Univerzitet u Sarajevu
Elektrotehnički fakultet
Sarajevo

DSAI
DATA SCIENCE & ARTIFICIAL INTELLIGENCE

# Evaluation Report

## Stress-level Management via Smartwatch Monitoring

**Course:** Machine Learning: Supervised Learning

**Professor:** Prof. Dr. Amila Akagić

**Students:** Emin Hadžiabdić, Muhammed Pašić, Armin Memišević

**Academic Year:** 2025/26

**Student ID:** 19960, 20109, 20016

Sarajevo, January 5th 2025

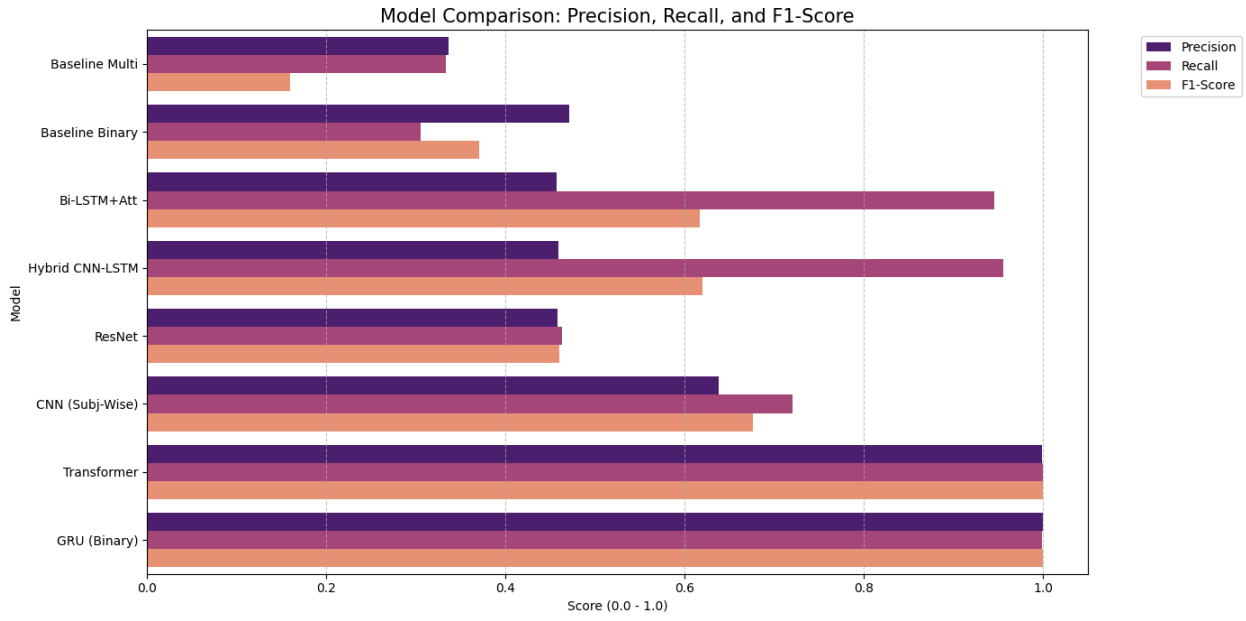
# Table of Contents:

# 1 Our strategy

This project followed an iterative development cycle, first having complex and high-dimensional research models and moving to a streamlined and an architecture optimized for a wearable device.

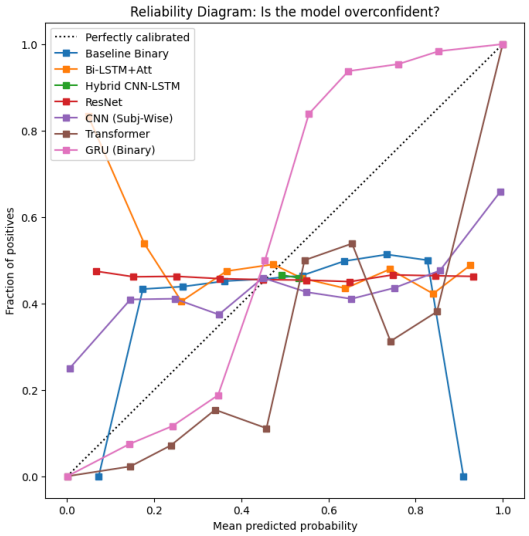
While the initial dataset had over 35 features, they were found to be prohibitive computationally for wearable hardware. This is why we reduced the input space firstly to 10 core HRV features, and then eventually settling for 5 (HR, MEAN_RR, SDRR, RMSSD and pNN50) for the prototype to ensure processing with low latency.

The process in development involved testing 10 distinct architectures. The earliest attempts which used baseline LSTMs yielded poor performance results (low recall). We then explored bidirectional LSTMs and xLSTM simulations, but ultimately in the end found that a hybrid CNN-GRU architecture provided the absolute best balance of both accuracy and efficiency:



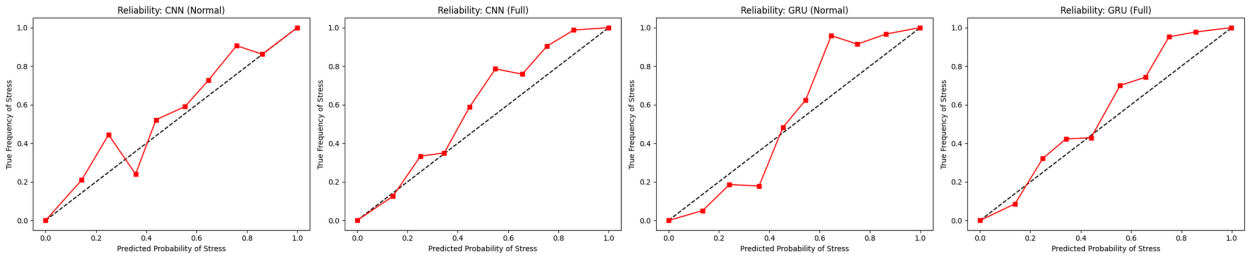
This bar chart compares the precision, recall and F1 score of models directly. GRU (Binary) and Transformer dominate near 1.0 across metrics, while CNN (Subj-Wise) while not perfect is the second best and the rest of the models are clustered much lower. Even though Transformer and GRU (Binary) are almost perfect, they are also computationally very heavy and less suited for efficient deployment on resource-constrained devices such as smartwatches.

In contrast the Hybrid CNN-GRU provides a strong balance between detection effectiveness and practicality for use in real time. The high recall gives certainty that the most stress events have been identified while remaining lightweight food continuous processing on smartwatches. In short, it provides the best trade-off between responsiveness and feasibility for our app.




This graph checks if the model's confidence matches reality. In this graph, we can see that the GRU is overconfident. This means for example that the model might say that it is 90% sure, but it is correct only 60% of the time. This can give false certainty. However, the CNN is calibrated better and is trustworthy.

In other words, the CNN solved the trust problem, because the GRU was overconfident and had a high chance of being incorrect, and the GRU solved the memory problem, because CNN, while trustworthy, is also forgetful, looking at a specific window of time. The GRU remembers the previous state. Therefore, by having both, we gained reliability from the CNN and memory from the GRU.




This calibration analysis further shows to us that while CNN and GRU individually perform reasonably well, each exhibits some weaknesses in different regions. For the CNN models, the predicted probabilities have fluctuations around the diagonal, with overestimation in lower probabilities and some underestimation in the middle. This indicates inconsistent confidence calibration.

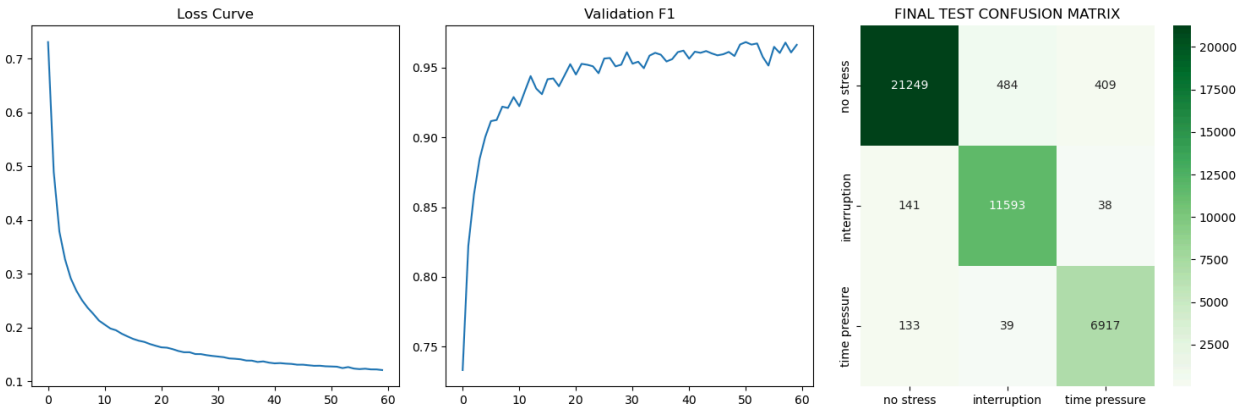
The GRU model that has normal features shows underestimation at the start and sharply transitions in the middle, while the GRU model with full features improves upon this but still shows overconfidence around intermediate stress levels.

Therefore, combining CNN and GRU in a hybrid architecture leads to smoother, more stable, more reliable results, and better calibrated predictions for real-world stress monitoring on a smartwatch. Through these tests we decided that the hybrid CNN-GRU route was our best option.

# 2 Model interpretability and explainability

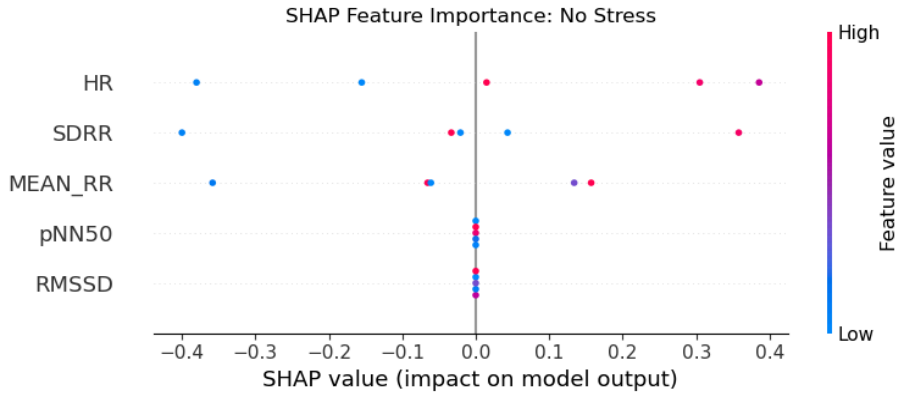
To understand our model's decision making process, we performed three checks:

## Confusion Matrix Analysis

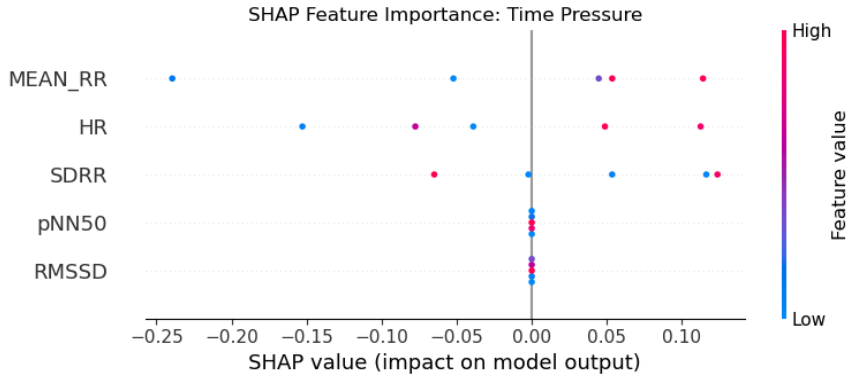



The confusion matrix validates the high precision of our CNN-GRU hybrid model. There is a strong diagonal trend, which confirms that the model correctly identifies the vast majority of time pressure and no stress states, achieving an overall F1-score of 0.97. By achieving these results on unseen test data, we have successfully demonstrated that our model generalizes well and is ready for a real-time detection on wearable devices.

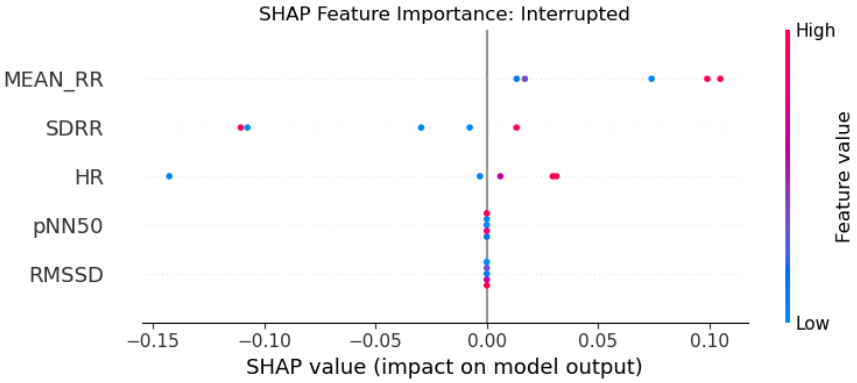
## SHAP Analysis



The SHAP analysis here shows that the CNN-GRU model mainly uses three core HRV features: HR, SDRR and MEAN_RR. The model uses them to identify non-stress states, providing interpretability and physiological validity for stress management on smartwatch based systems.




Here the SHAP analysis shows that the CNN-GRU model mainly uses HR, SDRR and MEAN_RR features for the time pressure class, identifying time-pressure stress, while pNN50 and RMSSD play a sort of secondary role, which confirms physiologically consistent and interpretable decision-making in the stress management system.

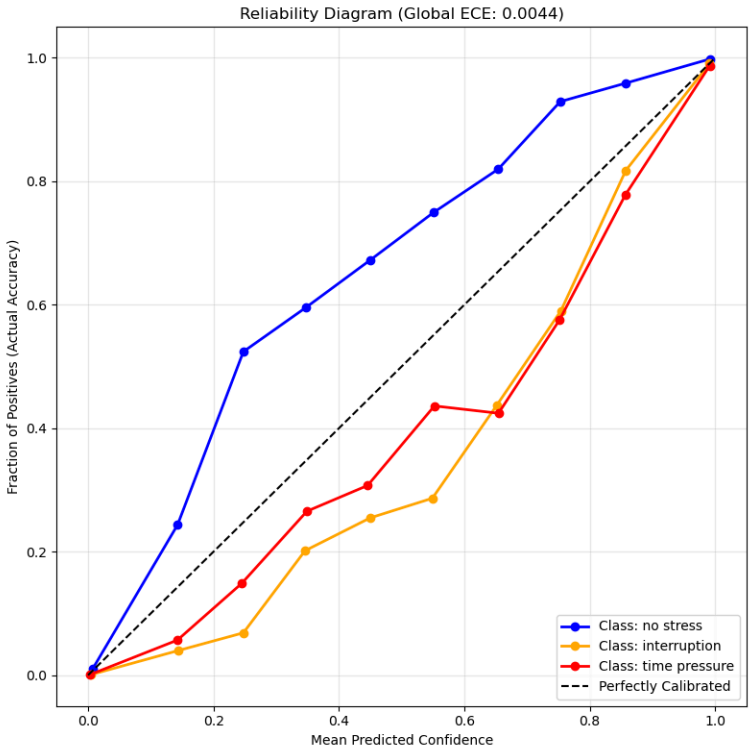



The SHAP analysis for the interrupted class indicates that the CNN-GRU model mainly relies on MEAN_RR, with secondary contributions from HR and SDRR. pNN50 and RMSSD have almost no impact. The mixed spread of positive and negative SHAP values reflects the irregular physiological response which is typically associated with interruptions.

## SHAP Summary

The SHAP analysis shows that our hybrid model consistently relies on the key physiological features, mainly HR, MEAN_RR and SDRR, to distinguish between the stress conditions. So, overall these results confirm that the model learns physiologically meaningful and interpretable representations, and this supports the model's reliability.

## Reliability Diagram




The reliability diagram shows how well the predicted confidence of the CNN-GRU model actually matches with the actual accuracy of classification. The low Global Expected Calibration Error (ECE = 0.0044) tells us that the model is calibrated great overall. The no stress class has slight under-confidence, while the interruption and time pressure classes show some over confidence at lower confidence levels. At higher confidence levels all classes closely follow the ideal calibration line, which shows the model' ability to provide reliable predictions when confident. This is very important for dependable stress monitoring particularly in a smartwatch application.

# 3 Technical limitations

Through evaluation, we found some key limitations that we should address:

- ➔ Our current model assumes a universal baseline. This means that a user with a naturally high resting heart rate or a highly athletic user with a low resting heart rate might receive drastically inaccurate readings.
- ➔ The version of Tizen on our smartwatches was too old for implementing the app onto the devices.

# 4 Biases and ethical considerations

There are some biases and ethical implications that should be discussed:

- ➔ There is a significant bias where physical exercise (ex. Running) looks identical to mental stress to our model.
- ➔ Users of the app must be informed that our app cannot distinguish between negative stress (distress) and positive excitement (eustress, eg. exercise).
- ➔ Real-time stress readings are carefully designed for not inducing additional anxiety with the alerts, which would create a feedback loop that worsens the physiological state of the user.
- ➔ Users must be informed that the physiological signals will be streamed to a remote dashboard in real-time before using the application.