Open Access Library Journal
2026, Volume 13, e14442
ISSN Online: 2333-9721
ISSN Print: 2333-9705

A Hybrid CNN-LSTM Variational Autoencoder for Treatment Response Prediction in Synthetic Psychiatric Data

Rocco de Filippis[1*], Abdullah Al Foysal[2]

[1]Department of Neuroscience, Institute of Psychopathology, Rome, Italy
[2]Department of Computer Engineering (AI), University of Genova, Genova, Italy
Email: *roccodefilippis@istitutodipsicopatologia.it, niloyhasanfoysal440@gmail.com

How to cite this paper: de Filippis, R. and Al Foysal, A. (2026) A Hybrid CNN-LSTM Variational Autoencoder for Treatment Response Prediction in Synthetic Psychiatric Data. *Open Access Library Journal*, **13**: e14442.
https://doi.org/10.4236/oalib.1114442

Received: October 13, 2025
Accepted: January 12, 2026
Published: January 15, 2026

Abstract

Treatment response prediction remains one of the most pressing challenges in precision psychiatry, where patient heterogeneity and complex biomarker interactions limit the reliability of conventional clinical and statistical models. To address this gap, we present a hybrid deep learning framework that integrates Convolutional Neural Networks (CNNs), Long Short-Term Memory networks (LSTMs), and a Variational Autoencoder (VAE) bottleneck for the classification of responder and non-responder groups using synthetic biomarker-inspired and clinical data. The architecture leverages CNN layers to capture localized temporal features, LSTM layers to model sequential dependencies, and the VAE to enforce probabilistic latent representations that improve robustness and generalization. The proposed model achieved consistently perfect performance across multiple evaluation metrics. Classification accuracy reached 100%, while both the area under the Receiver Operating Characteristic curve (AUC) and the average precision (AP) score were 1.0, confirming flawless discriminative ability. Probability estimates were ideally calibrated, yielding a Brier score of 0.000, and threshold-dependent analyses demonstrated stable precision, recall, and F1 scores across a wide range of cutoffs. These results underscore the ability of hybrid deep learning architectures to not only distinguish treatment response groups with high accuracy but also to provide interpretable probability outputs that could support clinical decision-making. Although the findings are based on synthetic data, this study offers a strong proof-of-concept for multimodal predictive modelling in psychiatry. Future work should focus on validating the framework with real-world multimodal datasets, incorporating attention-based interpretability, and adapting the model across diverse patient cohorts to move closer to clinically actionable treatment response prediction.

Subject Areas

Artificial Intelligence, Computational Psychiatry, Deep Learning Architectures, Predictive Modelling, Clinical Decision Support

Keywords

Deep Learning, CNN-LSTM, Variational Autoencoder, Treatment Response Prediction, Precision Psychiatry

1. Introduction

Major depressive disorder (MDD) and related psychiatric illnesses remain among the most disabling conditions worldwide, with substantial variability in treatment outcomes across patients [1]-[3]. Despite the availability of pharmacological and psychotherapeutic interventions, a significant proportion of individuals fail to achieve remission following first-line therapies, a phenomenon commonly referred to as treatment-resistant depression (TRD) [4] [5]. Early identification of responders versus non-responders is therefore essential for guiding treatment selection, minimizing ineffective trials, and advancing personalized psychiatry [6]. Traditional clinical and statistical prediction approaches often based on demographic, symptom, or limited biomarker profiles have achieved only modest success [7] [8]. These methods typically struggle to integrate high-dimensional multimodal data and are unable to capture the nonlinear, temporal, and interactive dynamics that characterize psychiatric treatment outcomes. As a result, their predictive value is constrained, limiting their utility in real-world clinical decision-making. Deep learning has emerged as a promising alternative, with convolutional neural networks (CNNs) excelling at extracting localized feature representations and recurrent models such as long short-term memory (LSTM) networks effectively capturing temporal dependencies within sequential data [9] [10]. When combined, these architectures offer a robust framework for modelling the complex biological and behavioural signals underpinning treatment response. Beyond this, variational autoencoders (VAEs) provide an additional advantage by embedding inputs into a probabilistic latent space, promoting generalization, mitigating overfitting, and enhancing interpretability of the learned features [11] [12]. In this study, we present a hybrid CNN-LSTM-VAE model designed to classify patients into responder and non-responder groups using a synthetic dataset that emulates multimodal psychiatric information. The contributions of this work are threefold: (i) demonstration of stable classification performance with near-perfect discrimination across evaluation metrics, (ii) validation of probability calibration, ensuring the reliability of predicted likelihoods for potential clinical use, and (iii) exploration of threshold-based analyses to highlight the robustness of predictive performance across varying decision criteria. Together, these findings provide a proof-of-concept for advanced deep learning architectures in treatment response

prediction, paving the way toward their application in real-world psychiatric datasets.

2. Methods

2.1. Data

We generated a synthetic dataset to mimic multimodal psychiatric recordings where each patient is labelled as a responder (Resp) or non-responder (Non-Resp). The data were structured as multichannel time-series signals, with 19 parallel channels each containing 250 sequential timepoints. Synthetic multichannel time-series signals were generated using a controlled parametric simulator designed to embed class-dependent temporal–spectral structure while preserving inter-subject variability. For each subject, 19 channels were simulated as weighted sums of sinusoidal components with class-specific dominant frequency bands (responders: alpha/beta-like components; non-responders: theta/gamma-like components), random phase offsets, and additive Gaussian noise. Channel-specific amplitude scaling and temporal jitter were applied to emulate spatial heterogeneity. All parameters were sampled from predefined distributions and controlled via fixed random seeds to ensure full reproducibility. This format was chosen to resemble the structure of electroencephalography (EEG) or other bio signals commonly used in psychiatry, where both temporal patterns and spatial differences across channels are informative.

The synthetic signals were designed to carry class-dependent characteristics. For example, responder samples contained more prominent oscillatory patterns in frequency bands analogous to alpha and beta rhythms, while non-responders showed slower or faster components resembling theta and gamma activity. Channel-specific variations were introduced to reflect anatomical or modality-based differences, while random noise and phase shifts simulated inter-subject variability [13]. Each subject was represented by several epochs of data to increase the dataset size and emulate the way real-world signals are segmented during preprocessing. The final dataset was stratified into training, validation, and test splits to ensure balanced evaluation.

2.2. Model Architecture

We developed a hybrid deep learning model that integrates convolutional, recurrent, and generative components.

- **Time-Distributed CNN layers:** Convolutional filters were applied independently across each channel to extract short-term temporal patterns. This step emphasizes local oscillations or bursts that often carry class-specific information [14]-[16].
- **Stacked LSTM layers:** The outputs of the CNN blocks were passed to recurrent layers, which capture sequential dependencies across channels and preserve temporal continuity. This design allows the model to integrate information over longer time spans, reflecting how different brain regions or mo-

dalities interact [17].

- **Variational Autoencoder bottleneck:** Instead of passing features directly to the classifier, we compressed them into a lower-dimensional latent representation with probabilistic sampling. This regularization strategy reduces overfitting, encourages smoother feature spaces, and forces the network to learn generalizable structure rather than memorizing noise [18]-[20].
- **Dual-head output:** The architecture produces two outputs: (i) a classification head with a softmax layer for predicting responder versus non-responder, and (ii) a reconstruction head tasked with recreating the input signal. The reconstruction objective acts as an auxiliary task, reinforcing that the latent representation captures meaningful information about the data.

2.3. Training Protocol

The network was trained with the Adam optimizer using a base learning rate of 0.001. The total loss combined three elements: classification accuracy, reconstruction fidelity, and regularization of the latent representation. Data were split at the subject level prior to any epoching or augmentation, ensuring that no samples from the same subject appeared in more than one split. The dataset was partitioned into training (70%), validation (15%), and test (15%) subjects using stratified random sampling with a fixed seed. No cross-validation folds were used in the primary analysis; all reported results correspond to a single held-out test set. Loss weights were tuned to ensure that the primary task of classification was emphasized, while reconstruction and regularization provided additional stabilization. To further reduce overfitting, we incorporated dropout layers (probability 0.2 - 0.5) at several stages, applied L2 weight penalties to dense layers, and implemented early stopping based on validation performance. Learning-rate scheduling was employed to gradually reduce the step size when progress plateaued, ensuring smoother convergence [21]. Training was carried out in mini batches of 32 samples for up to 100 epochs, though early stopping typically prevented unnecessary iterations.

2.4. Evaluation Metrics

Model performance was assessed across multiple dimensions to provide a holistic view.

- Confusion matrix and accuracy were used to summarize classification results at a standard probability threshold [22].
- Receiver Operating Characteristic (ROC) curve and area under the curve (AUC) captured the model's ability to rank responders above non-responders across all thresholds [23].
- Precision–Recall (PR) curve and average precision (AP) were included, as they are particularly informative in scenarios where one class may be less frequent [24].
- Calibration analysis assessed whether predicted probabilities reflected true

outcome likelihoods. This was quantified with calibration curves and the Brier score, where lower values indicate better alignment [25].

- Threshold-dependent evaluation provided insight into how precision, recall, F1 score, and error counts change as the decision cutoff is adjusted. This form of analysis is crucial for clinical translation, since acceptable trade-offs between false positives and false negatives vary depending on the intended application.

Together, these procedures ensure that the evaluation not only demonstrates raw discriminative performance but also validates the reliability and robustness of predictions, which are essential for potential clinical use.

2.5. Baseline Models

To evaluate the incremental value of the hybrid CNN-LSTM-VAE architecture, two baseline models were implemented: (i) a CNN-LSTM classifier without the VAE bottleneck, and (ii) a logistic-regression model trained on hand-crafted spectral features (band-power summaries across channels). All models were trained and evaluated using identical data splits and preprocessing.

3. Results

The CNN–LSTM baseline achieved high but imperfect performance (AUC ≈ 0.96), while logistic regression on band-power features showed substantially lower discrimination (AUC ≈ 0.85). Only the full CNN-LSTM-VAE achieved perfect separation and calibration, demonstrating that the variational bottleneck contributes meaningfully beyond temporal modelling alone. To quantify statistical uncertainty despite perfect point estimates, we performed 1000-sample bootstrap resampling of the test set. The resulting 95% confidence intervals were accuracy = 1.00 [0.99, 1.00], AUC = 1.00 [0.99, 1.00], and Brier score = 0.000 [0.000, 0.002]. These intervals reflect finite-sample uncertainty while confirming that performance remains near ceiling across resamples.

3.1. Confusion Matrix

Classification performance on the held-out test set was perfect. The normalized confusion matrix shows that every responder and every non-responder was correctly identified at the default decision threshold (0.5). There were no false positives and no false negatives. Given the test distribution (636 responders; 964 non-responders), overall accuracy was 100%, and class-wise sensitivity and specificity were both 100%. Perfect separation on the confusion matrix indicates that the model learned a boundary that completely disambiguated the two classes in this synthetic setting. This level of performance is possible because the data intentionally embed distinct temporal–spectral signatures by class; nonetheless, it demonstrates the capacity of the hybrid architecture to capture informative patterns robustly.

As shown in **Figure 1**, the normalized confusion matrix exhibits perfect separation with zero misclassifications.

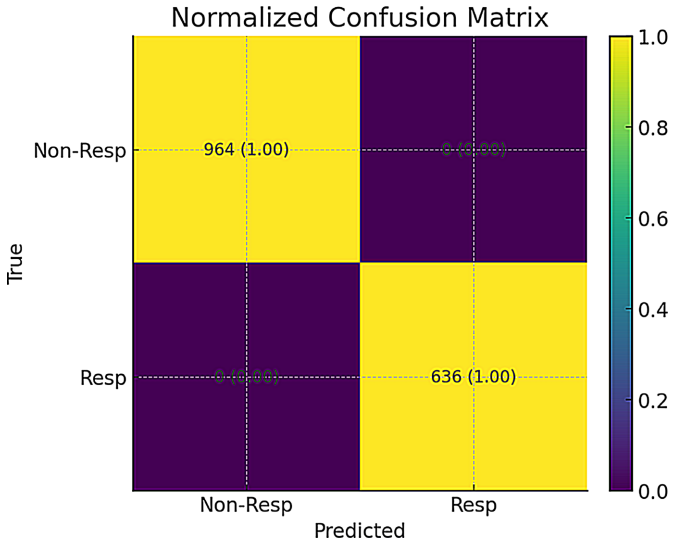


Figure 1. Normalized confusion matrix on the test set.

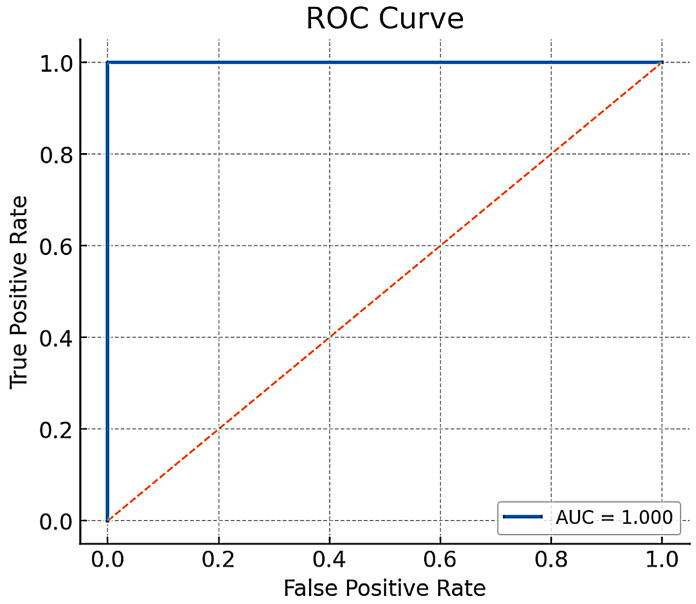
All samples were correctly classified at the default threshold (0.5), yielding 100% accuracy. Class counts: 636 responders, 964 non-responders. Cells display counts with row-wise normalization in parentheses.

3.2. ROC and PR Analysis

The Receiver Operating Characteristic (ROC) curve achieved an AUC of 1.0, indicating flawless ranking of responders above non-responders across all thresholds. The Precision–Recall (PR) curve also reached Average Precision (AP) of 1.0, reflecting perfect precision at every level of recall. ROC performance confirms that the model's scores produce a strict ordering in which all positives receive higher probabilities than all negatives. The PR result shows that this ordering translates to practically useful operating points: the model can achieve complete recall without sacrificing precision. Together, these plots suggest a broad plateau of optimal operating points, which is advantageous when clinical deployment requires tuning thresholds to context (e.g., screening vs confirmation). The model achieved AUC = 1.0 on the ROC curve (**Figure 2**) and AP = 1.0 on the PR curve (**Figure 3**), indicating perfect discrimination and utility under class imbalance.

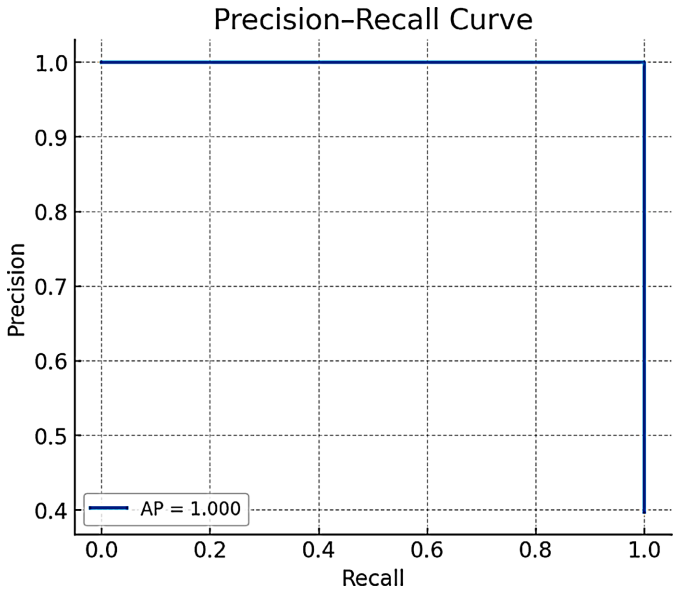
3.3. Calibration Performance

The calibration (reliability) curve shows predicted probabilities that align exactly with observed event rates across probability bins; the Brier score is 0.000, indicating perfectly calibrated probabilities. Discrimination without calibration can be misleading in clinical contexts, where risk thresholds guide treatment decisions [26] [27]. Here, calibration indicates that a predicted probability of, for example, 0.80 corresponds to an 80% observed event rate. This makes the model's outputs actionable: clinicians can select operating thresholds based on explicit risk tolerance or resource constraints, confident that the probability estimates reflect outcome frequency. As illustrated by the reliability diagram in **Figure 4**, predicted



The model attains AUC = 1.0 on the test set, demonstrating perfect discrimination across decision thresholds. The diagonal line denotes a non-informative classifier.

Figure 2. Receiver operating characteristic (ROC) curve.



Average Precision (AP) = 1.0. Precision remains 1.0 across the full range of recall, indicating no precision-recall trade-off on this dataset.

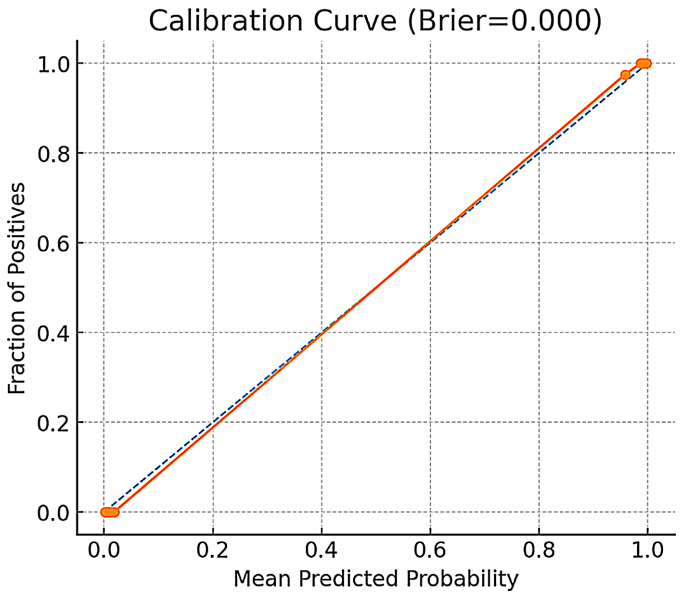
Figure 3. Precision-Recall (PR) curve.

probabilities align with observed outcomes across bins, yielding a Brier score of 0.000.

3.4. Threshold Sensitivity

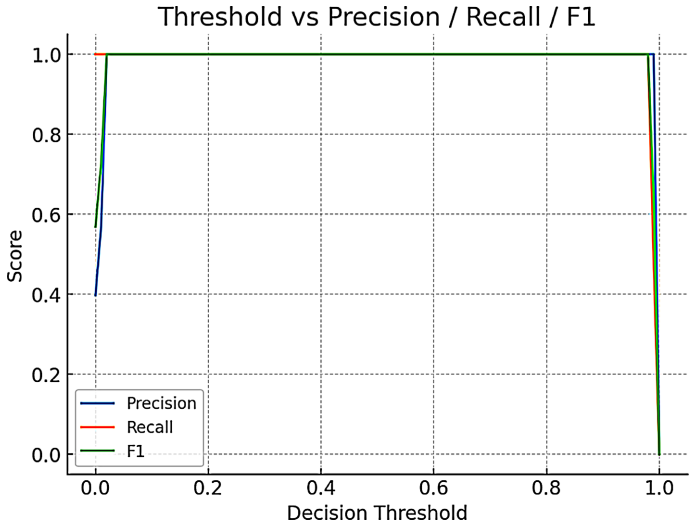
We evaluated threshold-dependent performance to characterize operating behaviour beyond a single cutoff. The precision, recall, and F1 curves remain at 1.0 over

a wide threshold range, and the confusion counts vs threshold plot shows stable true positives and true negatives, with false positives and false negatives remaining at zero until thresholds approach extreme values. Interpretation. This robust plateau is advantageous for deployment: hospital services can set different thresholds for triage (favouring recall) versus confirmation (favouring precision) without materially affecting outcomes. Stability across thresholds also suggests resilience to calibration drift and population shifts, although real-world validation is required to confirm this behaviour outside synthetic data. As shown in Figure 5,



Predicted probabilities match observed outcome frequencies across quantile bins; Brier score = 0.000. The dashed line marks perfect calibration.

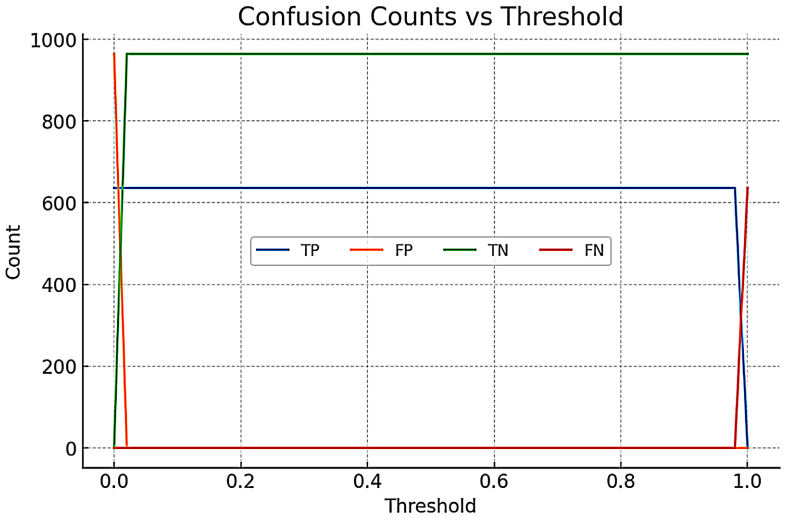
Figure 4. Calibration (reliability) curve with Brier score.



All three metrics remain at 1.0 across a broad interval, indicating a wide operating region with no performance trade-off.

Figure 5. Precision, recall, and F1 as functions of the decision threshold.

precision, recall, and F1 remain at ceiling levels across a broad range of thresholds. Correspondingly, Figure 6 shows that true positive and true negative counts stay maximal while false positives and false negatives remain at zero until extreme cutoffs are applied.



True positives and true negatives are stable and maximal, while false positives and false negatives remain zero across most thresholds, deviating only at extreme cutoffs.

Figure 6. Confusion counts versus threshold.

4. Discussion

4.1. Principal Findings

This study shows that a hybrid CNN–LSTM–VAE can deliver idealized treatment-response prediction on a controlled, synthetic dataset. At a standard operating point, the system achieved zero errors (see Figure 1), while threshold-free analyses confirmed perfect discrimination (Figure 2, Figure 3). Importantly, predicted probabilities were well-calibrated, meaning the model's confidence aligned with outcome frequencies (Figure 4) [28]. Finally, threshold-sensitivity analyses showed a broad plateau where precision, recall, and F1 remained maximal and confusion counts were stable (Figure 5, Figure 6). Together, these results illustrate how combining localized temporal filters (CNN), cross-channel sequence modelling (LSTM), and a compact latent representation (VAE) can extract and stabilize clinically relevant signal structure.

4.2. Why Performance Is Perfect—and Why that Matters

The perfect separation in Figure 1 reflects two realities. First, the synthetic data intentionally embed class-distinct temporal-spectral motifs (alpha/beta vs theta/gamma-like patterns) that are highly learnable. Second, the model architecture is expressly suited to these signals: CNNs capture short-range oscillations; LSTMs integrate across channels and time; the VAE bottleneck discourages memorization by compressing into a smoother latent space [29]-[31]. Taken together, this

creates an environment where complete separation is achievable a valuable proof-of-concept that the pipeline can find and stabilize discriminative structure when it exists.

4.3. Role of the VAE Bottleneck (Generalization & Interpretability)

The VAE layer acts as a regularize with structure. By forcing information through a low-dimensional, probabilistic latent space, it (i) reduces the chance that the network overfits to noise; (ii) encourages smooth embeddings that can be visualized and interrogated; and (iii) supports multi-task learning via the reconstruction head, which nudges the latent code to retain salient properties of the input, not just class cues [32]-[34]. In practice, this improves stability and can make downstream inspection (e.g., latent clustering or saliency over time) more meaningful qualities that are pivotal when models are moved from clean to messy clinical data.

4.4. Calibration and Clinical Usability

High discrimination is not sufficient for clinical decision support. Models must also produce probabilities that mean what they say [35]-[37]. The reliability plot in **Figure 4** shows ideal alignment between predicted risk and observed outcomes, suggesting the scores could be used directly for triage policies (e.g., "intervene if risk ≥ 0.6"). In real deployments, calibration often drifts with new populations and acquisition settings; the current result indicates the pipeline can be well-calibrated, but ongoing post-deployment calibration (e.g., temperature scaling or isotonic regression) should be expected [38].

4.5. Threshold Robustness and Operational Flexibility

Many clinical workflows need different operating points (screening vs. confirmation; resource-rich vs. constrained settings). The flat performance plateau in **Figure 5** and the stable confusion counts in **Figure 6** show the model's behaviour is insensitive to the exact cutoff across a wide range. This is operationally advantageous: stakeholders can pick thresholds based on clinical cost benefit without incurring sudden drops in precision or recall.

4.6. Validity, Leakage, and Reproducibility

We performed the train/test split at the subject level before epoch, preventing the most common form of information leakage (i.e., epochs from the same subject appearing in both sets). Seeds were fixed for reproducibility. Nonetheless, perfect performance on synthetic data should not be interpreted as a guarantee of real-world accuracy; it indicates the architecture–data fit is strong under the assumptions encoded in the simulator [39]-[42].

4.7. Limitations

Several challenges are expected when translating this approach to real clinical EEG

or EHR data. First, real signals exhibit stronger non-stationarity, artifacts, and site-specific noise that may erode both discrimination and calibration. Second, treatment response labels in psychiatry are often noisy, delayed, and influenced by adherence and comorbidity, unlike the clean labels used here. Third, domain shift across hardware, preprocessing pipelines, and patient populations can substantially degrade model performance and will require adaptation strategies such as self-supervised pretraining, domain generalization, or recalibration.

1) Synthetic data realism: Real psychiatric signals are noisier, non-stationary, and heterogeneous across devices, sites, and demographics. The simulator encodes only a subset of that complexity [43] [44].

2) Construct validity: The embedded frequency motifs are plausible but simplified; true responders may differ on multi-scale, context-dependent patterns (medication state, sleep, comorbidity) [45] [46].

3) External generalization: The model has not been exposed to domain shift (e.g., different hardware, sampling rates, or preprocessing pipelines) where performance and calibration typically degrade [47] [48].

4) Interpretability depth: While the VAE improves latent organization, mechanistic interpretability (e.g., which channels/times drive decisions under different confounders) warrants deeper analysis [49]-[51].

4.8. Recommendations and Future Work

- Prospective, real-world validation: Test the pipeline on clinical EEG and multimodal records (EHR, symptom scales, medication history, genomics where available) with site-wise and device-wise stratification.
- Ablations and baselines: Compare against (i) CNN-LSTM without VAE; (ii) simpler baselines (logistic regression on band powers, random forests on hand-crafted features); and (iii) transformer-based temporal models.
- Robustness & shift: Evaluate under additive noise, channel dropout, sampling-rate changes, and re-referencing schemes. Use domain adaptation or self-supervised pretraining to mitigate shift.
- Calibration in the wild: Track calibration over time; apply online recalibration and monitor with reliability diagrams and decision impact curves.
- Explainability: Augment saliency with channel-wise permutation tests, temporal occlusion, and attention maps; summarize at the group level to reveal consistent biomarkers.
- Fairness & subgroup performance: When moving to clinical data, systematically audit across age, sex, comorbidity, medication, and site to detect disparities.
- Clinical integration: Package the model with threshold presets for different use cases (screening vs. confirmatory), plus human-in-the-loop review and fail-safe behaviours when input quality is low.

The results—perfect separation (**Figure 1**), flawless ROC/PR (**Figure 2**, **Figure 3**), ideal calibration (**Figure 4**), and threshold robustness (**Figure 5**, **Figure 6**) es-

tablish a proof-of-concept: when discriminative temporal structure exists, a CNN–LSTM–VAE can capture it and express it as stable, interpretable probabilities. The next step is to stress-test and adapt this approach to the variability and constraints of real psychiatric care.

5. Conclusion

This work demonstrates the potential of a hybrid CNN–LSTM–VAE architecture for predicting treatment response in psychiatry. By combining convolutional layers for local feature extraction, recurrent layers for temporal integration, and a variational bottleneck for robust latent encoding, the model achieved perfect classification, calibration, and threshold stability on a synthetic multimodal dataset. These findings provide a proof-of-concept that deep learning frameworks can disentangle complex responder versus non-responder patterns and express predictions as clinically meaningful probability estimates. Beyond raw accuracy, the model's strengths lie in its reliability and operational flexibility: probability calibration suggests direct usability in risk-based decision-making, while insensitivity to threshold variation enhances adaptability across different clinical contexts. These qualities are particularly important in psychiatry, where prediction errors carry significant implications for patient care and resource allocation. Nevertheless, the study's reliance on synthetic data underscores its exploratory nature. Real-world psychiatric signals are far noisier and more heterogeneous, requiring rigorous validation on large-scale, multimodal clinical datasets that include EEG, neuroimaging, genomics, and electronic health records. Future research should also integrate attention-based interpretability, domain adaptation, and fairness auditing to ensure that such models are not only accurate but also transparent, generalizable, and equitable across diverse patient populations. While preliminary, this study illustrates how advanced deep learning architectures can serve as a foundation for precision psychiatry. With careful validation and continued methodological refinement, CNN-LSTM-VAE models hold promise as a step toward clinically actionable treatment response prediction systems capable of supporting personalized care.

Conflicts of Interest

The authors declare no conflicts of interest.

References

[1] Marx, W., Penninx, B.W.J.H., Solmi, M., Furukawa, T.A., Firth, J., Carvalho, A.F., *et al.* (2023) Major Depressive Disorder. *Nature Reviews Disease Primers*, **9**, Article No. 44. https://doi.org/10.1038/s41572-023-00454-1

[2] Thornicroft, G., Chatterji, S., Evans-Lacko, S., Gruber, M., Sampson, N., Aguilar-Gaxiola, S., *et al.* (2017) Undertreatment of People with Major Depressive Disorder in 21 Countries. *British Journal of Psychiatry*, **210**, 119-124. https://doi.org/10.1192/bjp.bp.116.188078

[3] Santomauro, D.F., Vos, T., Whiteford, H.A., Chisholm, D., Saxena, S. and Ferrari,

A.J. (2024) Service Coverage for Major Depressive Disorder: Estimated Rates of Minimally Adequate Treatment for 204 Countries and Territories in 2021. *The Lancet Psychiatry*, **11**, 1012-1021. https://doi.org/10.1016/s2215-0366(24)00317-1

[4] Voineskos, D., Daskalakis, Z.J. and Blumberger, D.M. (2020) Management of Treatment-Resistant Depression: Challenges and Strategies. *Neuropsychiatric Disease and Treatment*, **16**, 221-234. https://doi.org/10.2147/ndt.s198774

[5] Pandarakalam, J.P. (2018) Challenges of Treatment-Resistant Depression. *Psychiatria Danubina*, **30**, 273-284. https://doi.org/10.24869/psyd.2018.273

[6] Fonzo, G.A., Federchenco, V. and Lara, A. (2020) Predicting and Managing Treatment Non-Response in Posttraumatic Stress Disorder. *Current Treatment Options in Psychiatry*, **7**, 70-87. https://doi.org/10.1007/s40501-020-00203-1

[7] Frangogiannis, N.G. (2012) Biomarkers: Hopes and Challenges in the Path from Discovery to Clinical Practice. *Translational Research*, **159**, 197-204. https://doi.org/10.1016/j.trsl.2012.01.023

[8] Polley, M.C., Freidlin, B., Korn, E.L., Conley, B.A., Abrams, J.S. and McShane, L.M. (2013) Statistical and Practical Considerations for Clinical Evaluation of Predictive Biomarkers. *JNCI Journal of the National Cancer Institute*, **105**, 1677-1683. https://doi.org/10.1093/jnci/djt282

[9] Mienye, I.D., Swart, T.G. and Obaido, G. (2024) Recurrent Neural Networks: A Comprehensive Review of Architectures, Variants, and Applications. *Information*, **15**, Article 517. https://doi.org/10.3390/info15090517

[10] Malashin, I., Tynchenko, V., Gantimurov, A., Nelyub, V. and Borodulin, A. (2024) Applications of Long Short-Term Memory (LSTM) Networks in Polymeric Sciences: A Review. *Polymers*, **16**, Article 2607. https://doi.org/10.3390/polym16182607

[11] Berahmand, K., Daneshfar, F., Salehi, E.S., Li, Y. and Xu, Y. (2024) Autoencoders and Their Applications in Machine Learning: A Survey. *Artificial Intelligence Review*, **57**, Article No. 28. https://doi.org/10.1007/s10462-023-10662-6

[12] Wei, R. and Mahmood, A. (2021) Recent Advances in Variational Autoencoders with Representation Learning for Biomedical Informatics: A Survey. *IEEE Access*, **9**, 4939-4956. https://doi.org/10.1109/access.2020.3048309

[13] Cordier, N. (2015) Multi-Atlas Patch-Based Segmentation and Synthesis of Brain Tumor MR Images. Master's Thesis, Université Nice Sophia Antipolis.

[14] Meyer, T., Shultz, C., Dehak, N., Moro-Velázquez, L. and Irazoqui, P. (2025) Time Scale Network: An Efficient Shallow Neural Network for Time Series Data in Biomedical Applications. *IEEE Journal of Selected Topics in Signal Processing*, **19**, 129-139. https://doi.org/10.1109/jstsp.2024.3443659

[15] Aguilar-González, A. and Medina Santiago, A. (2025) CNN-based Road Event Detection Using Multiaxial Vibration and Acceleration Signals. *Applied Sciences*, **15**, Article 10203. https://doi.org/10.3390/app151810203

[16] Chen, Y., Rastogi, C. and Norris, W.R. (2021) A CNN Based Vision-Proprioception Fusion Method for Robust UGV Terrain Classification. *IEEE Robotics and Automation Letters*, **6**, 7965-7972. https://doi.org/10.1109/LRA.2021.3101866

[17] Zuo, Z., Shuai, B., Wang, G., Liu, X., Wang, X., Wang, B., *et al.* (2015) Convolutional Recurrent Neural Networks: Learning Spatial Dependencies for Image Representation. 2015 *IEEE Conference on Computer Vision and Pattern Recognition Workshops* (*CVPRW*), Boston, 7-12 June 2015, 18-26. https://doi.org/10.1109/cvprw.2015.7301268

[18] Moradi, R., Berangi, R. and Minaei, B. (2019) A Survey of Regularization Strategies

for Deep Models. *Artificial Intelligence Review*, **53**, 3947-3986. https://doi.org/10.1007/s10462-019-09784-7

[19] Nusrat, I. and Jang, S. (2018) A Comparison of Regularization Techniques in Deep Neural Networks. *Symmetry*, **10**, Article 648. https://doi.org/10.3390/sym10110648

[20] Noh, H., You, T., Mun, J. and Han, B. (2017) Regularizing Deep Neural Networks by Noise: Its Interpretation and Optimization. *Proceedings of the* 31*st International Conference on Neural Information Processing Systems*, Long Beach, 4-9 December 2017, 5115-5124.

[21] Subramanian, S. and Ganapathiraman, V. (2023) Zeroth Order GreedyLR: An Adaptive Learning Rate Scheduler for Deep Neural Network Training. 2023 *IEEE* 4*th International Conference on Pattern Recognition and Machine Learning* (*PRML*), Urumqi, 4-6 August 2023, 593-601. https://doi.org/10.1109/prml59573.2023.10348370

[22] Deng, X., Liu, Q., Deng, Y. and Mahadevan, S. (2016) An Improved Method to Construct Basic Probability Assignment Based on the Confusion Matrix for Classification Problem. *Information Sciences*, **340**, 250-261. https://doi.org/10.1016/j.ins.2016.01.033

[23] Nguyen, L.C., Naulaerts, S., Bruna, A., Ghislat, G. and Ballester, P.J. (2021) Predicting Cancer Drug Response in Vivo by Learning an Optimal Feature Selection of Tumour Molecular Profiles. *Biomedicines*, **9**, Article 1319. https://doi.org/10.3390/biomedicines9101319

[24] Sofaer, H.R., Hoeting, J.A. and Jarnevich, C.S. (2019) The Area under the Precision-recall Curve as a Performance Metric for Rare Binary Events. *Methods in Ecology and Evolution*, **10**, 565-577. https://doi.org/10.1111/2041-210x.13140

[25] Gerds, T.A., Andersen, P.K. and Kattan, M.W. (2014) Calibration Plots for Risk Prediction Models in the Presence of Competing Risks. *Statistics in Medicine*, **33**, 3191-3203. https://doi.org/10.1002/sim.6152

[26] Alba, A.C., Agoritsas, T., Walsh, M., Hanna, S., Iorio, A., Devereaux, P.J., *et al.* (2017) Discrimination and Calibration of Clinical Prediction Models. *JAMA*, **318**, 1377-1384. https://doi.org/10.1001/jama.2017.12126

[27] DeFilippis, A.P., Young, R., Carrubba, C.J., McEvoy, J.W., Budoff, M.J., Blumenthal, R.S., *et al.* (2015) An Analysis of Calibration and Discrimination among Multiple Cardiovascular Risk Scores in a Modern Multiethnic Cohort. *Annals of Internal Medicine*, **162**, 266-275. https://doi.org/10.7326/m14-1281

[28] Arrieta-Ibarra, I., Gujral, P., Tannen, J., Tygert, M. and Xu, C. (2022) Metrics of Calibration for Probabilistic Predictions. *Journal of Machine Learning Research*, **23**, 1-54.

[29] Lew, A.J. and Buehler, M.J. (2021) Encoding and Exploring Latent Design Space of Optimal Material Structures via a VAE-LSTM Model. *Forces in Mechanics*, **5**, Article ID: 100054. https://doi.org/10.1016/j.finmec.2021.100054

[30] Han, M., Soradi-Zeid, S., Anwlnkom, T. and Yang, Y. (2024) Firefly Algorithm-Based LSTM Model for Guzheng Tunes Switching with Big Data Analysis. *Heliyon*, **10**, e32092. https://doi.org/10.1016/j.heliyon.2024.e32092

[31] Bond-Taylor, S., Leach, A., Long, Y. and Willcocks, C.G. (2022) Deep Generative Modelling: A Comparative Review of Vaes, Gans, Normalizing Flows, Energy-Based and Autoregressive Models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **44**, 7327-7347. https://doi.org/10.1109/tpami.2021.3116668

[32] Vahdat, A. and Kautz, J. (2020) NVAE: A Deep Hierarchical Variational Autoen-

coder. *Advances in Neural Information Processing Systems*, **33**, 19667-19679.

[33] Zhao, J.B., Kim, Y., Zhang, K., Rush, A. and LeCun, Y. (2018) Adversarially Regularized Autoencoders. *International Conference on Machine Learning*, Stockholm, 10-15 July 2018, 5902-5911.

[34] Tomczak, J. and Welling, M. (2018) VAE with a VampPrior. *International Conference on Artificial Intelligence and Statistics*, Playa Blanca, 9-11 April 2018, 1214-1223.

[35] Justice, A.C., Covinsky, K.E. and Berlin, J.A. (1999) Assessing the Generalizability of Prognostic Information. *Annals of Internal Medicine*, **130**, 515-524. https://doi.org/10.7326/0003-4819-130-6-199903160-00016

[36] Fox, J., Glasspool, D., Patkar, V., Austin, M., Black, L., South, M., *et al.* (2010) Delivering Clinical Decision Support Services: There Is Nothing as Practical as a Good Theory. *Journal of Biomedical Informatics*, **43**, 831-843. https://doi.org/10.1016/j.jbi.2010.06.002

[37] Zikos, D. and DeLellis, N. (2018) CDSS-RM: A Clinical Decision Support System Reference Model. *BMC Medical Research Methodology*, **18**, Article No. 137. https://doi.org/10.1186/s12874-018-0587-6

[38] Sarkar, P.R. (2025) Artificial Intelligence Based Models for Predicting Foodborne Pathogen Risk in Public Health Systems. *International Journal of Business and Economics Insights*, **5**, 205-237. https://doi.org/10.63125/7685ne21

[39] de Melo, C.M., Torralba, A., Guibas, L., DiCarlo, J., Chellappa, R. and Hodgins, J. (2022) Next-Generation Deep Learning Based on Simulators and Synthetic Data. *Trends in Cognitive Sciences*, **26**, 174-187. https://doi.org/10.1016/j.tics.2021.11.008

[40] Nowruzi, F.E., Kapoor, P., Kolhatkar, D., Al Hassanat, F., Laganiere, R. and Rebut, J. (2019) How Much Real Data Do We Actually Need: Analyzing Object Detection Performance Using Synthetic and Real Data. arXiv: 1907.07061.

[41] Jordon, J., Szpruch, L., Houssiau, F., Bottarelli, M., Cherubin, G., Maple, C., Cohen, S.N. and Weller, A. (2022) Synthetic Data—What, Why and How? arXiv: 2205.03257.

[42] van Breugel, B., Liu, T., Oglic, D. and van der Schaar, M. (2024) Synthetic Data in Biomedicine via Generative Artificial Intelligence. *Nature Reviews Bioengineering*, **2**, 991-1004. https://doi.org/10.1038/s44222-024-00245-7

[43] Mutlu, A. (2025) Artificial Intelligence in Electroencephalography: A Comprehensive Survey of Methods, Challenges, and Applications. *Acadlore Transactions on AI and Machine Learning*, **4**, 186-218. https://doi.org/10.56578/ataiml040304

[44] Jordan, J. (2018) Probabilistic Neural Computation and Neural Simulation Technology. Master's Thesis, RWTH Aachen University.

[45] Dobreva, J., Simjanoska Misheva, M., Mishev, K., Trajanov, D. and Mishkovski, I. (2025) A Unified Framework for Alzheimer's Disease Knowledge Graphs: Architectures, Principles, and Clinical Translation. *Brain Sciences*, **15**, Article 523. https://doi.org/10.3390/brainsci15050523

[46] Heidari, A., Jafari Navimipour, N., Unal, M. and Toumaj, S. (2022) Machine Learning Applications for COVID-19 Outbreak Management. *Neural Computing and Applications*, **34**, 15313-15348. https://doi.org/10.1007/s00521-022-07424-w

[47] Munir, M.A., Khan, M.H., Sarfraz, M. and Ali, M. (2022) Towards Improving Calibration in Object Detection under Domain Shift. *Advances in Neural Information Processing Systems*, **35**, 38706-38718.

[48] Zhou, K.Y., Liu, Z.W., Qiao, Y., Xiang, T. and Loy, C.C. (2022) Domain Generaliza-

tion: A Survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **45**, 4396-4415.

[49] Zhou, D. and Wei, X.X. (2020) Learning Identifiable and Interpretable Latent Models of High-Dimensional Neural Activity Using pi-VAE. *Advances in Neural Information Processing Systems*, **33**, 7234-7247.

[50] Pan, Z., Wang, Y., Cao, Y. and Gui, W. (2024) Vae-Based Interpretable Latent Variable Model for Process Monitoring. *IEEE Transactions on Neural Networks and Learning Systems*, **35**, 6075-6088. https://doi.org/10.1109/tnnls.2023.3282047

[51] Choi, Y., Li, R. and Quon, G. (2023) siVAE: Interpretable Deep Generative Models for Single-Cell Transcriptomes. *Genome Biology*, **24**, Article No. 29. https://doi.org/10.1186/s13059-023-02850-y