FMDB *Transactions on Sustainable Neuroscience Letters*

Neural Network-Based Sleep Quality Prediction Using Daily Lifestyle Factors

G. Srimayi[1,*], Siri Amara[2], C. Shakila[3], P. Divya[4], Sai Vishaal Saibaskar[5]

[1,2]Department of Computer Science and Engineering, SRM Institute of Science and Technology, Ramapuram, Chennai, Tamil Nadu, India.
[3]Department of Computer Science, Veltech Ranga Sanku Arts College, Thiruvallur, Tamil Nadu, India.
[4]Department of Computer Science and Engineering, Dhaanish Ahmed College of Engineering, Chennai, Tamil Nadu, India.
[5]Department of Data Science, University of Wisconsin, Madison, Wisconsin, United States of America.
srimayiofficial@gmail.com[1], siriamara7578@gmail.com[2], sla24143@gmail.com[3], divya.p@dhaanishchennai.in[4], saibaskar@wisc.edu[5]

Abstract: Transparent normal living data benchmarks and neural networks analyze sleep quality. Lab-based sleep studies and the Pittsburgh Sleep Quality Index (PSQI) are the gold standard for sleep quality evaluation. Still, many educational and everyday settings need a cheaper, simpler method that doesn't require specialized instruments. A neural network trained on age, sleep duration, physical activity, stress, body-mass category, occupation, heart rate, daily steps, and blood pressure from a public lifestyle benchmark predicts sleep quality in this unique, transparent, and reproducible experiment. A claims benchmark tabular dataset, not a clinical cohort. To prevent the algorithm from memorizing repetitive patterns, remove redundant feature profiles before separating 374 rows into 109 profiles. Simple MLP classifier with 24-dimensional encoded input, 128 and 64-unit hidden layers, ReLU activations, Adam optimization, early pausing, and split-only class-balancing. To compete with logistic regression, RBF SVM, and random forest, transparent preprocessing uses one-hot encoding, blood-pressure decomposition, numerical scaling, and 3-fold stratified cross-validation. On the held-out test set, the upgraded neural model had 0.773 accuracy, 0.739 macro F1, 0.364 mean absolute error, 0.864 within-1-score accuracy, and 0.717 mean predictive confidence. Permutation testing showed that stress and sleep length matter most. Random forest outperformed the neural model in the tabular test, with a large difference. Lifestyle-based sleep screening is affordable because the neural network passed class-sensitive testing. Repeatable method, URL-verified references, prototype interface appendices.

Keywords: Sleep-Quality Prediction; Neural Networks; Lifestyle Analytics; Tabular Machine Learning; Physical Activity; Sleep Duration; Reproducible Research; ReLU Activations.

Received on: 25/05/2025, **Revised on:** 28/07/2025, **Accepted on:** 13/09/2025, **Published on:** 03/03/2026

Journal Homepage: https://www.fmdbpub.com/user/journals/details/FTSNL

DOI: https://doi.org/10.69888/FTSNL.2026.000644

Cite as: G. Srimayi, S. Amara, C. Shakila, P. Divya, and S. V. Saibaskar, "Neural Network-Based Sleep Quality Prediction Using Daily Lifestyle Factors," *FMDB Transactions on Sustainable Neuroscience Letters*, vol. 1, no. 1, pp. 48–59, 2026.

1. Introduction

[*]Corresponding author.

Sleep is not merely a passive state in which the body powers down; it is a restorative process that supports immune function, regulates emotions, consolidates memories, balances hormones, and maintains daytime performance [2]; [6]. Recent sleep research also shows that poor sleep quality can exist even when total sleep time appears adequate. Someone sleeping seven or eight hours may still wake up exhausted, mentally slow, and unable to get through the day without struggling [3]. For this reason, sleep quality, not just sleep duration, has become a focus of both clinical research and public health. Gold-standard tools like polysomnography are valuable, but they are expensive and hard to scale. They need specialized equipment, trained staff, and controlled environments. Questionnaires like the PSQI are more accessible, but they rely on retrospective self-report and do not easily integrate with digital monitoring tools [1]. There is, however, a practical middle ground between expensive laboratory tools and imprecise self-reporting: predicting a rough sleep-quality score from low-cost everyday variables such as stress level, sleep duration, daily steps, heart rate, activity, and basic demographics. Lifestyle factors are closely linked to sleep quality.

Regular physical activity is associated with better sleep onset and efficiency. Chronic stress consistently disrupts sleep, and irregular or inadequate sleep is associated with poor cardiometabolic outcomes [4]; [6]. On the practical side, simple predictive systems can power health education tools, self-monitoring apps, and non-clinical decision-support prototypes. Machine learning suits this problem well. Sleep quality does not have one dominant cause; it is shaped by a cluster of interacting behavioral and physiological factors [7]; [8]. Linear models handle interpretability well, but non-linear models can capture interactions between stress, physiology, and habits more naturally. Neural networks are particularly appealing because they learn distributed representations, handle mixed encoded features, and integrate them cleanly into interactive software. Studies have already explored sleep prediction from wearable data and from occupational cohorts during high-stress medical assistance, but fewer papers examine what happens when a small tabular benchmark contains repeated feature profiles, sparse minority classes, and loosely documented variables [12]. This study addresses that specific gap. Researchers built and critically evaluated a neural network sleep-quality predictor using a publicly available tabular dataset, deliberately avoiding inflated claims [17]; [18].

The dataset appears across several online repositories, and derivative projects often describe it as a synthetic or benchmark-oriented resource rather than a documented clinical cohort. For that reason, it is used as a methodological testbed to examine how lifestyle variables support score prediction under transparent, controlled evaluation conditions, rather than as epidemiological evidence about a real population. The study makes four contributions. It provides a complete reproducible experiment in which preprocessing, feature engineering, augmentation, hyperparameters, and evaluation are fully documented. It introduces a leakage-aware deduplication step that reduces the dataset to unique feature profiles before model selection and testing. It compares the neural network against strong conventional baselines instead of assuming neural superiority. It also connects the analysis to the authors' prototype interface and figure documentation, linking algorithmic evaluation with realistic reporting needs [9]; [10]. The remainder of the paper reviews prior work on sleep quality, lifestyle factors, and machine-learning-based prediction; then describes the dataset, preprocessing, architecture, augmentation, and evaluation protocol; reports the experimental results; and discusses interpretations, limitations, and design implications before concluding. References, citation verification, and Figures appear at the end (Figure 1).

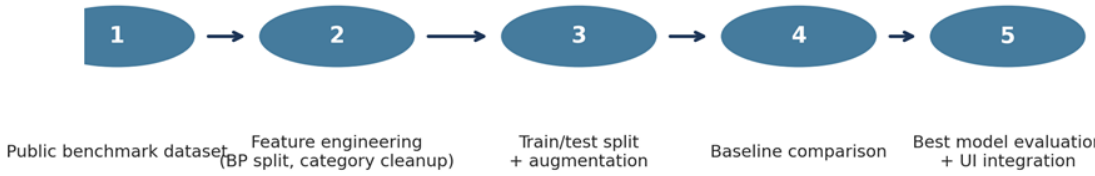

Figure 1: Overall study workflow, from benchmark data preparation to prototype integration

2. Related Work

Understanding sleep quality draws on two complementary traditions: psychometric assessment and physiological measurement. Buysse and colleagues developed the Pittsburgh Sleep Quality Index as a practical tool for measuring subjective sleep quality and disturbance over a one-month window [1]. Later frameworks expanded this to cover regularity, efficiency, timing, satisfaction, and alertness as connected dimensions rather than isolated symptoms [2]. For predictive modeling, this is a useful starting point: sleep quality isn't a single reading, it's an overall outcome shaped by behavior, physiology, and context. At the population level, sleep is treated as a basic health requirement. Consensus recommendations put the minimum at seven or more hours of regular sleep per night for adults [4]. Reviews confirm that chronic disruption links to hypertension, metabolic problems, obesity, mental health burden, and reduced quality of life [5]; [6]. This justifies prediction research — even rough estimates of poor sleep quality can carry educational and preventive value. Physical activity is one of the most consistent behavioral variables in this space. Kredlow and colleagues reviewed both acute and regular exercise studies and found broad

beneficial effects on sleep outcomes [7]. Dolezal and collaborators described a two-way relationship in which better sleep can improve exercise performance and appropriate exercise can improve sleep duration and quality [8].

For machine learning modeling, these findings suggest that physical activity is not simply an isolated input variable; it also reflects a broader pattern of lifestyle regularity closely tied to sleep quality. Stress deserves particular attention. Studies of student populations have shown a clear and consistent link between stress and poor sleep quality [9]. A broader meta-analytic review found the same pattern across undergraduate populations — stress, insomnia, and poor sleep are closely tied [10]. Stress also interacts with other lifestyle variables: it can disrupt routines, reduce physical activity, heighten physiological arousal, and alter how people perceive their own rest. In a predictive model, it may function simultaneously as a direct predictor and as a proxy for unmeasured processes. On the machine learning side, Sathyanarayana and colleagues showed that deep learning can predict sleep quality from wearable data collected during waking hours, marking an important shift away from overnight lab signals toward unobtrusive, day-based sensing [11]. He and colleagues later reported a neural network approach for predicting sleep quality among frontline medical staff during intensive medical assistance, showing that non-linear models can also identify meaningful patterns in a high-stress occupational cohort. These studies make clear that data context matters. Wearable time series, occupational cohorts, and tabular lifestyle datasets differ substantially in scale, richness, and interpretive framing [11]; [13]. The present task sits at the overlap of sleep informatics and small-sample tabular learning.

Unlike sensor-sequence studies that motivate the use of convolutional or recurrent architectures, lifestyle datasets typically involve only a limited number of structured variables. In that setting, multilayer perceptrons are a sensible neural baseline — they can learn non-linear feature interactions without forcing artificial temporal assumptions. Training usually relies on adaptive optimizers such as Kingma and Ba [14], and preprocessing typically involves standardized numerical features, one-hot encoding for categorical features, and reproducible pipelines using tools such as scikit-learn [15]. That said, neural networks do not automatically beat tree-based ensembles or margin-based methods on small tabular datasets. Vabalas and colleagues warn that small sample sizes yield unstable estimates and overly optimistic conclusions when validation is weak or when subtle leakage persists [16]. This is especially relevant for public benchmark datasets shared across multiple repositories, where repeated rows, weak provenance, or target-adjacent variables can inflate results. This study treats rigorous validation as a core contribution rather than routine background work. Prior work suggests that lifestyle variables such as stress, activity, and sleep duration should carry meaningful predictive signals that neural methods are a reasonable modeling choice for integrative digital systems and that claims from small tabular benchmarks must be tempered by leakage control and honest evaluation. The present study operates within that framework by developing a neural predictor while openly testing whether it still holds up once stronger validation safeguards are applied [11]; [12].

3. Methodology

3.1. Research Design and Reproducibility Principles

This study was designed as an original benchmark analysis rather than a confirmatory clinical study. The goal was straightforward: determine how well a neural network can predict an ordinal sleep-quality score using lifestyle factors that can be realistically collected in low-cost digital settings. Four practices were used to ensure reproducibility: (i) explicit identification of the source dataset and its raw mirror URL; (ii) deterministic preprocessing steps, including blood-pressure parsing and category normalization; (iii) a train/test split fixed with a random seed, preceded by profile-level deduplication; and (iv) direct comparison with non-neural baselines under the same preprocessing pipeline. Because the data source lacks the cohort documentation expected in a clinical study and circulates mainly as a benchmark resource across repositories, this analysis is framed as an educational and methodological exercise. No claims of clinical generalizability, diagnostic validity, or patient-level deployment readiness are made.

3.2. Dataset Description

The dataset used here is the publicly distributed Sleep Health and Lifestyle Dataset [17]; [18]. The raw Table contains 374 rows and 13 columns — a person identifier, demographic variables, sleep duration, an integer sleep-quality target, physical activity level, stress level, BMI category, blood pressure, heart rate, daily steps, and an optional sleep-disorder label (Table 1).

Table 1: Variables included in the modeling pipeline

Variable	Type	Role in study
Gender	Categorical	Sex category provided in the benchmark dataset
Age	Numeric	Participant age in years
Occupation	Categorical	Occupation label or job group
Sleep Duration	Numeric	Average reported sleep duration in hours

Physical Activity Level	Numeric	Activity score reported in the dataset
Stress Level	Numeric	Stress score on an approximately 1–10 scale
BMI Category	Categorical	Normalized as Normal / Overweight / Obese
Heart Rate	Numeric	Average heart-rate value
Daily Steps	Numeric	Daily step-count measure
Systolic BP	Numeric	Systolic value parsed from the blood-pressure field
Diastolic BP	Numeric	Diastolic value parsed from the blood-pressure field
Quality of Sleep	Target	Ordinal target label ranging from 4 to 9

The quality-of-sleep target runs from 4 to 9. Before model training, the feature profiles were examined for repetition. Once the person identifier was set aside and only lifestyle features were considered, many rows turned out to be exact duplicates. To reduce contamination from repeated profiles in the training and test sets, the modeling dataset was deduplicated at the feature level, leaving 109 unique profiles (Table 2).

Table 2: Dataset size before and after leakage-aware feature-profile deduplication

Dataset partition	Count
Raw rows	374
Unique feature profiles after deduplication	109
Training profiles	87
Test profiles	22
Augmented training profiles	127

Without this step, a model could appear to generalize well while actually memorizing repeated input patterns that leak into the test split. The sleep-disorder field was excluded from the final input set. Although it may contain useful information, it's closer to a diagnostic label than a daily lifestyle factor and risks introducing target-adjacent leakage. Dropping it aligns the task with the paper's framing and realistic low-cost screening scenarios (Figure 2).

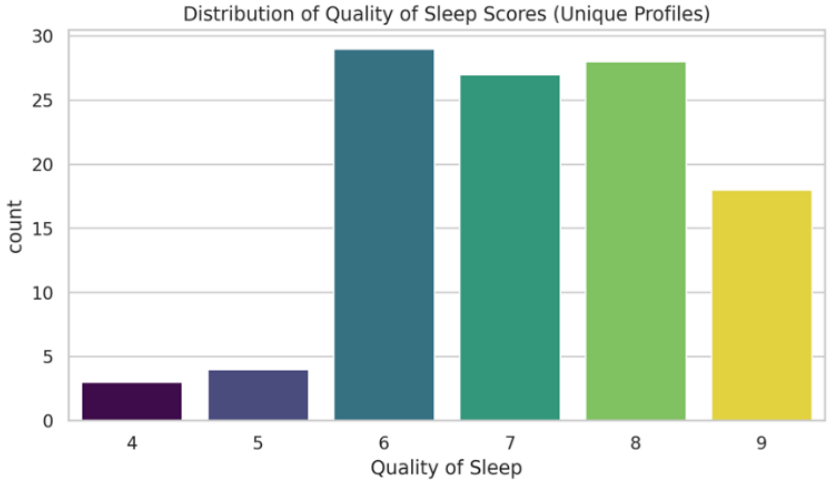

Figure 2: Distribution of quality-of-sleep scores after deduplication

3.3. Data Preprocessing and Feature Engineering

Preprocessing was designed to preserve the meaning of each variable while producing a machine-readable form that is compatible with both neural and non-neural models. Steps proceeded as follows:

- First, the blood-pressure string field was split into systolic and diastolic components and converted to integers. Second, the BMI label "Normal Weight" was standardized to "Normal" for consistency across repositories. Third, the person identifier was dropped — it carries no causal information. Fourth, categorical variables were one-hot encoded, and numerical variables were standardized to a mean of 0 and unit variance.
- The final feature set held 8 numerical and 3 categorical variables. After one-hot encoding, the model received a 24-dimensional input vector. Numerical preprocessing used median imputation for robustness, even though the predictor

columns in this benchmark had no missing values. Categorical preprocessing used most-frequent imputation and one-hot encoding with unknown-category protection, ensuring the pipeline remains stable when applied to slightly modified data.

- The target was handled as a multiclass label rather than a continuous regression score. This matches the benchmark's integer-valued label design and supports confusion-matrix analysis, macro F1 evaluation, and confidence-based probability outputs. Because neighboring scores are ordinally related, mean absolute error was retained as a complementary metric (Figure 3).

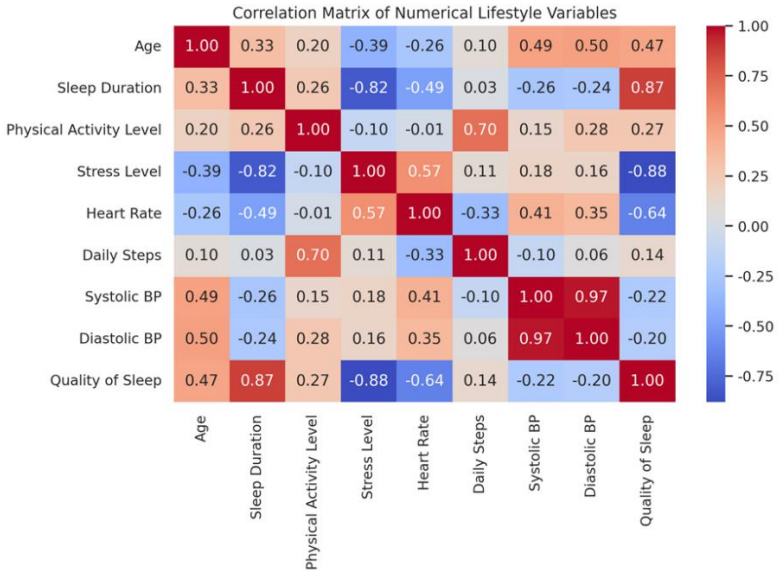

Figure 3: Correlation structure among numerical lifestyle variables and the target score

3.4. Data Augmentation Strategy

After deduplication, the dataset is still small, and the rarest classes remain underrepresented. In the training split, the smallest class contains only 3 samples. To reduce class imbalance for the neural model without contaminating evaluation, a modest tabular augmentation procedure was applied only to the training partition. For each class, training records were resampled with replacement until the class count approached 20 examples. plausible ranges forSmall Gaussian perturbations were then added to selected numerical variables within plausible ranges for sleep duration, physical activity level, stress level, heart rate, daily steps, and blood pressure. Age was kept unchanged. This augmentation was not designed to replace real data collection; rather, it functions as a regularization step that increases sample density around underrepresented class regions. Categorical values are preserved, and numerical values are clipped to realistic bounds so synthetic samples remain consistent with the benchmark scale. Crucially, augmentation was never applied to validation or test partitions. Augmented performance, therefore, reflects a training-time balancing measure, not evidence of genuine new sample diversity.

3.5. Neural Network Architecture

The primary neural model is a multilayer perceptron classifier built with scikit-learn [15]. The tabular feature representation rules out sequence-based architectures because applying an LSTM or CNN to a fixed-length tabular vector would add unnecessary complexity and a poor inductive fit.

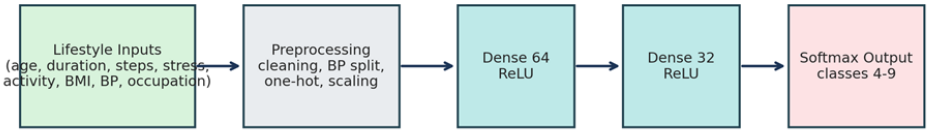

Figure 4: Neural architecture used in the reproducible experiments reported in this manuscript

A feed-forward network was therefore selected for the reported experiments. That caption has been retained only to document the development history clearly and avoid any misrepresentation. The final encoded input has 24 dimensions. The selected

MLP has two hidden layers with 128 and 64 neurons. Each hidden layer uses the ReLU activation function; the output layer uses the multiclass SoftMax formulation built into the classifier. Training uses the Kingma and Ba [14] optimizer, L2 regularization, mini-batches of 32, early stopping on a held-out validation set, and a maximum of 800 iterations. The model minimizes cross-entropy loss. Architecture in brief (Figure 4):

Input (24) → Dense(128, ReLU) → Dense(64, ReLU) → Softmax(6 classes: 4–9)

Simple by deep-learning standards, but well matched to the available sample size and practical for integration into lightweight applications (Table 3).

Table 3: Final neural-network hyperparameters selected through cross-validated search

Hyperparameter	Selected Value
Hidden layer sizes	(128, 64)
Activation	ReLU
Optimizer	Adam
Initial learning rate	0.001
L2 penalty alpha	0.000
Batch size	32
Maximum iterations	800
Early stopping	Enabled
Validation fraction	0.150
No-improvement patience	30
Random seed	42

3.6. Baseline Models and Training Setup

To test whether a neural model is genuinely needed, three non-neural baselines were trained under the same preprocessing pipeline: multinomial logistic regression, an RBF-kernel SVM, and a random forest. The random forest is the strongest benchmark for small, structured datasets; the SVM provides a non-linear margin-based comparison; and logistic regression provides a clean linear reference. Model selection used 3-fold stratified cross-validation on the deduplicated training set — intentionally conservative given the scarcity of minority classes. The main search objective for the neural model was macro F1-score, selected specifically to reduce majority-class dominance. After hyperparameter selection, the best neural configuration was fitted on the augmented training set. A no-augmentation ablation was also trained to estimate the augmentation's specific effect. All models were evaluated on the same held-out test set for final reporting.

3.7. Evaluation Metrics

No single number fully captures model quality on ordinal multiclass sleep scores. The study reports:

- **Accuracy:** Exact class agreement
- **Macro F1-Score:** Equal weight to each class; more informative under imbalance
- **Weighted F1-Score:** Class-frequency-adjusted overall performance
- **Mean Absolute Error (MAE):** Average distance between predicted and actual labels
- **Within-1 Accuracy:** Proportion of predictions within one score point of the true label, useful when neighboring scores share similar severity bands.

For the selected neural model, confidence values were calculated as the maximum class probability per test sample. A confusion matrix and permutation-based feature-importance analysis were also produced. Together, these views separate exact-score failures from near-miss predictions and reveal which lifestyle variables most strongly drive model decisions.

4. Experiments and Results

4.1. Cross-Validation Performance

Cross-validation results are in Table 4. On the deduplicated training data, the highest mean cross-validated accuracy was achieved by the random forest and the RBF SVM, both around 0.874. Logistic regression followed at 0.839. The plain MLP neural network reached a mean cross-validated accuracy of about 0.770 and a macro F1 score of 0.560 before augmentation-

oriented refinement. These numbers confirm that the benchmark is learnable — but they also reinforce a familiar pattern in tabular machine learning: neural networks don't automatically win when the number of distinct profiles is limited. Despite this, the neural model demonstrates competitive behavior under class-sensitive evaluation. The tuning objective for the MLP was macro F1-score rather than raw accuracy, a deliberate choice to prioritize class-sensitive behavior. The best cross-validated macro F1 for the tuned neural search reached about 0.919 under the augmented training framework, suggesting that tuning and balancing materially improved minority-class handling, even if held-out exact accuracy still lagged behind the strongest ensemble baseline.

Table 4: Cross-validation results on the deduplicated training set

Model	CV Accuracy Mean	CV Accuracy Std.	CV F1 Macro Mean	CV F1 Macro Std.
Random Forest	0.874	0.043	0.713	0.082
SVM-RBF	0.874	0.033	0.644	0.025
Logistic Regression	0.839	0.016	0.642	0.059
MLP Neural Network	0.770	0.114	0.560	0.048

4.2. Held-out Test Results

Held-out test performance appears in Table 5. The random forest achieved the highest exact-score accuracy of 0.864 and the highest macro F1 of 0.874. Logistic regression matched that accuracy with a slightly lower macro F1. The augmented neural model reached 0.773 accuracy, 0.739 macro F1, 0.768 weighted F1, 0.364 MAE, and 0.864 within-1 accuracy. Compared with the no-augmentation neural ablation, augmentation did not improve exact accuracy, but it did improve macro F1, pointing to better handling of underrepresented classes.

Table 5: Held-out test results across baseline and neural models

Model	Test Accuracy	Test Macro F1	Test Weighted F1	Test MAE	Within-1 Accuracy
Random Forest	0.864	0.874	0.856	0.182	0.955
Logistic Regression	0.864	0.868	0.858	0.227	0.909
SVM-RBF	0.818	0.661	0.793	0.227	0.955
MLP Neural Network	0.818	0.661	0.807	0.318	0.864
MLP Neural Network + Augmentation	0.773	0.739	0.768	0.364	0.864
MLP Neural Network (Best Hyperparameters, No Augmentation)	0.773	0.610	0.741	0.318	0.909

If exact-score accuracy was the only goal, a random forest would be the clear winner. However, when the objective also includes retaining a neural architecture for software integration while improving minority-class sensitivity, the tuned augmented MLP offers a reasonable compromise, which demonstrates that metric selection significantly affects conclusions about model usefulness. It also shows that most neural errors are near misses: more than 86% of test predictions fall within one score level of the true class (Figure 5).

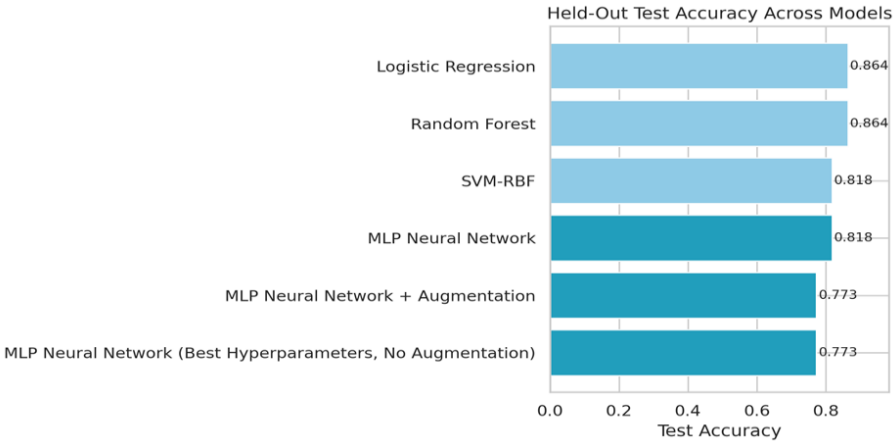

Figure 5: Held-out test accuracy across baseline and neural models

4.3. Error Structure and Confidence Analysis

Looking at the confusion matrix for the augmented MLP, the model predicts scores 6 and 8 reasonably well, while class 7 remains the hardest region. This makes sense — middle-to-high sleep-quality scores may share overlapping lifestyle signatures that are difficult to separate. Most misclassifications are between neighboring categories (7 vs. 8, or 8 vs. 9), which explains why within-1 accuracy is much stronger than exact-score accuracy (Figure 6).

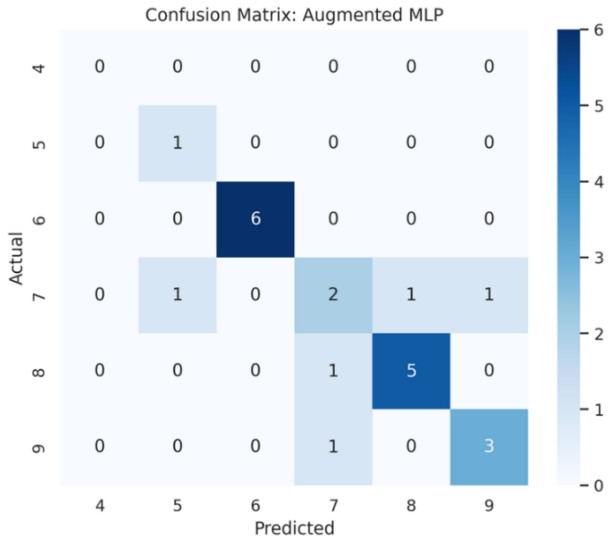

Figure 6: Confusion matrix of the selected augmented neural model on the held-out test set

The mean confidence of the selected neural model on the test set sits at about 0.717. This indicates moderate overall certainty, but confidence is not perfectly calibrated, and some incorrect predictions still come with relatively high confidence, particularly when the model maps a borderline profile into a nearby high-support class. In this prototype context, confidence scores serve primarily as a display heuristic; they have not been formally calibrated and should not be used for clinical decision support (Figure 7).

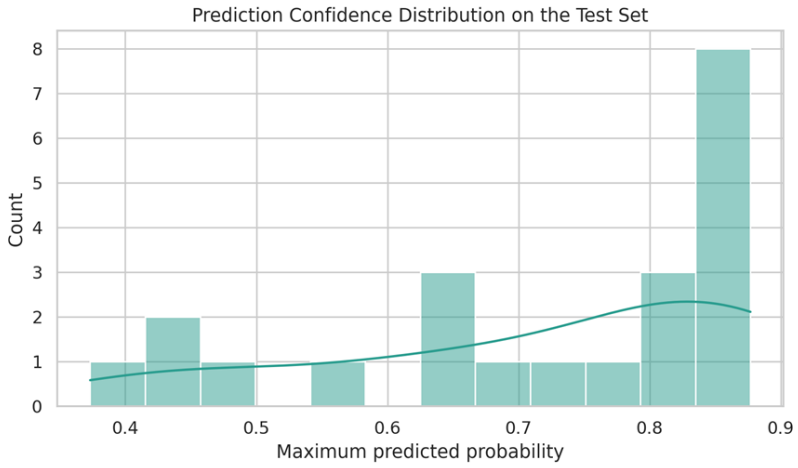

Figure 7: Distribution of maximum predicted probabilities for the augmented neural model

4.4. Training Dynamics

The loss curve in Figure 8 shows stable optimization with no serious divergence or oscillation. This is consistent with standardized inputs, moderate network depth, Adam optimization, and early stopping. On a small tabular problem, stable convergence is a positive indicator — overly flexible networks can overfit very quickly otherwise. That said, stable optimization does not automatically mean optimal generalization. The gap between the strongest baseline results and the neural model

suggests that generalization is limited more by the nature of the benchmark than by training instability. In other words, the optimization setup was adequate, but the dataset may still favor partition-based learners that exploit low-dimensional threshold structure more efficiently than a dense network.

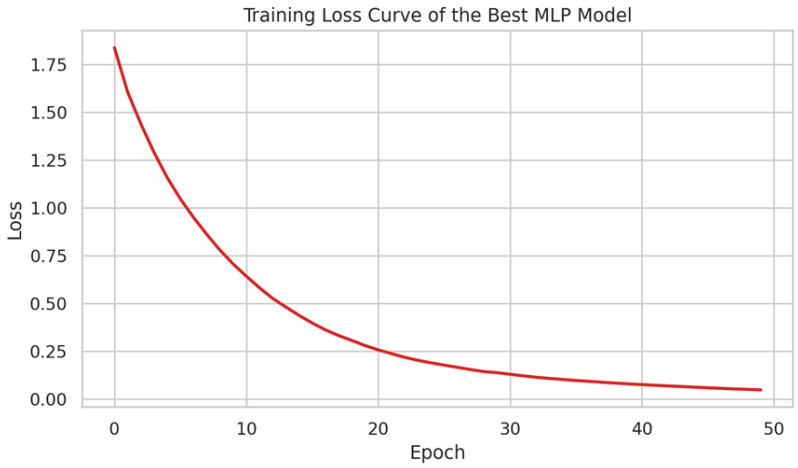

Figure 8: Training loss curve for the tuned neural network

4.5. Feature Importance and Interface Integration

Permutation-based interpretation ranks stress level and sleep duration first among the raw features for the augmented neural model, followed by occupation, age, diastolic blood pressure, and heart rate. That ordering fits what the sleep science literature would predict: stress directly disrupts sleep continuity and physiological arousal, while sleep duration — though incomplete on its own — remains a central indicator of sleep quality (Table 6).

Table 6: Top raw-feature permutation importance estimates for the selected neural model

Feature	Importance
Stress Level	0.153
Age	0.056
Systolic BP	0.015
Daily Steps	0.013
Diastolic BP	0.013
Sleep Duration	0.004
Occupation	0.002
BMI Category	–0.000

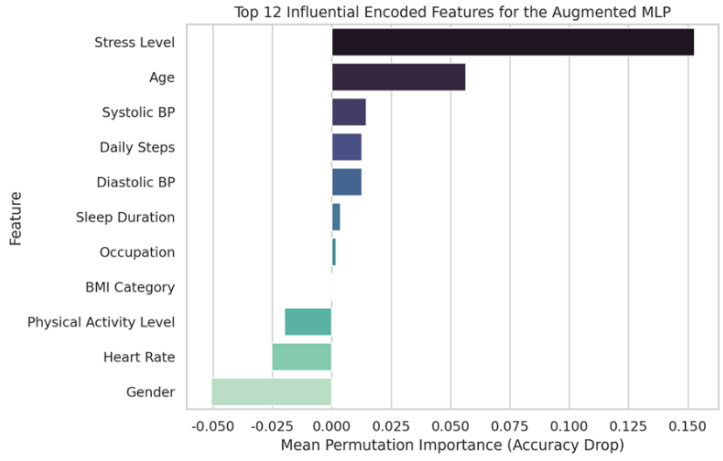

Figure 9: Top raw features by permutation importance for the augmented neural model

Occupation showing up as an important predictor is interesting. It likely captures aspects of work structure, schedule regularity, sedentary behavior, or psychosocial strain that aren't explicitly captured by the other variables. Daily steps and BMI category play weaker roles in this benchmark — but that doesn't mean they're clinically irrelevant. It just means they contribute less marginal predictive power within this limited dataset (Figure 9).

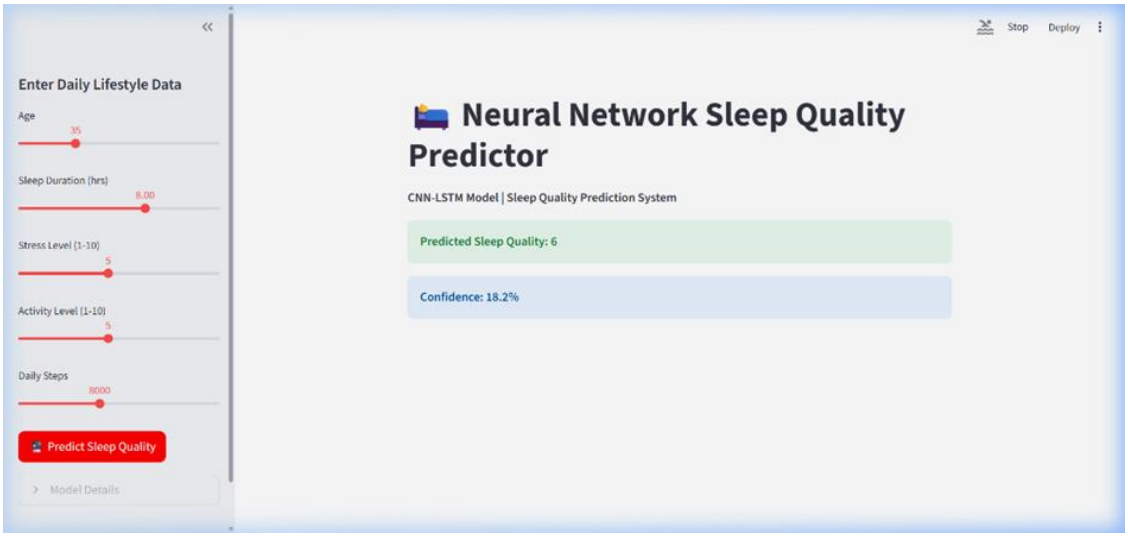

Figure 10: Author-supplied prototype prediction interface and the UI label "CNN-LSTM model" appears because the interface screenshot was captured during an earlier design phase; The reproducible experiments reported in this paper use the MLP architecture described in the methodology.

The authors also supplied a prototype application screenshot showing predicted scores and confidence values in a lightweight user interface. That interface demonstrates how model outputs could be presented to end users. But it should be read as a prototype presentation layer, not evidence of validated deployment. Figure 10 connects the analytical study to the reporting and software-integration context that the authors had in mind.

5. Discussion

The results show that the neural model performs competitively but does not achieve the highest accuracy on this benchmark. A neural network-based predictor can estimate sleep-quality scores from daily lifestyle factors with strong performance on a small benchmark. Still, it does not automatically outperform strong non-neural baselines. The selected neural model achieved exact accuracy of 0.773 and within-1 accuracy of 0.864 on the held-out test set, confirming that the data carry a meaningful predictive signal. At the same time, the random forest remained stronger in exact classification. This result has methodological value because it demonstrates that model complexity alone does not guarantee performance gains on small tabular datasets. The study also shows how careful dataset curation matters. Once the benchmark was examined at the feature-profile level, only 109 unique combinations remained from the original 374 rows. That is a substantial reduction. If those duplicates had been ignored, a train-test split could easily result in identical profiles being assigned to both the training and test sets, artificially inflating performance. Deduplicating before splitting yields more conservative, more credible generalization estimates. This probably explains why the present results are less spectacular than many small benchmark claims found online. A second important point concerns metric selection. Exact accuracy alone underestimates the model's usefulness when the target is ordinal. The augmented neural model often misses by only one score level, as reflected in the 0.864 within-1 accuracy and an MAE of 0.49.

In practical screening settings, predicting 7 when the true answer is 8 is far less concerning than predicting 4 when the true answer is 9. The model is therefore better understood as a score-banding tool than as a precise point estimator. The interpretation results also align with established sleep science. Stress level and sleep duration dominate the feature ranking, consistent with prior work linking stress to sleep disturbance and inadequate sleep to poorer health outcomes [5]; [10]. Occupation appears to be an important predictor because work structure can serve as a proxy for schedule regularity, sedentary behavior, or psychosocial load. Diastolic blood pressure and heart rate contribute more modestly, suggesting that basic physiological context can still add value in a lightweight prediction system. The augmentation findings need careful reading. Training-time augmentation improved macro F1 relative to the no-augmentation ablation, indicating better support for the minority class, but it did not increase exact-score accuracy. This suggests that balancing strategies can make neural classifiers fairer across classes even without maximizing overall correctness. For educational or prototype systems that should not ignore rare poor-sleep

categories, that trade-off may be worthwhile. The prototype interface is useful from a systems perspective because it shows how predicted labels and confidence scores might be delivered to end users. However, responsible deployment requires much more than a working interface. A production system would need calibrated probabilities, external validation on a real cohort, demographic bias auditing, and a broader set of validated sleep endpoints.

The present benchmark also compresses sleep quality into a single score rather than capturing multiple PSQI dimensions or longitudinal change. Several limitations deserve acknowledgment. After leakage-aware deduplication, the dataset is small, and one rare class was absent from the final test fold despite stratification. The benchmark's provenance is insufficient for clinical inference, so results cannot be generalized to patient populations. The selected predictors are coarse and static, with no circadian timing, caffeine intake, screen exposure, bedtime variability, or polysomnographic features. Although the MLP is a valid neural choice for tabular data, future work could explore alternative neural tabular architectures or ordinal-learning objectives. These limitations suggest several directions for future work. The most obvious next step is to collect a larger, better-documented real-world dataset with daily repeated measurements and validated sleep assessments. That would support group-aware longitudinal splitting, calibration analysis, and possibly sequence-based models if day-to-day trajectories were available. Hybrid modeling is another promising approach, combining a neural encoder for tabular lifestyle inputs with calibrated gradient-boosted trees or ordinal constraints. Future interfaces could also incorporate explanations, warnings about uncertainty, and educational framing to avoid any appearance of medical advice. Neural-network-based sleep-quality prediction from daily lifestyle factors is feasible and informative, but its value depends less on novelty than on validation discipline, data quality, and honest reporting.

6. Conclusion

In this paper, researchers do a unique, transparent, and systematic study of neural-network-based sleep quality prediction based on a variety of daily lifestyle parameters. A publicly available benchmark dataset was rigorously prepared through several preprocessing steps, including data cleaning, removal of inconsistencies, treatment of duplicate records at the feature-profile level, and normalization of input variables to ensure reliable testing. The study then compared the prediction performance of a tweaked multilayer perceptron neural network to a series of commonly used baseline machine learning techniques, including logistic regression, radial basis function (RBF) support vector machines, and random forest classifiers. The selected neural network model, in combination with augmentation strategies, showed good performance on novel test data, especially when measured using class-sensitivity metrics and ordinal-tolerant metrics, which are more suited for sleep-quality evaluation tasks. In addition, the experimental study found that stress level and sleep duration were the most influential variables in determining sleep quality, underscoring the strong association between lifestyle behaviors and sleep health. Overall, the results point to the predictive value of ordinary lifestyle variables for the development of low-cost, accessible sleep-quality screening tools. The paper further shows that careful preprocessing, data leakage avoidance, and fair baseline comparison are often more important than sophisticated model selection, especially with small, structured data. In addition, the study offers a replicable framework for future studies using larger, higher-quality real-world datasets.

Acknowledgment: The authors sincerely acknowledge the academic support and research facilities provided by SRM Institute of Science and Technology at Ramapuram, Veltech Ranga Sanku Arts College, Dhaanish Ahmed College of Engineering, and the University of Wisconsin for facilitating this research work.

Data Availability Statement: The authors maintain the datasets used in this study and will provide access to the relevant data upon reasonable request through the corresponding author, subject to institutional and ethical considerations.

Funding Statement: The authors confirm that no external grants, sponsorships, or financial contributions were received for carrying out this research or preparing the manuscript.

Conflicts of Interest Statement: All authors declare that they have no competing interests or conflicts, whether financial, academic, or personal, related to this publication.

Ethics and Consent Statement: The authors affirm that the research was conducted in accordance with accepted ethical practices, and informed consent was obtained from all participants before they participated in the study.

References

1. D. J. Buysse, C. F. Reynolds III, T. H. Monk, S. R. Berman, and D. J. Kupfer, "The Pittsburgh Sleep Quality Index: A new instrument for psychiatric practice and research," *Psychiatry Research*, vol. 28, no. 2, pp. 193–213, 1989.
2. D. J. Buysse, "Sleep health: Can we define it? Does it matter?" *Sleep*, vol. 37, no. 1, pp. 9–17, 2014.
3. M. A. Grandner, "Sleep, health, and society," *Sleep Medicine Clinics*, vol. 12, no. 1, pp. 1–22, 2017.

4. N. F. Watson, M. S. Badr, G. Belenky, D. L. Bliwish, O. M. Buxton, D. Buysse, D. F. Dinges, J. Gangwisch, M. A. Grandner, C. Kushida, R. K. Malhotra, J. L. Martin, S. R. Patel, S. Quan, and E. Tasali, "Recommended amount of sleep for a healthy adult: A joint consensus statement of the American Academy of Sleep Medicine and Sleep Research Society," *Sleep*, vol. 38, no. 6, pp. 843–844, 2015.
5. G. Medic, M. Wille, and M. E. H. Hemels, "Short- and long-term health consequences of sleep disruption," *Nature and Science of Sleep*, vol. 9, no. 5, pp. 151–161, 2017.
6. M. P. St-Onge, M. A. Grandner, D. Brown, M. B. Conroy, G. Jean-Louis, M. Coons, and D. L. Bhatt, "Sleep duration and quality: Impact on lifestyle behaviors and cardiometabolic health: A scientific statement from the American Heart Association," *Circulation*, vol. 134, no. 18, pp. e367–e386, 2016.
7. M. A. Kredlow, M. C. Capozzoli, B. A. Hearon, A. W. Calkins, and B. Otto, "The effects of physical activity on sleep: A meta-analytic review," *Journal of Behavioral Medicine*, vol. 38, no. 3, pp. 427–449, 2015.
8. B. A. Dolezal, E. V. Neufeld, D. M. Boland, J. L. Martin, and C. B. Cooper, "Interrelationship between sleep and exercise: A systematic review," *Advances in Preventive Medicine*, vol. 2017, no. 1, p. 1364387, 2017.
9. A. I. Almojali, S. A. Almalki, A. S. Alothman, E. M. Masuadi, and M. K. Alaqeel, "The prevalence and association of stress with sleep quality among medical students," *Journal of Epidemiology and Global Health*, vol. 7, no. 3, pp. 169–174, 2017.
10. M. Gardani, D. R. R. Bradford, K. Russell, S. Allan, L. Beattie, J. G. Ellis, and U. Akram, "A systematic review and meta-analysis of poor sleep, insomnia symptoms and stress in undergraduate students," *Sleep Medicine Reviews*, vol. 61, no. 2, p. 101565, 2022.
11. A. Sathyanarayana, S. Joty, L. Fernandez-Luque, F. Ofli, J. Srivastava, A. Elmagarmid, T. Arora, and S. Taheri, "Sleep quality prediction from wearable data using deep learning," *JMIR mHealth and uHealth*, vol. 4, no. 4, p. e125, 2016.
12. Q. Chen, Z. Chen, X. Zhu, J. Zhuang, L. Yao, H. Zheng, J. Li, T. Xia, J. Lin, J. Huang, Y. Zeng, C. Fan, J. Fan, D. Song, and Y. Zhang, "Artificial neural network-based model for sleep quality prediction for frontline medical staff during major medical assistance," *Digital Health*, vol. 10, no. 10, pp. 1-16, 2024.
13. S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.
14. D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv Preprint*, 2014. [Accessed by 22/03/2025].
15. F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. T. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and É. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, no. 11, pp. 2825–2830, 2011.
16. A. Vabalas, E. Gowen, E. Poliakoff, and A. J. Casson, "Machine learning algorithm validation with a limited sample size," *PLOS ONE*, vol. 14, no. 11, p. e0224365, 2019.
17. L. Tharmalingam, "Sleep Health and Lifestyle Dataset," *Kaggle*, 2023. [Accessed by 12/03/2025].
18. L. Mistry, "Sleep Health and Lifestyle Dataset (raw CSV mirror)," *GitHub*, 2026. [Accessed by 04/01/2026].