BlindMRI: Synthetic Dataset for Modeling Blind Patient Stress Responses During MRI Procedures

BlindMRI(this is the link for the dataset: https://github.com/waad64/BlindMRI) is a fully synthetic dataset designed to support research on stress modeling and response prediction in blind patients undergoing MRI procedures.
It was generated using an innovative pipeline combining LLMs and GANs.
We harness the power of LLMs not just as passive data crunchers but as intelligent co-creators in our generative pipeline. The LLM kickstarts the process by generating rich, context-aware data that guides our GAN in crafting precise, concept-driven outputs. But it doesn't stop there—the LLM doubles as a discerning critic, validating each GAN output to ensure fidelity and relevance. Every training episode becomes a feedback loop, punctuated by rigorous statistical and accuracy tests powered by prompt-driven evaluations. We keep a sharp eye on the generator-discriminator loss curves, steering the model towards convergence and stability. This synergy of LLMs and GANs is our glimpse into the future of adaptive, intelligent data generation.

We present our approach in three progressive stages centered on detecting stress in patient scenarios. The first stage focuses on learning the correlations between raw and engineered physiological features to detect stress from standard patient data. Building on this foundation, the second stage integrates MRI-contextual information, aiming to identify stress within MRI scanning scenarios by combining raw and engineered features. Finally, the third stage incorporates blindness-specific variables, enabling the model to detect stress in MRI contexts while accounting for the unique physiological and behavioral patterns present in blind patients. This staged approach ensures our model evolves from general stress detection to highly specialized, clinically nuanced applications.

1/ Baseline Stress Detection

The first stage lays the groundwork by focusing on the core physiological and demographic data needed to detect stress in patients. This stage integrates raw metadata with primary physiological signals . To deepen the insights, we engineer advanced features which capture subtle autonomic nervous system responses and cardiovascular dynamics, starting with :

Raw data: Our data collection process begins with gathering metadata (specifically age and gender) from a public dataset of patients undergoing X-ray scans (NIH Chest X-rays). The choice of this dataset was far from arbitrary; we deliberately avoided MRI-related datasets to uphold the highest standards of data privacy. Instead, we selected a similar yet less sensitive domain (X-rays) to maintain contextual relevance without compromising privacy.

Since age and gender are medically significant factors, including them was essential. However, collecting this metadata required a clever approach: before extraction, we shuffled the entire dataset thoroughly. This ensured that our samples spanned diverse segments, both in age and gender, rather than clustering around a narrow range or focusing on individual patients.

We validated this strategy by applying statistical tests to confirm an even distribution across different demographic segments. No patient was left behind, nor privacy breached: the result is a well-balanced, privacy-conscious dataset that the LLMs will later use to generate and validate synthetic data aligned with our clinical goals.

At this stage, we have a clean CSV file (888 samples) containing the metadata fully distributed across the spectrum.

In addition to the metadata CSV, we created separate CSV files for each physiological feature, each containing 888 samples extracted from the WESAD dataset. WESAD is a well-established public dataset specifically designed for stress detection tasks, making it highly aligned with our research goals. Importantly, it is distributed under the Creative Commons Attribution 4.0 International (CC BY 4.0) license, which provides a permissive and transparent framework for reuse and research.

Each physiological signal: heart rate (HR), electrodermal activity (EDA), body temperature, and inter-beat intervals (IBI) was extracted after shuffling the dataset at the sequence level, ensuring we sampled from diverse individuals rather than drawing contiguous sequences from a single subject. This approach avoids overfitting and promotes generalization by preventing patient-specific bias in the training data. By selecting WESAD, we not only ensure legal compliance and medical relevance but also benefited from its sequence-based structure to build a diverse and representative physiological dataset for baseline stress detection.

To complement the core physiological features from WESAD, we integrated additional health indicators relevant to stress and overall metabolic state. Glucose levels were sourced from diabetes prediction dataset which, while not originally designed for stress detection, offers high-quality glucose data from a clinically monitored population. Since stress has been shown to influence glucose metabolism via cortisol-mediated pathways, incorporating this feature allows the model to capture deeper physiological interactions.

Similarly, systolic (SBP) and diastolic blood pressure (DBP) readings were obtained from a specialized health metrics dataset , chosen for their reliability, public availability, and comprehensive cardiovascular profiling. Blood pressure is a well-established biomarker affected by both acute and chronic stress, making its inclusion critical to building a more robust and medically grounded stress detection model.

Although these features were sourced from different datasets, we carefully validated their statistical distributions using both numerical and graphical methods and applied consistent preprocessing to harmonize them with the rest of our pipeline. This multi-source strategy enhances the dataset's richness while maintaining a clear physiological and clinical rationale.

Once all individual feature files were prepared and validated, they were merged back into the primary metadata CSV, creating a comprehensive dataset that captures the full physiological and demographic profile for each sample. So now we have metadata + raw physiological signals data CSV.

Engineered data: Refers to features derived from raw measurements using specific formulas designed to capture deeper insights. In our case, we engineered two key physiological metrics: Heart Rate Variability (HRV), quantified by the Root Mean Square of Successive Differences (RMSSD), and Mean Arterial Pressure (MAP). These engineered features provide nuanced, clinically relevant signals: RMSSD reflects autonomic nervous system activity and stress response, while MAP offers a comprehensive view of blood pressure beyond just systolic and diastolic values. Incorporating these derived metrics strengthens our dataset's ability to model patient stress levels with greater precision.

We will apply the RMSSD and MAP formulas across all samples in the combined metadata and physiological signals CSV file. This comprehensive calculation ensures each patient record is enriched with these engineered features, enabling more insightful and clinically meaningful analysis downstream. By systematically embedding these derived metrics into our dataset, we set the stage for more accurate stress-level detection and robust model training.

At this stage, we consolidated all available data into a single CSV file containing 888 samples, combining metadata, raw physiological signals, and engineered features. While the dataset passed rigorous statistical distribution checks, we recognized that statistical balance alone doesn't guarantee semantic coherence or clinical plausibility across features. Subtle mismatches such as a low heart rate paired with high stress markers can compromise the dataset's integrity and downstream performance.

To address this, we leveraged OpenAI's GPT-4 model to perform a semantic validation step. Through carefully crafted prompts, we guided the LLM with medical domain rules, including the normal and abnormal ranges for each physiological feature in both baseline and stress conditions. The LLM was instructed to systematically review the entire dataset, identify inconsistencies, and reorganize the records to ensure that each sample was not only statistically sound, but also medically coherent and contextually meaningful.

The outcome is a refined version of the original CSV structurally unchanged, but semantically enhanced where the feature combinations better reflect real-world patient profiles, enabling more reliable training and generation in the next stages of the pipeline.

Following the semantic validation phase, the same GPT-4 model was leveraged not only for verification but also for data generation. Using the previously validated CSV file as context, the LLM was prompted to generate additional samples that maintain semantic integrity, clinical realism, and feature-level coherence. These synthetic samples were appended to the dataset and used to seed a cGAN for training.

During each training episode, the cGAN was tasked with learning the latent structure of the semantically validated data to augment the dataset with new, contextually accurate samples. We rigorously evaluated each episode using several criteria:

- Statistical distribution checks to ensure the generated data preserved the expected variability and normality across features.

- Generator-discriminator convergence, monitored via loss curves to verify stable adversarial training and avoid mode collapse.
- And most critically, we integrated a custom LLM-based stress classification model, a GPT-guided classifier trained to detect stress based on medically informed prompts and reference thresholds. This model was used to assess the generated samples at each iteration, determining whether the synthetic data accurately captured both stressed and non-stressed individual patterns.

Training continued in this iterative loop until we achieved a performance threshold where:

1. The cGAN consistently generated high-fidelity, statistically sound data,
2. The discriminator could no longer distinguish real from synthetic samples with high certainty.
3. The LLM-based stress model achieved a reliable accuracy, confirming that the generated data was sufficient and meaningful for the downstream classification task.

At the end of this stage, we obtained a final CSV file enriched with a significantly larger number of samples, each one statistically sound, semantically coherent, and medically meaningful. These samples were validated not just through traditional statistical tests and GAN convergence metrics, but also through our LLM-guided stress classifier, ensuring the data truly captured the variability and nuance of both stressed and non-stressed individuals.

The core objective at this stage was to enable the CGAN to learn generalized stress patterns from physiological signals and metadata independently of any specific context like MRI or blindness. This foundational learning phase ensures that the model grasps the underlying correlations between features that are indicative of stress responses. By mastering this baseline, we prepare the architecture for more complex, context-aware stress detection tasks in the following stages.

With a solid foundation in generalized stress pattern recognition, we now transition to the second stage:

2/ MRI-Enhanced Stress Profiling

In this second stage, we extend the foundational stress detection model by embedding contextual variables specific to MRI environments, a critical real-world scenario where patients often experience heightened stress levels. Here, we focus on two primary environmental stressors:

- Peak Decibel Level (PDB): the maximum acoustic intensity experienced during the scan,
- Noise Bursts per Minute (NBM): the frequency of sudden, loud acoustic events within the MRI suite.

To simulate this context, we generated separate CSV files for each feature, relying on LLM-guided generation. Prompts explicitly described an MRI scenario where patients could

exhibit varying levels of stress based on the acoustic environment. The generated values ranged across low, moderate, and high sound conditions, allowing the data to reflect a gradient of stress responses from calm to moderately or highly stressed individuals. The generation process emphasized variability, realism, and physiological plausibility.

Each of these features underwent statistical validation including Kernel Density Estimation (KDE), boxplots, and Q–Q plots to ensure their distribution aligned with what is medically expected in such environments.

The merging of MRI-context features into the Stage 1 CSV was performed iteratively and carefully:

- Each contextual feature (PDB, then NBM) was injected one at a time into the enriched dataset.
- After each injection, the LLM was used as a semantic validator, instructed to assess the coherence and correlation between the newly added feature and existing ones (like how high PDB levels relate to HR, EDA, or stress classifications).
- The LLM also reorganized and refined the new feature's values to align with the existing sample context and generated a limited set of new rows representing coherent MRI-based stress cases.

This updated dataset was then fed into a Recurrent Conditional GAN (RCGAN). The goal here was to train the GAN to model stress patterns in MRI contexts, producing new samples that incorporate environmental stressors and reflect nuanced physiological responses.

Each RCGAN training episode followed a rigorous multi-layered validation strategy:

- Quantitative performance metrics, including AUC, F1-score, Sensitivity (Recall), and Matthews Correlation Coefficient (MCC), were used to assess the utility of the generated data in a downstream classification task.
- Statistical distribution monitoring ensured each new batch maintained medically realistic ranges and avoided anomalies.
- GAN convergence behavior was tracked through Generator–Discriminator (G–D) loss trajectories to ensure adversarial balance and prevent mode collapse.
- Qualitative validation was conducted using the LLM-based stress detection model, now instructed with MRI-specific stress prompts to verify physiological coherence in the new context.

Once PDB was fully integrated and validated, the same structured process was applied to NBM:

- LLM-guided generation with MRI noise context,
- Statistical validation,
- Incremental merging and semantic checking,
- Sample augmentation through LLM,
- RCGAN training with comprehensive evaluation.

This cycle was repeated for each contextual feature until the combined dataset achieved performance thresholds across all metrics, confirming that the model had successfully learned to detect stress in MRI-specific environments using both physiological and contextual cues.

Now that we have a comprehensive CSV file consolidating metadata, physiological signals, engineered features, and MRI environmental variables capturing a rich and nuanced portrait of stress in typical and MRI-specific contexts, we are ready to take the next crucial stage.

3/ Blindness-Aware Stress Diagnosis

In this final stage, we fine-tune our stress detection framework to the unique challenges presented by blind patients undergoing MRI scans. Building on the rich, multi-dimensional dataset from Stage 2, we now inject blindness-specific features that capture critical aspects of the patient's sensory experience, psychological state, and medical history. This tailored approach acknowledges that blindness can significantly modulate stress responses, and adapting the model accordingly is essential for reliable, inclusive stress assessment.

The blindness-related features integrated into the dataset include:

- Blindness Duration: Number of months the patient has been blind, capturing potential long-term adaptations or sensitivities.
- First MRI Experience: Whether this is the patient's first MRI session (Yes=1, No=0), as novelty can heighten stress.

- Pre-procedure Briefing: Indicates if the patient received calming verbal explanations before the MRI (Yes=1, No=0), reflecting psychological preparation.
- Headphones Provided: Whether headphones were offered to reduce MRI noise exposure (Yes=1, No=0), a practical factor influencing stress levels.
- Cause of Blindness: Medical etiology of blindness (e.g., congenital, diabetic retinopathy, trauma), which may correlate with physiological responses.
- Mobility Independence: Level of physical autonomy, ranging from fully assisted (dependent=1) to independent (0), affecting overall patient comfort and anxiety.
- Anxiety Level: Baseline self-reported anxiety before the procedure (anxious=1, unanxious=0), providing direct insight into psychological state.

Following the established methodology from Stage 2, each blindness-specific feature is generated individually in separate CSV files, then thoroughly statistically validated to ensure medical plausibility and proper distribution. After validation, these features are incrementally merged one by one into the evolving dataset. At each merging step, the LLM is employed to check for semantic coherence and contextual consistency, reorganizing and refining the new data to maintain a meaningful and physiologically realistic structure. The LLM also performs limited data augmentation by generating additional rows that reflect plausible blind patient scenarios within the MRI stress context.

With this carefully enriched and semantically organized dataset, the cGAN resumes training, learning the intricate correlations between blindness-related variables and stress patterns in MRI environments. The training proceeds iteratively with:

- Rigorous statistical validation to ensure medical realism and distributional fidelity.
- Continuous monitoring of generator–discriminator convergence to maintain adversarial balance and prevent mode collapse.
- Quantitative evaluation using performance metrics such as AUC, F1-score, Sensitivity, and MCC to measure stress detection accuracy.
- Qualitative validation through our LLM-based stress detection model, now specifically tuned with blindness-aware prompts, verifying both physiological relevance and semantic consistency.

Now that we have a comprehensive CSV file consolidating metadata, physiological signals, engineered features, MRI environmental variables, and blindness-specific variables capturing a rich and nuanced portrait of stress in blind patients within MRI-specific contexts, the dataset generation process concludes for these features. These can, of course, be extended or modified as future needs arise.

However, generating data is only part of the journey; the critical next step is rigorous benchmarking. While we continuously monitor statistical distributions, model performance, accuracy, and GAN convergence metrics throughout training, a well-known challenge remains: LLMs can sometimes produce “hallucinated” or unrealistic data segments despite prompt engineering and validation efforts.

To safeguard the statistical integrity and clinical realism of our synthetic physiological signals, we implemented a comprehensive, robust benchmarking strategy tailored for our sizable dataset (12,500 samples). Classic normality tests like the Shapiro–Wilk were unsuitable here due to hypersensitivity with large sample sizes. Instead, we relied on a more informative suite of distributional diagnostics:

- The Anderson–Darling test assessed whether individual feature distributions adhered to expected theoretical profiles, ensuring each variable behaved as medically plausible.
- The Kolmogorov–Smirnov (K–S) test provided pairwise comparisons between real and synthetic feature distributions, quantifying alignment across raw and engineered data domains.
- Finally, the Wasserstein Distance (Earth Mover’s Distance) offered a geometric measure of how closely the synthetic samples approximated the true statistical structure of the original data, capturing subtle distributional nuances beyond classical tests.

Together, these complementary metrics ensured both univariate fidelity and holistic distributional coherence, quickly highlighting any deviations that might undermine the physiological plausibility essential for reliable stress modeling.

Importantly, these benchmarks were not a one-off check; they were executed iteratively at every GAN training stage: from initial cGAN stress signal generation, through MRI-context augmentation with RCGAN, to the final blindness-context fine-tuning. This continuous evaluation pipeline allowed us to monitor and maintain generation quality, ensuring each successive training phase enhanced, rather than degraded, data fidelity and clinical relevance.