Data Mining Project

Project Results Presentation

A multi-modal sensor dataset for continuous stress detection in nurses at a hospital

26 March, 2025

🔬BACKGROUND OF THE STUDY

- The data contains multimodal physiological signals gathered from a wearable device attached to the regular working nurses.
- A self-reported survey results completed daily by the nurses after their regular shift.

❓ Research Questions

- **How do physiological signals correlate with self-reported stress levels among nurses during the COVID-19 pandemic?**
- What are the primary contextual factors(e.g., COVID-related challenges, workload, and patient crises) that contribute to stress among nurses, and how are these factors reflected in physiological data?

🧠 What We Did?

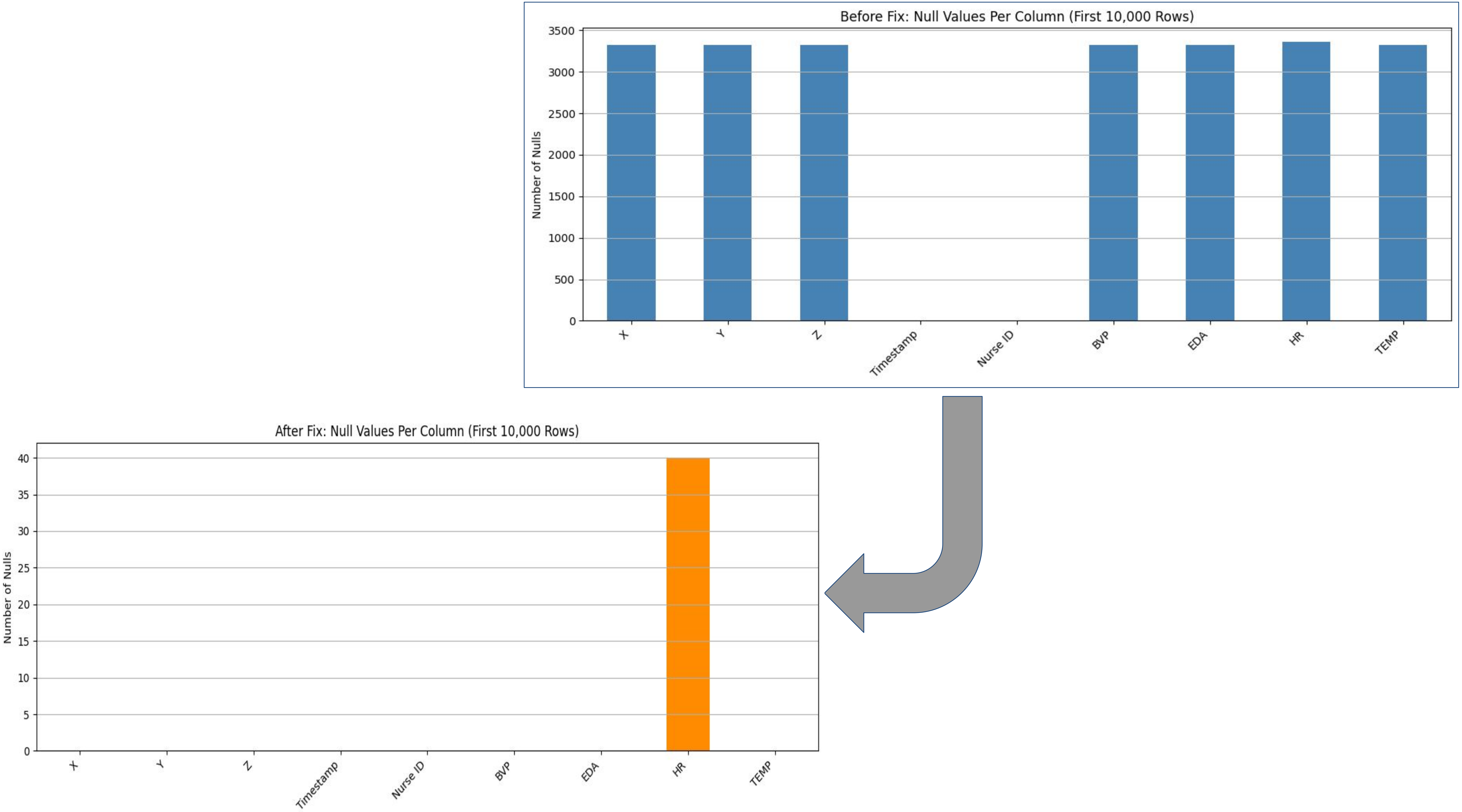
1. **Checked the Final Unified File for inconsistencies:**

   → Found duplicate rows caused by slight differences in timestamp precision (e.g., 1586788323.75 vs 1586788323.7500002)

   → One row had only HR, the other had missing HR but other values

   ✅ Solution: Rounded timestamps to 2 decimal places to normalize and fix merging issues

📊 The visual comparison of the data before and after fixing the issue ➡️ (Next slides)

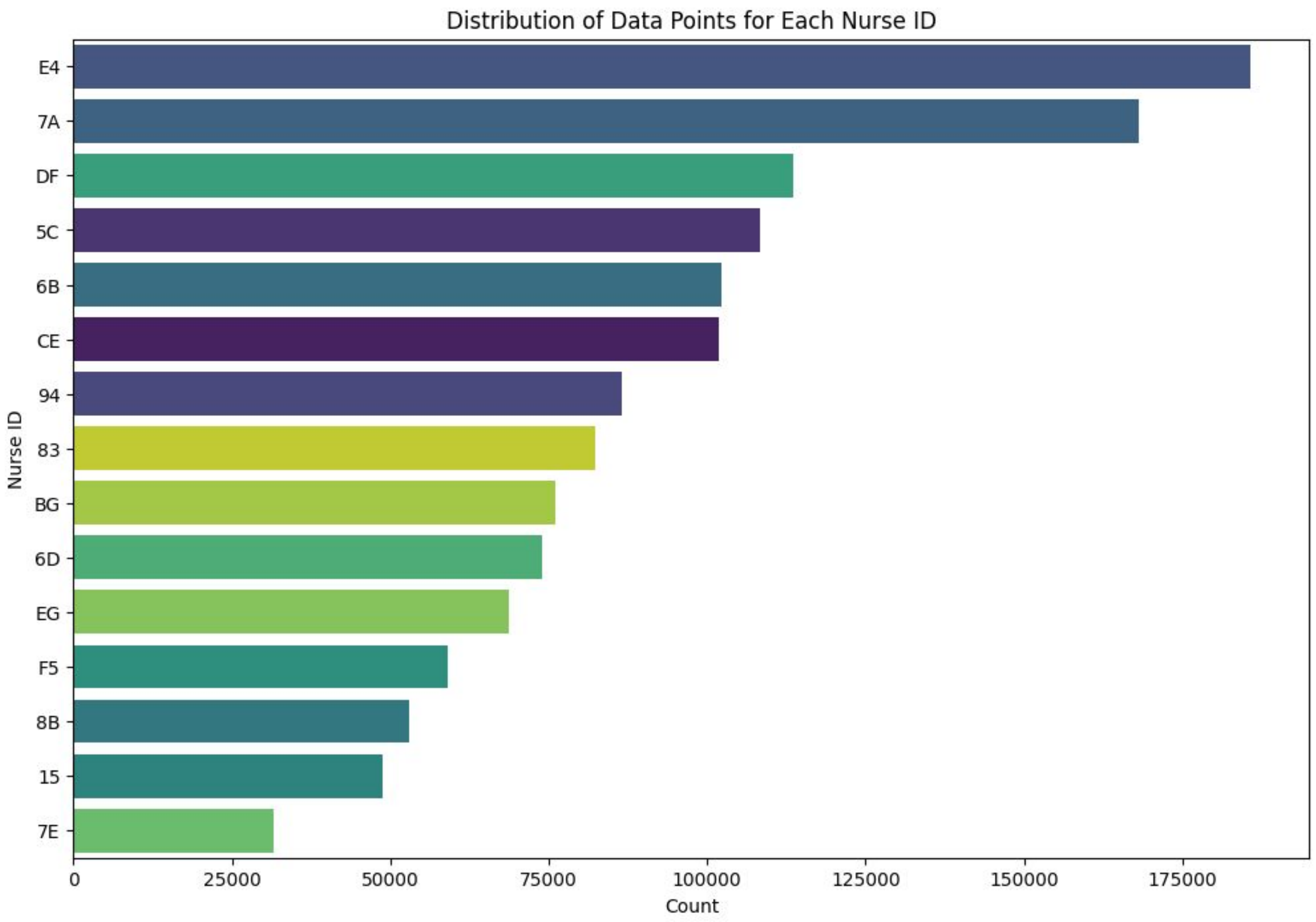

Before Fix: Null Values Per Column (First 10,000 Rows)
Number of Nulls
0
500
1000
1500
2000
2500
3000
3500
X
Y
Z
Timestamp
Nurse ID
BVP
EDA
HR
TEMP
After Fix: Null Values Per Column (First 10,000 Rows)
Number of Nulls
0
5
10
15
20
25
30
35
40
X
Y
Z
Timestamp
Nurse ID
BVP
EDA
HR
TEMP

🧠 What We Did?

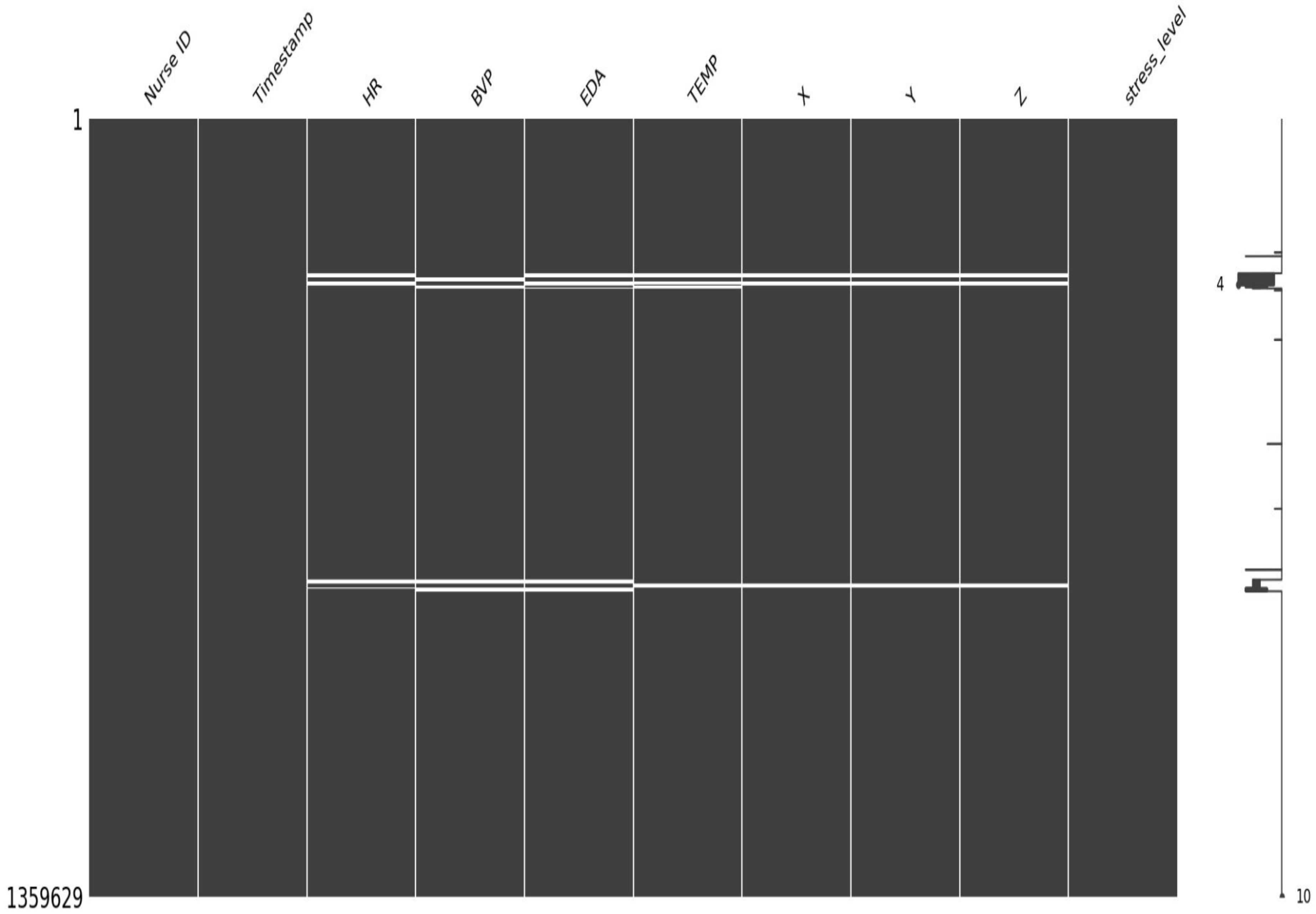
2. Align Physiological Data with Stress Events (Data Reduction):

- Merging the preprocessed survey results with the resampled physiological signal dataset 🧪📈
- Condition: Filter signals within the Start → End time window of each reported stress event 🕒

id	stress_level	covid_related	patient_in_crisis	description	start_timestamp	end_timestamp
6B	0	0.000	0.000	NaN	1592927160.000	1592928720.000
6B	2	0.000	1.000	NaN	1592929500.000	1592930940.000
7A	0	0.000	0.000	Walking across the street to get to important ...	1592933760.000	1592934600.000
5C	1	0.000	0.000	NaN	1592934720.000	1592935740.000
5C	0	0.000	0.000	Employee having personal crisis.	1592936520.000	1592946180.000
7A	1	0.000	0.000	Storming outside and needing to go to meeting	1592937600.000	1592938620.000
7A	1	0.000	0.000	Meeting	1592937600.000	1592938620.000
94	1	0.000	0.000	NaN	1592938200.000	1592939760.000
7A	2	0.000	0.000	In a meeting working on processes to streamlin...	1592942880.000	1592943300.000
6B	2	0.000	0.000	NaN	1592999520.000	1593000240.000



Without removing nulls



Identified null entries(HR, BVP,EDA and TEMP) after merge

🧠 What We Did?

3. Integrate Physiological Signals with self-reported Stress Levels to form a Complete Dataset:

- Handling Missing Signals & Null Entries:

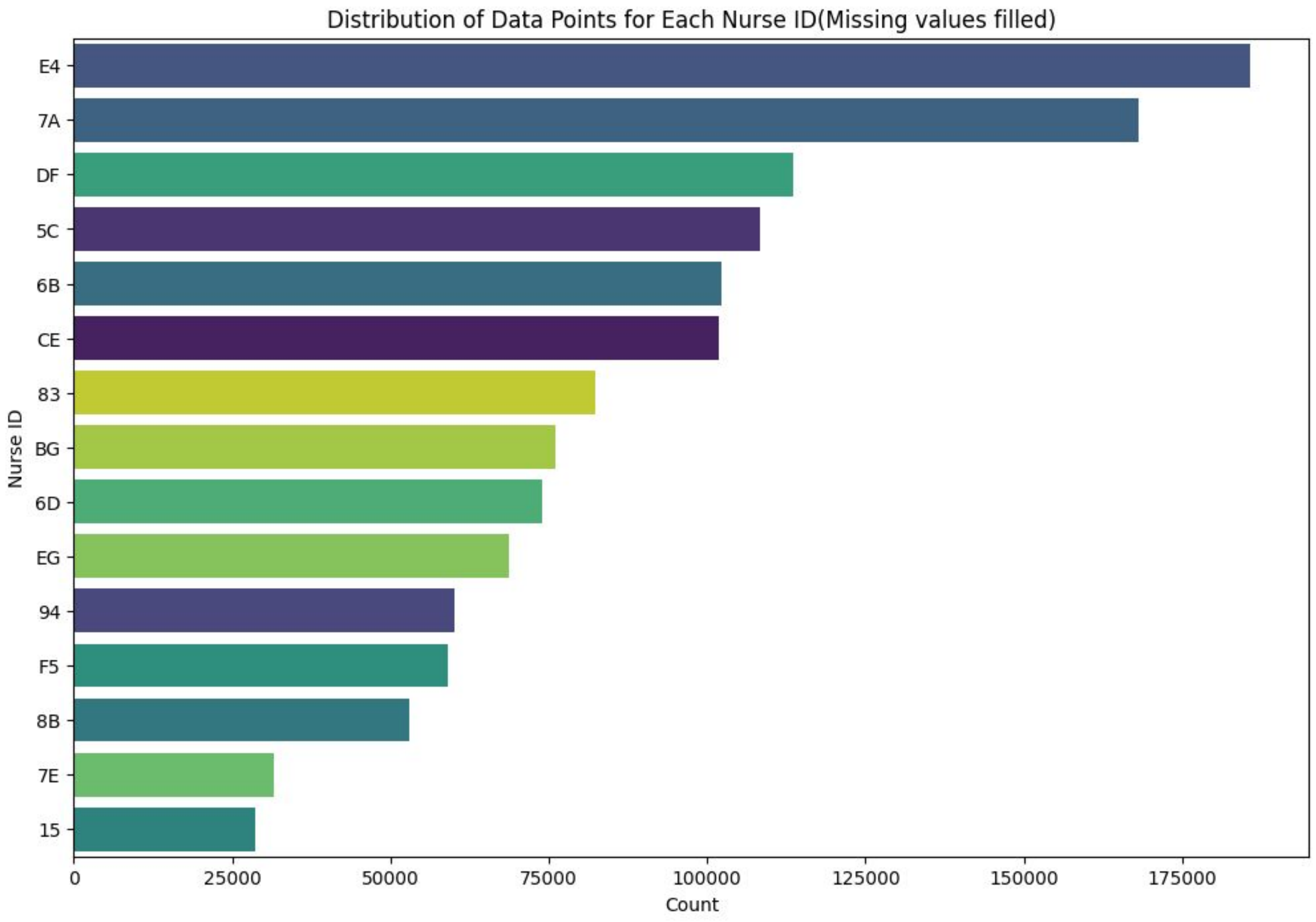
  → If all physiological signals are NaN, the record is removed without imputation ❌

  - Example: About 15–20% of data from Nurse ID 15 and 94 had continuous missing values across multiple fields

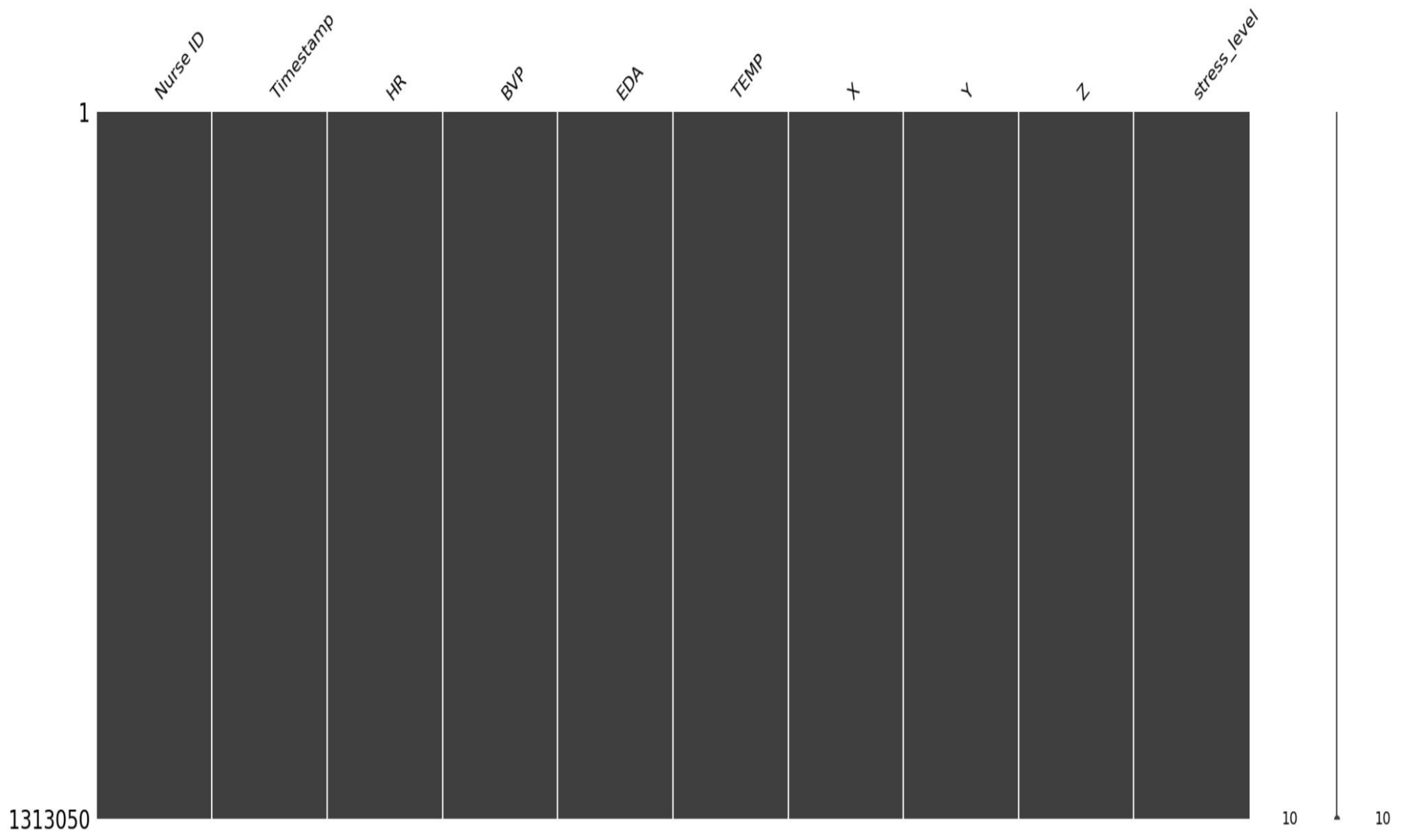
    → Dropping them was necessary since imputing with the mean could lead to unreliable or misleading outcomes ⚠️

→ The HR (Heart Rate) column had some missing values due to sampling frequency change (from 1Hz → 4Hz).

- These were filled using the average of adjacent values.
- Around 240 records (out of 100,000) were imputed this way.

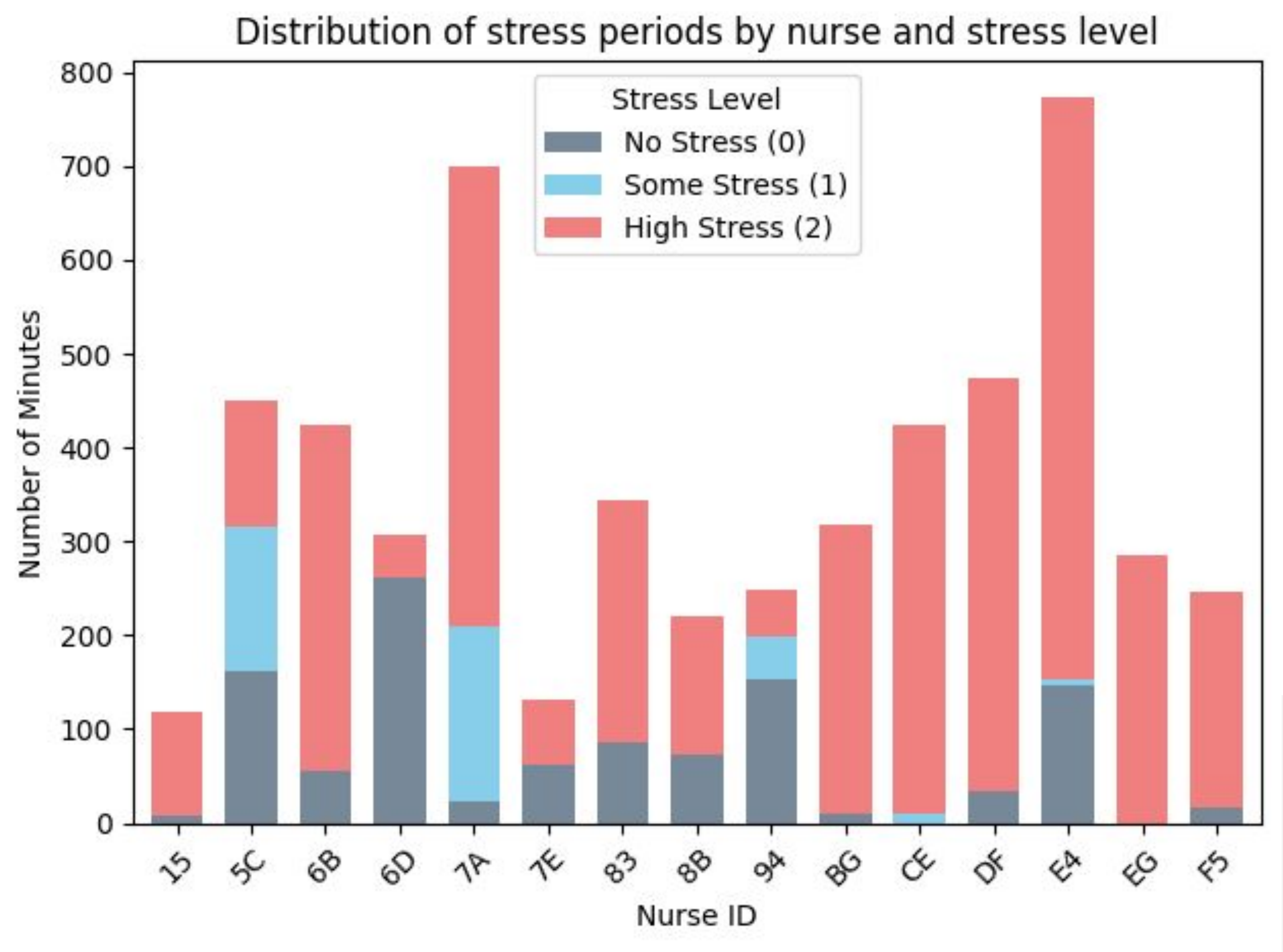


With removing nulls



All NaN removed from all fields

🧠 What We Did?



4. Structure Physiological Data into Time-Based Windows:

- <u>*Windowing Method:*</u>

  → Created 1-minute windows from the 4Hz sampling data

  → Grouped by NurseID and stress_level

  → Applied aggregation functions: mean and standard deviation (std)

  → Filtered out windows with only 240 samples to ensure full 250ms intervals are included per window

  → Final shape of windowed data: (5672, 18) 🗃️

Nurse ID	minute_window	stress_level	BVP_mean	BVP_std	EDA_mean	EDA_std	HR_mean	HR_std	TEMP_mean	TEMP_std	samples_per_window
15	2020-07-21 13:23:00	0.000	7.081	207.846	4.891	0.701	95.836	0.979	33.542	0.064	240
15	2020-07-21 13:24:00	0.000	1.414	120.570	4.867	0.719	101.032	1.998	33.706	0.172	240
15	2020-07-21 13:25:00	0.000	-1.036	99.918	4.025	0.591	98.304	2.346	34.037	0.036	240
15	2020-07-21 13:26:00	0.000	-1.201	88.186	4.527	0.216	109.153	2.654	34.140	0.038	240
15	2020-07-21 13:27:00	0.000	-2.112	103.255	4.359	0.236	99.423	1.224	34.360	0.079	240
...	...	...	...	...	...	...	...	...	...	...	...
F5	2020-07-23 17:25:00	2.000	-1.343	59.422	6.726	0.417	97.865	4.608	34.530	0.082	240
F5	2020-07-23 17:26:00	2.000	-1.129	51.674	5.494	0.526	100.974	4.536	34.581	0.043	240
F5	2020-07-23 17:27:00	2.000	-0.937	53.455	4.786	0.189	87.307	3.808	34.336	0.087	240
F5	2020-07-23 17:28:00	2.000	-1.319	68.752	3.713	0.367	84.026	2.591	33.922	0.126	240

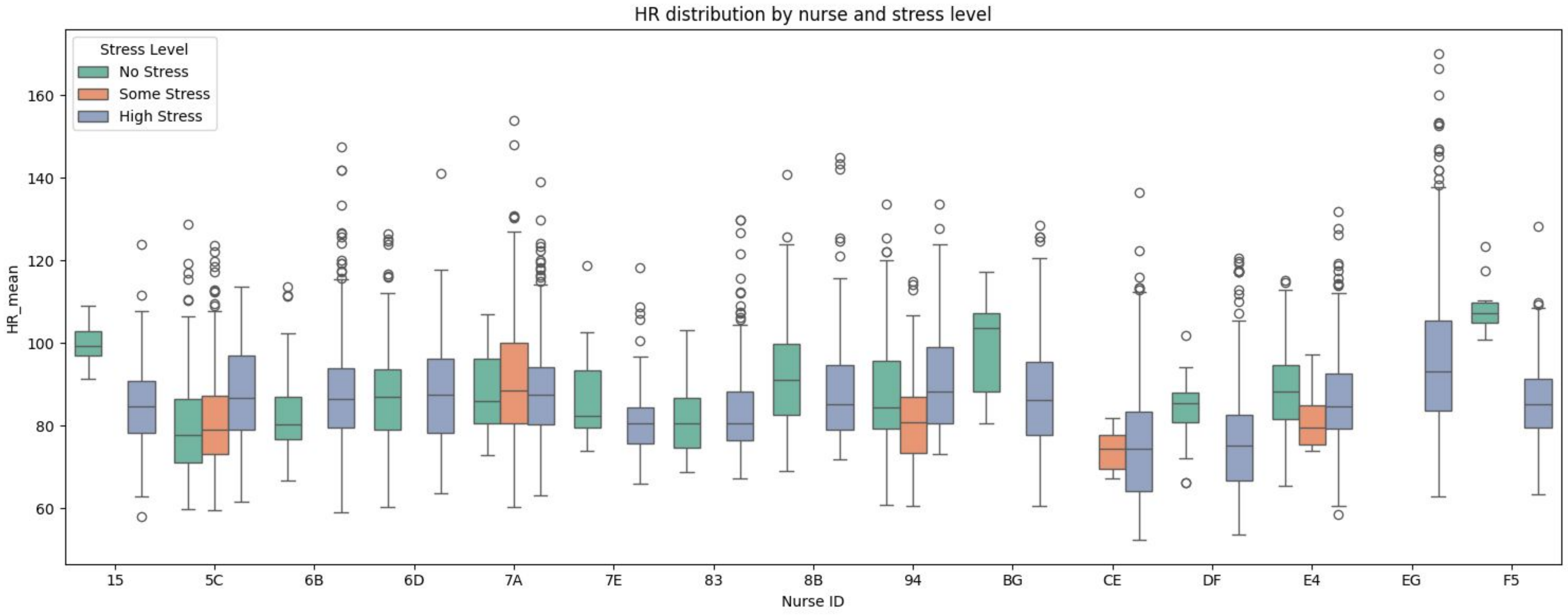
■

- **Boxplots for mean values of windowed physiological signals by NurseID:**

📖*[From research paper]:*

“The box plots show four individual physiological signals recorded by the wearable device—namely HR, EDA, skin temperature, BVP—and the reported stress levels from the survey.

Feature selection from within the signal streams at the user level improves performance, as biometric signals like EDA, HR, and temperature vary between subjects.”



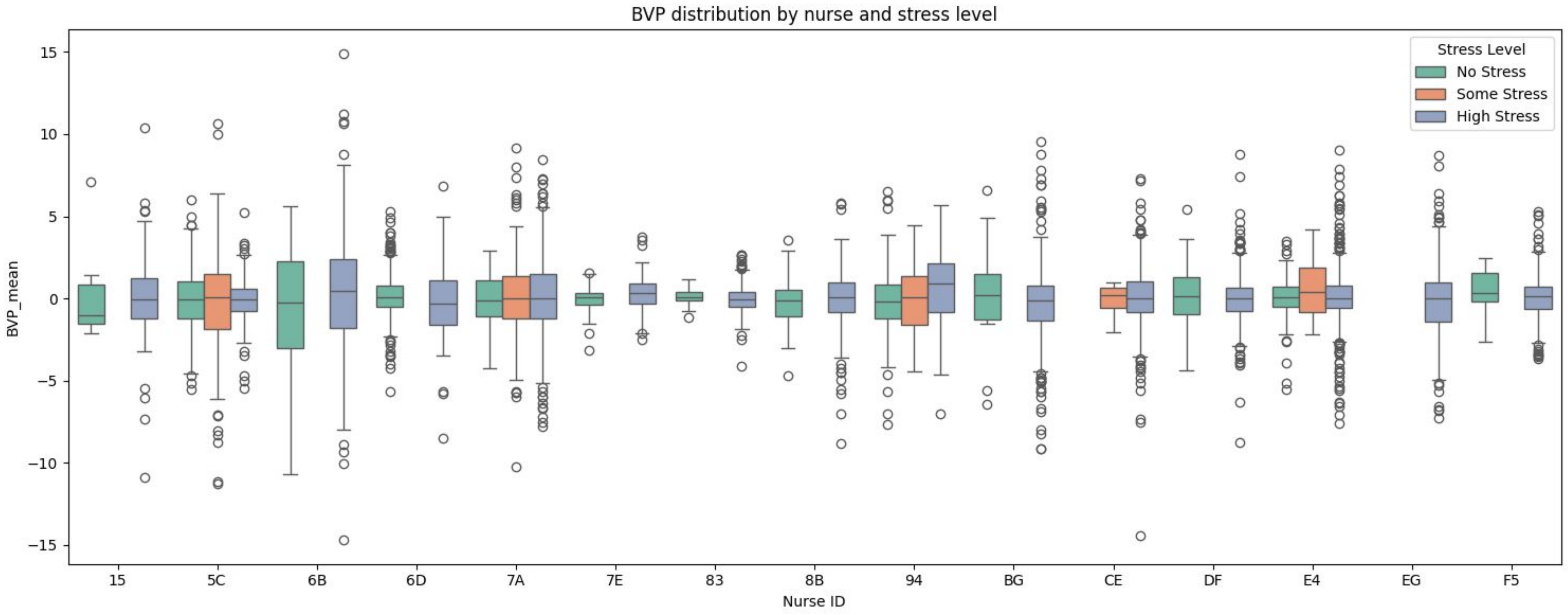
HR distribution by nurse and stress level

- **Boxplots for mean values of windowed physiological signals by NurseID:**

📖*[From research paper]:*

“The box plots show four individual physiological signals recorded by the wearable device—namely HR, EDA, skin temperature, BVP—and the reported stress levels from the survey.

Feature selection from within the signal streams at the user level improves performance, as biometric signals like EDA, HR, and temperature vary between subjects.”



BVP distribution by nurse and stress level

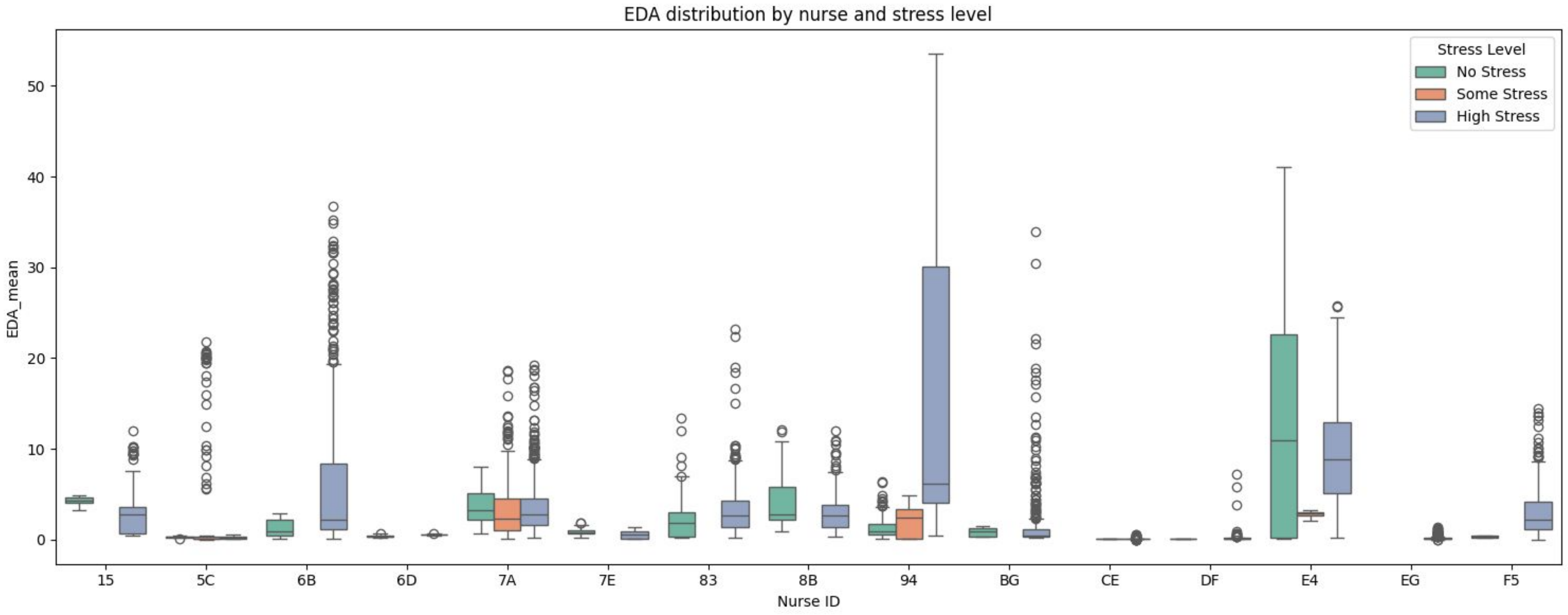
■

- **Boxplots for mean values of windowed physiological signals by NurseID:**

📖*[From research paper]:*

“The box plots show four individual physiological signals recorded by the wearable device—namely HR, EDA, skin temperature, BVP—and the reported stress levels from the survey.

Feature selection from within the signal streams at the user level improves performance, as biometric signals like EDA, HR, and temperature vary between subjects.”



EDA distribution by nurse and stress level

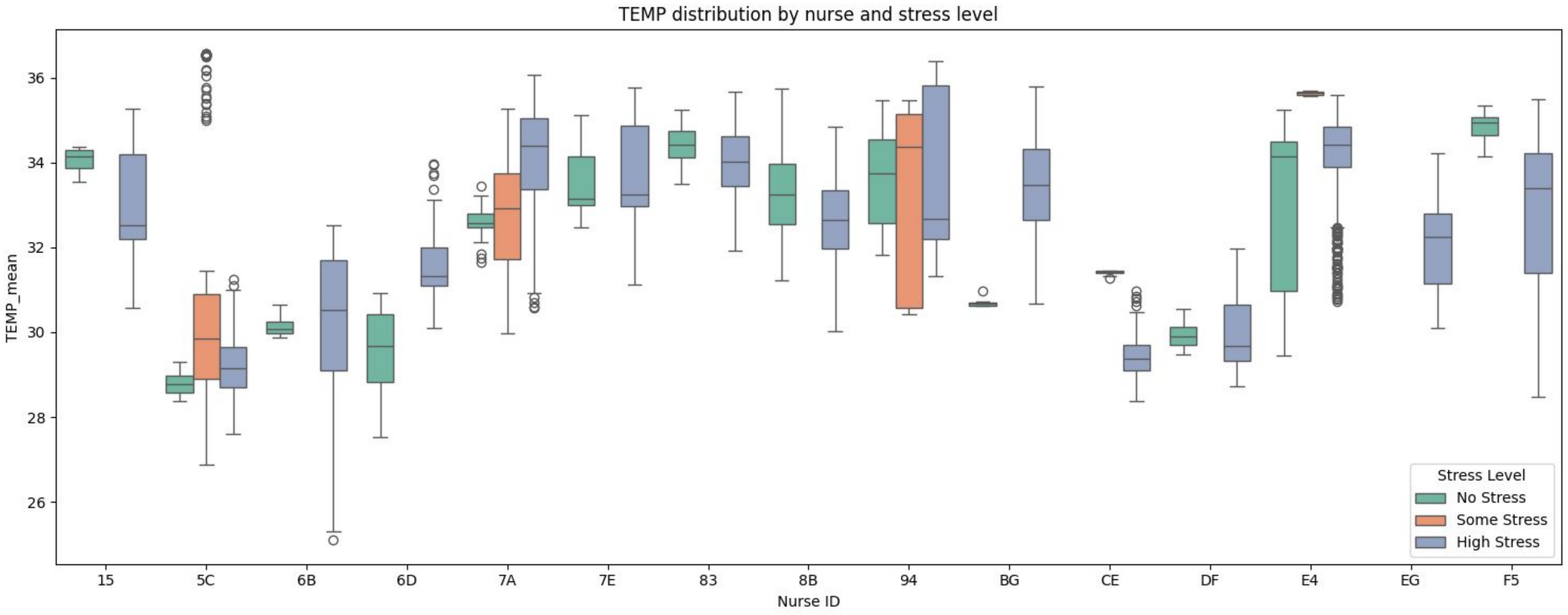
■

- **Boxplots for mean values of windowed physiological signals by NurseID:**

📖*[From research paper]:*

"The box plots show four individual physiological signals recorded by the wearable device—namely HR, EDA, skin temperature, BVP—and the reported stress levels from the survey.

Feature selection from within the signal streams at the user level improves performance, as biometric signals like EDA, HR, and temperature vary between subjects."



TEMP distribution by nurse and stress level

📌 Outcomes and Observations:

📖*[From research paper]:*

Based on the visual inspection of the plot, stress does not appear to have a strong correlation with Heart Rate. Heart Rate is generally associated with high-stress situations, but a high Heart Rate should not imply high stress, as it is more commonly influenced by non-stressful physical activity.

✅ ***Same trend seen in our HR boxplot.***

📖*[From research paper]:*

Distribution of EDA and associated stress levels for all the subjects. Based on visual inspection of the plot, stress has a positive correlation with EDA. The average EDA is higher in stressful situations for some participants. However, for some participants, EDA is not a good indicator of stress because it does not vary or is not positively correlated. There is high variability in EDA signals among various subjects in stressful events.

✅ ***Same pattern observed in our EDA boxplot.***

Skin Temperature (TEMP) and Blood Volume Pulse (BVP):

→ Do not show a clear difference in stress levels across individual nurses.

📖*[From research paper]:*

TEMP, this can vary quite widely based on the type and length of activity, as well as room temperature.

👍

■

📌 Outcomes and Observations:

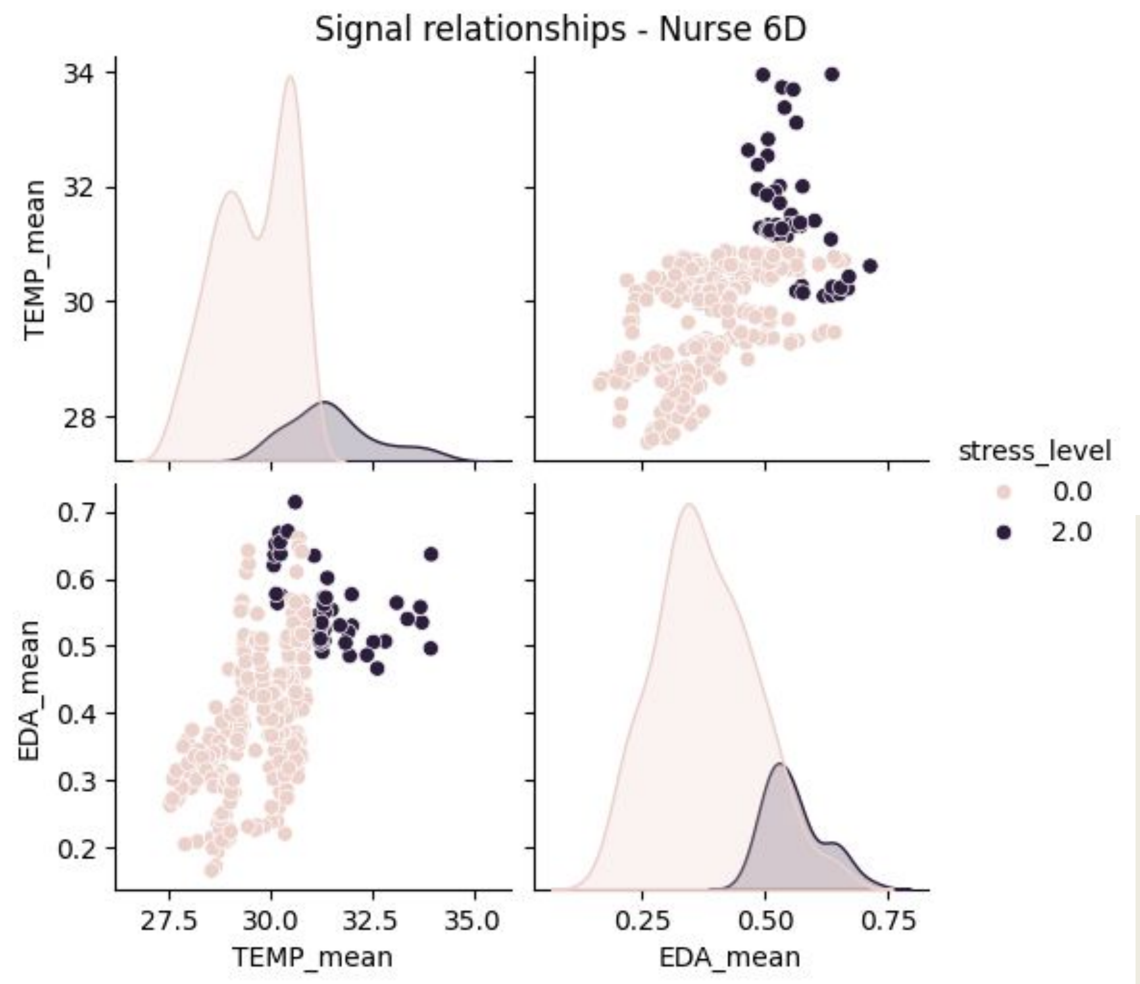
- Pairplots for Random NurseID’s:

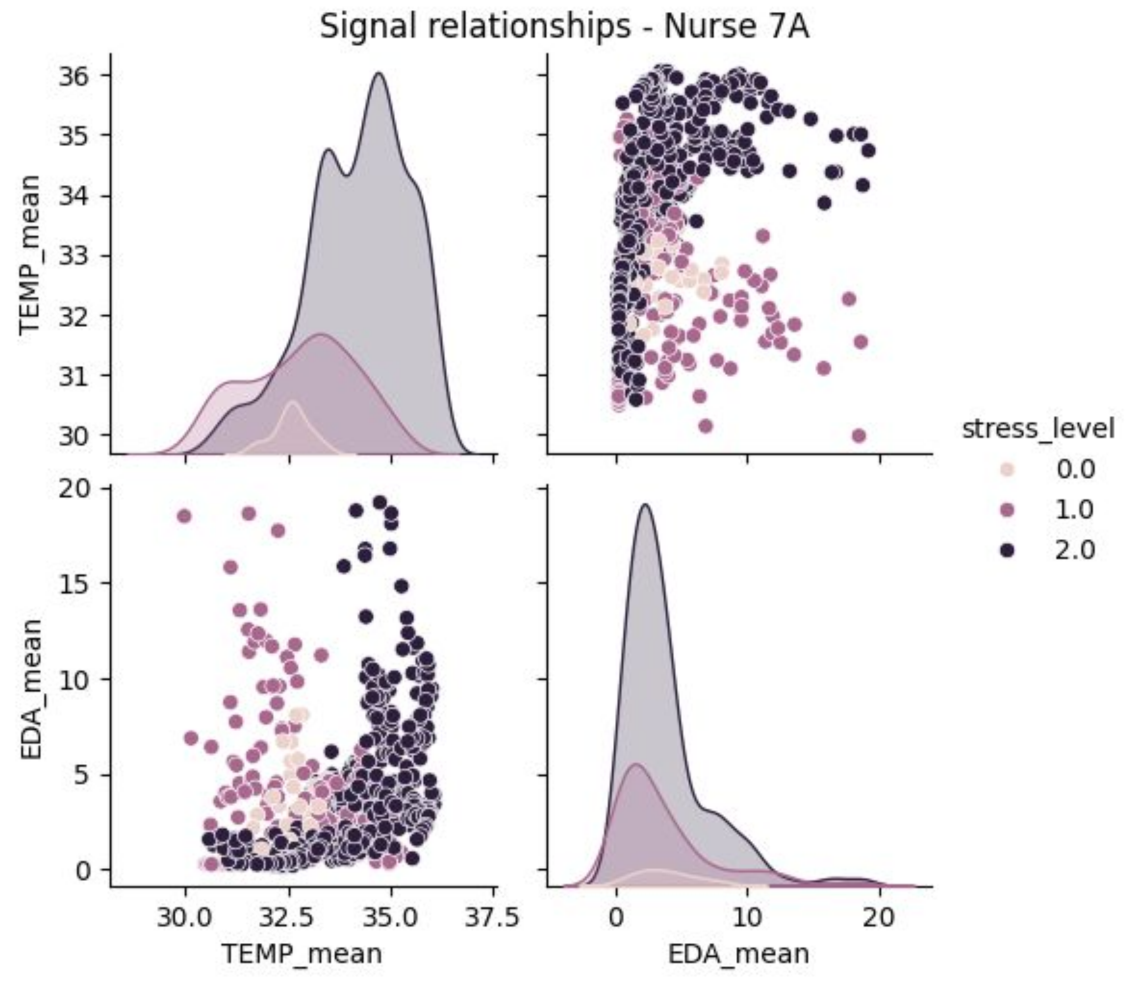
→ Observations from pair plots:

- Skewed distributions

- Non-linear data patterns

- Higher stress levels are associated with higher body temperature (TEMP) and greater electrodermal activity (EDA) (positive correlation).





🧠 What We Did?

5. Correlation Analysis:

→ *Pearson Correlation*

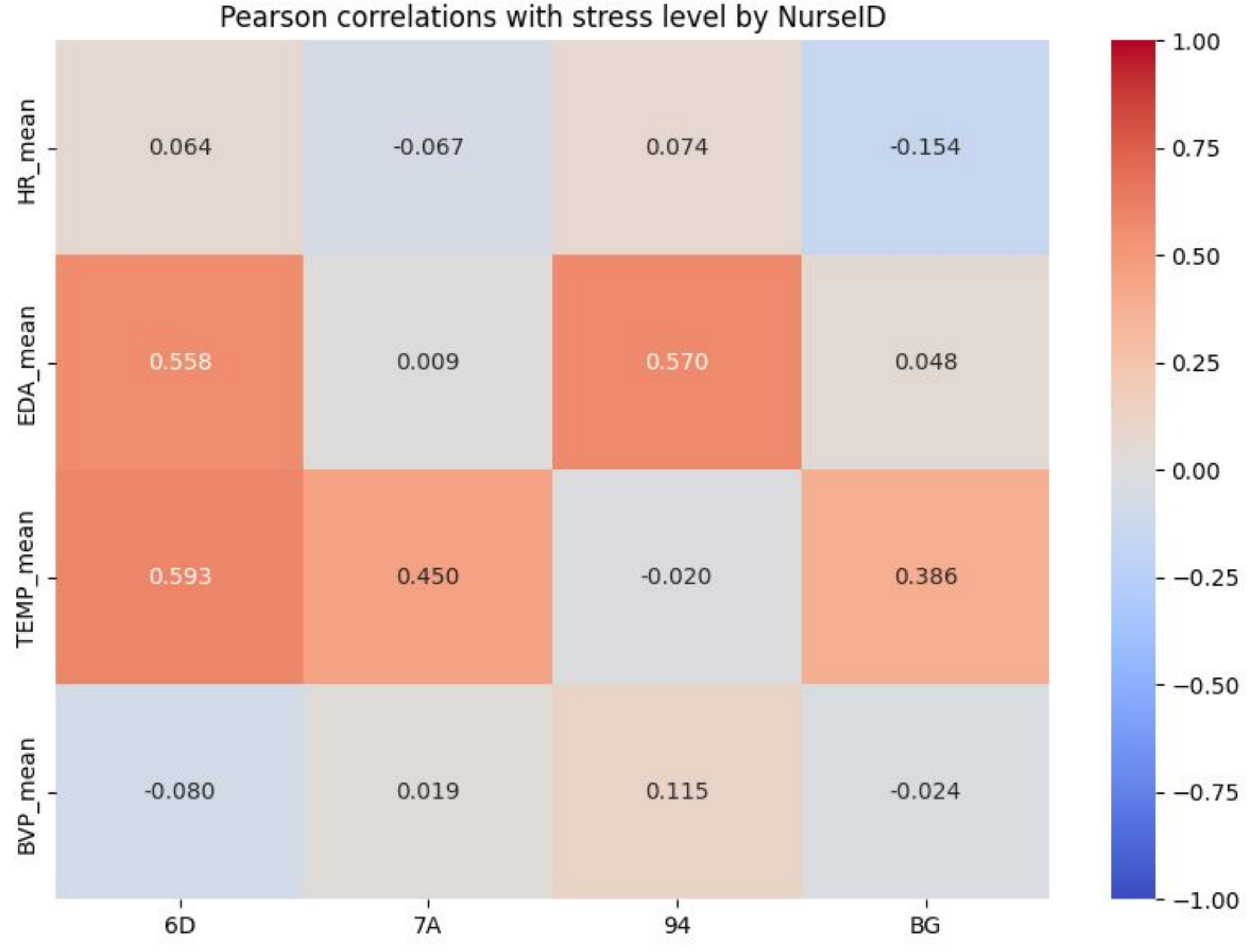
→ *Spearman Correlation*

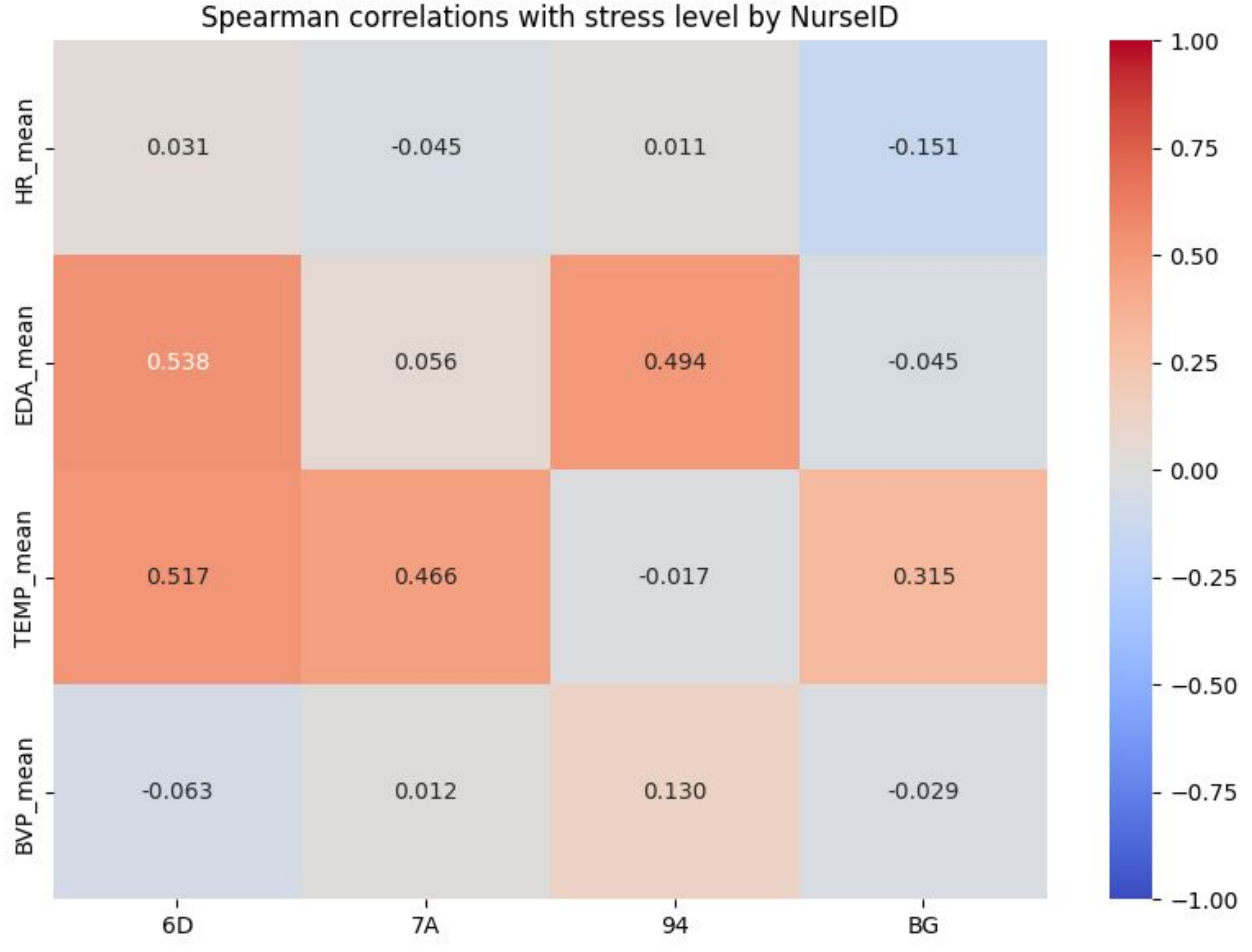
Pearson Correlation P-Values:

	6D	7A	94	BG
HR_mean	0.2664137068	0.0784311527	0.2442524341	0.0058606782
EDA_mean	0.0000000000	0.8209544365	0.0000000000	0.3899706516
TEMP_mean	0.0000000000	0.0000000000	0.7545586160	0.0000000000
BVP_mean	0.1606684511	0.6222456485	0.0710523405	0.6695527977

Spearman Correlation P-Values:

	6D	7A	94	BG
HR_mean	0.5824047549	0.2346371189	0.8586305579	0.0069088346
EDA_mean	0.0000000000	0.1372057593	0.0000000000	0.4244411368
TEMP_mean	0.0000000000	0.0000000000	0.7887241352	0.0000000097
BVP_mean	0.2718670106	0.7574426800	0.0408909315	0.6022265719





🧠 What We Did?

5. **Correlation Analysis:**

🔍 Observations from Correlation Analysis

→ HR shows weak correlation with stress levels

→ EDA for NurseID 6D and 94 shows positive correlation (weaker for others)

→ TEMP has moderate correlation for some subjects

→ BVP shows weak or no meaningful correlation with stress levels

🧠 What We Did?

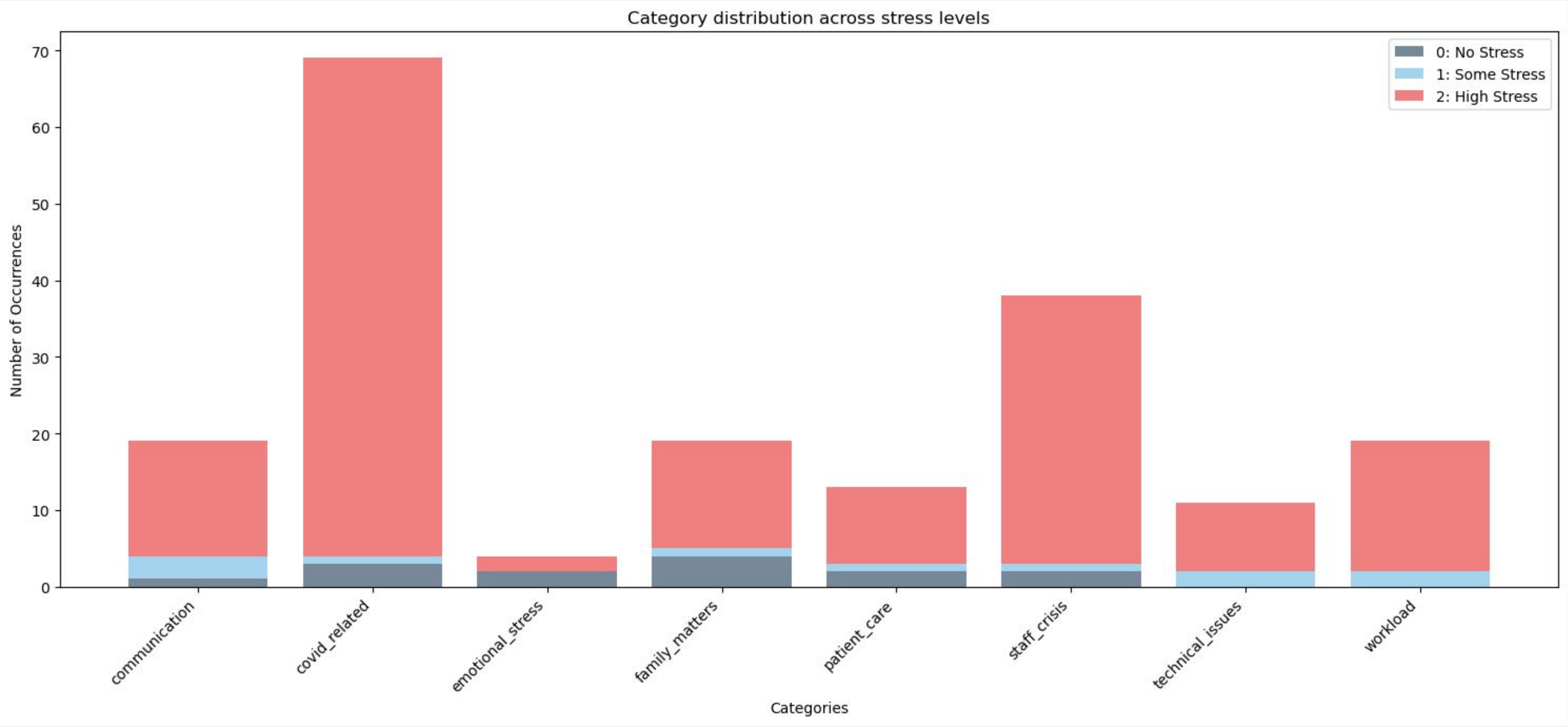
6. Survey Response Analysis with Word Clouds:

- Processing Survey Responses & Incident Descriptions:

→ Processed survey response categories related to recorded incidents 📝

→ Created subcategories for related stress events to improve visualization 🎯

→ Generated word clouds for each stress category to identify key stress-triggering factors ☁️ 🔍



🧠 What We Did?

6. Survey Response Analysis with Word Clouds:

- Processing Survey Responses & Incident Descriptions:

→ Processed survey response categories related to recorded incidents 📝

→ Created subcategories for related stress events to improve visualization 🎯

→ Generated word clouds for each stress category to identify key stress-triggering factors ☁️ 🔍

🔍 Findings:

→ High variability in physiological signal ranges across participants

→ Significant differences in individual responses — no uniform pattern detected ❗

→ Personalized assessment is essential for accurate stress monitoring

→ Correlations between signals can indicate positive/negative trends toward stress levels

→ Empatica E4 readings may be affected by environmental conditions

→ Contextual factors (e.g., workload, incidents) play a role in stress levels

→ Survey word clouds help visualize the major stress triggers

✅ Conclusion:

→ Personalized assessments are crucial for meaningful stress analysis

→ Physiological correlations provide valuable insights into stress trends

→ External conditions can distort signal accuracy and must be accounted for ⚠️

🚀 Future Directions / Next Steps:

→ Multimodal analysis: combine physiological + contextual data for prediction

→ Implement personalized models to adapt to individual variability

→ Use cross-validation with diverse data sources for improved accuracy

TEAM MEMBERS

Chau Nguyen
chau.nguyen@student.oulu.fi

Hans Karawitage
Hans.Madalagama@student.oulu.fi

Ata Jodeiri Seyedian
ata.jodeiri@student.oulu.fi

Fatemeh Soufian
Fatemeh.Soufian@student.oulu.fi
Leader

Who Did What?

- *Checking final signals csv file inconsistencies: Fatemeh*
- *Data Reduction (Align & Window Physiological Signals): Hans*
- *Windowing R&D: Ata*
- *Integrating: Hans*
- *Preparing Presentation: Hans & Fatemeh*
- *Preprocessing and Analysing Survey Results: Chau*
- *Correlation Analysis: All members*

THANK YOU

26 March, 2025