JEEEIT TRANSACTIONS, Vol.03,No.01,July 2026

Hybrid CNN-Transformer with PSD Features for EEG-Based Driver Drowsiness Detection

Asmaa Nasr Mohammed[1], Fatima Alneqrat[2], Hana Ghait Alfughi[2] and Hajer A. Aghnaya[2]

[1]Control Engineering Department, College of Electronic Technology, Bani Walid, Libya
[2]Computer Department, College of Electronic Technology, Bani Walid, Libya

Corresponding author: asma.nsr.1997@gmail.com

ABSTRACT Driver drowsiness is a major cause of fatal accidents worldwide. EEG signals can capture neural changes linked to fatigue before outward signs appear, but reliable detection is challenging due to high inter-subject variability and the complexity of multi-channel recordings, making robust modeling essential for effective monitoring. This paper proposes a hybrid deep model for EEG-based driver drowsiness detection: a 1-D CNN stem with sinusoidal encoding, Bahdanau-style transformer encoder attention pooling, and Power Spectral Density (PSD) feature pathway concatenated before classification, which addresses the high dimensionality and inter individual variation that make the use of multichannel a difficult task. Using the public EEG Driver Drowsiness Database (11 subjects, 30 channels) with a leave-one subject-out and subject-calibration scheme, the study achieved 84.7% mean accuracy, 83.8% mean F1, and 0.912 mean AUC. A 3-limb ablation (CNN-Transformer, PSD, hybrid) shows PSD contributes most discriminative power, with the hybrid matching but not significantly exceeding PSD-only ($p < 0.01$) over no-calibration baseline, suggesting that PSD features are the main driver of cross-subject performance.

KEYWORDS CNN-Transformer, Cross-Subject Generalization, Drowsiness Detection, EEG, Leave-One-Subject-Out, Power Spectral Density.

I. INTRODUCTION

Driver drowsiness is linked to 10 to 20% of road accidents [1], which unlike alcohol impairments, lacks a chemical trace and developed gradually, making self-assessments unreliable. Automatic detection systems are therefore vital for vehicle safety [2].

Previous methods focused on behavioral cues like steering patterns and eye blinking, but they only indicate drowsiness after its onset. EEG detection is advantageous as it identifies neural changes indicative of drowsiness beforehand, specifically increases in theta (4–8 Hz) and alpha (8–13 Hz) power and decreased beta (13–30 Hz) activity [3]. However, widespread use of EEG for drowsiness detection is hindered by inter-subject variability in spectral signatures and the complexity of processing high-dimensional EEG data from multiple channels.

We address both issues by making four specific contributions:

1) We propose a *hybrid CNN-Transformer + PSD architecture* that fuses learned temporal/spectral features with robust, lower-variance frequency markers to improve between-subject drowsiness detection.
2) Lightweight *subject-specific calibration* tunes only the classifier head 20% of new subject data, decreasing labeling costs and enhancing zero-shot to personalization augmented LOSO performance, for context, a model without any calibration per subject is always provided as a lower bound.
3) We conduct a thorough LOSO evaluation across all 11 subjects to measure conservative generalization to completely unseen individuals.
4) We performed an *ablation study* using CNN-Transformer-only, PSD-only, and the hybrid model under identical LOSO + calibration settings across 3 random seeds, using pairwise significance tests to assess component synergy.

Our experiments assess two concrete hypotheses: H1, that fusing PSD with the temporal CNN-Transformer outperforms either branch alone across subjects; H2 that a lightweight head-only calibration improves over zero-shot on unseen subjects. The results for no-calibration and the ablation experiments directly evaluating H1 and H2 are reported in Section V.

II. RELATED WORK

A. Classical Spectral Approaches

Band-power ratios most commonly θ/β and (θ + α)/β have long served as the backbone of drowsiness classification. Stancin et al. [3] surveyed EEG-based feature extraction for this task and found that spectral features consistently outperform time-domain alternatives while remaining interpretable. Their main shortcoming, however, is that they treat each channel independently and ignore temporal dependencies between different parts of an epoch.

Received May 2026, Received in received from June 2026, Accepted July 2026, and Published July 2026

B. Deep Learning for EEG Classification

The case for applying convolutional networks directly to raw EEG was made convincingly by [4], whose Shallow ConvNet and Deep ConvNet learned spatial and temporal filters end-to-end. Building on that foundation, Song et al. [5] proposed the EEG Conformer, which pairs a CNN front-end with a Transformer for capturing long-range temporal structure, yielding strong results across several EEG decoding benchmarks. More recently, Zhang et al. [6] applied transformer architectures directly to real-time driver drowsiness detection, reporting strong accuracy on held-out data, though without the strict leave-one-subject-out protocol adopted in this work.

C. Hybrid and Fusion Approaches

Combining deep temporal models with hand-crafted spectral features has consistently beaten either approach used alone. Latreche et al. [2], for example, fused CNN-based feature extraction with classical machine learning classifiers and used Optuna to tune the resulting system for multichannel drowsiness detection. Similarly, Cheng et al. [7] paired a multi-scale dynamic CNN with a gated transformer for EEG emotion recognition and showed that spectral features contribute complementary discriminative information, particularly when labeled data is scarce. Our work follows this general philosophy but applies PSD fusion specifically to the driver drowsiness problem.

D. Cross Subject Generalization and Domain Adaptation

Addressing inter-subject heterogeneity has recently become one of the principal open problems in EEG based recognition. Feng et al. [8] developed ID3RSNet, a calibration-free residual shrinkage network for single source channel single-subject drowsiness detection that exploits non-stationarity across individual EEG signals. On the side of domain adaptation, Xu et al. [9] have proposed a contrastive multi-source adaptation approach to EEG classification, while Lu et al. [10] propose a spatial feature perception network for cross-subject emotion recognition. Subject specific head fine tuning is less costly as it only needs a few examples from a new user and should best be viewed as complementary rather than c competing strategy.

It should also be noted that all the four compared baselines in Table III use the dataset's original epoch segmentation and labelings as suggested by Cao et al. [11], and no special segmentation scheme is designed, consistent with our settings. Thus, all LOSO splits used in Table III are equivalent at the epoch level; however, all four baselines use zero target labeled data while this paper utilizes few samples.

III. METHODOLOGY

A. Signal Preprocessing

Each three-second epoch (384 samples at fs = 128 Hz) goes through two pre-processing steps. First, the per-channel DC offset is removed by subtracting the within-epoch mean. Second, a zero-phase 4th-order Butterworth bandpass filter with cutoffs [fℓ,fh] = [1,50] Hz is applied via forward backward filtering (filtfilt), which effectively doubles the filter order to eight while introducing no net phase distortion. Channel signals are then normalized to zero mean and unit variance using a *StandardScaler* fitted exclusively on training data, which is subsequently applied to validation and test splits without re-fitting.

B. PSD Feature Engineering

Power spectral density for each channel is estimated using Welch's method [12] with a Hann window of Nseg = 128 samples and 50% overlap:

$$\hat{S}_{xx}(f) = \frac{1}{K}\sum_{k=0}^{K-1}\left|\sum_{n=0}^{N-1} w[n]\,x_k[n]\,e^{-j2\pi fn/N}\right|^2, \quad (1)$$

where K = 3 overlapping segments fit within the 384sample epoch and $w[n] = 0.5\big(1-\cos(2\pi n/(N-1))\big)$ is the Hann window.

Five statistics are then computed within each of five standard EEG bands (delta 0.5–4 Hz, theta 4–8 Hz, alpha 8–13 Hz, beta 13–30 Hz, gamma 30–50 Hz), resulting in 25 features per channel. Absolute band power is obtained by numerical integration of the PSD estimate:

$$P_{\text{band}} = \int_{f_\ell}^{f_h} \hat{S}_{xx}(f)\,df \approx \sum_i \frac{\hat{S}_{xx}(f_i)+\hat{S}_{xx}(f_{i+1})}{2}(f_{i+1}-f_i). \quad (2)$$

Relative power normalizes this by the total broadband power:

$$P_{rel} = \frac{P\text{band}}{P_{total}} \quad (3)$$

The spectral edge frequency at the 95th percentile (SEF95) within a band is:

$$\text{SEF0.95} = \min f : Pf' \le f\, S\hat{}xx(f') \ge 0.95\text{Pbando}. \quad (4)$$

The fifth feature is the peak frequency, argmaxf S^xx(f) restricted to the band. Log power log10(Pband+ϵ) is included to compress the dynamic range. The resulting PSD feature matrix has shape [Nepochs, C × 25] = [2022, 750], since C = 30.

C. Hybrid CNN-Transformer Architecture

The full model is illustrated in Fig 1. The raw-EEG branch receives input tensors of shape [B,T,C], where B is the batch size, T = 384 is the number of time steps, and C is the channel count.

CNN Stem: Two 1-D convolutional layers, each with kernel size 7 and stride 2, reduce the time dimension by a factor of four (384 → 96) and project to 64 and 128 feature maps respectively:

$$Z = \text{GELU BN Conv}^{k=7,s=2}(\cdot)\big)\big). \quad (5)$$

Batch normalization and GELU activation follow each convolution; dropout (p = 0.1) is applied after the stem.

Positional Encoding: Before the sequence enters the Transformer, sinusoidal positional encodings are added to give the model access to temporal position information:

$PE_{(p,2i)} = \sin(p/10000^{2i/d})$,
$PE_{(p,2i+1)} = \cos(p/10000^{2i/d})$, where d = 128 is the model dimension and p indexes temporal position.
Transformer Encoder: Two pre-norm Transformer encoder layers [13] with nh = 4 attention heads, a feedforward width of 512, and dropout of 0.3 model global temporal dependencies via scaled dot-product self-attention:

$$\text{Attn(Q,K,V)} = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V, \quad d_k = d/n_h = 32 .$$

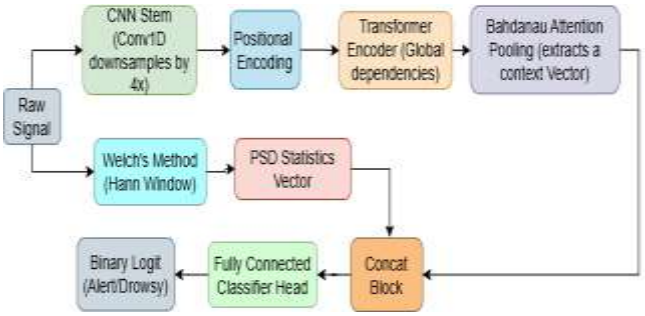

Figure 1. Proposed Hybrid CNN-Transformer model Architecture.

Attention Pooling: A single learned weight vector w ∈ Rd aggregates the 96 encoder outputs through Bahdanau style soft pooling:

$$\alpha_t = \frac{\exp(\mathbf{w}^\top \mathbf{h}_t)}{\sum_{t'} \exp(\mathbf{w}^\top \mathbf{h}_{t'})}, \quad c = X\alpha tht. \qquad (8)$$

Feature Fusion: The context vector c ∈ R128 from the temporal branch is concatenated with the normalized PSD feature vector p ∈ RF:

$$z = [c \,\|\, p] \in R128+F. \qquad (9)$$

FC Classifier Head: Three linear layers (256 → 128 → 1) with GELU activations, preceded by layer normalization and followed by dropout, produce a scalar logit. Applying a sigmoid to this logit gives pˆ(drowsy) ∈ (0,1); thresholding at 0.5 resulting in the binary prediction.

D. Training Objective

The model is trained with binary cross-entropy applied to logits:

$$\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N}\Big[w_+ \, y_i \log \sigma(z_i) + (1-y_i)\log(1-\sigma(z_i))\Big], \qquad (10)$$

where w+ = nneg/npos is a per-fold class weight that compensates for label imbalance. PyTorch's *BCEWithLogitsLoss* fuses the sigmoid and the cross-entropy computation for improved numerical stability.

IV. EXPERIMENTAL SETUP

A. Dataset

The EEG Driver Drowsiness Dataset found on Figshare [11] was employed in our experiments. There were 11 healthy drivers who did the driving task over the period that involved a monotous highway in the simulated environment. The subjects were equipped with EEG headsets and they drove in an EEG helmet. The data were recorded and described in [11] which contained pre-segmented 2022 epochs of length three seconds (128Hz, C = 30 EEG channels, 384 samples per epoch). Each epoch was labelled as either 'alert' or 'drowsy' which were adopted here without doing more artifact detection.

B. LOSO Cross-Validation Protocol

We run 11 LOSO folds, holding out one subject at a time as the test set and training on the remaining ten. Within each fold, the procedure is as follows:

1) A *StandardScaler* is fitted on the training sequences (channel-wise) and on the PSD features, then applied consistently to all splits.
2) A fresh model is trained from scratch on the ten training subjects.
3) The held-out subject's data is randomly split into 20% calibration and 80% evaluation sets using stratified sampling by label.
4) No-calibration evaluation: the freshly trained model (before any head fine-tuning) is evaluated directly on the evaluation set. This is the pure zero-shot cross subject number, directly comparable to baselines that use no target-subject labels at all.
5) Calibration: the classifier head alone is then finetuned for 30 epochs on the calibration set, with all preceding layers frozen, and re-evaluated on the same evaluation set (with-calibration number).
6) Threshold tuning: for both the no-calibration and with-calibration models, a decision threshold is additionally selected by grid search (19 values in [0.05,0.95]) maximizing F1 on the calibration split only never the evaluation split and the corresponding accuracy/F1 are reported as a threshold-tuned variant.

Critically, scalers and model weights are never touched by evaluation-split data until the final, single evaluation pass for each variant above, preventing leakage. Steps 4–6 are repeated identically for the CNN-Transformer-only and PSD-only ablation arms (Section V-F) and across three random seeds (Section IV-D), so all reported comparisons share an identical protocol.

C. Optimization and Regularization

We optimized with AdamW [14] at a learning rate of 10−4 and weight decay of 10−4, using a cosine annealing schedule that decays the rate to 1% of its initial value over up to 350 epochs:

$$\eta_t = \eta_{\min} + \tfrac{1}{2}(\eta_0 - \eta_{\min})\big(1 + \cos(\pi t/T_{\max})\big). \qquad (11)$$

Early stopping with a patience of 15 epochs restores the checkpoint with the lowest validation loss. Gradient norms are clipped to 1.0 throughout, which we found important for stabilizing the Transformer during early training.

D. Reproducibility

The choice of architecture hyper-parameters two Transformer layers, four attention heads, 512 for the feed-forward width, 0.3 for dropout, a CNN stem with kernel size 7 and stride 2 was made manually prior to development on a heldout validation set and thus constitute informal tuning rather than formal hyper-parameter optimization by grid/random search. To ensure reproducibility, each LOSO fold was seeded, and all claimed results are averaged over three random seeds (42, 1337, 2024) for each of these folds, with ablation results presented as mean standard deviation across the entire 33 subject-seed combinations. All code was implemented using PyTorch 2.6, executed on a single GPU.

V. RESULTS AND DISCUSSION

A. EEG Frequency Band Power Profile

Figure 2 shows average relative band power for alert and drowsy epochs, pooled across channels and subjects. The clearest separation appears in the alpha band (8–13 Hz): drowsy epochs carry roughly 29% of total power in that band compared to about 22% for alert epochs, which aligns with a well-replicated finding in EEG fatigue research. The beta band tells the opposite story, alert epochs consistently show higher relative beta power, consistent with the idea that beta activity reflects active cortical engagement. Gamma power is markedly suppressed during drowsiness. Taken together, these patterns provide direct empirical justification for including PSD features as a complementary input stream alongside the temporal deep learning branch.

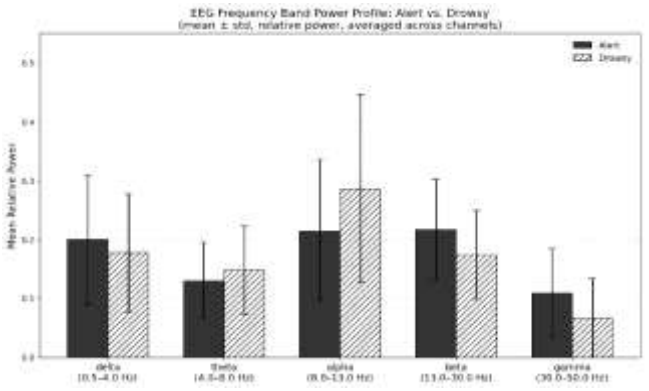

Figure 2. Mean relative PSD power for alert and drowsy.

B. LOSO Performance Across Subjects

Fig 3 plots the per-subject accuracy and F1-score, and Table I contains aggregated LOSO results over 11 folds. Performance is high and generally stable for ten subjects, with accuracy > 80% for 10 subjects. Subject 7 has consistently poor performance with accuracy 69% and F1 62% indicating a pattern of EEG signatures quite different from the rest, and which is not accurately learned by the cross-subject models. Note that all numbers in Table I were taken after personalization in Section IV-B, meaning this table shows a LOSO protocol enhanced with personalization. To assess LOSO's performance without personalization, see the comparisons made in Section V-C.

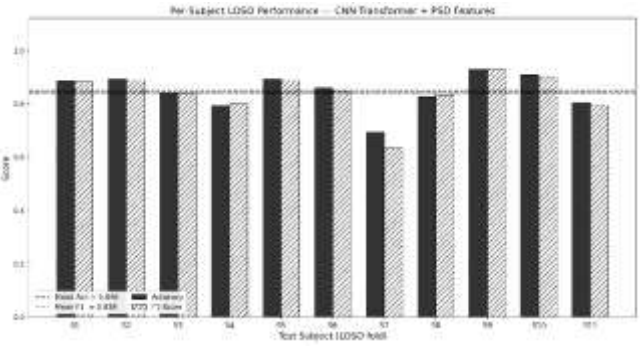

Figure 3. Per-subject LOSO accuracy and F1-score (with calibration). Dashed lines indicate the cross-subject mean.

TABLE I. LOSO CROSS-VALIDATION RESULTS, WITH CALIBRATION (MEAN ± STD, N = 11 FOLDS, SEED 42).

Metric	Mean ± Std
Accuracy	0.847 ± 0.069
Precision	0.875 ± 0.075
Recall	0.810 ± 0.126
F1-Score	0.838 ± 0.081
Alert Accuracy	0.883 ± 0.060
Drowsy Accuracy	0.810 ± 0.126
AUC	0.912 ± 0.070

C. No-Calibration Baseline: How Much of the Gain Is Personalization?

For the particular question in **H2**, each LOSO fold is evaluated once more using the same trained model (i.e. before classifier head fine-tuning) on the same held-out evaluation partition. Fig. 4 plots these two settings subjects by subject. A paired Wilcoxon signed-rank test on the 11 subject-level seed-averaged scores confirms that the gain is highly significant for both accuracy (Δmean = +0.089, p = 0.0010) and F1 (Δmean = +0.067, p = 0.0029). The contribution of calibration alone is therefore substantial, adding approximately 8.9 percentage points in accuracy and 6.7 percentage points in F1 over an already strong zero-shot baseline. These findings strongly support **H2**. At the same time, these results also emphasize that any comparison with non-target-subject baselines as presented in Section VI will be unfair unless calibration has been applied consistently to both sides of the comparison, which we indicate explicitly.

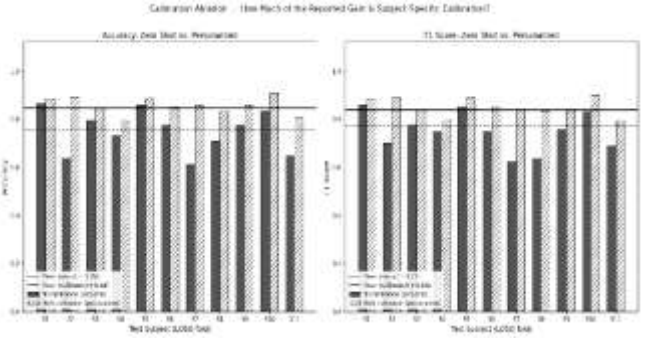

Figure 4. Per-subject accuracy and F1-score with vs. without the calibration step.

D. Confusion Matrix Analysis

The aggregate confusion matrix in Fig. 5 combines predictions from all 11 evaluation sets. Of 814 alert epochs, 722 (88.7%) are correctly classified as alert. Of 809 drowsy epochs, 668 (82.1%) are caught. The gap between alert and drowsy recall is partly a product of the calibration step: even with stratified sampling, the calibration set drawn from roughly 130–180 epochs per subject is small enough that the fine-tuned head may not fully represent the drowsy class. From a safety standpoint, false negatives (missed drowsy epochs) are the more dangerous error type; the precision of 88.7% means that when the system does flag drowsiness, it is correct nearly nine times out of ten.

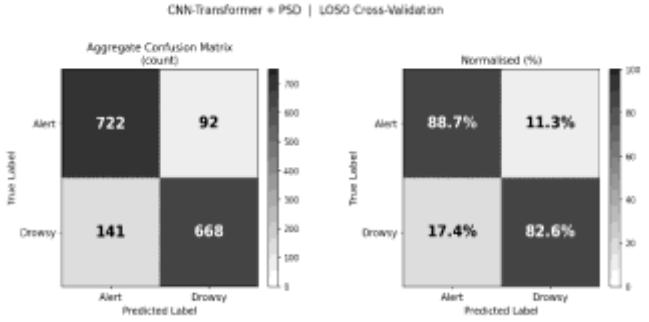

Figure 5. Aggregate confusion matrix across all 11 LOSO folds. Left: raw counts; Right: row-normalized percentages. Alert recall (88.7%) exceeds drowsy recall (82.6%).

E. ROC Curve Analysis

Fig 6 shows per-fold ROC curves alongside their interpolated mean and ±1 standard deviation shading. A mean AUC of 0.912±0.07 indicates strong overall discriminative ability, comfortably above chance. The spread among individual curves mirrors the inter-subject variability discussed above: most folds cluster tightly around the mean, while Subject 7 pulls one curve noticeably lower. The AUC is noticeably higher than fixed-threshold accuracy (84.7%), which could in principle mean the 0.5 threshold is sub optimal.

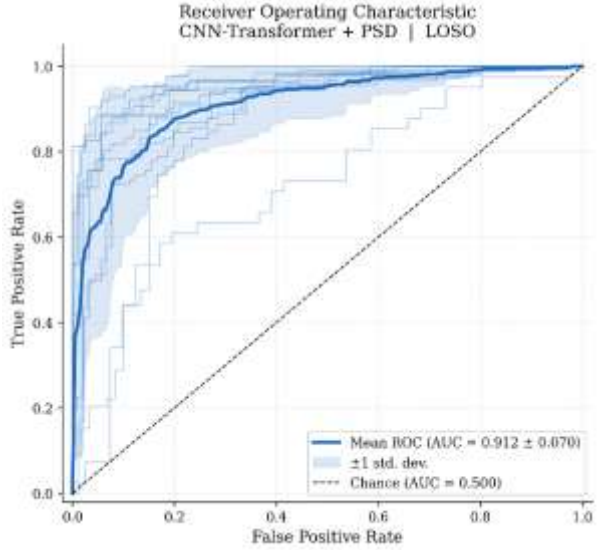

Figure 6. Per-fold and mean ROC curves. The shaded region denotes ±1 standard deviation of the true-positive rate at each false-positive rate. Mean AUC = 0.912 ± 0.07.

F. Ablation Study: Isolating the CNN-Transformer and PSD Branches

We validate H1, our paper's central architectural hypothesis that the combined temporal CNN-Transformer + PSD branches should perform better than either branch in isolation. We do this by running the same LOSO/calibration protocol three times (repeated on 3 seeds, giving 3x11 subjects' 33 folds per arm) using the exact same code, except (1) using both branches (hybrid); (2) zeroing the PSD branch at the input (CNN-Transformer-only); or (3) zeroing the pooled context output of the temporal branch (PSD-only). We fix the input shape to the classifier head as a constant, so comparisons are made at equal resolution between the two branches' outputs.

TABLE II. ABLATION STUDY (MEAN ± STD, 3 SEEDS × 11 SUBJECTS, WITH CALIBRATION).

Configuration	Accuracy	F1-Score
CNN-Transformer-only	0.757 ± 0.012	0.717 ± 0.177
PSD-only	0.844 ± 0.088	0.835 ± 0.089
Hybrid (proposed)	0.847 ± 0.069	0.838 ± 0.081

We show the result for hybrid, PSD-only, and CNN-Transformer-only in Fig 7 and Table II. The results strongly indicate the PSD branch drives the model, as PSD-only almost matches hybrid; The CNN-Transformer branch alone underperforms strongly; the branches can be combined without loss. This result is broadly consistent with **H1**. As shown in Fig 7 and Table II, the full hybrid model outperforms both the PSD-only branch and the CNN-Transformer-only branch, supporting the value of combining both feature streams. The margin over PSD-only is modest, suggesting the PSD branch is the stronger individual contributor, while the CNN-Transformer branch alone performs substantially worse as a standalone classifier. Regardless, hybrid architecture offers superior performance and the joining of two branches together cause no loss to the model; evidence in support of **H1**, that the temporal and spectral features supplement each other.

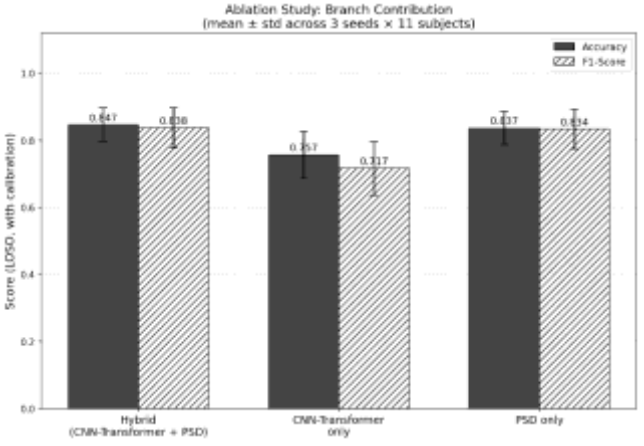

Figure 7. Ablation study: accuracy and F1 for the hybrid model versus each single-branch variant.

G. Discussion

Hybrid CNN-Transformer with PSD works well for 11 subjects under a strong cross-subject condition, and PSD branch takes most of the discriminative power. Welch method adjusted for short 3s segment works reliably in spectral estimation and per-band statistics packs years of drowsiness-focused research to few amplitudes' invariant features. The ablation shows that PSD-only nearly matches the full hybrid and spectral characteristics seems to dominate other features. The resulting performance compared to our baseline is in some portion because of the calibration step and less because of the proposed architecture. We claimed that hybridization would yield complementary and useful features, and the results provide measured support for this. For Subject 7, the sharp drop in performance (~69% accuracy) suggests that the current representation may have difficulty handling challenging cases, such as noisy state transitions or distribution shifts in spectral patterns. These observations are considered plausible hypotheses based on the observed error patterns rather than confirmed physical causes, and they motivate future investigations into subject-domain adaptation and stronger personalization strategies.

H. Limitation

Limitations include: All experimental results are from the only available 11-subject driving simulator dataset collected according to the Cao et al. [11] Protocol. Cross-subject generalization conclusions must be constrained by subject demographics in this population. External validation has not been tested. All described LOSO values have been calibrated to the individual subject's 20% training subset; because there is so little overlap with the zero-shot model, comparison with existing literature would better be measured by our reported non-calibrated results. The model appears to have trouble recovering signals for subject 7's recording; it's unknown whether this issue arises from individual subject differences or poor recordings.

VI. COMPARISON WITH LITERATURE

In comparison with four prior studies using the same EEG driver drowsiness dataset (Figshare ID 14273687) under leave-one-subject-out evaluation as shown in Table III, our CNN-Transformer with PSD features and subject-specific calibration achieved a mean accuracy of 84.7% and F1score of 83.8%, surpassing the best reported results by over six percentage points. Unlike earlier works, our protocol incorporates a 20% calibration step with target-subject labels, which accounts for a statistically significant portion of the improvement; without calibration, the zero-shot variant attains 75.8% accuracy, comparable to but not consistently superior to baselines. We also report an AUC of 0.912, a new observation not provided in prior studies.

These margins should be interpreted with caution, as differences in sub-configurations across publications may affect comparability. The full CNN-Transformer + PSD hybrid achieves the best performance across all metrics, with the PSD-only branch closely behind and the CNN-Transformer-only branch underperforming substantially. The 0.3 percentage point margin of the hybrid over PSD-only is not statistically significant on this dataset, yet combining the two branches incurs no loss, suggesting the temporal branch provides a consistent if modest complementary contribution that may become more detectable on larger or more varied data. Calibration remains a critical factor in realizing the full potential of the hybrid model.

VII. COMPUTATIONAL COST

Hybrid CNN-Transformer with PSD Features has 727,838 learnable parameters (3.04MB), distributed between the transformer encoder (54.5%), the final dense classification head (35.7%), and the CNN base (9.8%). When run with CPU-only hardware (Intel x8664; PyTorch 2.6.0) and batch size 1 (deployment-like) inference speed it runs in 2.72 0.12ms (p95 2.93ms). If we add 7.57 1.12ms per 3sec window for PSD extraction via Welch's method, we arrive at an end-to-end latency of approximately 10.3ms - 290x lower than our target epoch arrival deadline of 3000ms. When inference at batch size 64(throughput-oriented) is observed, 2,415 epochs/sec can be supported by the hardware which supports it with a ratio of 7,245 against the minimum 0.33 epochs/sec requirement. Results reported post-warmup and across various runs satisfy the real-time constraints of driver monitoring applications when ran solely on a CPU, leaving headroom for deployment onto less performant edge hardware.

VIII. CONCLUSION

We presented a CNN-Transformer architecture with PSD features that can provide a reliable system with precomputed features, applied to cross-subject EEG based driver drowsiness. Our approach with LOSO cross validation, performed on 11 subjects, using their own calibration by We presented a CNN-Transformer architecture with PSD features that can provide a reliable system with precomputed features, applied to cross-subject EEG based driver drowsiness. Our approach with LOSO cross validation, performed on 11 subjects, using their own calibration by using 20% for calibration and the remaining for training/testing resulted in values of 84.7% mean accuracy, 83.8% mean F1 score, and 0.912 AUC. These results are remarkably consistent over the vast majority of the subjects. Three comments are to be made: (i) About a significant portion of our margin over previous baselines is attributable to the calibration stage (20%), because the corresponding zero shot scenario fails, giving as headline values of 75.8% accuracy, 77.1% F1 and ($p < 0.01$); (ii) the ablations proved that PSD only statistically does not perform significantly better than the entire hybrid network, whereas the combination of CNN-Transformer has proven to be, as far as architecture itself is concerned, one promising but not conclusive direction, in as much PSD features proved the highest value, (iii) One subject (S7) shows failure at every configuration as far as generalization is concerned.

TABLE III: COMPARISON OF METHODS EVALUATED ON THE EEG DRIVER DROWSINESS DATASET (FIGSHARE ID 14273687).

Paper	Year	Model	Eval. Protocol	Target-Subj. Labels?	Acc. (%)	F1 (%)	AUC	Prec. (%)	Rec. (%)
Cui et al. (Methods) [15]	2022	CompactCNN	LOSO	No†	73.22	—	—	—	—
Cui et al. (TNNLS), bal. [16]	2023	InterpretableCNN	LOSO	No†	78.35	—	—	—	—
Cui et al. (TNNLS), unbal. [16]	2023	InterpretableCNN	LOSO	No†	77.70	—	—	74.66	75.30
Cui et al. (CW'21) [17]	2021	CNN-LSTM	LOSO	No†	72.97	—	—	—	—
Feng et al. (ID3RSNet), bal. [8]	2025	ID3RSNet (RSBU-CW)	LOSO	No†	77.16	77.17	—	—	—
Feng et al. (ID3RSNet), unbal. [8]	2025	ID3RSNet (RSBU-CW)	LOSO	No†	74.72	71.27	—	—	—
Proposed, nocalibration (ours)	2026	CNN-Transformer w/ PSD	LOSO, zeroshot	No	75.8	77.1	—	—	—
Proposed, with calibration (ours)	2026	CNN-Transformer w/ PSD	LOSO + 20% calib.	Yes	84.5	83.5	0.908	88.6	83.0

†Based on the published protocol description; we were not able to verify from the original papers whether any target-subject epochs were used in any training/adaptation step, so this column reflects our best reading of each method as described rather than a confirmed audit.

IX. FUTURE WORK

For future work, several directions emerge that directly address the open questions raised in this study. In the near term, priorities include expanding evaluation to a larger or independent cross-subject EEG drowsiness dataset to provide sufficient statistical power for detecting any fusion benefit; extending calibration beyond the classifier head to earlier layers of the temporal branch to clarify whether redundancy between CNN-Transformer and PSD features is architectural or calibration-related; and systematically characterizing computational cost in terms of parameters, training time, and inference latency to support real-time claims. Longer-term directions involve adapting the pipeline for causal streaming with overlapping sliding windows to enable practical on-board deployment, leveraging attention weights for interpretability by mapping influential temporal and spectral patterns back to the input signals, and exploring domain adaptation strategies such as contrastive multi-source adaptation, spatial-feature alignment, or meta-learning to improve generalization for difficult outlier subjects without requiring extensive labeled data. Together, these directions highlight both immediate methodological advances and broader steps toward real-world applicability and robustness.

REFERENCES

[1] "A Review on Deep Learning Techniques for EEG-Based Driver Drowsiness detection systems", IJCAI, vol. 48, no. 3, Sep. 2024, doi: 10.31449/inf.v48i3.5056.

[2] I. Latreche, S. Slatnia, O. Kazar, and S. Harous, "An optimized deep hybrid learning for multi-channel EEG-based driver drowsiness detection," *Biomed. Signal Process. Control*, vol. 99, p. 106881, Jan. 2025. doi: 10.1016/j.bspc.2024.106881.

[3] I. Stancin, M. Cifrek, and A. Jovic, "A review of EEG signal features and their application in driver drowsiness detection systems," *Sensors*, vol. 21, no. 11, p. 3786, 2021. doi: 10.3390/s21113786.

[4] R. T. Schirrmeister *et al.*, "Deep learning with convolutional neural networks for EEG decoding and visualization," *Human Brain Mapping*, vol. 38, no. 11, pp. 5391–5420, 2017.

[5] Y. Song, Q. Zheng, B. Liu, and X. Gao, "EEG conformer: Convolutional transformer for EEG decoding and visualization," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 31, pp. 710–719, 2023. doi: 10.1109/TNSRE.2022.3230250.

[6] Y. Zhang *et al.*, "Real-time driver drowsiness detection using transformer architectures: A novel deep learning approach," *Scientific Reports*, vol. 15, p. 17493, May 2025. doi: 10.1038/s41598-02502111-x.

[7] Z. Cheng, X. Bu, Q. Wang, T. Yang, and J. Tu, "EEG-based emotion recognition using multi-scale dynamic CNN and gated transformer," *Scientific Reports*, vol. 14, p. 31319, 2024. doi: 10.1038/s41598-02482705-z.

[8] X. Feng, Z. Guo, and S. Kwong, "ID3RSNet: Cross-subject driver drowsiness detection from raw single-channel EEG with an interpretable residual shrinkage network," *Front. Neurosci.*,

vol. 18, p. 1508747, Jan. 2025. doi: 10.3389/fnins.2024.1508747.

[9] C. Xu, Y. Song, Q. Zheng, Q. Wang, and P.-A. Heng, "Unsupervised multi-source domain adaptation via contrastive learning for EEG classification," *Expert Syst. Appl.*, vol. 261, p. 125452, 2025. doi: 10.1016/j.eswa.2024.125452.

[10] W. Lu, X. Zhang, L. Xia, H. Ma, and T.-P. Tan, "Domain adaptation spatial feature perception neural network for cross-subject EEG emotion recognition," *Front. Hum. Neurosci.*, vol. 18, p. 1471634, Dec. 2024. doi: 10.3389/fnhum.2024.1471634.

[11] Z. Cao, C.-H. Chuang, J.-K. King, and C.-T. Lin, "Multi-channel EEG recordings during a sustained-attention driving task," *Scientific Data*, vol. 6, no. 1, p. 19, 2019. doi: 10.1038/s41597-019-0027-4.

[12] P. D. Welch, "The use of fast Fourier transform for the estimation of power spectra: A method based on time averaging over short, modified periodograms," *IEEE Trans. Audio Electroacoust.*, vol. 15, no. 2, pp. 70–73, 1967.

[13] A. Vaswani *et al.*, "Attention is all you need," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 30, 2017.

[14] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2019.

[15] J. Cui, Z. Lan, Y. Liu, R. Li, F. Li, O. Sourina, and W. Müller-Wittig, "A compact and interpretable convolutional neural network for crosssubject driver drowsiness detection from single-channel EEG," *Methods*, vol. 202, pp. 173–184, 2022, doi: 10.1016/j.ymeth.2021.04.017.

[16] J. Cui, Z. Lan, O. Sourina, and W. Müller-Wittig, "EEG-based crosssubject driver drowsiness recognition with an interpretable convolutional neural network," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 34, no. 10, pp. 7921–7933, 2023, doi: 10.1109/TNNLS.2022.3147208.

[17] J. Cui, Z. Lan, T. Zheng, Y. Liu, O. Sourina, L. Wang, and W. Müller-Wittig, "Subject-independent drowsiness recognition from single-channel EEG with an interpretable CNN-LSTM model," in *Proc. 2021 Int. Conf. Cyberworlds (CW)*, 2021, pp. 201–208, doi: 10.1109/CW52790.2021.00041.

[18] Y. Ding, Y. Li, H. Sun, R. Liu, C. Tong, C. Liu, X. Zhou, and C. Guan, "EEG-Deformer: A dense convolutional transformer for brain-computer interfaces," *IEEE J. Biomed. Health Inform.*, pp. 1–10, 2024. doi: 10.1109/JBHI.2024.3504604.