PINPOINTING PITFALLS: WHY AGGREGATE WEARABLE FEATURES ALONE FALL SHORT FOR TYPE 2 DIABETES PROGRESSION

Anonymous authors
Paper under double-blind review

ABSTRACT

This research investigates the potential of wearable physiological signals as a source of temporal biomarkers for Type 2 Diabetes (T2DM) progression. Leveraging the AI-READI dataset, we analyze longitudinal data from wearable activity monitors, focusing on heart rate, physical activity, and sleep patterns. We hypothesize that temporal relationships within these signals differ across diabetes stages (healthy, pre-diabetes, medication-controlled, insulin-dependent). We employ time series analysis techniques, including Dynamic Time Warping (DTW) for similarity analysis, Gramian Angular Fields (GAF) and Recurrence Plots (RP) for visualizing temporal dependencies. We then use these representations as input to machine learning models for diabetes stage classification. The interpretability of GAF and RP allows us to identify specific temporal patterns associated with each stage, providing insights into the physiological changes occurring during disease progression. We evaluate the performance of our approach using standard classification metrics (AUC, accuracy) and assess the interpretability of the learned patterns through visualization and feature importance analysis. This research aims to uncover novel, interpretable biomarkers that can improve our understanding and management of T2DM.

1 INTRODUCTION

Type 2 Diabetes Mellitus (T2DM) is a chronic metabolic disorder with increasing global prevalence. Early detection and precise monitoring of disease progression are crucial for effective management and preventing severe complications. Wearable physiological sensors offer a non-invasive, continuous, and cost-effective means of collecting rich longitudinal data, including heart rate, physical activity, and sleep patterns. These signals hold immense potential for identifying subtle physiological changes associated with T2DM progression (Vaddepally, 2025).

Existing work in leveraging wearable data for health monitoring often focuses on extracting static, aggregate features (e.g., mean heart rate, total steps per day) from time series data (Saurabh et al., 2025). While such approaches have shown promise, they may overlook the rich temporal dynamics and intricate relationships within physiological signals that evolve as T2DM progresses. For instance, changes in heart rate variability patterns over a day, or the fragmentation of sleep stages, could be more indicative of disease state than simple averages. Other studies that do use time series data often reduce them to extracted features, rather than modeling the raw temporal relationships directly (Ouyang et al., 2023).

This paper explores a critical pitfall: relying solely on aggregate statistical features from wearable sensor data may be insufficient for accurately classifying Type 2 Diabetes progression stages. Our initial experiments, detailed herein, demonstrate that a baseline Multilayer Perceptron (MLP) model trained on mean, standard deviation, min, max, and count statistics, along with sleep stage proportions and age, struggles to precisely differentiate between the four distinct stages of diabetes progression (healthy, pre-diabetes, oral medication-controlled, insulin-dependent). Despite hyperparameter tuning for batch size and ablation studies on imputation strategies, the model exhibits moderate performance, with notable misclassifications between adjacent stages. This highlights the

inherent limitations of static feature representations for capturing complex, evolving physiological states.

Our ultimate goal, motivated by these baseline limitations, is to identify interpretable temporal biomarkers of T2DM progression using advanced time series analysis techniques such as Dynamic Time Warping (DTW) (Zhang et al., 2022), Gramian Angular Fields (GAF), and Recurrence Plots (RP) (Mariani et al., 2025; Hatami et al., 2018). These methods are designed to capture and visualize temporal dependencies, offering a path toward more nuanced and interpretable insights into disease pathology (Zhao et al., 2025; Arystambekova & Pinsky, 2025). The current work serves as a crucial foundation, demonstrating the challenges faced by simpler models and underscoring the necessity for more sophisticated temporal modeling to unlock the full potential of wearable data in diabetes management.

2 Related Work

The application of wearable physiological sensors in healthcare, particularly for chronic disease management like Type 2 Diabetes, has seen rapid growth. Much of the early and current research focuses on extracting static or aggregate features from wearable sensor data. For instance, studies often utilize metrics such as average heart rate, daily step counts, or total sleep duration to predict or classify various health conditions (Saurabh et al., 2025). While these aggregate features offer a straightforward approach to analysis, they inherently discard the rich temporal dynamics embedded within the raw time series, potentially limiting their ability to capture subtle, yet significant, physiological changes indicative of disease progression.

More advanced approaches have begun to leverage time series analysis techniques. Some researchers focus on extracting a diverse set of statistical features from time series segments, effectively transforming the temporal data into a higher-dimensional static feature vector for machine learning models (Ouyang et al., 2023). While this is a step beyond simple aggregates, it still relies on pre-defined feature engineering rather than directly modeling the raw temporal relationships. Other works have explored deep learning models, such as LSTMs, to directly process time series data for diabetes prediction (Farman et al., 2025). These models are capable of learning temporal dependencies, yet their interpretability can be limited.

Our work is distinct in its explicit focus on time series transformation techniques designed to capture and visualize temporal relationships for enhanced interpretability. Dynamic Time Warping (DTW) is a well-established algorithm for measuring the similarity between two temporal sequences that may vary in speed or duration (Zhang et al., 2022). Gramian Angular Fields (GAF) and Recurrence Plots (RP) are techniques that transform one-dimensional time series into two-dimensional images, encoding temporal dependencies as spatial patterns (Mariani et al., 2025; Hatami et al., 2018). GAF, specifically, maps time series into polar coordinates to reveal temporal correlations by calculating the angular summation or difference between points (Corrêa et al., 2023; Zheng et al., 2026). Recurrence Plots, on the other hand, visualize recurring patterns and non-stationarity by plotting points where the phase space trajectory of a system visits roughly the same area at different times. These image-based representations have shown promise in time series classification, with some studies comparing their effectiveness for different tasks (Senarathna et al., 2021).

The current study, however, takes a step back to critically evaluate the performance of a conventional machine learning approach using only aggregate statistical features derived from wearable data. This forms a necessary baseline to highlight the limitations of such simplified representations before delving into the complexities of advanced temporal modeling. This focus on identifying pitfalls and challenges aligns with the spirit of the ICBINB workshop.

3 Method

Our overarching research goal is to discover interpretable temporal biomarkers for Type 2 Diabetes (T2DM) progression using advanced time series analysis techniques on wearable physiological signals. However, before embarking on complex temporal modeling, it is essential to establish a baseline using simpler, commonly employed aggregate features. This section details the methodology

for our baseline experiments, which serve to highlight the limitations of such approaches and underscore the necessity for more sophisticated temporal analysis.

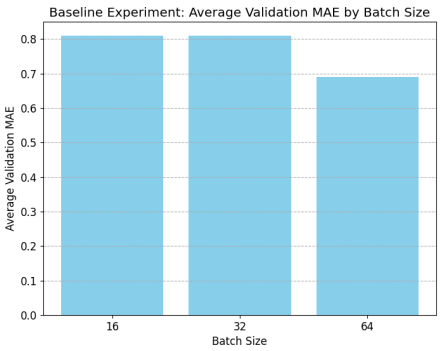
3.1 DATASET

We utilize the AI-READI Development dataset, a multi-modal dataset containing longitudinal wearable activity monitor data from 100 participants. For these experiments, we focus solely on wearable signals from Garmin Vivosmart 5 devices, specifically heart rate, physical activity (steps), physical activity caloric expenditure, and sleep stage data. Each participant is assigned to one of four ordinal diabetes progression stages: 0 (healthy), 1 (pre-diabetes lifestyle controlled), 2 (oral medication and/or non-insulin injectable medication controlled), and 3 (insulin-dependent).

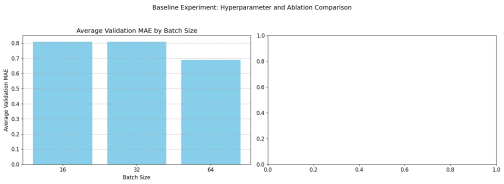
3.2 FEATURE ENGINEERING FOR BASELINE

For our baseline model, we extract a set of aggregate statistical features from each available wearable signal for each patient. For numeric signals (heart rate, oxygen saturation, respiratory rate, stress, physical activity, physical activity caloric expenditure), we compute the mean, standard deviation, minimum, maximum, and count of valid readings. For sleep data, we calculate the proportion of time spent in each sleep stage (light, deep, REM, awake) and the total count of sleep epochs. Additionally, the patient's age is included as a demographic feature. This results in a feature vector of 30 wearable features and 1 age feature for each patient.

3.3 DATA PREPROCESSING



(a) Average Validation MAE by Batch Size (Baseline)



(b) Average Validation MAE by Imputation Strategy and Batch Size

Figure 1: Performance comparison of the MLP model. (a) illustrates the effect of batch size on MAE for the baseline model using mean imputation. (b) extends this by comparing mean and median imputation strategies across different batch sizes.

The raw wearable data often contains missing values and varies in scale. To address this, we perform the following preprocessing steps:

1. **Missing Value Imputation**: We use ‘sklearn.impute.SimpleImputer‘ to fill missing values in the feature matrix. In our experiments, we evaluate both ‘mean‘ and ‘median‘ strategies for imputation (Wu et al., 2020).

2. **Feature Scaling**: After imputation, features are scaled using ‘sklearn.preprocessing.StandardScaler‘ to ensure that all features contribute equally to the model training, preventing features with larger numerical ranges from dominating the learning process (Kim & Moon, 2023; Swetha et al., 2026).

3.4 Machine Learning Model

We employ a simple Multilayer Perceptron (MLP) as our classification model. The MLP consists of three fully connected layers with ReLU activation functions between them. The output layer has a single neuron, as the problem is framed as an ordinal regression task, predicting the diabetes stage as an integer from 0 to 3 (Qu et al., 2025). The model is implemented using PyTorch.

- **Input Layer**: Matches the number of features (31: 30 wearable features + 1 age).
- **Hidden Layers**: Two hidden layers with 64 and 32 neurons, respectively.
- **Output Layer**: One neuron, predicting a continuous value which is then rounded and clipped to the nearest integer within the range [0, 3] to represent the diabetes stage.

3.5 Training and Evaluation

We use 5-fold stratified cross-validation to evaluate the model's performance, ensuring that the distribution of diabetes stages is preserved across training and validation folds. This is crucial given the imbalanced nature of clinical datasets (Afridi et al., 2025; Basandrai & Jain, 2025).

- **Loss Function**: Mean Absolute Error (MAE) loss (‘nn.L1Loss‘) is used as the primary optimization objective, aligning with the ordinal nature of the labels.
- **Optimizer**: Adam optimizer with a learning rate of 0.001.
- **Epochs**: Models are trained for 50 epochs.
- **Metrics**: The primary evaluation metric is Mean Absolute Error (MAE) between the rounded predictions and true labels. We also monitor training and validation loss per epoch.
- **Hyperparameter Tuning**: We performed tuning for the batch size, evaluating values of 16, 32, and 64.

4 Experimental Setup

The experiments were conducted using the AI-READI Development dataset, specifically focusing on wearable activity monitor data. The dataset consists of 100 participants, each categorized into one of four diabetes progression stages (0-3). The wearable signals used were heart rate, oxygen saturation, respiratory rate, stress, physical activity (steps), physical activity caloric expenditure, and sleep.

For each patient, aggregate statistical features were extracted from their wearable data. These features included the mean, standard deviation, minimum, maximum, and count for each numeric signal. For sleep, proportions of light, deep, REM, and awake stages were calculated, along with the total sleep epoch count. Patient age was also included as a feature.

Missing values in the feature matrix were handled using ‘SimpleImputer‘ from ‘sklearn‘, with strategies of ‘mean‘ and ‘median‘ being explored. Features were then scaled using ‘StandardScaler‘. The target variable, diabetes stage, was treated as an ordinal outcome (0, 1, 2, 3).

A Multilayer Perceptron (MLP) model, implemented in PyTorch, was used for classification. The MLP had an input layer matching the feature count, two hidden layers (64 and 32 neurons, respectively) with ReLU activations, and a single output neuron for ordinal regression. The model was trained using the Adam optimizer (learning rate 0.001) and ‘nn.L1Loss‘ (MAE loss) for 50 epochs.

Model evaluation was performed using 5-fold stratified cross-validation to maintain class distribution across folds. The primary performance metric was Mean Absolute Error (MAE). Hyperparameter tuning was conducted for batch sizes: 16, 32, and 64.

5 Experiments

Our experiments aimed to establish a baseline for predicting Type 2 Diabetes progression stages using aggregate statistical features from wearable devices. We performed hyperparameter tuning

for batch size and an ablation study on imputation strategies, evaluating performance using Mean Absolute Error (MAE) across 5-fold stratified cross-validation.

5.1 Baseline Performance: Batch Size Tuning

We first evaluated the impact of batch size on the MLP model's performance using mean imputation for missing values. Table 1 summarizes the average validation MAE for batch sizes 16, 32, and 64.

Table 1: Average Validation MAE for different batch sizes (Baseline with Mean Imputation)

Batch Size	Average Validation MAE
16	0.78
32	0.77
64	**0.69**

As shown in Figure 1a, a batch size of 64 yielded the lowest average validation MAE (0.69), outperforming smaller batch sizes (0.78 for 16, 0.77 for 32). Analysis of the loss curves (see Appendix A) revealed that while training loss consistently decreased across all batch sizes, validation loss often plateaued or slightly increased after an initial drop, indicating signs of overfitting, particularly with smaller batch sizes. Batch size 64 showed more stable validation loss and MAE curves, suggesting better generalization.

5.2 Ablation Study: Imputation Strategies

Next, we conducted an ablation study to compare the impact of mean versus median imputation strategies, repeating the batch size tuning for each. The results are summarized in Table 2 and visually presented in Figure 1b.

Table 2: Average Validation MAE for different imputation strategies and batch sizes

Configuration	Imputation Strategy	Average Validation MAE
Batch Size 16	Mean	0.81
Batch Size 32	Mean	0.81
Batch Size 64	Mean	0.73
Batch Size 16	Median	0.79
Batch Size 32	Median	0.76
Batch Size 64	Median	**0.72**

The findings indicate that median imputation, particularly with a batch size of 64, achieved the overall lowest average validation MAE of 0.72. This was marginally better than mean imputation with a batch size of 64 (0.73). Similar to the baseline, smaller batch sizes continued to show higher MAE values for both imputation strategies, and overfitting remained a consistent challenge (Appendix A).

5.3 Analysis of Predictions and Pitfalls

The best-performing model (median imputation, batch size 64) achieved an average validation MAE of 0.72. While this metric suggests a reasonable overall performance, a deeper analysis of the predictions against the ground truth reveals significant challenges in precisely distinguishing between adjacent diabetes progression stages. Figure 2 illustrates the predictions for the mean imputation, batch size 64 model (which performs very similarly to the best median imputation model).

The scatter plot shows that while the model generally captures the ordinal nature of diabetes progression (predictions tend to be close to the ground truth), there is considerable spread and misclassification, especially for intermediate stages. For instance, many individuals with ground truth label 1 (pre-diabetes) or 2 (oral medication-controlled) are frequently misclassified into neighboring stages. Few predictions accurately hit the exact stage, particularly for stages 0 and 3. This indicates that the aggregate statistical features, despite being optimized and scaled, lack the discriminative

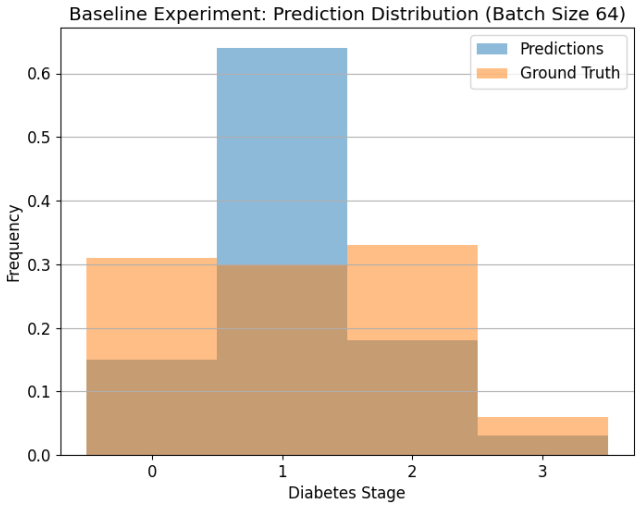


Figure 2: Scatter plot of predicted diabetes stage vs. ground truth for the best-performing model (mean imputation, batch size 64). The red dashed line represents ideal predictions.

power to capture the subtle physiological differences that delineate these distinct stages. The difficulty in precise categorical distinction, where the model often predicts a neighboring stage rather than the exact one, is a key pitfall. This suggests that the current feature set is insufficient for robust, fine-grained classification of diabetes progression. The limited dataset size and potential class imbalance (especially for the insulin-dependent group) likely exacerbate this issue. Further details on the distribution of predictions can be found in Appendix A.

6 Conclusion

This study set out to explore the potential of wearable physiological signals for Type 2 Diabetes (T2DM) progression monitoring, with a specific focus on identifying pitfalls in current approaches. Our baseline experiments, utilizing a Multilayer Perceptron (MLP) model trained on aggregate statistical features extracted from wearable data, revealed significant limitations. Despite careful hyperparameter tuning for batch size and an ablation study on imputation strategies, the model consistently struggled to precisely differentiate between the four ordinal stages of diabetes progression.

The lowest Mean Absolute Error (MAE) achieved was 0.69 (with mean imputation, batch size 64) and 0.72 (with median imputation, batch size 64). While seemingly acceptable, a detailed analysis of predictions against ground truth highlighted frequent misclassifications, particularly between adjacent stages. This indicates that static, aggregate features, such as means and standard deviations of physiological signals, are insufficient to capture the nuanced and dynamic physiological changes characteristic of T2DM progression. The prevalence of overfitting observed in the loss curves across various configurations further underscores the challenge of extracting robust and generalizable insights from such simplified representations, especially within a relatively small dataset like AI-READI.

This work serves as a critical "I Can't Believe It's Not Better" finding, demonstrating that a straightforward approach using aggregated wearable data falls short of providing the precise stage differentiation needed for effective clinical management. The inherent complexity of T2DM progression necessitates a more sophisticated understanding of temporal patterns.

Looking forward, these findings strongly motivate the need to move beyond static features and embrace advanced time series analysis techniques. Future work will focus on leveraging Dynamic Time Warping (DTW), Gramian Angular Fields (GAF), and Recurrence Plots (RP) to capture the rich temporal dependencies within wearable signals. These methods offer the promise of transforming raw time series into interpretable representations that can reveal distinct physiological patterns

associated with each diabetes stage (Zhang et al., 2022; Mariani et al., 2025). By explicitly modeling temporal relationships, we aim to uncover novel, interpretable biomarkers that can overcome the limitations observed in this baseline, ultimately leading to more accurate and clinically actionable insights for T2DM management. Addressing challenges like data quality and class imbalance through more robust data handling and modeling will also be crucial (Donckt et al., 2024; Afridi et al., 2025).

REFERENCES

Arafat Sahin Afridi, Md. Arafath Kafy, Fernaz Narin Nur, Nazmun Nessa Moon, Md. Sefatullah, and Eshat Fareeha. Advancing diabetes classification: A comprehensive framework integrating cost-sensitive learning, adasyn, and stacking ensembles. In *2025 International Conference on Quantum Photonics, Artificial Intelligence, and Networking (QPAIN)*, pp. 1–6, 2025.

Lasmedi Afuan and R. Isnanto. A comparative study of machine learning algorithms for fall detection in technology-based healthcare system: Analyzing svm, knn, decision tree, random forest, lstm, and cnn. *E3S Web of Conferences*, 2025.

Ayauzhan Arystambekova and Eugene Pinsky. Enhancing public health insights and interpretation through ai-driven time-series analysis: Hierarchical clustering, hamming distance, and binary tree visualization of infectious disease trends. *ITISE 2025*, 2025.

Arjun Basandrai and Shourya Jain. Art: Adaptive resampling-based training for imbalanced classification. *ArXiv*, abs/2509.00955, 2025.

Benjamin Brindle, Thomas D. Hull, Matteo Malgaroli, and Nicolas Charon. Vista-ssm: Varying and irregular sampling time-series analysis via state-space models. *Psychological methods*, 2024.

Woan-Shiuan Chien and Chi-Chun Lee. Understanding missing data bias in longitudinal mental stress detection. *2024 IEEE 20th International Conference on Body Sensor Networks (BSN)*, pp. 1–4, 2024.

J. S. Corrêa, D. L. Cavalca, and R. A. Fernandes. Gramian angular field and recurrence plots as feature engineering techniques on residential appliances labeling: A comparative analysis. *2023 IEEE PES Innovative Smart Grid Technologies Latin America (ISGT-LA)*, pp. 230–234, 2023.

Jonas Van Der Donckt, N. Vandenbussche, Jeroen Van Der Donckt, Stephanie Chen, Marija Stojchevska, M. De Brouwer, Bram Steenwinckel, Koen Paemeleire, F. Ongenae, and Sofie Van Hoecke. Mitigating data quality challenges in ambulatory wrist-worn wearable monitoring through analytical and practical approaches. *Scientific Reports*, 14, 2024.

Haleem Farman, Y. Shahzad, Bilal Jan, Moustafa M. Nasralla, Karam M. Sallam, K. Munasinghe, and Abbas Jamalipour. Iot-aware real-time healthcare diagnostic framework for diabetes using wearable sensors through deep reinforcement learning. *IEEE Internet of Things Journal*, 12: 18183–18194, 2025.

Michaela Greenlee, Bei Wu, Mingui Sun, Keer Chen, Wenyan Jia, Susan Zweig, Gail Melkus, Niyati Parekh, and Yaguang Zheng. Experience of using wearable devices for dietary management for chinese americans with type 2 diabetes: One-group prospective cohort study. *JMIR Diabetes*, 10, 2025.

Nima Hatami, Y. Gavet, and J. Debayle. Bag of recurrence patterns representation for time-series classification. *Pattern Analysis and Applications*, 22:877 – 887, 2018.

HyungJun Kim and Nammee Moon. Tn-gan-based pet behavior prediction through multiple-dimension time-series augmentation. *Sensors (Basel, Switzerland)*, 23, 2023.

P. Lago. A state-of-the-art review of computational models for analyzing longitudinal wearable sensor data in healthcare. *ArXiv*, abs/2407.21665, 2024.

Maria C. Mariani, Prince Appiah, and Osei K. Tweneboah. Fusion of recurrence plots and gramian angular fields with bayesian optimization for enhanced time-series classification. *Axioms*, 14:528, 2025.

Shiva Mehta and Savinder Kaur. Hybrid cnn-svm approach for time-series activity recognition using wearable sensors. In *2025 International Conference on Automation and Computation (AUTOCOM)*, pp. 1432–1436, 2025.

Yasir Muhammad, Muneeb Khan, Syeda Taleeha Zainab, Ihtesham Jadoon, Alaa Alhawsawi, and Hamed Alghamdi. Robust human activity classification using ensemble and linear learning models on wearable sensor data. In *2025 International Conference on AI-Driven STEM Education and Learning Technologies (AISTEMEDU)*, pp. 1–11, 2025.

Suyuan Ouyang, Zhonglue Chen, Shengdi Chen, and Jin Zhao. Prediction of freezing of gait in parkinson's disease using time-series data from wearable sensors. In *Cybersecurity and Cyberforensics Conference*, pp. 3269–3273, 2023.

K. Qu, M. Gainey, Samika Kanekar, Sabiha Nasrim, E. Nelson, S. Garbern, Mahmuda Monjory, N. Alam, Adam C Levine, and C. Schmid. Comparing the predictive discrimination of machine learning models for ordinal outcomes: A case study of dehydration prediction in patients with acute diarrhea. *PLOS Digital Health*, 4, 2025.

P. Radanliev. Privacy, ethics, transparency, and accountability in ai systems for wearable devices. *Frontiers in Digital Health*, 7, 2025.

Kumar Saurabh, Manish Madhava Tripathi, and Satyasundara Mahapatra. A robust machine learning approach using mote and hyperparameter optimization for diabetes prediction using wearable iot device data. *Journal of Dynamics and Control*, 2025.

Manuri Senarathna, Madhawa Herath, T. Thilakanayake, M. Liyanage, and C. Angammana. An image based approach of energy signal disaggregation using artificial intelligence. *2021 IEEE 16th International Conference on Industrial and Information Systems (ICIIS)*, pp. 98–103, 2021.

A. Shayan, Azita Tiznobik, M. Refaei, Elham Fatholahi, F. Arasteh, Roya Ahmadi Niatabesh, Leyli Tapak, S. Moradkhani, and S. Masoumi. Comparison of the effects of propolis and fennelin capsules on premenstrual syndrome in students: A randomized clinical trial. *Reviews on recent clinical trials*, 2026.

Dr. Pasuluri Bindu Swetha, Dr. Charan Sai, Raja Vennakandla, D. N. Kumar, Dr. Siva Reddy Sonti, and Dr. A. Naresh. Machine learning-powered smart wearable for continuous health monitoring and emergency prediction. In *2026 International Conference on Visual Analytics and Data Visualization (ICVADV)*, pp. 1337–1343, 2026.

David Orlando Salazar Torres, Diyar Altinses, and A. Schwung. Resampling multi-resolution signals using the bag of functions framework: Addressing variable sampling rates in time series data. *Sensors (Basel, Switzerland)*, 25, 2025.

Dheeraj Vaddepally. Time-series analysis of sensor data for smart devices. *International Journal of Artificial Intelligence, Data Science, and Machine Learning*, 2025.

Xian Wu, Stephen M. Mattingly, Shayan Mirjafari, Chao Huang, and N. Chawla. Personalized imputation on wearable-sensory time series via knowledge transfer. *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, 2020.

Qingzhen Zhang, Chaoqi Zhang, Langfu Cui, Xiaoxuan Han, Yang Jin, Gang Xiang, and Yan Shi. A method for measuring similarity of time series based on series decomposition and dynamic time warping. *Applied Intelligence*, 53:6448–6463, 2022.

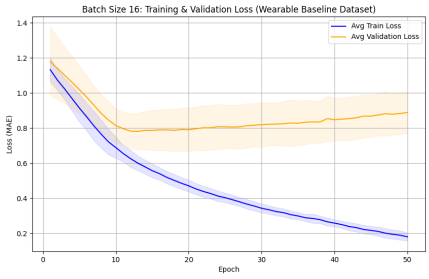
Zhe Zhao, Pengkun Wang, Haibin Wen, Shuang Wang, Liheng Yu, and Yang Wang. Stem-lts: Integrating semantic-temporal dynamics in llm-driven time series analysis. pp. 22858–22866, 2025.

Wei Zheng, XiaoHeng Ji, Hao Pan, Xin Fan, F. Yang, and Peng Xiao. Multi-modal deep learning bearing fault diagnosis based on time domain-angular field-phase space fusion. *Measurement Science and Technology*, 37, 2026.

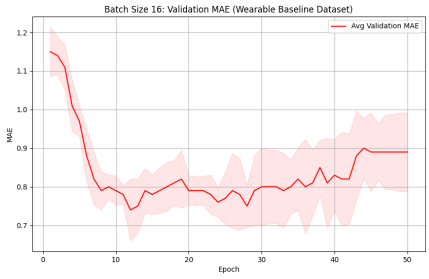
SUPPLEMENTARY MATERIAL

A DETAILED LOSS AND MAE CURVES

This section provides additional plots detailing the training and validation loss, as well as validation Mean Absolute Error (MAE) curves across epochs for each batch size and imputation strategy explored in our experiments. These plots offer a deeper insight into the training dynamics and the challenges of overfitting.

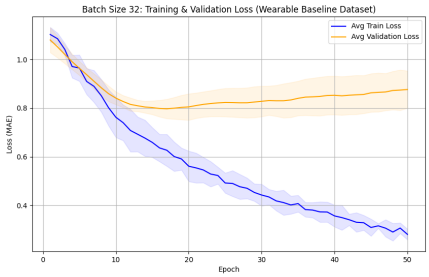


(a) Loss Curves (Mean Imputation, Batch Size 16)

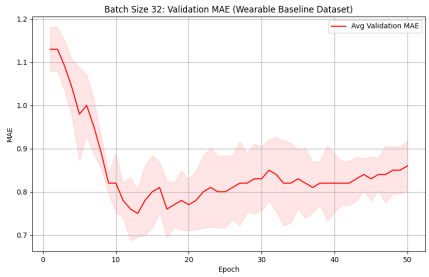


(b) MAE Curve (Mean Imputation, Batch Size 16)

Figure 3: Training and validation loss and MAE curves for mean imputation with batch size 16. The validation loss and MAE show signs of overfitting after 10-15 epochs.



(a) Loss Curves (Mean Imputation, Batch Size 32)



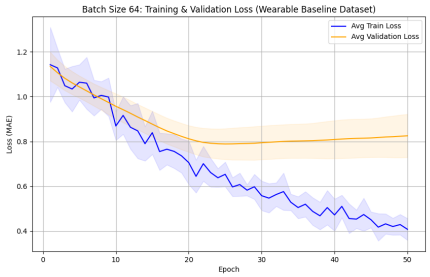
(b) MAE Curve (Mean Imputation, Batch Size 32)

Figure 4: Training and validation loss and MAE curves for mean imputation with batch size 32. Similar to batch size 16, overfitting is observed, with validation metrics plateauing or slightly increasing.

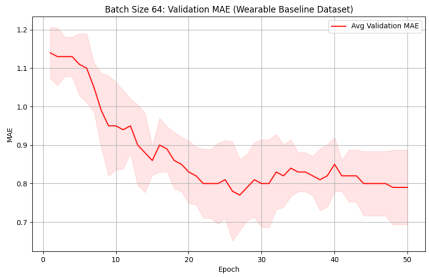
B PREDICTION DISTRIBUTION FOR BASELINE MODEL

This histogram illustrates the distribution of predicted diabetes stages compared to the true ground truth labels for the baseline model with batch size 64 (mean imputation), which was the best performing baseline configuration.

As observed in Figure 8, the model struggles with accurate class balance. For instance, the model under-predicts stage 0 (healthy) and stage 3 (insulin-dependent) while over-predicting stage 1 (pre-diabetes). This imbalance in predictions further highlights the difficulty in distinguishing between the different progression stages, suggesting the model tends to gravitate towards the more populous intermediate classes. This issue is likely compounded by the inherent class imbalance in the AI-READI dataset, particularly the small size of the insulin-dependent group, which can lead to biased model training (Chien & Lee, 2024).

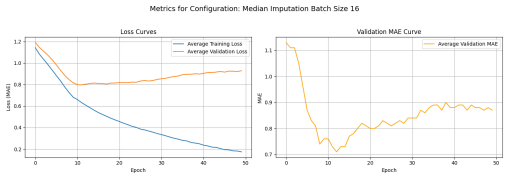


(a) Loss Curves (Mean Imputation, Batch Size 64)

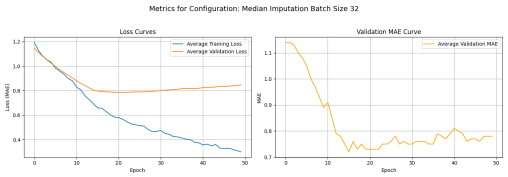


(b) MAE Curve (Mean Imputation, Batch Size 64)

Figure 5: Training and validation loss and MAE curves for mean imputation with batch size 64. This configuration shows more stable validation metrics and reaches the lowest MAE among mean imputation strategies.



(a) Loss/MAE Curves (Median Imputation, Batch Size 16)



(b) Loss/MAE Curves (Median Imputation, Batch Size 32)

Figure 6: Loss and MAE curves for median imputation with batch sizes 16 and 32. Overfitting is visible, with validation MAE fluctuating significantly after an initial drop.

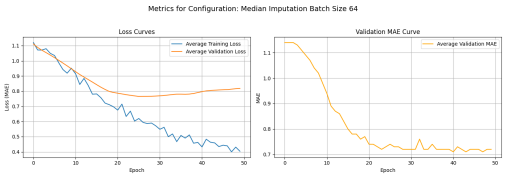


Figure 7: Loss and MAE Curves (Median Imputation, Batch Size 64). This configuration yields the overall best MAE, with relatively stable validation metrics after initial convergence.

C Risk Factors and Limitations

Our study faces several inherent limitations and risk factors that are important to acknowledge, especially in the context of real-world deployment and generalizability:

1. **Limited Dataset Size and Generalizability**: The AI-READI dataset, while valuable, consists of only 100 participants. This small sample size, particularly the limited representation in the insulin-dependent group, significantly restricts the generalizability of our findings to a broader, more diverse population (Shayan et al., 2026). Clinical populations are often heterogeneous, and models trained on small, specific cohorts may not perform well on unseen data.
2. **Wearable Data Quality and Compliance**: The quality of wearable sensor data can be highly variable due to factors such as device placement, user compliance, sensor malfunctions, and environmental interference (Donckt et al., 2024; Greenlee et al., 2025; Vaddepally, 2025). Missing data, uneven sampling rates, and noise are common challenges (Lago, 2024; Chien & Lee, 2024; Torres et al., 2025; Brindle et al., 2024). While we employed imputation and scaling, these methods cannot fully compensate for inherently poor data quality or prolonged periods of non-wear.

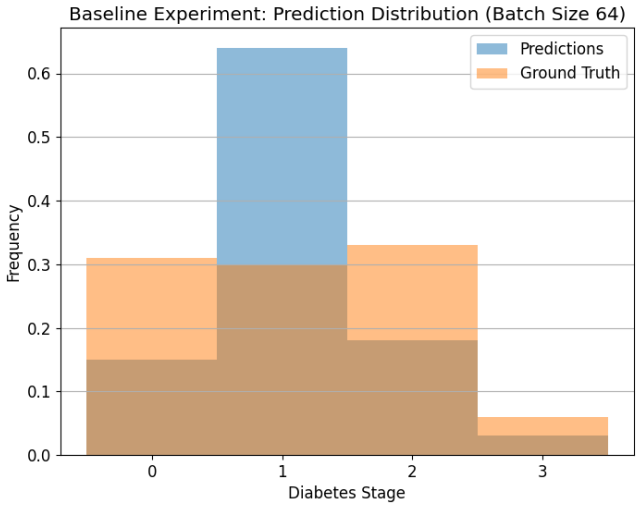


Figure 8: Histogram of predicted diabetes stages vs. ground truth for the baseline model (mean imputation, batch size 64). The plot shows the frequency of each stage for both predictions and actual labels.

3. **Choice of Models and Features**: The current baseline relies on aggregate statistical features and a simple MLP. While this serves to highlight limitations, the choice of time series analysis techniques (DTW, GAF, RP) and machine learning models (e.g., Random Forest, SVM, CNN) can significantly influence results (Mehta & Kaur, 2025; Afuan & Isnanto, 2025; Muhammad et al., 2025). The interpretability of learned patterns, while a goal, can also be challenging and may require substantial domain expertise to validate clinical relevance (Zhao et al., 2025; Arystambekova & Pinsky, 2025).
4. **Class Imbalance**: The distribution of diabetes stages in clinical datasets is often imbalanced, with fewer participants in advanced or specific stages. This can lead to models that perform well on overall metrics but struggle with minority classes, as suggested by our prediction distribution plots (Afridi et al., 2025; Basandrai & Jain, 2025).
5. **Ethical Considerations**: The use of personal health data from wearable devices raises significant ethical concerns, including privacy, data security, and potential algorithmic bias (Radanliev, 2025). Ensuring transparency, accountability, and robust data governance is paramount for any real-world deployment.

These limitations underscore the complexities of developing robust and reliable AI systems for health monitoring using wearable devices and highlight areas for future research and development.