To what extent can machine learning predict individual responses to temazepam using baseline polysomnographic sleep characteristics?

Briana Vasile

1. Abstract:

Sleep disorders affect millions of people worldwide, who then commonly get prescribed benzodiazepines such as temazepam, helping improve sleep quality. However, individual responses to these medications can vary substantially. This makes it difficult to identify which patients are most likely to benefit from treatment. Thus, this study aims to investigate the extent to which machine learning can be used to predict individual responses to temazepam using baseline polysomnographic (PSG) sleep characteristics.

Through the use of paired placebo and temazepam overnight sleep recordings from 22 participants, multiple sleep metrics were extracted. These sleep metrics were then compared between treatment conditions. Exploratory analyses evaluated changes in overall sleep architecture, variability between individuals, and demographic subgroups. AI model development followed an 80/10/10 train-validation-test split to maximise predictive performance while minimising overfitting. Ridge Regression and Random Forest Regression were tested using leave-one-out cross-validation (LOOCV) within the development cohort, with change in sleep efficiency defined as the primary treatment-response outcome.

Temazepam was associated with a significant reduction in N1 sleep ($p = 0.024$) and increase in N2 sleep ($p = 0.027$). Conversely, changes in overall sleep efficiency, total sleep time, and Wake After Sleep Onset (WASO) were not statistically significant. Ridge Regression achieved an LOOCV Mean Absolute Error (MAE) of 3.63 percentage points, compared with 3.79 for Random Forest Regression. However, Random Forest produced lower prediction errors in the validation set and was therefore chosen for independent testing, where it produced an MAE of 3.03 percentage points.

These findings suggest that baseline PSG characteristics do contain information relevant to individual temazepam response. Predictive performance, however, was relatively limited due to a smaller sample size and varied across evaluation stages. Thus, the findings should be considered preliminary and support the feasibility of individualised treatment-response modelling rather than be seen as a clinically reliable prediction.

2. Introduction:

2.1 - Sleep Disorders

Sleep is a fundamental biological process that maintains both physical and psychological health. During sleep, complex processes such as memory consolidation, cellular repair, and brain development take place, emphasizing that enough sleep is essential for normal physiological and cognitive function (De Gans et al., 2024). Sleep disruption could therefore have extensive consequences. Poor sleep has been often associated with anxiety, reduced alertness, and long-term metabolic and cardiovascular effects (De Gans et al., 2024).

Sleep disorders include a wide range of conditions that affect sleep quality, duration, or continuity. Among these, insomnia is one of the most prevalent. Insomnia is characterised by persistent difficulty initiating or maintaining sleep, even with proper opportunity to sleep (Naha et al., 2024; Kozole Smid et al., 2024). These nighttime difficulties commonly come hand in hand with daytime symptoms such as fatigue, irritability and reduced quality of life. Insomnia has also been associated with more severe conditions, such as depression, chronic pain, and obesity (Kozole Smid et al., 2024).

Treatment of insomnia may involve both non-pharmacological and pharmacological approaches. Cognitive behavioural therapy for insomnia (CBT-I) is generally recommended as a first-line treatment, while pharmacological interventions may be considered when additional or more immediate symptom management is required (Naha et al., 2024). Benzodiazepines represent one class of hypnotic medications used often in the treatment of insomnia. These drugs act on inhibitory signaling pathways within the central nervous system and can influence both sleep initiation and maintenance. However, benzodiazepine treatment is generally recommended for short periods due to risks associated with prolonged use including dependence and detrimental effects (Kozole Smid et al., 2024). One very common benzodiazepine used is temazepam.

2.2 - Temazepam

Temazepam is a benzodiazepine hypnotic that exerts its sedative effects primarily through modulation of gamma-aminobutyric acid type A (GABAA) receptors. GABA is the major inhibitory neurotransmitter in the central nervous system, and benzodiazepines act as positive allosteric modulators of the GABAA receptor complex, enhancing inhibitory signaling and reducing neuronal excitability (Naha et al., 2024).

Temazepam has been extensively investigated in sleep-laboratory studies. Earlier research reported improvements across several measures of sleep following temazepam administration, including total sleep time and sleep maintenance, although effects on sleep-onset latency were

less consistent in the laboratory tests (Jackson et al., 1982). Temazepam could consequently influence several dimensions of sleep rather than just increasing total sleep duration.

More importantly, the effects of a treatment observed at a population level may not reflect the response of every individual. Patients differ in their baseline characteristics, and these differences could contribute to heterogeneity in treatment effects (Jaki et al., 2025). Identifying characteristics associated with this variation may help support more personalized approaches to pharmacological treatment.

2.3 - Polysomnography and Sleep Measurement

Sleep is multidimensional, and changes in sleep cannot be fully distinguished by total sleep duration alone. Pharmacological interventions could influence both sleep continuity and sleep architecture, such as the time spent in each sleep stage. Polysomnography (PSG) provides an objective method for evaluating sleep and is considered the reference standard for sleep-stage assessment (De Gans et al., 2024).

PSG combines several physiological signals throughout the night, generally incorporating electroencephalography (EEG) to measure brain activity, electrooculography (EOG) to monitor eye movements, and electromyography (EMG) to assess muscle activity. These metrics are also done alongside additional cardiorespiratory measurements. Together, these recordings allow sleep to be divided into wakefulness, rapid eye movement (REM) sleep, and the non-REM stages N1, N2, and N3.

The different sleep stages represent physiologically distinct components of sleep. N1 represents the transition from wakefulness into light sleep, while N2 is a more established stage of non-REM sleep and typically accounts for a substantial proportion of total sleep time. N3, however, represents slow-wave or deep sleep and is characterized by high-amplitude, low-frequency EEG activity. REM sleep is distinguished by characteristic brain activity, rapid eye movements, and reduced skeletal muscle tone. Examining the distribution of these stages provides information about sleep architecture that cannot be obtained from sleep duration alone.

PSG recordings can also be used to derive quantitative measures of sleep continuity and quality. Total sleep time (TST) represents the overall duration of sleep, while sleep efficiency describes the proportion of the analysed recording period spent asleep. Sleep latency measures the time required to transition from wakefulness to sleep. Lastly wake after sleep onset (WASO) quantifies wakefulness taking place after sleep has begun. Together with sleep-stage composition, these measures provide a comprehensive profiling of an individual's sleep.

Although PSG requires more specialized equipment and analysis, its detailed physiological measurements make it particularly valuable for research investigating treatment response. Since individuals can differ considerably across these measures even before receiving

treatment, baseline PSG characteristics may contain information relevant to succeeding treatment response, highlighting the necessity of an integrated clinical assessment. Which may further lead to a more complete representation of an individual's sleep phenotype.

2.4 - Machine Learning in Sleep Medicine

The increasing availability of physiological and clinical data has led to growing interest in the application of machine learning (ML) to sleep medicine. Unlike more conventional approaches that often examine individual variables separately, ML methods can identify patterns across multiple variables and generate predictions based on relationships that may be difficult to detect through examination of individual variables alone. Applications of ML in sleep medicine have included automated sleep-stage classification, detection of sleep disorders, analysis of physiological signals, and prediction of clinical outcomes (Huang & Huang, 2026).

The potential of ML to extract clinically relevant information from physiological sleep data has also been demonstrated in original research. (Liu et al., 2024), for example, developed a machine-learning framework for the identification and severity classification of obstructive sleep apnea using physiological recordings. Their findings demonstrate the potential for computational models to identify predictive patterns within sleep-related physiological data and generate predictions for previously unseen individuals.

Beyond diagnosis and classification, ML may also provide a useful framework for investigating variability in treatment response. Conventional clinical analyses generally estimate the average effect of an intervention across a study population. It is quite well known however, that individual patients may respond differently to the same treatment. (Jaki et al., 2025) describe how ML approaches can incorporate the multiple baseline patient characteristics to investigate heterogeneity in treatment effects and potentially identify individuals who are more likely to benefit from a particular intervention.

Different ML algorithms can capture different forms of relationships within the data. Regularized linear approaches, such as Ridge Regression, can model relationships between multiple correlated predictors while reducing the risk of excessively large model coefficients. Conversely, ensemble methods such as Random Forest Regression can represent nonlinear relationships and interactions between variables. The more interesting part is that it can be done without requiring these relationships to be specified in advance. Comparing these approaches may therefore provide insight into whether treatment response can be predicted from baseline sleep characteristics.

Nonetheless, the development of predictive models using sleep data also presents methodological challenges. Small datasets increase the risk of overfitting. Meaning that a model could learn patterns specific to the development sample rather than relationships that generalize

to unseen individuals. Careful separation of training and evaluation data, together with appropriate cross-validation methods, is thus particularly important when investigating predictive performance in limited datasets.

2.5 - Research Gap & Aim

Previous research has demonstrated that temazepam can influence several aspects of sleep, while advances in machine learning have shown that physiological sleep data can be used for prediction and classification. However, much of the traditional literature on temazepam has focused on average treatment effects across groups, rather than explaining why individual response may differ. Equally, many applications of machine-learning in sleep medicine have only focused on tasks such as sleep-stage classification and sleep-disorder detection rather than predicting individual pharmacological response. The broader treatment-prediction literature suggests that baseline patient characteristics may help identify heterogeneity in treatment effects, but practical applications of these approaches remain relatively limited (Jaki et al., 2025).

This gap provides a premise for investigating whether objective sleep characteristics measured before treatment can provide information about an individual's later response to temazepam. Baseline PSG-derived measures provide a quantitative description of sleep continuity and architecture and could therefore serve as potential predictors of treatment response.

Thus, the primary aim of this study was to investigate whether baseline PSG-derived sleep characteristics could be used to predict individual response to temazepam using machine-learning methods. Treatment response was decided to be primarily defined according to the within-participant change in sleep efficiency between placebo and temazepam conditions. While Ridge Regression and Random Forest Regression were compared to assess the predictive value of baseline sleep characteristics using linear and nonlinear modelling approaches.

A secondary aim was to characterize the effects of temazepam on sleep continuity and architecture by comparing sleep efficiency, total sleep time, sleep latency, REM latency, WASO, and N1, N2, N3, and REM sleep between placebo and temazepam conditions. Moreover, exploratory analyses examined whether treatment-induced changes in sleep were associated with age or sex.

Ultimately, it was hypothesized that baseline PSG characteristics would contain information predictive of individual treatment response and that temazepam would produce measurable changes in sleep continuity and sleep-stage composition.

3. Methods:

3.1 - Participants and Dataset

Data was obtained from the Sleep Telemetry subset of the publicly available Sleep-EDF Expanded Database hosted on PhysioNet (Pollard et al., 2026). The Sleep Telemetry dataset is derived from a 1994 study investigating the effects of temazepam on sleep and contains 44 overnight polysomnographic (PSG) recordings from 22 participants. The participants were male and female adults who reported mild difficulty falling asleep but were otherwise healthy and were not taking any other medications. Each participant underwent approximately nine hours of PSG monitoring in a hospital setting on two separate nights, with one recording obtained following temazepam administration, and the other following placebo administration.

The PSG recordings contained electroencephalographic (EEG), electrooculographic (EOG), submental electromyographic (EMG), and event-marker signals. EEG, EOG, and EMG signals were sampled at 100 Hz. Corresponding hypnogram files provided manually scored sleep-stage annotations, including wake, rapid eye movement (REM) sleep, stages 1-4, and movement time. Sleep stages had been manually scored by trained technicians according to the Rechtschaffen and Kales criteria (Rechtschaffen & Kales, 1968).

For the present analysis, participants were separated at the subject level to prevent recordings from the same individual from appearing across model-development and evaluation datasets. Eighteen participants were included in the development/training group, while four participants were withheld from initial model development/training. Of these four, two were then used as a validation set for model comparison, and the other two were kept as an independent test set for final evaluation. The independent test participants remained excluded from model training and validation.

3.2 - PSG Processing and Sleep Feature Extraction

The PSG and corresponding hypnogram files were processed in Python. EDF files were read using the MNE-Python library, and recording metadata was extracted to verify that each PSG recording was correctly paired with its corresponding sleep-stage annotation file. Participants identifiers and treatment conditions were derived from the Sleep Telemetry file names. This allowed recordings from the placebo and temazepam nights to be matched at the participant level.

Sleep-stage annotations were converted into consecutive 30-second epochs. Original annotations corresponding to sleep stages 1 and 2 were mapped to N1 and N2, respectively, while stage 3 and 4 were combined into N3 to reflect the current classification of slow-wave

sleep. REM and wake annotations were retained as separate stages. Movement-time annotations were treated separately from physiological sleep stages, and unscored or unmapped periods were labelled as unknown. Unknown epochs were retained during quality-control procedures to prevent their incorrect classification as a sleep-stage.

Quality-control checks were performed before feature extraction. PSG and hypnogram files were assessed for successful loading and correct participant and treatment matching. At the same time, the duration and location of unknown epochs were reviewed for each recording. Only two placebo recordings contained unknown time and required additional inspection. ST7121 contained 0.5 minutes of unknown epochs, while ST7221 contained 32.5 minutes. These epochs were excluded from analysed recording time rather than imputed as a particular sleep stage.This approach avoided making assumptions about the participant's physiological state during the periods without valid scoring.

Valid scored recording time was calculated as the cumulative duration of all epochs with a recognized sleep or wake classification. Total sleep time (TST) was calculated as the cumulative duration of epochs only classified as N1, N2, N3, and REM. Sleep efficiency was calculated as the proportion of valid scored recording time spent asleep.

Sleep Efficiency (%) = (Total Sleep Time / Valid Scored Recording Time) x 100

Additional sleep measures included sleep latency, wake after sleep onset (WASO), REM latency, and the proportion of TST spent in each sleep stage. Sleep latency was calculated as the duration of valid scored epochs prior to the first sleep epoch. WASO was calculated as the cumulative duration of wake epochs occurring between the first and final sleep epochs. REM latency was calculated as the interval between sleep onset and the first REM epoch. Sleep-stage composition was expressed as the percentage of TST spent in N1, N2, N3, and REM sleep.

Following feature extraction, placebo and temazepam recordings were paired for each participant, allowing within-participant changes in sleep characteristics to be calculated and accordingly used for statistical and machine-learning analysis.

3.3 - Statistical Analysis

Statistical analyses were performed to evaluate changes in sleep characteristics between placebo and temazepam conditions and to explore whether treatment response was also associated with participant age or sex. Analyses were conducted using Python. Descriptive statistics, specifically means, standard deviations (SD), and coefficients of variation (CV), were calculated to characterise the distribution and variability of sleep measures. More exactly, the

CV was calculated as the standard deviation divided by the mean and expressed as a percentage. Statistical significance was defined as $p < 0.05$.

For each sleep outcome, within-participant treatment change was calculated as the temazepam value minus the corresponding placebo value. The analysed outcomes included sleep efficiency, TST, sleep latency, REM latency, WASO, and the percentages of total sleep time spent in N1, N2, N3 and REM sleep. Positive change values therefore represented an increase under temazepam, whereas negative values represented a decrease relative to placebo.

The normality of the within-participant differences for each outcome was assessed using the Shapiro-Wilk test. Placebo and temazepam conditions were compared using a paired-samples t-test when the paired differences did not significantly deviate from normality. For outcomes with non-normally distributed differences, the Wilcoxon signed-rank test was used. Descriptive statistics for each treatment condition and the magnitude of the intra-individual differences were also calculated.

Exploratory analyses were additionally conducted to investigate whether demographic characteristics were associated with treatment response. Associations between age and treatment-induced changes in each sleep outcome were assessed using Spearman's rank correlation coefficient. Differences in treatment response between male and female participants were evaluated using Welch's independent-sample t-tests, which do not assume equal variances between groups. These demographic analyses were considered strictly exploratory. This approach was necessary due to the small sample size and unequal sex distribution. To analyse the actual findings, Spearman's correlation was selected. This non-parametric method accounted for the limited sample size and minimised the influence of non-normal treatment-response distributions.

3.4 - Machine Learning Analysis

Machine-learning models were developed to investigate whether baseline sleep characteristics could predict individual response to temazepam. Ultimately, the within-participant change in sleep efficiency between the temazepam and placebo conditions was selected as the primary prediction target. Only variables available from the baseline placebo recording were used as predictors. This ensured that information from the treatment condition was not introduced into model training.

Baseline PSG-derived sleep characteristics used as predictor values included sleep efficiency, TST, sleep latency, REM latency, WASO, the proportions of total time spent in N1, N2, N3, and REM sleep, age and sex. Two regression algorithms were evaluated: Ridge Regression and Random Forest Regression. Ridge Regression was selected as a regularized linear model

capable of reducing the influence of correlated predictors. Conversely, Random Forest Regression was used to capture potentially nonlinear relationships and interactions between baseline sleep characteristics.

Participants were separated at the subject level into development, validation, and independent test sets. Eighteen participants formed the development cohort, while two participants were reserved for validation and two were retained as an independent test set. The validation and test participants were excluded from the initial model-development process to reduce the risk of data leakage and provide an assessment of model performance on unseen individuals.

Given the small development sample, leave-one-out cross-validation (LOOCV) was used within the 18-participant development cohort. In each iteration, one participant was held out while the model was trained on the remaining 17 participants. This procedure was repeated until each participant had served as the held-out observation once. Predictions from all 18 repetitions were combined to estimate cross-validated performance while allowing the maximum possible number of participants to contribute to model training.

Following cross-validation, the models were trained using the development cohort and evaluated on the separate two-participant validation set. Validation performance was used to compare the generalization of Ridge Regression and Random Forest Regression. The selected modelling approach was thereafter assessed using the independent two-participant test set, which had remained excluded from both model development and validation.

Model performance was evaluated primarily using mean absolute error (MAE) and root mean squared error (RMSE), both expressed in sleep-efficiency percentage points. MAE quantified the average absolute difference between predicted and observed treatment responses, whereas RMSE assigned greater weight to larger prediction errors. The coefficient of determination (R^2) was additionally calculated during model development where appropriate. Lower MAE and RMSE values indicated greater predictive accuracy. Given the small sizes of the validation and independent test sets, held-out performance estimates were interpreted cautiously.

4. Results:

4.1 - Participant Characteristics

The development cohort consisted of 18 participants with paired placebo and temazepam polysomnographic recordings. Participants ranged in age from 18 to 79 years, with a mean age of 40.67 years (SD = 18.39). Of the 18 participants, 13 were female and 5 were male. The remaining four participants were withheld from the development cohort, with two allocated to the validation set and two retained as an independent test set for the machine-learning analysis.

Characteristic	Development cohort
Participants, n	18
Age, mean ± SD (years)	40.67 ± 18.39
Age range (years)	18–79
Female, n (%)	13 (72.2%)
Male, n (%)	5 (27.8%)

Table 1. Participant characteristics of the development cohort (n=18).

4.2 - Effects of Temazepam on Sleep

Temazepam was associated with modest changes in overall sleep measures and sleep-stage composition. Mean sleep efficiency increased from 88.83 ± 9.11% under placebo to 90.30 ± 6.05% under temazepam. This change corresponded to a mean increase of 1.47 percentage points. However, this difference was not statistically significant ($p = 0.238$). Similarly, total sleep time (TST) increased from 426.94 ± 50.75 minutes to 440.31 ± 32.46 minutes, representing a mean increase of 13.36 minutes. Nevertheless, this difference was also not statistically significant ($p = 0.186$).

Wake after sleep onset (WASO) decreased from 33.47 ± 35.40 minutes under placebo to 22.89 ± 20.62 minutes under temazepam, corresponding to a mean reduction of 10.58 minutes. As the within-participant differences in WASO did not meet the normality assumption, the Wilcoxon signed-rank test was used. The observed reduction was close to, but still did not reach statistical significance ($p = 0.067$).

Changes were also observed in sleep-stage composition. The proportion of total sleep time spent in N1 decreased from 11.31 ± 6.56% under placebo to 8.31 ± 4.61% under temazepam, a mean reduction of 3.00 percentage points ($p = 0.024$). In contrast, N2 increased from 49.98 ± 8.33% to 53.78 ± 8.04%, corresponding to a mean increase of 3.80 percentage points ($p = 0.027$). These differences were statistically significant.

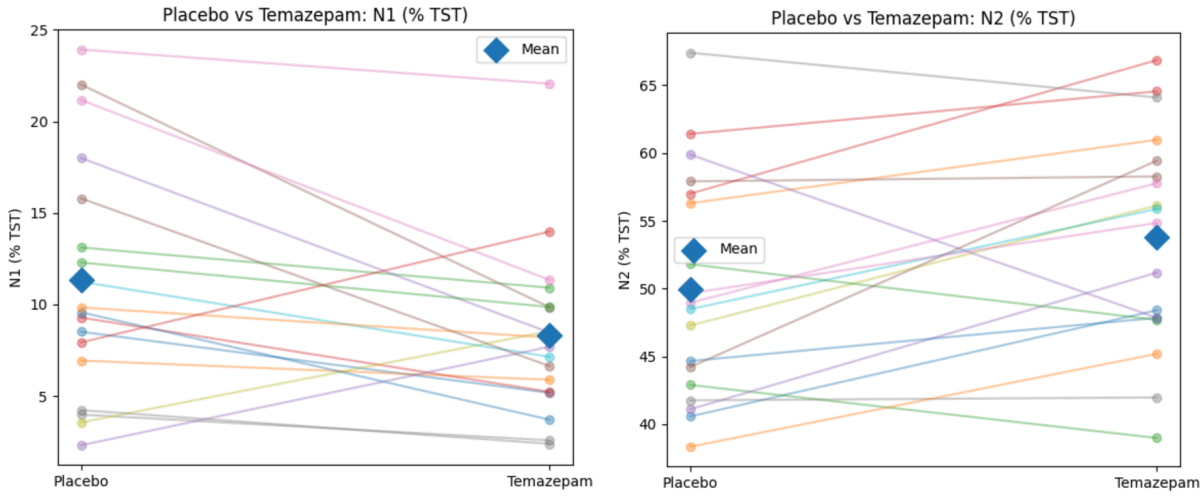

Figure 1. Individual changes in N1 and N2 sleep following temazepam administration.

Sleep parameter	Placebo mean	Temazepam mean	Mean change	SD change	Test	Statistic	*p*-value	Cohen's *d*	95% CI for change
Sleep efficiency (%)	88.83	90.30	+1.47	5.09	Paired *t*-test	1.22	.238	0.29	−1.06 to 4.00
Total sleep time (min)	426.94	440.31	+13.36	41.10	Paired *t*-test	1.38	.186	0.33	−7.08 to 33.80
Sleep latency (min)	11.08	12.00	+0.92	8.47	Wilcoxon	46.00	.683	0.11	−3.29 to 5.13
REM latency (min)	98.67	116.83	+18.17	39.90	Paired *t*-test	1.93	.070	0.46	−1.67 to 38.01
WASO (min)	33.47	22.89	−10.58	30.89	Wilcoxon	43.50	.067	−0.34	−25.94 to 4.78
N1 (% TST)	11.31	8.31	−3.00	5.14	Paired *t*-test	2.48	.024	0.58	−5.56 to −0.45
N2 (% TST)	49.98	53.78	+3.80	6.63	Paired *t*-test	2.43	.027	0.57	0.50 to 7.10
N3 (% TST)	16.27	16.79	+0.52	6.45	Paired *t*-test	0.34	.737	0.08	−2.69 to 3.73
REM (% TST)	22.43	21.12	−1.31	5.44	Paired *t*-test	−1.02	.320	−0.24	−4.02 to 1.39

Table 2. Changes in sleep outcomes between placebo and temazepam conditions.

4.3 - Demographic Associations with Treatment Response

Exploratory analyses were conducted to determine whether treatment-induced changes in sleep outcomes were associated with participant age or sex. Age was not significantly associated with the change in sleep efficiency following temazepam treatment (Spearman's ρ = 0.393, *p* = 0.107). Despite the lack of statistical significance, the positive correlation revealed a subtle trend. Specifically, older participants demonstrated a slight tendency toward greater improvements in sleep efficiency under the drug.

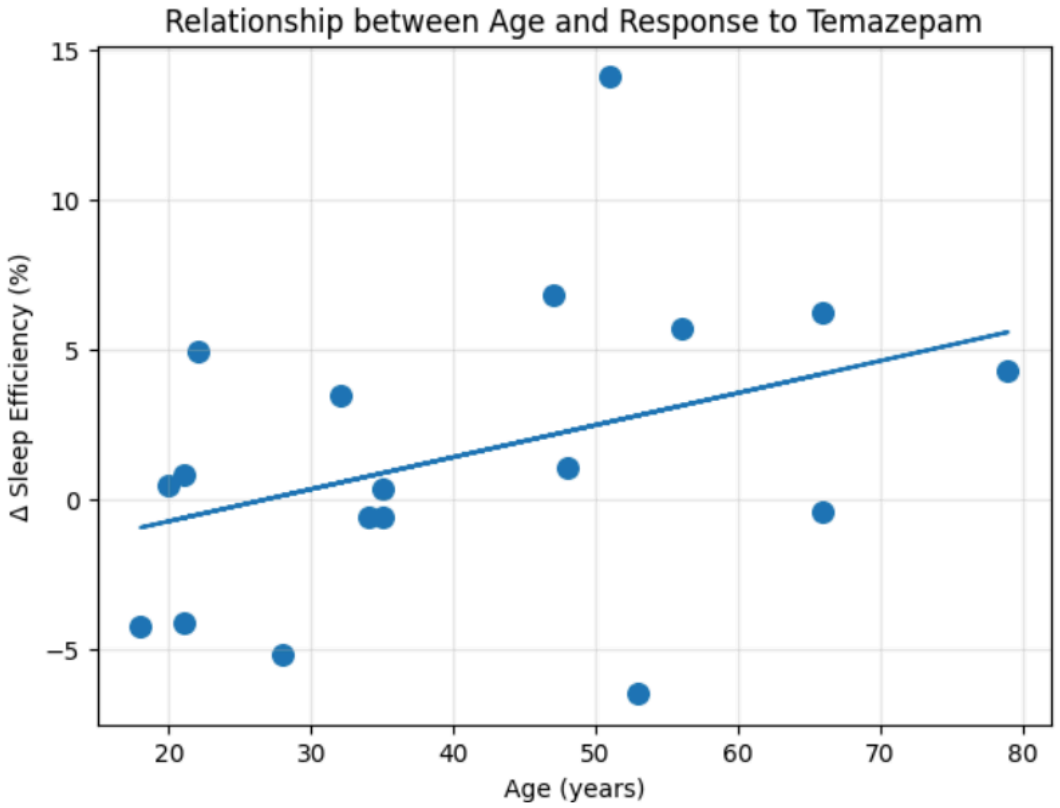

Figure 2. Association between participant age and change in sleep efficiency following temazepam administration.

Among the other sleep outcomes, age showed a moderate negative correlation with change in wake after sleep onset (WASO) (ρ = −0.518, *p* = 0.028). As treatment change was calculated as temazepam minus placebo, the negative correlation indicated that increasing age was associated with larger reductions in WASO under temazepam. This association should also be interpreted cautiously given the small sample size and exploratory nature of the demographic analysis.

Treatment response was then compared between male and female participants. Mean sleep-efficiency change was +2.76 percentage points in males and +0.97 percentage points in females. However, the difference between groups was not statistically significant (p = 0.497). No statistically significant sex differences were identified for any of the other assessed sleep outcomes. These findings were considered exploratory given the unequal group sizes of five males and thirteen females.

4.4 - Machine Learning Performance

Although mean sleep efficiency increased by 1.47 percentage points following temazepam administration, substantial variability was observed in individual responses. Some participants experienced reductions in sleep efficiency relative to placebo, whereas others showed considerable improvements.

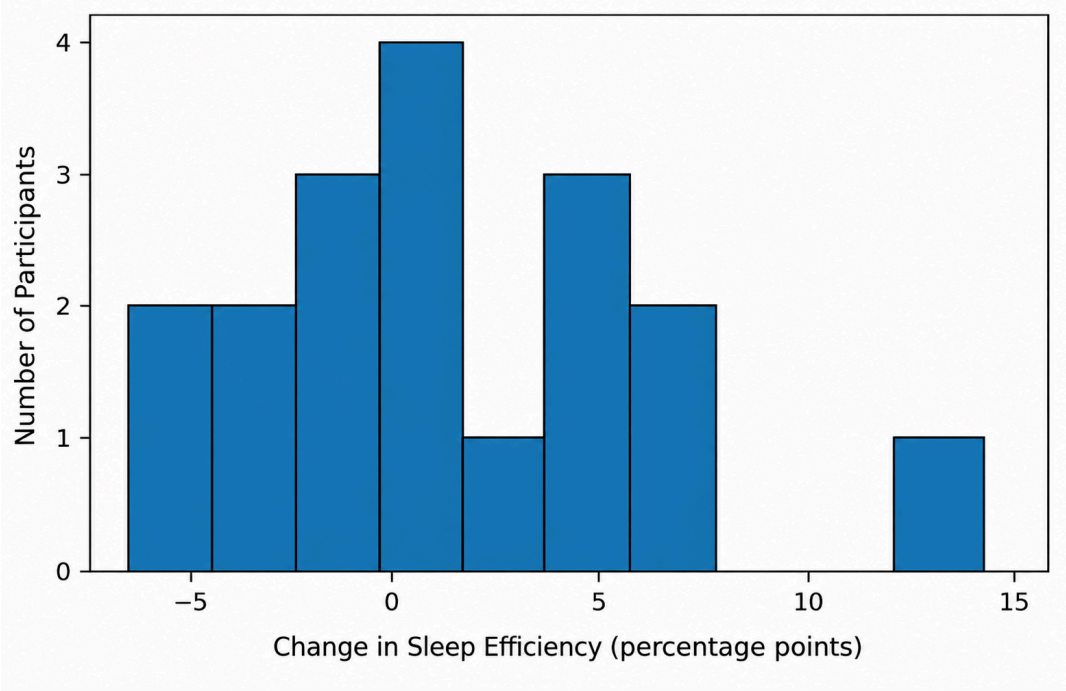

Figure 3. Distribution of individual changes in sleep efficiency following temazepam administration.

Ridge Regression and Random Forest Regression were evaluated for their ability to predict individual changes in sleep efficiency following temazepam treatment from baseline placebo sleep characteristics. Model performance was first assessed within the 18-participant development cohort using leave-one-out cross-validation (LOOCV). Wish was then followed by evaluation on the separately held-out validation and independent test sets.

During LOOCV within the development cohort, Ridge Regression achieved an MAE of 3.63 percentage points, an RMSE of 4.31 percentage points, and an R^2 of 0.239. Random Forest Regression achieved an MAE of 3.79 percentage points, an RMSE of 4.51 percentage points, and an R^2 of 0.168. Ridge Regression therefore showed slightly lower cross-validation prediction error than Random Forest Regression within the development cohort.

Model	MAE	RMSE	R^2
Ridge Regression	3.63	4.31	0.239
Random Forest Regression	3.79	4.51	0.168

Table 3. Leave-one-out cross-validation performance of Ridge Regression and Random Forest Regression in the development cohort (n=18). MAE and RMSE are expressed in sleep-efficiency percentage points.

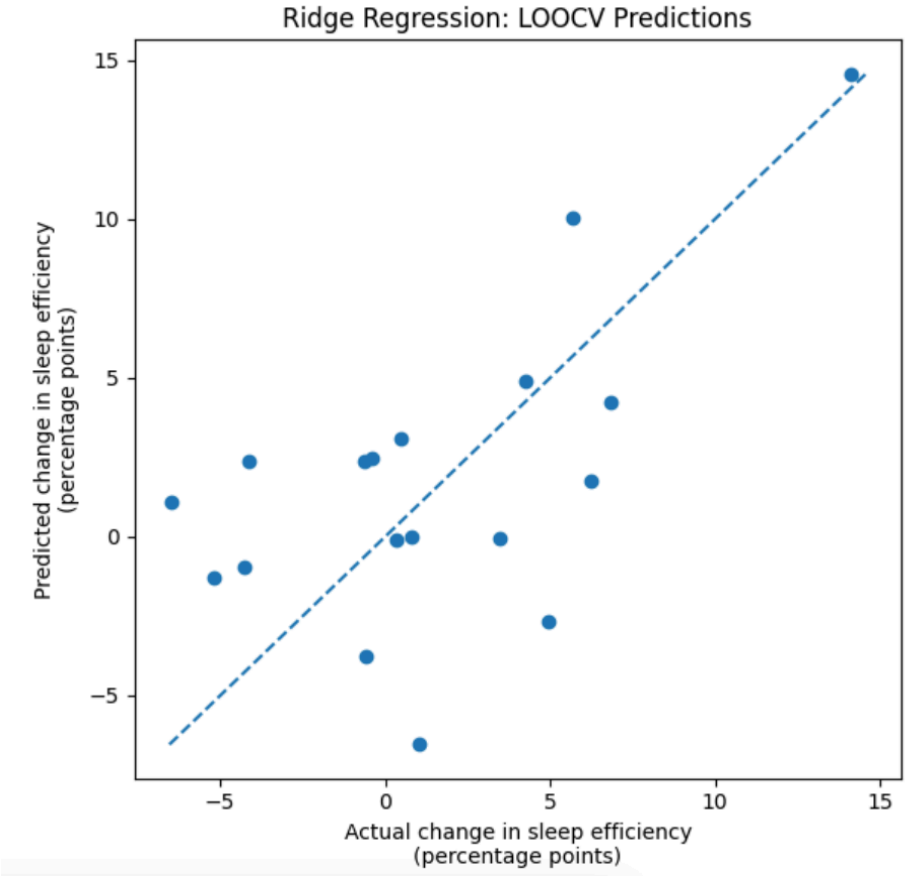

Figure 4. Observed versus LOOCV-predicted changes in sleep efficiency using Ridge regression in the development cohort.

Random Forest feature-importance analysis then indicated that baseline N1 percentage of total sleep time contributed most strongly to model predictions, followed by total sleep time, WASO, baseline sleep efficiency, and N2 percentage of total sleep time. Sex showed comparatively low feature importance. Given the small development cohort, these feature-importance estimates were considered exploratory.

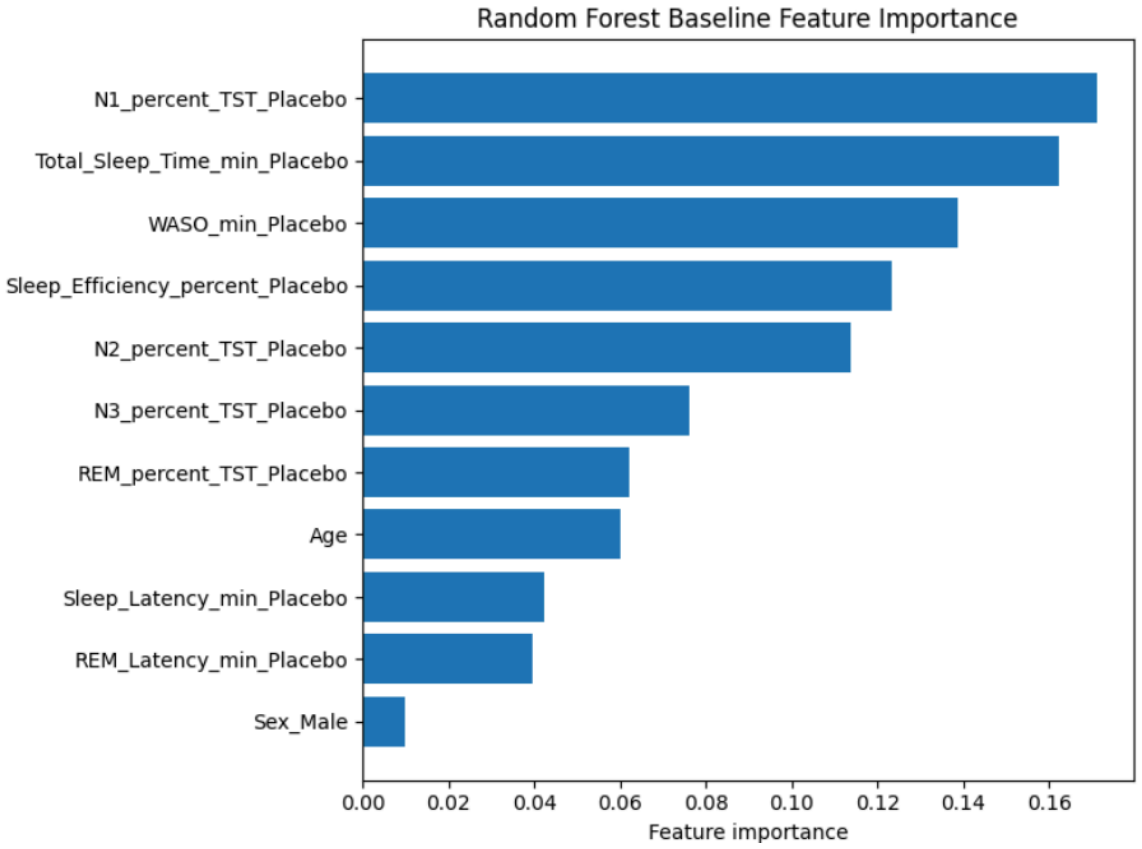

Figure 5. Random Forest baseline feature importance for prediction of temazepam-induced change in sleep efficiency from baseline characteristics. Feature importance represents the relative contribution of each predictor to the model and does not indicate a causal relationship.

When subsequently evaluated on the two-participant validation set, Random Forest Regression produced lower prediction errors than Ridge Regression. Random Forest achieved an MAE of 6.15 percentage points and an RMSE of 6.18 percentage points, compared with an MAE of 8.38 and RMSE of 8.42 percentage points for Ridge Regression. Based on this validation performance, Random Forest Regression was selected for evaluation on the independent test set.

Model	Validation MAE	Validation RMSE
Ridge Regression	8.38	8.42
Random Forest Regression	6.15	6.18

Table 4. Predictive performance of Ridge Regression and Random Forest Regression on the held-out validation set (n=2). MAE and RMSE are expressed in sleep-efficiency percentage points.

The selected Random Forest model was then evaluated on the independent two-participant test set, which had remained excluded from both model development and validation. The model achieved an MAE of 3.03 percentage points and an RMSE of 3.31 percentage points. Given that both the validation and test sets contained only two participants, these performance estimates were treated as preliminary measures of model generalizability.

Participant	Actual response	Predicted response	Prediction error	Absolute error
ST717	+9.40	+5.04	−4.36	4.36
ST720	−0.50	−2.20	−1.70	1.70

Table 5. Observed and Random Forest-predicted changes in sleep efficiency for the two participants in the independent test set.

5. Discussion:

5.1 - Principal Findings

This study investigated the extent to which machine learning could predict individual responses to temazepam using baseline polysomnographic characteristics. The findings suggest that baseline sleep characteristics contained some predictive information, but the overall predictive ability of the models was modest and varied across stages of evaluation. Ridge Regression performed slightly better during cross-validation within the development cohort, whereas Random Forest Regression produced lower errors in the separate validation set and was therefore chosen for independent testing. These differences indicate that no single modelling approach demonstrated a consistently superior performance across all available data.

The analysis also revealed considerable variability in individual sleep-efficiency responses to temazepam. Although the average change in sleep efficiency was relatively small, some participants experienced improvements, while others showed little change or even reductions. This heterogeneity was particularly relevant to the research question since it demonstrated that there was meaningful individual variation for the models to attempt to predict. Furthermore, exploratory demographic analyses also did not provide strong enough evidence that age or sex alone explained variability. This finding suggested that response may depend on a more complex combination of baseline characteristics.

The secondary analysis of sleep outcomes suggested that temazepam may influence aspects of sleep architecture and continuity even when changes in overall sleep efficiency are modest. The most apparent changes involved a reduction in N1 sleep and an increase in N2 sleep, suggesting that temazepam may influence sleep-stage distribution even in the absence of a significant change in overall sleep efficiency. Overall, the results provide preliminary evidence that machine learning could capture a limited proportion of individual variation in temazepam response from baseline PSG characteristics, but they do not demonstrate sufficiently reliable prediction for clinical application. The study should therefore be interpreted as an exploratory proof of concept rather than evidence of an established predictive model.

5.2 - Temazepam and Sleep Architecture

The observed reduction in N1 sleep alongside an increase in N2 suggests that temazepam may influence the organization of sleep even when changes in overall sleep efficiency are relatively small. Rather than simply increasing the amount of time spent asleep, the drug may alter how sleep is distributed across different stages. This is broadly consistent with previous research demonstrating that benzodiazepine hypnotics can influence sleep architecture in addition to measures of sleep duration and continuity (Naha et al., 2024). Earlier sleep-laboratory investigations of temazepam have similarly reported improvements across several sleep parameters, including measures related to sleep maintenance (Jackson et al., 1982).

The reduction in WASO observed in the present study, although not statistically significant, is also consistent with a possible improvement in sleep maintenance. Together with the shift from N1 towards N2 sleep, this suggests that the effects of temazepam in this group may have been more apparent in the continuity and organization of sleep than in overall sleep duration. Differences in the magnitude of treatment effects between the present study and previous research could reflect variation in participant characteristics, treatment conditions, or sample size.

The absence of clear changes in N3 and REM sleep further suggests that temazepam did not affect all components of sleep equally within this cohort. This further highlights the value of PSG-based assessment, as examining sleep efficiency or total sleep time alone may overlook more subtle changes in the organization of sleep.

5.3 - Predicting Individual Treatment Response

The considerable variability in sleep-efficiency response provides an important rationale for investigating individualized prediction. This reflects a broader challenge in clinical research identified by (Jaki et al., 2025), whereby average treatment effects may conceal substantial variability between individuals. Their work emphasizes the potential for machine-learning approaches to incorporate multiple baseline characteristics when investigating individual treatment effects. The variability observed in the present study provides a practical example of why such an approach may be relevant to pharmacological sleep treatment.

Machine learning has already demonstrated potential for extracting clinically relevant information from physiological sleep data. Previous applications have included automated sleep staging, identification of sleep disorders, and prediction of clinical outcomes (Huang & Huang, 2026, Liu et al., 2024). Building on these applications, the present study explored its use in individualized pharmacological response prediction. Specifically, investigating whether information contained in baseline PSG measurements can contribute to predicting how sleep efficiency changes following temazepam administration.

The variability observed across evaluation stages additionally highlights the challenges of establishing stable predictive relationships from a small sample. Individual participants may have had a significant influence on model performance, thus limiting confidence in the generalisability of the identified pattern. Larger independent groups would therefore be necessary to determine whether these relationships remain consistent across unseen individuals.

The relatively modest cross-validated R^2 values also suggest that the baseline characteristics included in the models explained only part of the observed variability in treatment response. This may indicate that important predictors were not represented in the available dataset or that the relationships between baseline characteristics and response could not be reliably learned from the limited number of participants. Treatment response may therefore depend on a broader combination of physiological, demographic, and clinical factors than those included in the present analysis.

Nevertheless, exploratory feature-importance analysis provided some insight into which baseline sleep characteristics may have contributed to the model's predictions. Baseline N1

percentage showed the greatest contribution, followed by total sleep time, WASO, and sleep efficiency. This finding suggested that several aspects of an individual's baseline sleep profile may work collectively and contain useful information relevant to treatment response. However, these rankings should be interpreted cautiously, as feature importance reflects contribution to model predictions rather than causal relationships. The rankings may also be unstable given the small development sample.

Evaluation on participants excluded from model development further highlighted the uncertainty surrounding predictive performance. Random Forest Regression performed better than Ridge Regression in the validation set and was consequently selected for independent testing. On the independent test set, the model captured the direction of response in both participants but underestimated the magnitude of their responses. Specifically, with a larger error for the participant who experienced the greater improvement in sleep efficiency. One possible explanation is that more pronounced responses are difficult to predict when few comparable examples are available during model development. However, with only two validation and two test participants, individual observations have an extensive influence on performance metrics, preventing firm conclusions about generalizability.

Taken together, these findings suggest that baseline PSG characteristics contain some crucial predictive signal relevant to individual temazepam response, but that this signal was not sufficient to produce consistently accurate predictions across the available evaluation stages. The results therefore support the feasibility of individualized treatment-response modelling rather than demonstrating that temazepam response can currently be predicted with clinically useful accuracy.

5.4 - Limitations and Future Directions

The primary limitation of this study was the small sample size. Only 18 participants were available for model development, while the validation and independent test sets each contained two participants. This limited the statistical power available to detect treatment and demographic effects and increased the uncertainty surrounding estimates of machine-learning performance. In particular, performance metrics calculated from only two participants are highly sensitive to individual prediction errors and cannot provide a stable estimate of how the model would perform in a broader population. The difference in relative performance between Ridge Regression and Random Forest Regression across evaluation stages should therefore be interpreted cautiously.

The demographic composition of the cohort also restricted interpretation. The unequal numbers of male and female participants reduced the ability to reliably investigate sex-related differences in treatment response. Although participants covered a broad age range, relatively few individuals represented different portions of this range. Hence, the demographic findings should

be interpreted cautiously and should not be taken as definitive evidence regarding the influence of age or sex on temazepam response.

Another limitation concerns the number and type of predictors available to the models. The analysis focused primarily on baseline PSG-derived characteristics together with age and sex. Individual response to a pharmacological intervention is likely to be influenced by additional biological, clinical, and behavioural factors that were not represented in the available dataset. The modest predictive performance observed in the present study may therefore partly reflect missing information rather than an absence of predictable variation in treatment response.

The number of statistical comparisons performed across sleep outcomes also requires consideration. Although the analyses focused on within-participant changes in predefined sleep parameters and unadjusted p-values were reported, several of the outcomes were examined within the same relatively small cohort. The statistically significant changes observed in N1 and N2 would also require replication in larger independent samples.

Future studies should evaluate this approach in notably larger and more diverse cohorts, with sufficiently sized development, validation, and independent test sets. Larger samples would allow more stable estimates of predictive performance and feature importance, as well as stronger comparisons between modelling approaches. They would also permit more extensive model optimization. Incorporating additional relevant baseline characteristics may further improve prediction and help determine whether PSG-derived measures provide predictive value beyond demographic and clinical information alone.

Future work could also investigate whether alternative definitions of treatment response provide stronger or more clinically meaningful prediction targets. Although change in sleep efficiency provided a continuous measure of individual response in the present study, response to insomnia treatment is usually multidimensional. Incorporating multiple outcomes related to sleep maintenance, sleep-stage composition, or combinations of PSG-derived measures may provide a more complete representation of treatment benefit.

Conclusion:

This study investigated the extent to which machine learning could predict individual responses to temazepam using baseline polysomnographic sleep characteristics. The findings demonstrate major interindividual variability in treatment response and suggest that baseline sleep characteristics contain some information relevant to predicting this variability. However, it is important to note that predictive performance was modest and slightly inconsistent across evaluation stages. This indicates that the models captured only part of the variation in individual response.

Despite these limitations, the results provide preliminary support for the feasibility of using machine learning to investigate individualized responses to sleep medication. Larger and more diverse datasets, additional predictors, and more extensive independent validation will be necessary to determine whether this approach can achieve clinically meaningful predictive accuracy. Ultimately, this study provides an initial step toward moving beyond average treatment effects and exploring whether an individual's baseline sleep profile can help predict how they will respond to treatment.

References:

[1] De Gans, C. J., Burger, P., Van Den Ende, E. S., Hermanides, J., Nanayakkara, P. W. B., Gemke, R. J. B. J., Rutters, F., & Stenvers, D. J. (2024). Sleep assessment using EEG-based wearables – A systematic review. *Sleep Medicine Reviews*, *76*, 101951. https://doi.org/10.1016/j.smrv.2024.101951

[2] Huang, A. A., & Huang, S. Y. (2026). Revolutionizing Sleep Medicine: The Impact of Machine Learning on Diagnosis, Treatment, and Personalized Care. *Health Science Reports*, *9*(6), e72323. https://doi.org/10.1002/hsr2.72323

[3] Jackson, E. A., Cardoni, A. A., McElnay, J. C., Jones, M. E., & Alexander, B. (1982). Temazepam (Restoril, Sandoz Pharmaceuticals). *Drug Intelligence & Clinical Pharmacy*, *16*(9), 650–656. https://doi.org/10.1177/106002808201600902

[4] Jaki, T., Chang, C., Kuhlemeier, A., The Pooled Resource Open-Access ALS Clinical Trials Consortium, & Van Horn, M. L. (2025). Predicting Individual Treatment Effects: Challenges and Opportunities for Machine Learning and Artificial Intelligence. *KI - Künstliche Intelligenz*, *39*(1), 27–32. https://doi.org/10.1007/s13218-023-00827-4

[5] Kozole Smid, A. K., Mlakar, A., & Štukovnik, V. (2024). Toxicity of benzodiazepines in the treatment of insomnia disorders in older adults: A systematic literature review. *Croatian Medical Journal*, *65*(2), 146–155. https://doi.org/10.3325/cmj.2024.65.146

[6] Liu, M.-H., Chien, S.-Y., Wu, Y.-L., Sun, T.-H., Huang, C.-S., Hsu, K.-C., & Hang, L.-W. (2024). EfficientNet-based machine learning architecture for sleep apnea identification in clinical single-lead ECG signal data sets. *BioMedical Engineering OnLine*, *23*(1), 57. https://doi.org/10.1186/s12938-024-01252-w

[7] Naha, S., Sivaraman, M., & Sahota, P. (2024). Insomnia: A Current Review. *Missouri Medicine*, *121*(1), 44–51.

[8] Pollard, T., Moody, B. E., Lehman, L. H., Gow, B. J., Fernandes, C., Xie, C., Johnson, A., Mark, R. G., & Heldt, T. (2026). PhysioNet as a global platform for biomedical research. *Nature Health*, *1*(8), 792–795. https://doi.org/10.1038/s44360-026-00096-z

[9] Rechtschaffen, A., & Kales, A. (Eds). (1968). *A Manual of Standardized Terminology, Techniques and Scoring System for Sleep Stages of Human Subjects* (NIH Publication No. 204). U.S. National Institute of Neurological Diseases and Blindness.