Hybrid VAE-SVM Framework for Improved Imbalanced Data Classification

Minsung Kim
Department of Electronic Enginerring
Hanbat National University
Daejeon, Republic of Korea
30242761@edu.hanbat.ac.kr

Minseok Lee
Department of Electronic Enginerring
Hanbat National University
Daejeon, Republic of Korea
wi5224196@gmail.com

Minyong Shin
Department of Electronic Enginerring
Hanbat National University
Daejeon, Republic of Korea
smy768@naver.com

Jimin Jeon
Department of Electronic Enginerring
Hanbat National University
Daejeon, Republic of Korea
wjswlals5399@gmail.com

Sooyeong Kwak *
Department of Electronic Enginerring
Hanbat National University
Daejeon, Republic of Korea
sykwak@hanbat.ac.kr

***Abstract*—Recent advances in wearable technology have enabled the prediction of sleep states using diverse physiological signals. However, inconsistencies in data collection cycles across sensor types and reliance on subjective assessments often result in long-term data scarcity and severe class imbalance, which degrade learning stability and classification performance. To address these challenges, we propose a hybrid classification framework that predicts both subjective and objective sleep-related indicators from lifelog data collected via wearable devices. The framework leverages a Variational Autoencoder (VAE) to capture the nonlinear distribution of minority classes and generate synthetic samples, which are combined with original data and processed by dual SVM classifiers. A meta-classifier is then applied to resolve conflicting predictions by exploiting probability estimates as higher-level inputs. Experimental results demonstrate that the proposed approach improves minority class representation and yields notable gains in predictive performance, particularly in F1-score and sensitivity. These results highlight the effectiveness of integrating generative modeling with meta-learning for reliable sleep state prediction in imbalanced data environments.**

Keywords—Wearable sensors, Sleep state prediction, Lifelog data, Variational Autoencoder (VAE), Class imbalance, Support Vector Machine (SVM), Meta-learning, Data augmentation

I. INTRODUCTION

Advances in wearable technology have significantly improved the convenience and accessibility of daily sleep monitoring. By capturing physiological signals such as heart rate, gait patterns, and activity data, wearable devices provide continuous and quantitative assessments of sleep quality, offering important opportunities for early detection of sleep disorders and personalized health management. However, accurate prediction of sleep-related indicators remains challenging due to (1) variability in data collection cycles across heterogeneous sensors, which hinders reliable long-term acquisition, and (2) severe class imbalance in subjectively assessed indicators, which reduces model stability and accuracy.

To investigate these issues, we utilize a lifelog dataset [1] that characterizes daily states through six indicators, including three subjective measures (sleep quality, pre-sleep fatigue, pre-sleep stress) and three objective measures (total sleep time, sleep efficiency, sleep latency). However, this dataset present inherent challenges: limited sample size, imbalance across classes, and noise in subjective evaluations. Previous approaches have attempted to address class imbalance through replication-based oversampling or statistical augmentation such as SMOTE [2]. While these methods partially mitigate imbalance, they introduce problems such as overfitting [3], inter-sample correlation, and distortion of temporal dependencies [4], [5]. More recent generative approaches based on GANs and VAEs [6], [7] have shown potential but still suffer from distributional distortions and incomplete preservation of original data structures. These limitations highlight the need for a more robust framework capable of

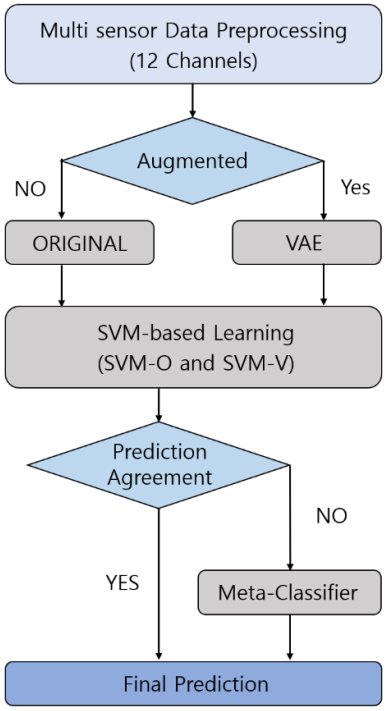

Fig. 1. Overview of the proposed VAE-SVM hybrid classification framework.

*Corresponding author: Sooyeong Kwak (email: sykwak@hanbat.ac.kr).

handling both data scarcity and class imbalance in wearable-based sleep prediction.

In this work, we propose a hybrid classification framework that integrates Support Vector Machines (SVMs), Variational Autoencoder (VAE)-based data augmentation, and a dual meta-classification strategy. Specifically, two independent SVMs are trained separately on original and VAE-augmented data, with a meta-classifier—comprising an Artificial Neural Network (ANN) and a Decision Tree (DT)—resolving prediction conflicts and refining probability estimates. This architecture not only alleviates issues caused by synthetic data distortions but also enhances classification stability for boundary-region samples.

The main contributions of this study are summarized as follows:

- We present a hybrid classification framework that explicitly addresses both class imbalance and small-sample constraints in wearable-based sleep prediction.
- We integrate VAE-driven data augmentation with dual SVM models to improve minority class representation while preserving generalization.
- We design a dual meta-classifier combining ANN and DT to refine prediction probabilities, thereby reducing errors introduced by synthetic data.
- We validate the proposed method on a multi-device lifelog dataset, demonstrating improved performance in terms of stability and predictive accuracy compared to conventional oversampling and generative approaches.

The remainder of this paper is organized as follows. Section II describes the methodology, including dataset characteristics and preprocessing procedures. Section III presents the experimental setup and results. Finally, Section IV concludes with a summary of the findings and directions for future research.

II. Methodology

A. Dataset

This study employs a lifelog dataset collected in 2024, encompassing approximately 40 days of continuous records per participant, with a total coverage of about 450 participant-days. Data acquisition was conducted using Android-based smartphones, smartwatches, sleep sensors, and self-reporting applications, resulting in 12 sensor modalities. These modalities captured a wide range of physiological and environmental measurements related to sleep states, with collection intervals varying from 1 to 10 minutes depending on sensor-specific characteristics.

Several challenges emerged during data collection. Device non-compliance factors—including non-wear periods, charging cycles, and system reboots—introduced substantial missing values and noise artifacts. To protect privacy, sensitive information such as GPS coordinates was transformed from absolute to relative reference frames, thereby preventing personal identification.

In addition, the six sleep-related indicators (Q1: sleep quality, Q2: pre-sleep fatigue, Q3: pre-sleep stress, S1: total sleep time, S2: sleep efficiency, S3: sleep latency) derived from self-reporting applications exhibited severe class imbalance across all target variables, as shown in Fig. 2. To address these issues, we adopted a two-stage strategy: (1) preprocessing techniques were applied to mitigate missing values and noise artifacts in raw sensor data, and (2) class imbalance in sleep indicators was alleviated during training through dedicated data reconstruction and model calibration methods (detailed in Section II-B).

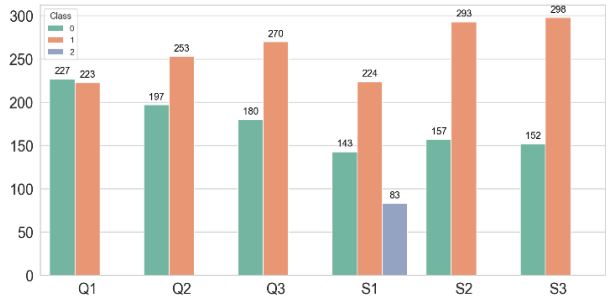

Fig. 2. Class distributions of sleep-related metrics.

B. Dataset Preprocessing

To mitigate missing values and measurement noise while ensuring reliable representation of temporal dynamics, we selected a subset of sensors from the 12 available modalities listed in Table I. The selection criteria prioritized sensors with one-minute sampling intervals, as these provided sufficient granularity for capturing temporal variations relevant to sleep-state monitoring. Sensors with low interpretability or weak relevance to sleep prediction were excluded or underwent specialized preprocessing.

Specifically, the *mActivity* sensor was recategorized into three discrete activity levels: low, medium, and high. The *mWifi* sensor was transformed into a binary home-presence indicator (0 or 1) based on BSSID measurements collected between 11 p.m. and midnight. After preprocessing, a total of seven sensor modalities were retained for subsequent analysis. The selected modalities are summarized in Table I, where preprocessed sensors are indicated by an asterisk (*).

TABLE I. Summary Table of Sensor Types and Variable Usage

Items	Dataset Description		
	Feature	***Used***	***Frequency(Hz)***
mACStatus	m_charging	✓	1/60
mActivity	m_activity	✓*	1/60
mAmbience	m_ambience		1/120
mBle	address		1/600
	device_class		
	rssi		
mGPS	altitude	✓	1/60
	latitude		
	longitude		

Items	Dataset Description		
	Feature	*Used*	*Frequency(Hz)*
	speed		
mLight	m_light		1/600
mScreenStatus	m_screen_use	✓	1/60
mUsageStats	app_name		1/600
	total_time		
mWifi	bssid	✓*	1/600
	rssi		
wHr	heart_rate	✓	1
wLight	w_light	✓	1/60
wPedo	step		1/60
	step_frequency		
	running_step		
	walking_step		
	distance		
	speed		
	burned_calories		

C. Oversampling

In binary classification tasks, imbalanced class distributions often degrade model performance by biasing decision boundaries toward majority classes. Traditional oversampling techniques such as SMOTE address this issue by generating synthetic minority samples through Euclidean distance–based interpolation. Although effective in balancing training sets, SMOTE relies on linear interpolation and is therefore limited in capturing complex data structures or nonlinear boundary regions. In high-dimensional spaces, this can lead to distributional distortions and blurred inter-class separations.

To overcome these limitations, we adopt a Variational Autoencoder (VAE), a probabilistic generative model that learns latent representations of input data and reconstructs them into the original space. Unlike interpolation-based methods, VAE captures nonlinear structures and intrinsic variability within minority classes, enabling the generation of diverse and representative synthetic samples. By augmenting sparse samples near decision boundaries, VAE improves classifier sensitivity to boundary-region patterns, thereby reducing misclassification and enhancing generalization.

D. Support Vector Machine

Support Vector Machines (SVMs) are well suited for imbalanced and small-scale datasets, as they construct separating hyperplanes that maximize class margins and rely on a limited number of support vectors. This margin-based learning mechanism provides stable generalization, even with limited training samples.

The dataset in this study consists of high-dimensional sensor-based time-series data with nonlinear distributions arising from complex inter-variable interactions. These characteristics require models that can handle nonlinear decision boundaries. To address this, we employed kernel-based SVMs, which project input data into higher-dimensional feature spaces. In particular, the Radial Basis Function (RBF) kernel was used to transform linearly inseparable samples into a space where effective linear separation becomes feasible, thereby enhancing classification robustness.

E. Hybrid Framework

As discussed in the previous sections, VAE-based oversampling enhances minority class representation, while SVM provides robust classification under small and nonlinear data conditions. However, each method has limitations when applied independently: SVM trained only on original data may underrepresent minority classes, and SVM trained with VAE-augmented data may be affected by distributional distortions.

To address these issues, we propose a hybrid classification framework, illustrated in Fig. 1. The architecture comprises two models: SVM-Original (SVM-O), trained solely on observed data to preserve authentic distributional characteristics, and SVM-VAE (SVM-V), trained with both original and VAE-generated data to enhance minority class sensitivity. Final predictions are produced using a conservative consensus strategy, where outcomes are accepted only when both classifiers agree.

This hybrid design leverages the stability of SVM-O and the augmentation strength of SVM-V, thereby improving boundary-region classification, reducing misclassification of minority samples, and enhancing overall prediction reliability.

F. Meta-Classifier

Although the hybrid architecture improves overall performance, SVM-O and SVM-V may yield conflicting predictions in boundary regions due to their different training distributions. To resolve these inconsistencies, we introduce a meta-classifier that refines final outputs. Since SVM-O is trained solely on original data and thus provides more reliable distributional representation, its class-wise prediction probabilities are employed as inputs to the meta-classifier.

Single-structure meta-classifiers, however, exhibit limitations: Artificial Neural Networks (ANNs) effectively capture nonlinear correlations but are prone to overfitting and limited interpretability, while Decision Trees (DTs) offer interpretable rule-based partitions but cannot adequately model complex high-dimensional relationships. To leverage the strengths of both, we propose a dual ANN–DT meta-classifier, illustrated in Fig. 3. In this design, the ANN first extracts nonlinear feature representations from SVM-O probabilities, and the DT subsequently performs interpretable rule-based classification on these features. This combination simultaneously ensures expressive modeling capacity and transparent decision-making, thereby enhancing prediction stability in high-uncertainty regions.

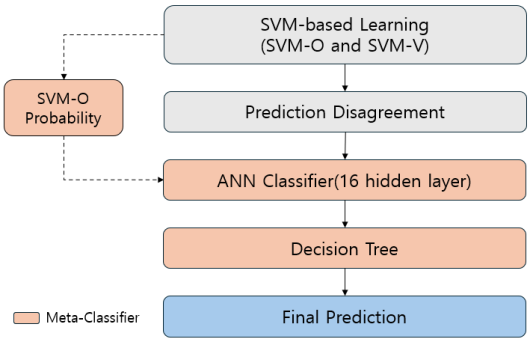

Fig. 3. Structural diagram of the meta-classifier

III. Experiment

A. Experiment Setup

All experiments were conducted in a local Windows 10 environment equipped with an Intel(R) Core(TM) i9-13900K CPU @ 3.00GHz, 64GB RAM, and an NVIDIA GeForce RTX 4090 GPU. The software environment was based on Python 3.10.16, with primary dependencies including scikit-learn 1.0.2, PyTorch 2.5.1+cu118, imbalanced-learn 0.9.0, pandas 2.0.3, and NumPy 1.24.4.

B. Training Setup

To prevent overfitting and ensure robust generalization in limited data scenarios, stratified 10-fold cross-validation was adopted, with each fold preserving uniform class distributions for stable performance evaluation. The classification model employed an RBF kernel-based Support Vector Machine (SVM), with the *class_weight* parameter set to balanced for class imbalance correction. Other SVM hyperparameters followed scikit-learn defaults: regularization parameter C=1.0, RBF kernel γ=scale, tolerance tol=10^{-3}, and iteration limit max_iter=−1.

For data augmentation, the VAE utilized a 16-dimensional latent space and a loss function combining mean squared error (MSE) reconstruction loss with KL divergence. The VAE was trained using the Adam optimizer with a learning rate of 0.001 over 100 epochs. Generated synthetic samples were combined with original data to construct balanced datasets for training the SVM-V model.

C. Evaluation Metric

To accurately assess classification performance under imbalanced data conditions, this study employed Precision, Recall, and F1-score as the primary evaluation metrics.

- Precision, defined as TP/(TP+FP), represents the proportion of correctly identified positive samples among all predicted positives. It reflects the model's ability to suppress false positives, which is crucial in applications such as anomaly detection and alert systems.
- Recall, calculated as TP/(TP+FN), measures the proportion of actual positives correctly identified by the model, indicating its effectiveness in detecting minority classes without omission.
- F1-score, the harmonic mean of Precision and Recall, provides a balanced measure by jointly considering both metrics. Therefore, F1-score serves as a balanced and reliable evaluation metric in imbalanced data scenarios.

Accordingly, this study quantitatively compared model performance using Precision, Recall, and F1-score.

D. Comparison of Oversampling Techniques

The purpose of this study is to investigate oversampling techniques for imbalanced data environments and systematically evaluate their impact on the generalization performance of classification models. For comparison, we adopted SMOTE, a representative interpolation-based method, and a proposed VAE-based data augmentation approach. While SMOTE generates synthetic samples through linear interpolation, the VAE produces nonlinear synthetic data via probabilistic sampling in the latent space. These differing mechanisms expand the minority class distribution in distinct ways, thereby influencing decision boundary formation and classifier stability.

In the experiments, three models with different oversampling techniques but identical classifier structures were compared:

1) Baseline model: a single classifier trained on the original imbalanced dataset.

2) Hybrid SMOTE model: data balancing with SMOTE, followed by reclassification of prediction mismatches using a meta-classifier.

3) Hybrid VAE model: data balancing using VAE-generated samples, combined with the same meta-classifier.

Performance was evaluated separately for each target variable (Q1, Q2, Q3, S1, S2, S3), with F1-score adopted as the primary evaluation metric to equally reflect the performance across classes. The comparative results of the three models—baseline, hybrid SMOTE, and hybrid VAE—are summarized in Table II, which presents the F1-scores obtained under different oversampling techniques.

TABLE II. Comparison of F1-Scores for SVM Models Using Different Oversampling Techniques

Method	Target					
	Q1	Q2	Q3	S1	S2	S3
Baseline SVM	0.5984	0.5726	0.5868	0.4188	0.5569	0.6144
Hybrid SMOTE	0.5988	0.6029	0.6480	0.4314	0.6143	0.6495
Hybrid VAE	**0.6733**	**0.6247**	**0.6564**	**0.4396**	**0.6302**	**0.6571**

E. Evaluation of Meta-Classifier Input Candidatesble

To determine the most suitable input for the meta-classifier, we compared the predicted probabilities from SVM-O (Pred O), trained on the original data, and SVM-V (Pred V), trained on VAE-augmented data. When tested on the same dataset, both models produced identical outputs for samples where their predictions were consistent; however,

for inconsistent samples, either Pred O or Pred V was used as input to the meta-classifier to generate the final prediction. Since the choice of input probabilities is critical to classification performance, we evaluated the meta-classifier using each type of input separately. As shown in Table III, Pred O consistently outperformed Pred V across most evaluation metrics. This can be attributed to the fact that a considerable portion of inconsistent samples was misclassified by SVM-V, owing to distributional distortions introduced during the oversampling process. In contrast, SVM-O, trained solely on the original data, better preserved the true class distribution and produced more reliable predictions near the decision boundary. Therefore, Pred O was selected as the final input to the meta-classifier in subsequent experiments.

F. Comparison with Baseline Models

To quantitatively evaluate the effectiveness of the proposed Hybrid-VAE framework, we conducted comparative experiments against representative machine learning classifiers, including LightGBM (LGBM) [12], XGBoost [13], and CatBoost [14]. Each model was independently trained and evaluated on the six target variables (Q1, Q2, Q3, S1, S2, S3), and the results are summarized in Table IV.

The results show that all Hybrid-VAE models outperformed their corresponding single classifiers, with the SVM-based Hybrid-VAE achieving the best overall performance. This improvement is largely attributed to its margin-based decision boundary learning, which enhances sensitivity to minority classes under imbalanced conditions.

A detailed analysis highlights two representative cases. For S3 (sleep onset latency), all models performed similarly (F1-score: 0.572–0.657), reflecting the reliability of objective sensor-based measurements and well-defined binary criteria from the National Sleep Foundation. Even in this case, however, the Hybrid-VAE model achieved the highest score. In contrast, S1 (total sleep time) presented greater difficulty due to its three-class structure and the variability of individual sleep patterns. Despite this challenge, the proposed model achieved a notable improvement over the baseline SVM (F1-score: 0.4188 → 0.4396, ≈5%).

These findings confirm that integrating VAE-based data augmentation with meta-classifiers enhances robustness across both binary and multi-class imbalanced environments, consistently outperforming conventional baselines.

IV. Conclusion

This study proposed a hybrid classification framework for sleep-related state prediction under imbalanced data conditions. The approach combines an SVM trained on original data with an SVM trained on VAE-augmented data, while a meta-classifier resolves prediction inconsistencies between the two. Experimental results showed that the proposed model consistently outperformed conventional oversampling methods and baseline machine learning classifiers, achieving notable improvements in F1-score and recall, particularly for minority classes.

Future work includes enhancing the meta-classifier architecture (e.g., through deep ensemble strategies), assessing scalability and generalizability on larger and more diverse datasets across heterogeneous domains, and reducing computational complexity to enable real-time prediction on wearable devices. These efforts are expected to improve both the clinical and practical applicability of the proposed framework in healthcare monitoring.

TABLE III. Comparison of Input Probability Values For The Meta-Classifier

Input	Target					
	Q1	Q2	Q3	S1	S2	S3
Pred O	**0.673**	**0.625**	**0.656**	**0.440**	**0.630**	**0.657**
Pred V	0.601	0.543	0.574	0.410	0.583	0.628

TABLE IV. Comparison of F1-score of The Proposed Model and ConventionalClassifiers

Model	Metrix	Target					
		Q1	Q2	Q3	S1	S2	S3
LGBM	Rec.	0.602	0.568	0.638	0.427	0.575	0.578
	Prec.	0.607	0.583	0.672	0.505	0.617	0.641
	F1	0.596	0.555	0.634	0.429	0.557	0.572
XGBoost	Rec.	0.601	0.573	0.638	0.444	0.594	0.608
	Prec.	0.606	0.591	0.674	0.521	0.673	0.699
	F1	0.599	0.563	0.636	0.432	0.583	0.607
Catboost	Rec.	0.624	0.567	0.608	0.426	0.550	0.581
	Prec.	0.628	0.601	0.660	0.506	0.643	0.752
	F1	0.621	0.545	0.600	0.405	0.511	0.559
Our Model	Rec.	**0.676**	**0.627**	**0.658**	**0.447**	**0.636**	**0.661**
	Prec.	**0.683**	**0.631**	**0.669**	**0.449**	**0.634**	**0.665**
	F1	**0.673**	**0.625**	**0.656**	**0.440**	**0.630**	**0.657**

Acknowledgment

The authors would like to thank the Electronics and Telecommunications Research Institute (ETRI) for providing the ETRI Lifelog Dataset 2024 used in this study.

References

[1] S. W. Oh, H. Jeong, S. Chung, J. M. Lim, K. J. Noh, S. Lee, and G. Jung, "Understanding human daily experience through continuous sensing: ETRI Lifelog Dataset 2024," arXiv preprint arXiv:2508.03698, Jul. 2025.

[2] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," J. Artif. Intell. Res., vol. 16, pp. 321–357, 2002.

[3] Y. Nam, Y. Kim, and J. Lee, "InsightSleepNet: interpretable and uncertainty-aware deep learning network for sleep staging using continuous photoplethysmography," BMC Med. Inform. Decis. Mak., vol. 24, no. 47, 2024.

[4] R. Blagus and L. Lusa, "SMOTE for high-dimensional class-imbalanced data," BMC Bioinf., vol. 14, no. 106, 2013.

[5] S. Zhao, Y. Li, S. Wang, and F. Xiao, "T-SMOTE: Temporal-oriented synthetic minority oversampling technique for imbalanced time series

classification," in Proc. 31st Int. Joint Conf. Artif. Intell. (IJCAI-22), 2022, pp. 4707–4713.

[6] C. Fan, F. Sun, L. Ju, M. Ning, and J. Zhang, "EEG data augmentation: Towards class imbalance problem in sleep staging tasks," J. Neural Eng., vol. 17, no. 5, Art. no. 056017, 2020.

[7] D. Hazra and Y. C. Byun, "SynSigGAN: Generative adversarial networks for synthetic biomedical signal generation," Biology (Basel), vol. 9, no. 9, Art. no. 301, 2020.

[8] B. E. Boser, I. M. Guyon, and V. N. Vapnik, "A training algorithm for optimal margin classifiers," in Proc. 5th Annu. Workshop Comput. Learn. Theory, 1992, pp. 144–152.

[9] D. P. Kingma and M. Welling, "Auto-encoding variational Bayes," in Proc. 2nd Int. Conf. Learn. Represent. (ICLR), Banff, AB, Canada, Apr. 2014.

[10] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," Nature, vol. 323, no. 6088, pp. 533–536, Oct. 1986.

[11] J. R. Quinlan, "Induction of decision trees," Mach. Learn., vol. 1, no. 1, pp. 81–106, Mar. 1986.

[12] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, and T.-Y. Liu, "LightGBM: A highly efficient gradient boosting decision tree," in Adv. Neural Inf. Process. Syst., vol. 30, pp. 3146–3154, 2017.

[13] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discov. Data Min., 2016, pp. 785–794.

[14] L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, and A. Gulin, "CatBoost: Unbiased boosting with categorical features," in Adv. Neural Inf. Process. Syst., vol. 31, pp. 6638–6648, 2018.