# Multimodal machine-learning–based stress detection in a firefighter driving simulator

## Wielomodalne wykrywanie stresu z wykorzystaniem uczenia maszynowego w symulatorze wozu strażackiego

**EWELINA CHOŁODOWICZ**
**PAWEŁ LISIECKI***

DOI: https://doi.org/10.17814/mechanik.2026.1.2

**Stress is a major factor contributing to road-traffic accidents, highlighting the need for reliable and non-invasive methods to monitor driver stress. Existing research relies largely on Random Forest and XGBoost, whereas newer gradient-boosting algorithms such as LightGBM and CatBoost remain underexplored. This study develops a multimodal stress-detection pipeline using eye-tracking, electrodermal activity, HR/IBI, and driving-behaviour features collected in a firefighter emergency-driving simulator with controlled stress-inducing events. Four models–Random Forest, XGBoost, LightGBM, and CatBoost–were evaluated with comprehensive hyperparameter optimisation. LightGBM achieved the strongest overall performance, offering the highest precision and AUC, particularly in scenarios S2–S4, whereas S1 showed weak separability due to low physiological activation. Key predictive features included EDA phasic/tonic activity, pupil-dilation measures, speed dynamics, and steering variability. The results demonstrate that modern boosting methods combined with multimodal sensing provide robust and generalizable stress detection in operational emergency-driving conditions.**
**KEYWORDS: driver stress detection, wearable sensing, machine learning, electrodermal activity, eye tracking, hyperparameter tuning**

**Stres jest znaczącym czynnikiem prowadzącym do wypadków drogowych, co podkreśla potrzebę niezawodnych i nieinwazyjnych metod monitorowania poziomu stresu u kierowcy. Większość dotychczasowych badań wykorzystuje klasyczne algorytmy, takie jak Random Forest i XGBoost, podczas gdy nowsze algorytmy wzmacniania gradientowego, takie jak LightGBM i CatBoost, pozostają niedostatecznie zbadane w kontekście wykrywania stresu u kierowców na podstawie danych multimodalnych. W niniejszej pracy opracowano wielomodalny system detekcji stresu oparty na danych okulograficznych, elektrodermalnych, HR/IBI oraz telemetrycznych cechach zachowania kierowcy, zarejestrowanych w symulatorze jazdy pojazdu strażackiego wyposażonym w kontrolowane zdarzenia stresogenne. Oceniono cztery modele: Random Forest, XGBoost, LightGBM i CatBoost, poddając je optymalizacji parametrów. LightGBM osiągnął najlepszą ogólną skuteczność, uzyskując najwyższą precyzję i wartość AUC, zwłaszcza w scenariuszach S2–S4, podczas gdy w scenariuszu S1 separowalność klas była słaba ze względu na niską aktywność fizjologiczną. Do kluczowych cech predykcyjnych należały fazowa i toniczna aktywność EDA, miary poszerzenia źrenic, dynamika prędkości oraz zmienność kierowania. Uzyskane wyniki wskazują, że nowoczesne metody wzmacniania gradientowego w połączeniu z wielomodalnym systemem czujników, umożliwiają skuteczną detekcję stresu w warunkach jazdy interwencyjnej.**
**SŁOWA KLUCZOWE: detekcja stresu kierowcy, urządzenia ubieralne, uczenie maszynowe, aktywność elektrodermalna, okulografia, strojenie hiperparametrów.**

## Introduction

Stress is recognized as a major contributing factor in road-traffic incidents, with elevated driver stress levels linked to increased crash risk, injuries, and fatalities [13]. Consequently, reliable and non-invasive stress monitoring has become an important research focus for enhancing road safety. Current literature on physiological or behavioral stress detection predominantly employs classical machine learning models. Random Forest (RF) and XGBoost frequently emerge as top performers. For instance. [14] evaluated stress during emotion-evoking smartphone use using only RF and XGBoost. In studies on anxiety and stress classification [1], three widely used machine learning algorithms: Support Vector Machine (SVM), K-Nearest Neighbour (KNN), and RF were evaluated based on Galvanic Skin Response (GSR) signals. Additionally, [12] showed that RF outperformed both SVM and KNN in predicting anxiety, depression, and stress. By contrast, modern gradient-boosting frameworks such as LightGBM and CatBoost remain under-represented in stress-detection research. For instance, Nägeli et al. [11] noted that RF and SVM dominate earlier stress-detection research, while boosting-based methods remain uncommon. It was highlighted that LightGBM, in particular, has only recently begun to appear in this domain. Therefore, one of our contributions is to address this gap by exploring LightGBM and CatBoost not only with HRV and behavioral features, as in [11], but also by incorporating additional modalities such as EDA and eye-tracking signals. CatBoost appears even

* Dr inż. Ewelina Chołodowicz – cholodowicz.ewelina@gmail.com, https://orcid.org/0000-0001-8221-4309 – Autocomp Managment Sp. z o.o., Szczecin, Polska
mgr inż. Paweł Lisiecki – plisiecki@autocomp.com.pl, https://orcid.org/0009-0002-0713-4434 – Autocomp Management Sp. z o.o., Szczecin, Polska

more rarely, with only a few recent examples – such as Lalwani and Ferdowsi [8], who evaluated CatBoost (alongside RF, SVM, and XGBoost) for exam-related stress detection from wearable bio signals.

It is important to note that most stress-assessment studies are conducted in office, exam, or everyday-interaction contexts rather than controlled, repeatable stress-inducing driving scenarios. Structured stress-evoking events in a driving environment, particularly emergency-response scenarios, remain insufficiently explored. For example, in study [6] the volunteers underwent a series of tasks which were about solving some tests. Similarly, in another study [8] attempted to predict student exam performance from physiological stress signals. Although several studies have investigated stress and workload during simulated or real driving, the literature consistently highlights that driver-stress detection remains considerably under-explored, with only a limited number of multimodal or scenario-based investigations [9].

Beyond model selection and application, optimization of model settings also plays a crucial role in achieving reliable stress-detection performance. The benefits of hyperparameter optimization for improving stress-prediction accuracy are emphasized in recent work [15], reinforcing the need for systematic tuning rather than relying on default configurations. Addressing this requirement is also part of our contribution, as comprehensive hyperparameter optimization was applied to all evaluated models in this study to ensure fair comparison and maximize predictive performance.

This study addresses the aforementioned gaps by developing a predictive model capable of reliably distinguishing stress and no-stress states using GSR, eye-tracking features, HR/IBI, and simulator-derived behavioral data collected during firefighter-vehicle driving simulations. The proposed pipeline includes data preprocessing, feature extraction, hyperparameter optimization, and supervised classification. In addition to Random Forest, commonly cited as a strong baseline for stress-related tasks, we evaluate gradient-boosting models (XGBoost, LightGBM, CatBoost) to investigate whether these newer methods provide performance gains in multimodal stress detection during high-demand emergency-driving scenarios. Model performance is assessed using AUROC, precision, recall, F1 score, and accuracy.

This paper is organized as follows. Section 2 describes the experimental setup, including the simulator environment, stress-inducing events, and data acquisition. Section 3 outlines the methodology used for preprocessing, feature extraction, and model optimization. Section 4 presents and discusses the results. Section 5 concludes the work and outlines its contributions to advancing scenario-based research in driver stress detection tasks.

## Experimental setup

The study employed a Firefighter Driving Simulator to emulate realistic emergency driving conditions. A cohort of 29 adult volunteers was recruited through convenience sampling, spanning ages 25–50 years (mean driving-licence tenure ~20 years). The sample was predominantly male (59%) and none of the participants had prior exposure to this simulator platform. Most volunteers were mid-career professionals who contribute to the simulator programme itself (e.g., optics fabrication and simulator production teams, administrative staff, railway-route designers, and project managers), giving them above-average domain familiarity while still lacking hands-on simulator driving practice. Each participant completed two sessions that were recorded under identical logging and synchronization procedures:

A) Baseline drive – a hazard-free route used to characterize each participant's tonic physiological state and typical gaze behavior.

B) Stress session – a route containing predefined hazardous events intended to elicit acute stress responses.

### *Driving simulator*

The firefighter driving simulator is a computer-controlled training environment that reproduces the cabin, controls, and telemetry of an emergency vehicle. It was developed by *Autocomp Management Ltd.* at the request of the State Fire Service to train firefighters in safe and correct driving during emergency and high-pressure situations. The system offers configurable scenarios such as urban traffic, obstacles, and high-speed response routes, providing realistic steering, braking, and motion feedback. It enables performance evaluation and testing under controlled, repeatable, and risk-free conditions.

### *Stress-inducing events*

During the stress session, predefined hazardous driving scenarios were presented within the simulator. Each event was software-triggered and its onset timestamp was recorded enabling precise temporal alignment with the multimodal physiological signals. The scenario set was designed to reflect safety-critical situations that emergency vehicle drivers may encounter in urban traffic. Each scenario represented a distinct, safety-critical traffic event, including sudden pedestrian incursions (S1), abrupt barrier closures at crossings (S2), falling-branch obstacles triggered upon entering a predefined zone (S3), pedestrians entering the roadway within a marked area (S4), and right-of-way violations by another vehicle (S5).

### *Data acquisition*

Multimodal physiological and behavioural data were acquired from four synchronized data streams: (A) eye tracking, (B) electrodermal activity, (C) cardiac activity, and (D) vehicle telemetry. All recordings were obtained during the simulated driving task. The characteristics of each modality are summarized below:

A. Eye tracking – Pupil Labs Pupil Core headset

The binocular Pupil Core headset recorded pupil-diameter estimates together with continuous pupil-detection confidence values (0–1 scale), at an effective sampling rate of approximately 60 Hz. Calibration was performed according to the manufacturer's procedure described in [17].

B. Electrodermal Activity (EDA) – Shimmer3 GSR+

EDA was captured using a Shimmer3 GSR+ sensor with two electrodes attached to the palmar surfaces of the participant's non-dominant hand. The device recorded skin-conductance signals at approximately 51 Hz, following standard manufacturer configuration and calibration guidelines [18].

C. Cardiac activity – Garmin Fenix 7 Solar Pro

Heart Rate (HR) and Inter-Beat Intervals (IBI) was monitored with a PPG-based wearable. Instantaneous heart rate was recorded at 1 Hz, and inter-beat intervals (IBI) were extracted as beat-timed events from the device's photoplethysmographic (PPG) signal.

D. Vehicle telemetry – Simulator measurements

Vehicle telemetry included vehicle speed, steering angle, and brake-pedal state, recorded at an effective update rate of approximately 20 Hz.

Fig. 1 presents a participant equipped with all sensors A-C during driving using simulator described in chapter 2.1.

## Methods

In this section, the methods used for data processing, feature extraction, hyperparameter tuning, and the machine-learning classification task were outlined.


Fig. 1. Setup with all sensors and simulator screens
Rys. 1. Konfiguracja ze wszystkimi sensorami i ekranem symulatora

1. Data processing

As a prerequisite comprehensive data preprocessing to clean, align, and standardize all sensor streams, forming a reliable foundation for feature extraction and model development.

1.1. Eye-tracking data

Pupil Labs Pupil Core eye-tracking logs were preprocessed according to the staged artefact-handling procedure presented in [7]. Samples were initially marked as valid when pupil-detection confidence exceeded 0.85 and pupil diameter lay within the physiological range 1.5–9.0 mm. Artefact rejection then removed out-of-range diameters and isolated single-sample islands, applied a median absolute deviation (MAD)–based dilation-speed filter. Short gaps (<=0.3 s) within valid segments were reconstructed using shape-preserving Piecewise Cubic Hermite Interpolating Polynomial interpolation, while longer gaps remained missing. The cleaned trajectories were then smoothed with a Savitzky–Golay filter (0.38 s window, polynomial order 3) to attenuate sensor noise without introducing temporal lag [5].

1.2. Electrodermal activity data

Electrodermal activity signals were processed to extract both tonic and phasic components relevant to psychophysiological stress analysis. Signal preprocessing was performed using the NeuroKit2 Python library [16]. The EDA signal was subjected to artefact-aware cleaning to suppress motion-induced and technical noise. Subsequently, convex optimization–based decomposition (cvxEDA) was applied to separate the slow-varying tonic component, reflecting baseline skin conductance, from the fast-varying phasic component, which captures event-related skin conductance responses (SCRs). The algorithm also automatically detected and quantified individual SCR peaks, providing measures of their amplitude and timing.

1.3. Heart Rate (HR) and Inter-Beat Intervals (IBI)

The IBI series was sorted by time and screened for physiological plausibility, retaining only intervals between 300 ms and 2000 ms, with additional median – and MAD-based outlier rejection for PPG-derived data. Missing or rejected intervals were reconstructed by time-based interpolation. Instantaneous heart rate was calculated by converting each inter-beat interval (IBI) value to beats per minute. For alignment with the multimodal dataset, heart rate was further down sampled to the 10 Hz analysis grid and smoothed using a second-order Butterworth low-pass filter with a 0.5 Hz cutoff.

1.4. Vehicle telemetry

Vehicle telemetry data, including steering angle, speed, and brake pedal position, were first resampled to a uniform 10 Hz temporal grid to ensure alignment with other modalities. Short gaps in the data were linearly interpolated to maintain continuity. A simple moving average filter with a window of 3 to 5 samples was applied to each signal to reduce mechanical jitter while preserving rapid control changes.

2. Feature extraction

In this section the features extracted from four physiological and behavioral data streams were described. The set of extracted features is presented in Table 1. Below, the rationale behind selecting each group of features is outlined. Pupil size, mean pupil dilation and blink rate were included due to their strong and repeatedly shown to vary systematically between no-stress and stress states, providing measurable separation between the two conditions [3], [10]. What's more, both the total blink count and the blink rate (blinks per minute) were computed.

For EDA, features including mean, standard deviation, minimum, maximum, and peak-to-peak amplitude were extracted using the PyEDA toolkit [2], following established practice for representing tonic and phasic components of skin conductance [1]. Vehicle telemetry features (mean and standard deviation of speed, steering variability, brake-pedal behavior) followed the feature design applied in previous driver stress-detection study [10].

Heart rate (HR) and inter-beat intervals (IBI) were used to compute time-domain heart-rate-variability (HRV) indicators such as RMSSD and pNN50, which are widely recognized as reliable markers of autonomic regulation under stress. These HRV indicators are widely used in stress detection research and are highlighted as reliable physiological markers in [4].

All features were resampled and time-aligned with the multimodal dataset, enabling direct integration into stress classification models that are described in sections 3 and 4.

3. Hyperparameters tuning

Bayesian optimization was applied to tune model hyperparameters presented in tab. II with respect to the F1 score, which served as the primary performance criterion. All tuning was performed on the training set (dataset split described in chapter 4) after preprocessing, feature extraction steps (described in chapter 1 and 2) and data split. Like the rationale behind the feature set, the learning algorithms were selected based on their prevalence in stress and workload-detection studies. Traditional ensemble methods such as Random Forest and XGBoost are commonly used, while LightGBM and CatBoost have been explored less frequently in driving-related stress research. Therefore, the following models were chosen for presented research:

- XGBoost (Extreme Gradient Boosting) – boosted decision trees, where each tree corrects residual errors of the previous one,
- RF (Random Forest) – an ensemble of decorrelated trees trained on bootstrap samples,
- CatBoost (Categorical Boosting) – gradient boosting with ordered boosting,
- LightGBM (Light Gradient Boosting Machine) – a gradient boosting framework optimized for speed and leaf-wise tree growth.

Bayesian optimization was run separately for each algorithm. The resulting optimal configurations obtained in the driving-simulator task are shown in tab. II.

4. Machine learning

After completing all preprocessing (chapter 1) and feature extraction procedures (chapter 2), chosen machine-learning models (chapter 3) were developed to classify driver stress states from the resulting feature-based datasets. All models were trained using hyperparameters obtained in the dedicated tuning stage (section 3), ensuring consistent optimization across algorithms. Because the dataset exhibited a strong imbalance between stress and no-stress samples, algorithm-specific mitigation strategies were applied.

**Table I. Overview of features used for stress classification, grouped by sensor modality**
**Tabela I. Przegląd cech używanych do klasyfikacji stresu, pogrupowanych według rodzaju czujników**

| Sensor category | Features |
|---|---|
| EDA / GSR | Mean and standard deviation of tonic skin conductance over 16 s and 60 s windows, maximum tonic level (16 s) and SCR peak count and amplitudes (min/max), phasic EDA component, count of SCR peaks and distribution of SCR amplitudes, including minimum and maximum within 16 s windows |
| Vehicle telemetry | Vehicle speed (mean, variability; 16–60 s windows); steering angle (mean, variability; 16–60 s), brake pedal position (mean, variability; 16–60 s) |
| Eye tracking | Pupil size (mean, deviation from baseline, filtered, % of baseline; 5–15 s windows), pupil dilation (mean, maximum, outlier scores per eye and combined), blink (rate per eye and combined over 5–15 s, total blink count) |
| Cardiac (HR / IBI) | Pulse rate from IBI (filtered), short-window HRV indices over 60 s (total power, very-low-frequency power, SDNN, RMSSD, pNN50, average NN interval) |

**Table II. Obtained optimized parameters for the considered machine-learning algorithms in the driving-simulator task**
**Tabela II. Uzyskane zoptymalizowane parametry dla rozważanych algorytmów uczenia maszynowego w zadaniu z symulatorem jazdy**

| Model | Tuned parameters |
|---|---|
| CatBoost | iterations=800, learning_rate=0.03, depth=6, l2_leaf_reg=7.0 |
| LightGBM | n_estimators=800, learning_rate=0.06, num_leaves=63, min_child_samples=40 |
| XGBoost | n_estimators=600, learning_rate=0.07, max_depth=4 |
| Random Forest | n_estimators=800, max_depth=5, max_features=0.3, min_samples_split=10, min_samples_leaf=4 |

These included class-weight adjustments and, where appropriate, oversampling of the minority class, ensuring that misclassification of stress samples incurred a higher penalty during training. Hyperparameter search relied on cross-validation with participant-wise grouped folds, ensuring that all data from a given participant appeared exclusively in either the training or testing subset. To prevent information leakage, the data were organized into driving-sessions-level samples and then split by driving session into training and testing subsets. Approximately 90% of sessions were assigned to the training set and the remaining 10% to the testing set, yielding 281 795 fully processed samples derived from 132 simulator-driving sessions. The split was performed after all preprocessing steps, so the reported sizes correspond to the final ML-ready dataset. Model development, hyperparameter search, final training, and evaluation were executed on a workstation equipped with an NVIDIA GeForce RTX 5070 Ti GPU, 64 GB RAM, and an AMD Ryzen 7 9800X3D CPU.

## Results and discussion

This section presents the comparative performance of the four machine-learning algorithms evaluated for driver stress detection in a firefighter driving simulator. The models included LightGBM, CatBoost, XGBoost, and Random Forest. Overall metrics (accuracy, precision, recall, F1, macro-averages, and ROC AUC) are summarized in tab. III, while tab. IV provides scenario-specific results for S1–S5. Fig. 3 shows the ROC curves for all algorithms, and fig. 2 presents the stress-class precision across scenarios S1–S5.

Tab. III summarises the overall classification performance of the evaluated algorithms on the test dataset. LightGBM achieves the highest precision for detecting the stress class, exceeding CatBoost by 3.1%, while XGBoost is very close with only a 0.36% lower precision. The accuracy differences within the boosting group are similarly small – remaining below 0.6% – which indicates no substantial variation in accuracy among these algorithms. In contrast, LightGBM has 3% higher accuracy than Random Forest, demonstrating a clear performance advantage. Although LightGBM shows lower recall than Random Forest, its F1-score remains higher (as do the F1-scores of XGBoost and CatBoost), indicating that the boosting models achieve a more favorable balance between correctly identifying stress class and avoiding false positives. Overall, the gradient-boosting algorithms show strong and closely aligned performance, with LightGBM providing the best overall metric profile. What is more, the macro-averaged precision, recall, and F1 scores remain high for both classes, indicating that the boosting models – and LightGBM in particular – perform well not only on the majority no-stress class but also on the minority stress class. By comparison, Random Forest exhibits substantially lower macro-precision (by 6.5–7.9%) and macro-F1 (by 4.7–5.2%).

Tab. IV summarizes the per-scenario precision for detecting the stress class across all algorithms. In the stress-inducing scenarios (S2–S4), the gradient-boosting models consistently achieve high precision, with values exceeding 0.72 in all cases, indicating that stress-labelled predictions are rarely false alarms. LightGBM attains the highest precision in each of these scenarios. XGBoost performs comparably in S3 and S4, while CatBoost remains slightly lower while remaining competitive precision (0.9194 and 0.8508), confirming that all boosting models maintain robust performance in scenarios that are associated with elevated stress responses. It is evident that scenario S1 remains the most challenging for every algorithm: precision ranges from 0.2577 to 0.4774, reflecting sparse positive samples and label noise in this low-intensity condition.

The scenario-wise accuracy shown in fig. 2 illustrates how effectively each algorithm distinguishes between stress and non-stress states under the operational demands imposed by events S1-S5. The weakest

**Table III. Overall classification performance metrics per algorithm on the test dataset**
**Tabela III. Ogólne metryki skuteczności klasyfikacji dla poszczególnych algorytmów na zbiorze testowym**

| Algorithm | Accuracy | Precision | Recall | F1 | Precision (macro) | Recall (macro) | F1 (macro) | ROC AUC |
|---|---|---|---|---|---|---|---|---|
| CatBoost | 0.8935 | 0.7350 | 0.7427 | 0.7389 | 0.8347 | 0.8373 | 0.8360 | 0.8700 |
| LightGBM | 0.8993 | 0.7660 | 0.7247 | 0.7448 | 0.8485 | 0.8342 | 0.8410 | 0.8805 |
| XGBoost | 0.8978 | 0.7624 | 0.7204 | 0.7408 | 0.8461 | 0.8316 | 0.8386 | 0.8715 |
| Random Forest | 0.8501 | 0.6024 | 0.7671 | 0.6748 | 0.7694 | 0.8192 | 0.7887 | 0.8495 |

**Table IV. Per-scenario precision scores in stress detection class for each algorithm on the test dataset**
**Tabela IV. Wyniki precyzji (precision) dla wykrywania stresu w poszczególnych scenariuszach, dla każdego algorytmu, na zbiorze testowym**

| Algorithm | S1 | S2 | S3 | S4 | S5 |
|---|---|---|---|---|---|
| CatBoost | 0.2577 | 0.7267 | 0.9194 | 0.8508 | 0.7606 |
| LightGBM | 0.2931 | 0.8659 | 0.9283 | 0.8888 | 0.7789 |
| XGBoost | 0.4774 | 0.7609 | 0.9224 | 0.8580 | 0.7358 |
| Random Forest | 0.2500 | 0.6787 | 0.7129 | 0.7936 | 0.7590 |

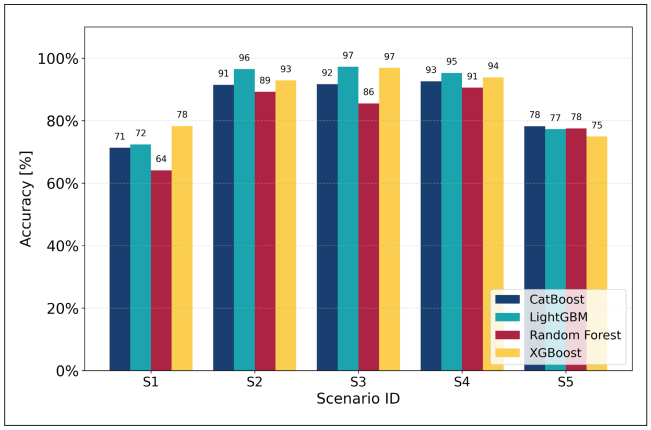


Fig. 2. Scenario-wise classification accuracy in stress detection task on test dataset
Rys. 2. Dokładność klasyfikacji w zadaniu wykrywania stresu dla poszczególnych scenariuszy – wyniki na zbiorze testowym

performance is observed in S1 and S5**,** indicating that these scenarios provide poor separability between baseline and stressed conditions. In contrast, all algorithms display a significant increase in accuracy in S2-S4. What is more, S3 yields the highest accuracy for the boosting models, reaching 92-97%. Going deeper, LightGBM and XGBoost lead across scenarios S1-S4, reflecting their strong capability to learn from diverse multimodal inputs. Random Forest remains the weakest, especially in S1-S4. Although XGBoost shows slightly lower accuracy in S5 than S1, S1 remains the least reliable scenario overall because its recall and F1 scores are extremely low, indicating that the model correctly identifies very few stress instances.

The ROC curves in fig. 3 align with the metrics reported in tab. III, confirming that LightGBM provides the most reliable and robust overall classification performance, as reflected by its highest AUC. The ROC curves of XGBoost and CatBoost remain very close to that of LightGBM across most FPR intervals, and their AUC values differ only marginally, indicating highly comparable discriminative behavior among the gradient-boosting models. Nevertheless, all boosting-based approaches clearly outperform the Random Forest, whose ROC curve lies noticeably closer to the random baseline, demonstrating its weaker ability to separate the two classes.

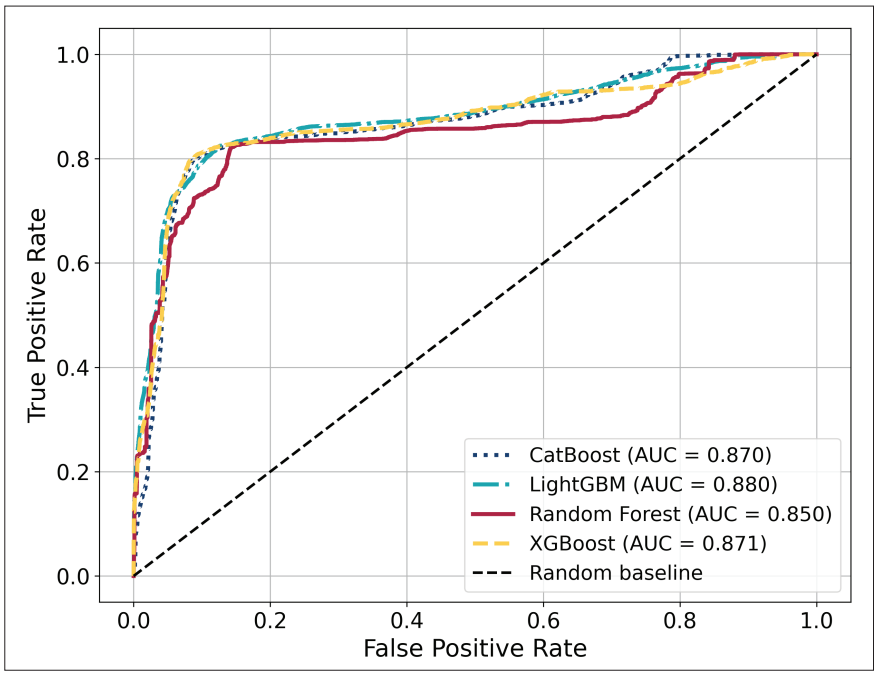


Fig. 3. Receiver Operating Characteristic (ROC) Curves with Area Under the Curve (AUC) Values for given stress detection tasks
Rys. 3. Krzywe ROC (Receiver Operating Characteristic) wraz z wartościami AUC (Area Under the Curve) dla danych zadań wykrywania stresu

## Conclusions

This study demonstrated that multimodal machine-learning models can reliably distinguish stress from no-stress states in a firefighter emergency-driving simulator. By integrating eye-tracking, EDA, HR/IBI, and vehicle-behavior features within a unified pipeline that included preprocessing, feature extraction, and comprehensive hyperparameter optimization, the work addressed several gaps identified in the literature – particularly the limited use of modern gradient-boosting methods and the scarcity of controlled, stress-inducing driving scenarios.

Across all evaluated algorithms, gradient-boosting models consistently outperformed Random Forest. LightGBM provided the strongest overall performance, yielding the highest precision, F1, and AUC. Scenario-level analysis confirmed that S2–S4 produced clear physiological separation between stress and baseline, enabling precision values exceeding 85% for the boosting models in S3–S4. In contrast, S1 remained the most challenging, reflecting its mild workload and weak physiological activation, which aligns with prior findings that low-intensity events yield limited separability. S3 produced the most stable stress-related multimodal patterns, enabling peak accuracy.

Overall, the results demonstrate that gradient boosting generalises well across diverse stress-inducing scenarios and that multimodal integration substantially enhances sensitivity to stress-related physiological and behavioural changes. The pipeline contributes to the field scenario-based evaluation, and multimodal sensing in a controlled emergency-driving context.

## Acknowledgements

The theses and results presented in the article are the outcome of the project “Simulator for Improving the Skills of Fire Service Emergency Vehicle Drivers and Studying the Impact of Drivers’ Psychophysical State During and After an Operation,” carried out by Autocomp Management under the Regional Operational Program of the West Pomeranian Voivodeship 2014–2020, Priority Axis 1:

Economy, Innovation, Modern Technologies, Action 1.1: R&D Projects of Enterprises, Project Type 2: R&D Projects of Enterprises Aimed at Implementing R&D Results into Business Activities.

REFERENCES

[1] Al-Nafjan, A., Aldayel, M.: *Anxiety Detection System Based on Galvanic Skin Response Signals*, Appl. Sci., vol. 14, no. 10788, 2024.
[2] Aqajari, S.A.H. et al.: *PyEDA: An Open-Source Python Toolkit for Pre-Processing and Feature Extraction of Electrodermal Activity*. In: Procedia Computer Science. pp. 99–106 (2021).
[3] Baltaci, S., Gokcay, D.: *Stress detection in human-computer interaction: Fusion of pupil dilation and facial temperature features*, Int. J. Human--Computer Interact., vol. 32, no. 12, pp. 956–966, 2016.
[4] Dalmeida, K.M., Masala, G.L.: *HRV features as viable physiological markers for stress detection using wearable devices*, Sensors, vol. 21, no. 8, pp. 2873, 2021.
[5] Gesztesi, G., Pajkossy, P.: *Wink or blush? Pupil-linked phasic arousal signals both change and uncertainty during assessment of changing environmental regularities*, Cognition, vol. 264, pp. 106256, 2025.
[6] Iqbal, Talha and Simpkin, Andrew J and Roshan, Davood and Glynn, Nicola and Killilea, John and Walsh, Jane and Molloy, Gerard and Ganly, Sandra and Ryman, Hannah and Coen, E. and others: *Stress Monitoring Using Wearable Sensors: A Pilot Study and Stress-Predict Dataset*, 2022.
[7] Kret, M.E., Sjak-Shie, E.E.: *Preprocessing pupil size data: Guidelines and code*, Behav. Res. Methods, vol. 51, no. 3, pp. 1336–1342, 2019.
[8] Lalwani, S., Ferdowsi, S.: *Predictive Modelling of Exam Outcomes Using Stress-Aware Learning from Wearable Biosignals*, Sensors, 2025.
[9] Mateos-Garc\'\ia, N. et al.: *Driver stress detection from physiological signals by virtual reality simulator*, Electronics, vol. 12, no. 10, pp. 2179, 2023.
[10] Mou, L. et al.: *Driver stress detection via multimodal fusion using attention-based CNN-LSTM*, Expert Syst. Appl., vol. 173, pp. 114693, 2021.
[11] Naegelin, M. et al.: *An interpretable machine learning approach to multimodal stress detection in a simulated office environment*, J. Biomed. Inform., vol. 139, pp. 104299, 2023.
[12] Priya, A. et al.: *Predicting anxiety, depression and stress in modern life using machine learning algorithms*, Procedia Comput. Sci., vol. 167, pp. 1258–1267, 2020.
[13] Rastgoo, M.N. et al.: *Automatic driver stress level classification using multimodal deep learning*, Expert Syst. Appl., 2019.
[14] Rupp, Lydia Helene and Kumar, Akash and Sadeghi, Misha and Schindler-Gmelch, Lena and Keinert, Marie and Eskofier, Bjoern M and Berking, M.: *Stress can be detected during emotion-evoking smartphone use: a pilot study using machine learning*, Front. Digit. Heal., vol. 7, pp. 1578917, 2025.
[15] Suraj Arya, A., Ramli, N.A.: *Predicting the stress level of students using supervised machine learning and artificial neural network (ANN)*, Indian J. Eng., vol. 21, pp. e9ije1684, 2024.
[16] *NeuroKit2 Python library*, https://neurokit2.readthedocs.io/.
[17] *Pupil Labs*, https://docs.pupil-labs.com/core/ .
[18] *Shimmer Sensing*, https://shimmersensing.com/wp-content/docs/support/documentation/GSR_User_Guide_rev1. 13. pdf .