Towards a personalized pregnancy experience: Forecasting symptoms using graph neural networks and digital health technologies

Rui Zhu* V.ZHU@MAIL.UTORONTO.CA
Department of Computer Science, University of Toronto, Toronto, Canada
Hospital for Sick Children, Toronto, Canada
Vector Institute, Toronto, Canada
Jennifer Yu* JENNIFERJIE.YU@MAIL.UTORONTO.CA
Department of Computer Science, University of Toronto, Toronto, Canada
Hospital for Sick Children, Toronto, Canada
Vector Institute, Toronto, Canada
Stephen H. Friend FRIEND@4YOUANDME.ORG
4YouandMe, Seattle, WA, USA
Department of Psychiatry, University of Oxford, Oxford, UK
Sarah M. Goodday SARAH@4YOUANDME.ORG
4YouandMe, Seattle, WA, USA
Department of Psychiatry, University of Oxford, Oxford, UK
Bo Wang BOWANG@VECTORINSTITUTE.AI
Department of Medical Biophysics, University of Toronto, Toronto, Canada
Vector Institute, Toronto, Canada
University Health Network, Toronto, Canada
Anna Goldenberg ANNA.GOLDENBERG@UTORONTO.CA
Department of Computer Science, University of Toronto, Toronto, Canada
Department of Laboratory Medicine and Pathobiology, University of Toronto, Toronto, Canada
Hospital for Sick Children, Toronto, Canada
Vector Institute, Toronto, Canada

Abstract

Pregnancy is an intricate process involving substantial physiological changes that impact both maternal and fetal health. In this study, we demonstrate the ability to predict two common symptoms during the third trimester of pregnancy—edema and fatigue—using physiological measures from wearable devices and self-reported daily surveys. Our approach employs a Graph Neural Network (GNN) framework, enhanced with a modified weighted Cross-Entropy loss to improve prediction performance. The model achieved AUC scores of 77.27% for edema and 69.70% for fatigue. Additionally, we aligned data from self-reported pregnancy symptoms with clinical examinations to carefully select participant cohorts for our experiments. We also explored how various features identified by the GNN are linked to these symptoms, gaining deeper insights into the relationship between physiological measures and pregnancy symptoms. Our findings indicate that heart rate variability plays a significant role in predicting symptoms of edema and fatigue, and features related to low-intensity activity also have a notable impact. Some of our findings closely align with previous studies on pregnancy. Our research serves as a proof of concept that symptoms can be predicted using wearable data, which may enhance the immediate well-being of expectant mothers and potentially personalize the overall pregnancy experience.

Keywords: Time-series forecasting, Graph Neural Networks, Wearable technology, Maternal Health.

* Equal contribution

Data and Code Availability The dataset used in this study comes from the Better Understanding the Metamorphosis of Pregnancy (BUMP) study (Goodday et al., 2022). The BUMP study is a digital health longitudinal study. It aims to gain a deeper understanding of the pre-pregnancy and pregnancy individual-level experience through wearable devices, smart scales, and smartphone apps. The BUMP dataset will be published soon on the Synapse platform at Sage Bionetworks. Please check the 4YouandMe website or contact the corresponding author for an update on when study data will be available and how to access it. The code is available here.

Institutional Review Board (IRB) The study was approved by the IRB, Advarra (Pro00047893).

1. Introduction

Despite major advancements in prenatal care over the past several decades, maternal health remains a significant global public health problem associated with considerable morbidity, mortality (Gon et al., 2018) and economic burden (Moran et al., 2020). One of the reasons is the complex and highly heterogeneous physiological and psychological changes during pregnancy. Current approaches, such as periodic clinical check-ups, are limited by assessing symptoms in cross-sectional snapshots in time and retrospective recall. Wearables offer a solution by enabling dynamic, longitudinal monitoring of objective health measures specific to pregnancy. Paired with mobile apps and machine learning methods, wearables data could enable a more personalized approach to monitoring maternal health, addressing the high heterogeneity of health trajectories over pregnancy.

This research serves as a proof of concept that symptoms can be predicted using wearable data, demonstrating the capabilities of predictive models in health monitoring. Here, we focus on two common pregnancy symptoms: edema and fatigue. Edema, often severe in the third trimester, affects up to 80% of pregnant women (Sanghavi and Rutherford, 2014) and is characterized by pain, nighttime cramps, numbness, tingling, and leg heaviness (Smyth et al., 2015). Fatigue, another frequent symptom, is an overwhelming tiredness that reduces physical and mental activity (Wan et al., 2017). Overall, being aware of these pregnancy-related symptoms beforehand could enhance the immediate well-being of expectant mothers, while the approach used here could be applied to other symptoms of pregnancy.

While leveraging wearables to predict pregnancy-related symptoms holds promise, this approach comprises some challenges. Firstly, there is significant variability in physiological signals from wearables and daily self-reported survey responses (Sanghavi and Rutherford, 2014). Secondly, low adherence to completing app-based surveys and continuous use of wearables is often observed in digital health technology (DHT) research that results in large amounts of missing data (Sanghavi and Rutherford, 2014). These challenges increase the complexity of machine learning analyses for this type of prediction task.

In this study, we aim to navigate these challenges and evaluate the potential of wearable data in forecasting pregnancy-related symptoms. We used multimodal personal DHT data from the Better Understanding the Metamorphosis of Pregnancy (BUMP) study (Goodday et al., 2022) - a US-based digital health, participant-centric research study that tracks individuals from preconception to the fourth trimester (three-month postpartum). Based on measured wearable features from the previous week(s) in the third trimester, we aim to predict self-reported pregnancy symptoms (edema and fatigue) from mobile app surveys for the next day using a dynamic graphical neural network (GNN). Dynamic GNNs have been proposed to handle both structural and temporal information over dynamic data and achieved remarkable results in many predictive tasks (Skarding et al., 2020; Zhu et al., 2022). The strong capability of learning relational dependencies within physiological signals enables us to learn the change in feature dynamics over time; therefore improving the performance of symptom prediction and our understanding of the underlying feature interactions.
The main contributions of this project are as follows:

1. Leveraged Graph Neural Networks (GNN) to predict the next day's symptom-related survey responses using both self-reported survey data and objective physiological measures collected from the preceding week(s) using wearables.

2. Performed a careful study on the impact of various loss functions and GNN model architectures and examined their effectiveness and robustness in forecasting self-reported edema and fatigue. Our final model architecture uses Multivariate Time-series GNN architecture (MTGNN) with a weighted cross-entropy loss function.

3. Investigated the interplay between self-reported survey symptoms (edema and fatigue) and objective physiological measures within the graph structure. Our findings indicate that heart rate variability (HRV) and certain low-intensity activity features are key factors associated with both symptoms, consistent with previous studies on pregnancy.

2. Related Work

Prior works have explored the use of wearable technologies to track patient health and provide valuable insights to healthcare providers. Runkle et al. (2019) conducted an observational cross-section study to evaluate how women of child-bearing age perceive the use of remote fetal electrocardiogram (ECG) monitoring technologies. The study found that participants were willing to use these devices during pregnancy, had no privacy concerns, and were comfortable sharing data with physicians. This positive response supports our research focus on meeting essential healthcare needs for pregnant women.

Souza et al. (2019) studied non-invasive tools to identify predictors of maternal complications, focusing on an actigraphy device to estimate physical activity and sleep-wake patterns among pregnant women. They found links between changes in these patterns and chronic conditions. While similar to the BUMP study, our research includes a broader range of wearable devices and survey data, offering a wider perspective on the pregnancy experience.

A similar study by Guo-Hung et al. (2021) developed a prediction system using lifestyle data, environmental factors, and patient symptoms to detect acute exacerbations of chronic obstructive pulmonary disease (AECOPD) in the upcoming 7 days. This study suggests that combining physiological measures with recent survey responses could enhance the prediction of pregnancy symptom severity.

3. Features and Cohort

3.1. Feature Choices

This study utilized datasets from the BUMP study (Goodday et al., 2022). Objective physiological data from the Oura ring were analyzed due to higher adherence rates, along with survey responses on edema and fatigue as subjective features. Three demographic features—age, number of previous births and prior pregnancies—were also included. Detailed feature information is in Appendix C.

Objective features The Gen2 Oura ring, a health-monitoring device worn on the finger, provided daily summaries of physiology signals such as heart rate, respiratory rate, movement, and skin temperature. These features, related to daytime activity and nighttime sleep quality, were used for forecasting fatigue and edema in third trimester of pregnancy.

Subjective features Participants completed daily symptom surveys via the BUMP study app, rating symptoms like edema and fatigue on a scale of 1 (no symptom) to 7 (severe). Fatigue was reported more frequently than edema. However, participants may have developed their own criteria for these responses, which could vary over time. To ensure the reliability of these responses, we compared them with a more detailed survey (see Section 3.2). Additional details on survey response distributions are in Appendix A.

3.2. Cohort Selection

Out of 498 participants, 265 met the criteria for inclusion in the edema survey analysis, and 59 were included for fatigue symptoms. We evaluated the alignment between self-reported edema and fatigue symptoms with rigorous edema and fatigue examinations (pitting edema and fatigue assessment) by measuring how correlated they are using Intraclass Correlation Coefficient (ICC) tests, a widely used reliability index for interrater reliability analysis. Please refer to Appendix E and Appendix F to see the details on pitting edema and fatigue assessments.

Edema survey selection criteria Participants were assigned pitting edema surveys (Seidel et al., 1995) during regular check-ins, conducted via video or phone every two weeks to a month. Participants were excluded from the edema symptom forecasting task if they showed inconsistencies in their self-reports, such as reporting no clinical edema in the pitting edema survey but rating daily edema above 2, or indicating clinical edema but no symptoms in the daily edema survey on the same day.

Fatigue survey selection criteria Fatigue Assessment Scale (FAS) (Shahid et al., 2012) were performed on everyone weekly via the BUMP study app, providing enough data points to perform an ICC test (Fisher, 1921). We considered the FAS score and the aggregated daily score as two raters and conducted

an ICC test using a two-way mixed model to assess each participant's consistency, as recommended by Koo and Li (2016). Participants with ICC values below 0.5, indicating poor consistency, were excluded from the fatigue symptom forecasting task.

To assess consistency between surveys before and after selecting reliable participants, ICC tests were conducted for all and for the selected group. These tests compared the pitting edema survey with the daily edema survey, and the FAS with the daily fatigue survey. Unlike the fatigue survey selection, which focused on individual consistency, this approach evaluated groups of participants. As shown in Table 1, the overall ICC was low for edema (0.27) and fatigue (0.13). However, after applying inclusion criteria, ICC improved (edema: 0.62, fatigue: 0.18). Notably, edema responses showed better alignment than fatigue responses, suggesting that fatigue survey data may be more subjective and noisy.

3.3. Data preprocessing

Since different subgroups were selected for the forecasting tasks, the experiments were conducted separately. Our forecasting experiments employed a live-update evaluation setting, in contrast to the conventional approach of utilizing the initial 60% of the entire participant time series for training and reserving the remaining portion for evaluation. The rolling basis training and evaluation reflects the evolving nature of data and models (You et al., 2022). Specifically, we conducted model training and evaluation on consecutive 28-day windows within the third trimester of pregnancy. In each window, the initial 19 days were allocated for training purposes, followed by 2 days for validation and 7 days for testing with an input sequence length of 5 and forecasting horizon of 1, as illustrated in Figure 1. Static features were broadcast across the time dimension.

Moreover, it's important to note that not every 28-day window was available for analysis due to missingness from periods when the wearable device was not used or surveys were not completed. To ensure data quality, we excluded windows with more than 3 consecutive days of missing data. We conducted linear searches in a greedy manner to identify 28-day windows that met the criteria. The remaining missing data was imputed using the Hyperimpute library (Jarrett et al., 2022). Since the third trimester spans 91 days, each participant could have up to three 28-day windows for analysis, with the first window aligning with the first month of the third trimester.

4. Methods

4.1. Models

We evaluated the performance of two symptom forecasting tasks using three different models: a vanilla LSTM, LTSF-NLinear (Zeng et al., 2022), and MTGNN model (Wu et al., 2020). The LTSF-NLinear is a simple linear baseline model, yet it outperforms some state-of-the-art transformers in long-term forecasting. The MTGNN model is notable for its graph learning layer, which captures the relationships between features and includes temporal convolution and graph convolution layers with skip connections.

Moreover, we modified the MTGNN model by altering its graph learning module. Instead of relying on random initialization, this modification utilized the last hidden state of an LSTM to generate embeddings from input time series, aiming to learn more dynamic embeddings. The modified graph learning layer is illustrated as follows:

$$\mathbf{E}_1 = \text{LSTM\_1}(\mathbf{X}) \tag{1}$$
$$\mathbf{E}_2 = \text{LSTM\_2}(\mathbf{X}) \tag{2}$$
$$\mathbf{M}_1 = \tanh(\alpha \mathbf{E}_1 \mathbf{\Theta}_1) \tag{3}$$
$$\mathbf{M}_2 = \tanh(\alpha \mathbf{E}_2 \mathbf{\Theta}_2) \tag{4}$$
$$\mathbf{A} = \text{ReLU}(\tanh(\alpha(\mathbf{M}_1\mathbf{M}_2^\intercal - \mathbf{M}_2\mathbf{M}_1^\intercal))) \tag{5}$$
$$\textit{for } i = 1, 2, \ldots, N \tag{6}$$
$$\mathbf{idx} = \textit{argtopk}(\mathbf{A}[i, :]) \tag{7}$$
$$\mathbf{A}[i, -\mathbf{idx}] = 0, \tag{8}$$

where $\mathbf{X}$ represents model input, $\mathbf{\Theta}_1$, $\mathbf{\Theta}_2$ are model parameters, α is the saturation rate hyperparameter, and *argtopk* returns the index of the top-k largest values of a vector. In the original graph learning layer, $\mathbf{E}_1$ and $\mathbf{E}_2$ are randomly initialized node embeddings. The modified MTGNN is denoted as MTGNN-LSTM. However, this approach did not yield the anticipated results, and our analysis focuses on understanding the reasons behind this outcome.

4.2. Loss functions

We experimented with 3 loss functions: cross-entropy (CE) loss ($\mathcal{L}_{CE}$), CE with change-aware weight ($\mathcal{L}_{CE\_weighted}$), CE with adversarial loss ($\mathcal{L}_{adv}$).

Cohort	ICC (Edema)	ICC (Fatigue)
All	0.27	0.13
Selected	0.62	0.18

Table 1: Intraclass correlation coefficient (ICC) before and after filtering participants for two comparisons: edema (daily symptom survey vs. pitting edema survey) and fatigue (daily fatigue survey vs. fatigue assessment scale). All p-values ≤ 0.1.

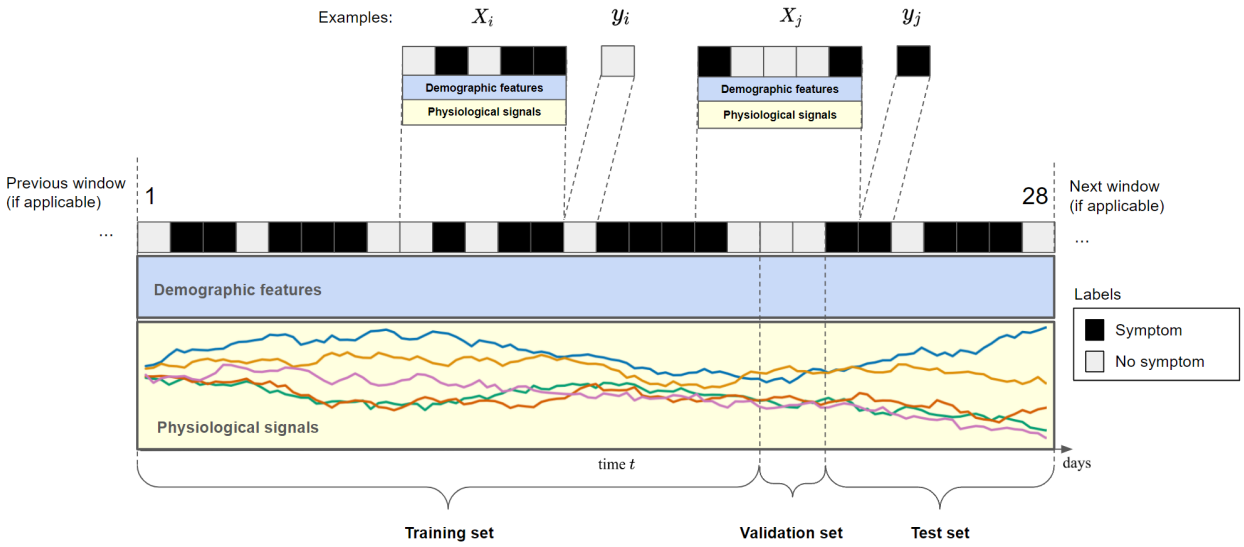


Figure 1: Illustration of training, validation, and test set split within one 28-day window for one participant. The initial 19 days were for training, followed by 2 days for validation and 7 days for testing. Each data point consisted of X and y, where y was the symptom label of the day and X was the combination of the previous 5 day's symptom survey responses, physiological signals, and demographics data of the participant. Demographic features included the age of the participant, the total number of times the participant became pregnant, and the number of prior pregnancies that resulted in a live birth.

CE The baseline loss function used in the study was the cross-entropy (CE) loss between the forecasted outcome and the expected outcome of the binarized symptom label. The weights were adjusted to account for class imbalance, with weights of 0.15 and 0.85 for cases with edema and cases without edema respectively, while for the symptom of fatigue, the weights were kept equal.

CE with change-aware weight CE with change-aware weight was designed to highlight changes in symptom status. An extra weight factor γ, where $\gamma > 1$, on top of the class weights, was applied to data points where the target label differs from the previous day, as illustrated in Figure 2.

CE with adversarial loss To mitigate the inter-participants difference and help MTGNN to learn a more generalized graph structure, the domain-adversarial training technique proposed by Ganin et al. (2015) was applied for inter-participants domain adaptation. The second-to-last layer of the MTGNN model was extracted and input into a discriminator model D, which was trained to classify participants' identities. The discriminator minimized a CE loss between the predicted participants' IDs and the actual IDs for 20 epochs within each training epoch of the

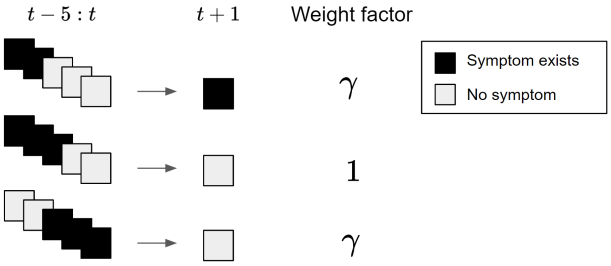


Figure 2: Illustration of assigning a higher weight $\gamma > 1$ to data points where the target label at $t+1$ differs from the label at t.

MTGNN model. The parameters of the discriminator were re-initialized once every 5 training epochs of the MTGNN model training to stabilize the gradients. The CE loss that was minimized by the discriminator is denoted as:

$$\mathcal{L}_d = \text{CE}(D(M_{2nd\_to\_last\_layer}(\boldsymbol{X}); \theta_D), \boldsymbol{Y}_{\text{ID}}) \quad (9)$$

where $M_{2nd\_to\_last\_layer}$ is the second-to-last layer of the MTGNN model, and $\boldsymbol{Y}_{\text{ID}}$ are the ID of each participants. Training model against the discriminator under domain-adversarial training yielded the loss function below:

$$\mathcal{L}_{adv} = \mathcal{L}_{CE} - \beta \mathcal{L}_d \quad (10)$$

where β is the weight of the adversarial loss.

In this research, we have selectively employed a series of GNN model and loss function combinations to improve the forecasting performance, including the LSTM model, the LTSF-NLinear, and the MTGNN model with standard CE loss. Additionally, we explored the MTGNN model with an integrated LSTM (MTGNN-LSTM) with the same standard CE loss. We also experimented with the MTGNN model using the change-aware weighted CE loss ($\mathcal{L}_{CE\_weighted}$), and MTGNN model with a combination of standard CE and adversarial loss ($\mathcal{L}_{adv}$). The modified CE loss and the adversarial loss were not combined together due to the increased complexity in hyperparameter tuning. We used a saturation rate $\alpha = 3$ as the original MTGNN model(Wu et al., 2020). Hyperparameter γ and β were used MTGNN with change-aware weight and MTGNN with adversarial loss, respectively. They were tuned separately using the validation set from a random selection range from 0.001 to 1. In short, the MTGNN model with a weighted CE loss function proved to be the most effective.

5. Results

5.1. Overview

Overall, we forecasted next-day pregnancy symptoms from the previous 5 days, specifically edema and fatigue, using various baseline models and different modifications of the MTGNN model. We divided the third trimester (the last three months of pregnancy) into three 28-day windows (approximately one month each), with each period designated for separate training and evaluation. We conducted 5 repetitions of the experiment with different initialization parameters. We reported the AUC scores of these models averaged over the five runs for each window, and illustrated the relationships between the symptoms and physiological features using a heatmap.

5.2. Forecasting results

We present the performance of various models across the three months, where each month was represented by a 28-day window, averaged over 5 runs. The AUC scores for edema and fatigue symptoms are shown in Figure 3. In general, all MTGNN-based models performed better than the RNN-based and linear models (LSTM and LTSF-NLinear). The LSTM and LTSF-NLinear models perform relatively poorly, with an AUC score hovering around the 0.5 mark, which is indicative of random guessing. In addition, the MTGNN with a weighted CE loss $\mathcal{L}_{CE\_weighted}$ (MTGNN_weighted_CE) consistently shows more robust performance across the last three months of pregnancy. In contrast, the MTGNN with the adversarial loss $\mathcal{L}_{adv}$ (MTGNN_adv) exhibits fluctuating results, performing less effectively during certain periods. Notably, the MTGNN model with LSTM integration (MTGNN-LSTM) did not show consistent improvement either, but we did see a significant improvement in the last month for the fatigue symptom, suggesting its effectiveness may be context-dependent. Overall, MTGNN_weighted_CE stood out as the most stable and effective model for predicting the symptoms considered in this study. A more in-depth analysis of the strengths and weaknesses of the various models can be found in Section 6.2.

5.3. Learned graph structure of the GNN

We delved into the graph structure learned by the MTGNN models trained using different loss functions and different windows. We focused on exploring

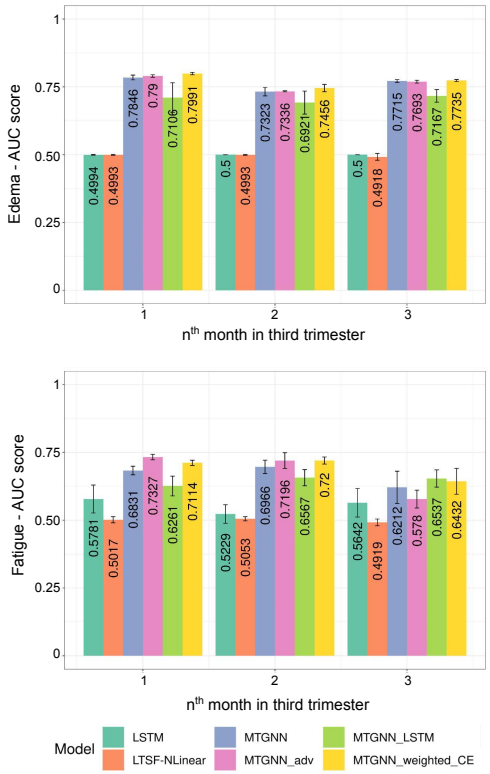


Figure 3: Forecasting results

the relationship between physiological measures and self-reported symptoms to identify which features are most strongly associated with each symptom. Understanding these relationships may provide clinical insights, highlighting which features to investigate further for specific symptoms. In the graph structure developed by MTGNN, we considered the directed edges from the objective physiological measures to self-reported symptoms as the inflow. This indicates how these features contribute to predicting the next day's symptoms. Since MTGNN does not involve causality, the directed edges were not considered as causal relationships. The heatmap in Figure 4 illustrates the strength of the directed edges, averaged over 5 runs, acquired by the MTGNN_weighted_CE model. The connections represent the information flow, with the colour of each link matching the colour of its source node. In summary, we found that:

- Sleep features: In both prediction tasks, heart rate variability (rmssd) was the most important sleep-related feature. For fatigue, the total restless time and the lowest heart rate value at night are moderately important.
- Activity features: The number of steps and time of low-intensity activity were also relatively important features for predicting both symptoms. These features were more important in edema than fatigue forecasting.
- Demographics features: were key predictors of edema but less important for fatigue.

Overall, we have found that the MTGNN_weighted_CE model was the most stable and effective model. Heart rate variability, number of steps, and time of low-intensity activity were important factors for predicting edema and fatigue.

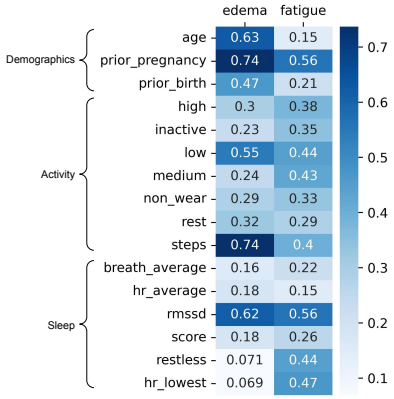


Figure 4: Overall inflow edge weights from objective features into the subjective symptoms of the graph structure learned by MTGNN_weighted_CE, averaged over all windows and all runs.

6. Discussion

Our paper is the first to use wearable data with MTGNN to predict next-day pregnancy symptoms, specifically edema and fatigue. We investigated several MTGNN architectures such as incorporating an LSTM into the graph learning module, implementing data point-specific weighted loss function, and employing domain adaptation through adversarial loss. These modifications were designed to mitigate inter-participant discrepancies caused by different biases such as how participants interpreted the survey questions. Empirically, we discovered that integrating weighted Cross Entropy loss into the MTGNN model leads to a more accurate representation of daily changes in pregnancy symptoms, resulting in robust performance. Moreover, we examined the adjacency matrices learned by the MTGNN models, exploring

the intricate relationships between self-reported survey responses and physiological measures.

6.1. Fatigue is more subjective than edema

Through intraclass correlation analysis, we determined that self-reported fatigue is more subjective than self-reported edema. The analysis shows that, among all the participants, the correlation between self-reported edema and clinical pitting edema assessments was stronger than the correlation between self-reported fatigue and fatigue assessment tests. In each selected cohort, as detailed in Table 1, there was moderate agreement (ICC=0.6208) for edema responses, whereas the agreement for fatigue was significantly lower (ICC=0.1810) among the selected cohorts. This discrepancy may be because edema being more associated to physical conditions, allows participants to more easily agree on the criteria than fatigue does. Regardless of whether participants provided true responses to the daily fatigue survey, there was still low agreement among them.

Given that fatigue is more subjective than edema, the self-reported fatigue responses are expected to be noisier than those for edema. This may partly explain why predicting edema gives more accurate results than predicting fatigue. Moreover, since fatigue is closely tied to the participant's mental state, it may be influenced by external factors not captured by the physiological measures used in this BUMP study. Therefore, despite the edema surveys appearing more unbalanced than those for fatigue, they still achieved a higher AUC score than fatigue.

6.2. GNN model performance and participant heterogeneity

We explored how various models, namely the original MTGNN, MTGNN-LSTM, MTGNN_adv, and MTGNN_weighted_CE, influence the predictive results across different third-trimester windows. Our observations suggest that the degree of heterogeneity among participants may be crucial. Specifically, we note that edema, being less subjective and thus having less heterogeneity among participants than fatigue, gains smaller benefits from performance improvement. This suggests that when participant responses are more consistent and less noisy, the benefits of implementing sophisticated loss functions or model architectures might be marginal. In contrast, with labels that are more subjective and noisy, complex loss function may achieve a better performance.

The MTGNN-LSTM and MTGNN_adv models modify the original MTGNN to handle data heterogeneity differently. The MTGNN-LSTM model customizes the graph structure for each data point, while the adversarial loss in MTGNN_adv aims to make feature embeddings more homogenized. Given that MTGNN_adv generally outperforms MTGNN-LSTM in most scenarios, it suggests that MTGNN-LSTM might be more prone to overfitting, whereas MTGNN_adv appears to provide better generalization across different situations. However, in the final month of pregnancy, MTGNN-LSTM excelled, likely due to the increased diversity of experiences during this period (as shown in Figure 4).

The MTGNN_weighted_CE model, which assigns additional weight to data points with label changes from the previous day, achieved the best performance. This suggests that focusing on significant individual pattern changes, rather than accommodating cross-individual heterogeneity, may enhance generalization.

6.3. Exploring insights from feature usefulness

Several insights on the learned graph structures align with findings from existing pregnancy research. We noticed that nighttime heart rate variability (HRV) emerged as a significant variable for both edema and fatigue symptoms. Typically, HRV tends to decrease as pregnancy proceeds (Sarhaddi et al., 2022), which serves as a marker of progression in pregnancy. Furthermore, emotional stress and anxiety about the impending birth can also affect HRV (Kim et al., 2018), indicating a heightened stress response which may affect edema and fatigue symptoms. More sleep-related features show greater edge weight towards forecasting fatigue symptoms such as restlessness. The total amount of restless sleep time is an important indicator of sleep quality. This finding matches the study conducted by Effati-Daryani et al. (2021), which indicates a negative correlation between fatigue and sleep quality. In other words, poorer sleep quality is associated with a higher likelihood of experiencing fatigue the next day.

Daily activity also demonstrates an association with both symptoms. Specifically, step count, a key metric of daily activity, has proven to be an important prediction factor for edema, as shown in Figure 4. This finding aligns with previous research suggesting that combining compression therapy with low-intensity exercise such as walking can help prevent

and treat lower limb edema (Ochalek et al., 2017). This also aligns with another relatively important feature which represents the total amount of time spent on the low-intensity activity.

Age and the number of previous pregnancies are notable factors in predicting edema, showing a higher edge weight in the learned graph structure. However, a direct interdependence between the frequency of pregnancies and the onset of edema has not been established in existing research. When interpreting the adjacency graph produced by the models, it is tempting to derive deeper insights about pregnancy from the dataset. We must be cautious to avoid over-interpreting the models' outputs, which may lead to unwarranted assumptions.

6.4. Limitations

The focus of this work was on short-term forecasting, specifically predicting only the next day's symptoms. Future studies could extend the forecasting window to cover a longer term. Researchers might also analyze how seasonal patterns and trends in physiological signals impact symptoms. However, using inaccurate imputation methods on long-term data may distort the original signals and impair forecasting results. Methods that manage both imputation and forecasting, such as those proposed by (Alcaraz and Strodthoff, 2023), could be considered. Moreover, this study faces limitations due to the lack of access to raw data; all physiological measures were processed using wearables manufacturers' backend algorithms, which prevents analyses of physiological signals at a more fine-grained level. The tailored loss improves generalization within our dataset but may not fully address variability in entirely new patient populations. Future work will focus on enhancing model robustness for unseen patients, potentially through more diverse training data or transfer learning techniques. Additionally, caution must be exercised in interpreting results due to the inherent heterogeneity in the study population, as individual differences in physiology and symptom experience are expected. Despite these limitations, the research establishes a connection between physiological measures and self-reported symptoms by setting up predictive models. We hope that being able to predict pregnancy symptoms could help develop interventions that could alleviate symptoms and enhance the pregnancy experience in the future.

7. Conclusion

Our research demonstrates the feasibility of predicting the following day's pregnancy symptoms, edema and fatigue, using physiological signals from wearables and symptoms data. We found that GNN with weighted CE loss achieved the best overall results. We also explored the relationships between symptoms and physiological features through heatmaps derived from the GNN's learned graph structure. Our findings highlight that HRV plays an important role in predicting both symptoms. Additionally, certain low-intensity activity features significantly influence fatigue but have less impact on edema.

References

Juan Miguel Lopez Alcaraz and Nils Strodthoff. Diffusion-based time series imputation and forecasting with structured state space models, 2023.

Jolanda de Vries, Helen Michielsen, Guus L. Van Heck, and Marjolein Drent. Measuring fatigue in sarcoidosis: The fatigue assessment scale (fas). *British Journal of Health Psychology*, 9(3):279–291, Sep 2004. doi: 10.1348/1359107041557048.

Fatemeh Effati-Daryani, Sakineh Mohammad-Alizadeh-Charandabi, Azam Mohammadi, Somayeh Zarei, and Mojgan Mirghafourvand. Fatigue and sleep quality in different trimesters of pregnancy. *Sleep Sci*, 14(Spec 1):69–74, January 2021.

Ronald Aylmer Fisher. On the "Probable Error" of a Coefficient of Correlation Deduced from a Small Sample. *Metron*, 1:3–32, 1921.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *Advances in Computer Vision and Pattern Recognition*, 17:189–209, 5 2015. ISSN 21916594. doi: 10.1007/978-3-319-58347-1_10. URL https://arxiv.org/abs/1505.07818v4.

Giorgia Gon, Andreia Leite, Clara Calvert, Susannah Woodd, Wendy J. Graham, and Veronique Filippi. The frequency of maternal morbidity: A systematic review of systematic reviews. *International Journal of Gynecology Obstetrics*, 141:20–38, 5 2018. ISSN 1879-3479. doi: 10.1002/IJGO.12468. URL https://onlinelibrary.wiley.com/doi/full/10.1002/ijgo.12468https://onlinelibrary.wiley.com/doi/abs/10.1002/ijgo.12468https://obgyn.onlinelibrary.wiley.com/doi/10.1002/ijgo.12468.

S M Goodday, E Karlin, A Brooks, C Chapman, D R Karlin, L Foschini, E Kipping, M Wildman, M Francis, H Greenman, Li Li, E Schadt, M Ghassemi, A Goldenberg, F Cormack, N Taptiklis, C Centen, S Smith, and S Friend. Better understanding of the metamorphosis of pregnancy (bump): protocol for a digital feasibility study in women from preconception to postpartum. *npj Digital Medicine*, 5:40, 2022. ISSN 2398-6352. doi: 10.1038/s41746-022-00579-9. URL https://doi.org/10.1038/s41746-022-00579-9.

Guo-Hung, Huang Chun-Ta, Cheng Yu-Chieh, Chen Chi-Hsien, Chien Jung-Yien, Kuo Ping-Hung, Kuo Lu-Cheng, Lai Feipei Wu Chia-Tung, and Li. Acute exacerbation of a chronic obstructive pulmonary disease prediction system using wearable device data, machine learning, and deep learning: Development and cohort study. *JMIR Mhealth Uhealth*, 9:e22591, 5 2021. ISSN 2291-5222. doi: 10.2196/22591. URL http://www.ncbi.nlm.nih.gov/pubmed/33955840.

Daniel Jarrett, Bogdan Cebere, Tennison Liu, Alicia Curth, and Mihaela van der Schaar. Hyperimpute: Generalized iterative imputation with automatic model selection. 6 2022. doi: 10.48550/arxiv.2206.07769. URL https://arxiv.org/abs/2206.07769v1.

Hye-Geum Kim, Eun-Jin Cheon, Dai-Seg Bai, Young Hwan Lee, and Bon-Hoon Koo. Stress and heart rate variability: A Meta-Analysis and review of the literature. *Psychiatry Investig*, 15(3):235–245, February 2018.

Terry K Koo and Mae Y Li. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *J Chiropr Med*, 15(2):155–163, March 2016.

Patrick S. Moran, Francesca Wuytack, Michael Turner, Charles Normand, Stephanie Brown, Cecily Begley, and Deirdre Daly. Economic burden of maternal morbidity – a systematic review of cost-of-illness studies. *PLoS ONE*, 15, 1 2020. ISSN 19326203. doi: 10.1371/JOURNAL.PONE.0227377. URL /pmc/articles/PMC6964978//pmc/articles/PMC6964978/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC6964978/.

Katarzyna Ochalek, Katarzyna Pacyga, Marta Curyło, Aleksandra Frydrych-Szymonik, and Zbigniew Szygula. Risk factors related to lower limb edema, compression, and physical activity during pregnancy: A retrospective study. *Lymphat Res Biol*, 15(2):166–171, March 2017.

Jennifer Runkle, Maggie Sugg, Danielle Boase, Shelley L. Galvin, and Carol C. Coulson. Use of wearable sensors for pregnancy health and environmental monitoring: Descriptive

findings from the perspective of patients and providers. *Digital health*, 5, 2 2019. ISSN 2055-2076. doi: 10.1177/2055207619828220. URL https://pubmed-ncbi-nlm-nih-gov.myaccess.library.utoronto.ca/30792878/.

Monika Sanghavi and John D. Rutherford. Cardiovascular physiology of pregnancy. *Circulation*, 130(12):1003–1008, Sep 2014. doi: 10.1161/circulationaha.114.009029.

Fatemeh Sarhaddi, Iman Azimi, Anna Axelin, Hannakaisa Niela-Vilen, Pasi Liljeberg, and Amir M Rahmani. Trends in heart rate and heart rate variability during pregnancy and the 3-month postpartum period: Continuous monitoring in a free-living context. *JMIR Mhealth Uhealth*, 10(6):e33458, June 2022.

HM Seidel, JW Ball, JE Dains, and GW Benedict. Heart and blood vessels. *Mosby's Guide to Physical Examination, 3rd ed. St. Louis, MO: Mosby*, 419, 1995.

Azmeh Shahid, Kate Wilkinson, Shai Marcu, and Colin M. Shapiro. Fatigue Assessment Scale (FAS). In Azmeh Shahid, Kate Wilkinson, Shai Marcu, and Colin M Shapiro, editors, *STOP, THAT and One Hundred Other Sleep Scales*, pages 161–162. Springer New York, New York, NY, 2012. ISBN 978-1-4419-9893-4. doi: 10.1007/978-1-4419-9893-4_33. URL https://doi.org/10.1007/978-1-4419-9893-4_33.

Joakim Skarding, Bogdan Gabrys, and Katarzyna Musial. Foundations and modelling of dynamic networks using dynamic graph neural networks: A survey. *CoRR*, abs/2005.07496, 2020. URL https://arxiv.org/abs/2005.07496.

Rebecca M D Smyth, Nasreen Aflaifel, and Anthony A Bamigboye. Interventions for varicose veins and leg oedema in pregnancy. *Cochrane Database Syst Rev*, 2015(10):CD001066, October 2015.

Renato T. Souza, Jose Guilherme Cecatti, Jussara Mayrink, Rafael Bessa Galvão, Maria Laura Costa, Francisco Feitosa, Edilberto Rocha Filho, Debora F. Leite, Janete Vettorazzi, Ricardo P. Tedesco, Danielly S. Santana, Joao Paulo Souza, Carina B. Luiz, Luiza C. Brust, Danilo Anacleto, Lívia C. Nascimento, Daisy Lucena, Denise Ellen F. Cordeiro, and Mariana B. Rogerio. Identification of earlier predictors of pregnancy complications through wearable technologies in a brazilian multicentre cohort: Maternal actigraphy exploratory study i (maes-i) study protocol. *BMJ open*, 9, 4 2019. ISSN 2044-6055. doi: 10.1136/BMJOPEN-2018-023101. URL https://pubmed-ncbi-nlm-nih-gov.myaccess.library.utoronto.ca/31005906/.

Jing-Jing Wan, Zhen Qin, Peng-Yuan Wang, Yang Sun, and Xia Liu. Muscle fatigue: general understanding and treatment. *Exp Mol Med*, 49(10):e384, October 2017.

Zonghan Wu, Shirui Pan, Guodong Long, Jing Jiang, Xiaojun Chang, and Chengqi Zhang. Connecting the dots: Multivariate time series forecasting with graph neural networks. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 20:753–763, 5 2020. doi: 10.48550/arxiv.2005.11650. URL https://arxiv.org/abs/2005.11650v1.

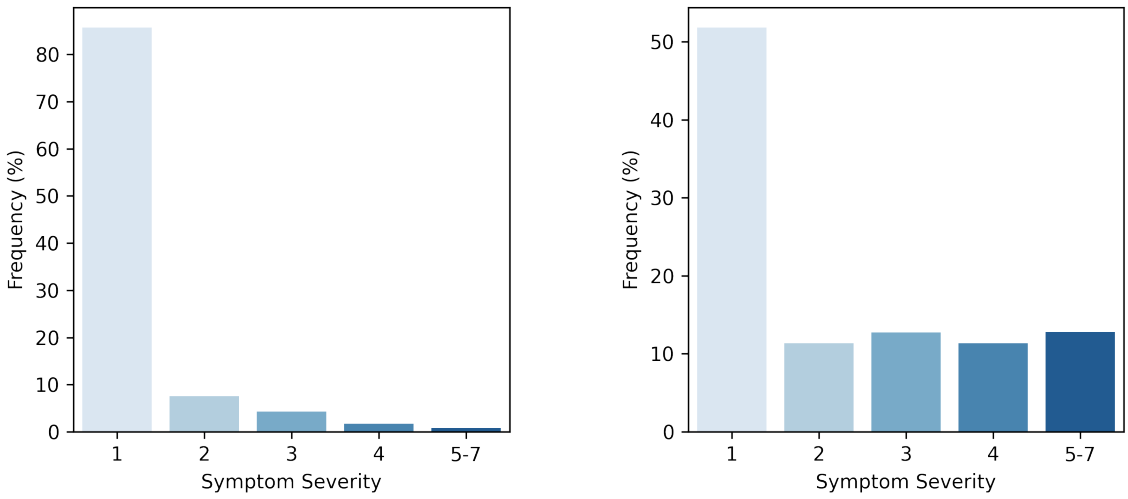
Steven H. Yale and Joseph J. Mazza. Approach to diagnosing lower extremity edema. *Comprehensive Therapy*, 27(3):242–252, Sep 2001. doi: 10.1007/s12019-001-0021-5.

Jiaxuan You, Tianyu Du, and Jure Leskovec. Roland: Graph learning framework for dynamic graphs, 2022.

Ailing Zeng, Muxi Chen, Lei Zhang, and Qiang Xu. Are transformers effective for time series forecasting?, 2022.

Yuecai Zhu, Fuyuan Lyu, Chengming Hu, Xi Chen, and Xue Liu. Encoder-decoder architecture for supervised dynamic graph learning: A survey, 2022.

Appendix A. Symptom distribution



(*a*) Edema responses (*b*) Fatigue respones

Figure 5: Percentage of severity for each survey response, where 1 indicates no symptom and 7 indicates very severe symptom.

Appendix B. Cohort Overview

Demographics	Edema (N=250)	Fatigue (N=59)	Total (N=563)
Age (years)			
Mean (SD)	33.7 (3.58)	33.1 (3.83)	33.4 (3.90)
Median (IQR)	34 (31-36)	33.5 (31-36)	34 (31-36)
Prior pregnancy, n (%)			
0	158 (61.72%)	42 (71.19%)	346 (61.35%)
1-3	61 (23.83%)	10 (16.95%)	141 (25%)
3-5	29 (11.33%)	5 (8.47%)	60 (10.64%)
≥ 5	8 (3.13%)	2 (3.39%)	17 (3.01%)
Prefer not to say	0 (0%)	0 (0%)	0 (0%)
Prior birth, n (%)			
0	192 (75%)	50 (84.75%)	438 (77.66%)
1	43 (16.80%)	5 (8.47%)	85 (15.07%)
2	16 (6.25%)	4 (6.78%)	30 (5.32%)
≥ 3	5 (1.95%)	0 (0%)	10 (1.77%)
Prefer not to say	0 (0%)	0 (0%)	1 (0.18%)
Education, n (%)			
Doctorate degree	16 (6.25%)	3 (5.08%)	34 (5.98%)
Professional degree	25 (9.77%)	4 (6.78%)	46 (8.08%)
Master's degree	95 (37.11%)	20 (33.90%)	191 (33.57%)
Bachelor's degree	87 (34.99%)	23 (38.98%)	208 (36.56%)
Associate's degree	11 (4.30%)	3 (5.08%)	34 (5.98%)
Non-degree college	11 (4.30%)	3 (5.08%)	38 (6.68%)
High school diploma	4 (1.56%)	3 (5.08%)	9 (1.58%)
Did not complete high school	1 (0.39%)	0 (0%)	3 (0.53%)
Prefer not to say	6 (2.34%)	0 (0%)	6 (1.05%)
Race, n (%)			
White	109 (42.58)	34 (57.63%)	240 (42.18%)
Black	7 (2.73%)	0 (0%)	20 (3.52%)
Asian	17 (6.64%)	1 (1.69%)	38 (6.68%)
Other	0 (0%)	0 (0%)	3 (0.53%)
Perfer not to say	123 (48.04%)	24 (40.68%)	268 (47.10%)

Table 2: Patient Demographics

Appendix C. Feature details

The descriptions of the features are in Table 3.

Variable name	Source	Explanation
high	Activity	Number of minutes during the day with high-intensity activity
inactive	Activity	Number of inactive minutes during the day
low	Activity	Number of minutes during the day with low-intensity activity
medium	Activity	Number of minutes during the day with medium-intensity activity
non_wear	Activity	Number of minutes during the day when the participant was not wearing the ring
rest	Activity	Number of minutes during the day spent resting i.e. sleeping or lying down
steps	Activity	The total number of steps during the day
breath_average	Sleep	The average respiratory rate during the sleep period at night
hr_average	Sleep	The average heart rate during the sleep period at night
rMSSD	Sleep	The average heart rate variability calculated with rMSSD method during the sleep period at night
score	Sleep	A sleep score representing the overall sleep quality during the sleep period
restless	Sleep	The restlessness during the sleep period, i.e. the percentage of sleep time when the participant was moving
hr_lowest	Sleep	The lowest heart rate (5 minutes sliding average) during the sleep period
prior_birth	Demographic	Number of prior pregnancies that resulted in a live birth
prior_pregnancy	Demographic	Total number of times becoming pregnant
age	Demographic	Age
Edema	Survey	The severity on a scale of 1-7 on the symptom of swelling in the body, particularly the hands and/or feet
Fatigue	Survey	The severity on a scale of 1-7 on the symptom of feeling fatigued or easily tired

Table 3: The details of features.

Appendix D. All inflow weights

Inflow edge weights from objective features into the fatigue of window 0 learned by all MTGNN-based models.

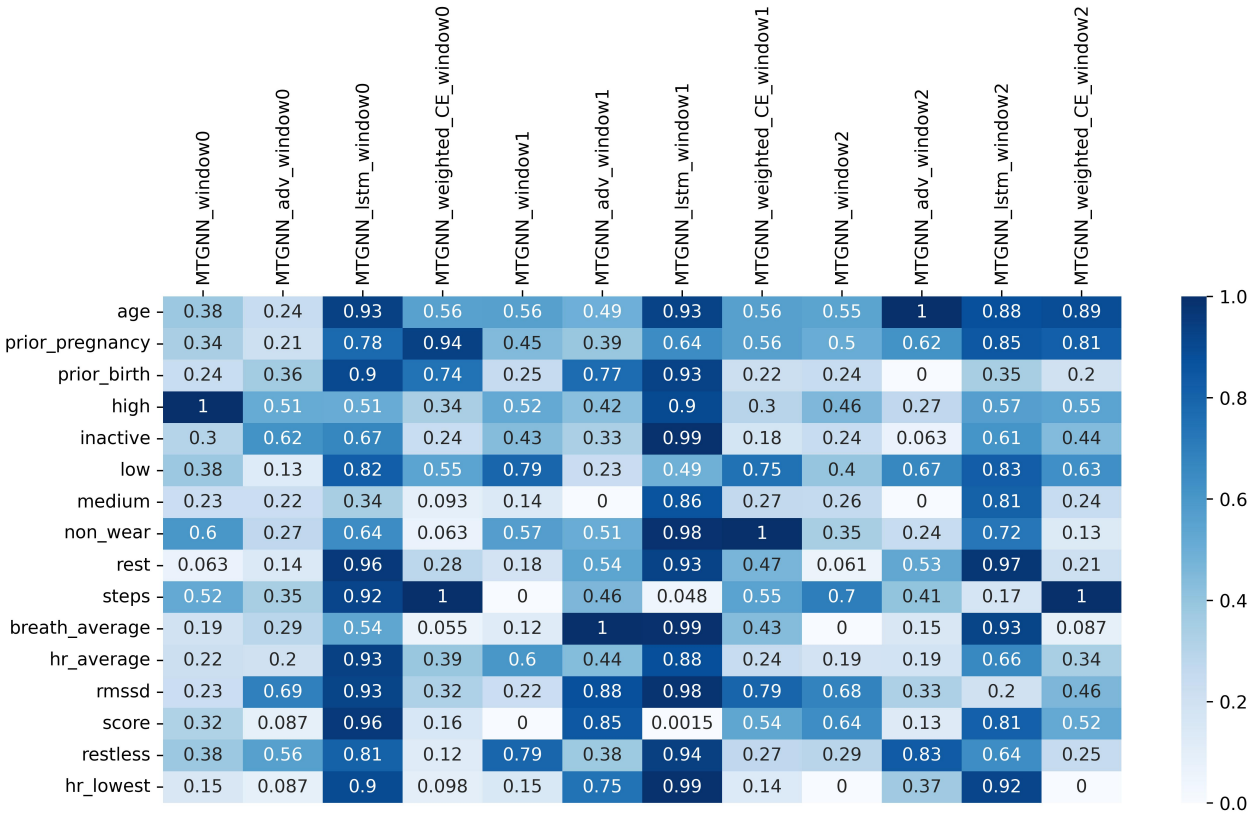


Figure 6: Inflow edge weights from objective features into the edema of all windows learned by all MTGNN-based models. The models were ordered based on the order shown in Figure 3.

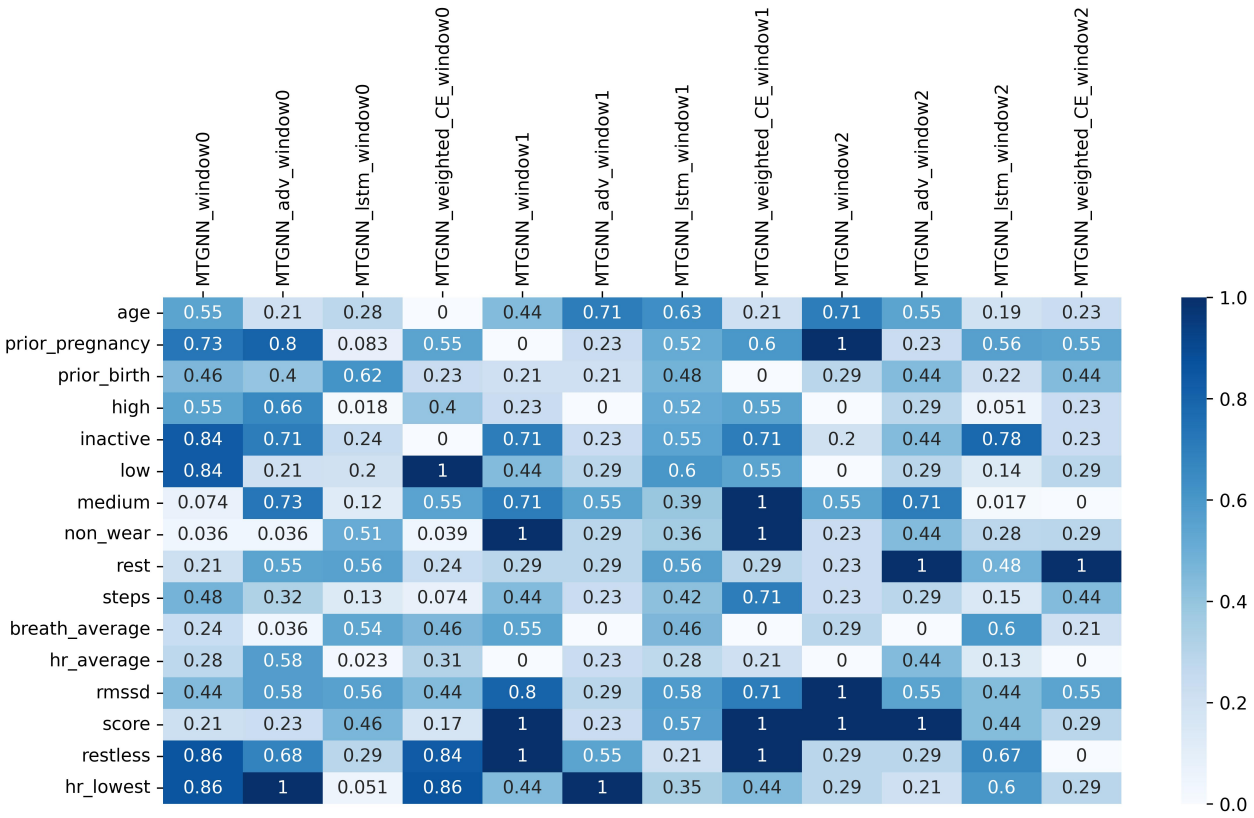


Figure 7: Inflow edge weights from objective features into the fatigue of all windows learned by all MTGNN-based models. The models were ordered based on the order shown in Figure 3.

Appendix E. Pitting edema assessment

Pitting Edema Assessment is a measurement taken by pushing on a person's skin, measuring the depth of the indention, and record how long it takes for the skin to rebound back to its original position. This measurement will be taken at each study visit. The pitting edema assessment question is shown below in the table below.

Perform the pitting edema assessment and provide score below.	Scoring System
No clinical edema	0
$\leq$ 2 mm indentation	1
2-4 mm indentation	2
4-6 mm indentation	3
6-8 mm indentation	4

Table 4: Pitting edema assessment survey (Yale and Mazza, 2001)

Appendix F. Fatigue assessment

The Fatigue Assessment Scale (FAS) is a 10-item self-report scale evaluating symptoms of chronic fatigue. The FAS survey is shown below:

You are about to start a survey that asks about how much energy you've had over the past week. It is 10 questions and will take about 1 minute to complete.

Questions	Answer choices
During the past week, did you experience any fatigue (feeling physically or mentally exhausted)?	1=no, 2 =yes
During the past week, I have found that:	never, sometimes, regularly, often, always
I was bothered by fatigue	never, sometimes, regularly, often, always
I got tired very quickly	never, sometimes, regularly, often, always
I didn't do much during the day	never, sometimes, regularly, often, always
I had enough energy for everyday life	never, sometimes, regularly, often, always
Physically, I felt exhausted	never, sometimes, regularly, often, always
I had problems to start things	never, sometimes, regularly, often, always
I had problems to think clearly	never, sometimes, regularly, often, always
I felt no desire to do anything	never, sometimes, regularly, often, always
Mentally, I felt exhausted	never, sometimes, regularly, often, always
When I was doing something, I could concentrate quite well	never, sometimes, regularly, often, always

Table 5: Fatigue assessment survey (de Vries et al., 2004)