given that different EEG paradigms (resting-state vs. cognitive tasks) may have different discriminative power, the application should log which paradigm types yield highest classification confidence across users, informing future model refinement.

6.2 Critical Deployment Considerations

Several technical and methodological considerations are essential for responsible web deployment:

- **Subject-Level Cross-Validation Integrity:** During model validation on the web platform, ensure train/test splits remain at the subject level. If multiple sessions from the same subject are uploaded for testing, all sessions must be evaluated together—never split across training/testing phases—to prevent identity leakage artifacts.
- **Domain Shift Monitoring:** EEG recordings uploaded to the web application may come from different hardware, populations, or recording conditions than the ASZED training data. Implement logging and anomaly detection to identify inputs that fall outside the training distribution, flagging predictions that may be unreliable due to domain mismatch.
- **Confidence Thresholding:** Predictions with probability scores near 0.5 (e.g., 0.4–0.6) should be flagged as "uncertain" and communicated appropriately to users. Binary classification without uncertainty quantification provides false confidence unsuitable for clinical screening.
- **Model Versioning:** The serialized model file should include version metadata, training date, and dataset provenance. As the model is updated with additional training data or refined feature extraction, version tracking ensures reproducibility and enables performance monitoring across deployments.

6.3 Validation Roadmap

Before clinical deployment, the following validation milestones should be completed:

1. **Internal Validation:** Test the web application with held-out ASZED recordings to confirm the model reproduces expected classification metrics.
2. **Hardware Validation:** Evaluate model performance on EEG recordings acquired using the low-cost prototype hardware, quantifying the domain shift between research-grade and affordable acquisition systems.
3. **External Validation:** Test on at least one geographically and demographically independent EEG schizophrenia dataset to assess cross-population generalizability.
4. **Prospective Pilot:** Conduct a small-scale prospective study comparing web application predictions against clinical diagnoses in a real-world screening setting.

7 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 and Code Availability

The ASZED-153 dataset is publicly available through Zenodo (DOI: 10.5281/zenodo.14178398).

Reproducibility Statement. Complete analysis code is available at (to be released upon publication). The analysis environment comprised:

- Python 3.10.12
- MNE-Python 1.5.1 (preprocessing)
- scikit-learn 1.3.2 (classification)
- NumPy 1.24.3, SciPy 1.11.4
- Random seed: 42 (used for all cross-validation splits and model initialization)

All random number generators were seeded to ensure exact reproducibility of results. The preprocessing and feature extraction pipeline processes the full ASZED dataset in approximately 45 minutes on a standard workstation (Intel i7, 32GB RAM).

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] Mosaku SK, Olateju EO, Ayodele KP, et al. An open-access EEG dataset from indigenous African populations for schizophrenia research. *Data in Brief*. 2025;62:111934. DOI: 10.1016/j.dib.2025.111934.

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

[13] 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.

[14] 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.

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

[16] 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.