Data Mining Project

Preprocessing Procedure Presentation

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

26 February, 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 OBJECTIVES

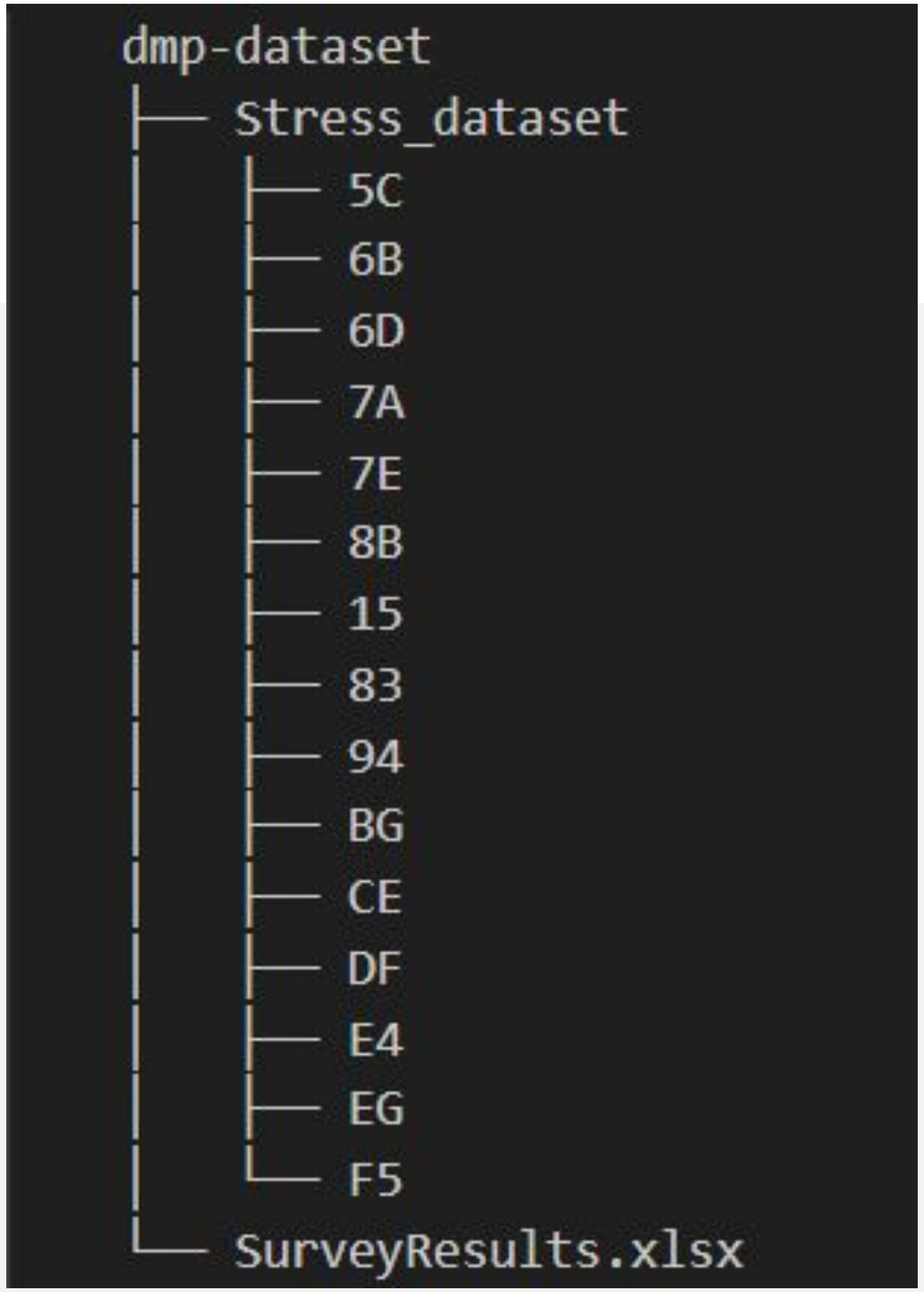
- How do physiological signals correlate with self-reported stress levels among nurses during the COVID-19 pandemic?
- What are the primary contextual factors(e.g., COVID-related challenges, workload, and patient crises) that contribute to stress among nurses, and how are these factors reflected in physiological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

Preprocessing Steps:

1. 📂 Multi-file Dataset → 🔗 Data Integration

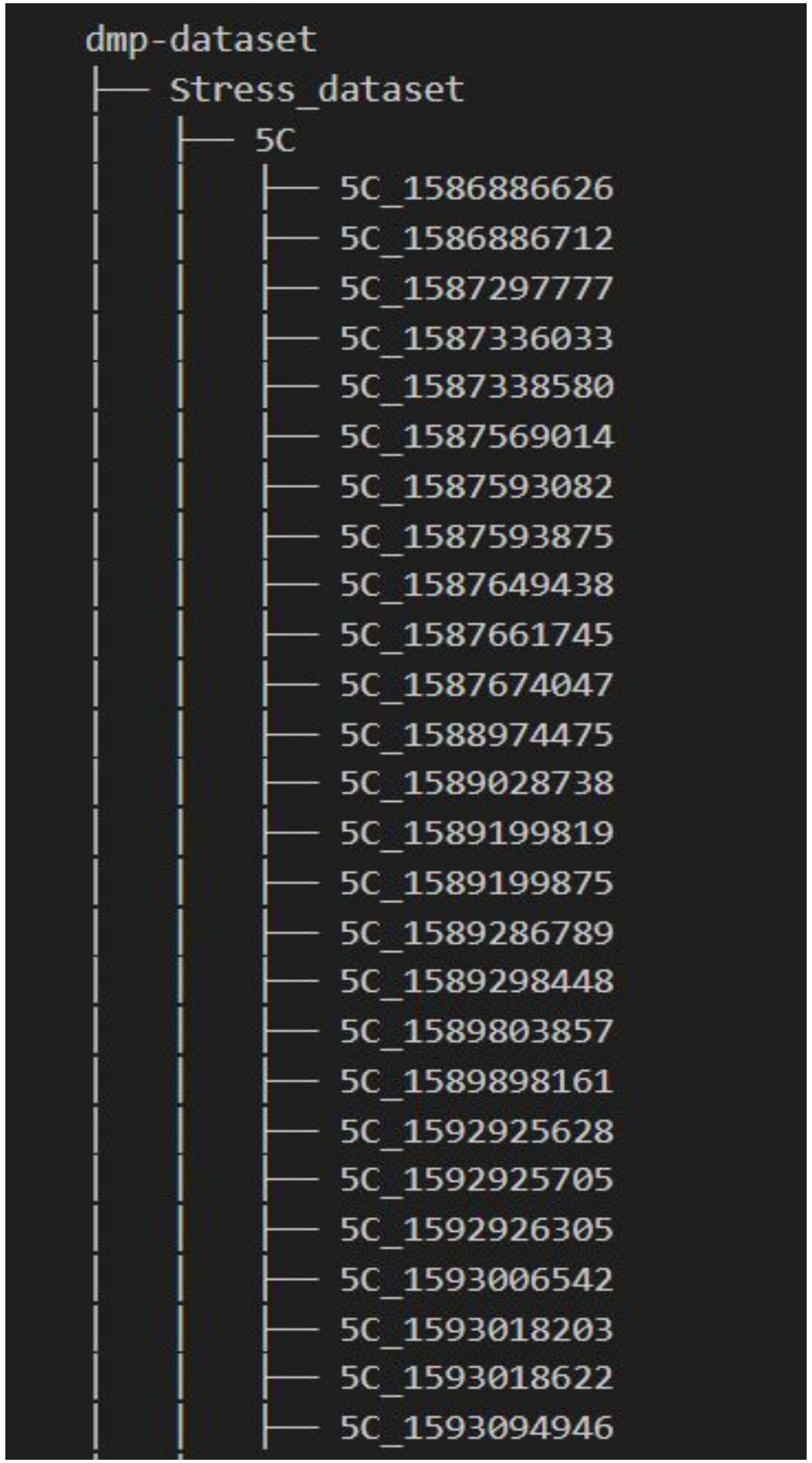
   ✨ Brings everything together in one place so we can work with it more easily.

- Our dataset consists of multiple files, each containing different physiological signals (HR, BVP, ACC, etc.).
- Signals for each nurse are spread across multiple files.

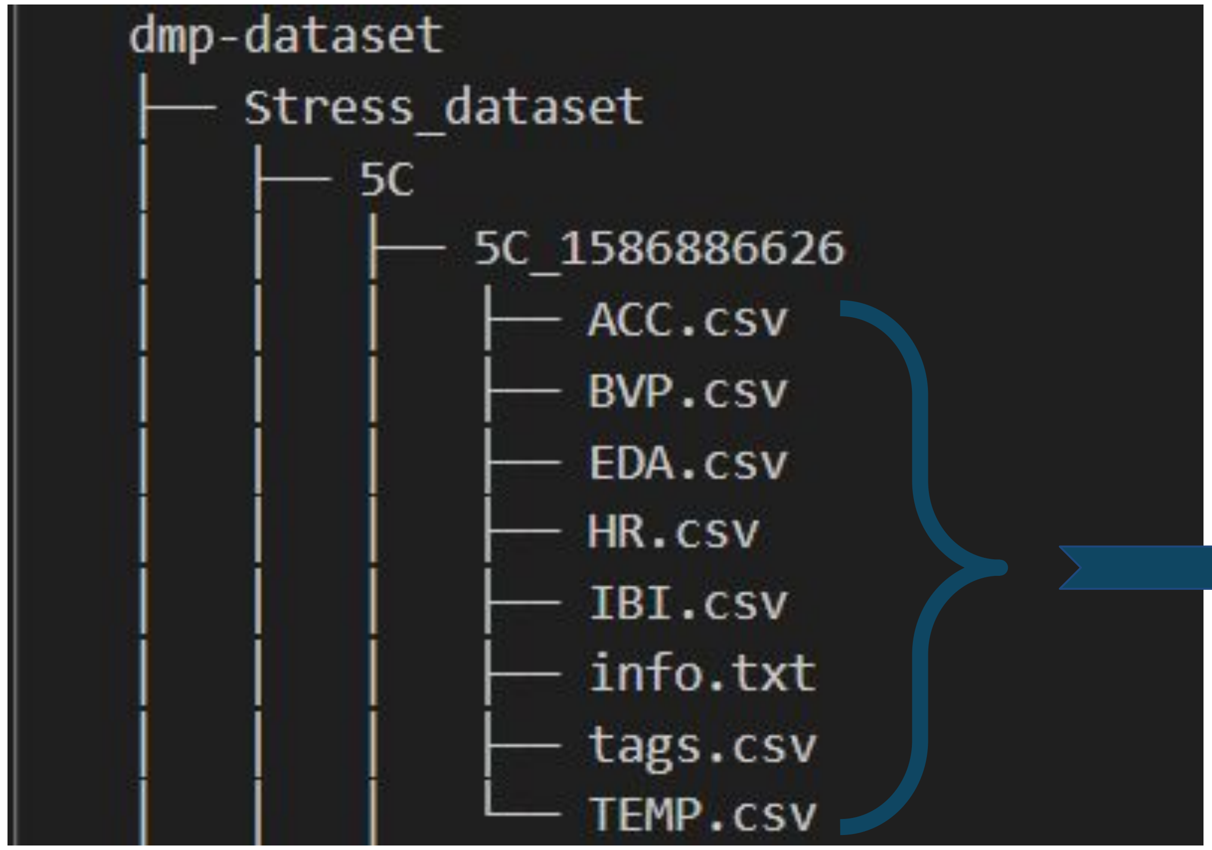


```
dmp-dataset
├── Stress_dataset
│   ├── 5C
│   ├── 6B
│   ├── 6D
│   ├── 7A
│   ├── 7E
│   ├── 8B
│   ├── 15
│   ├── 83
│   ├── 94
│   ├── BG
│   ├── CE
│   ├── DF
│   ├── E4
│   ├── EG
│   └── F5
└── SurveyResults.xlsx
```

- ## Dataset Files Structure



```
dmp-dataset
├── Stress_dataset
│   ├── 5C
│   │   ├── 5C_1586886626
│   │   ├── 5C_1586886712
│   │   ├── 5C_1587297777
│   │   ├── 5C_1587336033
│   │   ├── 5C_1587338580
│   │   ├── 5C_1587569014
│   │   ├── 5C_1587593082
│   │   ├── 5C_1587593875
│   │   ├── 5C_1587649438
│   │   ├── 5C_1587661745
│   │   ├── 5C_1587674047
│   │   ├── 5C_1588974475
│   │   ├── 5C_1589028738
│   │   ├── 5C_1589199819
│   │   ├── 5C_1589199875
│   │   ├── 5C_1589286789
│   │   ├── 5C_1589298448
│   │   ├── 5C_1589803857
│   │   ├── 5C_1589898161
│   │   ├── 5C_1592925628
│   │   ├── 5C_1592925705
│   │   ├── 5C_1592926305
│   │   ├── 5C_1593006542
│   │   ├── 5C_1593018203
│   │   ├── 5C_1593018622
│   │   ├── 5C_1593094946
```



```
dmp-dataset
├── Stress_dataset
│   ├── 5C
│   │   ├── 5C_1586886626
│   │   │   ├── ACC.csv
│   │   │   ├── BVP.csv
│   │   │   ├── EDA.csv
│   │   │   ├── HR.csv
│   │   │   ├── IBI.csv
│   │   │   ├── info.txt
│   │   │   ├── tags.csv
│   │   │   └── TEMP.csv
```

Signals for the nurse with 5C ID at timestamp 1586886626

What We Did?

```
def extract(directory_path):
    for dir in os.scandir(directory_path):
        if dir.is_file() and
dir.path.endswith(".zip"):
            extract_path = dir.path[:-4]
            with zipfile.ZipFile(dir.path, 'r')
as zip_ref:
                print(f'extracting {dir.path}')

zip_ref.extractall(extract_path)

            os.remove(dir.path)
            extract(extract_path)

        elif dir.is_dir():  # If it's a
directory, go inside to check for zip files
            extract(dir.path)
```

🗂️ ***Extract all ZIP files recursively, handling nested structures efficiently.***

➕ *Add Nurse ID and Timestamp to each extracted CSV file, ensuring each record is uniquely identifiable.*

➡️ *Move all processed CSV files to the root folder for easy access.*

🗑️ *Delete empty subfolders after extraction to keep the workspace clean and organized.*

What We Did?

→ *In each CSV file, the first two rows contain metadata:*

- *Initial timestamp – The start time of the recording.*
- *Sampling frequency – The rate at which data was recorded (Hz).*

Using these values, we calculate the timestamp for each record.

→ *For Nurse ID, we extract it from the filename, which follows the format: `NurseID_InitialTimestamp`*

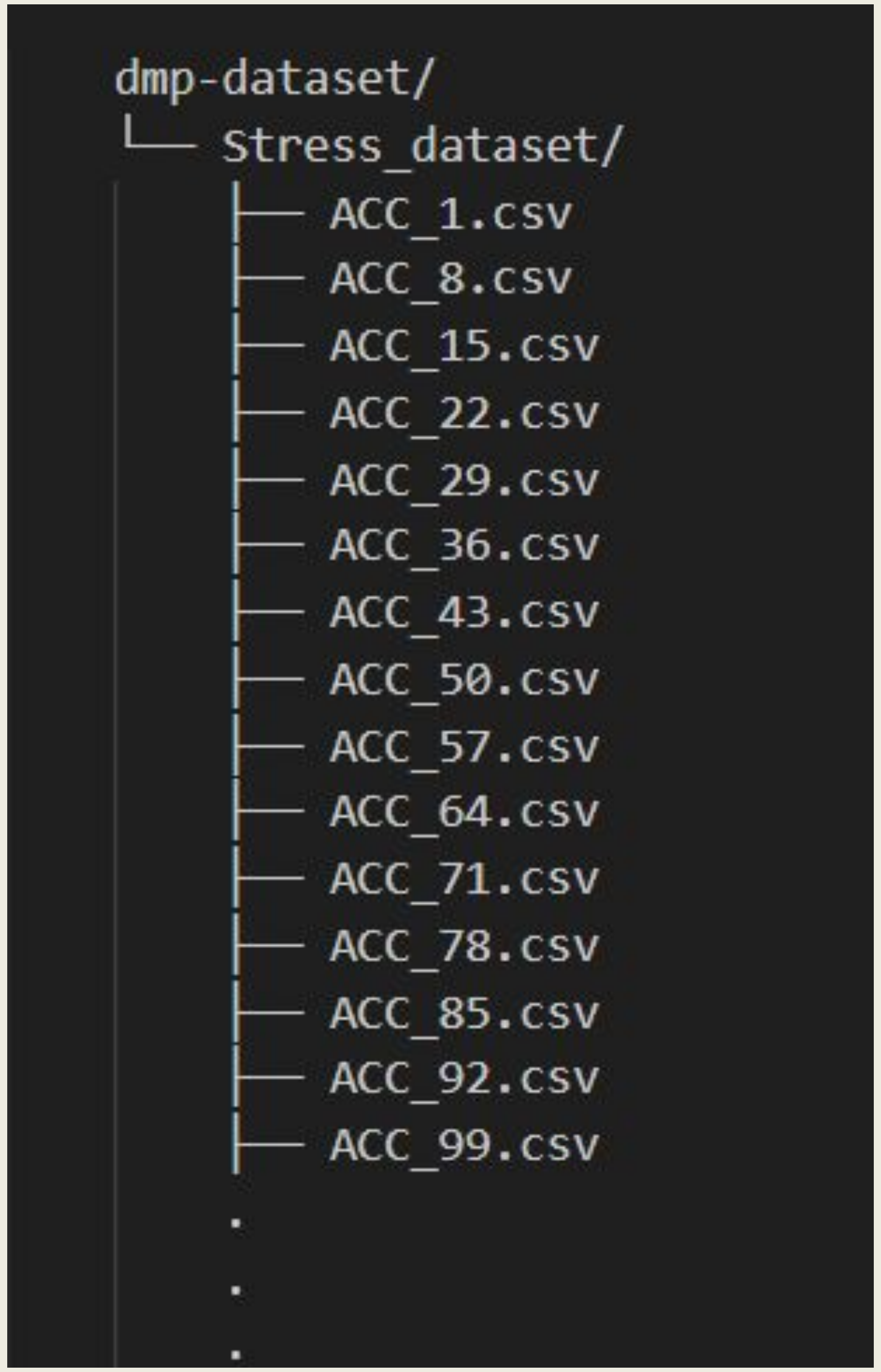
🗂️ *Extract all ZIP files recursively, handling nested structures efficiently.*

➕ ***Add Nurse ID and Timestamp to each extracted CSV file, ensuring each record is uniquely identifiable.***

➡️ *Move all processed CSV files to the root folder for easy access.*

🗑️ *Delete empty subfolders after extraction to keep the workspace clean and organized.*

New Dataset Files Structure



```
dmp-dataset/
└── Stress_dataset/
    ├── ACC_1.csv
    ├── ACC_8.csv
    ├── ACC_15.csv
    ├── ACC_22.csv
    ├── ACC_29.csv
    ├── ACC_36.csv
    ├── ACC_43.csv
    ├── ACC_50.csv
    ├── ACC_57.csv
    ├── ACC_64.csv
    ├── ACC_71.csv
    ├── ACC_78.csv
    ├── ACC_85.csv
    ├── ACC_92.csv
    ├── ACC_99.csv
    .
    .
    .
```

What We Did?

🗂️ *Extract all ZIP files recursively, handling nested structures efficiently.*

➕ *Add Nurse ID and Timestamp to each extracted CSV file, ensuring each record is uniquely identifiable.*

➡️ ***Move all processed CSV files to the root folder for easy access.***

🗑️ ***Delete empty subfolders after extraction to keep the workspace clean and organized.***

Preprocessing Steps:

2. 🎚️ Different sampling frequency → 🔁 Resampling

Empatica E4 Signals and Frequency chart

Signal	Abbreviation	Frequency (Hz)
Heart Rate	HR	1.0
Electrodermal Activity	EDA	4.0
Skin Temperature	TEMP	4.0
Accelerometer	ACC	32.0
Inter-Beat Interval	IBI	64.0
Blood Volume Pulse	BVP	64.0

🚨Some Problems Caused by Different Sampling Rates:

- Very large file sizes – Just BVP data alone is 7GB, slowing down processing.
- If we don't resample, for a specific timestamp, not all measurements will be available, resulting in a huge number of records, most of which will contain multiple null values.

Preprocessing Steps:

2. 🎚️ Different sampling frequency → 🔁 Resampling

📊 Resampling Process:

1. Choosing a Common Frequency:

   ✅ Chosen Common Frequency: 4 Hz

2. Upsampling:

   Low-frequency signals (HR at 1 Hz) → Increased to 4 Hz using interpolation to fill missing values.

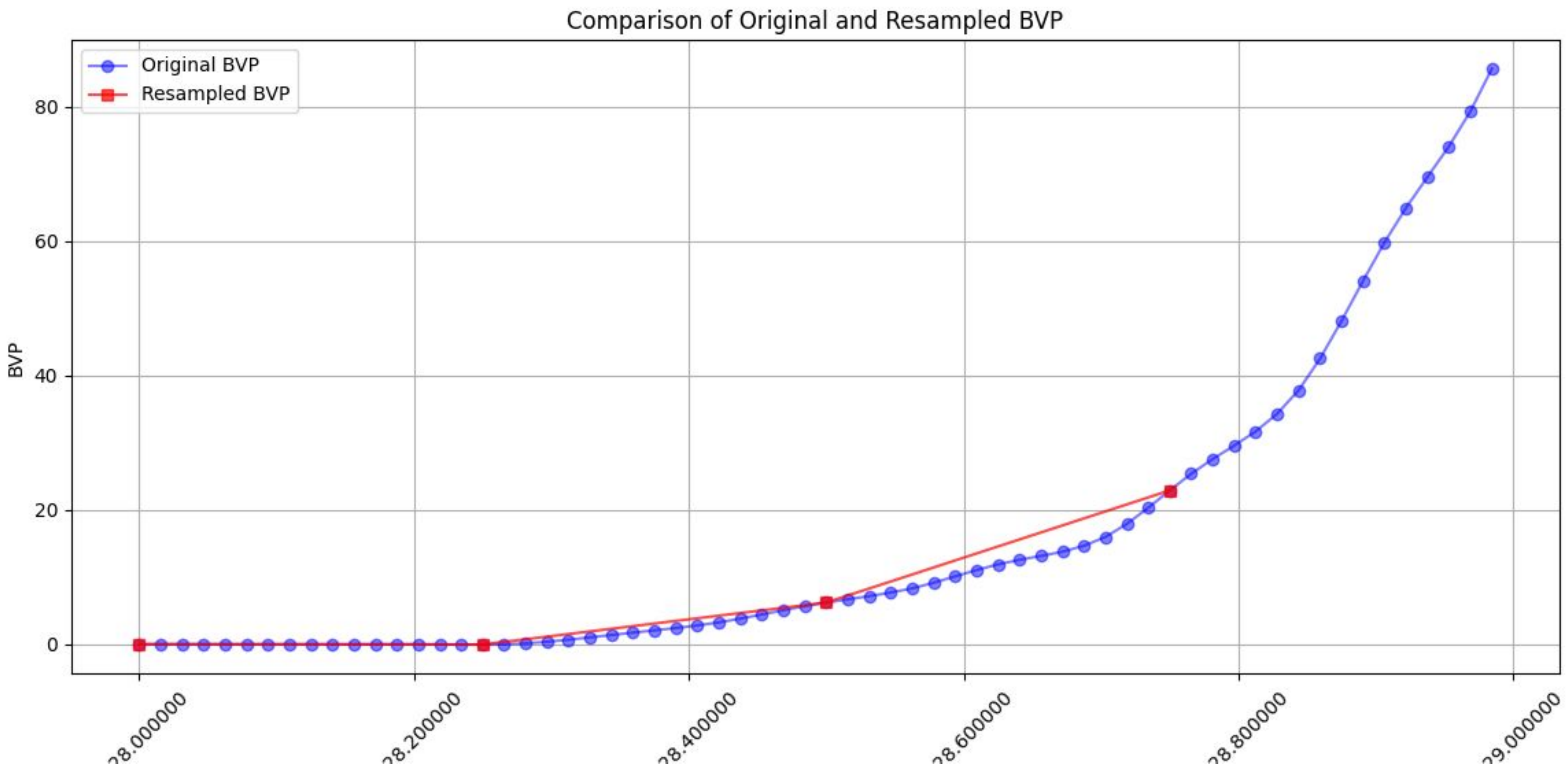
3. Downsampling:

   High-frequency signals (BVP at 64 Hz, ACC at 32 Hz) → Reduced to 4 Hz using .asfreq(), selecting values at the exact resampled timestamps.

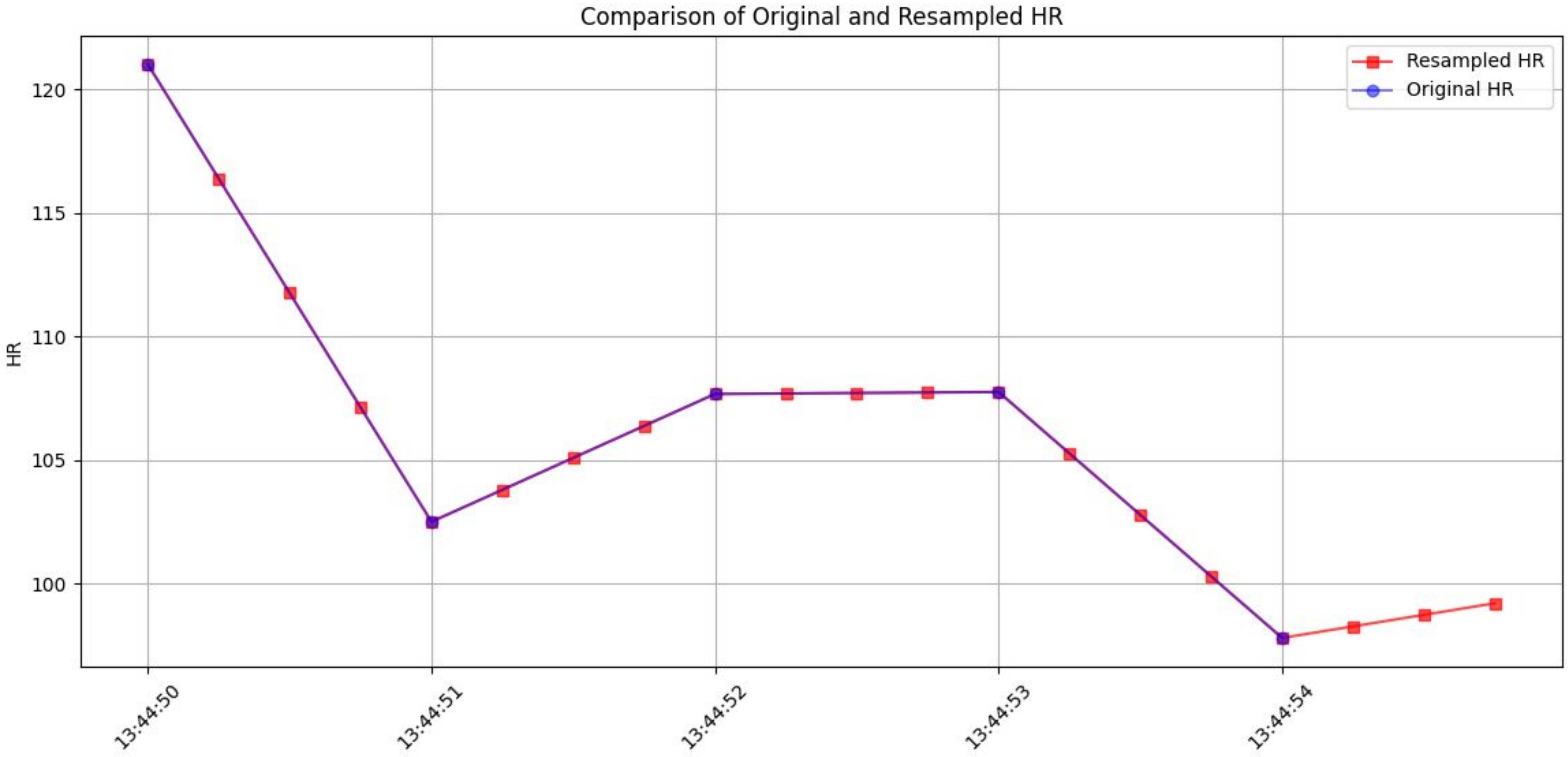
Empatica E4 Signals and Frequency chart

Signal	Abbreviation	Frequency (Hz)
Heart Rate	HR	1.0
Electrodermal Activity	EDA	4.0
Skin Temperature	TEMP	4.0
Accelerometer	ACC	32.0
Inter-Beat Interval	IBI	64.0
Blood Volume Pulse	BVP	64.0

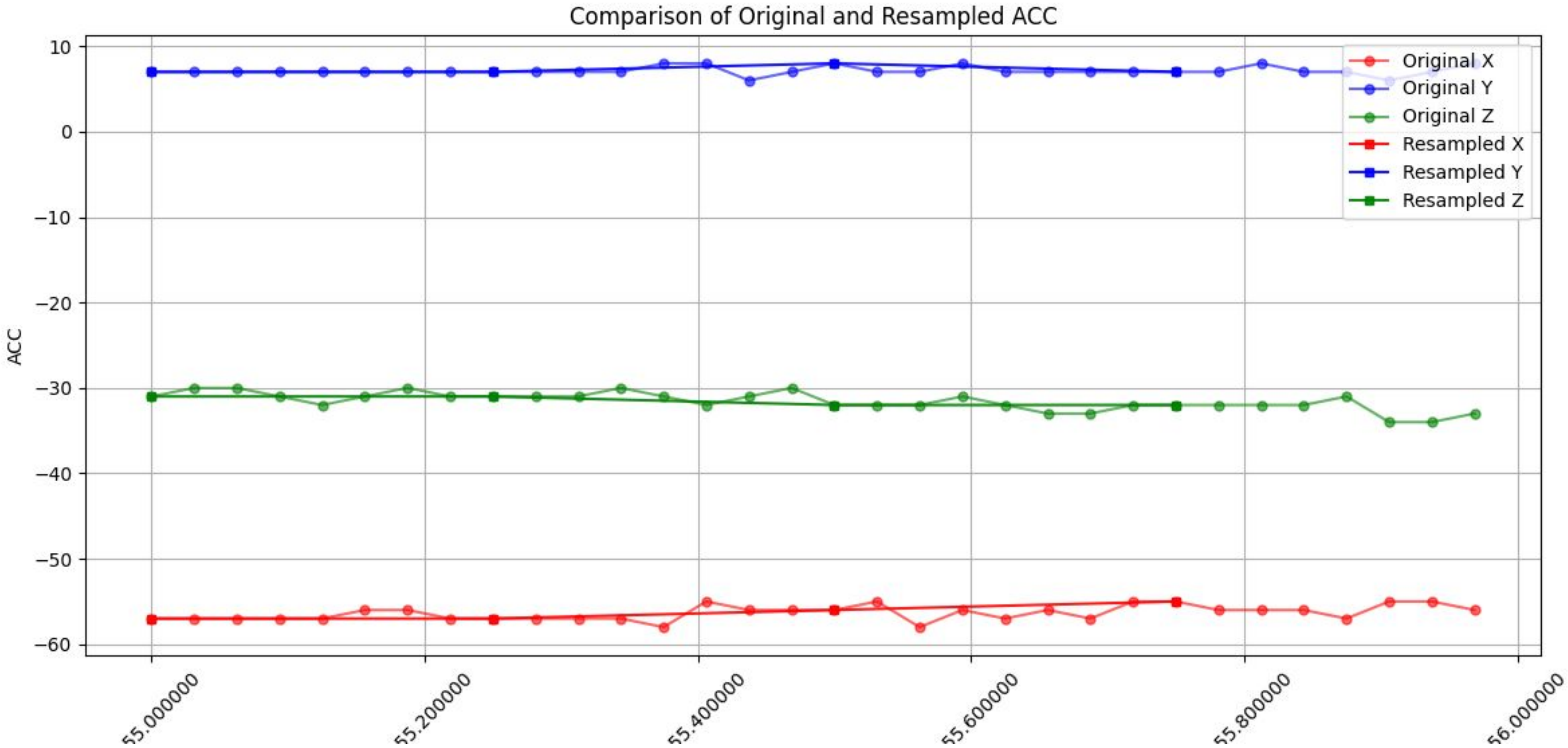
Comparison of Original vs. Resampled BVP Data

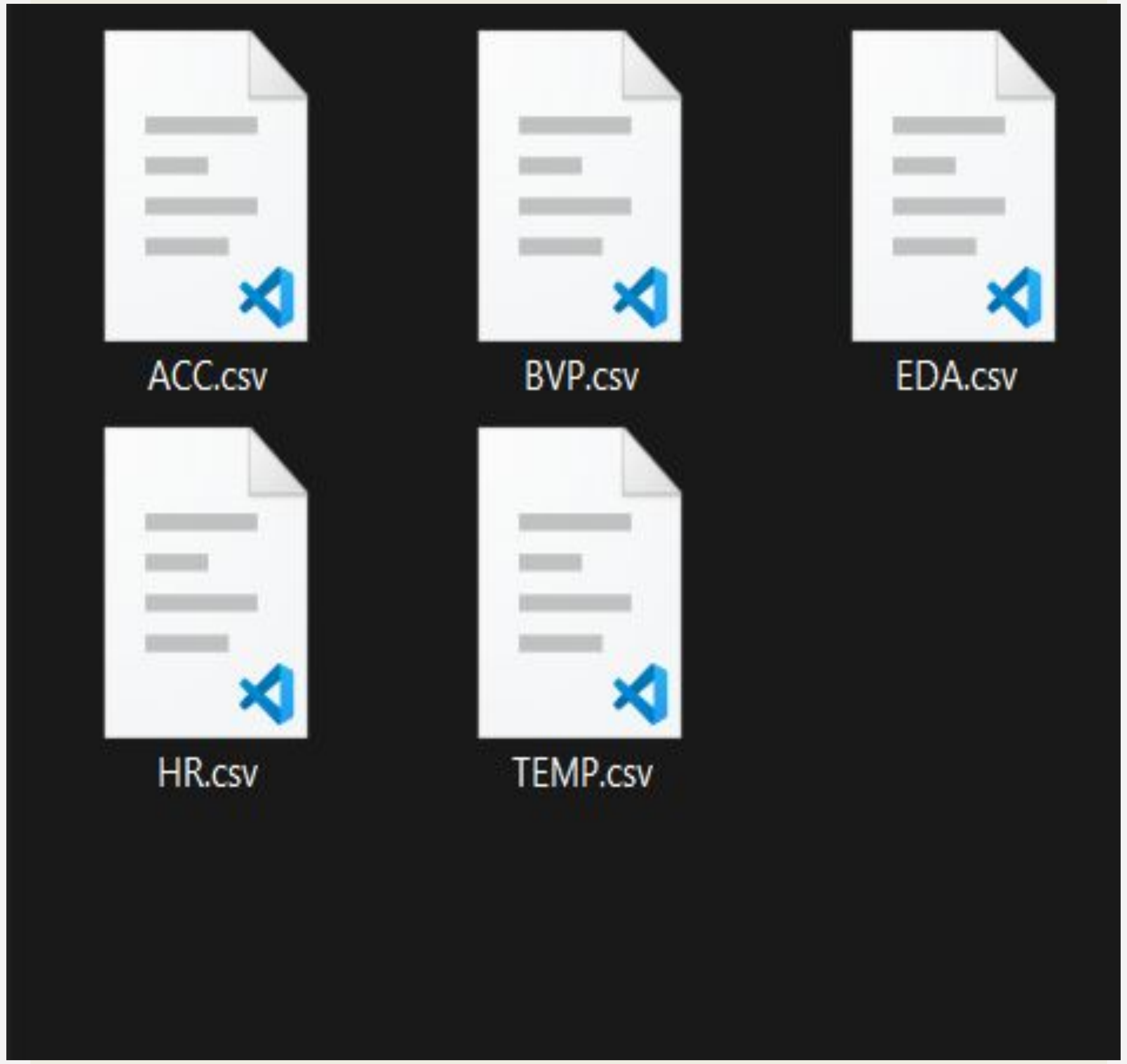


Comparison of Original vs. Resampled HR Data



Comparison of Original vs. Resampled ACC Data





Building the Final Unified Dataset

❑ *Step 1: Signal-Based Concatenation:*

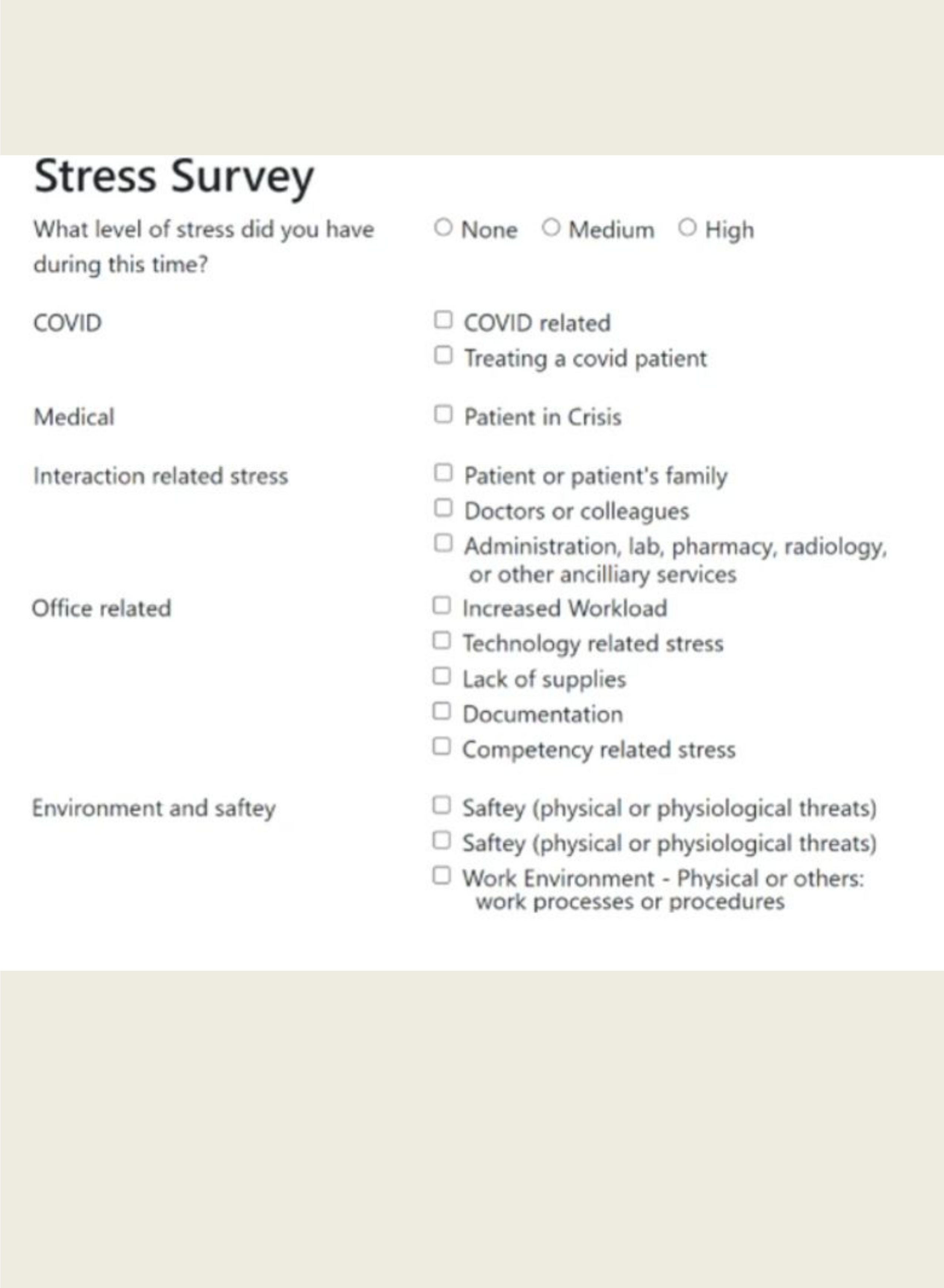
Individual measurement files were concatenated per signal type (HR, BVP, ACC, EDA, TEMP).

➔ *This resulted in five final CSV files, each containing all data for a specific physiological signal.*

❑ *Step 2: Merging All Signals:*

The five signal-based CSVs were merged using Timestamp and Nurse ID as the composite key.

This created a single unified dataset, ensuring all physiological signals are aligned for analysis.



Stress Survey

What level of stress did you have during this time?	○ None ○ Medium ○ High
COVID	□ COVID related □ Treating a covid patient
Medical	□ Patient in Crisis
Interaction related stress	□ Patient or patient's family □ Doctors or colleagues □ Administration, lab, pharmacy, radiology, or other ancilliary services
Office related	□ Increased Workload □ Technology related stress □ Lack of supplies □ Documentation □ Competency related stress
Environment and saftey	□ Saftey (physical or physiological threats) □ Saftey (physical or physiological threats) □ Work Environment - Physical or others: work processes or procedures

Preprocess SurveyResult File

❑ *Handling Missing Values*

❑ *Converting Datetime to Timestamp*

❑ *Still in progress..* 🛠️⏳

Next Steps:

1. Check Final Unified File for Noise & Missing Values

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

   - Filter physiological signals within the Start and End time window of each reported stress event.

3. Merging Physiological Data with Survey Data

   - Integrate physiological signals with self-reported stress levels for a complete dataset.

4. Correlation & Statistical Analysis:

   - Pearson/Spearman Correlation: Measure how stress levels and physiological signals (HR, BVP, etc.) are related.

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?

- *Data Integration: Fatemeh & Ata*
- *Resampling R&D: Hans*
- *Resampling Implementation: Fatemeh*
- *Preparing Presentation: Hans & Fatemeh*
- *Preprocessing Survey Results: Chau*

THANK YOU

26 February, 2025