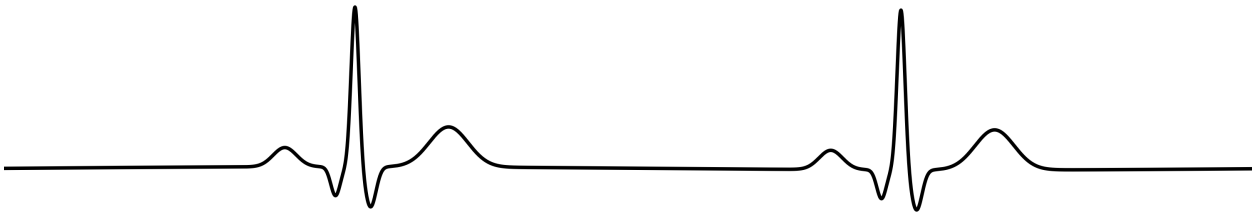

Detection of Alcohol-Induced Driving Impairment Using Heart Signal Monitoring and Machine Learning

Master's thesis in Biomedical Engineering

Teodor Svensson
Viggo Eriksson

DEPARTMENT OF MECHANICS AND MARITIME SCIENCES

CHALMERS UNIVERSITY OF TECHNOLOGY
Gothenburg, Sweden 2026
www.chalmers.se

Master's Thesis in Biomedical Engineering

Detection of Alcohol-Induced Driving Impairment Using Heart Signal Monitoring and Machine Learning

Teodor Svensson
Viggo Eriksson

CHALMERS
UNIVERSITY OF TECHNOLOGY

Department of Mechanics and Maritime Sciences
Division of Vehicle Safety
Chalmers University of Technology
Gothenburg, Sweden 2026

Detection of Alcohol-Induced Driving Impairment Using Heart Signal Monitoring and Machine Learning
Teodor Svensson
Viggo Eriksson

Supervisor: Johan Stjernholm, Sightic (Smart-eye)
Examiner: Jonas Bärgman, Department of Mechanics and Maritime Sciences

Master's Thesis 2026
Department of Mechanics and Maritime Sciences
Chalmers University of Technology
SE-412 96 Gothenburg
Sweden
Telephone +46 31 772 1000

Cover: Python-generated visualization of two PQRST electrocardiogram curves.

Typeset in LaTeX
Gothenburg, Sweden 2026

Detection of Alcohol-Induced Driving Impairment Using Heart Signal Monitoring and Machine Learning
Teodor Svensson
Viggo Eriksson
Department of Mechanics and Maritime Sciences
Division of Vehicle Safety
Chalmers University of Technology

Abstract

Alcohol-related driving impairment remains a significant safety challenge, contributing to approximately 25% of traffic-related fatalities in Europe. This master's thesis investigates whether heart signal monitoring, in the form of heart rate (HR) and heart rate variability (HRV), together with machine learning, can detect alcohol-induced driver impairment. A within-subject experimental study was conducted with 37 participants in a controlled driving simulator under three conditions: sober, low-to-moderate blood alcohol concentration (BAC), and high-BAC ($\geq$0.7‰). Additionally, an external real-world driving dataset from 17 participants (sober and high-BAC) was used to evaluate model generalizability. Although a contactless 24 GHz radar-based heart signal monitoring system was initially evaluated, the radar-derived signals lacked the reliability needed for robust HRV analysis in this thesis. Therefore, the main classification pipeline utilized reference data from a Polar H10 chest-worn sensor. Several machine learning classifiers, including Support Vector Machines (SVM) and Extreme Gradient Boosting (XGBoost), were evaluated using leave-one-subject-out cross-validation. Tested on the real-world dataset, the models achieved a session-level area under the receiver operating characteristic curve (AUC) of 0.75. The best-performing model evaluated on simulator data reached a session-level AUC of 0.81 in the high-BAC setting. An exploratory late-fusion approach that combines simulator-based heart signals and camera-based driver monitoring system (DMS) outputs further improved performance, achieving an AUC of 0.93 in the high-BAC setting. These findings highlight the potential of heart signal monitoring to complement camera-based DMS, thereby supporting multimodal impairment-detection approaches aligned with Euro NCAP 2026 safety requirements.

Keywords: Alcohol driving impairment, HRV, Machine learning, real-world, DMS fusion.

Preface

This report presents the outcome of a master's thesis project carried out at the Department of Mechanics and Maritime Sciences at Chalmers University of Technology during the spring of 2026. The work was conducted in collaboration with Sightic Analytics and was initially carried out at their office in Gothenburg. During the course of the project, Sightic Analytics was acquired by Smart Eye, and from March 2026 the work was instead carried out at the Smart Eye office in Gothenburg.

Acknowledgements

We would like to thank our examiner and supervisor, Jonas Bärgman, for his guidance and support throughout this project. We would also like to thank our supervisor at Sightic Analytics, Johan Stjernholm, for always being available and for assisting us with any questions that arose during the work.
Finally, we would like to thank all participants who took part in the data collection events and contributed valuable data to this project.
Their contributions were essential for making this study possible.

Teodor Svensson, Viggo Eriksson, Gothenburg, June 2026

List of Acronyms

The following is the list of acronyms that have been used throughout this thesis, listed in alphabetical order:

ANS	Autonomic Nervous System
ANOVA	Analysis of Variance
AUC	Area under the receiver operating characteristic curve
BAC	Blood Alcohol Concentration
BPM	Beats per minute
BR	Breathing Rate
CV	Cross-Validation
DMS	Driver Monitoring System
ECG	Electrocardiogram
EDA	Electrodermal Activity
FMCW	Frequency-Modulated Continuous Wave
FN	False Negative
FP	False Positive
GB	Gradient Boosting
GDPR	General Data Protection Regulation
HF	High-Frequency
HistGB	Histogram-Based Gradient Boosting
HR	Heart Rate
HRV	Heart Rate Variability
IBIs	Inter-Beat Intervals
KNN	K-Nearest Neighbors
LF	Low-Frequency
LOSO	Leave-One-Subject-Out
LR	Logistic Regression
ML	Machine Learning
mmWave	Millimeter-wave
NN	Normal-to-Normal Interval
pNN50	Percentage of successive RR interval differences greater than 50 ms
PNS	Parasympathetic Nervous System
PPG	Photoplethysmogram
RBF	Radial Basis Function
RF	Random Forest

RMSSD	Root Mean Square of Successive Differences
ROC	Receiver Operating Characteristic
RR	R–R Intervals (time between consecutive R-peaks)
SDNN	Standard Deviation of RR Intervals
SDSD	Standard deviation of successive RR-interval differences
SNS	Sympathetic Nervous System
SVM	Support Vector Machines
TN	True Negative
TP	True Positive
XGBoost	Extreme Gradient Boosting

Nomenclature

Below is the nomenclature of mathematical symbols that have been used throughout this thesis.

Indices

n	Index of cardiac interval

Sets

–	No mathematical sets are explicitly defined in this thesis

Parameters

N	Total number of intervals within the analysis window
$NN50$	Number of successive NN interval pairs differing by more than 50 ms
k	Number of folds in k-fold cross-validation or number of nearest neighbors in KNN, depending on context

Variables

RR_n	nth RR interval; time between consecutive R-peaks
NN_n	nth normal-to-normal interval; artifact-corrected RR interval
$\overline{RR}$	Mean RR interval
$\overline{NN}$	Mean NN interval
RR_{n+1}	RR interval following RR_n

Contents

List of Figures

List of Tables

1

Introduction

This is a master's thesis within the master's program in Biomedical Engineering at Chalmers University of Technology. The project investigates the feasibility of using heart signal monitoring to detect alcohol-related driver impairment, with an initial focus on whether such monitoring can be performed using radar-based technology. The main physiological analysis was based on reference heart rate (HR) and heart rate variability (HRV) data collected during both simulator-based and real-world driving, while radar data were collected and evaluated in the simulator setting. The degree project was conducted at the company *Sightic* in collaboration with *Emsense Technology*, which provided the radar-based sensing technology used during data collection.

At the outset of the project, the radar system was intended to play a central role in the study. It was initially planned to serve as the primary source of physiological data. The system enables non-contact estimation of physiological parameters, such as HR and HRV, which are considered highly relevant to impairment detection.

However, as the project progressed, limitations in the usability of the radar-derived signals prevented their inclusion in the final classification pipeline. As a result, the radar component remained an important part of the original research motivation and experimental setup, while the final analysis focused on HR and HRV-based features extracted from reference physiological recordings.

Toward the end of the project, an additional exploratory comparison with camera-based driver monitoring system (DMS) data was included. This analysis was added to investigate how physiological features compare with camera-based behavioral indicators, and whether the two modalities may provide complementary information for alcohol-related impairment detection.

The report is structured as follows. Chapter 1 introduces the background, purpose, research questions, and scope of the study. Chapter 2 presents the theoretical foundation, including cardiovascular physiology, heart rate variability, radar-based heart signal monitoring, camera-based DMS, and machine learning (ML) methods. Chapter 3 describes the methodology, including the experimental design, data collection, preprocessing, feature extraction, dataset construction, and model evaluation. Chapter 4 presents the results, which are discussed in Chapter 5, and Chapter 6 concludes the work.

1.1 Background

Alcohol is a toxic psychoactive substance that adversely affects the central nervous system, leading to impaired cognitive function, reduced attention, and decreased psychomotor performance [1]. These effects can significantly compromise an individual's mental and physical ability to perform complex tasks safely.

Alcohol impairment is a major contributor to road traffic accidents and is estimated to account for approximately 25% of all traffic-related deaths and serious injuries in Europe [2].

To mitigate alcohol-impaired driving, emerging in-vehicle technologies aim to detect early signs of impairment by monitoring driver state and behavior. Current approaches rely on signals such as eye-closure patterns, head and eye movements from DMS, and monitoring of steering dynamics and overall driving patterns. For such systems, both accuracy and rapid detection are essential. However, factors such as lighting conditions, occlusions, and driver posture can negatively affect the performance of camera-based DMS. Incorporating additional sources of driver information therefore has the potential to improve robustness and reliability.

Physiological measures such as HR and HRV have been shown to be affected by acute alcohol ingestion. Several studies report a reduction in HRV following alcohol intake, often accompanied by an increase in HR, indicating altered autonomic regulation [3, 4, 5]. Although these studies differ in experimental protocols and alcohol dosage, their findings consistently demonstrate similar autonomic responses to alcohol consumption.

Since alcohol-related impairment can manifest itself in physiological signals, ML offers a potential way to detect such patterns automatically. A study on alcohol-impaired driving collected multiple physiological signals, including electrocardiogram (ECG), photoplethysmogram (PPG), and electrodermal activity (EDA), using non-invasive sensors with minimal driver restraint in a driving simulator [6]. The study evaluated a broad set of physiological features and showed promising results for detecting alcohol-related impairment using ML. However, the evaluation was limited to a simulator setting, and the findings were therefore not validated during real-world driving.

In addition, [7] investigated alcohol-impairment detection using ECG signals and reported promising classification performance. However, their experimental setup involved participants sitting with their hands on a steering wheel rather than performing an actual driving task. These previous studies therefore support the feasibility of detecting alcohol-related impairment from physiological measurements, but they also highlight the need for further evaluation in more realistic driving conditions.

For practical driver monitoring applications, it is also important to consider how model performance is balanced between detecting impaired states and avoiding false alarms. Unnecessary warnings could reduce user acceptance and trust in the system. This is particularly relevant since impairment detection is becoming an increasingly important part of modern DMS development. While DMS have traditionally focused on driver distraction and fatigue, recent assessment frameworks such as Euro NCAP's 2026 Safe Driving assessment also include non-fatigue-related impairments,

such as alcohol- and drug-induced impairment [8]. The Euro NCAP protocol allows a learning period of up to 10 minutes during driving before impairment must be detected.

The present study contributes to this by investigating HR- and HRV-based impairment detection using data collected during both simulated and real-world driving. This provides an opportunity to evaluate whether a reduced set of vital parameters is informative under more practically relevant conditions, while also considering the importance of false-alarm reduction in a potential driver-monitoring application.

Recent advances in radar-based vital sign monitoring now promise non-contact measurement of HR, HRV, and breathing rate (BR) [9, 10, 11]. These physiological indicators offer a fundamentally different perspective on driver state and may provide valuable complementary information for identifying alcohol-related impairment compared to camera-based DMS.

ML approaches have shown potential for interpreting radar-based vital sign data for driver-state monitoring applications [12]. Related work on radar-based drowsiness detection has evaluated ML classifiers using respiration-related radar data and reported promising classification performance [13, 14]. This suggests that radar-derived physiological information may be suitable for automated driver-state inference, although its application to alcohol-related impairment remains less explored.

Although these findings suggest that physiological signals such as HR and HRV may be useful for automated driver-state monitoring, they are not specific markers of alcohol-related impairment. This is important for the present study, since the classification models are based on HR- and HRV-derived features rather than alcohol-specific biomarkers.

HR and HRV are also influenced by several factors other than alcohol. For example, previous studies investigating the acute effects of caffeine on HRV have reported mixed results, with some studies observing a significant reduction in HRV following caffeine ingestion, while others found no clear association [15, 16, 17]. In contrast, both acute and long-term nicotine exposure have consistently been associated with a significant decrease in HRV, indicating altered autonomic regulation [18, 19]. Psychological stress has also been shown to negatively affect HRV, primarily through increased sympathetic activation and reduced parasympathetic activity, reflecting a shift in autonomic balance [20]. Similarly, drowsiness and fatigue have been associated with altered HRV patterns, indicating changes in autonomic nervous system regulation [21].

A relatively recent study investigated whether heart rate measured by a physician reflects a patient's true resting heart rate [22]. The authors referred to this phenomenon as white-coat heart rate and related it to the well-established concepts of white-coat syndrome and white-coat hypertension, which describe increases in blood pressure in clinical settings. The study concluded that resting HR measured by a physician during consultation does not necessarily reflect the patient's true resting heart rate. This is an important consideration in experiments involving HR measurements, since the clinical setting itself may influence the observed values.

1.2 Purpose

The purpose of this thesis is to investigate whether heart signal monitoring can be used to detect alcohol-induced impairment in drivers.

More specifically, the study evaluates whether HR- and HRV-based features can be used to classify impaired and unimpaired driving states using ML models.

An additional purpose is to evaluate the feasibility of radar-based heart signal monitoring for this application. Radar data were therefore processed and assessed as part of the project. However, due to limitations in the quality of the radar-derived heart signals, the final classification analysis was based on pulse band data.

Toward the end of the project, the scope was extended to include an exploratory comparison of camera-based DMS data to investigate whether physiological and camera-based features may provide complementary information for alcohol-induced impairment detection.

1.3 Overarching research questions

- Can HR- and HRV-based features be used to classify sober and alcohol-impaired driving states?

- To what extent can radar-derived heart signals be used as a non-contact alternative for extracting physiological features relevant to impairment classification?

- As an exploratory extension, how do models based on physiological features compare with models based on camera-based DMS data, and to what extent may these modalities complement each other for detecting alcohol-induced impairment?

At the outset of the project, radar-based vital sign monitoring was intended to be a central part of the study and was initially planned to serve as the primary source of physiological data. However, as the project progressed, limitations in the radar-derived signals meant that this data could not be used in the final classification models. The radar component therefore remained an important part of the original research motivation and data collection design, but the final analysis focused on HR- and HRV-based features extracted from the reference physiological recordings.

In the final stage of the project, an additional exploratory analysis was conducted to compare the physiological classification models with camera-based DMS models. Since this objective was introduced late in the project, it was not intended to replace the thesis's original focus, but rather to provide an initial investigation into how camera-based behavioral indicators and physiological features may complement each other in the context of alcohol-induced driver impairment detection.

1.4 Scope

This study is a proof-of-concept investigation into whether HR- and HRV-based features can be used to classify sober and alcohol-impaired driving states, and whether radar-based vital sign monitoring can support this application. The study is based on data collected from a limited group of participants during both simulator-based and real-world driving.
The scope of the thesis is limited to the experimental setup used in this project, including the available participant sample, the driving conditions, the fixed sensor placement, and the use of a chest-worn reference sensor. These factors should be considered when interpreting the generalizability of the results.
The comparison with camera-based DMS data was included as an exploratory extension in the final stage of the project. It was therefore limited to an initial comparison between physiological and camera-based models, rather than a full multimodal system development or validation.

2
Theory

This chapter presents the theoretical background needed to understand the use of physiological and behavioral signals for alcohol-induced driver impairment classification. The chapter begins by introducing alcohol-related impairment and its relevance to driver monitoring systems. It then describes how alcohol can affect the autonomic nervous system and cardiovascular regulation, with particular focus on HR, RR intervals, and HRV.
The chapter further introduces the HRV features used in this thesis. The basic principles of radar-based vital sign monitoring are then presented, followed by an introduction to camera-based DMS. Finally, the chapter describes the ML framework used for impairment classification, including feature selection, classification models, ensemble learning, cross-validation strategies, and evaluation metrics.

2.1 Alcohol-Induced Driver Impairment

Alcohol consumption can impair several functions that are important for safe driving, including attention, reaction time, decision-making, motor coordination, and risk assessment [1]. These effects can reduce a driver's ability to respond appropriately to changes in the traffic environment. In driver monitoring applications, alcohol-induced impairment is therefore relevant not only as a legal issue, but also as a measurable change in driver state that may be reflected in behavioral, physiological, or performance-related signals.
Alcohol-induced impairment is also becoming increasingly relevant in the context of modern DMS. Traditionally, DMS development has focused primarily on detecting driver distraction and fatigue. However, recent assessment frameworks indicate a broader shift toward monitoring impaired driver states more generally. For example, Euro NCAP, which produces vehicle safety ratings, includes non-fatigue-related impairment in its 2026 Safe Driving assessment. Within this assessment, DMS can receive credit for detecting impairment caused by alcohol and/or drugs [8]. This inclusion reflects the growing importance of in-vehicle systems capable of identifying safety-critical driver states beyond conventional distraction and drowsiness monitoring.

2.1.1 Blood Alcohol Concentration

Blood alcohol concentration (BAC) refers to the concentration of ethanol in the bloodstream and is commonly expressed in parts per thousand (‰). In practice,

alcohol levels are often estimated using breath alcohol measurements, where the concentration of alcohol in exhaled breath is measured and related to the corresponding blood alcohol concentration.
BAC is commonly used as an indicator of alcohol-related impairment. In Sweden, the legal limit for driving a motor vehicle is 0.2‰ BAC, corresponding to 0.10 mg of alcohol per liter of exhaled breath [23]. However, legal limits vary between countries; for example, the general limit is 0.8‰ in most of the United States and in England, Wales, and Northern Ireland [24, 25], while Scotland and most of Europe uses a lower limit of 0.5‰ [26].

2.1.2 Introduction to the Autonomic Nervous System

The human body is continuously regulated by the nervous system. To understand how alcohol-induced impairment affects physiological functions, it is important to first introduce the autonomic nervous system (ANS). The ANS is commonly divided into two main branches: the sympathetic nervous system (SNS) and the parasympathetic nervous system (PNS) [27]. These branches are continuously active and contribute to the regulation of involuntary bodily functions, including HR, blood pressure, respiration, and digestion. The PNS is primarily associated with "rest and digest" responses, supporting recovery, relaxation, and digestive processes. In contrast, the SNS is associated with the "fight or flight" response, preparing the body for stress, arousal, or physically demanding situations.

2.1.3 Effects of Alcohol on the Autonomic Nervous System

Acute alcohol intake can affect cardiovascular regulation by altering the balance between sympathetic and parasympathetic activity. Since HR is modulated by both branches of the ANS, changes in this balance can be reflected in HR and HRV. Increased sympathetic activity and/or reduced parasympathetic modulation generally lead to increased HR and reduced HRV.
Several studies have reported that acute alcohol ingestion is associated with increased HR and reduced HRV, indicating altered ANS regulation following alcohol consumption [3, 4, 5]. These physiological effects provide the rationale for using HR and HRV as potential indicators of alcohol-induced impairment in this thesis.

2.2 Cardiovascular Physiology

The electrocardiogram (ECG) is a non-invasive measurement of the heart's electrical activity and reflects the electrical processes that drive each cardiac cycle [28]. One heartbeat in an ECG signal can be divided into several characteristic waves, commonly denoted P, Q, R, S, and T. These can be seen in Figure 2.1.

2.2.1 RR Intervals

The RR interval is defined as the time difference between two consecutive R-peaks in an ECG signal. The R-peak is the most prominent feature of the ECG waveform

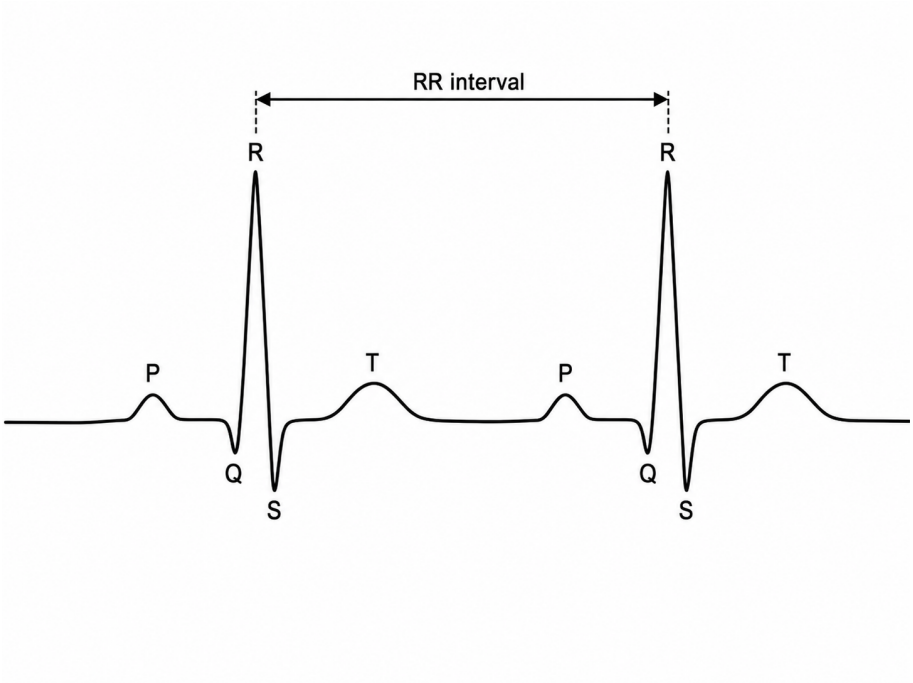

Figure 2.1: Illustration of the RR interval between two consecutive R-peaks in an ECG signal. The image was generated using ChatGPT.

and provides a robust reference point for identifying individual cardiac cycles. The RR interval, therefore, represents the duration between successive heartbeats. An illustration of the RR interval is shown in Figure 2.1.
The mean RR interval is defined as

$$\overline{RR} = \frac{1}{N} \sum_{n=1}^{N} RR_n \tag{2.1}$$

where N denotes the total number of intervals and RR_n represents the nth RR interval [29].
HR is inversely related to the RR interval, and RR intervals are commonly expressed in either milliseconds or seconds. When RR is expressed in seconds, the HR corresponding to a given RR interval can be calculated in beats per minute (bpm) as

$$HR = \frac{60}{RR}. \tag{2.2}$$

This inverse relationship implies that shorter RR intervals correspond to higher HR, whereas longer RR intervals indicate lower HR.
The sequence of consecutive RR intervals forms the basis for HRV analysis. However, HRV analysis is commonly based on normal-to-normal (NN) intervals rather than raw RR intervals. NN intervals refer to intervals between successive normal heartbeats after removal or correction of abnormal beats and measurement artifacts, such as missed beats, falsely detected beats, or signal disturbances. If not corrected, such artifacts may introduce artificial variability and distort HRV measures.

Since the RR interval data used in this thesis were artifact-corrected before HRV feature extraction, the term RR interval is used to refer to artifact-corrected NN intervals unless otherwise stated.

2.2.2 Heart Rate Variability

Although HR provides a measure of the average cardiac frequency, it does not capture the beat-to-beat fluctuations that occur between cardiac cycles. HRV refers to the variation in the time intervals between successive normal heartbeats and is therefore derived from RR intervals rather than from HR alone. Analysis of HRV provides more detailed information about cardiac regulation than average HR alone [30, 31].
As stated above, HRV reflects the ANS ability to adapt to internal and external demands. This adaptability is important both physically, for example during recovery after exercise when HR needs to decrease efficiently, and mentally, for example when returning to a relaxed state after a stressful situation. Since HRV can be influenced by external factors including alcohol, nicotine, caffeine, stress, and physical activity, these factors should be considered when interpreting HRV outcomes [31, 32].

2.3 Heart Rate Variability Features

In the context of ML, features are quantitative descriptors extracted from time series. This type of feature-based representation is commonly used for time-series classification, where statistical properties of a signal are transformed into structured inputs for data-driven models [33].
HRV features are typically computed over smaller segments of the RR interval time series. These segments are referred to as windows. Windowing is particularly important because classification algorithms require inputs of consistent size. By extracting HRV metrics from consecutive signal segments, the continuous RR interval series can be transformed into a sequence of feature vectors suitable for statistical analysis and data-driven modeling. To handle autocorrelation between nearby windows, various cross-validation (CV) methods can be implemented. These methods are explained in more detail in Section 2.6.5.
In this thesis, HRV-derived features are used to quantify short-term changes in cardiac regulation that may be related to alcohol-induced impairment during driver monitoring. HRV can be interpreted using several metrics in both the time and frequency domains. Commonly used HRV features include time-domain measures such as SDNN, RMSSD, and pNN50; frequency-domain measures such as low-frequency power (LF), high-frequency power (HF), and the LF/HF ratio; and Poincaré-plot-derived measures such as SD2 [30, 31]. These features are defined below.

2.3.1 Time-Domain Metrics

Time-domain metrics describe HRV directly from the RR interval time series, without transforming the signal into the frequency domain. These metrics quantify different aspects of the temporal variation between successive heartbeats, ranging

from average HR to overall variability and short-term beat-to-beat changes. In this section, the time-domain metrics used in this thesis are defined.
The mean HR represents the average HR over a specified time period, typically calculated from the corresponding mean RR interval. When the mean RR interval is expressed in seconds, mean HR can be calculated as

$$\overline{HR} = \frac{60}{\overline{RR}} \tag{2.3}$$

[29]
RMSSD is the square root of the mean of the squared differences between successive RR intervals. RMSSD reflects HRV and is widely considered the primary time-domain metric of parasympathetic activity [30, 31]. It also correlates strongly with HF-power and pNN50, and is considered robust for short-duration recordings due to its reduced sensitivity to slower HR trends. RMSSD is defined as

$$RMSSD = \sqrt{\frac{1}{N-1}\sum_{n=1}^{N-1}\left(RR_{n+1} - RR_n\right)^2} \tag{2.4}$$

[29]
where N denotes the number of intervals and RR_n represents the nth interval.
pNN50 is defined as the proportion of consecutive RR interval pairs that differ by more than 50 ms, expressed as a percentage. Similar to RMSSD, pNN50 reflects parasympathetic activity [30, 31]. pNN50 is calculated as

$$pNN50 = \frac{NN50}{N-1} \times 100\% \tag{2.5}$$

[29]
SDNN is another commonly used metric that represents the standard deviation of all RR intervals over the recording period. It reflects the overall variability in HR, capturing both branches of the ANS [30, 31]. SDNN is defined as

$$SDNN = \sqrt{\frac{1}{N-1}\sum_{n=1}^{N}\left(RR_n - \overline{RR}\right)^2} \tag{2.6}$$

[29]

2.3.2 Poincaré-Based Metric

In addition to standard time-domain HRV measures, SD2 was included as a Poincaré-based feature [31]. A Poincaré plot is created by plotting each RR interval against the following RR interval, i.e., RR_n against RR_{n+1}. The resulting point cloud can be described by an ellipse, where SD2 represents the standard deviation of the points along the line of identity and therefore reflects both parts of the ANS [31, 34].
SD2 can be expressed in relation to SDNN and the standard deviation of successive RR-interval differences, SDSD, according to [34], as

$$SD2 = \sqrt{2SDNN^2 - \frac{1}{2}SDSD^2} \tag{2.7}$$

where SDNN is the standard deviation of the RR intervals, and SDSD is the standard deviation of the successive differences between adjacent RR intervals. SDSD was not included as a separate feature in this thesis, since it captures short-term HRV similarly to RMSSD.

2.3.3 Frequency-Domain Metrics

HRV can also be analyzed in the frequency domain by decomposing the RR interval time series into its spectral components. For short-term recordings, the most commonly used frequency bands are the LF and HF components, which can typically be estimated from recordings of at least two minutes in duration [30, 31].
The LF band, typically defined as 0.04-0.15 Hz, reflects a combination of sympathetic and parasympathetic modulation. The HF band, typically defined as 0.15-0.40 Hz, is primarily associated with parasympathetic activity and reflects variations in HR related to the breathing cycle. The LF/HF ratio is a typically used frequency-domain HRV feature calculated from the relative contribution of the LF and HF components [30, 31].

2.3.4 Automated Feature Extraction

Automated feature extraction is particularly useful in ML applications, as it enables a systematic exploration of potentially informative features without requiring manual feature definition [33]. This can improve model performance by capturing subtle and potentially non-linear relationships in the data.
One example of an automated feature extraction package is tsfresh, a Python package for extracting features from time-series data [33]. Tsfresh applies a large set of predefined feature calculators to each time series, converting the signal into a fixed-length feature vector. The extracted features describe various properties of the signal, including statistical characteristics, temporal dependencies, frequency-related information, entropy, and temporal changes.
Tsfresh provides different feature extraction settings that control which feature calculators are applied and, consequently, the number and complexity of the extracted features. The comprehensive setting extracts a large feature set, while the efficient setting excludes computationally expensive calculators to reduce extraction time. A minimal setting is also available for extracting only basic features. The choice of setting represents a trade-off between feature richness and computational cost [33].

2.4 Radar and FMCW Principles

Radar systems operate by transmitting electromagnetic waves and analyzing the reflected signals to obtain information about surrounding objects, such as their range, motion, or direction. Millimeter-wave (mmWave) radar is a class of radar technology that uses short-wavelength electromagnetic waves [35]. One commonly used technique in mmWave radar is frequency-modulated continuous wave (FMCW) radar.

In FMCW radar, the transmitted signal varies continuously in frequency, often as a linear chirp. When the signal is reflected by an object and received back at the radar, it is delayed relative to the transmitted signal. By comparing the transmitted and received signals, a difference frequency is obtained, which is related to the target distance. By transmitting multiple chirps in sequence, the radar can also estimate relative velocity, and with multiple antennas, angle information can be extracted. The FMCW radar is therefore well-suited for applications where distance, motion, and position need to be measured simultaneously.

2.4.1 Radar-Based Vital Sign Monitoring

Radar can be used for contactless vital sign monitoring by detecting small movements of the chest surface [36]. These movements are mainly caused by respiration and cardiac activity, both of which modulate the reflected radar signal. Since no physical contact with the body is required, radar-based monitoring is an attractive approach for applications that require unobtrusive sensing.
Previous studies have shown that FMCW radar can be used to estimate BR, HR, and inter-beat intervals (IBIs) under controlled conditions [37]. In this thesis, radar-derived IBIs are treated as estimates of the corresponding ECG-derived RR intervals, since both describe the time between successive heartbeats. However, the term RR interval strictly refers to intervals between R-peaks in an ECG signal.
The chest displacement caused by respiration is typically larger than the motion caused by cardiac activity, which makes the extraction of cardiac micro-motions challenging. By separating the cardiac component from respiration and other motion sources, it can become possible to estimate heartbeat timing and derive radar-based IBIs for HRV analysis.

2.4.2 Limitations of Radar-Derived Inter-Beat Interval Estimation

When heartbeat-related signals are derived from radar data, several factors can affect the accuracy and reliability of the estimated IBIs [38]. These include motion artifacts, posture, signal-to-noise ratio, calibration sensitivity, and clothing or material effects.
Motion artifacts are a major limitation in practical measurement conditions. Large body movements can produce signal components that are much stronger than the small chest micro-motions associated with cardiac activity. As a result, the cardiac component may be difficult to isolate, especially when the subject is not stationary. This can reduce the reliability of radar-derived heartbeat timing and limit the robustness of subsequent HRV analysis.

2.5 Camera-Based Driver Monitoring Systems

DMS are systems designed to estimate a driver's state and behavior. In camera-based DMS, in-cabin cameras are used alongside computer vision and artificial intelligence to extract information about the driver's visual and behavioral state. Such

systems are commonly used to detect indicators of driver impairment, including distraction, drowsiness, gaze direction, eye closure, and head pose [39].
In the present thesis, DMS-derived features were obtained from a Smart Eye camera-based DMS. Smart Eye is an industrial provider of DMS technology, where in-cabin cameras are used together with computer vision and artificial intelligence to estimate aspects of the driver's state and behavior. According to Smart Eye, such systems can be applied to monitoring distraction, alcohol impairment, and drowsiness [40].

2.6 Machine Learning for Impairment Classification

ML is a data-driven approach in which algorithms learn patterns from observed data in order to make predictions or classifications on new, unseen samples [41]. This section introduces the supervised classification framework used in this thesis, including feature selection, classification algorithms, ensemble learning, CV strategies, and evaluation metrics.

2.6.1 Supervised Classification Framework

Supervised classification is an ML approach in which models are trained using observations with known class labels. Each observation is represented by a set of input features, and the corresponding label indicates the class to which the observation belongs [41]. During training, the model learns patterns that relate input features to class labels, enabling it to correctly classify new, unseen observations.
In binary classification, the task is to distinguish between two possible classes. In the context of impairment classification, this corresponds to separating observations into sober and impaired states. The learned model therefore estimates whether a given feature representation is more consistent with one class or the other.

2.6.2 Feature Selection

Feature selection is used to reduce the number of input variables before model training. This can improve interpretability, reduce computational complexity, and decrease the risk of overfitting, particularly when the number of features is large relative to the number of samples [42].
A common filter-based feature selection method is the ANOVA F-test. Filter methods evaluate features independently of the final classifier, often by assigning each feature a statistical score that reflects its relationship with the target variable [42]. In a classification setting, the ANOVA F-test evaluates each feature individually by comparing the variation between classes with the variation within classes. Features with higher F-values are considered more informative, since they show larger differences between the class groups relative to the variability within each group [43].
In ML pipelines, feature selection should be performed using only the training data. If feature selection is performed before CV using the full dataset, information from

the test data may influence which features are selected. This can lead to overly optimistic performance estimates due to information leakage [44].

2.6.3 Machine Learning Classifiers

Within the supervised classification framework, different algorithms can be used to learn the relationship between input features and class labels. The classifiers used in this thesis were selected to represent different types of decision rules, including linear models, distance-based methods, margin-based methods, tree-based ensembles, and boosting methods.
The following sections briefly describe the main principles of each classifier.

2.6.3.1 Logistic Regression

Logistic Regression (LR) is a statistical model commonly used for binary classification problems [41, 45]. The model estimates the probability that an observation belongs to a given class based on a linear combination of the input features. The resulting probability can then be converted into a class label using a decision threshold.
Since LR assumes a linear relationship between the features and the log-odds of the outcome, it provides a relatively simple and interpretable baseline model. However, this also limits its ability to capture complex non-linear relationships between features and class labels.

2.6.3.2 Random Forest

Random Forest (RF) is an ML algorithm that combines multiple decision trees to improve predictive performance and robustness [41, 46]. Each tree is trained on a randomly sampled subset of the training data, a process commonly referred to as bootstrap sampling. In addition, at each split in the tree, a random subset of features is considered when determining the optimal split. This combination of bootstrap sampling and random feature selection increases the diversity among the trees and helps reduce overfitting. The final prediction of the model is obtained by aggregating the predictions from all individual trees, typically through majority voting in classification tasks.

2.6.3.3 K-Nearest Neighbors

K-Nearest Neighbors (KNN) is a non-parametric, instance-based classification method [47, 41]. Instead of learning an explicit model during training, KNN classifies a new observation based on the labels of the K most similar observations in the training data. Similarity is commonly measured using a distance metric such as Euclidean distance.
The choice of K affects the flexibility of the model. A small value of K can make the classifier sensitive to noise, while a larger value produces a smoother decision boundary but may increase bias.

2.6.3.4 Support Vector Machines

Support Vector Machines (SVM) are supervised learning models that classify observations by constructing a decision boundary between classes [41]. The model aims to find the boundary that maximizes the margin between the classes, where the margin is determined by the closest training observations, known as support vectors.
SVM can use different kernel functions to construct either linear or non-linear decision boundaries. In this thesis, this is relevant since the relationship between physiological or behavioral features and impairment may not be strictly linear.

2.6.3.5 Gradient Boosting

Gradient Boosting (GB) is an ensemble ML technique similar to RF, but based on the boosting strategy [41]. In contrast to RF, where trees are trained independently, GB constructs decision trees sequentially. Each new tree is trained to correct the errors made by the previous model by minimizing a differentiable loss function using gradient descent. The final model is formed by combining the predictions of all trees in an additive manner. This sequential learning process enables GB to capture complex nonlinear relationships in the data. However, due to its iterative nature and high model flexibility, GB can be prone to overfitting, particularly when applied to small datasets without appropriate regularization or validation strategies.

2.6.3.6 Histogram-Based Gradient Boosting

Histogram-Based Gradient Boosting (HistGB) is an efficient implementation of GB where continuous feature values are discretized into a fixed number of bins before tree construction [48]. The model searches for optimal splits across these bins rather than evaluating all possible split points. With this approach, the computational and memory usage decreases, making it suitable for datasets with many samples or features.

2.6.3.7 Extreme Gradient Boosting

Extreme gradient boosting (XGBoost) is an optimized variant of GB designed for high predictive performance and computational efficiency [49]. The key difference between XGBoost and traditional GB implementation is that XGBoost includes several regularization techniques to reduce overfitting. The objective function includes both a loss term, which measures prediction error, and a regularization term, which penalizes model complexity. In addition, XGBoost uses second-order gradient information during optimization, which can improve both the efficiency and the accuracy of the learning process.

2.6.4 Ensemble Learning

Ensemble learning refers to a class of ML methods in which multiple models are combined to produce a final prediction [50]. The central idea is that a collection of models often can achieve better generalization than a single model, particularly when

individual models make different types of errors. By combining several predictors, ensemble methods can reduce variance and improve robustness.
Ensemble models can generally be described using two main steps: filtering and fusion. The filtering step involves selecting or removing base models before the final ensemble is created, for example, by excluding models with lower performance. The fusion step then combines the outputs of the selected base models into a final prediction [50].
In this thesis, ensemble learning is relevant through the multi-modal fusion of Polar and DMS models. The multi-modal ensemble follows a late fusion strategy, in which modality-specific models are trained independently and their predicted impairment scores are averaged at the decision level. This corresponds to a uniform score-averaging approach, in which each modality contributes equally to the final prediction.

2.6.5 Cross-Validation Strategies

CV is commonly used in ML to estimate how well a model generalizes to unseen data. This is particularly important when the available dataset is small, since a single train-test split may give a performance estimate that depends strongly on which samples are included in each subset [41].
In standard k-fold CV, the dataset is divided into k subsets, or folds. The model is trained on $k - 1$ folds and evaluated on the remaining fold. This process is repeated until each fold has been used once as a test set, and the final performance is calculated as the average across the folds.
However, when multiple samples originate from the same participant, randomly assigning samples to folds can lead to information leakage. In such cases, samples from the same individual may appear in both the training and test sets. For impairment detection, this could allow the model to learn subject-specific characteristics rather than alcohol-induced impairment patterns.
To address this, a leave-one-subject-out CV (LOSO) can be used. In LOSO CV, all samples from one participant are held out as the test set, while the model is trained on data from the remaining participants [51]. This process is repeated until each participant has been used once as the held-out test subject. LOSO CV therefore provides an estimate of subject-independent generalization, meaning how well the model performs on a previously unseen participant.

2.6.6 Evaluation Metrics

The performance of the classification models can be evaluated using several metrics derived from the classification outcomes. In this thesis, the impaired class is treated as the positive class, while the sober class is treated as the negative class. In a binary classification setting, the predictions can therefore be divided into four categories: true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) [52].
These outcomes can be summarized in a confusion matrix, which compares the true and predicted class labels. The rows of the matrix represent the true classes, while

the columns represent the predicted classes. This provides an overview of how many samples were correctly or incorrectly classified for each class.
The four outcomes describe whether the model's prediction agrees with the true impairment state. Correct classifications are obtained when impaired samples are assigned to the impaired class, or when sober samples are assigned to the sober class. These cases correspond to TP and TN, respectively. Incorrect classifications occur when the predicted class does not match the true class. In this thesis, an FP means that a sober sample is classified as impaired, while an FN means that an impaired sample is classified as sober.
Accuracy is an evaluation metric that measures the proportion of correctly classified samples among all samples [52]:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}. \tag{2.8}$$

However, accuracy alone can be misleading, especially for small or imbalanced datasets. For example, if one class is more common than the other, a model may achieve high accuracy by mainly predicting the majority class while still performing poorly on the minority class [52]. Therefore, in this thesis, additional evaluation metrics were used, including precision, sensitivity, specificity, and the area under the receiver operating characteristic curve (ROC-AUC), hereafter referred to as AUC.
Precision measures the proportion of samples predicted as impaired that were actually impaired [52]:

$$Precision = \frac{TP}{TP + FP}. \tag{2.9}$$

Sensitivity, also referred to as the true positive rate, measures the proportion of impaired samples that were correctly identified as impaired [52]:

$$Sensitivity = \frac{TP}{TP + FN}. \tag{2.10}$$

Specificity, also referred to as the true negative rate, measures the proportion of sober samples that were correctly identified as sober [52]:

$$Specificity = \frac{TN}{TN + FP}. \tag{2.11}$$

Accuracy, precision, sensitivity, and specificity are threshold-dependent metrics, since the predicted class labels depend on the chosen decision threshold. In addition to these metrics, the AUC can be used to evaluate a model's performance [53]. AUC reflects how well the model separates the two classes across all possible classification thresholds. A high AUC indicates that positive samples generally receive higher prediction scores than negative samples, suggesting good class separability. An AUC value of 1 indicates perfect discrimination, whereas an AUC value of 0.5 corresponds to random classification [53].
To assess the uncertainty of the evaluation metrics, bootstrapping can be used. Bootstrapping is a non-parametric resampling method in which repeated samples are drawn with replacement from the original data set [54]. For each bootstrap

sample, the performance metrics of interest are recalculated, producing an empirical distribution of the metrics. This distribution can then be used to estimate confidence intervals without making strong assumptions about the underlying data distribution.

3
Methods

This chapter describes the methodological framework of the study. It first presents the overall study design, participant recruitment, ethical considerations, and experimental protocols used for data collection. The main data collection was conducted in a controlled driving simulator environment, where physiological reference data, radar-based vital sign measurements, and camera-based DMS data were collected. In addition, an external real-world driving dataset was used to evaluate model generalizability outside the simulator setting.
The chapter further describes the measurement equipment, including the Polar H10 chest-worn sensor, the radar-based sensing system, and the camera-based DMS. The procedures for signal preprocessing, feature extraction, dataset construction, and labeling are then presented. Finally, the ML models, hyperparameter tuning, evaluation strategy, and the exploratory ensemble approach combining physiological and DMS-based model outputs are described.

3.1 Study Design and Overview

The primary objective of this study was to develop and evaluate ML models to detect alcohol-induced impairment using physiological signals. An experimental within-subjects study design was used, in which physiological signals were recorded from each participant under sober conditions and at increasing BAC levels. This design enabled the assessment of alcohol-related physiological changes within the same individual.
The main data collection was conducted during controlled driving sessions in a physical driving simulator environment. During these sessions, physiological reference data, radar-based vital sign measurements, and camera-based DMS data were collected. In addition, an external real-world driving dataset was used as an independent test set to evaluate the physiological models' performance outside the simulator environment.
Radar-based vital sign monitoring was included to evaluate the feasibility of unobtrusive, non-contact physiological sensing in this application. The radar-based sensing system was provided by Emsense Technologies, which supported both the hardware implementation and the software pipeline for vital sign extraction. Emsense was responsible for installing and validating the radar system in the simulator environment. The radar unit was mounted in the backrest of the driver's seat, enabling unobtrusive monitoring of physiological signals during simulator driving.
As a reference measurement, participants wore a chest-mounted Polar H10 pulse

band. The reference recordings were used both to compare against the radar-derived HR measurements and to extract HR- and HRV-based features for the final physiological classification models. Radar-derived vital signs were evaluated against the Polar H10 reference measurements. Since the radar-derived IBIs were deemed insufficiently reliable for HRV-based classification, the final physiological ML pipeline was based on Polar H10 recordings.
In addition to the physiological measurements, data from a camera-based DMS developed by Sightic were available during the simulator sessions. The DMS was used to assess driver impairment based on image-derived behavioral indicators. These data provided a second sensing modality complementary to the HR- and HRV-based physiological features. As an exploratory extension, the camera-based DMS model outputs were later combined with the physiological model outputs to evaluate whether a late-fusion ensemble approach could improve the detection of alcohol-induced impairment.

3.2 Ethical Considerations

A written informed consent form was developed in accordance with GDPR. In addition, an information document was prepared describing the study procedure, data handling, and participants' rights.
The information document was distributed to the participants several days before the scheduled data collection session to ensure that they had sufficient time to review the study details. On the day of the experiment, the informed consent form was signed in writing by each participant. A copy of the signed consent form was provided to the participant for their records.
Participants also signed Sightic's consent form regarding safe alcohol consumption and consented to being recorded by the Driver Monitoring System (DMS). The Swedish data collection was covered by ethical approval from the Swedish Ethical Review Authority (Etikprövningsmyndigheten), approved on 8 March 2026, diary number Dnr 2026-00090-02. This approval constitutes an amendment to Sightic's previously approved ethical review application, diary number Dnr 2023-03755-01, which was approved on 12 December 2023. The amendment includes the use of a wearable physiological sensor, radar-based measurement of vital signs, and questions regarding medical conditions. These additions were approved to improve the ability to distinguish alcohol-impaired participants from non-impaired participants.
An additional amendment to the Swedish ethical approval was obtained for the use of externally collected data from Serbia. This amendment was approved by the Swedish Ethical Review Authority on 22 August 2025, diary number Dnr 2025-05668-02. The amendment concerns the use of ML models to detect alcohol-impaired drivers in naturalistic driving environments, including the analysis and processing of data collected abroad in Sweden. The approved amendment allowed the Serbian data collection to be expanded and the corresponding data to be analyzed in Sweden.
For the external data collection conducted in Serbia, ethical approval was also obtained from the Ethics Commission of the Faculty of Transport and Traffic Engineering at the University of Belgrade. The approval was issued on 26 December 2024, with reference number 1732/3. The approval covers research on detecting drivers

under the influence of alcohol using image analysis, conducted in collaboration with Sightic Analytics AB. The document states that the research may be used for scientific and professional publications, technical reports, and the development of new technologies in the automotive industry.

Copies of the relevant ethical approvals are available upon request.

3.3 Preliminary Study

Prior to the main data collection, a small-scale preliminary study was conducted during a mini-event involving voluntary participants. Physiological data were recorded using a radar system together with a chest-worn pulse band serving as a reference, and the participants consumed approximately three standard units of alcohol.

The preliminary analysis showed that the radar-derived HR and HRV estimates were insufficiently reliable relative to the reference measurements. In addition, the observed physiological changes were modest, indicating that the resulting level of impairment was limited.

Based on these findings, the radar data were considered unsuitable for quantitative HRV analysis at this stage, and the target BAC for the main study was adjusted to approximately 0.8‰. The preliminary dataset was also used to explore initial ML models and potentially informative features, but the results highlighted the need for a larger dataset to ensure more robust model performance.

3.4 Driving Simulator Setup in the Main Study

The main data collection was conducted using a controlled physical driving simulator located at Sightic's office. The simulator provided a standardized, repeatable driving environment, allowing all participants to complete the same scenario under comparable conditions. The setup consisted of a driver's seat, a steering wheel with blinkers, and pedals mounted on a motion platform. The motion platform was disabled during all data collection sessions. For visualization, three 55-inch screens were used to provide an immersive driving environment.

An overview of the simulator setup is shown in Figure 3.1 below.

Figure 3.1: A picture of the simulator setup.

To ensure consistency across participants and measurement sessions, the same predefined driving route was used throughout the study. The route was designed to include different traffic environments and levels of driving demand. It consisted of three main sections: an urban section, a highway section, and a rural road section.

3.5 Measurement Equipment

The radar system used in the study was a compact 24 GHz continuous-wave radar system. The radar system sampled in-phase (I) and quadrature (Q) baseband signals at a rate of 12,000 Hz. The raw I/Q data were acquired by the radar system; however, signal processing and heart signal extraction were performed by an external partner. The radar unit was mounted inside the backrest of the driving simulator seat to enable unobtrusive monitoring of physiological signals during the simulator sessions. The system was configured to capture micro-movements associated with cardiac and respiratory activity.
As a reference measurement, a chest-worn Polar H10 pulse belt (Polar Electro Oy,

Finland) was used to record HR and inter-beat intervals. The Polar H10 is widely used in research settings and has been shown to provide reliable HRV measurements under resting and moderate activity conditions [55].
Participants wore the Polar H10 chest strap according to the manufacturer's instructions, positioned around the chest just below the pectoral muscles, as illustrated in Figure 3.2.

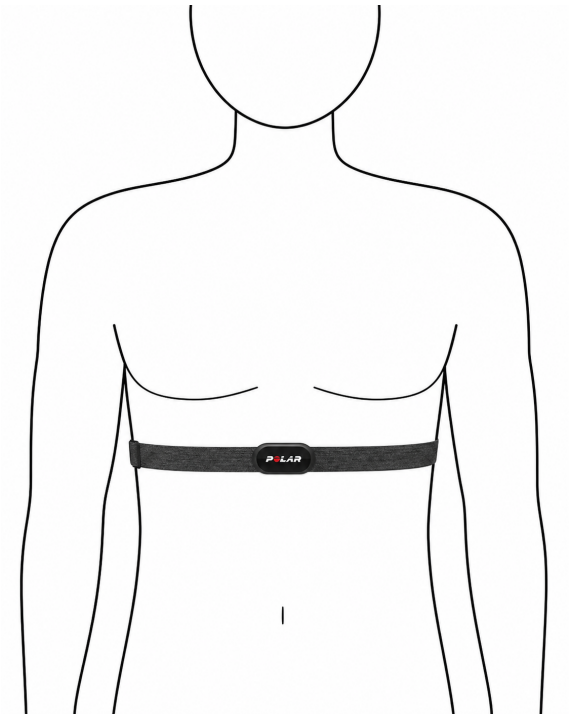

Figure 3.2: Illustration of the recommended placement of the Polar H10 chest strap. The illustration was generated using ChatGPT.

In addition to the physiological sensors, camera-based DMS data were recorded at 30 Hz using a Tobii system. The system was equipped with an Omnivision OV2311 2 MP image sensor and 940 nm infrared illumination. These data were used as an additional sensing modality for the camera-based driver monitoring system and were later included in the multimodal impairment classification analysis.

3.6 Experimental Protocol

The next step of the project was to acquire experimental data for ML training and evaluation. The data collection sessions were conducted at Sightic's office. Each session began with a five-minute resting period to ensure a calm and standardized starting condition. Participants then performed the predefined driving task in the simulator described in Section 3.4, during which the study measurements were recorded for ten minutes. This first driving session was conducted while sober.
After the sober driving session, participants consumed alcohol in a controlled manner intending to reach a lower impaired BAC level of approximately 0.4‰. Before the

next driving session, participants had another short five-minute rest and water-drinking period. They then performed a second ten-minute driving session at the lower impaired BAC level.
Following the second driving session, participants consumed additional alcohol with the aim of reaching a higher impaired BAC level of approximately 0.8‰. This was again followed by a short rest and water-drinking period before the participants completed a third driving session at the higher impaired BAC level.
BAC levels were monitored using a breathalyzer after each driving session. Breath alcohol measurements were performed using a Dräger Alcotest 6000 breath alcohol tester (type Alcotest 5820, Dräger Safety AG & Co. KGaA, Lübeck, Germany).
Participants were not explicitly instructed to restrict natural body movement or conversation during the simulator sessions. Use of nicotine was noted, as well as the time since the participant last consumed caffeine, but these factors were not further investigated.
This protocol enabled the collection of data under three experimental conditions: sober driving, low-to-moderate-BAC impaired driving, and high-BAC impaired driving.

3.6.1 Simulator Dataset

A total of 37 participants were recruited for the study. Participants included voluntary colleagues, friends, and family members. The participants ranged in age from 20 to 36 years, with a mean age of 25 years and a median age of 24 years. The sample consisted of 27 male and 10 female participants. All participants met the inclusion criteria of being over 18 years of age and holding a valid driver's license. Individuals with known cardiovascular conditions or other medical conditions that could affect physiological measurements were excluded.
Table 3.1 shows the resulting BAC levels observed under the three experimental conditions across all 37 participants. For the low-to-moderate BAC level, the lowest recorded BAC was 0.24 ‰.

Table 3.1: Mean BAC levels and number of recordings across experimental conditions.

Condition	Mean BAC	Number of recordings
Sober	0.00 ‰	37
Impaired 1 (Low-to-moderate BAC)	0.47 ‰	35
Impaired 2 (High-BAC)	1.11 ‰	36

The imbalance in the number of recordings across conditions was mainly due to differences in the experimental setup during one smaller data collection event. During this event, three participants were only scheduled to complete one impaired driving session rather than two. In addition, one participant did not reach the BAC level required for impaired session 2. As a result, the sober condition included 37 recordings, impaired session 1 included 35 recordings, and impaired session 2 included 36 recordings.

3.7 External Data Collection

An external data collection event was conducted to acquire real-world physiological driving data using a Polar H10 HR sensor. During this event, participants drove a vehicle under controlled conditions on a designated test track, enabling safe, consistent data acquisition in a real-world driving environment. For the present thesis, only pulse band data from the Polar H10 were available from this external data collection; no radar-based measurements were collected or analyzed for the real-world driving dataset.
The study was conducted on a test track in Serbia, where the experimental protocol and overall data collection methodology were developed in collaboration with the Serbian researchers *Emir Smailovic* and *Dalibor Pesic* from the Faculty of Transport and Traffic Engineering, University of Belgrade, in partnership with Sightic. Their contributions included the design of the experimental setup, driving scenarios, and safety procedures, which formed the foundation for the real-world impairment data collection.
The purpose of this data collection was primarily to evaluate the generalizability of models trained on simulator data. Specifically, the real-world dataset was used as an independent test set to assess whether the developed models could distinguish between sober and impaired driving conditions outside the simulated environment. In addition, an exploratory analysis was performed in which a subset of the real-world data was included during model training to investigate whether limited exposure to real-world driving data could improve performance on the external dataset.

3.7.1 External Real-World Dataset

The external real-world dataset consisted of 26 participants with data collected under both sober and impaired real-world driving conditions.
This resulted in a final external test dataset consisting of 17 participants, each contributing one sober and one impaired driving session. The participants in the final dataset ranged in age from 20 to 57 years, with a mean age of 36.5 years and a median age of 35 years. The sample consisted of 10 male and 7 female participants. Table 3.2 summarizes the mean BAC levels and the number of sessions included in the final external test dataset.

Table 3.2: Mean BAC levels and number of sessions included in the final external real-world test dataset.

Condition	Mean BAC	Number of sessions
Sober	0.00 ‰	17
Impaired	1.03 ‰	17

3.8 Implementation Pipeline

All data processing and ML analyses were implemented in Python. Data handling was primarily performed using pandas and NumPy, while scikit-learn was used for

preprocessing, feature selection, model training, CV, and evaluation. In addition, XGBoost was used for gradient-boosted tree-based classification, and tsfresh was used for automated time-series feature extraction.
The overall analysis pipeline consisted of several sequential steps: signal preprocessing, window-based segmentation, dataset construction and labeling, feature extraction, model training, and model evaluation. The same general pipeline was applied to the simulator dataset and the external real-world dataset, with the external dataset used primarily for independent testing.
For the physiological data, RR-interval recordings from the Polar H10 were preprocessed and segmented into fixed-length windows before HR- and HRV-based features were extracted. The radar-derived signals were evaluated separately against the reference measurements, but were not included in the final ML pipeline due to insufficient signal quality. Camera-based DMS outputs were processed as a separate modality and later used in the exploratory ensemble analysis.
Each step is described in more detail in the following sections.

3.8.1 Preprocessing

During the data collection events, Polar H10 data from each participant was recorded and stored as files containing timestamps and corresponding RR intervals. A metadata JSON file was also created for each recording, including participant ID and BAC.
Preprocessing of the RR-interval data from the Polar H10 was performed to remove artifacts and obtain a more reliable physiological signal. First, RR intervals shorter than 300 ms or longer than 1600 ms were removed, as such values were considered physiologically implausible and likely to reflect measurement artifacts or signal corruption.
To further improve signal quality, a jump filter was applied. This filter processed the RR-interval sequence sequentially and compared each new interval with the median of the most recent accepted intervals. After an initial warm-up period, the local reference was based on the ten most recent accepted intervals. An interval was rejected if its deviation from this local median exceeded an adaptive tolerance, defined as the maximum of 30% of the local median. Such abrupt deviations were considered unlikely to reflect true physiological variation and were instead assumed to originate from noise, motion artifacts, missed beats, or sensor errors. The removed RR intervals were excluded from feature extraction.
For the external real-world dataset, an additional trimming step was applied. Each recording originally contained approximately 24 minutes of driving data but did not include a separate resting period consistent with the simulator protocol. To make the external recordings more comparable to the simulator data and to reduce potential transient effects at the beginning and end of each drive, the first 7 minutes and the last 2 minutes of each recording were excluded from the analysis. The remaining middle segment was then processed using the same preprocessing, segmentation, and feature-extraction procedures as the simulator data.
All preprocessing steps were applied before feature extraction.
For the radar data, no additional signal processing was performed within this the-

sis. Instead, preprocessed vital sign signals were provided by Emsense, which was responsible for the radar signal acquisition and processing pipeline.
For the camera-based DMS data, no additional image processing was performed within this thesis. Instead, Sightic provided preprocessed time-series outputs from the DMS. The DMS data were time-synchronized with the Polar H10 recordings and segmented into fixed-length time windows.

3.8.2 Dataset Construction and Labeling

To construct the ML dataset, the continuous RR-interval signals were segmented into non-overlapping 45-second windows after preprocessing. A window length of 45 seconds was chosen to provide short-term physiological estimates while retaining a sufficient number of RR intervals for feature extraction. Since BAC was measured at the session level, all windows within a given recording were assigned the same label.
Labeling was performed based on the BAC value associated with each recording. Recordings with BAC equal to 0 were labeled as *sober*, while recordings with BAC greater than 0 were considered impaired recordings. To evaluate model performance under different impairment definitions, three simulator-based classification settings were used, as summarized in Table 3.3.

Table 3.3: Definition of the simulator-based classification settings used in the analysis.

Setting	Definition
High-BAC	Sober vs. recordings with BAC $\geq 0.7‰$
Low-to-moderate BAC	Sober vs. recordings with $0 <$ BAC $< 0.7‰$
All-data	Sober vs. all recordings with BAC > 0

In the High-BAC setting, recordings with BAC values below 0.7‰ were excluded, while recordings with BAC values greater than or equal to 0.7‰ were labeled as *impaired*. A threshold of 0.7‰ was selected because several participants in the simulator dataset did not reach the intended BAC level of 0.8‰.
In the Low-to-moderate BAC setting, only impaired recordings with BAC values above 0 and below 0.7‰ were included. In the All-data setting, all available impaired simulator recordings were included, regardless of BAC level.
The High-BAC setting resulted in an approximately balanced dataset, since each participant generally contributed one sober recording and one high-BAC impaired recording. In contrast, the All-data setting was imbalanced because each participant could contribute one sober recording and up to two impaired recordings at different BAC levels.
To account for class imbalance, weights were applied during model training within each cross-validation fold, using only the training data from that fold. For XGBoost, imbalance was handled using the built-in `scale_pos_weight` parameter. For LR, SVM, and RF, balanced class weights were used. For models where direct class weighting was not applied, such as KNN, the model was retained as an unweighted

baseline. The held-out test participant was always evaluated using the original class distribution.
The final dataset consisted of feature vectors extracted from individual time windows, each associated with a class label, BAC value, and participant ID.
For the external dataset evaluation, the same high-BAC impairment criterion as in the simulator-based high-BAC analysis was applied. Participants were included if they had one sober recording and one impaired recording in which they reached a BAC level of 0.7‰ or higher. Using the same BAC threshold for the external real-world dataset allowed the external evaluation to be performed under a comparable impairment definition.

3.8.3 Feature Extraction and Selection

Manual features were first extracted from each segmented 45-second window of the Polar H10 RR-interval data. These features were defined using established HRV measures, including SDNN, RMSSD, pNN50, SD2, and mean HR. LF and HF features were also defined manually. A detailed description of the extracted features is provided in the theory chapter.
Two additional ratio-based features were derived from the manually extracted HRV features. The first feature was defined as the logarithm of the ratio between mean HR and RMSSD:

$$\log\left(\frac{\overline{HR}}{RMSSD}\right) \tag{3.1}$$

The second feature was defined as the logarithm of the ratio between RMSSD and SDNN:

$$\log\left(\frac{RMSSD}{SDNN}\right) \tag{3.2}$$

These ratio-based features were included to capture relative changes in HR level and HRV, rather than relying only on their absolute values. These two features were included as exploratory engineered features designed specifically for this study. They were not taken directly from established HRV feature sets, but were introduced to investigate whether relative relationships between HR level and HRV measures could provide additional discriminatory information.
In addition to the manually defined HRV features, automated feature extraction was performed using the Python package tsfresh with the efficient setting. For each RR-interval window, tsfresh computed a large set of general time-series features. These features were intended to complement the manually defined HRV features by capturing additional patterns in the RR-interval signal. This approach is useful for physiological signals, where relevant patterns may not be captured by a small set of manually defined features. At the same time, automated extraction can produce many redundant or irrelevant features. Therefore, the extracted tsfresh features were used as candidate features before subsequent feature selection and model training.
The resulting tsfresh features were combined with the manual features and used as input to the physiological ML models.

For the camera-based DMS data, no manual features were defined within this thesis. Instead, feature extraction was performed using tsfresh on the preprocessed DMS time-series outputs, following the same window-based approach as for the Polar H10 data.
Feature selection was performed separately within each LOSO-CV fold using an ANOVA F-test. In each fold, the ANOVA F-test was fitted only on the training subjects, excluding the held-out subject to prevent information leakage. Features were then ranked according to their F-values, and different numbers of top-ranked features were selected for model training.

3.8.4 Machine Learning Models

Several ML models were evaluated to compare their ability to classify alcohol-induced impairment based on physiological features. The evaluated models included LR, RF, GB, HistGB, XGBoost, SVM with both linear and radial basis function kernels, and KNN.
These models were selected to represent different types of classification approaches, including linear models, distance-based methods, margin-based classifiers, tree-based ensembles, and boosting methods. This made it possible to evaluate how different model assumptions and levels of model complexity affected classification performance.
To ensure a fair comparison, all models were trained and evaluated using the same feature sets and validation strategy.

3.8.5 Hyperparameter Tuning

To optimize model performance, some hyperparameters were manually tuned. Rather than relying on default settings, key hyperparameters were varied within predefined ranges based on prior knowledge and common practices in the literature.
For instance, in the KNN model, the number of neighbors k was varied to investigate the trade-off between sensitivity to noise and model generalization. Similarly, for tree-based models such as XGBoost, parameters including the number of estimators, maximum tree depth, and learning rate were adjusted. For SVM, both kernel choice and regularization parameters were explored.

3.8.6 Model Evaluation

Model performance was evaluated using accuracy, precision, sensitivity, specificity, and AUC. These metrics are described in detail in the theory chapter.
Evaluation was performed at both the window and session levels. At the window level, each 45-second segment was treated as a single sample and assigned a predicted impairment score. At the session level, all window-level prediction scores for the same recording session were averaged to obtain a single session-level prediction score.
To evaluate subject-independent generalization, LOSO CV was used. In each fold, all data from one participant were held out as the test set, while the model was trained on data from the remaining participants. This ensured that windows from the same participant did not appear in both training and test data.

After completing all LOSO folds, the predictions from the held-out test sets were pooled. This produced one out-of-sample prediction score for each window, and one aggregated prediction score for each session. The final evaluation metrics were then computed from these pooled out-of-fold predictions. This means that each reported prediction was generated by a model that had not been trained on data from the corresponding participant.
For the session-level evaluation, uncertainty was estimated using bootstrap resampling. Resampling was performed at the participant level, rather than at the individual session level, in order to preserve the pairing between sober and impaired sessions from the same participant. For each bootstrap sample, the evaluation metrics were recalculated. The 95% confidence intervals were then obtained from the 2.5th and 97.5th percentiles of the resulting bootstrap distribution.

3.8.7 Exploratory Ensemble Approach

After evaluating the individual modality-specific models, an ensemble-based late fusion approach was implemented to investigate whether combining physiological and camera-based information could improve impairment classification. The motivation for this approach was that HR/HRV and DMS features may capture different physiological and behavioral effects of alcohol-induced impairment, and may therefore provide complementary information.
The ensemble was implemented as a score-level fusion method. Separate classifiers were trained for each modality: one using the Polar-derived features and one using the DMS-derived features. For each held-out participant in the LOSO CV, the modality-specific models were trained only on the remaining participants and then used to generate impairment probability scores for the held-out participant. This ensured that the fusion procedure followed the same subject-independent evaluation protocol as the individual models.
The final ensemble score was computed by averaging the predicted impairment scores from the modality-specific models. In this approach, both modalities contributed equally to the final decision. The averaged score was then compared with the classification decision threshold to assign the final predicted class label. This late fusion strategy allowed the two modalities to remain independent during feature extraction and model training, while still combining their predictions at the decision stage.

4

Results

This chapter presents the results of the alcohol-induced impairment classification analysis. The chapter begins with an evaluation of radar-based vital sign monitoring, as the radar-derived HR estimates were insufficiently reliable to be included in the final classification pipeline. These results are therefore presented separately to illustrate the limitations encountered in this study's use of radar-based measurements.

The main classification results are then presented using HR- and HRV-related features extracted from the Polar H10 pulse belt recordings. These results include simulator-based classification across different BAC settings, as well as an external real-world test in which models trained on simulator data were evaluated on independent driving data. The selected features and descriptive HR and HRV changes across alcohol conditions are also presented to support the interpretation of the classification results.

Finally, the chapter presents results from models based on features derived from the image-based driver monitoring system. These predictions are evaluated both separately and in combination with the Polar H10-based predictions using a late-fusion approach to examine whether physiological and image-based information provide complementary information for impairment classification.

4.1 Radar-Based Vital Sign Monitoring

All RR interval recordings obtained from the chest-worn Polar H10 sensor were compared with RR intervals derived from the radar-based system.

Figure 4.1 shows a representative subset of the data and illustrates the differences between the two measurement methods. The green line represents the mean HR over time derived from the Polar H10, while the red line represents the corresponding estimate from the radar-based system. Since the Polar H10 was used as the reference measurement in this study, the radar-derived estimates were expected to follow the same general pattern.

The radar-based measurements did not show sufficient agreement with the Polar H10 reference signal in the analyzed recordings. Large deviations were observed, and the radar-derived HR estimates did not reliably follow the reference trend. Therefore, the radar data were excluded from the subsequent classification analysis, as described previously.

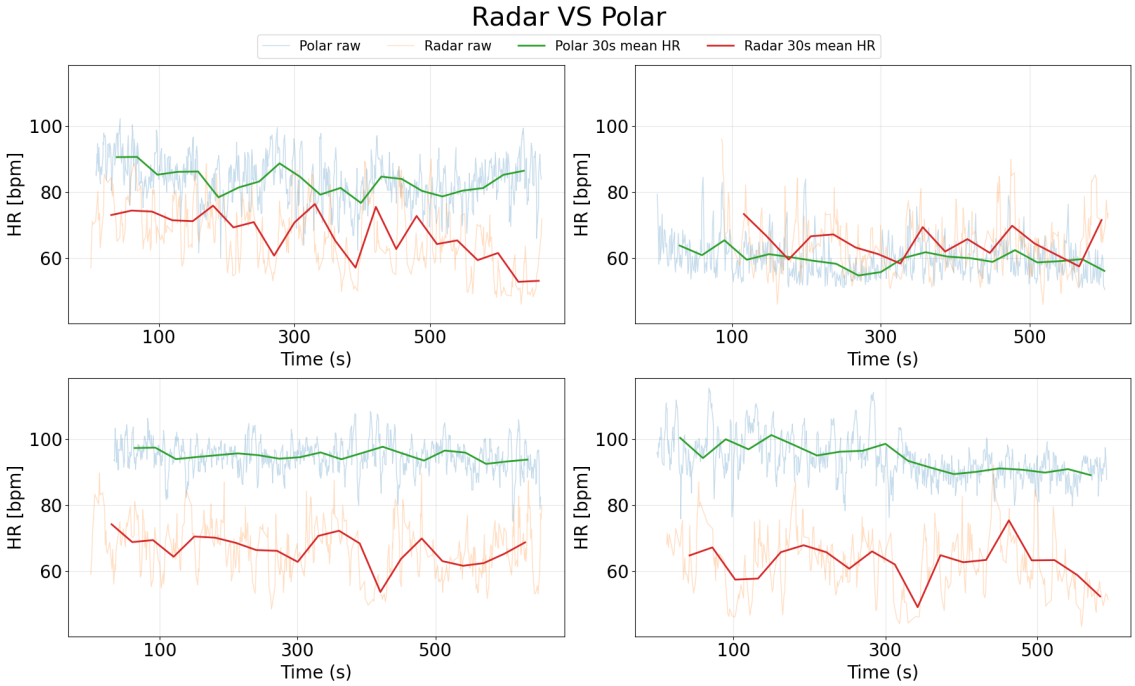

Figure 4.1: Comparison of raw HR and 30-second mean HR estimated from the radar system and the Polar H10.

4.2 Polar H10-Based Classification Performance

This section presents the classification results based on HR- and HRV-related features extracted from the Polar H10 recordings. For clarity, the simulator-based analyses are grouped into three classification settings: high-BAC, low-to-moderate BAC, and all-data, as summarized in Table 3.3. In addition, the external real-world driving data were used as an independent test set, with models trained on the simulator data.

The classification performance is reported at both the window and session levels. Window-level performance refers to predictions made for individual 45-second windows. The 45-second window size yielded the best results compared to the 30- and 60-second window sizes in the high-BAC setting. As a result, 45 seconds were used for the entire analysis. Session-level performance was obtained by averaging the window-level prediction scores within each recording session, then applying the decision threshold to the resulting average. Thus, the session-level results represent the final classification of an entire driving session rather than individual short-term windows.

All simulator-based evaluations were performed using LOSO CV. Decision-threshold-dependent metrics were computed from pooled out-of-fold predictions, meaning that the held-out predictions from all folds were combined before calculating the confusion matrices. AUC was computed from the pooled prediction scores.

4.2.1 Feature Set Comparison

Three different feature sets were evaluated: manually defined HR- and HRV-based features, automatically extracted tsfresh features, and a combined feature set containing both manual and tsfresh features. This comparison was performed to evaluate whether the automatically extracted features provided additional information beyond the manually selected physiological features.

Table 4.1 summarizes the best session-level performance obtained for each feature set in the simulator classification task on the high-BAC setting. The comparison focuses on models trained using three selected features, since this provided strong performance while keeping the models relatively simple. Only the best-performing model for each feature set is shown to keep the comparison concise. The full model results for each feature set are provided in A.

Table 4.1: Best session-level performance for each feature set in the simulator high-BAC classification task using three selected features.

Feature set	Best model	AUC	Accuracy	Sensitivity	Specificity
Manual	SVM linear	0.78	0.73	0.71	0.75
tsfresh	SVM linear	0.81	0.78	0.74	0.83
Combined	SVM linear	0.81	0.78	0.74	0.83

The manual feature set achieved lower performance than the feature sets containing tsfresh features. The tsfresh-only and combined feature sets produced almost identical session-level performance, both reaching an AUC of 0.81 and an accuracy of 0.78. This indicates that the automatically extracted time-series features captured information that was useful for the classification task. The combined feature set was therefore used for the main analysis, as it retained the performance of the tsfresh-only feature set while also including manually derived physiological features that were easier to interpret.

The following classification results are based on models trained using the combined feature set with three selected features.

4.2.2 Summary of Main Classification Results

Table 4.2 summarizes the best session-level results from the main Polar H10-based classification experiments. The table also includes bootstrap confidence intervals to estimate the variability and uncertainty of the evaluation metrics.

For the all-data setting and the external real-world test, two models are included because they represent different trade-offs between sensitivity and specificity. In the all-data setting, the linear SVM achieved the highest accuracy and sensitivity but had low specificity. XGBoost, on the other hand, produced a lower accuracy but substantially higher specificity.

Table 4.2: Best session-level performance across the main Polar H10-based classification settings.

Set.	Model	AUC	Acc.	Sens.	Spec.
HB	SVM linear	0.81 [0.70, 0.90]	0.78	0.74 [0.60, 0.87]	0.83 [0.69, 0.94]
LMB	XGBoost	0.68 [0.55, 0.81]	0.72	0.66 [0.49, 0.80]	0.78 [0.64, 0.92]
AD	SVM linear	0.74 [0.65, 0.84]	0.72	0.90 [0.84, 0.96]	0.33 [0.17, 0.50]
AD	XGBoost	0.73 [0.63, 0.83]	0.71	0.68 [0.58, 0.79]	0.75 [0.58, 0.89]
EXT	XGBoost	0.75 [0.58, 0.91]	0.74	0.71 [0.47, 0.88]	0.76 [0.53, 0.94]
EXT	SVM linear	0.72 [0.53, 0.88]	0.74	0.59 [0.35, 0.82]	0.88 [0.71, 1.00]

Note. Values in square brackets denote 95% bootstrap confidence intervals based on 2000 resamples. Performance metrics were computed at a decision threshold of 0.5 for HB, LMB, and AD, and 0.6 for EXT. Abbreviations: HB = high-BAC; LMB = low-to-moderate BAC; AD = all-data; EXT = external real-world; Acc. = accuracy; Sens. = sensitivity; Spec. = specificity.

4.2.3 High-BAC Classification

The high-BAC setting was used as the main simulator-based classification task, since it represented the clearest separation between sober and impaired driving. The best-performing model in this setting was the linear SVM trained on three selected combined features, which achieved an AUC of 0.81 with a confidence interval between 0.70 and 0.90, indicating class separability. At a decision threshold of 0.5, the model achieved a session-level accuracy of 0.78 and a specificity of 0.83, with six FP classifications. The confusion matrix is shown in Figure 4.2.

Table 4.3 summarizes the pooled out-of-fold performance for all evaluated models in the high-BAC. Several models achieved similar session-level performance. In particular, the SVM with RBF kernel, XGBoost, and LR performed comparably to the linear SVM, indicating that the selected features contained information that could be captured by several different classifiers.

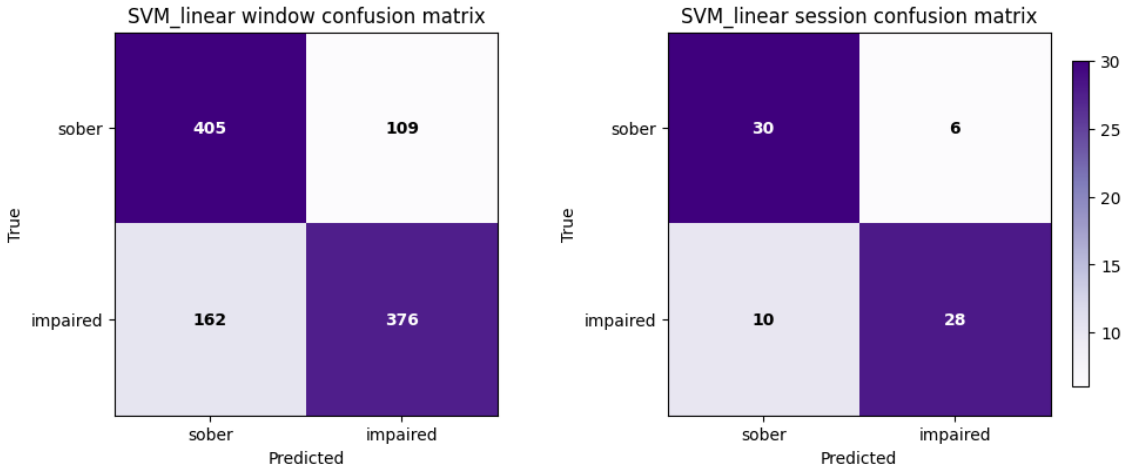

Figure 4.2: Window-level and session-level confusion matrices for the linear SVM in the high-BAC setting using three combined features at threshold 0.5.

Table 4.3: Pooled out-of-fold model performance in the high-BAC setting using three selected combined features at decision threshold 0.5.

Model	Level	AUC	Accuracy	Precision	Sensitivity	Specificity
SVM RBF	Window	0.75	0.75	0.77	0.72	0.78
	Session	0.79	0.78	0.82	0.74	0.83
SVM linear	Window	0.79	0.74	0.78	0.70	0.79
	Session	0.81	0.78	0.82	0.74	0.83
XGBoost	Window	0.75	0.74	0.77	0.69	0.79
	Session	0.79	0.77	0.80	0.74	0.81
Logistic regression	Window	0.79	0.74	0.77	0.70	0.78
	Session	0.81	0.77	0.80	0.74	0.81
Random forest	Window	0.75	0.70	0.71	0.68	0.71
	Session	0.80	0.74	0.77	0.71	0.78
HistGB	Window	0.73	0.68	0.70	0.68	0.69
	Session	0.79	0.73	0.75	0.71	0.75
KNN	Window	0.75	0.70	0.71	0.68	0.71
	Session	0.78	0.70	0.71	0.71	0.69

Additional results for this high-BAC classification setting, including model performance and the most consistently selected features when using the top 5 and top 10 features, are provided in A.

4.2.4 Low-to-Moderate BAC Classification

To evaluate the model performance for lower levels of alcohol impairment, an additional experiment was performed where sober data was compared against low-to-moderate BAC levels. Table 4.4 shows the performance for every evaluated model. The best-performing model in this setting was an XGBoost classifier, achieving a session accuracy of 0.72 at a decision threshold of 0.5. It also achieved a session-level AUC of 0.68 with a confidence interval between 0.55 and 0.81, suggesting that some separability was present, but more difficult compared to the previous setting. The confusion matrix for XGBoost is shown in Figure 4.3, indicating that the number of FPs and FNs increased by 2 compared to the high-BAC setting.

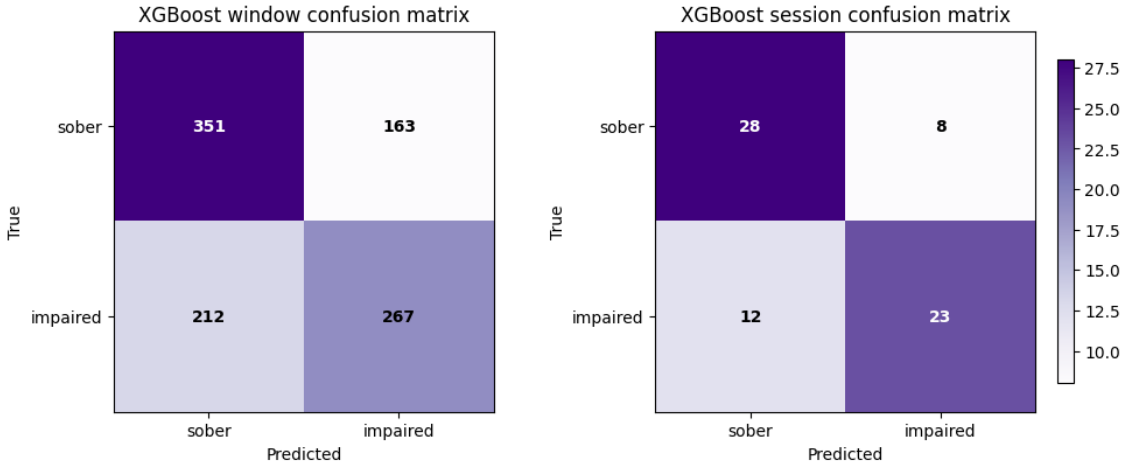

Figure 4.3: Window-level and session-level confusion matrices for XGBoost in the low-to-moderate BAC setting using three combined features at decision threshold 0.5.

Overall, these results indicate that physiological differences between sober driving and lower BAC impairment levels are more difficult to detect than for higher BAC levels. The relatively high number of FN, especially at the window level, suggests that light alcohol impairment may not always produce sufficiently distinct HR and HRV patterns within short time windows. This is consistent with the expectation that classification becomes more challenging when the impaired class includes lower BAC values closer to the sober condition.

Table 4.4: Pooled out-of-fold model performance in the low-to-moderate BAC setting using three combined features at decision threshold 0.5.

Model	Level	AUC	Accuracy	Precision	Sensitivity	Specificity
XGBoost	Window	0.64	0.62	0.62	0.56	0.68
	Session	0.68	0.72	0.74	0.66	0.78
SVM RBF	Window	0.63	0.63	0.62	0.60	0.65
	Session	0.69	0.69	0.69	0.69	0.69
Logistic regression	Window	0.67	0.63	0.63	0.58	0.69
	Session	0.70	0.68	0.70	0.60	0.75
KNN	Window	0.60	0.55	0.54	0.49	0.62
	Session	0.65	0.68	0.69	0.63	0.72
Random forest	Window	0.60	0.58	0.57	0.55	0.61
	Session	0.66	0.68	0.69	0.63	0.72
SVM linear	Window	0.66	0.63	0.64	0.52	0.73
	Session	0.69	0.65	0.71	0.49	0.81
HistGB	Window	0.58	0.54	0.52	0.52	0.56
	Session	0.62	0.63	0.65	0.57	0.69

Additional results for this low-to-moderate BAC classification setting, including model performance and the most consistently selected features for the top 5 and top 10 feature sets, are provided in A.

4.2.5 All-Data Classification

The all-data setting included sober recordings and all available impaired simulator recordings, regardless of BAC level. This made the impaired class less consistent, since it contained recordings with both lower and higher BAC values. In addition, the all-data setting introduced class imbalance, since more impaired than sober recordings were available per participant.

Class imbalance was handled during training for models with suitable built-in weighting mechanisms. These models include XGBoost, LR, SVM, and RF. Models without equivalent weighting support in the current setup, such as KNN, were retained as unweighted baselines. Class weighting did not uniformly improve model performance. Instead, it mainly altered the trade-off between sensitivity and specificity.

Table 4.5 shows the model performance in the all-data setting. The models generally showed high sensitivity but reduced specificity, indicating that they often identified impaired recordings at the cost of more FP classifications. This was most evident for the linear SVM, which achieved the highest session-level accuracy of 0.73 and sensitivity of 0.90, but only a specificity of 0.33 at a decision threshold of 0.5.

XGBoost exhibited more conservative decision-making, and its confusion matrix is shown in Figure 4.4. Although the session-level accuracy was lower at 0.71, it achieved the highest specificity of 0.75, indicating fewer FP classifications. However, this came at the cost of lower sensitivity than several other models. XGBoost achieved a session-level AUC of 0.73, with a confidence interval of 0.63 to 0.83.

The best low-to-moderate BAC model yielded a more balanced session-level performance, with an accuracy of 0.72, a sensitivity of 0.66, and a specificity of 0.78. In contrast, the all-data setting either produced a highly imbalanced sensitivity-specificity trade-off, as seen for the linear SVM, or lower overall accuracy, as seen for XGBoost. Overall, these results suggest that combining all impaired recordings into one class increased the variability of the impaired class and made the classification task less stable.

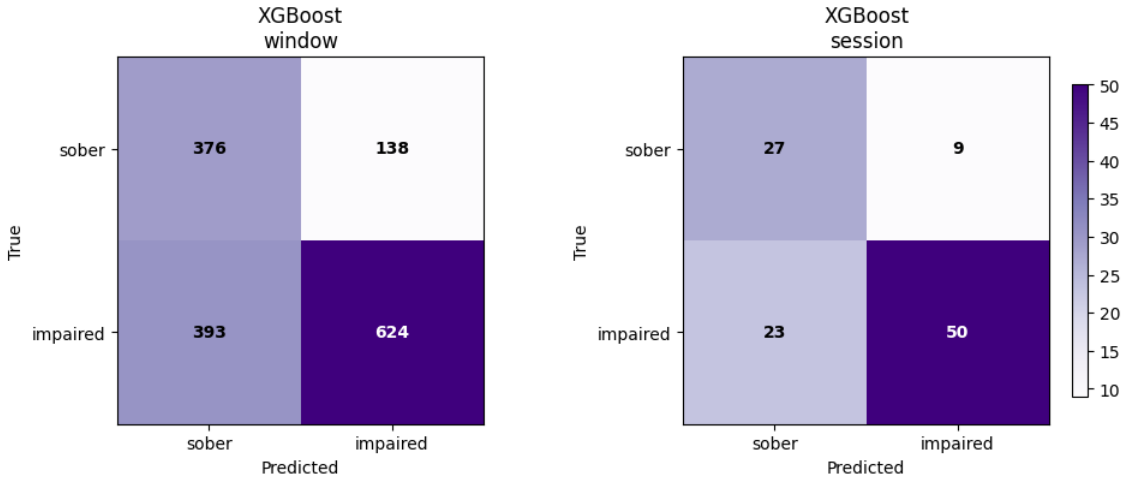

Figure 4.4: Window-level and session-level confusion matrices for XGBoost in the all-data setting using three combined features at threshold 0.5.

Table 4.5: Pooled out-of-fold model performance in the all-data setting using three combined features at decision threshold 0.5.

Model	Level	AUC	Accuracy	Precision	Sensitivity	Specificity
SVM linear	Window	0.72	0.70	0.72	0.89	0.31
	Session	0.74	0.72	0.73	0.90	0.33
Random forest	Window	0.69	0.68	0.73	0.81	0.41
	Session	0.77	0.69	0.72	0.88	0.31
HistGB	Window	0.69	0.67	0.73	0.79	0.42
	Session	0.76	0.70	0.73	0.88	0.33
KNN	Window	0.69	0.67	0.72	0.83	0.35
	Session	0.75	0.65	0.69	0.88	0.19
XGBoost	Window	0.70	0.65	0.82	0.61	0.73
	Session	0.73	0.71	0.85	0.68	0.75
SVM RBF	Window	0.69	0.70	0.77	0.78	0.53
	Session	0.73	0.66	0.73	0.79	0.39
Logistic regression	Window	0.71	0.66	0.77	0.68	0.61
	Session	0.72	0.66	0.78	0.68	0.61

4.2.6 External Real-World Test

For the following analysis, the external real-world driving data was used solely as an independent test set, while the simulator data was used for training. The external test set was evaluated using the high-BAC setting, resulting in 17 sober and 17 impaired sessions.

Table 4.6 shows the results obtained for each model. For the external dataset, results are reported using a decision threshold of 0.6, since this threshold provided a better balance between sensitivity and specificity in this evaluation. At the session level, the linear SVM and XGBoost achieved the highest accuracy, both at 0.74, with XGBoost achieving the best AUC of 0.75, with a confidence interval between 0.58 and 0.91. The best models for the external dataset with their corresponding confidence intervals can be seen in Table 4.2.
The two models showed different trade-offs between sensitivity and specificity. XGBoost produced a more balanced result, with a sensitivity of 0.71 and a specificity of 0.76. In contrast, the linear SVM achieved a higher specificity of 0.88, corresponding to only two FPs, but a lower sensitivity of 0.59. The confusion matrices for these two models are shown in Figures 4.6 and 4.5, illustrating differences in their classification behavior.

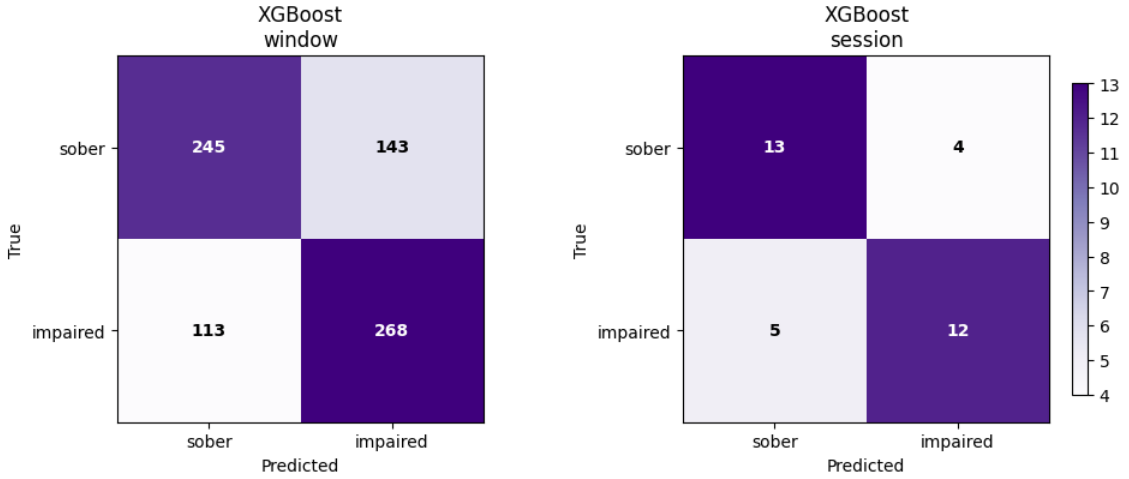

Figure 4.5: Confusion matrix for XGBoost trained on simulator data and tested on external real-world driving data at decision threshold 0.6.

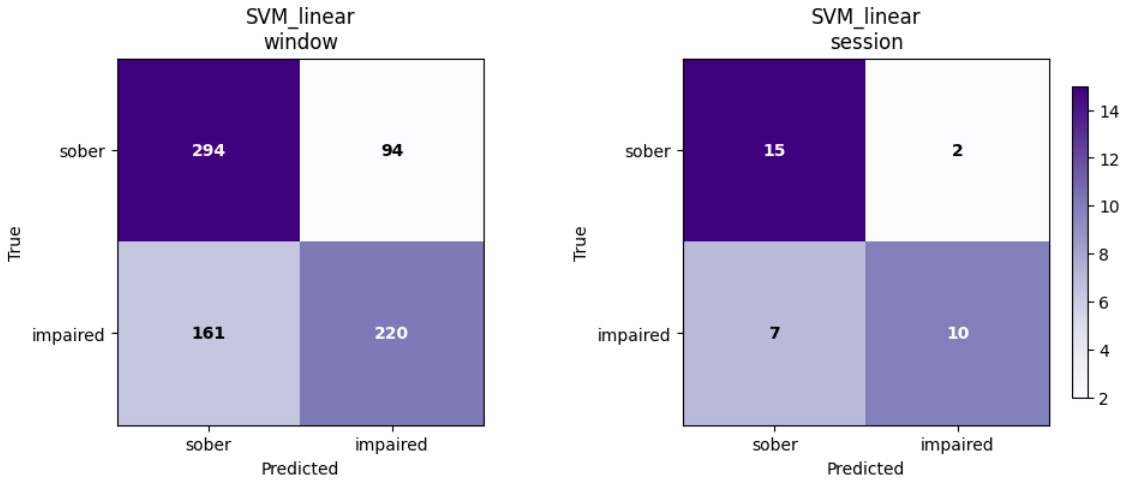

Figure 4.6: Confusion matrix for the linear SVM trained on simulator data and tested on external real-world driving data at decision threshold 0.6.

Table 4.6: Pooled model performance on the external real-world test set using three combined features at decision threshold 0.6.

Model	Level	AUC	Accuracy	Precision	Sensitivity	Specificity
SVM RBF	Window	0.67	0.66	0.64	0.70	0.61
	Session	0.76	0.65	0.63	0.71	0.59
XGBoost	Window	0.70	0.67	0.65	0.70	0.63
	Session	0.75	0.74	0.75	0.71	0.76
KNN	Window	0.67	0.63	0.61	0.71	0.55
	Session	0.75	0.68	0.65	0.76	0.59
HistGB	Window	0.64	0.60	0.58	0.69	0.51
	Session	0.74	0.62	0.60	0.71	0.53
Random forest	Window	0.64	0.61	0.59	0.66	0.55
	Session	0.72	0.65	0.63	0.71	0.59
SVM linear	Window	0.69	0.67	0.70	0.58	0.76
	Session	0.72	0.74	0.83	0.59	0.88
Logistic regression	Window	0.69	0.67	0.68	0.64	0.71
	Session	0.72	0.68	0.71	0.59	0.76

Results using five and ten selected features are provided in A.

4.3 Selected Features

Feature selection was performed separately within each CV fold using only the training data, thereby avoiding information leakage from the held-out participant. The selected features were then compared across folds to identify the features that were most consistently selected. For the external real-world test set, the final feature set was selected using only the simulator training data and was then applied unchanged to the external data.

Table 4.7 summarizes the selected features for the main classification settings. The comparison between manual features, tsfresh features, and the combined feature set was performed for the high-BAC setting. Based on this comparison, the combined feature representation was used for the remaining classification settings.

Table 4.7: Selected features across the main classification settings using three selected features.

Setting	Feature set	Selected feature

High-BAC	Combined	`log_hr_rmssd_ratio`
		`ibi_ms__fft_aggregated__aggtype_"skew"`
		`ibi_ms__absolute_sum_of_changes`

High-BAC	Manual only	`log_hr_rmssd_ratio`
		`sdnn`
		`sd2`

High-BAC	tsfresh only	`ibi_ms__fft_aggregated__aggtype_"skew"`
		`ibi_ms__absolute_sum_of_changes`
		`ibi_ms__standard_deviation`

Low-to-moderate BAC	Combined	`ibi_ms__fft_aggregated__aggtype_"skew"`
		`ibi_ms__cid_ce__normalize_False`
		`log_hr_rmssd_ratio`

All-data	Combined	`ibi_ms__cid_ce__normalize_False`
		`rmssd`
		`log_hr_rmssd_ratio`

External real-world	Combined	`log_hr_rmssd_ratio`
		`ibi_ms__fft_aggregated__aggtype_"skew"`
		`ibi_ms__absolute_sum_of_changes`

Across the classification settings, several features were selected repeatedly. The manually derived `log_hr_rmssd_ratio` feature was included in all combined settings indicating that the relationship between HR level and short-term HRV was informative for the classification task.

Among the automatically extracted features, `ibi_ms__fft_aggregated__aggtype_"skew"` and `ibi_ms__absolute_sum_of_changes` were selected in the high-BAC and external real-world settings.

The tsfresh feature `ibi_ms__cid_ce__normalize_False` was selected in both the low-to-moderate BAC and all-data settings, suggesting that signal complexity or irregularity measures became more relevant when lower BAC levels were included. In the all-data setting, RMSSD was also selected, indicating that conventional short-term HRV information remained relevant when all impaired simulator recordings were combined.

Overall, the selected features indicate that both manually derived HRV features and automatically extracted IBI-based time-series features contributed useful information.

4.4 Descriptive HR and HRV Changes

To provide a descriptive overview of how the manually extracted physiological features varied with BAC level, mean feature values were calculated for the sober, low-to-moderate BAC, and high-BAC conditions from the simulator data. The impaired conditions were separated into low-to-moderate BAC and high-BAC recordings, corresponding to the classification settings used in the model evaluation. The results are presented in Table 4.8.
Mean HR increased from 88 ± 19 bpm in the sober condition to 92 ± 18 bpm in the low-to-moderate BAC condition and 100 ± 16 bpm in the high-BAC condition. Most HRV-related features showed a decreasing trend with increasing BAC. RMSSD decreased from 37.0 ± 21.5 ms in the sober condition to 27.2 ± 17.3 ms in the low-to-moderate BAC condition and 16.3 ± 11.0 ms in the high-BAC condition. Similar decreases were observed for SDNN, SD2, pNN50, LF power, and HF power.
In general, these descriptive changes indicate that higher BAC levels were associated with increased HR and reduced short-term HRV in the recorded data.

Table 4.8: Descriptive statistics for BAC, HR, and HRV features across the three experimental conditions. Values are reported as mean ± standard deviation.

Feature	**Sober**	**Low-to-moderate BAC**	**High-BAC**
Participants (n)	37	35	36
Mean BAC (‰)	–	0.47 ± 0.12	1.1 ± 0.25
HR mean (bpm)	88 ± 19	92 ± 18	100 ± 16
HR median (bpm)	88 ± 20	93 ± 18	100 ± 16
RR-interval mean (ms)	719 ± 176	677 ± 137	617 ± 102
SDNN (ms)	49.6 ± 18.0	41.0 ± 19.8	27.8 ± 13.7
RMSSD (ms)	37.0 ± 21.5	27.2 ± 17.3	16.3 ± 11.0
SD2 (ms)	64.3 ± 21.5	54.3 ± 25.4	37.4 ± 17.8
pNN50 (%)	15.7 ± 16.3	10.1 ± 13.4	3.1 ± 5.5
LF power (ms^2)	1130 ± 701	891 ± 864	443 ± 393
HF power (ms^2)	475 ± 414	302 ± 303	134 ± 149
LF/HF ratio	3.6 ± 2.0	4.5 ± 2.4	5.2 ± 2.1

4.5 Effect of BAC and Decision Thresholds

The results showed that classification performance depended on the BAC threshold used in each setting. Performance was better when the impaired class was restricted to high-BAC recordings and worse when low-to-moderate BAC recordings were included. This indicates that the distinction between the sober and impaired classes was more pronounced at higher BAC levels. When low-to-moderate BAC recordings were included, the separation between the two classes was less distinct.
The classification results were primarily evaluated using a fixed decision threshold of 0.5, which was used to convert predicted model scores into binary class labels. Consequently, decision-based metrics such as accuracy, sensitivity, and specificity

depended on this threshold. In contrast, AUC was threshold-independent because it evaluates the ranking of predicted scores across all possible thresholds.
Changing the decision threshold affected the balance between FP and FN. A higher decision threshold made the classifiers more conservative when assigning the impaired class, generally reducing FP while increasing FN. This trade-off was particularly relevant in the external dataset evaluation and the DMS fusion analysis, where adjusted decision thresholds were also evaluated.

4.6 Late Fusion of Polar H10 and Image-Based DMS Predictions

A late fusion approach was used to combine the prediction scores from the best-performing Polar-based and DMS-based models. Several candidate models were first evaluated separately for each modality, after which the best model from each modality was selected for the fusion experiment. For each window, the final fusion score was computed as the average of the Polar and DMS prediction scores, giving both modalities equal influence on the final decision. For the DMS models, only features derived using tsfresh were used.

4.6.1 High-BAC Classification

By fusing the prediction scores from the best-performing model of each modality in the high-BAC setting, the confusion matrices in Figure 4.7 were obtained. The late fusion model combined the Polar-based and DMS-based SVM linear models' prediction scores using equal weighting, with a decision threshold of 0.5.

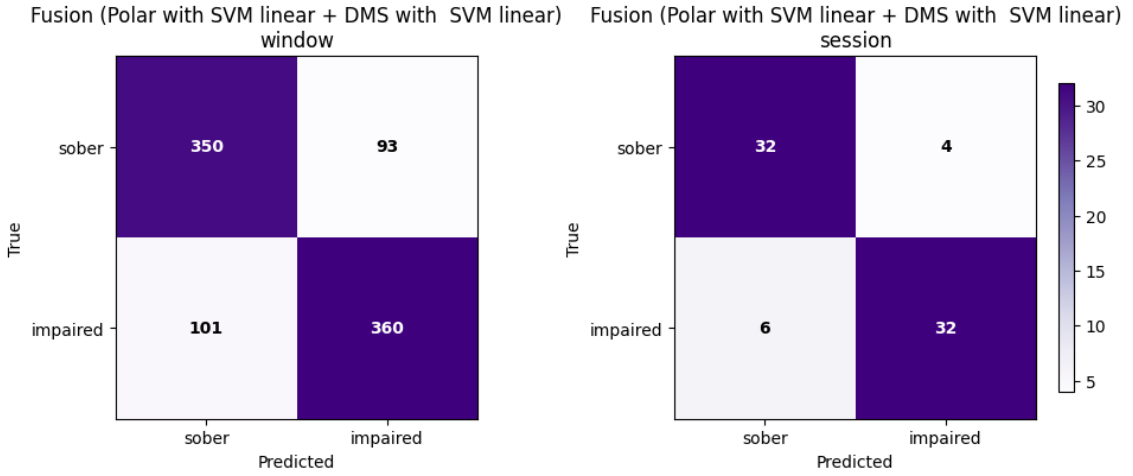

Figure 4.7: Late fusion decision based on modality-specific ensemble models at a decision threshold of 0.5 using three features.

The performance of the best single-modality models and the late fusion model is summarized in Table 4.9. The late fusion model achieved the highest session-level scores for all metrics, indicating that combining the two modalities improved the

overall classification performance. For example, the session-level AUC increased from 0.81 on Polar and 0.87 on DMS to a fused AUC of 0.93 with a confidence interval between 0.87 and 0.98, indicating an overall increase in separability between the two classes.

Table 4.9: Comparison between the best-performing single-modality models and the late fusion model for the high-BAC setting at decision threshold 0.5. Session-level AUC, sensitivity, and specificity are reported with 95% bootstrap confidence intervals.

Model	Level	AUC	Accuracy	Sensitivity	Specificity
Polar	Window	0.79	0.74	0.70	0.79
Polar	Session	0.81 [0.70, 0.90]	0.78	0.74 [0.60, 0.87]	0.83 [0.69, 0.94]
DMS	Window	0.78	0.71	0.75	0.66
DMS	Session	0.87 [0.70, 0.98]	0.82	0.87 [0.61, 0.95]	0.78 [0.56, 0.94]
Fusion	Window	0.87	0.79	0.78	0.79
Fusion	Session	0.93 [0.87, 0.98]	0.86	0.84 [0.71, 0.95]	0.89 [0.78, 0.97]

Note. Values in square brackets denote 95% bootstrap confidence intervals based on 2000 resamples. Performance metrics were computed at a decision threshold of 0.5. Linear SVM models were used for both single-modality inputs. The fusion model combined the prediction scores from the Polar and DMS models.

Table 4.10 summarizes the session-level agreement and disagreement patterns between the Polar, DMS, and late fusion model. The majority of sessions were correctly classified by both modalities. However, several sessions showed disagreement between the Polar and DMS models, indicating that the two modalities made errors on different sessions.

Table 4.10: Session-level agreement and disagreement cases between the Polar, DMS, and the late fusion model for the high-BAC setting.

Case type	Number of sessions
All correct	46
Fusion and Polar correct, DMS wrong	10
Fusion wrong, DMS correct, Polar wrong	7
Fusion and DMS correct, Polar wrong	6
Fusion wrong, Polar correct, DMS wrong	2
All wrong	1

To examine how the modalities contributed to the fused decision, selected session-level cases were analyzed. Table 4.11 shows examples where one modality was correct while the other was incorrect, as well as one case where both modalities failed. Prediction scores above 0.5 correspond to an impaired prediction, while scores below 0.5 correspond to a sober prediction. The examples show that the fused model could produce a correct final decision even when one of the individual modalities

was incorrect. This supports the interpretation that Polar and DMS provide complementary information, since they did not always fail in the same sessions. The rest of the classification cases can be seen in A.

Table 4.11: Examples of session-level classification cases illustrating agreement and disagreement between the Polar-based and DMS-based models in the late fusion experiment for the high-BAC setting when using a decision threshold of 0.5.

Case type	BAC	Polar score	DMS score	Fusion score	Fusion
Polar correct, DMS wrong	0	0.40	0.58	0.49	Correct
Polar correct, DMS wrong	0	0.29	0.67	0.48	Correct
Polar correct, DMS wrong	1.2	0.68	0.43	0.55	Correct
DMS correct, Polar wrong	0	0.66	0.25	0.46	Correct
DMS correct, Polar wrong	0	0.56	0.40	0.48	Correct
DMS correct, Polar wrong	0.96	0.39	0.70	0.54	Correct
Both wrong	1.1	0.42	0.49	0.45	Wrong

An attempt to reduce the number of FP was made by adjusting the decision threshold. In this setting, the Polar-based model was changed from a linear SVM to LR. When the decision threshold was increased from 0.50 to 0.55, the number of FP was reduced to zero, as shown in Figure 4.8.

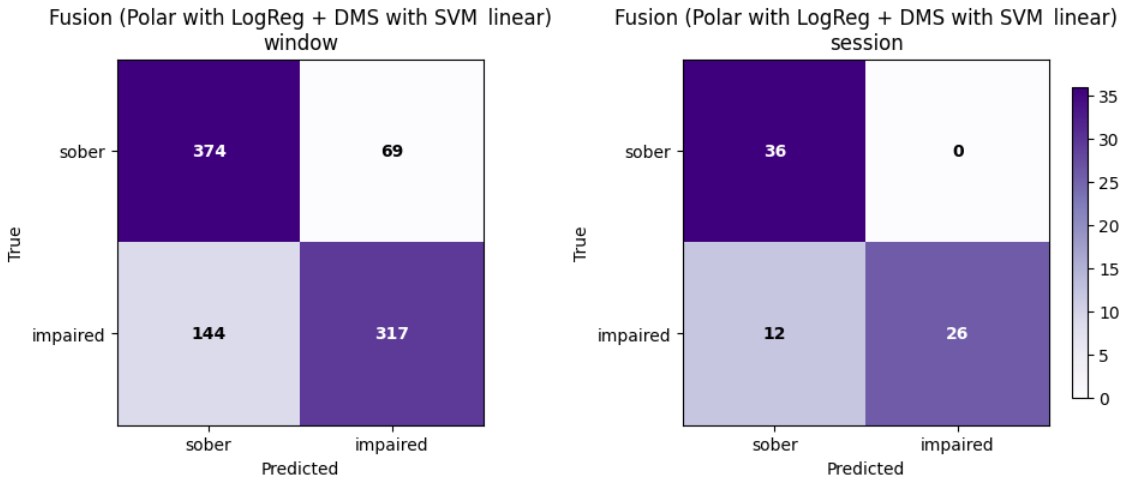

Figure 4.8: Late fusion decision based on modality-specific ensemble models at decision threshold 0.55 using three features.

This result indicates that fine-tuning the decision threshold on the fusion model can reduce FP without substantially reducing the overall accuracy.

4.6.2 Low-to-Moderate BAC Classification

To evaluate whether the same multi-modal fusion strategy was also beneficial for lower levels of alcohol impairment, the late fusion approach was repeated for the low-to-moderate setting. The fused confusion matrix based on the two modality-specific

models is shown in Figure 4.9. The same fusion logic was used as for high-BAC, but now XGBoost performed best for the two modalities when using a decision threshold of 0.5.

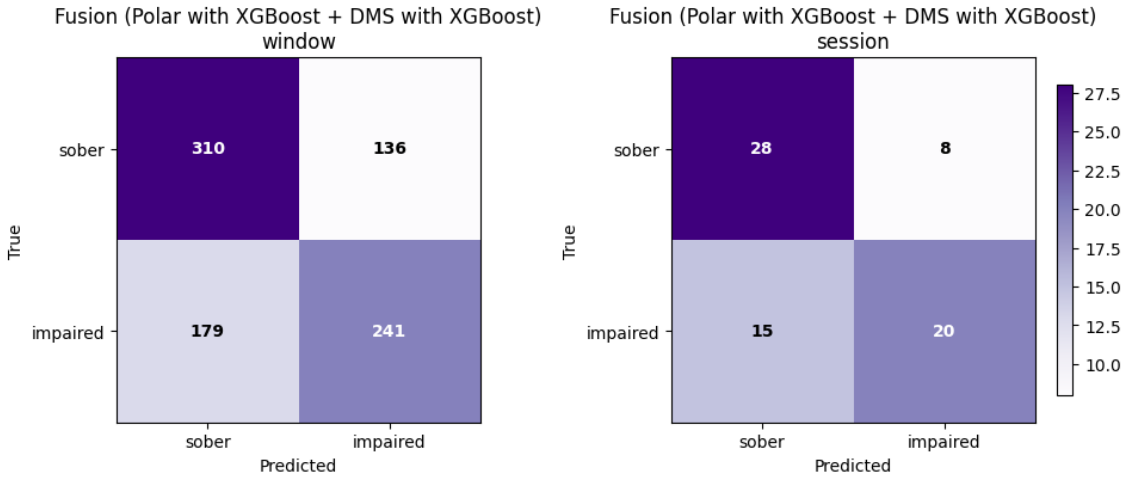

Figure 4.9: Late fusion decision based on modality-specific ensemble models for low-to-moderate BAC setting, using three features and a decision threshold of 0.5.

The performance of the selected Polar model, DMS model, and late fusion model is summarized in Table 4.12. The fused model achieved the highest window-level accuracy, increasing from 0.62 to 0.64, but at the session level, both Polar and DMS accuracy were higher individually than the fused model. This indicates that the benefit of late fusion was more apparent at the window level for the low-to-moderate BAC task. Although the session-level AUC increased from 0.68 for the individual modalities to 0.74 for the late fusion model, this improvement was not reflected in session-level accuracy.

Table 4.12: Comparison of single-modality and late fusion performance for the low-to-moderate BAC setting, using three features and a decision threshold of 0.5.

Model	Level	AUC	Accuracy	Sensitivity	Specificity
Polar	Window	0.64	0.62	0.55	0.68
Polar	Session	0.68 [0.55, 0.81]	0.72	0.66 [0.49, 0.80]	0.78 [0.64, 0.92]
DMS	Window	0.64	0.62	0.55	0.68
DMS	Session	0.68 [0.55, 0.81]	0.72	0.69 [0.51, 0.83]	0.75 [0.61, 0.89]
Fusion	Window	0.69	0.64	0.57	0.69
Fusion	Session	0.74 [0.61, 0.85]	0.68	0.57 [0.40, 0.71]	0.78 [0.64, 0.92]

Note. Values in square brackets denote 95% bootstrap confidence intervals based on 2000 resamples. Performance metrics were computed at a decision threshold of 0.5. XGBoost was used for both the Polar model and the DMS model. The fusion model combined the prediction scores from the Polar and DMS models.

Table 4.13 summarizes how often the modalities agreed or disagreed at the session level. Although both modalities were correct in 29 sessions, there were also several

cases in which only one modality was correct. This shows that the two modalities captured partly different information, but also that the fused decision was not always able to recover the correct class when one modality produced a strongly incorrect score.

Table 4.13: Session-level agreement and disagreement cases between the Polar, DMS, and late fusion models for the low-to-moderate BAC setting.

Case type	Number of sessions
All correct	29
Fusion wrong, Polar correct, DMS wrong	6
Fusion and DMS correct, Polar wrong	3
Fusion and Polar correct, DMS wrong	16
Fusion wrong, DMS correct, Polar wrong	11
All wrong	6

Selected examples of session-level classification cases are shown in Table 4.14. Prediction scores above 0.5 correspond to an impaired prediction, while scores below 0.5 correspond to a sober prediction. The examples show that the two modalities did not always fail in the same sessions. In some cases, the fused score was shifted toward the correct decision by the modality with the more reliable prediction. The rest of the classification cases can be seen in A.

Table 4.14: Examples of session-level classification cases illustrating agreement and disagreement between the Polar-based and DMS-based models using the low-to-moderate setting when using a decision threshold of 0.5.

Case type	BAC	Polar score	DMS score	Fusion score	Fusion
Polar correct, DMS wrong	0	0.39	0.51	0.45	Correct
Polar correct, DMS wrong	0.30	0.67	0.50	0.58	Correct
Polar correct, DMS wrong	0.38	0.65	0.38	0.52	Correct
DMS correct, Polar wrong	0	0.53	0.40	0.46	Correct
DMS correct, Polar wrong	0	0.51	0.41	0.46	Correct
DMS correct, Polar wrong	0.50	0.49	0.63	0.56	Correct
Both wrong	0	0.55	0.55	0.55	Wrong
Both wrong	0.55	0.31	0.46	0.38	Wrong
Both wrong	0.54	0.29	0.47	0.38	Wrong

5
Discussion

This master's thesis investigated whether alcohol-induced driving impairment can be detected using heart-related signals and ML methods. Driving data was collected in both a simulator environment and a real-world driving setup under varying levels of alcohol impairment, while physiological measurements were obtained using a Polar H10 chest strap. In addition to physiological signals, multimodal approaches that combine heart-related features with DMS-based features were evaluated using a late-fusion strategy.

Overall, the results indicate that alcohol-induced impairment is associated with measurable changes in heart-related signals, and that HR- and HRV-derived features contain relevant information for classification. However, classification performance depended on the BAC level, feature representation, and evaluation setting. The highest performance was generally achieved when distinguishing sober driving from higher BAC levels, whereas lower BAC levels resulted in weaker separability between classes. The external real-world evaluation further suggested that some simulator-learned physiological patterns may transfer to real-world driving, although these findings should be interpreted with caution due to the limited test data.

The original experimental setup also included radar-derived vital sign measurements. However, the radar measurements obtained using the available sensor and processing pipeline were insufficiently reliable to serve as a primary physiological input in this study. Consequently, the main modeling was based on pulse belt data from the Polar H10, while the radar results are discussed separately as a methodological limitation.

This chapter discusses challenges with radar-based measurements, the physiological relevance of the observed HR- and HRV-based changes, the influence of BAC level and feature representation on classification performance, and the potential value of combining physiological information with DMS-based features. Methodological considerations related to measurement conditions, labeling strategy, evaluation procedure, and real-world applicability are also addressed, followed by the study's main limitations and suggestions for future work.

5.1 Challenges with Radar-Based Measurements

Although previous studies have shown that radar-based systems can be used for physiological monitoring [9, 10, 11], this performance was not observed in the present experimental setup. The extracted radar-based HR was highly inconsistent when compared with the Polar H10 reference measurements and showed little to no correspondence with the reference HR.

One possible explanation is that the provided processing pipeline was originally developed for real-vehicle conditions, where vibration and accelerometer data may be important for signal interpretation. The present study was conducted in a simulator setting, and it was not possible within the project scope to determine how the processing should be adapted to this environment. Differences between the present setup and the conditions under which the radar system was originally calibrated may also have contributed, although this was not expected to be the main issue since the radar was installed by Emsense specifically for this study.
An attempt was made to improve the radar processing independently. However, this was not pursued further due to time constraints and because the raw radar signal appeared too noisy for reliable analysis within the available time frame.
Consequently, the radar data could not be used as a reliable source of physiological information in this study, and all subsequent analysis was therefore based on the Polar H10 reference data. This means that the original objective of evaluating radar-based heart signal monitoring for alcohol-related driver impairment could not be fulfilled. The radar results should therefore be interpreted as a limitation of the specific radar implementation and processing pipeline used in this thesis, rather than as evidence against radar-based heart signal monitoring in general.

5.2 Main Classification Results

The main classification results indicate that alcohol-induced impairment could be detected from HR and HRV-related features, particularly when the impaired class is restricted to higher BAC levels. The strongest simulator-based performance was obtained in the high-BAC setting, where the best model was SVM linear and achieved a session-level AUC of 0.81 with a confidence interval between 0.70 and 0.90, and a prediction accuracy of 0.78. This suggests that the higher BAC levels were associated with the most distinct physiological changes between sober and impaired classes.
This result is consistent with HR and HRV statistics, in which higher BAC was associated with higher HR and lower HRV. This is also evident in Table 4.8, where the trend between HR and HRV aligns with the HR and HRV statistics. These physiological changes likely contributed to the improved separability observed in the high-BAC classification task.
In contrast, classification performance decreased when the impaired class consisted of low-to-moderate BAC recordings. The best model in this setting achieved a session-level AUC of 0.68 with a confidence interval between 0.55 and 0.81, and a prediction accuracy of 0.72, which indicates a moderate classification ability but weaker separability than in the high-BAC setting. This result is expected, since lower BAC levels are likely to produce smaller and less physiological effects compared to high BAC levels, making them harder to distinguish from sober. This was also reflected in the sensitivity, which decreased from 0.74 in the high-BAC setting to 0.66 in the low-to-moderate BAC setting. The lower sensitivity suggests that some mildly impaired sessions were classified as sober. From an application perspective, this is important because early or mild impairment may be the most difficult to detect reliably using heart features alone.

The all-data setting did not improve performance, despite including more impaired recordings. This may partly be explained by the class imbalance introduced in this setting, where the impaired class contained more recordings than the sober class. As a result, some models may have been biased toward predicting the majority class, leading to high sensitivity but low specificity. This was particularly evident for the linear SVM, which detected most impaired sessions but incorrectly classified many sober sessions as impaired, resulting in a low specificity of 0.33. This indicates that the model was biased toward the impaired class in this setting. However, class imbalance was likely not the only reason for the reduced performance. The impaired class also became more varied, since it included recordings from low, moderate, and high BAC levels. Lower BAC recordings may be physiologically closer to sober driving, whereas higher BAC recordings showed more distinct changes in HR and HRV. This increased within-class variability may have made the classification boundary less stable and reduced the overall separability between sober and impaired driving.

The external real-world test is particularly relevant, since it evaluates whether models trained in the simulator environment can transfer to a different driving domain. The best real-world model was SVM linear, which reached a session-level AUC of 0.72 with a confidence interval between 0.53 and 0.88, and a prediction accuracy of 0.74. Compared with XGBoost, the linear SVM showed higher specificity but lower sensitivity, indicating a more conservative classification behavior with fewer FP but more missed impaired sessions. XGBoost showed more balanced sensitivity and specificity but produced more FPs. This highlights that accuracy alone is insufficient for evaluating model performance, as the balance between FP and FN is highly relevant in driver monitoring applications.

These results suggest that some alcohol-related physiological patterns learned from the simulator data were transferable to real-world driving conditions. However, the external test should be interpreted with caution due to the limited sample size of the test data, wide confidence intervals, and domain shift between the simulator and real-world driving. Therefore, the external test should be viewed as a preliminary study of generalization rather than a definitive validation.

By using bootstrap resampling to estimate confidence intervals, the uncertainty of the performance estimates could be assessed more clearly than from point estimates alone. These intervals should not be interpreted as definitive evidence of how the models would perform in a larger population, but they indicate how sensitive the results are to variations in the available data.

As presented in Table 4.2, there is a large variance in the AUC for several of the settings, leading to an overlap between models. This could indicate that the models could perform relatively similarly on a different dataset. For example, the confidence interval for the AUC in the low-to-moderate BAC setting at the session level for the best model was between 0.55 and 0.81, while in the real-world dataset it was between 0.53 and 0.88. These results are very similar, which suggest that the two different setting may have relatively similar discriminative ability depending on which observations are included in each bootstrap sample. The small differences in point estimates should not be overinterpreted when the uncertainty intervals are wide and overlapping.

The difference between the confidence intervals for the high-BAC and real-world

driving settings indicates that the high-BAC setting exhibited more stable classification performance under bootstrap resampling. The high-BAC model achieved an AUC of 0.81 with a confidence interval of [0.70, 0.90], whereas the real-world driving setting with SVM had an AUC of 0.72 with a wider interval of [0.53, 0.88]. This suggests that the high-BAC setting had a more consistent representation of alcohol-related physiological differences compared to the real-world setting.
This suggests that the definition of the classification task, the BAC threshold, and the quality of the physiological input may have had a greater influence on performance than the choice among the tested model types.

5.2.1 Models

Several models achieved similar performance in the high-BAC settings, including linear SVM, SVM with RBF kernel, LR, and XGBoost. This suggests that the selected features contained information that could be captured by both linear and nonlinear classifiers. The fact that simpler models, such as LR and linear SVM, performed comparably to more flexible models indicates that the separation between sober and high-BAC recordings was partly captured by relatively simple decision boundaries.
More flexible models, such as RF and HistGB, did not consistently outperform the simpler classifiers. One possible explanation could be the limited dataset, which may restrict complex models' ability to learn robust patterns without overfitting. Another explanation is that the input data was based solely on RR-interval windows and features derived from them. Although RR intervals contain relevant information about HR and HRV, they represent a relatively limited physiological input space. Therefore, the more complex models may not have had access to sufficient additional information to learn more detailed patterns than simpler classifiers. An exception to this is the XGBoost classifier, which, despite the small dataset, still performs comparably to the simpler classifiers. This could be because XGBoost employs robust regularization mechanisms that prevent overfitting, making it highly effective even on smaller, less diverse datasets.
The results also show that the choice of model should depend on the intended application. A model with high sensitivity may be preferred if the goal is to detect as many impaired drivers as possible, even at the cost of more false alarms. However, in a real driver monitoring system, too many FPs could reduce user acceptance.

5.2.2 Effect of BAC and Decision Threshold

One important methodological consideration is how the impaired class was defined. In this study, all windows with BAC above 0.0‰ were labeled as impaired in the all-data analysis. However, low BAC levels may not necessarily correspond to clear functional impairment, and could alternatively be described as alcohol-positive or alcohol-exposed rather than impaired. The lowest recorded BAC in the present dataset was 0.24‰, meaning that the alcohol-positive class did not include near-zero BAC values. Still, the interpretation of low-BAC classifications should be made with caution, since physiological differences from the sober condition may be smaller

at these levels.
The decision threshold was not extensively optimized in the main analysis, since changing it does not alter the underlying separability of the model's predictions. Instead, it changes how predicted scores are converted into binary classifications. For this reason, AUC was used as an important threshold-independent metric, whereas sensitivity and specificity were interpreted relative to the chosen threshold.
Threshold adjustment was mainly relevant for the external dataset and the DMS fusion analysis. In the external dataset, differences between simulator and real-world driving may have shifted the distribution of predicted scores, making the default threshold of 0.5 less suitable. In the DMS fusion analysis, small changes in the decision threshold affected the FP-FN trade-off. For example, increasing the threshold from 0.5 to 0.55 eliminated the FPs in that setting and achieved a specificity of 1.0. However, this should be interpreted with caution, since the dataset was small and threshold tuning can easily overfit the available observations.

5.2.3 Features

The results showed that some features performed uniformly best across all models and datasets.
One of the exploratory ratio-based features, `log_hr_rmssd_ratio`, showed promising results and was selected in several of the evaluated models and data splits. This suggests that the relative relationship between HR level and short-term HRV contained relevant information for the classification task. Physiologically, this is reasonable, since increased HR is often associated with reduced HRV, which is reflected by lower RMSSD values. The feature therefore combines two related but complementary aspects of cardiac regulation into a single measure.
In contrast, the second ratio-based feature, `log_rmssd_sdnn_ratio`, did not appear to provide the same predictive value. This may indicate that the relationship between short-term variability and overall variability was less informative for distinguishing between the investigated conditions in this dataset.
The results obtained using only automatically extracted tsfresh features were similar to those obtained when manual and automatic features were combined. This suggests that some of the information captured by the manually engineered ratio feature may also have been represented indirectly by one or more of the automatically extracted features. However, since the exact correspondence between individual tsfresh features and the manually derived ratio was not investigated in detail, this interpretation should be considered tentative.

5.3 Late Fusion of Polar H10 and Image-Based DMS Predictions

The late fusion results suggest that the two modalities contained partly complementary information for detecting alcohol-induced impairment. In the high-BAC setting, combining the modality-specific prediction scores improved the overall classification performance, and the number of FP could be reduced to zero after only a

minor adjustment of the decision threshold. This indicates that the fusion approach provided a more robust estimate than either modality alone in this setting.
However, the same improvement was not observed in the low-to-moderate BAC setting. This is likely because the physiological and behavioral differences between sober and mildly impaired driving were less pronounced, making the classification task more difficult. In addition, the individual modality models performed more weakly in this setting, limiting the benefit of combining their predictions. Since late fusion relies on modality-specific models to provide informative and reasonably stable scores, fusion is unlikely to improve performance when both modalities are uncertain or noisy.
These results indicate that simple score-level fusion may be useful when the underlying modalities contain sufficiently discriminative information, as observed in the high-BAC setting. For lower impairment levels, however, more advanced fusion strategies, additional features, or larger datasets may be required to achieve a clear improvement.

5.4 Relation to Previous Research

The physiological trends observed in the present study are generally consistent with previous research on acute alcohol intake and ANS response. During the data collection sessions, HR tended to increase, while HRV tended to decrease as BAC levels increased. This agrees with earlier studies showing that alcohol consumption is associated with increased HR and reduced HRV [3, 4, 5]. These findings support the physiological rationale for using HR- and HRV-based features as indicators of alcohol-related impairment.
When comparing the classification results with previous studies on alcohol-impairment detection, one particularly relevant study is [6]. Similar to the present study, they investigated ML-based detection of alcohol-related impairment using physiological signals. However, their input data included a broader set of physiological modalities, namely ECG, PPG, and EDA. PPG and EDA are additional physiological measurement modalities that can capture information on cardiovascular activity and ANS responses, respectively [56, 57]. This multimodal setup may explain why they reported stronger classification performance, with both XGBoost and GB achieving 88% accuracy. ECG, PPG, and EDA can capture different aspects of the physiological response to alcohol. In contrast, the present study mainly focused on HR- and HRV-derived information, which provides a more limited representation of the physiological state.
A further difference is the experimental setting. The study by [6] was conducted in a driving simulator, whereas the present study included both simulated and real-world driving. Simulator data allow for more controlled measurement conditions, but may not fully capture the variability, motion artifacts, environmental changes, and behavioral complexity present during real-world driving. The lower classification performance in the present study may therefore partly reflect the more challenging and practically relevant measurement conditions. From an application perspective, this is important, since a driver monitoring system must operate reliably outside controlled simulator environments.

Another relevant study is [7], which investigated alcohol-impairment detection using ECG signals and reported promising classification performance. However, their experimental setup differed substantially from both [6] and the present study. Participants were seated and held their hands on a steering wheel, rather than performing a complete driving task. The measurements were therefore obtained under relatively stationary and controlled conditions. In addition, chest-mounted wet ECG electrodes were used, which provide high-quality cardiac signals but are less representative of unobtrusive in-vehicle monitoring. These conditions likely made heartbeat detection and feature extraction more reliable than in the present study.
The level at which classification was performed also differs between studies. In [7], classification was performed on heartbeat-level samples, meaning that each heartbeat or short cardiac segment could contribute as an individual sample. This can increase the number of available samples substantially, but it may also make the classification task less comparable to subject-independent impairment detection. Heartbeats from the same individual are not fully independent, and heartbeat-level classification may partly capture subject-specific physiological characteristics rather than generalizable alcohol-related effects. In the present study, leave-one-subject-out cross-validation was used to evaluate subject-independent generalization, which is a stricter and more practically relevant evaluation strategy for detecting impairment in previously unseen drivers.
The feature representations also differed between studies. The study by [7] included detailed ECG morphology features, which can describe changes in waveform shape in addition to timing-related information. Such features require high-quality ECG recordings. In contrast, the present study focused on HR- and HRV-based features, which are more directly transferable to unobtrusive sensing approaches but may contain less detailed physiological information.
Overall, previous studies support the feasibility of detecting alcohol-related impairment using physiological measurements, but the results are not directly comparable due to differences in signal modalities, experimental conditions, feature extraction, classification levels, and validation strategies. Compared with earlier work, the present study used a more limited physiological input but evaluated impairment detection under more practically relevant conditions, including real-world driving and subject-independent cross-validation. The results therefore provide a more conservative estimate of performance, while highlighting both the potential and limitations of HR- and HRV-based impairment detection for future driver monitoring systems.

5.5 Influence of Measurement Conditions on HR and HRV

Upon arrival, many participants reported feeling nervous, partly because they did not know what to expect. This seemed to be related to both the office environment and the simulator setup. The nervousness may have resulted from being in a new setting with many impressions, as well as from the feeling that they needed to behave seriously in the experimental environment.
This stress and nervousness likely affected the measured signals. As discussed in the

background section, stress can influence HR and HRV in ways that may resemble some of the physiological effects associated with alcohol consumption. In addition, a white-coat HR effect, as described in the background section, may also have influenced the results, since the measurement context itself can affect cardiovascular signals.
The relatively high HR values observed across all conditions suggest that the experimental setting may have influenced the absolute physiological measurements. However, the extent and timing of this effect cannot be determined from the present data. It may have affected all recordings to some degree, resulting in generally elevated HR values. Alternatively, the effect may have been strongest during the sober recording, since this was the first driving session and participants were likely most unfamiliar with the setup. If so, the physiological difference between the sober and impaired conditions may have been underestimated, making classification more difficult and potentially contributing to poorer model performance.
Furthermore, the recordings were obtained during simulator driving rather than during rest, meaning that task-related factors such as attention, concentration, and mental workload may also have contributed to elevated HR and altered HRV.
Despite this uncertainty, the descriptive results in Table 4.8 show a consistent pattern of increasing HR and decreasing HRV with higher BAC levels. This suggests that although stress-related variability may have influenced the absolute values, the overall physiological trends were still consistent with the expected alcohol-related changes described in the literature [3, 4, 5]. The results should therefore be interpreted primarily as relative changes across BAC levels rather than as exact estimates of resting physiological values.
One possible approach to reduce stress-related variability is to introduce an acclimatization period before the first driving task, allowing participants to adapt to the environment before physiological measurements are used for analysis. Another option would be to include a second sober session later in the protocol, making it possible to distinguish initial stress effects from BAC-related changes more clearly. However, due to time constraints during data collection, neither approach was implemented.

5.6 Challenges with Impairment Detection

One of the main challenges in impairment detection using heart-related signals is the substantial inter-individual variation among participants. Heart-related signals are also affected by several external and physiological factors, such as stress, caffeine, and nicotine [20, 17, 15, 16, 18].
One possible approach to reduce the effect of inter-individual variation would be to collect a sober baseline for each driver before driving. This would allow the system to compare the driver to their own normal physiological state, rather than relying solely on absolute HR- and HRV-derived values from the general population. However, the baseline measurement itself could still be influenced by factors such as stress, caffeine, nicotine, sleep quality, physical activity, and the measurement context. Even so, an individual baseline could make these factors easier to handle, since later measurements could be interpreted relative to the driver's own sober

reference rather than as isolated absolute values. This was not implemented in this thesis, as the aim was to evaluate a more general approach applicable to new drivers without requiring individual calibration or prior baseline data.
Another challenge is that alcohol-related impairment is not only determined by BAC, but also by individual tolerance, current state, and contextual factors. Two drivers with similar BAC levels may therefore show different levels of functional impairment and different physiological responses. This makes it difficult to define a universal physiological threshold for impairment based only on HR- and HRV-derived features. In practice, such models should therefore be interpreted as estimating the probability of alcohol-related physiological deviation, rather than providing a definitive measurement of impairment.

5.7 Study Limitations

The study included 37 participants in the simulator dataset, a relatively small sample size that limits the generalizability of the findings. In addition, the participants had a narrow age range, and differences in factors such as age, sex, ethnicity, health status, and fitness level were not fully represented. While the dataset was sufficient for the analyses carried out in this thesis, a larger and more diverse study population would be needed to draw stronger conclusions about the broader population.
The external real-world evaluation was based on a smaller subset of 17 participants, which further limits the strength of the conclusions that can be drawn from this part of the study. The limited sample size also means that variation related to age, sex, individual physiology, driving behavior, and alcohol response may not be fully captured in the real-world dataset. Therefore, the real-world results should be interpreted as a preliminary indication of transferability from the simulator to real-world driving, rather than as definitive validation of model performance under naturalistic driving conditions.
The main experiments in this study were conducted in a stationary driving simulator. This provided a controlled and repeatable environment, but also limited the realism of the driving conditions. Compared to real-world driving, the simulator did not include vehicle motion, road vibrations, traffic complexity, environmental variation, or other disturbances that may affect both physiological signals and driver behavior. These differences may limit the transferability of the simulator-based results to real-world driving, although the separate real-world evaluation provided an initial assessment of this transfer.
The radar sensor was integrated into the seat backrest, providing a controlled and repeatable sensor position relative to the participant. However, since the radar signals were not used in the final analysis, the robustness of this setup under more variable real-world conditions could not be assessed.
While the chest-worn Polar H10 served as a suitable reference sensor in this study, it is not a clinical multi-lead ECG system, which may introduce minor measurement uncertainties. Such uncertainties may have introduced minor noise into the HR- and HRV-based classification models.
This study was based solely on RR intervals, which represent a limited physiological input signal. Additional physiological measurements could potentially have provided

a broader basis for classification.
Another limitation is that several factors that may influence HR and HRV were not systematically controlled or analyzed. These include sleep quality, caffeine and nicotine intake, stress levels, physical activity before the experiment, medication use, hydration, and individual fitness level. Since such factors can affect heart-related signals independently of alcohol intake, they may have contributed to variability in the classification results.

5.8 Relevance for Euro NCAP and Practical Implementation

A relevant aspect of the Euro NCAP 2026 Driver Engagement protocol is that impairment detection is treated as a scored driver monitoring function rather than as an immediate start-up requirement. According to the protocol, the DMS shall be active by default and continuously monitor the driver state. However, for impairment-related driver states, a learning period of up to 10 minutes at a speed of at least 50 km/h is permitted from the start of each journey [8].
From a safety perspective, this raises an important limitation. Alcohol-induced impairment may already be present before the vehicle starts moving, meaning that the driver may enter traffic while impaired. If a system requires up to ten minutes of driving before impairment can be detected, the vehicle may already have been exposed to an elevated crash risk during the initial part of the journey. This is particularly relevant to the present study, which is motivated by the growing emphasis on impairment detection in driver-monitoring assessment frameworks. Although the Euro NCAP protocol permits a learning period for scoring purposes, an ideal impairment detection system should aim to identify impairment as early as possible, preferably before driving begins.
In this context, the 10-minute learning period should not be interpreted as the desired detection time for alcohol-induced impairment, but rather as an allowance within the assessment framework. The present thesis follows the Euro NCAP recommendations by investigating driver impairment detection using driver-state-related signals. However, the long-term safety objective is to reduce detection delays and enable earlier identification of impaired drivers.
Another practical consideration is how a driver monitoring system should respond once alcohol-related impairment has been detected. From an industrial and safety perspective, detection alone is insufficient; the system also requires a safe and acceptable intervention strategy. An immediate stop may not be appropriate in all situations, since stopping abruptly or in an unsafe location could introduce new risks. A more realistic approach could be a gradual intervention strategy in which the driver is first warned and encouraged to stop safely. If the driver does not respond, the system could limit vehicle speed, increase the intensity of warnings, or guide the vehicle toward a safer stopping location if automated driving functions are available.
This highlights that impairment detection should be considered as part of a broader safety system rather than as an isolated classification task. The appropriate response

would likely depend on the driving context, road environment, vehicle automation level, and confidence of the impairment detection. Therefore, future work should not only focus on improving classification performance but also on translating detected impairments into safe, reliable, and user-acceptable vehicle interventions.

5.9 Implications for Driver Monitoring and Future Directions

For driver monitoring applications, classification accuracy alone is insufficient to evaluate practical usefulness. FP are particularly important because incorrectly classifying a sober driver as impaired may lead to unnecessary warnings and reduced trust in the system. Therefore, differences in FP rate and specificity should be considered when comparing the present study with previous work. A model with high accuracy but many FPs may be less suitable for deployment than a model with slightly lower sensitivity but better control of false alarms. This is especially relevant in real-world driving, where the system may be used repeatedly and false alarms could negatively affect user acceptance.

The results of this study indicate that physiological signals may provide useful complementary information for detecting alcohol-related impairment, but HR-based features cannot be used as the sole source for intoxication assessment; there would be an unacceptable level of FP (i.e., situations where the vehicle would be required to stop if alcohol impairment is detected). When combined with existing DMS, heart signals could improve robustness and reliability by adding information that is less dependent on visibility conditions, head pose, or camera occlusions. The DMS fusion results showed improved classification performance, primarily reflected in a reduced number of FP.

At the same time, the study highlights that classification robustness is crucial for real-world applications. A practical DMS must be reliable enough to minimize both FP and FN. Drivers should not feel anxious about being incorrectly classified as impaired when sober, while truly impaired cases must also not be overlooked. From this perspective, contactless heart signal monitoring is particularly attractive, since requiring drivers to wear a pulse belt or electrodes is not realistic in everyday driving. Although the radar-derived signals in the present study were not sufficiently reliable, alcohol-impairment detection based on radar-derived heart signals may still be feasible with more robust sensing and signal processing.

Future work should therefore focus on improving contactless heart signal monitoring and validating such methods under more realistic driving conditions. This includes improved radar-based sensing and signal processing, larger and more diverse datasets, better control of confounding factors, and further integration of physiological sensing with camera-based DMS systems. In particular, future studies should include a broader participant population with greater variation in age, sex, physiology, fitness level, and alcohol response to improve the generalizability of the results. The experimental design could also be improved by including an acclimatization or practice driving session before the first recorded sober session. This could reduce the influence of initial nervousness and white-coat HR effects on the physiological

measurements.
Future studies should also consider more balanced strategies for dataset construction, especially when combining multiple BAC levels into a single impaired class. This could include collecting a more equal number of sober and impaired recordings, using class-weighting or resampling methods during model training, and evaluating how different BAC thresholds affect the balance between sensitivity and specificity. Such analyses would make it easier to assess whether model performance reflects robust alcohol-related patterns rather than class imbalance or threshold-specific effects.

6
Conclusion

This master's thesis provides a proof of concept for detecting alcohol-induced driver impairment using heart signal monitoring and machine learning. The study highlights both the potential and the technical challenges of integrating heart signal analysis into future driver monitoring systems (DMS).

The classification results show that high blood alcohol concentrations ($\geq$0.7‰) can be detected with reasonable performance, achieving a session-level AUC of 0.81 using a linear SVM model. However, identifying low-to-moderate impairment remains challenging, since heart signal deviations at these levels are less distinct from sober baselines. Models trained on simulator data also showed promising generalizability when tested on an external real-world driving dataset of 17 participants, reaching an AUC of 0.75.

The strongest result in this study was obtained through late fusion of multimodal data. Combining simulator-based heart signals with camera-based DMS outputs improved performance to an AUC of 0.93 and indicated that false positives could be reduced to zero with minor tuning of the decision threshold. This suggests that heart signals may not be sufficient as a standalone tool for intoxication assessment, but could provide a valuable complement to existing DMS. Such robustness is relevant for future systems aligned with Euro NCAP 2026 safety requirements for alcohol-induced impairment detection.

Future work should focus on developing reliable contactless heart signal monitoring systems, since wearable sensors are not practical for everyday driving applications. Future studies should also include larger, more diverse participant populations, along with better control of confounding factors, to improve the generalizability of the results.

Bibliography

[1] World Health Organization. (2023, Jan.) No level of alcohol consumption is safe for our health. Accessed: 2026-01-21. [Online]. Available: https://www.who.int/europe/news/item/04-01-2023-no-level-of-alcohol-consumption-is-safe-for-our-health

[2] European Commission, "Alcohol - mobility & transport - road safety," https://road-safety.transport.ec.europa.eu/eu-road-safety-policy/priorities/safe-road-use/alcohol_en, 2024, accessed: 2026-01-19.

[3] P. Koskinen, J. Virolainen, and M. Kupari, "Acute alcohol intake decreases short-term heart rate variability in healthy subjects," *Clinical Science (London)*, vol. 87, no. 2, pp. 225–230, 1994.

[4] P. van de Borne, A. L. Mark, N. Montano, D. Mion, and V. K. Somers, "Effects of alcohol on sympathetic activity, hemodynamics, and chemoreflex sensitivity," *Hypertension*, vol. 29, no. 6, pp. 1278–1283, 1997.

[5] F. Weise, D. Krell, and N. Brinkhoff, "Acute alcohol ingestion reduces heart rate variability," *Drug and Alcohol Dependence*, vol. 17, no. 1, pp. 89–91, 1986.

[6] S. H. Kim, H. W. Son, T. M. Lee, and H. J. Baek, "Drunk driver detection using multiple non-invasive biosignals," *Sensors*, vol. 25, no. 5, p. 1281, 2025. [Online]. Available: https://doi.org/10.3390/s25051281

[7] C. K. Wu, K. F. Tsang, H. R. Chi, and F. H. Hung, "A precise drunk driving detection using weighted kernel based on electrocardiogram," *Sensors*, vol. 16, no. 5, p. 659, 2016. [Online]. Available: https://www.mdpi.com/1424-8220/16/5/659

[8] Euro NCAP, "Safe driving: Driver engagement protocol, version 1.1," Oct. 2025, implementation from 2026. [Online]. Available: https://cdn.euroncap.com/cars/assets/euro_ncap_protocol_safe_driving_driver_engagement_v11_a30e874152.pdf

[9] P. Chen, Y. Du, W. Huang, Y. Bai, S. Wei, and J. Wang, "Vehicle occupant detection and vital sign monitoring based on random body movement correction using millimeter-wave radar," *IEEE Sensors Journal*, vol. 25, no. 12, pp. 21 945–21 957, 2025.

[10] O. Andersson and J. Nystrand, "Radar based in-vehicle health monitoring: Using fmcw radar for breathing and heartbeat detection," Master's thesis, Chalmers University of Technology, Gothenburg, Sweden, 2021, master's Thesis in Biomedical Engineering. [Online]. Available: https://odr.chalmers.se/server/api/core/bitstreams/4e845613-fc22-496e-85f3-74bce3bea26d/content

[11] Cadence, "Radar-based vital signs detection for in-cabin applications," Cadence Design Systems, White Paper, Oct 2024. [Online].

Available: https://www.cadence.com/en_US/home/resources/white-papers/radar-based-vital-signs-detection-for-in-cabin-wp.html

[12] A. Kanakapura Sriranga, Q. Lu, and S. A. Birrell, "A deep learning-based contactless driver state monitoring radar system for in-vehicle physiological applications," *IEEE Transactions on Intelligent Transportation Systems*, vol. 26, no. 7, pp. 9491–9499, 2025.

[13] H. U. R. Siddiqui, A. A. Saleem, R. Brown, B. Bademci, E. Lee, F. Rustam, and S. Dudley, "Non-invasive driver drowsiness detection system," *Sensors*, vol. 21, no. 14, 2021.

[14] H. U. R. Siddiqui, A. A. Saleem, D. Brown, and M. Faisal, "Ultra-wide band radar empowered driver drowsiness detection with convolutional spatial feature engineering and artificial intelligence," *Sensors*, vol. 24, no. 12, p. 3754, 2024.

[15] J. Koenig, M. N. Jarczok, W. Kuhn, K. Morsch, A. Schäfer, T. K. Hillecke, and J. F. Thayer, "Impact of caffeine on heart rate variability: A systematic review," *Journal of Caffeine Research*, vol. 3, no. 1, pp. 22–37, 2013.

[16] J. L. Flueck, F. Schaufelberger, M. Lienert, D. Schäfer Olstad, M. Wilhelm, and C. Perret, "Acute effects of caffeine on heart rate variability, blood pressure and tidal volume in paraplegic and tetraplegic compared to able-bodied individuals: A randomized, blinded trial," *PLOS ONE*, vol. 11, no. 10, p. e0165034, 2016.

[17] H. P. Sondermeijer, A. G. J. van Marle, P. Kamen, and H. Krum, "Acute effects of caffeine on heart rate variability," *The American Journal of Cardiology*, vol. 90, no. 8, pp. 906–907, 2002.

[18] N. Sjoberg and D. A. Saint, "A single 4 mg dose of nicotine decreases heart rate variability in healthy nonsmokers: implications for smoking cessation programs," *Nicotine & Tobacco Research*, vol. 13, no. 5, pp. 369–372, 2011, accessed: 2026-01-21. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/21350044/

[19] C. B. Harte and C. M. Meston, "Effects of smoking cessation on heart rate variability among long-term male smokers," *International Journal of Behavioral Medicine*, vol. 21, pp. 302–309, 2014, accessed: 2026-01-22. [Online]. Available: https://link.springer.com/article/10.1007/s12529-013-9295-0

[20] H.-G. Kim, E.-J. Cheon, D.-S. Bai, Y. H. Lee, and B.-H. Koo, "Stress and heart rate variability: A meta-analysis and review of the literature," *Psychiatry Investigation*, vol. 15, no. 3, pp. 235–245, 2018, accessed: 2026-01-23. [Online]. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC5900369/

[21] Z. AlArnaout, C. Zaki, Y. Kotb, M. AlAkkoumi, and N. Mostafa, "Exploiting heart rate variability for driver drowsiness detection using wearable sensors and machine learning," *Scientific Reports*, vol. 15, p. 24898, 2025, accessed: 2026-01-23. [Online]. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC12246425/

[22] B. Lequeux, C. Uzan, and M. B. Rehman, "Does resting heart rate measured by the physician reflect the patient's true resting heart rate? White-coat heart rate," *Indian Heart Journal*, vol. 70, no. 1, pp. 93–98, Jan. 2018. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0019483217301141

[23] Polismyndigheten, "Ratt- och sjöfylleri," 2025, accessed: 2026-02-16. [Online]. Available: https://polisen.se/lagar-och-regler/trafik-och-fordon/

ratt--och-sjofylleri/
[24] National Highway Traffic Safety Administration, "Drunk driving," https://www.nhtsa.gov/risky-driving/drunk-driving, accessed: 2026-04-27.
[25] GOV.UK, "The drink drive limit," https://www.gov.uk/drink-drive-limit, accessed: 2026-04-27.
[26] "Blood Alcohol Content (BAC) Drink Driving Limits across Europe." [Online]. Available: https://etsc.eu/issues/drink-driving/blood-alcohol-content-bac-drink-driving-limits-across-europe/
[27] J. Tindle and P. Tadi, "Neuroanatomy, parasympathetic nervous system," in *StatPearls*. Treasure Island (FL): StatPearls Publishing, 2022, available from: NCBI Bookshelf. [Online]. Available: https://www.ncbi.nlm.nih.gov/books/NBK553141/
[28] J. W. Hurst, "Naming of the waves in the ecg, with a brief account of their genesis," *Circulation*, vol. 98, no. 18, pp. 1937–1942, 1998.
[29] "HRV analysis methods - How is HRV calculated." [Online]. Available: https://www.kubios.com/blog/hrv-analysis-methods/
[30] T. F. O. T. E. S. O. C. T. N. A. Electrophysiology, "Heart Rate Variability: Standards of Measurement, Physiological Interpretation, and Clinical Use," *Circulation*, vol. 93, no. 5, pp. 1043–1065, Mar. 1996. [Online]. Available: https://www.ahajournals.org/doi/10.1161/01.CIR.93.5.1043
[31] F. Shaffer and J. P. Ginsberg, "An overview of heart rate variability metrics and norms," *Frontiers in Public Health*, vol. 5, p. 258, 2017. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5624990/
[32] Y. Liu *et al.*, "Heart rate variability as a potential biomarker for alcohol consumption: A systematic review," *International Journal of Environmental Research and Public Health*, vol. 19, no. 6, p. 3337, 2022. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8950456/
[33] M. Christ, N. Braun, J. Neuffer, and A. W. Kempa-Liehr, "Time Series FeatuRe Extraction on basis of Scalable Hypothesis tests (tsfresh – A Python package)," *Neurocomputing*, vol. 307, pp. 72–77, Sep. 2018. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0925231218304843
[34] A. M. van Roon *et al.*, "Overview of mathematical relations between poincaré plot parameters and time domain parameters of heart rate variability," *Entropy*, vol. 27, no. 8, p. 861, 2025.
[35] C. Iovescu and S. Rao, "The fundamentals of millimeter wave radar sensors," Texas Instruments, Tech. Rep. SPYY005A, Jul. 2020, accessed: 2026-04-21. [Online]. Available: https://www.ti.com/lit/spyy005
[36] M. Kebe, R. Gadhafi, B. Mohammad, M. Sanduleanu, H. Saleh, and M. Al-Qutayri, "Human vital signs detection methods and potential using radars: A review," *Sensors*, vol. 20, no. 5, p. 1454, 2020.
[37] E. Turppa, J. M. Kortelainen, O. Antropov, and T. Kiuru, "Vital sign monitoring using fmcw radar in various sleeping scenarios," *Sensors*, vol. 20, no. 22, p. 6505, 2020.
[38] A. Smiley and J. Finkelstein, "Feasibility of Radar-Based Heart Rate Variability Measurement: A Comparative Study with Electrocardiography," in *2025 IEEE 15th Annual Computing and Communication Workshop and*

Conference (CCWC), Jan. 2025, pp. 00 376–00 379. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/10903703

[39] A. Guettas, S. Ayad, and O. Kazar, "Driver state monitoring system: A review," in *Proceedings of the 4th International Conference on Big Data and Internet of Things*. Association for Computing Machinery, 2019, pp. 1–7. [Online]. Available: https://doi.org/10.1145/3372938.3372966

[40] Smart Eye, "Driver Monitoring System," https://smarteye.se/solutions/automotive/driver-monitoring-system/, 2026, accessed: 2026-05-15.

[41] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning: With Applications in R*, 2nd ed. Springer, 2021.

[42] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *Journal of Machine Learning Research*, vol. 3, pp. 1157–1182, 2003.

[43] scikit-learn developers, "sklearn.feature_selection.f_classif," https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.f_classif.html, accessed: 2026-05-08.

[44] A. Demircioğlu, "Measuring the bias of incorrect application of feature selection when using cross-validation in radiomics," *Insights into Imaging*, vol. 12, no. 1, p. 172, 2021.

[45] E. Y. Boateng and D. A. Abaye, "A Review of the Logistic Regression Model with Emphasis on Medical Research," *Journal of Data Analysis and Information Processing*, vol. 7, no. 4, pp. 190–207, Oct. 2019. [Online]. Available: https://www.scirp.org/journal/paperinformation?paperid=95655

[46] M. Belgiu and L. Drăguţ, "Random forest in remote sensing: A review of applications and future directions," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 114, pp. 24–31, 2016.

[47] ScienceDirect, "K-nearest neighbor," https://www.sciencedirect.com/topics/immunology-and-microbiology/k-nearest-neighbor, 2024, accessed: 2026-04-20.

[48] N.-D. Hoang and V.-D. Tran, "Comparison of histogram-based gradient boosting classification machine, random forest, and deep convolutional neural network for pavement raveling severity classification," *Automation in Construction*, vol. 148, p. 104767, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0926580523000274

[49] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," Jun. 2016, arXiv:1603.02754. [Online]. Available: http://arxiv.org/abs/1603.02754

[50] B. Naderalvojoud and T. Hernandez-Boussard, "Improving machine learning with ensemble learning on observational healthcare data," *AMIA Annual Symposium Proceedings*, vol. 2023, pp. 521–529, 2023. [Online]. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC10785929/

[51] A. U. Haq, J. P. Li, J. Khan, M. H. Memon, S. Nazir, S. Ahmad, G. A. Khan, and A. Ali, "Intelligent Machine Learning Approach for Effective Recognition of Diabetes in E-Healthcare Using Clinical Data," *Sensors (Basel, Switzerland)*, vol. 20, no. 9, p. 2649, May 2020. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7249007/

[52] O. Rainio, J. Teuho, and R. Klén, "Evaluation metrics and statistical tests for machine learning," *Scientific Reports*, vol. 14, 2024.

[53] T. Fawcett, "An introduction to roc analysis," *Pattern Recognition Letters*, vol. 27, no. 8, pp. 861–874, 2006.

[54] Penn State Eberly College of Science, "4.3 - introduction to bootstrapping," https://online.stat.psu.edu/stat200/lesson/4/4.3, n.d., accessed: 2026-05-20.

[55] M. Schaffarczyk, B. Rogers, R. Reer, and T. Gronwald, "Validity of the polar h10 sensor for heart rate variability analysis during resting state and incremental exercise in recreational men and women," *Sensors*, vol. 22, no. 17, p. 6536, 2022.

[56] J. Allen, "Photoplethysmography and its application in clinical physiological measurement," *Physiological Measurement*, vol. 28, no. 3, pp. R1–R39, 2007.

[57] C. Tronstad, G. E. Gjein, S. Grimnes, Ø. G. Martinsen, A.-L. Krogstad, E. Fosse, and H. Kalvøy, "Current trends and opportunities in the methodology of electrodermal activity measurement," *Physiological Measurement*, vol. 43, no. 2, p. 02TR01, 2022.

A

Appendix

Use of Artificial Intelligence (AI) Tools

AI-based tools, including ChatGPT, were used during this project to assist with language refinement, code suggestions, and general problem-solving.
All generated content was carefully reviewed, verified, and adapted by the authors to ensure correctness and relevance to the project.

A.1 Additional Results

Here, additional results are stored.

A.1.1 Simulator Data using Combined Features at the high-BAC setting

Table A.1: Pooled out-of-fold model performance using three combined features at threshold 0.5.

Model	Level	AUC	Accuracy	Precision	Sensitivity	Specificity
SVM RBF	Window	0.75	0.75	0.77	0.72	0.78
	Session	0.79	0.78	0.82	0.74	0.83
SVM linear	Window	0.79	0.74	0.78	0.70	0.79
	Session	0.81	0.78	0.82	0.74	0.83
XGBoost	Window	0.75	0.74	0.77	0.69	0.79
	Session	0.79	0.77	0.80	0.74	0.81
Logistic regression	Window	0.79	0.74	0.77	0.70	0.78
	Session	0.81	0.77	0.80	0.74	0.81
Random forest	Window	0.75	0.70	0.71	0.68	0.71
	Session	0.80	0.74	0.77	0.71	0.78
HistGB	Window	0.73	0.68	0.70	0.68	0.69
	Session	0.79	0.73	0.75	0.71	0.75
KNN	Window	0.75	0.70	0.71	0.68	0.71
	Session	0.78	0.70	0.71	0.71	0.69

Table A.2: Selected combined features across all evaluated models.

Feature	Selected folds	Total folds
`log_hr_rmssd_ratio`	37	37
`ibi_ms__fft_aggregated__aggtype_"skew"`	37	37
`ibi_ms__absolute_sum_of_changes`	32	37

The selected features were identical across HistGB, KNN, LR, RF, SVM linear, SVM RBF, and XGBoost.

Table A.3: Pooled out-of-fold model performance using five combined features at threshold 0.5.

Model	Level	AUC	Accuracy	Precision	Sensitivity	Specificity
SVM linear	Window	0.78	0.74	0.77	0.71	0.78
	Session	0.81	0.78	0.82	0.74	0.83
Logistic regression	Window	0.79	0.74	0.76	0.71	0.77
	Session	0.81	0.77	0.80	0.74	0.81
XGBoost	Window	0.75	0.74	0.78	0.69	0.80
	Session	0.79	0.77	0.80	0.74	0.81
SVM RBF	Window	0.73	0.74	0.75	0.72	0.75
	Session	0.78	0.77	0.80	0.74	0.81
Random forest	Window	0.73	0.68	0.71	0.66	0.71
	Session	0.77	0.74	0.77	0.71	0.78
KNN	Window	0.74	0.71	0.72	0.70	0.71
	Session	0.78	0.73	0.75	0.71	0.75
HistGB	Window	0.72	0.67	0.68	0.66	0.68
	Session	0.78	0.72	0.73	0.71	0.72

Table A.4: Selected combined features for the five features.

Feature	Selected folds	Total folds
`ibi_ms__fft_aggregated__aggtype_"skew"`	37	37
`log_hr_rmssd_ratio`	37	37
`ibi_ms__absolute_sum_of_changes`	37	37
`sdnn`	37	37
`ibi_ms__standard_deviation`	35	37

Table A.5: Pooled out-of-fold model performance using ten combined features at threshold 0.5.

Model	Level	AUC	Accuracy	Precision	Sensitivity	Specificity
XGBoost	Window	0.74	0.74	0.78	0.69	0.79
	Session	0.78	0.77	0.80	0.74	0.81
Logistic regression	Window	0.78	0.73	0.76	0.71	0.76
	Session	0.81	0.77	0.80	0.74	0.81
SVM linear	Window	0.78	0.75	0.79	0.69	0.80
	Session	0.81	0.77	0.82	0.71	0.83
SVM RBF	Window	0.74	0.73	0.75	0.72	0.74
	Session	0.77	0.77	0.80	0.74	0.81
KNN	Window	0.74	0.71	0.73	0.71	0.72
	Session	0.77	0.74	0.77	0.71	0.78
Random forest	Window	0.73	0.69	0.71	0.65	0.72
	Session	0.76	0.73	0.75	0.71	0.75
HistGB	Window	0.72	0.66	0.68	0.64	0.68
	Session	0.77	0.72	0.73	0.71	0.72

Table A.6: Selected combined features for the ten features.

Feature	Selected folds	Total folds
`ibi_ms__fft_aggregated__aggtype_"skew"`	37	37
`log_hr_rmssd_ratio`	37	37
`ibi_ms__absolute_sum_of_changes`	37	37
`ibi_ms__standard_deviation`	37	37
`sdnn`	37	37
`ibi_ms__cid_ce__normalize_False`	36	37
`ibi_ms__mean_abs_change`	35	37
`ibi_ms__change_quantiles__f_agg_"mean" __isabs_True__qh_1.0__ql_0.2`	34	37
`sd2`	32	37
`ibi_ms__change_quantiles__f_agg_"mean" __isabs_True__qh_1.0__ql_0.0`	32	37

A.1.2 Simulator Data using Manual Features in the high-BAC setting

Table A.7: Pooled out-of-fold model performance using three manual features at threshold 0.5.

Model	Level	AUC	Accuracy	Precision	Sensitivity	Specificity
SVM RBF	Window	0.70	0.69	0.70	0.68	0.70
	Session	0.73	0.73	0.75	0.71	0.75
SVM linear	Window	0.77	0.70	0.71	0.72	0.69
	Session	0.78	0.73	0.75	0.71	0.75
XGBoost	Window	0.74	0.69	0.71	0.65	0.73
	Session	0.76	0.70	0.74	0.66	0.75
Logistic regression	Window	0.77	0.71	0.71	0.71	0.70
	Session	0.78	0.73	0.75	0.71	0.75
Random forest	Window	0.70	0.64	0.65	0.64	0.64
	Session	0.74	0.72	0.76	0.66	0.78
HistGB	Window	0.69	0.65	0.67	0.65	0.66
	Session	0.74	0.69	0.71	0.66	0.72
KNN	Window	0.72	0.65	0.66	0.64	0.65
	Session	0.74	0.72	0.74	0.68	0.75

Table A.8: Selected manual features across all evaluated models.

Feature	Selected folds	Total folds
`log_hr_rmssd_ratio`	37	37
`sdnn`	37	37
`sd2`	36	37

The selected features were identical across HistGB, KNN, LR, RF, SVM linear, SVM RBF, and XGBoost.

Table A.9: Pooled out-of-fold model performance using five manual features at threshold 0.5.

Model	Level	AUC	Accuracy	Precision	Sensitivity	Specificity
SVM RBF	Window	0.72	0.70	0.71	0.71	0.70
	Session	0.75	0.76	0.76	0.76	0.75
Random forest	Window	0.73	0.68	0.70	0.65	0.71
	Session	0.78	0.73	0.78	0.66	0.81
Logistic regression	Window	0.76	0.70	0.71	0.70	0.71
	Session	0.77	0.73	0.75	0.71	0.75
HistGB	Window	0.72	0.67	0.68	0.65	0.69
	Session	0.78	0.73	0.77	0.68	0.78
XGBoost	Window	0.74	0.72	0.76	0.68	0.77
	Session	0.76	0.72	0.74	0.68	0.75
SVM linear	Window	0.76	0.69	0.70	0.68	0.69
	Session	0.77	0.72	0.74	0.68	0.75
KNN	Window	0.73	0.66	0.67	0.63	0.68
	Session	0.76	0.70	0.74	0.66	0.75

Table A.10: Selected manual features for the five-feature Polar model comparison.

Feature	Selected folds	Total folds
`log_hr_rmssd_ratio`	37	37
`sdnn`	37	37
`sd2`	37	37
`rmssd`	37	37
`pnn50`	37	37

Table A.11: Pooled out-of-fold model performance using ten manual features at threshold 0.5.

Model	Level	AUC	Accuracy	Precision	Sensitivity	Specificity
Logistic regression	Window	0.79	0.71	0.73	0.69	0.74
	Session	0.81	0.77	0.80	0.74	0.81
SVM linear	Window	0.78	0.71	0.74	0.67	0.76
	Session	0.81	0.76	0.81	0.68	0.83
XGBoost	Window	0.73	0.71	0.74	0.66	0.76
	Session	0.76	0.72	0.76	0.66	0.78
SVM RBF	Window	0.74	0.69	0.69	0.69	0.68
	Session	0.76	0.70	0.72	0.68	0.72
Random forest	Window	0.73	0.66	0.67	0.65	0.67
	Session	0.78	0.70	0.72	0.68	0.72
HistGB	Window	0.71	0.65	0.66	0.65	0.65
	Session	0.76	0.70	0.72	0.68	0.72
KNN	Window	0.72	0.66	0.67	0.65	0.67
	Session	0.76	0.70	0.72	0.68	0.72

Table A.12: Selected manual features for the ten-feature Polar model comparison.

Feature	Selected folds	Total folds
`log_hr_rmssd_ratio`	37	37
`sdnn`	37	37
`sd2`	37	37
`rmssd`	37	37
`hf_power`	37	37
`pnn50`	37	37
`hr_mean`	37	37
`ibi_mean`	37	37
`hr_median`	37	37
`lf_power`	36	37

A.1.3 Simulator Data using tsfresh features at the high-BAC setting

Table A.13: Pooled out-of-fold model performance using three tsfresh features at threshold 0.5.

Model	Level	AUC	Accuracy	Precision	Sensitivity	Specificity
SVM RBF	Window	0.73	0.74	0.76	0.73	0.77
	Session	0.78	0.77	0.80	0.74	0.81
SVM linear	Window	0.79	0.74	0.77	0.70	0.78
	Session	0.81	0.78	0.82	0.74	0.83
XGBoost	Window	0.75	0.74	0.78	0.70	0.80
	Session	0.80	0.77	0.80	0.74	0.81
Logistic regression	Window	0.79	0.74	0.76	0.71	0.77
	Session	0.81	0.78	0.82	0.74	0.83
Random forest	Window	0.74	0.69	0.71	0.67	0.72
	Session	0.78	0.77	0.80	0.74	0.81
HistGB	Window	0.73	0.68	0.69	0.66	0.69
	Session	0.78	0.77	0.80	0.74	0.81
KNN	Window	0.75	0.73	0.75	0.69	0.77
	Session	0.77	0.77	0.80	0.74	0.81

Table A.14: Selected tsfresh features across all evaluated models.

Feature	Selected folds	Total folds
`ibi_ms__fft_aggregated__aggtype_"skew"`	37	37
`ibi_ms__absolute_sum_of_changes`	37	37
`ibi_ms__standard_deviation`	35	37

The selected features were identical across HistGB, KNN, LR, RF, SVM linear, SVM RBF, and XGBoost.

Table A.15: Pooled out-of-fold model performance using five tsfresh features at threshold 0.5.

Model	Level	AUC	Accuracy	Precision	Sensitivity	Specificity
Logistic regression	Window	0.78	0.73	0.75	0.71	0.76
	Session	0.80	0.78	0.82	0.74	0.83
XGBoost	Window	0.75	0.74	0.77	0.69	0.79
	Session	0.78	0.77	0.80	0.74	0.81
SVM linear	Window	0.78	0.74	0.78	0.69	0.79
	Session	0.80	0.77	0.82	0.71	0.83
SVM RBF	Window	0.73	0.73	0.75	0.71	0.75
	Session	0.77	0.77	0.80	0.74	0.81
Random forest	Window	0.73	0.68	0.71	0.64	0.72
	Session	0.77	0.76	0.78	0.74	0.78
KNN	Window	0.75	0.71	0.73	0.69	0.74
	Session	0.79	0.74	0.77	0.71	0.78
HistGB	Window	0.72	0.66	0.68	0.64	0.68
	Session	0.77	0.74	0.76	0.74	0.75

Table A.16: Selected tsfresh features for the five-feature Polar model comparison.

Feature	Selected folds	Total folds
`ibi_ms__fft_aggregated__aggtype_"skew"`	37	37
`ibi_ms__standard_deviation`	37	37
`ibi_ms__absolute_sum_of_changes`	37	37
`ibi_ms__cid_ce__normalize_False`	36	37
`ibi_ms__variation_coefficient`	33	37

Table A.17: Pooled out-of-fold model performance using ten tsfresh features at decision threshold 0.5.

Model	Level	AUC	Accuracy	Precision	Sensitivity	Specificity
Logistic regression	Window	0.79	0.73	0.75	0.70	0.75
	Session	0.81	0.77	0.82	0.71	0.83
SVM RBF	Window	0.74	0.73	0.75	0.71	0.75
	Session	0.77	0.77	0.80	0.74	0.81
SVM linear	Window	0.79	0.73	0.75	0.70	0.76
	Session	0.81	0.77	0.82	0.71	0.83
XGBoost	Window	0.75	0.74	0.77	0.69	0.79
	Session	0.78	0.76	0.79	0.71	0.81
Random forest	Window	0.73	0.69	0.71	0.66	0.72
	Session	0.76	0.74	0.77	0.71	0.78
KNN	Window	0.75	0.71	0.73	0.68	0.73
	Session	0.79	0.74	0.77	0.71	0.78
HistGB	Window	0.71	0.65	0.67	0.65	0.66
	Session	0.75	0.73	0.75	0.71	0.75

Table A.18: Selected tsfresh features for the ten features.

Feature	Selected folds	Total folds
`ibi_ms__fft_aggregated__aggtype_"skew"`	37	37
`ibi_ms__variation_coefficient`	37	37
`ibi_ms__standard_deviation`	37	37
`ibi_ms__absolute_sum_of_changes`	37	37
`ibi_ms__cid_ce__normalize_False`	37	37
`ibi_ms__change_quantiles__f_agg_"mean" __isabs_True__qh_1.0__ql_0.4`	37	37
`ibi_ms__change_quantiles__f_agg_"mean" __isabs_True__qh_1.0__ql_0.2`	37	37
`ibi_ms__change_quantiles__f_agg_"mean" __isabs_True__qh_1.0__ql_0.0`	37	37
`ibi_ms__mean_abs_change`	36	37
`ibi_ms__fft_aggregated__aggtype_"kurtosis"`	35	37

A.1.4 Simulator Data using the low-to-moderate BAC setting

Table A.19: Pooled out-of-fold model performance using five combined features at the low-to-moderate BAC setting at decision threshold 0.5.

Model	Level	AUC	Accuracy	Precision	Sensitivity	Specificity
Logistic Regression	Window	0.68	0.64	0.64	0.60	0.68
	Session	0.71	0.72	0.73	0.69	0.75
SVM (RBF)	Window	0.65	0.63	0.62	0.62	0.65
	Session	0.70	0.72	0.71	0.71	0.72
SVM (Linear)	Window	0.67	0.63	0.63	0.55	0.70
	Session	0.70	0.70	0.75	0.60	0.81
XGBoost	Window	0.64	0.62	0.61	0.56	0.67
	Session	0.68	0.70	0.72	0.66	0.75
Random Forest	Window	0.61	0.56	0.55	0.53	0.60
	Session	0.66	0.66	0.67	0.63	0.69
KNN	Window	0.62	0.60	0.59	0.55	0.64
	Session	0.67	0.65	0.66	0.60	0.69
HistGB	Window	0.59	0.55	0.53	0.51	0.58
	Session	0.64	0.63	0.64	0.60	0.67

Table A.20: Pooled out-of-fold model performance using ten combined features at the low-to-moderate BAC setting at decision threshold 0.5.

Model	Level	AUC	Accuracy	Precision	Sensitivity	Specificity
SVM (RBF)	Window	0.64	0.62	0.60	0.60	0.63
	Session	0.68	0.70	0.71	0.69	0.72
XGBoost	Window	0.63	0.61	0.60	0.56	0.65
	Session	0.66	0.70	0.72	0.66	0.75
SVM (Linear)	Window	0.67	0.63	0.63	0.56	0.69
	Session	0.71	0.69	0.74	0.57	0.81
KNN	Window	0.61	0.57	0.55	0.55	0.59
	Session	0.66	0.68	0.68	0.66	0.69
Random Forest	Window	0.62	0.57	0.55	0.54	0.60
	Session	0.66	0.66	0.67	0.63	0.69
Logistic Regression	Window	0.67	0.64	0.63	0.59	0.68
	Session	0.70	0.65	0.65	0.63	0.67
HistGB	Window	0.60	0.56	0.54	0.53	0.58
	Session	0.65	0.63	0.64	0.60	0.67

A.1.5 Simulator Data using the all-data setting

Table A.21: Pooled out-of-fold model performance using all data with five selected combined features at decision threshold 0.5.

Model	Level	AUC	Accuracy	Precision	Sensitivity	Specificity
Random forest	Window	0.70	0.69	0.74	0.82	0.43
	Session	0.77	0.72	0.75	0.88	0.42
SVM linear	Window	0.72	0.69	0.71	0.89	0.29
	Session	0.74	0.72	0.73	0.90	0.33
XGBoost	Window	0.70	0.65	0.82	0.61	0.73
	Session	0.72	0.71	0.85	0.68	0.75
HistGB	Window	0.71	0.70	0.75	0.82	0.46
	Session	0.78	0.72	0.75	0.86	0.42
Logistic regression	Window	0.72	0.67	0.79	0.67	0.65
	Session	0.73	0.66	0.79	0.67	0.64
SVM RBF	Window	0.71	0.71	0.77	0.79	0.54
	Session	0.75	0.65	0.72	0.78	0.39
KNN	Window	0.69	0.68	0.72	0.84	0.35
	Session	0.74	0.66	0.70	0.88	0.22

Table A.22: Pooled out-of-fold model performance using all data with ten selected combined features at decision threshold 0.5.

Model	Level	AUC	Accuracy	Precision	Sensitivity	Specificity
SVM linear	Window	0.73	0.69	0.73	0.85	0.37
	Session	0.76	0.72	0.74	0.89	0.36
Logistic regression	Window	0.74	0.67	0.82	0.64	0.72
	Session	0.76	0.72	0.86	0.68	0.78
HistGB	Window	0.69	0.66	0.73	0.78	0.42
	Session	0.75	0.71	0.74	0.86	0.39
XGBoost	Window	0.70	0.66	0.82	0.63	0.73
	Session	0.73	0.70	0.86	0.66	0.78
SVM RBF	Window	0.71	0.70	0.77	0.78	0.54
	Session	0.75	0.68	0.74	0.79	0.44
KNN	Window	0.68	0.67	0.72	0.83	0.35
	Session	0.72	0.68	0.71	0.89	0.25
Random forest	Window	0.69	0.66	0.72	0.81	0.36
	Session	0.73	0.68	0.71	0.88	0.28

A.1.6 Real-world driving data at the high-BAC setting

All results here are made using combined features and by having only simulator data in the training data.

Table A.23: Pooled model performance using five features on the real-world driving test at threshold 0.6.

Model	Level	AUC	Accuracy	Precision	Sensitivity	Specificity
SVM RBF	Window	0.67	0.65	0.64	0.71	0.60
	Session	0.76	0.68	0.67	0.71	0.65
KNN	Window	0.68	0.64	0.62	0.70	0.59
	Session	0.76	0.65	0.63	0.71	0.59
XGBoost	Window	0.70	0.66	0.65	0.70	0.63
	Session	0.75	0.74	0.75	0.71	0.76
Logistic regression	Window	0.70	0.68	0.68	0.66	0.69
	Session	0.73	0.71	0.73	0.65	0.76
SVM linear	Window	0.70	0.68	0.71	0.60	0.76
	Session	0.72	0.74	0.83	0.59	0.88
HistGB	Window	0.64	0.59	0.57	0.66	0.52
	Session	0.72	0.62	0.60	0.71	0.53
Random forest	Window	0.63	0.59	0.59	0.61	0.58
	Session	0.70	0.62	0.61	0.65	0.59

Table A.24: Pooled model performance using ten combined features on the real-world driving test at threshold 0.6.

Model	Level	AUC	Accuracy	Precision	Sensitivity	Specificity
SVM RBF	Window	0.68	0.65	0.63	0.70	0.60
	Session	0.76	0.65	0.63	0.71	0.59
KNN	Window	0.68	0.64	0.63	0.70	0.59
	Session	0.75	0.65	0.63	0.71	0.59
XGBoost	Window	0.70	0.67	0.66	0.69	0.65
	Session	0.75	0.74	0.75	0.71	0.76
Logistic regression	Window	0.70	0.66	0.65	0.67	0.65
	Session	0.73	0.68	0.69	0.65	0.71
SVM linear	Window	0.69	0.67	0.68	0.63	0.70
	Session	0.73	0.68	0.71	0.59	0.76
Random forest	Window	0.65	0.61	0.60	0.64	0.58
	Session	0.73	0.68	0.67	0.71	0.65
HistGB	Window	0.65	0.59	0.57	0.70	0.48
	Session	0.73	0.68	0.67	0.71	0.65

A.1.7 Late Fusion results using the high-BAC setting

Table A.25: Late fusion performance using five selected combined features on the high-BAC setting at decision threshold 0.5.

Model	Level	AUC	Accuracy	Sensitivity	Specificity
Fusion	Window	0.87	0.78	0.78	0.79
Fusion	Session	0.93	0.86	0.84	0.89

Table A.26: Late fusion performance using ten selected combined features on high-BAC setting at decision threshold 0.5.

Model	Level	AUC	Accuracy	Sensitivity	Specificity
Fusion	Window	0.87	0.79	0.78	0.80
Fusion	Session	0.93	0.86	0.84	0.89

A.1.8 Late Fusion results using the Low-to-Moderate BAC Setting

Table A.27: Late fusion performance using five selected combined features on low-to-moderate BAC settings at decision threshold 0.5.

Model	Level	AUC	Accuracy	Sensitivity	Specificity
Fusion	Window	0.68	0.73	0.55	0.71
Fusion	Session	0.73	0.68	0.57	0.78

Table A.28: Late fusion performance using ten selected combined features on low-to-moderate settings at decision threshold 0.5.

Model	Level	AUC	Accuracy	Sensitivity	Specificity
Fusion	Window	0.66	0.60	0.52	0.68
Fusion	Session	0.70	0.62	0.51	0.72

DEPARTMENT OF MECHANICS AND MARITIME SCIENCES
CHALMERS UNIVERSITY OF TECHNOLOGY
Gothenburg, Sweden 2026
www.chalmers.se

CHALMERS
UNIVERSITY OF TECHNOLOGY