A Machine Learning Approach to Predict Emotional Arousal and Aggressive Driving from EDA and Heart Rate Signals

A THESIS
SUBMITTED TO THE FACULTY OF THE GRADUATE
SCHOOL
OF THE UNIVERSITY OF MINNESOTA
BY

Md Sakibul Hassan Nahid

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
MASTER OF SCIENCE

Supervised by
Dr. Turuna Seecharan

Committee Members
Dr. Ping Zhao
Dr. Jomara Sandbulte

June 2025

Abstract

Aggressive driving behaviors, such as sudden acceleration and hard braking, abrupt steering, significantly increase the risk of road accidents, especially among young drivers. This study aims to predict aggressive driving behavior and emotional arousal using physiological signals, Electrodermal Activity (EDA) and Heart Rate (HR), collected via the Empatica E4 wristband in real-world driving conditions. A total of 25 drivers participated in this study. From the collected Electrodermal activity (EDA) and Heart Rate signals, 22 features were extracted, including tonic and phasic components, difference-based features, and heart rate statistics. Feature selection was performed using Random Forest feature importance and L1-regularized Logistic Regression feature importance method to identify the most relevant features for three prediction tasks: emotional arousal, aggressive acceleration, and aggressive braking. These features were used to train four classification models, Random Forest, Logistic Regression, Support Vector Machine (SVM), and XGBoost, as well as an ensemble VotingClassifier. Model performance was evaluated using five-fold cross-validation across accuracy, precision, recall, and F1-score. The ensemble model achieved the best overall performance across tasks. For arousal prediction using logistic regression feature importance, the ensemble achieved an accuracy of 0.65, a precision of 0.69, a recall of 0.60, and an F1-score of 0.57. In aggressive acceleration prediction, the highest-performing model (XGBoost) achieved 0.62 accuracy, while ensemble models reached 0.55. For aggressive braking prediction, both SVM and the ensemble achieved 0.70 accuracy. These findings indicate that physiological signals can reliably predict emotional arousal and driving behaviour. This research highlights the potential of integrating wearable sensor data with machine learning to build real-time driver monitoring systems aimed at improving road safety.

Table of Contents

List of Figures:

List of Tables:

Nomenclature

ANN	Artificial Neural Network
EDA	Electrodermal Activity
EKF	Extended Kalman Filter
GSR	Galvanic Skin Response
HR	Heart Rate
HRV	Heart Rate Variability
IRB	Institutional Review Board
OBD	On Board Diagnostics
RF	Random Forest Classifier
SCL	Skin Conductance Level
SCR	Skin Conductance Response
SVM	Support Vector Machine
XGBOOST	Extreme Gradient Boosting

Table of Features

Feature Code	Description	Signal Type	Category
Base Phasic Mean	Base Phasic Mean	EDA	Baseline Features
Base Phasic Max	Base Phasic Maximum	EDA	Baseline Features
Base Phasic Std	Base Phasic Standard Deviation	EDA	Baseline Features
High Phasic Mean	High Phasic Mean	EDA	High Activity Features
High Phasic Max	High Phasic Maximum	EDA	High Activity Features
High Phasic Std	High Phasic Standard Deviation	EDA	High Activity Features
Base Tonic Mean	Base Tonic Mean	EDA	Baseline Features
Base Tonic Max	Base Tonic Maximum	EDA	Baseline Features
Base Tonic Std	Base Tonic Standard Deviation	EDA	Baseline Features
High Tonic Mean	High Tonic Mean	EDA	High Activity Features
High Tonic Max	High Tonic Maximum	EDA	High Activity Features
High Tonic Std	High Tonic Standard Deviation	EDA	High Activity Features
PhMeanDiff	Phasic Mean Difference	EDA	Difference Features
PhMaxDiff	Phasic Maximum Difference	EDA	Difference Features
PhStdDiff	Phasic Standard Deviation Difference	EDA	Difference Features
TnMeanDiff	Tonic Mean Difference	EDA	Difference Features
TnMaxDiff	Tonic Maximum Difference	EDA	Difference Features
TnStdDiff	Tonic Standard Deviation Difference	EDA	Difference Features
Mean	Heart Rate Mean	Heart Rate	Statistical Features
Median	Heart Rate Median	Heart Rate	Statistical Features
SKEW	Heart Rate Skewness	Heart Rate	Statistical Features
STD	Heart Rate Standard Deviation	Heart Rate	Statistical Features

Chapter 1: Introduction

A major road safety issue is aggressive driving. Aggressive driving, including speeding, harsh braking, or tailgating, played a role in 56% of fatal collisions in the United States (U.S) [1]. Young drivers are particularly at risk. Motor vehicle crashes are the second leading cause of death among U.S. teens, largely due to aggressive driving [2]. A Minnesota traffic safety survey reported that 22% of respondents viewed speeding as a top road safety issue (second only to distracted driving), and a further subset (8%) specifically flagged "aggressive driving" as a major concern [3]. State crash data from 2020–2023 reveal that "Reckless or Aggressive Driving" was cited as a factor in about 1.8% of crashes for drivers aged 20–29, a higher rate than for any older age group [4]. This overrepresentation helps explain why the 18–30 age group has such high crash involvement. It also highlights the need to understand the factors that contribute to aggressive driving in young drivers.

Aggressive driving is associated with drivers' emotional state (i.e., fatigue, stress, anger, etc.) [5]. Emotions can have a significant impact on driving behavior. In fact, intense emotional stress is known to degrade driver concentration and increase the likelihood of risky driving behavior [6]. Physiological signals such as heart rate and EDA give a measurable window into those states, and recent work, including real-world trials, demonstrates that machine learning can use these signals to predict aggressive driving behavior. This combination of wearable sensing and predictive modeling offers a novel approach to improve driver safety, particularly for high-risk groups such as young drivers, by lowering emotionally induced driving risks through real-time identification and feedback.

1.1 Problem Statement

Current research uses driving simulators, vehicle data, or surveys that ask drivers how they feel. Driving simulators are useful to observe driving behaviors in controlled environments but they are

expensive and driving behavior observed in a simulated environment may not translate to real-world driving. Surveys are easy to get a large amount of data, but these don't give a clear picture of what happened during the driving session. This work uses a combination of surveys and naturalistic driving to understand the role of emotional arousal on the driver's driving behaviors. To support survey responses, bio signals, such as Electrodermal Activity and Heart Rate, are used. Electrodermal activity shows how emotionally aroused someone is through the changes in their skin's sweat.

This research uses data collected from the Empatica E4 wristband, which records EDA and HR signals during real-world driving sessions, paired with driving data obtained directly from the vehicle using Geotab Go. Unlike simulated or retrospective data, this naturalistic dataset allows for the exploration of driver state and behavior in uncontrolled, real driving conditions, providing a more realistic foundation for prediction.

A key focus of this work is the identification of the most relevant physiological features for three prediction tasks: emotional arousal, aggressive acceleration, and aggressive braking. To reduce noise, improve model performance, and enhance interpretability, two feature selection techniques were applied—Random Forest feature importance and L1-regularized Logistic Regression. These methods help isolate the physiological indicators most predictive of the target behaviors, which are then used to train and evaluate multiple machine learning models, both individually and in ensemble form.

Through this approach, this study aims to develop accurate, data-driven models capable of predicting aggressive driving behaviors and emotional arousal, ultimately contributing to the development of real-time monitoring systems that enhance road safety and driver well-being.

1.2 Electrodermal Activity (EDA)

Electrodermal Activity (EDA), also known as galvanic skin response (GSR), refers to the change in the skin's electrical conductance due to sweat gland activity. This physiological response is closely linked to the autonomic nervous system, particularly the sympathetic branch, which is responsible for our emotional response.

When a person experiences emotional arousal, such as stress, excitement, or fear, the body activates the sympathetic nervous system as part of the autonomic response. This activation eventually increases the activity in the eccrine sweat glands, which are densely situated on the palms of the hands, soles of the feet, and forehead. These glands play an important role in emotional sweating.

This emotional sweating is sometimes on a microscopic level and often not visible to the naked eye, but it can significantly alter the electrical properties of the skin. Sweat contains electrolytes like sodium and chloride, which increase the skin's electrical conductivity. The more sweat is produced, the more conductive the skin becomes.

EDA sensors are used to measure this change in skin. These EDA sensors commonly consist of two electrodes and are placed on the skin’s surface. A small, constant voltage is applied between the electrodes, and the resulting current is measured. The magnitude of the current flow corresponds to the skin conductance, which varies in response to physiological arousal.

Typically, these sensors are positioned on the fingertips or inner wrists, where eccrine glands are abundant and signal quality is high. Modern wearable devices, like Empatica E4 wristbands, use this principle to track EDA in real time, enabling researchers and practitioners to monitor psychological and emotional states during various activities, such as driving, public speaking, or cognitive tasks.

This makes EDA a powerful and practical biomarker for studying stress, attention, workload, and emotional engagement across different work.

1.3 Aggressive Driving

Aggressive driving is a major problem that affects road safety, human life, and traffic efficiency. Aggressive driving involves dangerous behaviour such as speeding, hard braking, hard acceleration, abrupt steering, rapid lane changes, and ignoring traffic rules. Studies have found that these actions are often influenced by the emotional state of the driver. Aggressive driving increases the risk of a collision that can lead to serious injuries, even deaths. It not only endangers the driver's life but also puts the passengers, pedestrians, and other vehicles at risk.

In this research, driving behavior was monitored using acceleration, braking, and speeding data collected from the vehicle. This data reflects how a driver drives or controls the vehicle. These data give us information on how drivers accelerate or decelerate their vehicles and the speed at which they operate them.

1.4 Study Scope and Novel Contribution

There is growing research interest in using physiological signals to understand and predict driver behavior, especially in relation to stress, arousal, and aggression. While many existing studies are based in simulation or lab settings, this research focuses on real-world driving in Duluth, Minnesota.

In this study, data were collected from 25 participants over two weeks, combining the physiological signals - Electrodermal Activity (EDA) and heart rate - using Empatica E4 wristbands with driving behavior data from telematics systems. The goal is to develop models that predict emotional arousal and aggressive driving behaviors.

Due to the high dimensionality of the extracted features, this study employs feature selection techniques rather than dimensionality reduction methods like Principal Component Analysis (PCA). Unlike PCA, which creates linear combinations of original variables and transforms them into uncorrelated principal components, feature selection techniques retain the original features, thereby preserving interpretability and physiological relevance. Understanding which specific physiological features contribute most to aggressive driving is essential for both model transparency and future system design.

To achieve this, the study applies Random Forest feature importance and L1-regularized Logistic Regression. These techniques were selected for their ability to not only reduce feature space but also highlight the most influential individual predictors. Random Forest is well-suited for capturing nonlinear relationships and provides a robust ranking of feature importance, while L1-regularized Logistic Regression can capture the linear relationship and offers sparsity by shrinking less relevant coefficients to zero, making it particularly useful for identifying a minimal set of predictive variables. To date, these techniques have rarely been applied in this context, especially with physiological data collected during naturalistic driving.

Unlike controlled laboratory studies, this work captures everyday driving conditions, making it more relevant for real-world applications in intelligent driver monitoring systems. While the geographical scope of the study is limited, the findings provide important insights into how physiological signals can be used to enhance road safety.

Chapter 2 : Literature Review

It is well known that human factors, such as distracted driving, impaired driving, and speeding, are the main reasons for road traffic accidents. Although vehicle technologies exist, recent studies focus on the driver, such as investigating the role of the driver's personality traits and physiological states in their driving style and, thus, their propensity toward being involved in road traffic accidents.

2.1 Emotional Arousal

Driving is a complex cognitive activity that is deeply influenced by a driver's emotional and psychological state. Emotions such as anger, stress, fatigue, calmness, and even happiness significantly affect how drivers perceive risk, keep their focus, and make decisions on the road. These emotional states often arise spontaneously in response to personal experiences, traffic conditions, or external stressors and can influence driving behavior, sometimes leading to dangerous consequences.

Studies have been conducted on the impact of emotions on driving behavior [8] [9] [10]. Emotions can influence driving behavior [8]. However, most of those studies used a driving simulator and a questionnaire. Notably, a paper by Najafi et al. investigated the mental attention state of individuals using their EDA, Electrocardiogram (ECG), and Electroencephalogram (EEG) [32].

Anger has been shown to impact decision-making by increasing risk-taking behavior, making the driver more impulsive, and reducing attention to safety cues. Anger tends to induce a sense of certainty and personal control, making individuals more likely to blame others and take aggressive actions, which can be considered aggressive driving or road rage incidents [10]. In driving contexts, angry drivers are more prone to speeding, tailgating, or abrupt lane changes due to their emotional arousal and reduced cognitive control.

Stress, especially when chronic or triggered by work-related conditions, has an immense effect on driving performance. It elevates cognitive load, impairs attentional control, and promotes aggressive responses to perceived threats or delays. Studies found that the stress experienced by professional drivers was significantly associated with traffic violations and caused by driving anger [11]. Similarly, the drivers experiencing high levels of stress were more likely to engage in hostile or aggressive behavior, which can result in dangerous behaviors such as abrupt braking, overtaking, or confrontational gestures [12].

Fatigue, often overlooked, is another crucial factor. It also reduces cognitive performance, slows reaction times, and increases the likelihood of distraction. In environments with high demands and irregular schedules, such as professional driving, fatigue combined with stress can further compromise judgment, increasing the risk of accidents [4].

Happiness and calmness, though positive in nature, also influence driving. While calmness generally enhances focus, patience, and rule-following behavior, excessive happiness can lead to overconfidence and inattention. Drivers in an elevated mood may underestimate risks or overestimate their capabilities, resulting in delayed braking or ignoring signs.

.

Negative emotions, especially those directly related to traffic events such as being cut off or delayed, tend to influence decision speed and conservativeness. In an eye movement study, it was found that traffic-related negative emotions led to quicker decisions and more frequent accelerations and decelerations, but with reduced visual attention to risky areas, suggesting a potentially problematic cognitive response [13].

That’s why in this research, the stressed, happy, fatigued, and angry responses were labeled as “Aroused” sessions, and the calm responses were labeled as “Not Aroused” sessions.

2.2 Personality and Driving Style

It is important to understand how an individual's behaviour and personality influence their driving. It is crucial to improving road safety. Research in traffic psychology has long suggested that personality traits such as anger, anxiety, and risk tolerance can shape a driver's responses to road situations. Drivers' interpretations of other drivers' behaviors, particularly cautious or erratic actions, can influence their own driving style and safety-related decisions. This section explores how demographic, and personality factors affect the perception of timid and erratic braking behaviors, highlighting implications for designing more effective driver safety interventions.

An online survey of 439 drivers (ages 18 to 89) examined the perceived road safety risks posed by timid driving behaviors and their links to driver traits (anxiety, anger) and demographics (age, mileage, gender) [14]. According to this research work, drivers believed that unexpected and overly cautious braking behaviors posed moderate risk levels. Additionally, multiple linear regression studies revealed that older, male, and irritable drivers considered the hazards of slow and overly cautious behavior higher. According to this research, driver traits and the dangers of erratic braking were not shown to be significantly associated. The results imply that safety programs aimed at reducing aggressive behavior may benefit by focusing on how other drivers' behaviors are perceived.

Dangerous, anxious, and irate–high-speed styles were positively associated with unsafe driving, while cautious driving was negatively related to hazardous conduct and positively related to good driving [15].

According to the multiple regression analysis, aggressive driving was negatively related to age and positively associated with neuroticism [16]. Age, driving experience, and depression, as well as age, driving experience, and neuroticism, were significantly associated. Simple slope testing

indicated that perhaps depression may make older, experienced drivers more aggressive. Neuroticism led to more aggressive driving in those with more experience than those with less. Hasaninasab et al. investigated the use of psychology in traffic safety to find the effect that personality traits have on the number of accidents [17]. Two hundred college students completed a questionnaire divided into three sections: personal information, questions related to the subject under review, and the Neo test. Using structural equations, the maximum likelihood method, and testing the hypotheses using the AMOS software, the study found significant associations between personality traits and the number of fines and between personality traits and the number of accidents [17].

Recent studies show low sleep efficiency (<80%) and elevated Karolinska Sleepiness Scale (KSS) scores, which suggest that fatigue is a key factor contributing to aberrant driving behavior (ADB) [18]. Also, heart rate variability showed a significant change before the ADB events, which makes it a potential psychological marker. The Gated Recurrent Unit (GRU) model achieved an accuracy of 81.16%–84.22% in predicting ADB using heart rate variability.

Furthermore, research on driver ticket frequency also gives a better understanding of traffic violations and risk factors associated with aggressive driving behavior. A study found that younger drivers, particularly males, are more likely to receive tickets. Also, aggressive driving and frequent speeding were strongly linked to higher ticket frequencies, proving the relationship between drivers' behavior and traffic offenses. The study also revealed that personality traits such as impulsivity and low conscientiousness were associated with an increased likelihood of receiving traffic violations, as drivers with these characteristics were more prone to risky driving behaviors [19].

Recent research has also explored the data-driven approach to driver classification using psychophysiological data and driving performance. A study involving 90 drivers used a

questionnaire and a driving simulator to classify driving characteristics. The results showed that the psychometric questionnaire before driving can effectively predict driving behavior [20].

2.3 Driving Style and Emotional Arousal

Driving behavior depends on several parameters: drivers' mood, focus, road conditions, weather, traffic, etc. Emotional arousal is one of the reasons behind aggressive driving behavior. Poor driving habits are associated with aggressive behavior. To examine the conceptual model of variables and discover the link between aggressive behavior and risky driver conduct, structural equations and the decision tree approach were used to determine the different degrees of dangerous and aggressive driving behavior [21]. Age, attitude, personality traits, hazardous and aggressive driving, and negative emotions and perceptions strongly affected aggressive and dangerous driving behavior. Consequently, the most important variable in the structural equation model, attitude, is associated with aggressive conduct with 84% accuracy for neural networks and unsafe driving with 75% accuracy for SVM.

A recent study used a driving simulator and the Multidimensional Driving Style Inventory (MDSI) to examine the correlation between driving style and biological responses to negative emotions. These findings demonstrated a significant association, with EEG showing the highest correlation. The participants with a nervous and anxious driving style exhibited substantial biological changes, while those with a reckless driving style showed minimal variations [22].

A comprehensive review of fatigue studies involving bus and truck drivers revealed significant findings. The review encompassed various approaches, including biochemical and psychophysiological tests, acoustic speech analysis, continuous heart rate monitoring, and questionnaire studies. These approaches indicated notable psychophysiological, speech, and subjective changes associated with driver fatigue, shedding light on symptoms experienced during prolonged driving as attributable to fatigue [23].

Another study examined the impact of fatigued and drunk driving on drivers' physical characteristics. It involved 25 participants and measured parameters such as systolic blood pressure, heart rate, eyesight, and reaction times. The results showed that fatigue driving had a greater impact on heart rate and reaction times, while drunk driving affected parameters such as blood pressure, reaction times, and depth perception [24].

Various methods for estimating vehicle acceleration could enhance Advanced Driving Assistance Systems (ADAS) for improved traffic safety. However, many models overlook individual driver characteristics and driving behavior diversity. This study proposes a machine learning model that incorporates driving behavior analysis to improve acceleration estimation accuracy [25].

A study found that higher pre-driving arousal led to better driving behavior, while frequent braking increased post-drive arousal. Sunshine and having passengers boosted well-being. Low pre-driving arousal negatively impacted driving behavior, and monotonous driving situations at high speeds with low cognitive demand increased the risk of tiredness and impaired driving, suggesting the need for timely interventions to enhance well-being while driving [26].

The research found that drivers' tension increases in specific road segments, including the transition to tunnel entrances, tunnel themselves, and open road portions. The analysis of speed profiles revealed that vehicles often accelerated before exiting a tunnel and slowed down before entering one, with the transition and tunnel segments showing the highest levels of stress [27].

2.4 Wearable sensors to monitor physiological data

Stress represents the brain's responses to some factors: an individual's traits and surroundings [28]. Stress can be measured in several ways. The conventional approach for evaluating stress is via cortisol level measurements [29] or via self-reports (e.g., Perceived Stress Scale or PSS) [30]. These approaches cannot be used to quantify stress continually over a lengthy period. In addition,

these treatments may involve a visit to a therapist or psychologist [31]. However, recent developments in the detection capabilities of wearable sensors have been employed by researchers for the continuous monitoring of stress [32] [33]. In past research works, several sensors and devices were used to monitor physiological responses; for example, Gjoreski et al. used the E4 Empatica device, which is commercially available. This device was used to detect stress both in the lab and outside the lab [34]; similarly, Mishra et al. used a sensor, Polar H7, with an EDA sensor to detect stress both in and outside the lab [35] [36].

Other studies have also used wearable sensors to detect stress rather than determine stress during driving. Under challenging circumstances, student pilots frequently experience stress [37]. As a result, human factors often serve as the primary contributing component in aircraft accidents. In this study, E4 Empatica was used to collect real-time EDA during several flight simulator sessions. According to this study, high demand-related performance correlates with high stress levels. Additionally, it was seen that students needed help to complete tough jobs despite the fact that they may have performed well overall for the session. The urban stresses discovered using the E4 Empatica in real-world investigations were caused by congested metropolitan locations, risky driving conditions, and traffic jams [3].

A proof-of-concept study continuously monitored electrodermal activity (EDA) and heart rate variability (HRV) to detect stress using a wearable sensor in an ambulatory setting [19]. The participants wore the E4 Empatica for up to 14 days, completing daily surveys on stress, emotions, cravings, pain, and discomfort. The study found that 87.86% of the EDA signals were clean, with strong associations between physiological signals and self-reported outcomes in individuals with alcohol consumption disorder, showing data quality comparable to past research. Recent emotion-recognition studies have utilized wearable sensors. A real-time mobile biofeedback system using the E4 Empatica sensor displayed five basic emotions and provided users with emotional feedback [18]. The iAware system, tested in controlled and uncontrolled conditions, increased emotional

self-awareness and decreased prediction error by 3.333% for women and 16.673% for men, highlighting its value in offering real-time biofeedback based on physiological data.

2.5 Use of Machine Learning in Predicting Arousal

Recent studies indicate a significant increase in the use of devices that measure physiological arousal via electrodermal activity (EDA) [38]. Although EDA has been recently used to measure emotions in educational contexts, substantial methodological variations exist, revealing inconsistent associations between physiological arousal and learning outcomes, thus indicating the need for explicit guidelines [38]. Detecting stress beforehand can prevent long-term conditions like depression and anxiety [39]. An overview of stress detection techniques based on EDA and machine learning (ML) shows that support vector machines and artificial neural networks achieve high accuracy in predicting stress [24].

Another study used E4 Empatica, where participants played Tetris at different difficulty levels to induce boredom, tension, and flow. The system was able to detect low-flow vs. high-flow states with 67.50% accuracy and differentiate all three affective states with a percentage of 49.23%, using an end-to-end deep learning architecture [40]. Several other studies have used E4 Empatica for real-time data collection in detecting migraine attacks, infection, health monitoring, emotion recognition, etc. [41] [42] [43] [44].

Another advancement has been made using EEG signals to detect arousal. A study employing a 64-channel EEG system with 18 healthy participants extracted ten frequency characteristics from each channel, resulting in a 640-dimensional feature vector. The genetic algorithm was used to optimize the feature selection. The method achieved high accuracy (93.11% and 81.35%), demonstrating its effectiveness in detecting arousal levels [45]. Research on risky driving behaviors in expressway tunnels introduced a driving behavior spectrum approach. It analyzed EEG, heart rate, eye movement, speed, and steering data from 30 drivers. The study showed that

speed and steering data were the most important factors in identifying driving risks at tunnel entrances and exits, where drivers experience more psychological stress. Using machine learning models (ADASYN-RF), the study achieved the highest accuracy (AUPRC = 0.978) in detecting risky behaviors [46].

A recent study on driving stress in high-altitude expressways used naturalistic driving tests and HRV data to classify stress levels with K-means clustering. It also used DeepLabv3 for semantic segmentation and extracting environmental features like trucks, buses, mountains, and sky. The model used various machine learning models like Decision Tree, XGBoost, CatBoost, and LightGBM. However, the XGBoost model outperformed all the models with an F1 score of 0.855 and a G-mean score of 0.890, showing that high traffic increases stress while open landscapes reduce it [47].

Another study also investigated driver fatigue classification in conditionally automated driving using physiological signals and machine learning. The study showed that sleep deprivation, driving environment, and sleepiness could be predicted with 99%, 85% and 73% accuracy. The study also found that EDA features were the most important features. These findings emphasize the need for real-time driver monitoring in automated vehicles to enhance safety by detecting fatigue and issuing timely warnings [48].

A recent study explored the use of electrodermal activity (EDA) signals for predicting emotional arousal and aggressive driving in young drivers. Using data from an Empatica E4 wristband and a telematics device, they extracted tonic and phasic EDA features and trained neural network models. The model predicted emotional arousal with 65% accuracy using 18 EDA features. A separate model using the top 10 features and 6 difference-based features showed moderate success in predicting aggressive behaviors like hard braking and rapid acceleration. The study highlights the potential of wearable EDA-based systems for real-time driver state monitoring and suggests the inclusion of additional physiological signals in future work [49].

Our study was conducted in the naturalistic driving environment to capture data in real-world driving conditions, allowing for the natural variability of both road environments and driver behavior. We focused on evaluating the effectiveness of machine learning models in predicting three key outcomes: emotional arousal, aggressive acceleration, and aggressive braking. Each is based on physiological features derived from Electrodermal Activity (EDA) signals.

To support the modeling of emotional arousal, participants also completed post-trip surveys assessing their emotional states. These self-reported responses were used to validate and label physiological changes, providing ground truth for training and evaluating the predictive models.

2.6 Feature Selection Techniques

Feature selection is an important step in the machine learning process, especially when working with datasets that contain a large number of features. Using all available features can be computationally expensive and time-consuming, and may even reduce model performance due to redundancy or irrelevant information. By identifying and selecting only the most relevant features—those that have the strongest relationship with the target variable—the model can be trained more efficiently and often with improved accuracy, while also improving interpretability and reducing the risk of overfitting [49].

There are many methods for feature selection. However, these methods can be categorized into three groups: filter, wrapper, and embedded methods [50]. Each category selects features in a different way:

Filter Method: These methods rank, or score features by some statistical criterion, independently of any learning algorithm [51]. Common filters include measures like correlation coefficients, chi-

square tests, mutual information, or ANOVA F-values. Features are selected or discarded based on these scores before any model is trained. Filter approaches are simple and computationally fast, making them suitable for high-dimensional data. However, because they evaluate each feature in isolation, they ignore feature interdependence and the effect on a specific predictor. So, a filter might miss combinations of features that are jointly informative.

Wrapper methods: Wrappers evaluate subsets of features by training and testing a predictive model (the "wrapper") on those features [45]. They use a search strategy, for example forward selection, backward elimination, or genetic algorithms) to find the feature subset that gives the best model performance. Wrappers can capture feature interactions better than filters because they consider the actual model outcome as the evaluation criterion. However, wrapper methods are computationally intensive and can also overfit the data.

Embedded Methods: Embedded methods perform feature selection as part of the model training process itself [45]. In these approaches, the learning algorithm has built-in mechanisms to select features. Typical examples include decision tree algorithms and regularization-based models like LASSO (L1-regularized) regression. Embedded methods incorporate interaction with the classifier (like wrappers) but are much more computationally efficient. They are less prone to overfitting than wrappers because the selection process is constrained by the model's regularization or structure.

For this research which includes driver behavior modeling with physiological signals, this work employs two well-established embedded methods for feature selection and importance analysis: Random Forest and LASSO-regularized logistic regression.

Chapter 3 : Materials and Methods

This chapter presents the methodological framework that was adopted to develop predictive models for identifying emotional arousal, aggressive braking, and aggressive acceleration. The process includes five key stages. Data collection, signal processing, feature extraction, feature ranking and selection, and model development.

3.1 Study Design

This study was conducted as a naturalistic driving experiment with 25 licensed drivers (21 male ans 4 female) between the ages of 18 and 30 with a valid US driving license, over a period of 2 weeks. Participants used their vehicle on public roads, in Duluth, during the Fall season, while performing their regular driving routines to ensure realistic driving conditions. Each participant was asked to complete multiple trips at least 5 miles in length for two-weeks period, with the goal of covering a total of 50 miles of driving over two weeks. Each participant received 50 dollars compensation for participating in the study. After each trip, drivers self-reported their emotional state before the trip, while driving, and after the trip via a short survey. Throughout the study, Empatica E4 wristbands and in-vehicle telematics devices were used to collect physiological data and driving data. Participants' own cars and public roads were chosen to capture genuine driving behaviours and stress responses.

3.2 IRB Protocol

Before the data collection was started, ethical approval was obtained from the university's Institutional Review Board (IRB). The study was reviewed and approved under the University of Minnesota's human research guidelines (IRB Study Code: STUDY00021898). All participants gave their informed consent after being briefed on the study's procedures and purpose. They were assured that their data would remain confidential. Participants received instructions about the safe

use of the recording devices and were informed of any potential risks (which were minimal, as the study was non-invasive). Figure 3.1 will give an overview of the methodology

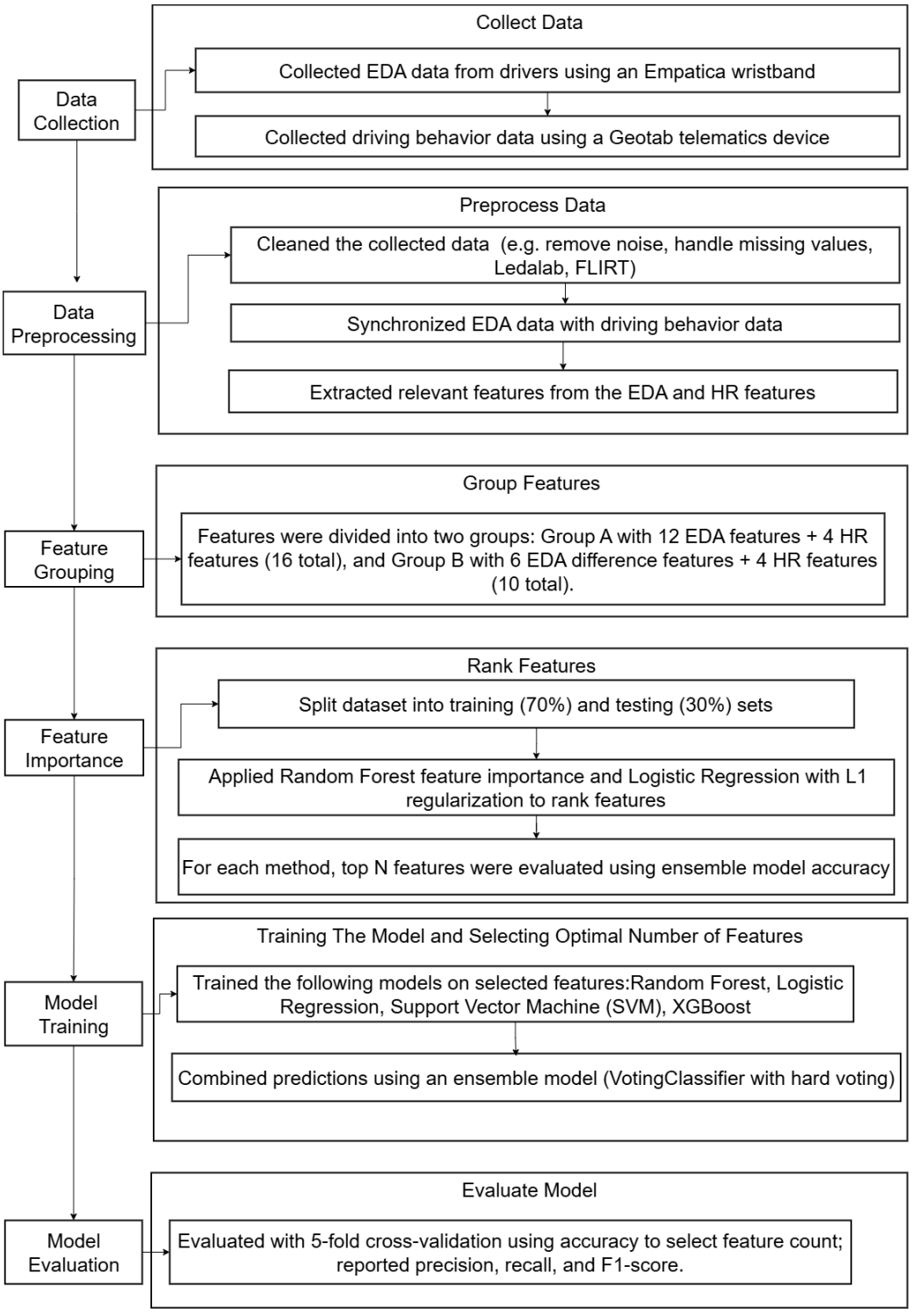

Figure 3.1: Flowchart of the Methodology

3.3 Data Collection

3.3.1 Physiological Data

Every participant wore an Empatca E4 wristband (Figure 3.2) while they were driving. Empatica E4 wristband can record Blood Volume Pulse (BVP), which leads to the calculation of HRV (Heart Rate Variability) and IBI (Inter Beat Interval), as well as SKT (skin temperature). And variations in the electrical characteristics of the skin, such as electrodermal activity (EDA). However, for this study, only EDA and Heart Rate measurements were used to create the dataset. The Empatica E4 is a clinical-grade wearable device that is equipped with multiple sensors, including a photoplethysmography (PPG) sensor for blood volume pulse and an electrodermal activity sensor. Heart rate (HR) can be derived from the PPG signals. Each participant had to set the E4 device to its recording mode during each trip; participants activated the device via a long-press button and connected it to the Empatica E4 smartphone app (E4 Realtime) before starting each driving session. The physiological data was saved to Empatica's secure cloud platform, "E4 Connect".
The Empatica E4 device samples the EDA at 4 Hz. It also measures skin conductance in microSimens. It also measures the PPG at 64 Hz, which provides high resolution data on the driver's sympathetic nervous system and cardiovascular activity.

Figure 3.2: Empatica E4 Wristband

3.3.2 Driving Data

A telematics device (Geotab Go) (Figure 3.3) was installed in each participant's vehicle to collect the driving data. This device is plugged into the driver's OBD II port. It can record various things, including acceleration, braking, and speeding data. Participants kept the telematics unit installed for the entire two-week period, allowing all trips to be logged automatically. The acceleration and braking data were saved in Geotab's online platform. The acceleration and braking data were later retrieved from the website. All telematics data were time-stamped, enabling alignment with the physiological data timeline.

Figure 3.3: Geotab Go Device

3.3.3 Self-Reported Survey

After completing each trip, participants immediately filled out a brief survey (Appendix F) about their emotional state during the drive, before driving, and after driving. The survey asked them to rate five emotions—Happy, Calm, Stressed, Fatigued, and Angry—using a 5-point scale:
Not at all, Mild, Moderate, Very, and Extreme. For analysis, these responses were binarized into two categories: “aroused” and “non-aroused.” Specifically, any response indicating a high-emotion state (such as feeling stressed, angry, fatigued, or happy) was classified as “aroused,” while a response of feeling calm was classified as “non-aroused.”

3.4 Data Processing

3.4.1 EDA Data Processing:

EDA signals are composed of two components: a tonic component and a phasic component.

Tonic EDA: Tonic EDA, also known as the skin conductance level (SCL), represents the slow-varying baseline level of skin conductance over a longer period of time. Changes in the tonic level gradually increase and decrease and are thought to reflect general autonomic arousal.

Phasic EDA: The phasic EDA component corresponds to rapid, transient changes in skin conductance known as skin conductance responses (SCRs). Phasic SCR events are short-lived peaks triggered by stimuli or sudden arousal, and their amplitudes indicate the intensity of momentary sympathetic activation.

3.4.2 EDA Signal Preprocessing using FLIRT Toolkit

The raw EDA data needs to be carefully processed before analysis. To remove the noise and separate the signal into meaningful features, an open-source FLIRT(0.0.3) was used [50]. FLIRT (Feature Learning and Importance Recognition Toolkit) is a Python-based toolkit designed for physiological signals from wearables. It provides algorithms for cleaning EDA signals and extracting standardized features for machine learning models.

Noise Filtering and Artifact Correction: EDA signals from wearable sensors often contain motion artifacts, abrupt spikes, and missing data, which can skew analysis. FLIRT addresses this by applying advanced filtering and artifact detection techniques. In particular, it leverages Extended Kalman Filter (EKF) and particle filter methods to remove high-frequency noise and smooth the signal.

- **Extended Kalman Filter:** EKF is a recursive estimation technique that combines observed measurements with a theoretical model to infer the true signal. It estimates a five-dimensional state vector:

$$x = [SC_H, k_{diff}, SC_0, SCR, S]^T \quad \text{.....(3.1)}$$

Here,

SC_H is the hydration-related component of skin conductance,

SC_0 is the baseline skin conductance,

SCR denotes the skin conductance response,

k_{diff} is inverse sweat diffusion time constant,

S is sudomotor nerve activity

The nonlinear state-space model is discretized and linearized at each step. Kalman filter prediction and correction steps are then applied to estimate the true signal. The final EDA signal is reconstructed as:

$$SC = SC_H + SC_0 + SCR \quad \text{.....(3.2)}$$

- **Particle Filter:** The PF is a Bayesian filtering method that does not assume Gaussian noise distributions, making it suitable for real-world wearable data. The PF operates on a simple linear state-space model:

$$x_{k+1} = x_k + v_k, \quad z_k = x_k + w_k \quad \text{.....(3.3)}$$

Here, x_k is the estimated skin conductance, z_k is the observed EDA measurement, and v_k and w_k are the process and measurement noise term respectively.

Furthermore, there is also a modular approach. A low-pass filter to suppress high-frequency noise above the EDA signal's typical physiological range (usually >0.5 Hz). FLIRT also includes

automated artifact detection algorithms to identify segments corrupted by movement or sensor detachment. For example, the EDAExplorer algorithm checks the rise time, slope, etc. of an EDA and flags segments as “artifact” or “clean” accordingly. Similarly, a logistic regression model (from the Ideas Lab UT) classifies each 5-second window as clean vs. noisy, so that bad segments can be removed or interpolated. Through this combination of filtering and artifact correction, the EDA signal is cleaned.

3.4.3 Tonic/Phasic Separation using cvxEDA:

After removing the noise, the next preprocessing step is to decompose the EDA into its tonic and phasic components. To do this separation, FLIRT uses the cvxEDA algorithm [51]. The cvxEDA method models the observed EDA time series as the sum of a tonic component (SCL), a phasic component (SCR activity), and additive noise. By determining this as a convex optimization problem, cvxEDA estimates the most likely underlying tonic and phasic signals that could produce the observed EDA data. cvxEDA finds a smooth underlying SCL curve (tonic level) and a series of sparse SCR impulses that, when combined, reconstruct the original signal.

Using cvxEDA via FLIRT, the E4 EDA data were split into a low-frequency tonic EDA signal and a high-frequency phasic EDA signal. This decomposition is very important for feature extraction, as it allows computing the tonic and the phasic components. Figure 3.4 will give a signal processing pipeline that FLIRT uses to clear the raw data and extract the features.

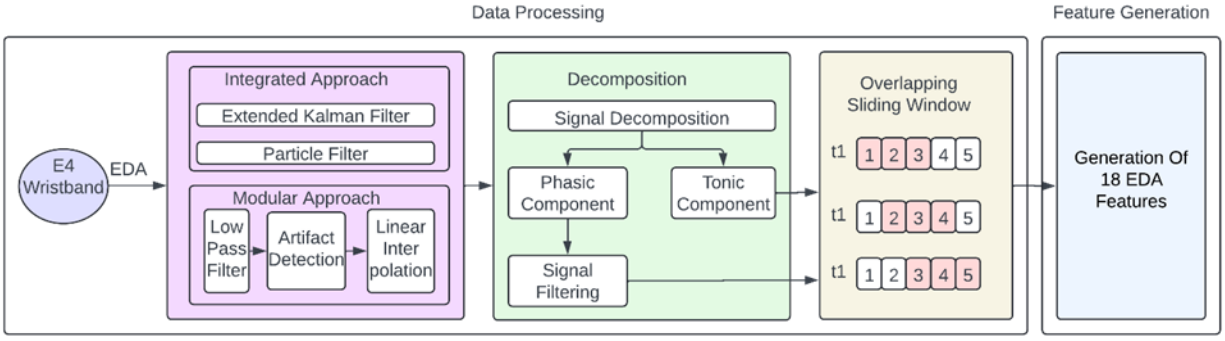

Figure 3.4: EDA signal processing pipeline using FLIRT toolkit

3.4.4 Feature Extraction from Tonic and Phasic EDA

After cleaning and separating the phasic and tonic parts of the EDA signals the quantitative features from both the tonic and phasic components were extracted to characterize the drivers' emotional state. Following prior work on EDA-based stress analysis, this work was focused on simple statistical features (mean, maximum, and variability) of each component. In this analysis, these features were computed in two ways: a baseline period (representing low arousal or rest, where the driver was calm) and a high arousal period (e.g., during a stress-inducing driving event). In total, 12 feature variables were derived (6 metrics × 2 conditions) and six difference-based features that represent the change in EDA from a baseline state. Below, the 18 extracted EDA features are defined:

1. **Base Tonic Mean:** The average tonic skin conductance level during the baseline period.
2. **High Tonic Mean:** The average tonic EDA level during the high arousal period. An elevated high tonic mean (relative to baseline) would indicate a higher overall autonomic arousal under stress
3. **Base Phasic Mean:** The average phasic EDA (SCR) amplitude during the baseline period.
4. **High Phasic Mean:** The average phasic component amplitude during the high-arousal period.
5. **Base Tonic Max:** The highest value of the tonic EDA signal in the baseline period. This reflects the maximum baseline skin conductance level reached when the driver is calm.
6. **High Tonic Max:** The highest tonic EDA value during the high-arousal period.

7. **Base Phasic Max:** The largest phasic SCR amplitude observed in the baseline period.
8. **High Phasic Max:** The maximum phasic EDA peak during the high-arousal period.
9. **Base Tonic Std:** The variability of the tonic EDA during baseline, measured as the standard deviation.
10. **High Tonic Std:** The standard deviation of the tonic EDA during the high arousal interval.
11. **Base Phasic Std:** The variability in phasic EDA amplitudes during baseline.

12. **High Phasic Std:** The variability of phasic EDA amplitudes during the high-arousal period.
13. **Phasic Mean Difference:** Phasic Mean Difference is the difference between the baseline and high phasic mean values.
14. **Tonic Mean Difference:** Tonic Mean Difference is the difference between the baseline and high tonic mean values.
15. **Phasic Max Difference:** Phasic Max Difference is the difference between the baseline and high phasic max values.
16. **Tonic Max Difference:** Tonic Max Difference is the difference between the baseline and high tonic max values.
17. **Phasic Std Difference:** Phasic Std Difference is the difference between the baseline and high phasic standard deviation values.
18. **Tonic Std Difference:** Tonic Std Difference is the difference between the baseline and high tonic standard deviation values.

Each of these features provides insight into different aspects of the driver's electrodermal response. By comparing baseline and peak values, we can quantify how much a driver's physiological arousal increases from a resting state to a high-stress state.

3.4.5 Heart Rate Features Extraction

From the heart rate time series of each trip, we extracted four statistical features: the **HR Mean**(average beats per minute, reflecting overall cardiovascular load), **HR Median** (to mitigate the effect of outliers and skew in distribution), **HR standard deviation** (variability of heart rate, indicating how much the heart rate fluctuated, which can increase under stress), and **HR skewness** (the skew of the heart rate distribution over the trip, which can reveal whether there were occasional high heart rate surges – a positively skewed distribution might indicate short bursts of elevated heart rate). These HR features complement the EDA features by capturing the cardiovascular aspect of arousal.

3.4.6 Driving Data Analysis

Driving behavior data was collected using the Geotab Go telematics device. The data were used to calculate longitudinal acceleration and braking. Longitudinal acceleration values (AccX) were analyzed to detect aggressive acceleration and braking events.

For each trip, all positive and negative AccX values were categorized into three levels of severity [55]:

Table 3.1: Levels of Driving Metrics

Range	Metric	Level
AccX<2.83 ms^2	Acceleration	Mild (2)
2.83 ms^2<AccX<3.83 ms^2		Moderate (1)
Acc>3.83 ms^2		Severe (0)
AccX>-2.83 ms^2	Braking	Mild (2)
-3.83 ms^2<AccX<-2.83 ms^2		Moderate (1)
AccX<-3.83 ms^2		Severe (0)

A weighted scoring system was applied: Level 2 contributed more positively to the score, while Level 0 contributed none. For each trip, an acceleration score and a brake score were computed as:

$$Score = \left(\frac{Total\ weighted\ points}{2 \times Total\ acceleration\ or\ braking\ events}\right) \times 10 \quad \text{.....(3.4)}$$

This resulted in scores between 0 and 10, where 10 indicates smoother driving with fewer aggressive events.

Finally, we categorized each trip as either “aggressive acceleration” or “non-aggressive acceleration” by comparing its score to the driver-specific median acceleration score. This approach introduces a personalized baseline for aggressiveness. For a given driver, if a trip’s acceleration score fell below that driver’s median score, it implied the trip was more aggressive than the driver’s usual behavior (since a lower score means more time spent in aggressive acceleration/braking). Such trips were labeled as exhibiting aggressive acceleration. Conversely, if a trip’s score exceeded the driver’s median, the trip was relatively smoother than their norm and was labeled as non-aggressive acceleration. By using each driver’s median as a benchmark, this method accounts for individual differences in driving style when flagging aggressive trips.

3.5 Feature Ranking

Given the relatively large number of features (22) and the small dataset, we performed a feature selection procedure to identify the most informative features and avoid overfitting. For subsequent analysis, we organized these features into two groups to investigate their effectiveness:

Group A contained the original EDA features plus heart rate features (i.e., the 12 direct EDA features from tonic/phasic + 4 HR features = 16 features total)

Group B contained the difference-based EDA features plus heart rate features (6 difference EDA + 4 HR = 10 features). This was done to remove redundancy and to test whether relative changes in EDA (differences from baseline) combined with HR might outperform the absolute EDA measures in predicting aggressive driving.

Two different feature importance analysis methods were applied on the training data: Random Forest feature importance and L1-regularized logistic regression.

Random Forest feature importance: Random Forest is an ensemble-based decision tree model that can be used not only for classification but also for assessing feature importance. It evaluates

how much each feature contributes to reducing impurity (e.g., Gini impurity or entropy) across all trees in the forest. In practice, feature importance is computed by averaging the decrease in impurity each time a feature is used to split a node.

This method is very useful for several reasons. First, due to its nonlinearity, it can capture complex interactions between features. Second, it is robust to multicollinearity, and finally, it provides a ranking of features that can guide dimensionality reduction and model simplification.
Random Forest is particularly useful in biomedical and behavioral datasets, where patterns may be nonlinear and affected by multiple interrelated variables.

L1-regularized logistic regression (Lasso Logistic Regression): L1-regularized logistic regression, also known as Lasso Logistic Regression, introduces a sparsity-inducing penalty to the standard logistic regression cost function. The L1 penalty encourages the model to shrink less informative feature coefficients to exactly zero, effectively performing feature selection.
L1 regularization is important it yields sparse models, improving interpretability by retaining only the most informative features. It helps prevent overfitting, especially with high-dimensional physiological data. Finally, it is suitable when feature selection and classification need to be performed simultaneously.

By comparing the feature rankings and subsets selected by both Random Forest and Lasso Logistic Regression, this study gains complementary insights into which physiological markers (e.g., tonic/phasic features, heart rate metrics) are most predictive of aggressive driving and emotional arousal.

3.6 Feature Selection

We applied both feature selection methods to each of the two feature sets (Group A and Group B) and considered both target outcomes (emotional arousal vs. non-arousal, and aggressive vs. non-aggressive driving).

Using each method, we obtained a ranked list of features from most important to least. To determine an optimal subset of features, we then employed a sequential feature addition approach (wrapper method with cross-validation). The procedure was as follows:

1. **Rank Features:** For a given feature ranking for example: the order given by Random Forest importance on Group A features, list the features from highest importance to lowest.

2. **Incremental Training:** Starting with the single most important feature and then adding features one by one according to the ranking. For each incremental feature set (top 1 feature, top 2 features, top 3, and so on), train a classifier (in this case, we used the ensemble model described in Section 3.7 as the classifier for consistency) on the training data and evaluate its performance using cross-validation. 5-fold cross-validation was used on the training set for these evaluations: the data was split into 5 folds, and in each iteration the model was trained on 4 folds and validated on the 1 remaining fold, so that every data point gets predicted once. The average accuracy across the 5 folds was recorded for each feature set size.

3. **Early Stopping Criterion:** The validation accuracy of the model was monitored as features were added. If adding a new feature did not improve the accuracy or if the accuracy decreased, we stopped adding more features. This point indicated the optimal number of features.

3.7 Model Training

To predict emotional arousal and aggressive driving behaviors based on physiological features, four supervised machine learning models were implemented: Logistic Regression, Support Vector

Machine (SVM), Random Forest (RF), and XGBoost. These models were chosen to balance interpretability, performance, and diversity in learning mechanisms.

3.7.1 Logistic Regression

Logistic Regression is a linear model used for binary classification. It estimates the probability that a given input belongs to a specific class by applying a sigmoid function. From equation 3.5,

$$P(y = 1|X) = \frac{1}{1+e^{-(\beta_0+\beta_1 x_1+\dots\dots\dots+\beta_n x_n)}} \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\text{(3.5)}$$

where,

- $P(y = 1|X)$ is the predicted probability that output y belongs to class 1, given a feature vector X.
- β_0 is the intercept.
- $\beta_1, \dots\dots \beta_n$ are the model weights.
- $x_1, x_2, \dots\dots x_n$ are the input features,

Logistic regression is easy to implement and interpret, and it performs well when the relationship between features and the target variable is approximately linear. It also served as the base model for L1-regularized feature selection in this study.

3.7.2 Support Vector Machine (SVM)

Support Vector Machines are margin-based classifiers that find the hyperplane which maximally separates data points of different classes. It is described by equation 3.6

$$maximize\ \frac{2}{||w||}\ subject\ to\ y_n(w^T x_n + b) \geq 1 \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\text{(3.6)}$$

where,

- w is the weight vector that defines the orientation of the hyperplane
- b is the bias (intercept) term that shifts the hyperplane
- ||w|| is the Euclidean norm (magnitude) of the weight vector

- x_n is the feature vector for the nth training sample
- y_n is the class label for the nth sample
- $y_n(w^T x_n + b) \geq 1$ ensures that all data points are correctly classified and lie on the correct side of the margin.

SVMs are effective in high-dimensional spaces and are robust to overfitting, particularly with smaller datasets. The use of kernels allows SVMs to capture nonlinear decision boundaries.

3.7.3 Random Forest

Random Forest is a popular machine learning algorithm used for both classification and regression tasks. It's an ensemble learning method that combines multiple decision trees to improve accuracy and reduce overfitting. Each decision tree in the forest is trained on a random subset of the data and features, and their predictions are then combined, typically through majority voting for classification or averaging for regression.

Key advantages of Random Forest include:

- High accuracy with minimal hyperparameter tuning
- Robustness to noisy and missing data
- Built-in feature importance estimation

3.7.4 XGBoost

Extreme Gradient Boosting (XGBoost) is a highly efficient and scalable implementation of gradient boosting. It builds an additive model in a forward stage-wise fashion, where each new tree corrects the residuals (errors) of the previous trees.

The final prediction function in XGBoost is expressed by equation 3.6

$$f(x) = \sum_{k=1}^{K} f_k\,(x), f_k \in \xi \quad \text{.....(3.7)}$$

where,

- $f(x)$ is the final value for the input x
- K is the total number of tresses
- $f_k(x)$ is the prediction from k-th decision tree
- ξ is the space of all possible regression trees
-

A decision tree in this context is a structure composed of internal nodes which test a specific feature value, branches which represent the outcomes of those tests, and leaf nodes which output a prediction. In XGBoost, each tree tries to correct the errors made by the previous ones by learning the residuals.

In this additive model, trees are added one at a time. Each new tree $f_k(x)$ is trained to minimize residual errors.

Each model was trained using the selected features (as determined in Section 3.6) for each feature set scenario (Group A and Group B) and each prediction task (arousal, aggressive acceleration, aggressive braking).

Training was done in Python (using scikit-learn for Logistic, SVM, RandomForest, and XGBoost library for the gradient boosting model). We ensured that within each cross-validation fold, feature selection was performed on the training portion only, to avoid any look-ahead bias in feature selection

3.7.5 Ensemble Model

After training the individual models, an ensemble model was created using a hard voting classifier. The ensemble combined the predictions of the four classifiers mentioned above. In hard voting,

each base model casts a vote for the predicted class (aggressive or not, for instance), and the majority vote determines the final output. Equal weight was given to each model's vote.

The intuition behind ensembling was that each algorithm might capture different patterns in the data, for example, the SVM might detect a particular nonlinear combination of features while the Random Forest might focus on different feature interactions, so the ensemble could improve overall robustness and accuracy by using the strengths of all models. The VotingClassifier from scikit-learn was used to implement this ensemble. We trained the voting ensemble on the training folds during cross-validation and also on the full training set when preparing the final model. The final output of the ensemble is the predicted class of driving behavior for each trip.

3.8 Evaluation

The performance of the models was evaluated using cross-validation. All model training was embedded in a 5-fold cross-validation framework. This means the entire set of 116 trips was split into five folds of roughly equal size. Each model was trained five times, and each time one fold was used as a validation set. Through this cross-validation, we obtained an estimate of how well the model can perform with unseen data. However, for feature selection using Random Forest feature importance and L1-regularized Logistic Regression, the dataset was split into 70% for training and 30% for testing to evaluate feature relevance.

The primary metric used during feature selection was classification accuracy (the percentage of correct predictions) on the validation folds. Accuracy was chosen to guide feature selection because it provides a single summary of model performance, and we wanted to maximize overall correctness when deciding how many features to include. Once the feature subsets were finalized, we assessed the final models (each algorithm and the ensemble) using a broader set of metrics to capture different aspects of performance: Precision, Recall, and F1-score. These metrics are derived from the confusion matrix, which summarizes the number of:

- True Positives (TP): Correctly predicted positive cases.
- True Negatives (TN): Correctly predicted negative cases.
- False Positives (FP): Negative cases predicted as positives
- False Negatives (FN): Positive cases predicted as negative

Accuracy

Accuracy represents the ratio of totally correct predictions on both positive and negative cases out of all predictions:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad \text{(3.8)}$$

Precision

Precision measures how many predicted positive cases are actually positive.

$$Precision = \frac{TP}{TP+FP} \quad \text{(3.9)}$$

Recall

Recall measures how many actual positive cases are correctly predicted.

$$Recall = \frac{TP}{TP+FN} \quad \text{(3.10)}$$

F1- Score

The F1-score is the harmonic mean of precision and recall. It provides a balance between them and is especially useful when class distribution is imbalanced.

$$F1 - Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad \text{(3.11)}$$

In summary, the methodology involved collecting synchronized physiological and driving data in a real-world setting, extracting meaningful features of driver arousal, selecting an efficient feature

subset, and training multiple machine learning models (and an ensemble) to predict aggressive driving behavior. The models were validated with cross-validation, and their effectiveness was measured using accuracy, precision, recall, and F1-score.

Chapter 4 : Results

A total of 25 drivers participated in this study, during which 116 driving sessions were recorded. Based on the pre-drive and post-drive surveys, 50 of these sessions were classified as "aroused," while the remaining 66 were labeled as "not aroused." Each participant was involved in the study for approximately two weeks and drove an estimated 50 miles during the data collection period. A few sessions were excluded from the analysis due to missing data. For the final set of 116 valid trips, 18 Electrodermal Activity (EDA) features and four heart rate features were extracted and used for modeling. These features were divided into two separate groups to prevent redundancy, as several features were derived from others.

The first feature group consisted of the 12 EDA features representing the base, high, and max values of the phasic and tonic components. The other feature group consists of the six “difference features” derived from the base and high values of the phasic and tonic components. The Heart rate features are included in both sets.

In this study, we investigated the relationship between physiological responses and aggressive driving behaviors by analyzing Electrodermal Activity (EDA) signals collected using the Empatica E4 wristband. Specifically, we aimed to predict three target outcomes: emotional arousal, aggressive acceleration, and aggressive braking.

A total of 22 features were extracted from the EDA signals and heart rate data, including tonic and phasic components (mean, max, and standard deviation), difference features, and statistical descriptors of heart rate (mean, median, standard deviation, and skewness). To ensure meaningful input representation, these features were separated into two groups for comparison: one with the raw tonic/phasic values and heart rate features, and another with the difference features and heart rate.

For feature selection, two methods were employed: Random Forest feature importance and L1-regularized Logistic Regression (Lasso). In each case, the dataset was split into 70% training and 30% testing during the feature selection process to identify the most informative inputs.
Then, in each feature selection approach, features were ranked based on their importance scores. Following this, we employed an iterative model evaluation strategy to determine the optimal number of features. Starting with only the top-ranked feature that we got from the feature importance chart, we incrementally added one feature at a time and trained an ensemble model (VotingClassifier) at each step. The ensemble model combined the predictions of four classifiers—Random Forest, Logistic Regression, Support Vector Machine (SVM), and XGBoost—using a hard voting mechanism.

For every subset of features, the ensemble model's accuracy was evaluated using five-fold stratified cross-validation. This process continued until the ensemble accuracy reached its peak. Once the maximum accuracy was identified, the corresponding number of top features at that point was selected as the optimal feature subset for each prediction task.

This section presents results from the following feature selection techniques:

1. Arousal Prediction Based on Random Forest Feature Importance
2. Arousal Prediction Based on L1-Regularized Logistic Regression Feature Importance
3. Aggressive Acceleration Prediction Based on Random Forest Feature Importance
4. Aggressive Acceleration Prediction Based on L1-Regularized Logistic Regression Feature Importance
5. Aggressive Braking Prediction Based on Random Forest Feature Importance
6. Aggressive Braking Prediction Based on L1-Regularized Logistic Regression Feature Importance

4.1 Predicting Arousal

4.1.1 Arousal Prediction Based on Random Forest Feature Importance

To predict arousal, the random forest feature selection was utilized to get the important features, and two distinct sets of input features were evaluated separately. In the first approach, the first set of data was used with 12 features and the four heart rate features to predict the Arousal.

Figure 4.1 displays the ranked importance of these features, highlighting 'High Tonic std', 'High phasic max' and 'SKEW' as among the most influential in arousal prediction.

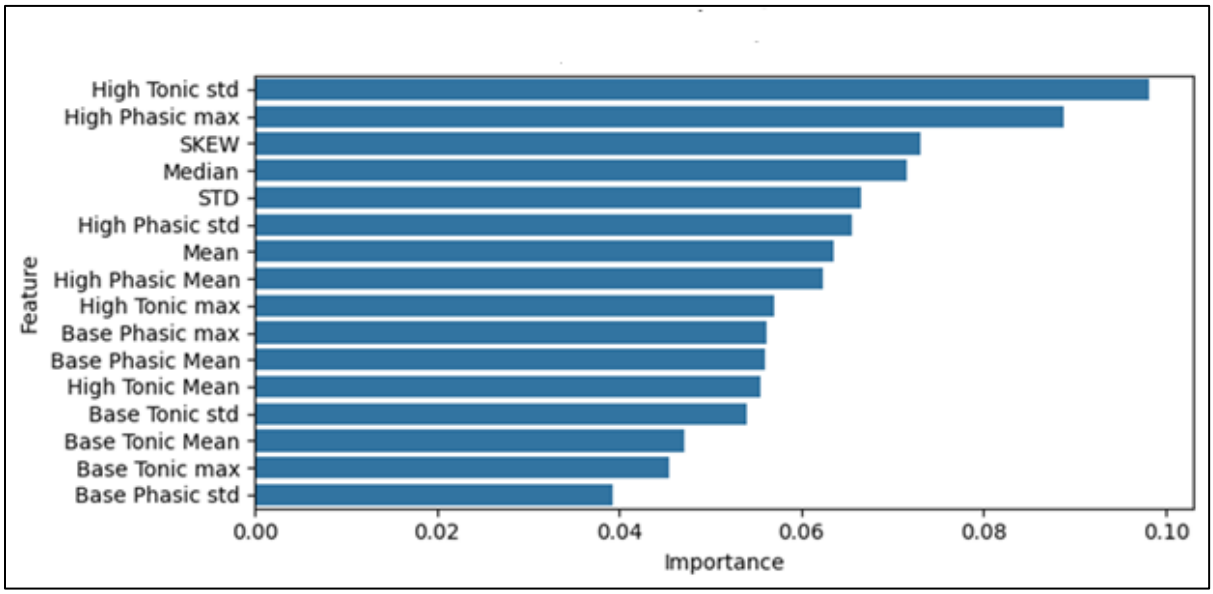

Figure 4.1: Feature importance chart for 16 initial features in arousal prediction using Random Forest

Based on the feature importance ranking, the optimal number of features for predicting arousal was determined through a systematic evaluation process. This involved training an ensemble model, specifically a VotingClassifier, which used different subsets of top-ranked features, ranging from 1 to 16. For each subset, the model's accuracy was evaluated using five-fold cross-validation. The highest accuracy was observed when only the top two features were used; beyond this point,

the performance began to decline. To prevent overfitting, an early stopping criterion with a patience threshold was applied, and once no further improvement was observed, the process was halted. The top two features were then used as inputs for arousal prediction using RF, SVM, LR, XGBoost, and an ensemble model.

The VotingClassifier combined the predictions of four base classifiers—Random Forest, Logistic Regression, XGBoost, and SVM—using a hard voting strategy. In this method, each model cast a vote for a predicted class, and the final output was determined by majority vote. This ensemble approach leveraged the strengths of multiple algorithms to improve prediction accuracy and model robustness.

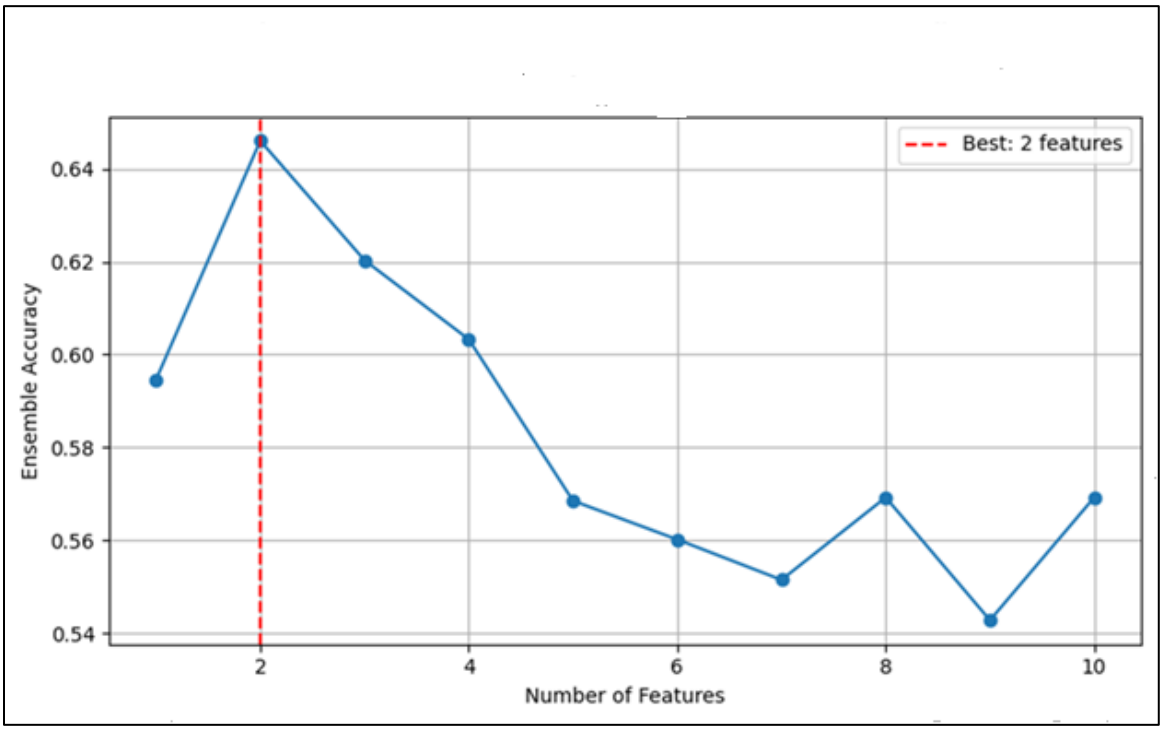

Figure 4.2: Chart showing Ensemble model accuracy for arousal prediction using different numbers of top features ranked by Random Forest feature importance from the original 16 features.

Table 4.1 summarizes the classification performance of all models using the first two features. SVM and Logistic Regression showed the best individual performance, while the ensemble classifier slightly outperformed individual models across all metrics.

The VotingClassifier now combines the predictions of four base classifiers—Random Forest, Logistic Regression, SVM, and XGBoost —using a hard voting strategy. In this method, each model cast a vote for a predicted class, and the final output was determined by majority vote. This ensemble approach leveraged the strengths of multiple algorithms to improve prediction accuracy and model robustness. The ensemble achieved an accuracy of 0.65, a precision of 0.69, a recall of 0.60, and an F1-score of 0.57.

Table 4.1: Evaluation metrics of different machine learning models for predicting arousal using 2 features selected from the original 16 based on Random Forest feature importance.

Model	Accuracy	Precision	Recall	F1 Score
Logistic Regression	0.61	0.66	0.56	0.52
Random Forest	0.58	0.56	0.56	0.56
XGBoost	0.50	0.49	0.49	0.49
SVM	0.64	0.69	0.59	0.55
Ensemble	0.65	0.69	0.60	0.57

Figure 4.3 visually compares the average model scores. Here, the ensemble model offers a balanced trade-off across evaluation metrics and performs robustly compared to individual classifiers. The ensemble model outperforms individual models by achieving the highest accuracy

(~0.65), precision (~0.69), and recall (~0.60), with a balanced F1-score (~0.57). In contrast, individual models like SVM and logistic regression showed slightly lower scores, and XGBoost had the weakest performance across all metrics.

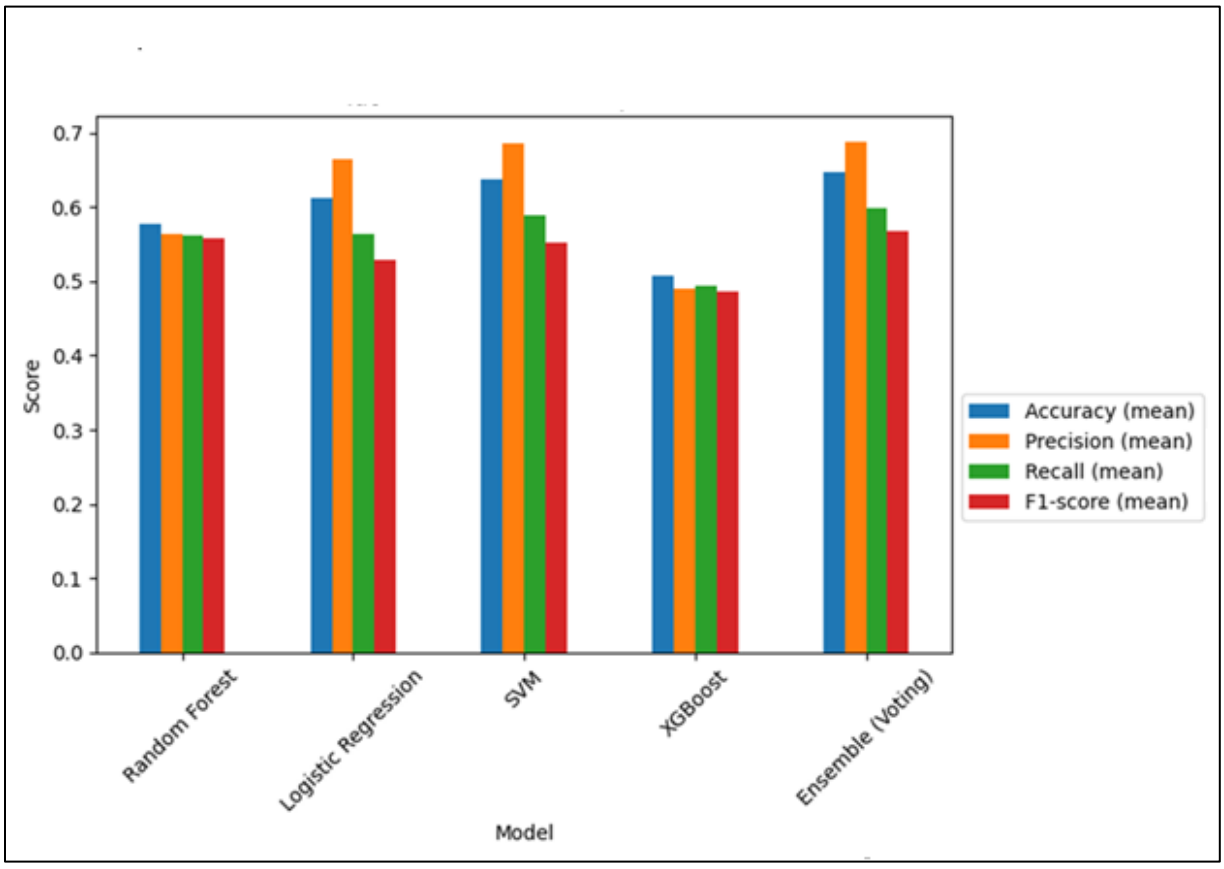

Figure 4.3: Comparison of performance metrics for different machine learning models in predicting arousal using two features selected from the original 16 based on the Random Forest feature importance method.

After that, the second set of features, consisting of 10 variables, was used to predict arousal. This set included six difference-based EDA features, derived from the base and high values of phasic and tonic components, combined with the same four heart rate features. Like the previous model, Random Forest was employed to compute feature importance before model training, rather than directly using all features.

Figure 4.4 illustrates the feature importance ranking for this feature set. The Phasic Max Difference emerged as the most influential feature, demonstrating the highest correlation with arousal prediction.

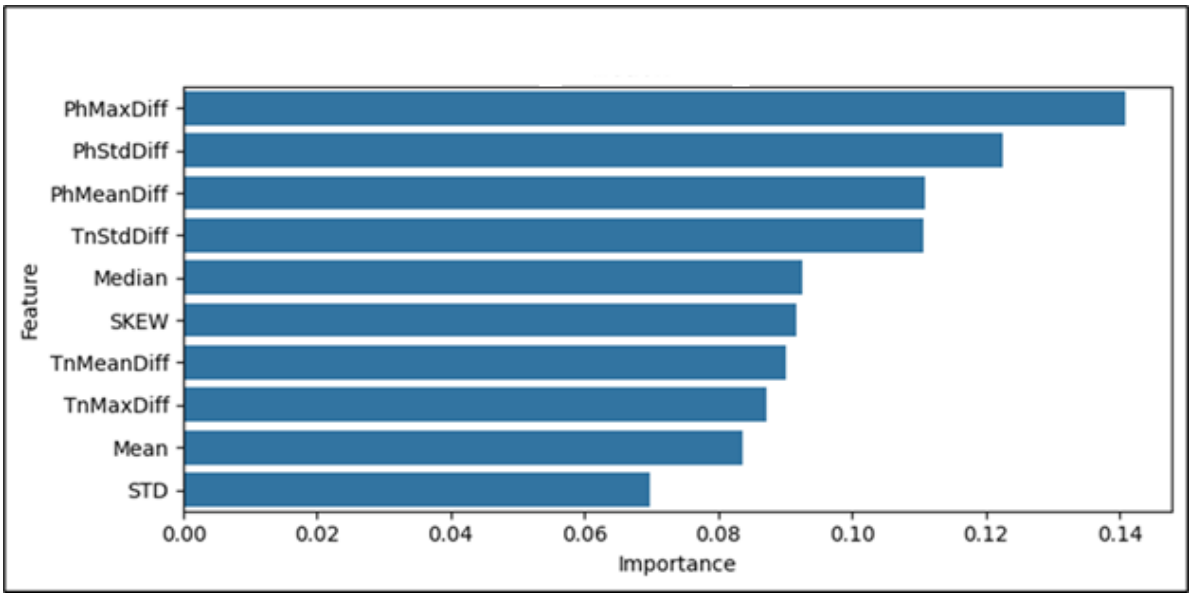

Figure 4.4: Feature importance chart for 10 features in arousal predicting using Random Forest

From this ranking of the features, to determine the optimal number of features, again, an ensemble model (VotingClassifier) was evaluated across a range of top-ranked features from 1 to 10. As shown in Figure 4.5, the ensemble model accuracy was maximum with two features. After that, the accuracy decreased as the number of features increased. An Early stopping with a patience threshold was applied to avoid overfitting. The highest ensemble accuracy was achieved with two features.

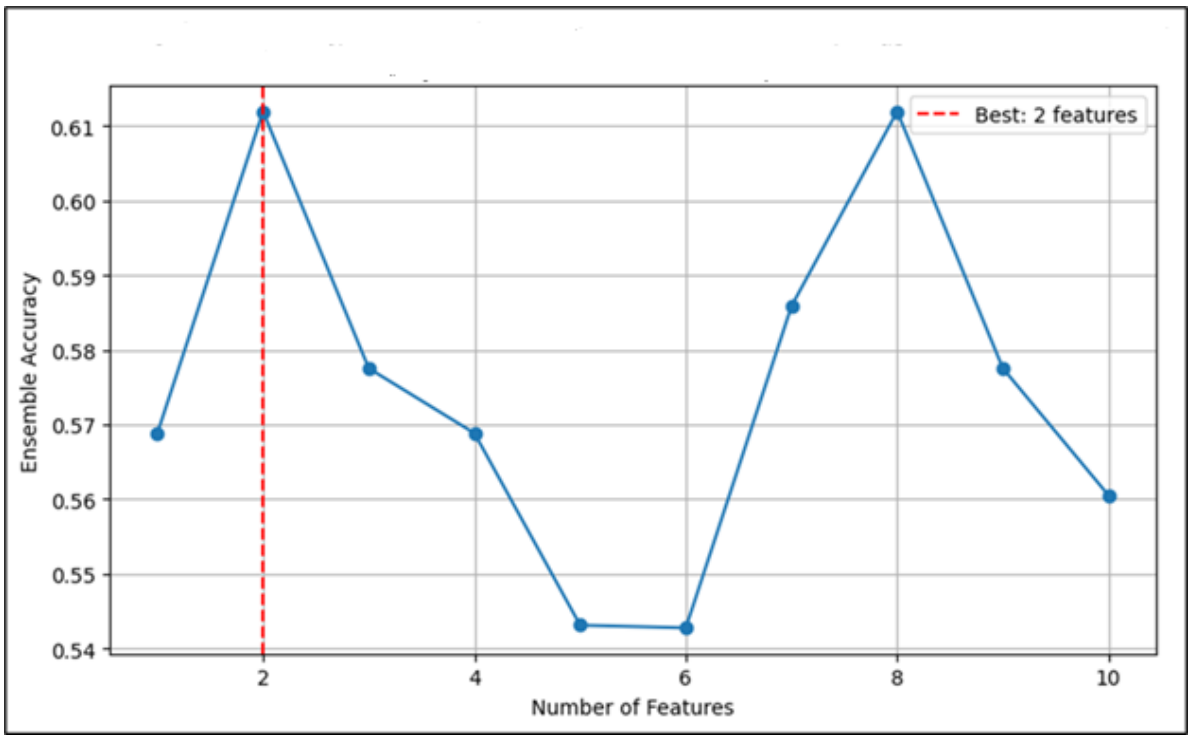

Figure 4.5: Chart showing ensemble model accuracy for arousal prediction using different numbers of top features ranked by Random Forest feature importance from the 10 features.

Table 4.2 presents the classification performance of all models using the top two features obtained through the feature selection method. Among the individual classifiers, Logistic Regression, Random Forest, and SVM demonstrated relatively strong results, with all achieving an accuracy of 0.61. However, the ensemble model continued to show a slight edge in overall performance.

By combining the strengths of multiple algorithms, the ensemble approach gave the most balanced results, achieving the highest precision (0.65) and F1-score (0.51), along with consistent accuracy (0.61) and recall (0.56). These results further reinforce the advantage of ensemble learning, which, even with a minimal feature set, provided more robust and reliable performance compared to any single model.

Table 4.2: Evaluation metrics of different machine learning models for predicting arousal using 2 features selected from 10 features based on Random Forest feature importance.

Model	Accuracy	Precision	Recall	F1 Score
Logistic Regression	0.61	0.63	0.57	0.55
Random Forest	0.61	0.55	0.55	0.54
XGBoost	0.57	0.56	0.55	0.55
SVM	0.61	0.63	0.55	0.53
Ensemble	0.61	0.65	0.56	0.51

Figure 4.6 visually compares all the models, where it can be clearly seen that the ensemble model is performing better than the individual models in predicting arousal.

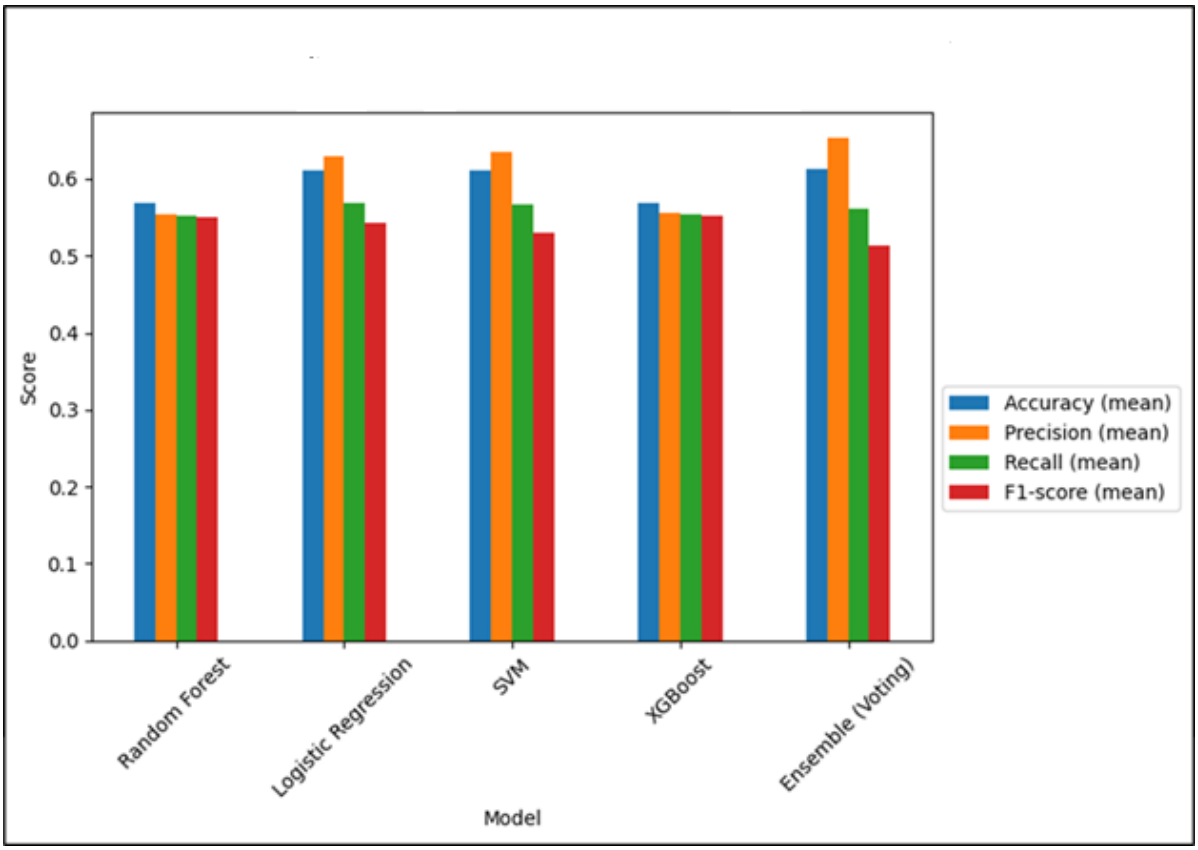

Figure 4.6: Comparison of performance metrics for different machine learning models in predicting arousal using two features selected from the 10 features based on the Random Forest feature importance method

4.1.2 Arousal Prediction Based on L1-Regularized Logistic Regression Feature Importance

In this approach, arousal prediction was performed using a set of 16 input features comprising 12 EDA and 4 heart rate features. Feature importance was first computed using L1-regularized logistic regression, as shown in Figure 4.7.

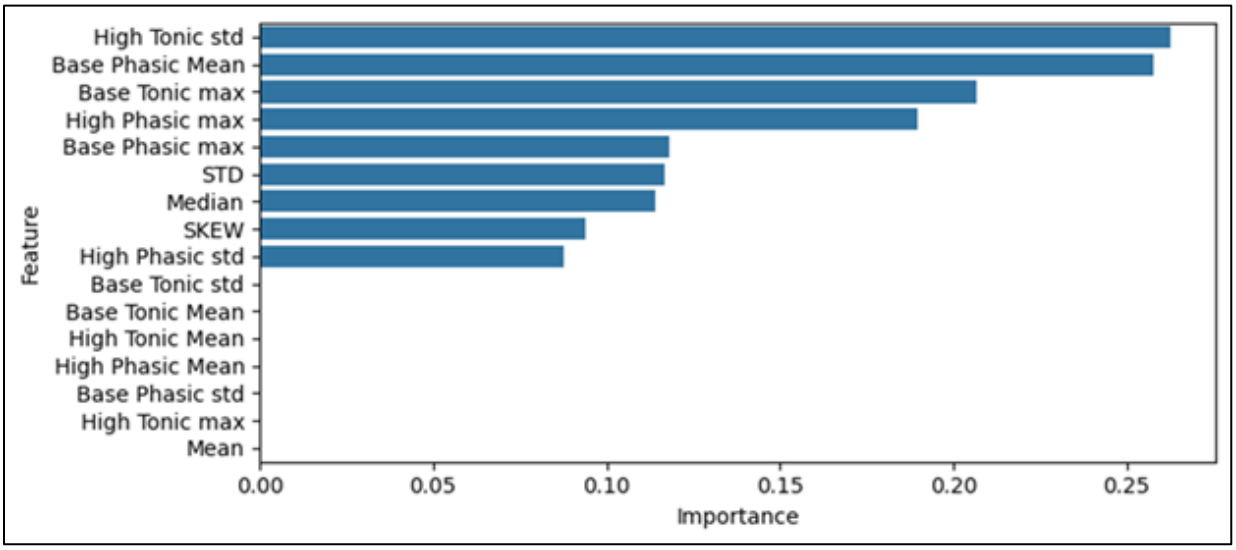

Figure 4.7: Feature importance chart for 16 initial features in arousal prediction using L1-regularized Logistic Regression

The top three features—High Tonic std, Base Phasic Mean, and Base Tonic Max—were found to have the highest importance scores, indicating a strong correlation with emotional arousal. These were then used in an ensemble classifier to determine the optimal feature subset for prediction. As shown in Figure 4.8, ensemble accuracy peaked with the top three features, achieving a maximum of 64.5%. Beyond this point, the accuracy declined, suggesting that adding more features introduced noise or redundancy rather than improving predictive performance.

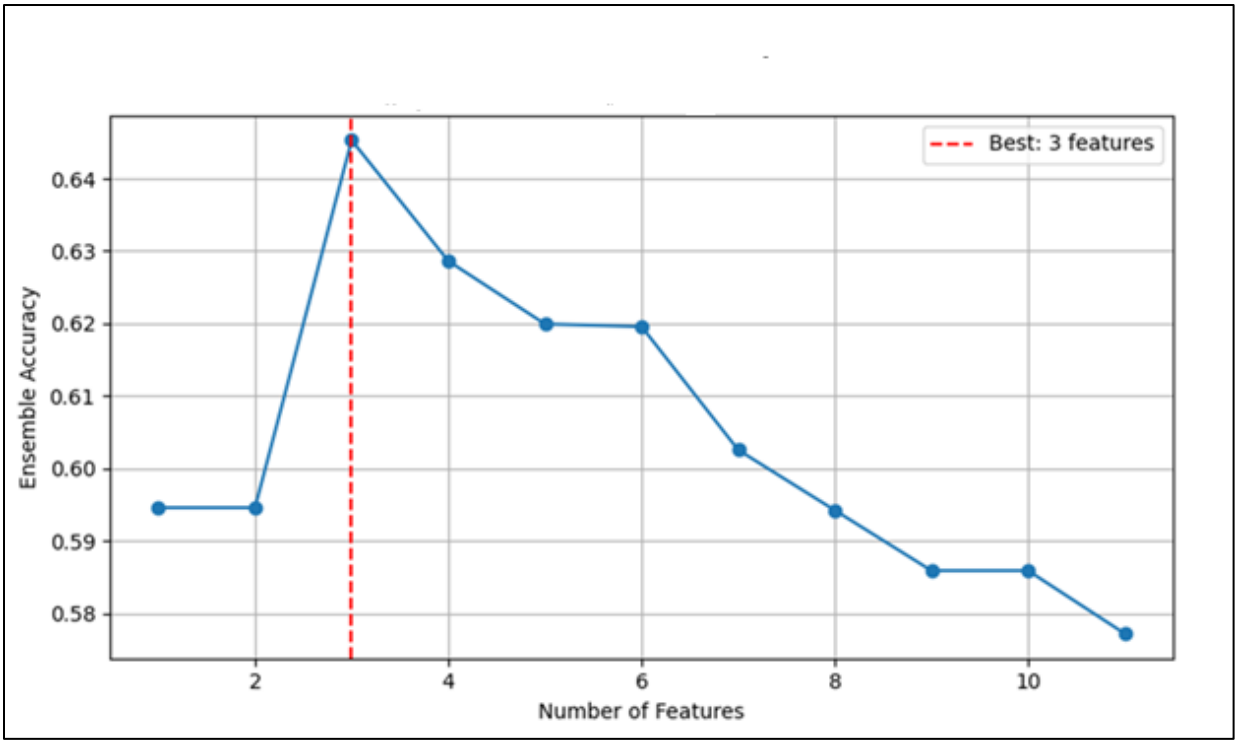

Figure 4.8: Chart showing ensemble model accuracy for arousal prediction using different numbers of top features ranked by L1-regularized Logistic Regression feature importance from the original 16 features

Table 4.3 presents the classification results using these three selected features. The ensemble model outperformed all individual classifiers, achieving the highest overall accuracy (0.65) and strong performance across other metrics, including precision (0.64) and recall (0.59). Logistic regression and SVM also performed well individually, with logistic regression achieving F1-score (0.55) same as the ensemble (0.57), while SVM maintained a balanced profile with an accuracy of 0.62 and F1-score of 0.51.

Table 4.3: Evaluation metrics of different machine learning models for predicting arousal using 3 features selected from the original 16 based on L1-regularized Logistic Regression feature importance

Model	Accuracy	Precision	Recall	F1 Score
Logistic Regression	0.63	0.64	0.54	0.55
Random Forest	0.56	0.54	0.53	0.53
XGBoost	0.48	0.46	0.47	0.46
SVM	0.62	0.61	0.57	0.51
Ensemble	0.65	0.64	0.59	0.55

Figure 4.9 visually compares all the models, where it can be clearly seen that the ensemble model is performing better than the individual models in predicting arousal.

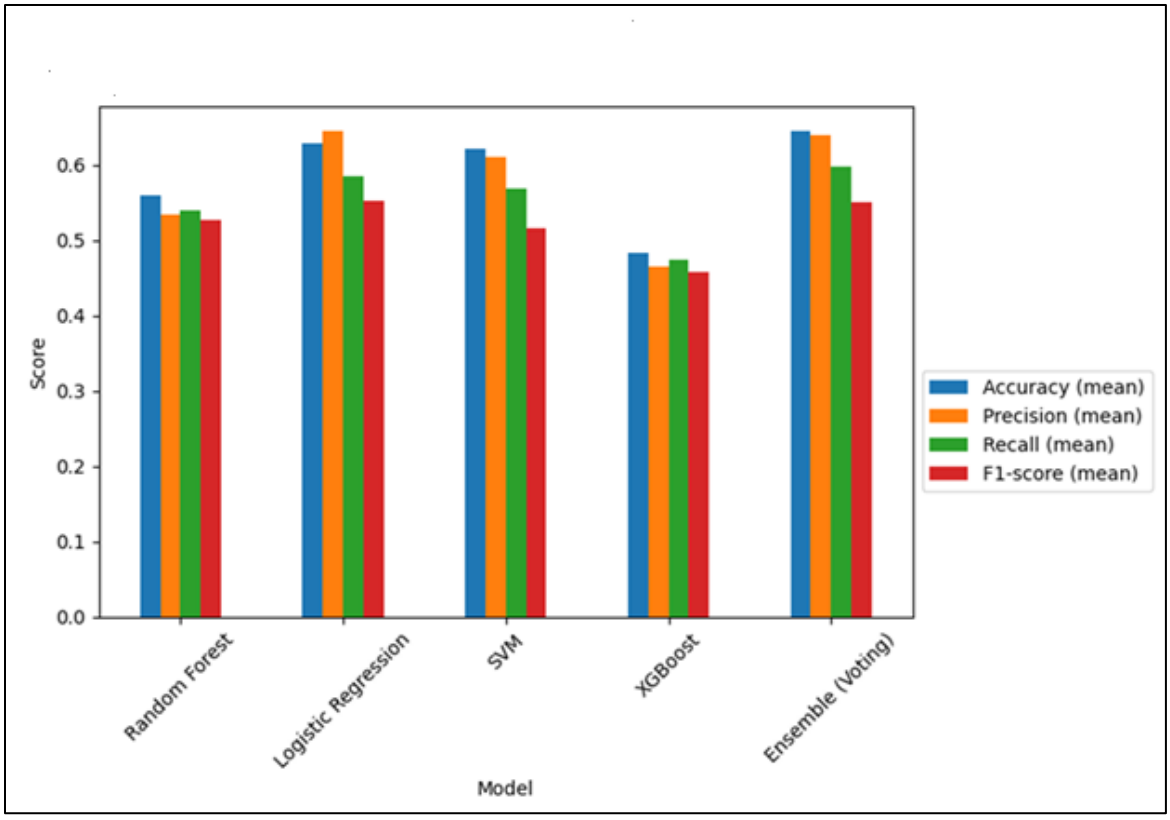

Figure 4.9: Comparison of performance metric for different machine learning models in predicting arousal using three feature selected from the original 16 features based on the L1-regularized Logistic Regression feature importance method

Now for the arousal prediction, the 6 difference features were used with the four heart rate features. Again, the logistic regression feature importance was used as shown in figure 4.10.

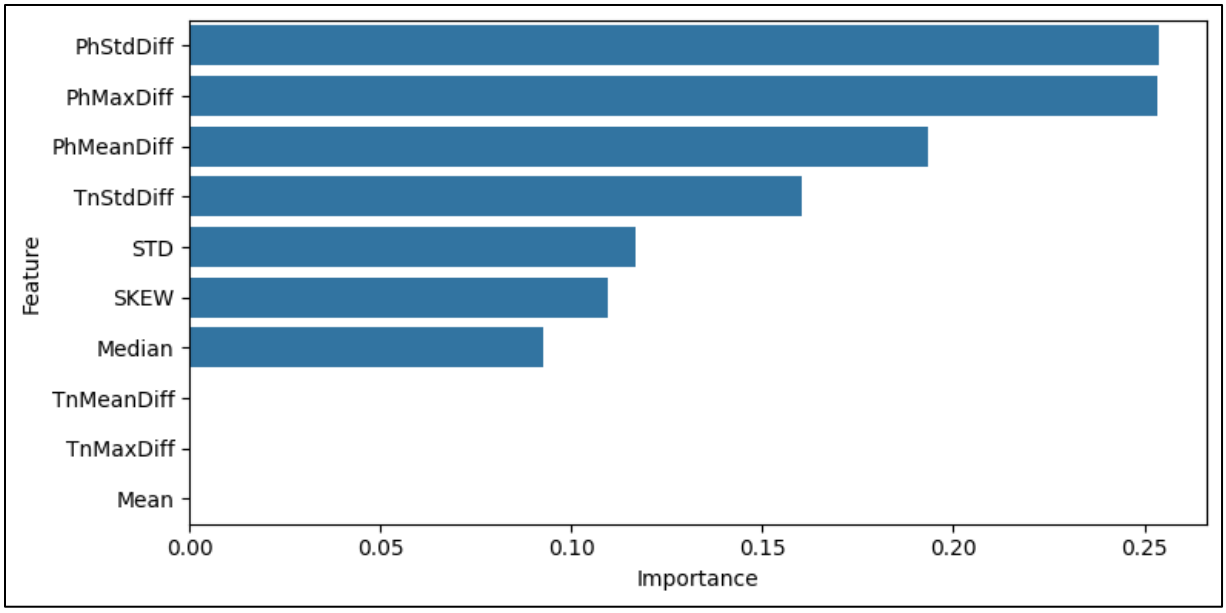

Figure 4.10: Feature importance chart for 10 features in arousal prediction using L1-regularized Logistic Regression

As shown in Figure 4.10, the features Phasic Std Difference (PhStdDiff) and Phasic Max Difference (PhMaxDiff) were assigned the highest importance weights. To identify the optimal subset of features for modeling, a VotingClassifier ensemble was evaluated across a range of top-ranked features (1 to 10), with performance measured using five-fold cross-validation. As shown in Figure 4.11, ensemble accuracy peaked at 0.61 with just the top two features.

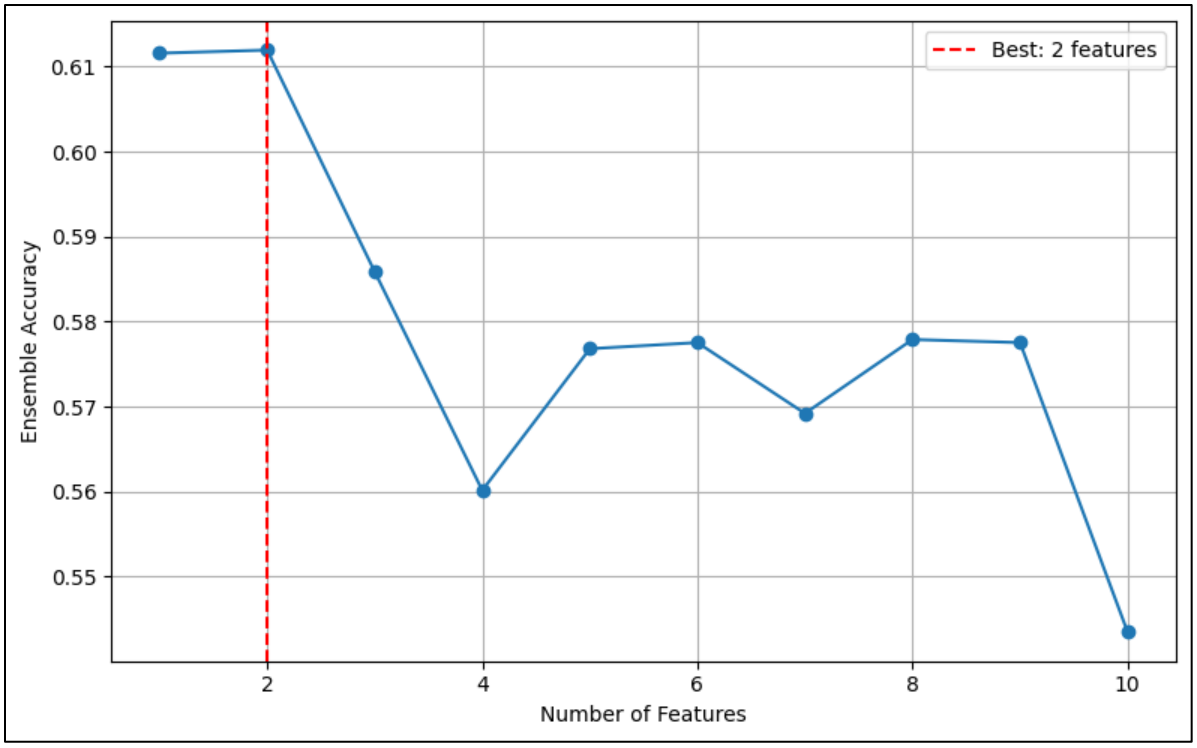

Figure 4.11: Chart showing ensemble model accuracy for arousal prediction using different numbers of top features ranked by L1-regularized Logistic Regression feature importance from the 10 features

Table 4.4 presents the classification results using the top two features. The ensemble model achieved the highest precision (0.65) while maintaining strong overall accuracy (0.61). Logistic regression and SVM also performed consistently well, each achieving 0.61 accuracy and balanced precision and recall scores. In contrast, Random Forest and XGBoost performed moderately, both yielding similar F1-scores and accuracy like the other models.

Table 4.4: Evaluation metrics of different machine learning models for predicting arousal using 2 features selected from the 10 features based on L1-regularized Logistic Regression feature importance

Model	Accuracy	Precision	Recall	F1 Score

Logistic Regression	0.61	0.63	0.57	0.53
Random Forest	0.56	0.54	0.54	0.54
XGBoost	0.57	0.56	0.55	0.55
SVM	0.61	0.63	0.57	0.53
Ensemble	0.61	0.65	0.56	0.52

Figure 4.12 visually compares all the models, where it can be clearly seen that the ensemble model is performing better than the individual models in predicting arousal.

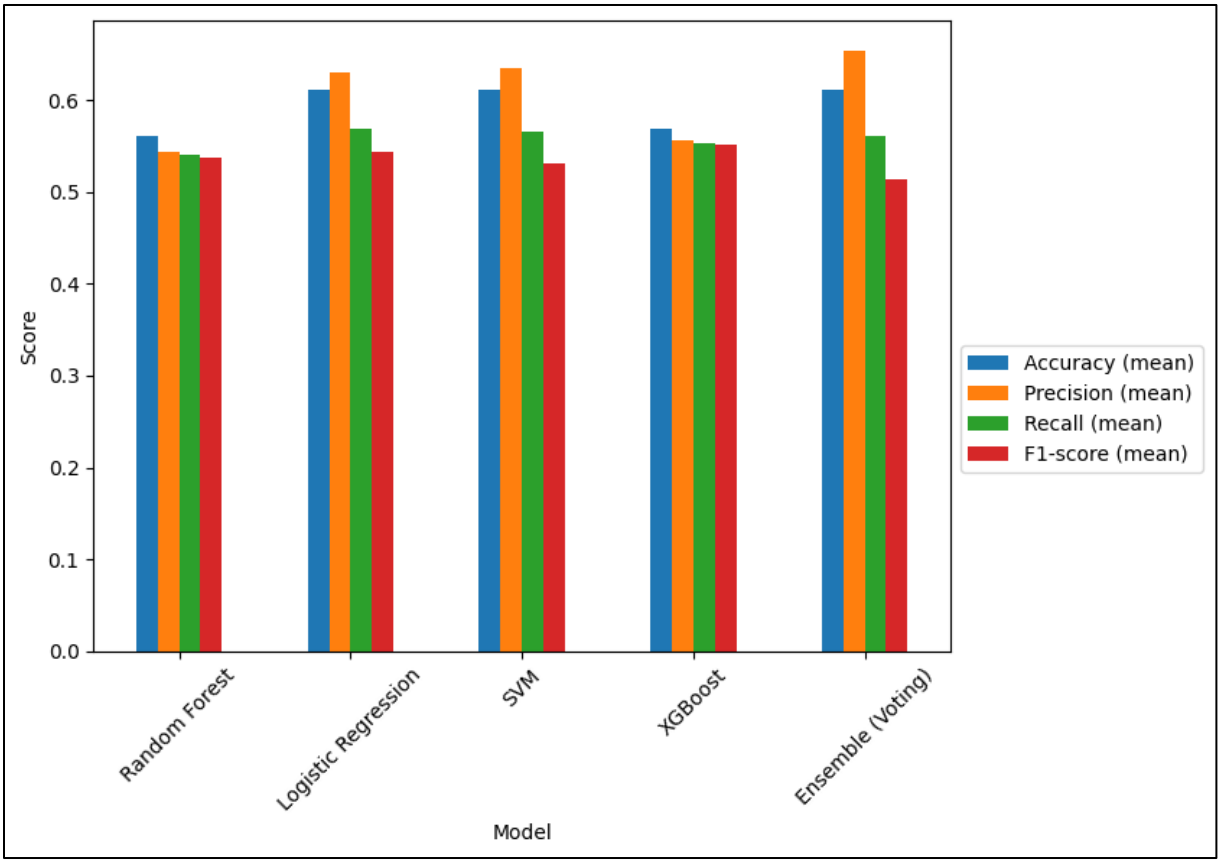

Figure 4.12: Comparison of performance metrics for different machine learning models in predicting arousal using two features selected from the 10 features based on the L1-regularized Logistic Regression feature importance method

4.2 Predicting Aggressive Acceleration

To predict aggressive acceleration, two different feature selection methods were used for comparative analysis: Random Forest and Logistic Regression with L1 regularization. The full set of 22 features was again divided into two groups to prevent redundancy. The first group included the high and base values of phasic and tonic EDA signals, along with four heart rate features. The second group consisted of six difference-based EDA features combined with the same heart rate features. The findings from both feature selection approaches, applied to each feature group, are discussed in the following sections.

4.2.1 Aggressive Acceleration Prediction Based on Random Forest Feature Importance

Like the arousal prediction, Random Forest was first used to get the feature importance for aggressive acceleration. The resulting feature importance ranking is shown in Figure 4.13. Among all features, High Phasic Max emerged as the most strongly associated with aggressive acceleration. Additionally, several heart rate features—specifically Median, Skew, and Standard Deviation (STD)—also ranked highly, indicating that physiological heart rate responses play a significant role in predicting aggressive acceleration behavior. These findings suggest that both EDA and heart rate signals contribute meaningfully to identifying aggressive driving patterns.

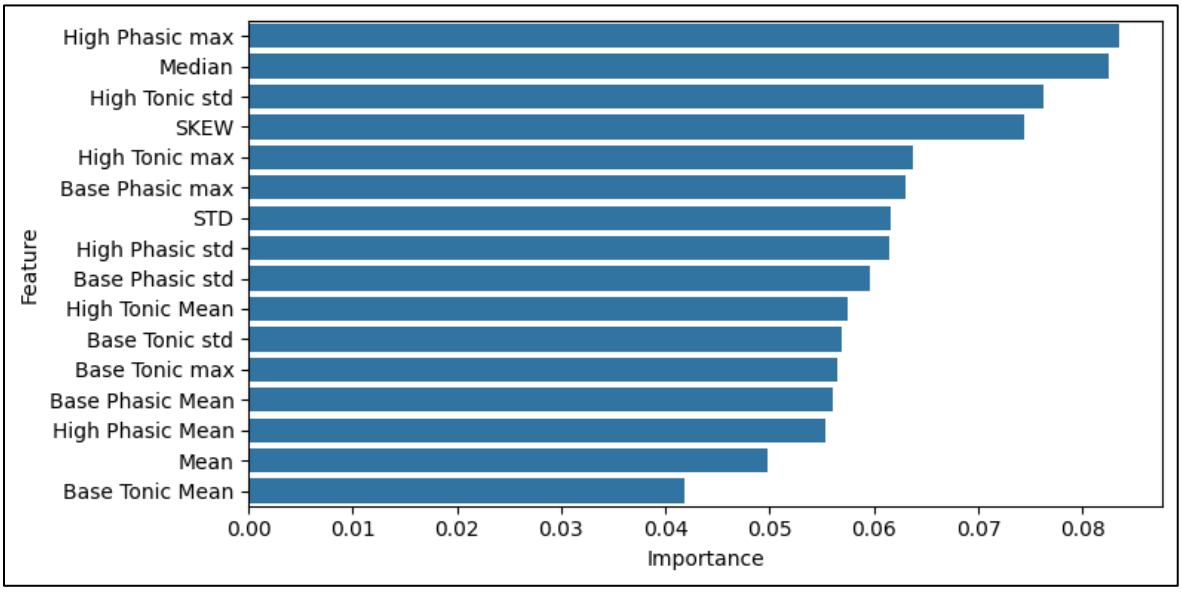

Figure 4.13: Feature importance chart for 16 features in aggressive acceleration prediction using Random Forest

Then, from this ranking of the features, to determine the optimal number of features, again, an ensemble model (VotingClassifier) was evaluated across a range of top-ranked features from 1 to 16. As shown in Figure 4.14, the ensemble model accuracy was maximum with six features. After

that, the accuracy decreased as the number of features increased. And these six features were used to evaluate the model.

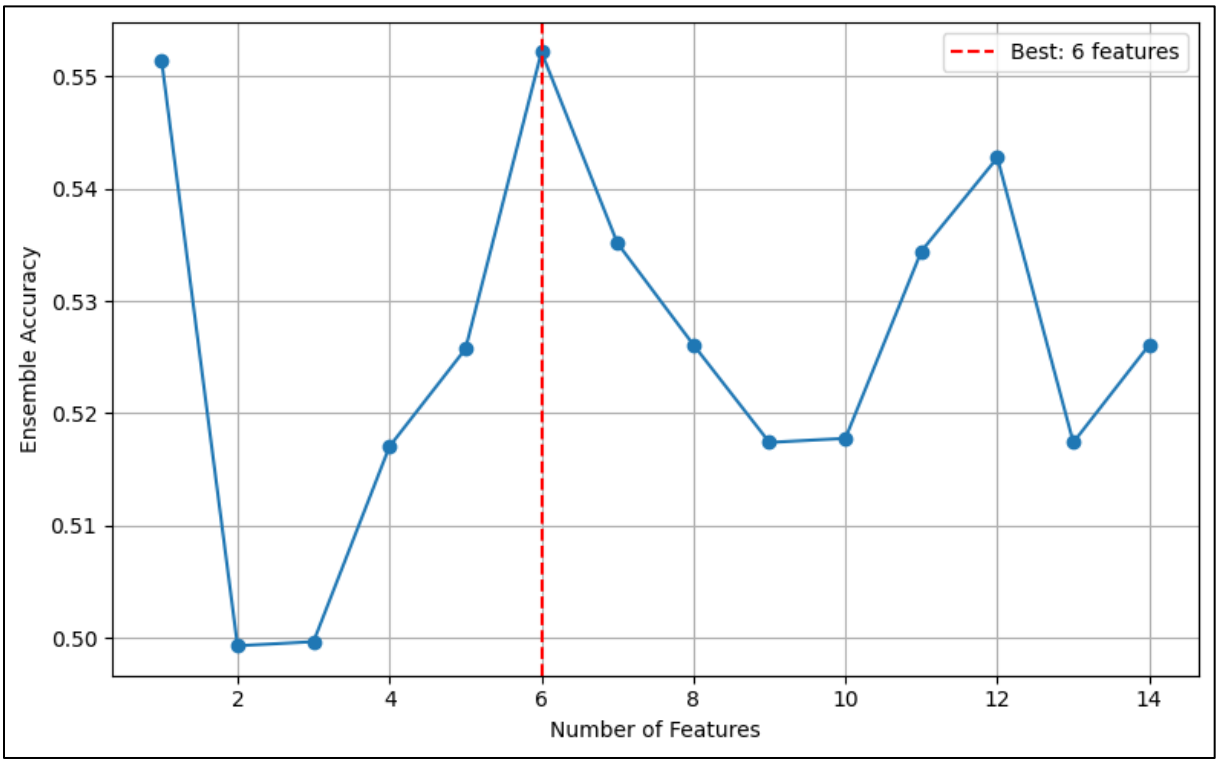

Figure 4.14: Chart showing ensemble model accuracy for aggressive acceleration prediction using different numbers of top features ranked by Random Forest feature importance from the original 16 features.

Table 4.5 presents the classification performance of various models in predicting aggressive acceleration using the top six features selected through Random Forest feature importance. Among the individual classifiers, XGBoost demonstrated the best performance, achieving the highest accuracy (0.62), precision (0.62), recall (0.59), and F1-score (0.59), indicating its strong capability in capturing the relevant patterns from the selected features.

The ensemble model, while integrating the strengths of all classifiers, did not surpass the performance of XGBoost. It achieved moderate results, with an accuracy of 0.55 and an F1-score

of 0.45. These findings suggest that while ensemble models can offer good performance, in this case, XGBoost independently provided the most reliable predictions for aggressive acceleration when using the selected features.

Table 4.5: Evaluation metrics of different machine learning models for predicting aggressive acceleration using 2 features selected from the original 16 based on Random Forest feature importance

Model	Accuracy	Precision	Recall	F1 Score
Logistic Regression	0.54	0.56	0.48	0.43
Random Forest	0.53	0.51	0.50	0.49
XGBoost	0.62	0.62	0.59	0.59
SVM	0.47	0.38	0.43	0.40
Ensemble	0.55	0.53	0.50	0.45

Figure 4.15 provides a visual comparison of the performance metrics across all models used for predicting aggressive acceleration using the top six features out of the 16 features.

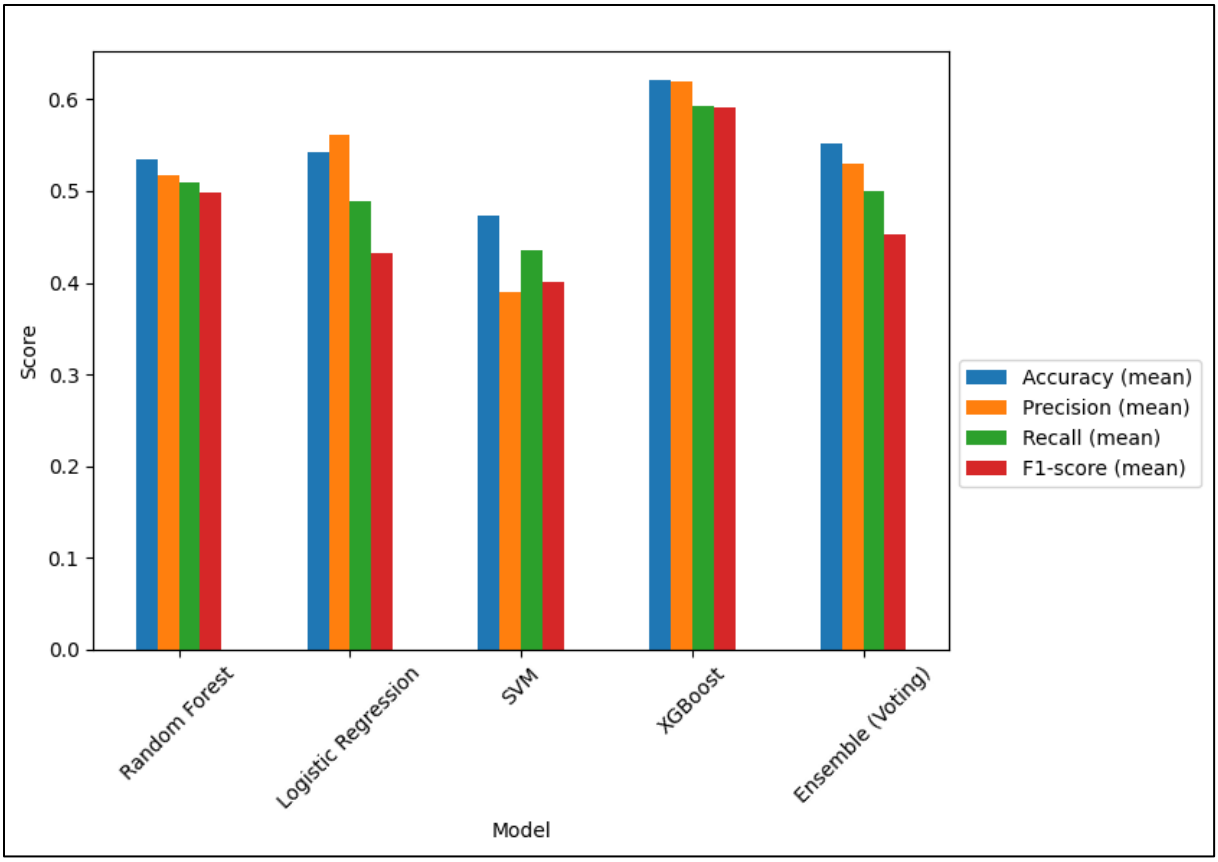

Figure 4.15: Comparison of performance metrics for different machine learning models in predicting aggressive acceleration using six features selected from the original 16 features based on the Random Forest feature importance method

Then, the second set of features, consisting of 10 variables, was used to predict aggressive acceleration. This set included six difference-based EDA features, derived from the base and high values of phasic and tonic components, combined with the same four heart rate features. Like the previous model, Random Forest was employed to compute feature importance before model training, rather than directly using all features.

Figure 4.16 illustrates the feature importance ranking for this feature set. This time, the Tonic Standard Difference emerged as the most influential feature, demonstrating the highest correlation with aggressive acceleration prediction.

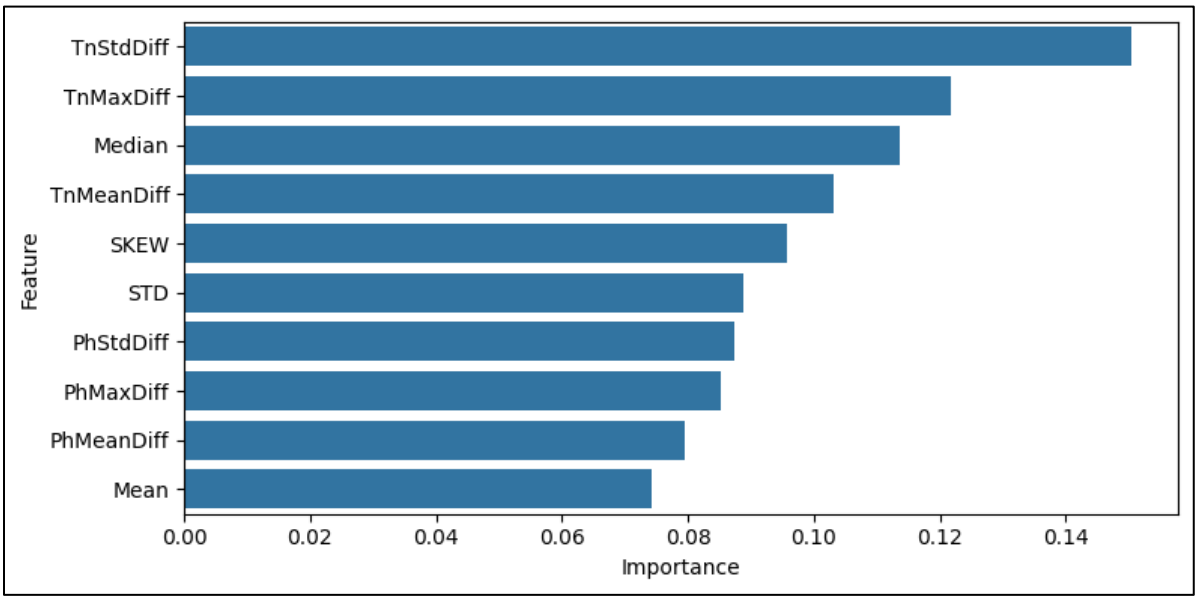

Figure 4.16: Feature importance chart for 10 features in aggressive acceleration prediction using Random Forest

From this ranking of the features, to determine the optimal number of features, again, an ensemble model (VotingClassifier) was evaluated across a range of top-ranked features from 1 to 10. As shown in Figure 4.17, ensemble model accuracy was maximum with the top 5 features. After that, the accuracy decreased as the number of features increased. An Early stopping with a patience threshold was applied to avoid overfitting. The highest ensemble accuracy was achieved with the top 5 features. And these top 5 features were used to evaluate the model.

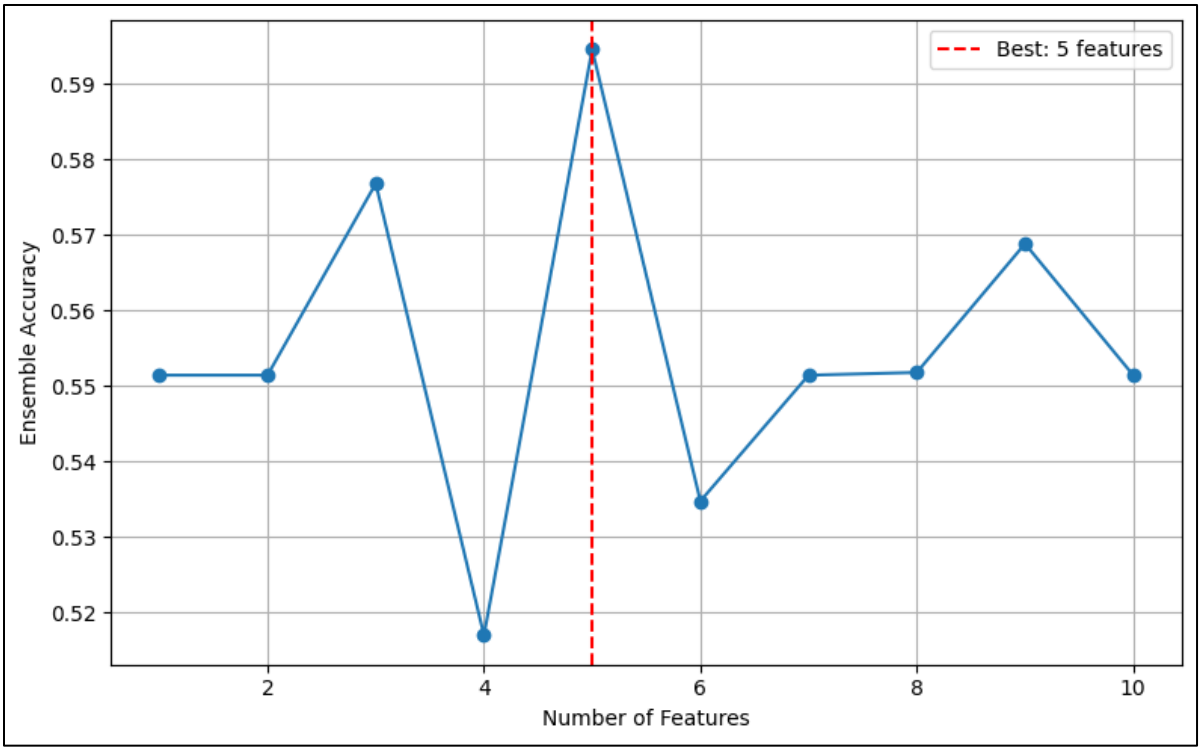

Figure 4.17: Chart showing ensemble model accuracy for aggressive acceleration prediction using different numbers of top features ranked by Random Forest Feature importance from 10 features

Table 4.6 presents the classification performance of various models in predicting aggressive acceleration using the top 5 features selected through Random Forest feature importance. Among the individual classifiers, the Random Forest model demonstrated the best performance, achieving the highest accuracy (0.60), precision (0.60), recall (0.58), and F1-score (0.59), indicating its strong capability in capturing the relevant patterns from the selected features.

The ensemble model did not surpass the performance of the Random Forest. However, it achieved moderate results, with an accuracy of 0.59 and an F1-score of 0.58, which is similar to the Random Forest model.

Table 4.6: Evaluation metrics of different machine learning models for predicting aggressive acceleration using 5 features selected from the 10 features based on Random Forest feature importance.

Model	Accuracy	Precision	Recall	F1 Score
Logistic Regression	0.53	0.28	0.46	0.34
Random Forest	0.60	0.60	0.58	0.59
XGBoost	0.58	0.56	0.56	0.55
SVM	0.57	0.56	0.55	0.55
Ensemble	0.59	0.61	0.55	0.52

Figure 4.18 provides a visual comparison of the performance metrics across all models used for predicting aggressive acceleration using the top five features out of the 10 features.

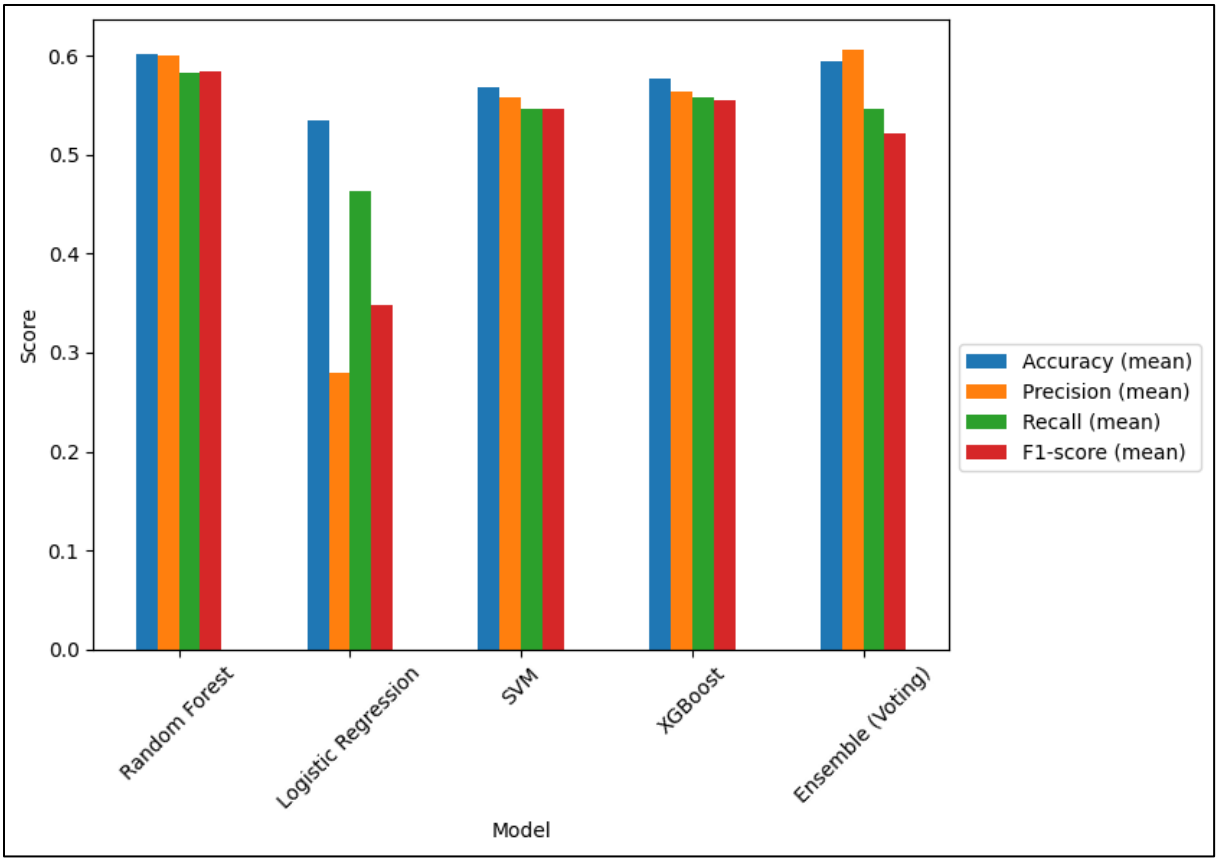

Figure 4.18: Comparison of performance metrics for different machine learning models in predicting aggressive acceleration using five features selected from the 10 features based on the Random Forest feature importance method

4.2.2 Aggressive Acceleration Prediction Based on L1- Regularized Logistic Regression Feature Importance

In this approach, Logistic Regression with L1 regularization was used to perform feature selection for predicting aggressive acceleration. Figure 4.19 illustrates the feature importance ranking obtained from applying this method to the full set of 16 features.

Figure 4.19 shows the feature importance ranking for this feature set. This time, the Base phasic standard deviation emerged as the most influential feature, demonstrating the highest correlation with aggressive acceleration prediction.

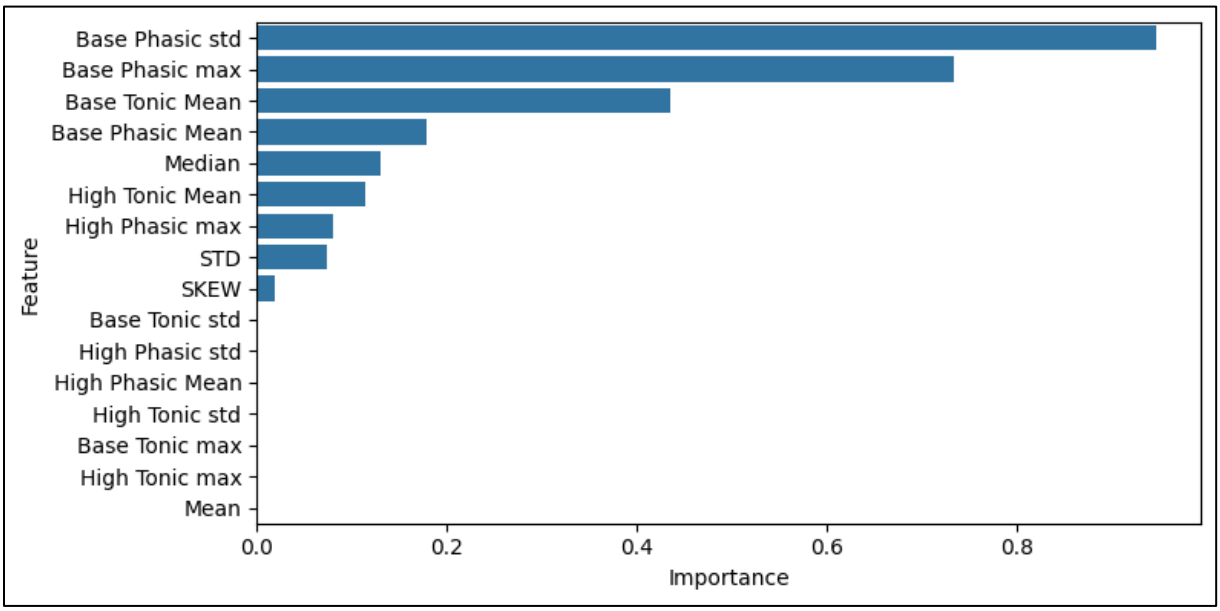

Figure 4.19: Feature importance chart for 16 features using the logistic regression feature importance method for aggressive acceleration prediction

Figure 4.20 illustrates the relationship between the number of top-ranked features and the ensemble model's accuracy when predicting aggressive acceleration using features selected through Logistic Regression with L1 regularization.

As shown in the figure, the ensemble model achieved its highest accuracy when the top three features were used. After this point, accuracy decreased as additional features were included.

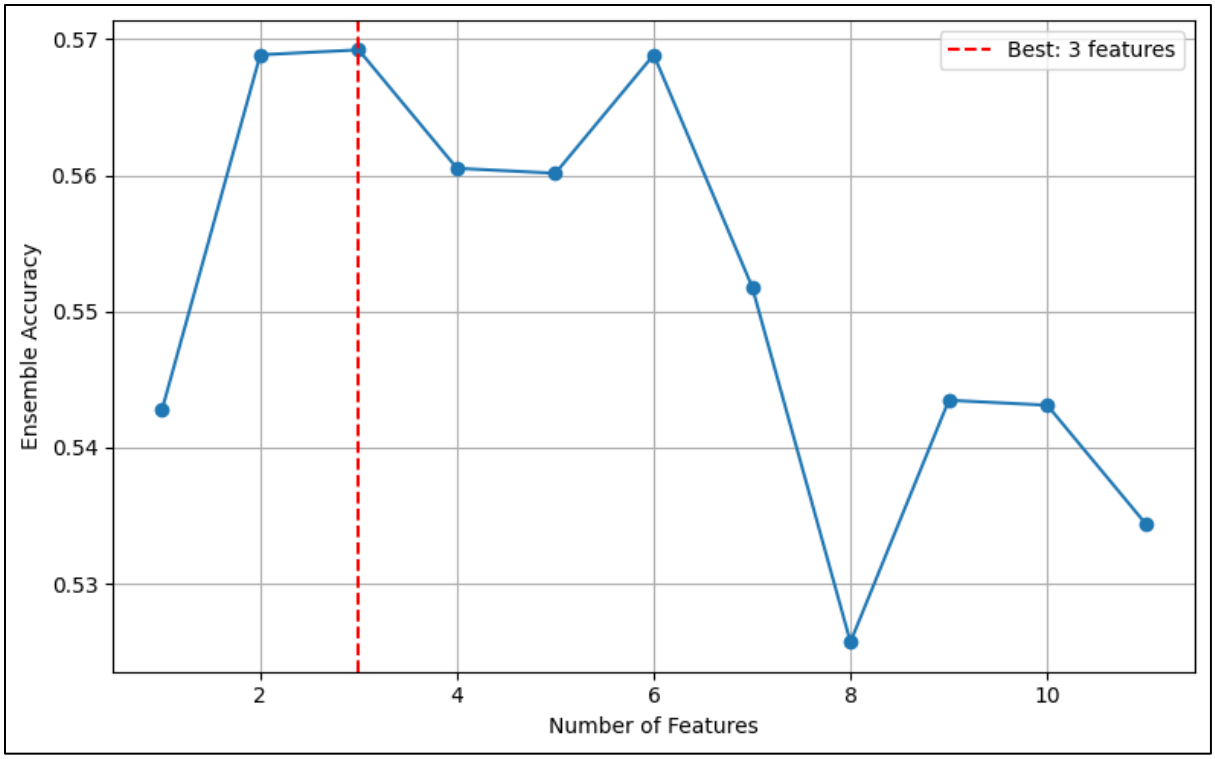

Figure 4.20: Chart showing ensemble model accuracy for aggressive acceleration prediction using different numbers of top features ranked by L1-regularized Logistic Regression feature importance from the original 16 features.

Table 4.7 summarizes the performance of various models in predicting aggressive acceleration using the top three features identified through Logistic Regression with L1 regularization.

Here, among the individual models, logistic regression and SVM achieved the highest accuracy of 59%. Although the ensemble model, which combined the strengths of all base classifiers, achieved an accuracy of 0.57. Although it did not surpass the best-performing individual models in this case, the ensemble still maintained competitive precision (0.61).

Table 4.7: Evaluation metrics of different machine learning models for predicting aggressive acceleration using 3 features selected from the original 16 based on L1-regularized Logistic Regression feature importance

Model	Accuracy	Precision	Recall	F1 Score
Logistic Regression	0.59	0.69	0.53	0.49
Random Forest	0.53	0.51	0.51	0.51
XGBoost	0.54	0.53	0.53	0.52
SVM	0.59	0.61	0.52	0.45
Ensemble	0.57	0.61	0.51	0.44

Figure 4.21 provides a visual comparison of the performance metrics across all models used for predicting aggressive acceleration using the top 3 features out of the 16 features.

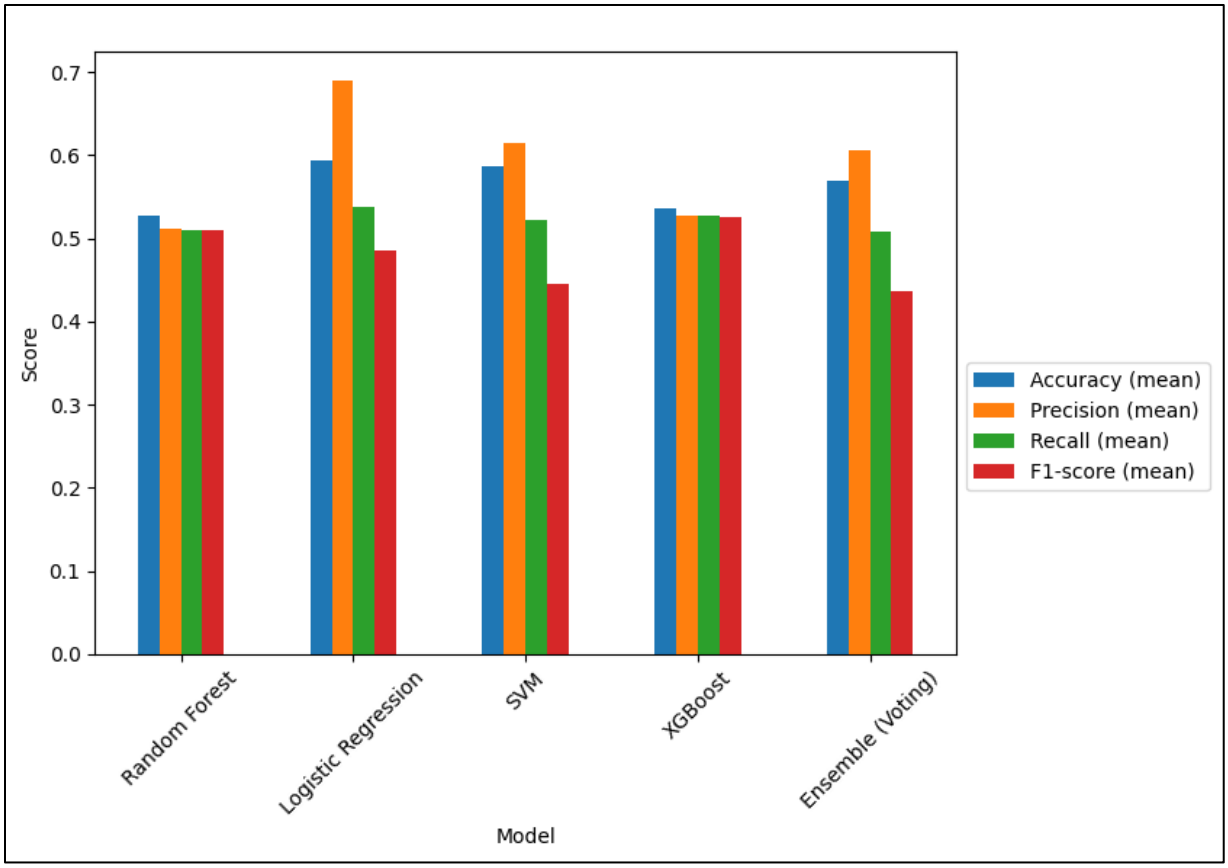

Figure 4.21: Comparison of performance metrics for different machine learning models in predicting aggressive acceleration using three features selected from the original 16 features based on the L1-regularized Logistic Regression feature importance method.

Logistic Regression with L1 regularization was used to evaluate feature importance for predicting aggressive acceleration based on the difference-based feature set and heart rate variables. Figure 4.22 presents the resulting feature importance ranking with set B dataset which includes ten features.

The phasic standard deviation difference has emerged as the most significant feature in determining the aggressive acceleration.

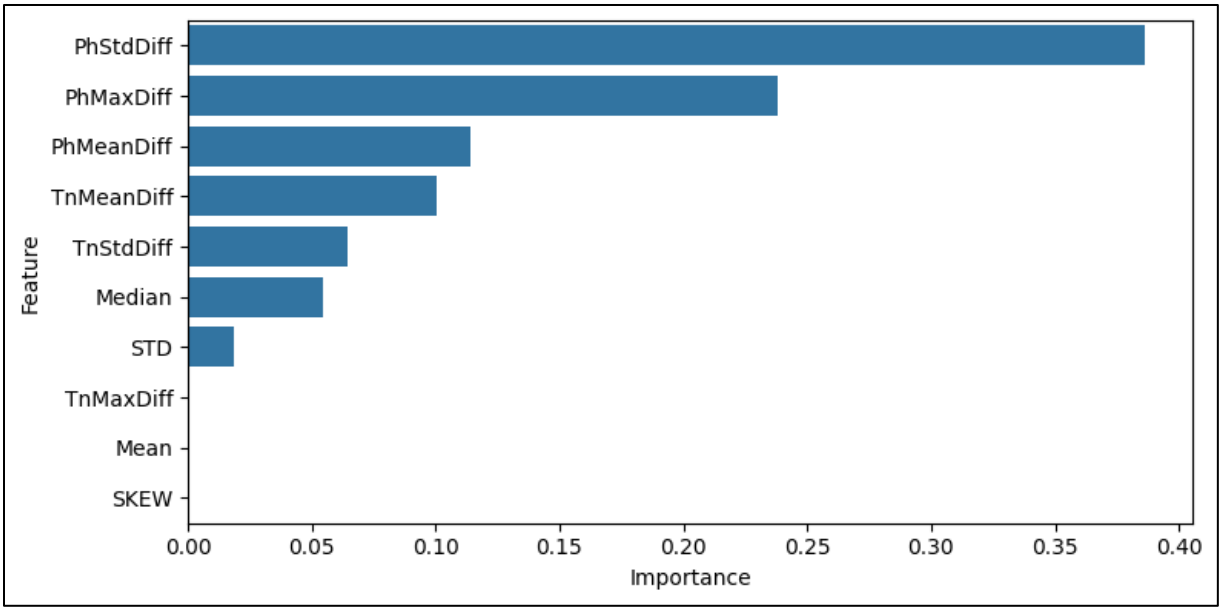

Figure 4.22: Feature importance chart for 10 features in aggressive acceleration prediction using L1-regularized Logistic Regression

Figure 4.23 shows that the ensemble model achieved its highest accuracy when the top three features were used. After this point, accuracy decreased as additional features were included.

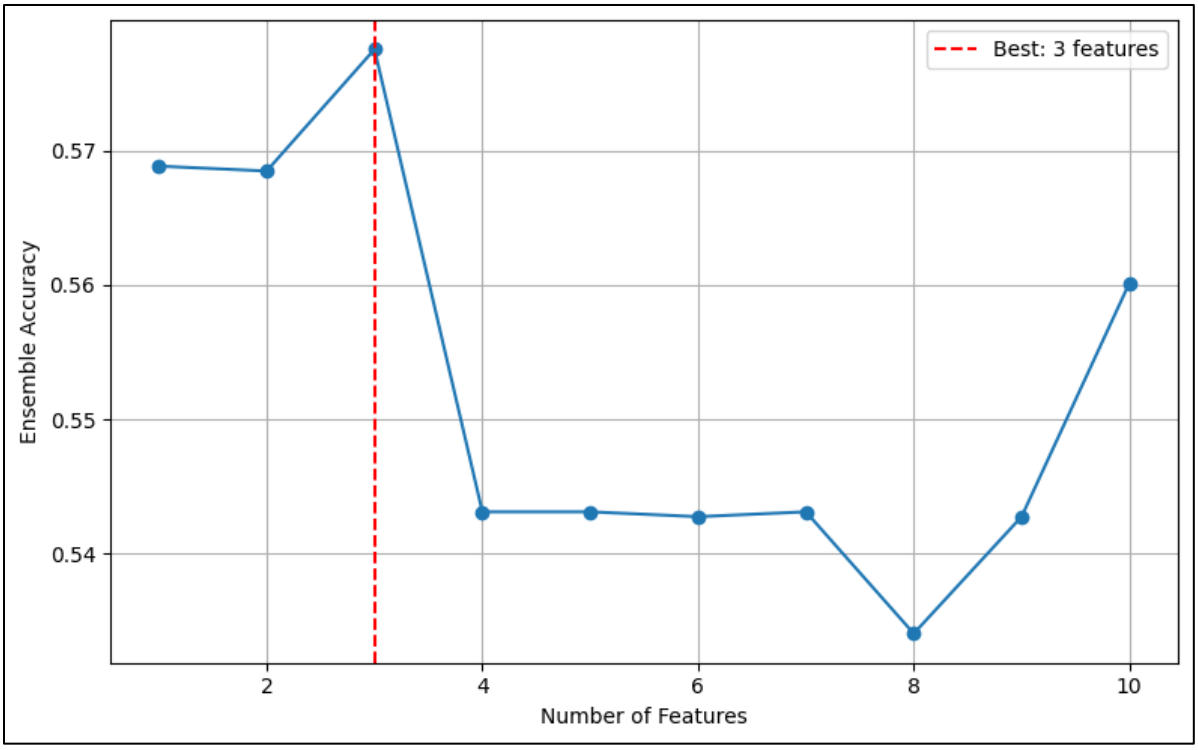

Figure 4.23: Chart showing ensemble model accuracy for aggressive acceleration prediction using different numbers of top features ranked by L1-regularized Logistic Regression feature importance from the 10 features.

Table 4.8 summarizes the classification performances of different models. Among the individual models, the SVM model had the highest accuracy, and the ensemble model also reached the accuracy of the SVM model. However, both models showed relatively low precision values, suggesting a tendency toward false positive predictions despite achieving balanced recall (0.50).

Table 4.8: Evaluation metrics of different machine learning models for predicting aggressive acceleration using 3 features selected from the 10 features based on L1-regularized Logistic Regression feature importance.

Model	Accuracy	Precision	Recall	F1 Score

Logistic Regression	0.55	0.38	0.48	0.37
Random Forest	0.38	0.36	0.36	0.36
XGBoost	0.44	0.42	0.43	0.42
SVM	0.58	0.28	0.50	0.37
Ensemble	0.58	0.29	0.50	0.37

Figure 4.24 provides a visual comparison of the performance metrics across all models used for predicting aggressive acceleration using the top 3 features out of the 10 features**.**

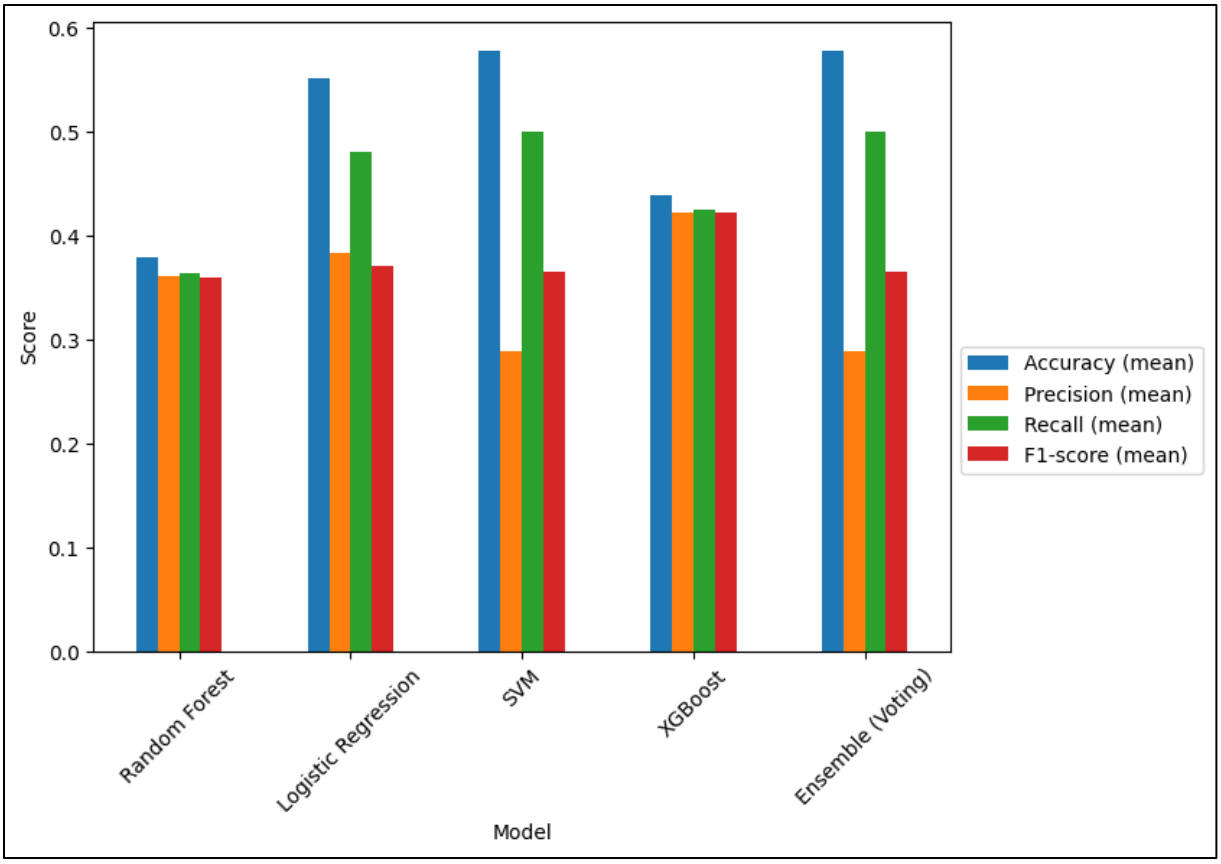

Figure 4.24: Comparison of performance metrics for different machine learning models in predicting aggressive acceleration using three features selected from the 10 features based on the L1-regularized Logistic Regression feature importance method.

4.3 Predicting Aggressive Braking

Similar to the approach used for aggressive acceleration, the prediction of aggressive braking followed the same methodology. Two different feature selection methods—Random Forest and Logistic Regression with L1 regularization—were used for comparative analysis. The complete set of 22 features was again divided into two groups to avoid redundancy. The first group included the high and base values of phasic and tonic EDA signals along with four heart rate features, while the second group consisted of six difference-based EDA features combined with the same heart

rate features. The results from both feature selection methods, applied to each feature group, are presented in the following sections.

4.3.1 Aggressive Braking Prediction Based on Random Forest Feature Importance

Before predicting arousal on all 16 features, random forest feature selection was employed to find out the relationship between the features and aggressive braking. As shown in Figure 4.25, the heart rate features demonstrated a strong relation with aggressive braking. In particular, Heart Rate Standard Deviation (STD) and Heart Rate Mean (Mean) were identified as the two most important features in the prediction.

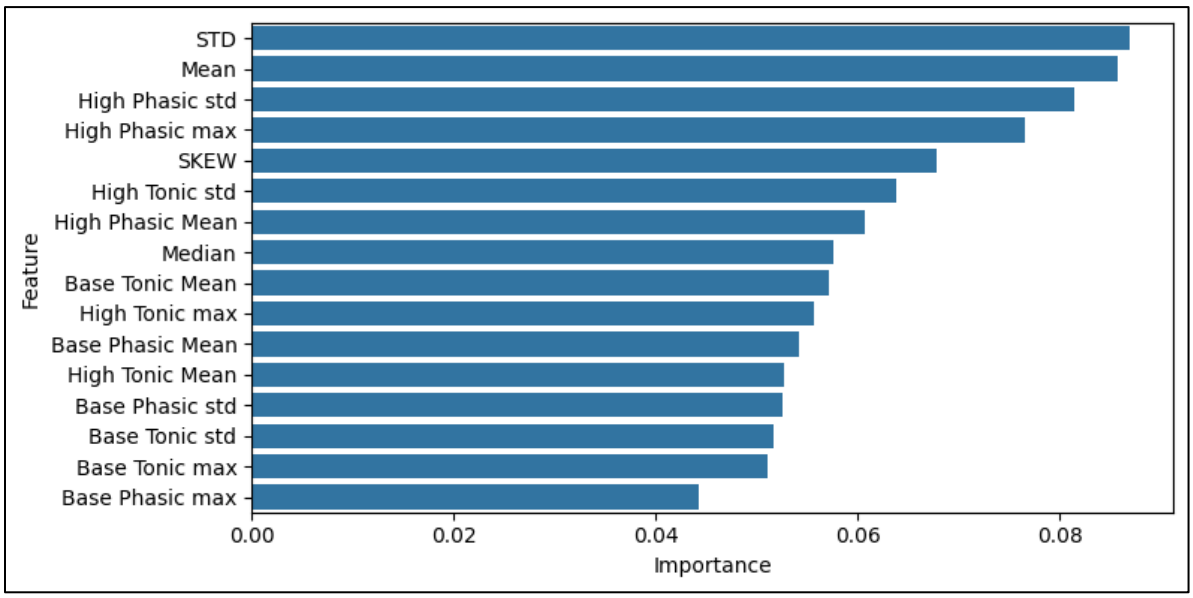

Figure 4.25: Feature importance chart for 16 initial features in aggressive braking prediction using Random Forest

Figure 4.26 shows that the ensemble model achieved its highest accuracy using only the top feature and maintained this maximum accuracy up to the addition of the top five features. After this point, accuracy began to decrease as more features were added.

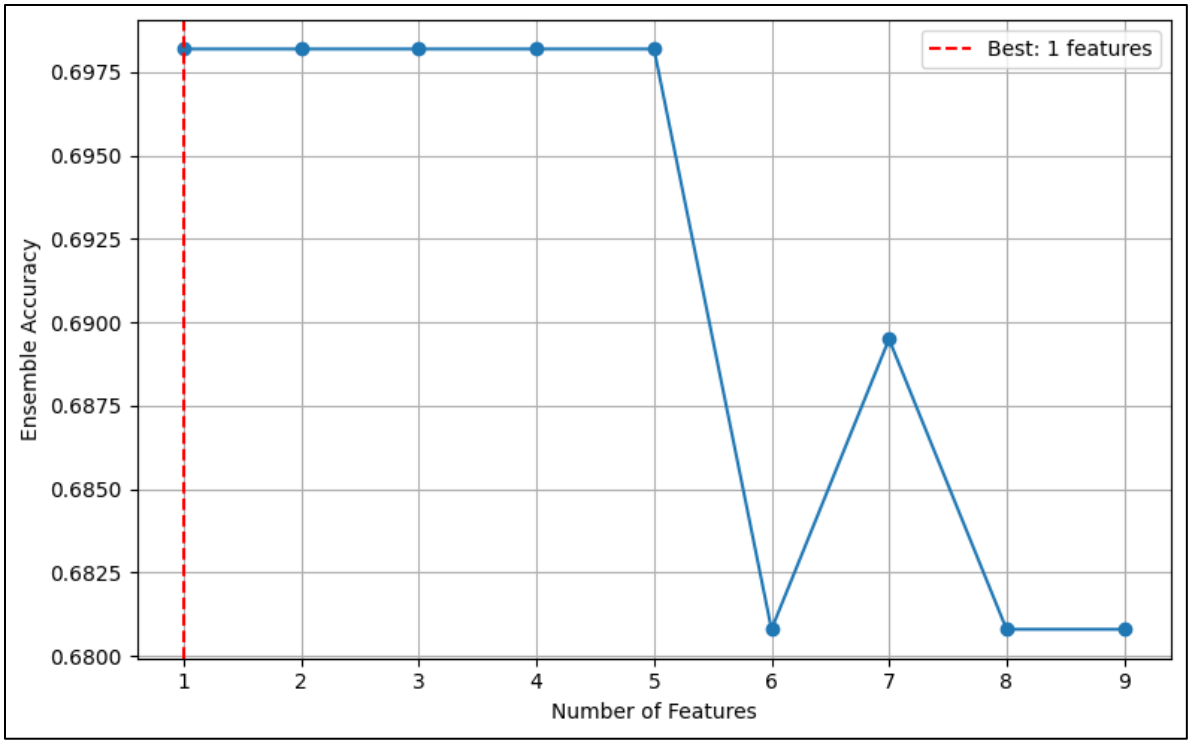

Figure 4.26: Chart showing ensemble model accuracy for aggressive braking prediction using different numbers of top features ranked by Random Forest feature importance from the 16 features.

Table 4.9 summarizes the model performance for predicting aggressive braking using the single most important feature identified through Random Forest feature importance. Among the individual models, SVM and Logistic Regression achieved the highest accuracy (0.70 and 0.70, respectively), with the ensemble model matching the performance of SVM at 0.70 accuracy.

While accuracy was relatively high, precision remained low across all models, ranging from 0.34 to 0.53, indicating a tendency toward false positive predictions. Recall values were consistent at 0.50 across all models, suggesting that half of the actual aggressive braking events were correctly identified.

Table 4.9: Evaluation metrics of different machine learning models for predicting aggressive braking using 1 feature selected from the original 16 based on Random Forest feature importance.

Model	Accuracy	Precision	Recall	F1 Score
Logistic Regression	0.70	0.34	0.5	0.41
Random Forest	0.58	0.50	0.51	0.50
XGBoost	0.59	0.53	0.50	0.53
SVM	0.70	0.35	0.5	0.41
Ensemble	0.70	0.35	0.5	0.41

Figure 4.27 provides a visual comparison of the performance metrics across all models used for predicting aggressive acceleration using the top 1 feature out of the 16 features.

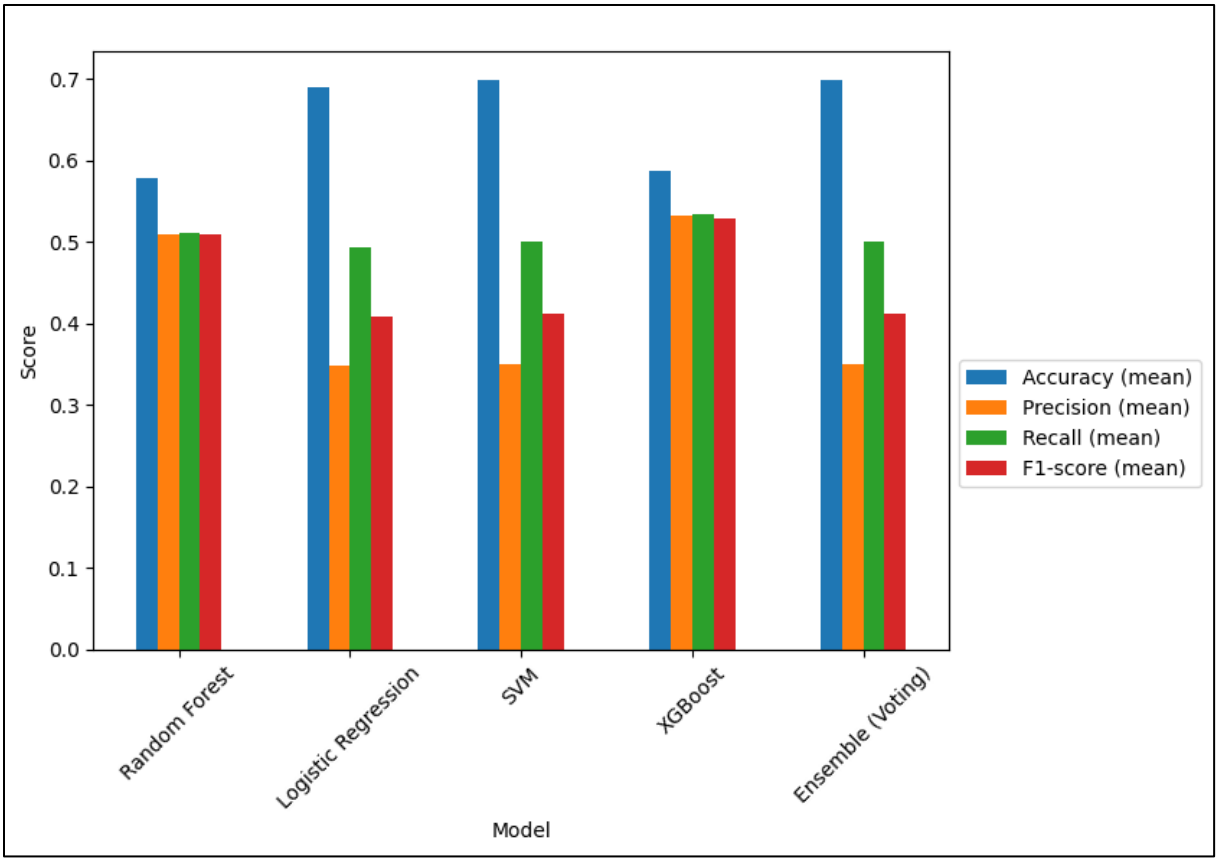

Figure 4.27: Comparison of performance metrics for different machine learning models in predicting aggressive braking using five features selected from the original 16 features based on the Random Forest feature importance method

Then, the second set of features, consisting of 10 variables, was used to predict aggressive Braking. This set included six difference-based EDA features, derived from the base and high values of phasic and tonic components, combined with the same four heart rate features. Like the previous model, Random Forest was employed to compute feature importance before model training, rather than directly using all features.

Figure 4.28 illustrates the feature importance ranking for this feature set. Again, a heart rate feature, Mean Heart Rate (Mean), emerged as the most influential feature, demonstrating the highest correlation with aggressive braking prediction.

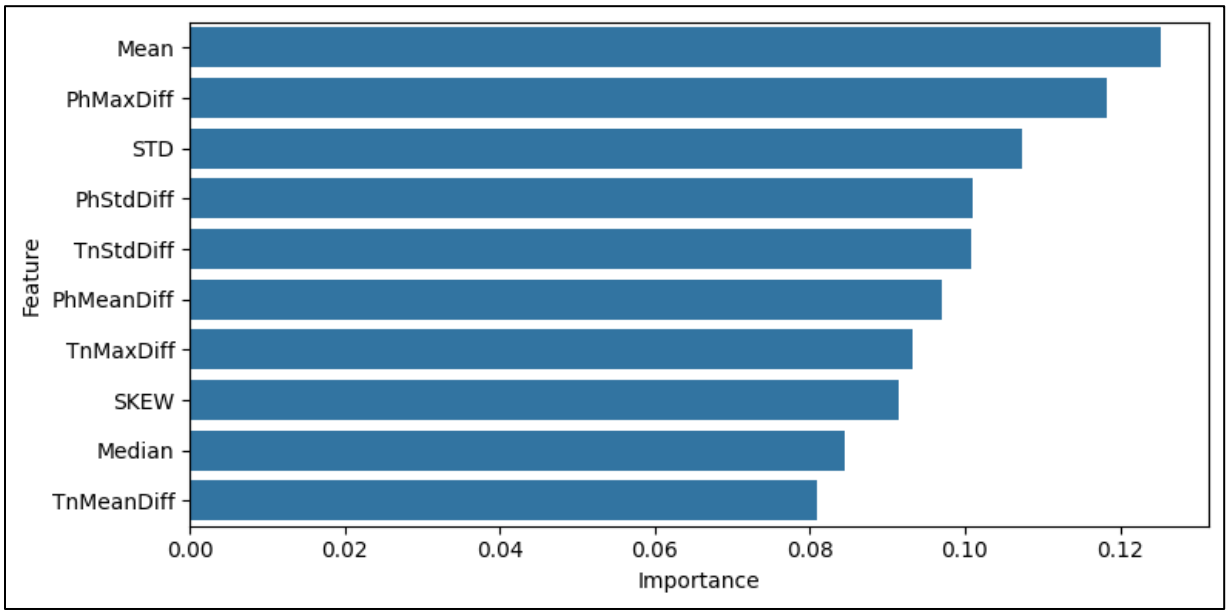

Figure 4.28: Feature importance chart for 10 features in aggressive braking prediction using Random Forest

Figure 4.29 shows that the ensemble model achieved its highest accuracy when the top two features were used. After this point, accuracy decreased as additional features were included.

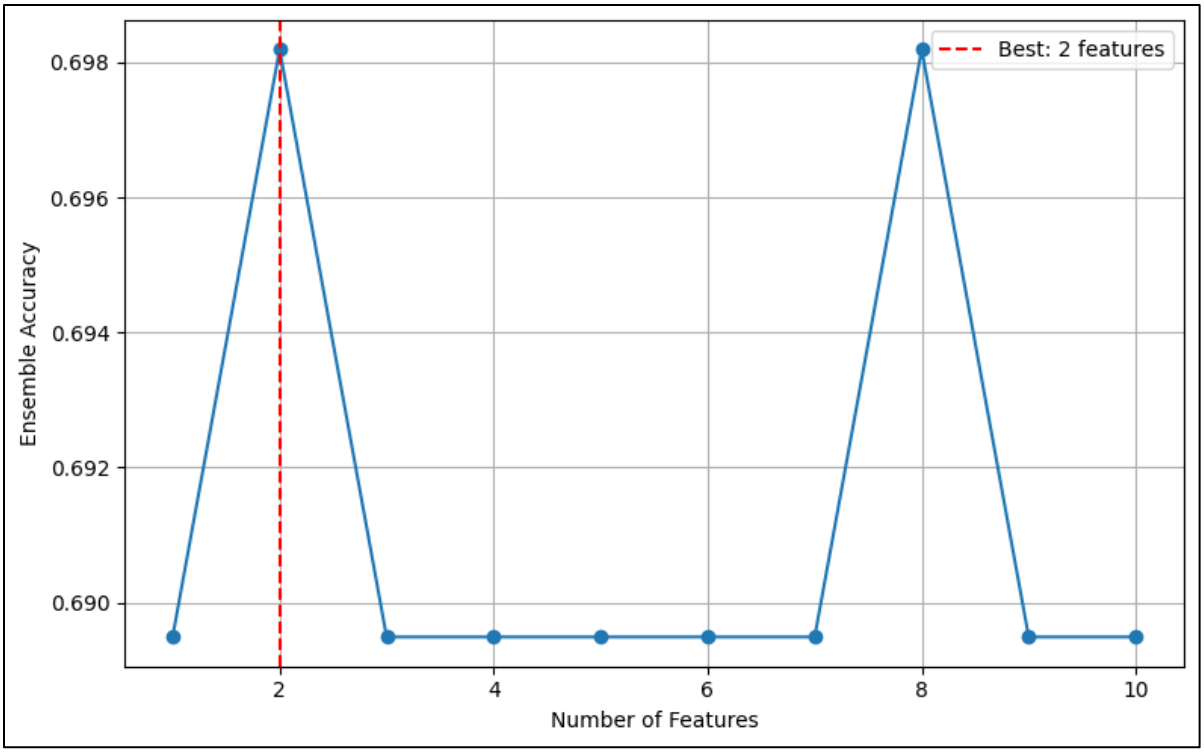

Figure 4.29: Chart showing ensemble model accuracy for aggressive braking prediction using different numbers of top features ranked by Random Forest feature importance from the 10 features

Table 4.10 summarizes the model performance for predicting aggressive braking using the top two most important features identified through Random Forest feature importance. Among the individual models, SVM and Logistic Regression achieved the highest accuracy (0.70, 0.70, respectively), with the ensemble model matching the performance of SVM and logistic regression at 0.70 accuracy.

While accuracy was relatively high, precision remained low across all models, ranging from 0.35 to 0.55, indicating a tendency toward false positive predictions. Recall values were consistent around 0.50 across all models, suggesting that half of the actual aggressive braking events were correctly identified.

Table 4.10: Evaluation metrics of different machine learning models for predicting aggressive braking using 2 features selected from the 10 features based on Random Forest feature importance.

Model	Accuracy	Precision	Recall	F1 Score
Logistic Regression	0.70	0.35	0.49	0.41
Random Forest	0.60	0.48	0.48	0.47
XGBoost	0.54	0.55	0.53	0.53
SVM	0.70	0.35	0.50	0.41
Ensemble	0.70	0.35	0.50	0.41

Figure 4.30 provides a visual comparison of the performance metrics across all models used for predicting aggressive braking using the top 2 features out of the 10 features.

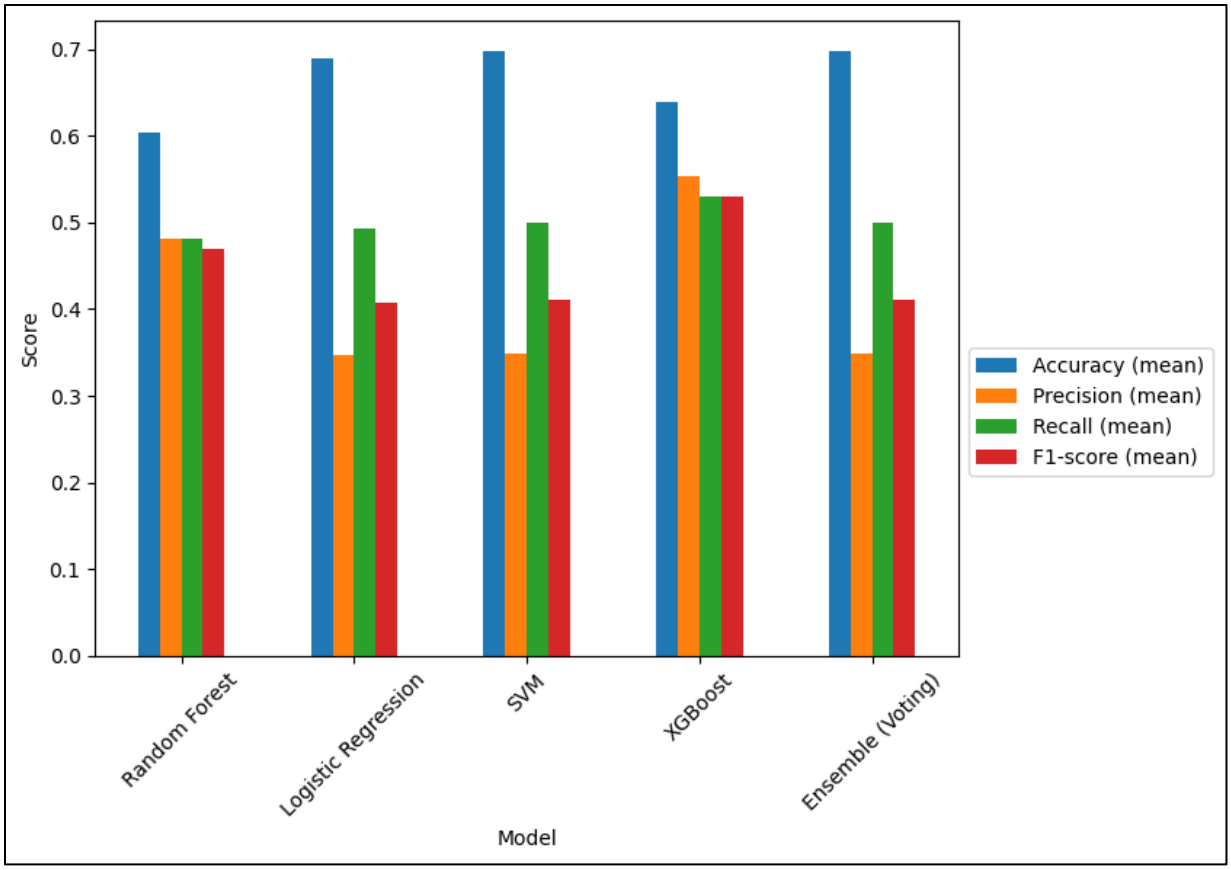

Figure 4.30: Comparison of performance metrics for different machine learning models in predicting aggressive braking using two features selected from the 10 features based on the Random Forest feature importance method

4.3.2 Aggressive Braking Prediction Based on L1- Regularized Logistic Regression Feature Importance

Finally, to evaluate the relationship between input features and aggressive braking, Logistic Regression with L1 regularization (Lasso) was applied to the full set of 16 features. The resulting feature importance rankings are illustrated in Figure 4.31.

As shown, Base Tonic Max and High Tonic Max were identified as the most influential features, suggesting that tonic EDA components are highly relevant in predicting aggressive braking behavior.

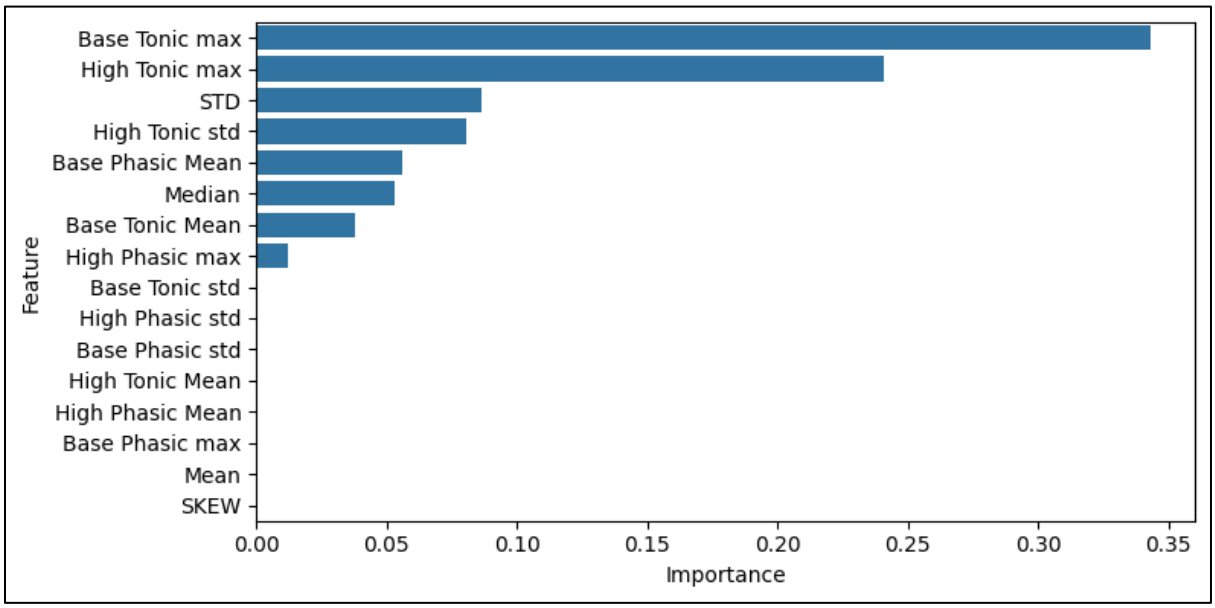

Figure 4.31: Feature importance chart for 16 initial features in aggressive braking prediction using L1-Regularized Logistic Regression

Figure 4.32 shows that the ensemble model achieved its highest accuracy using only the top feature and maintained this maximum accuracy up to the addition of the top 1 feature. It remained same with the top two features too. After this point, the accuracy began to decrease as more features were added.

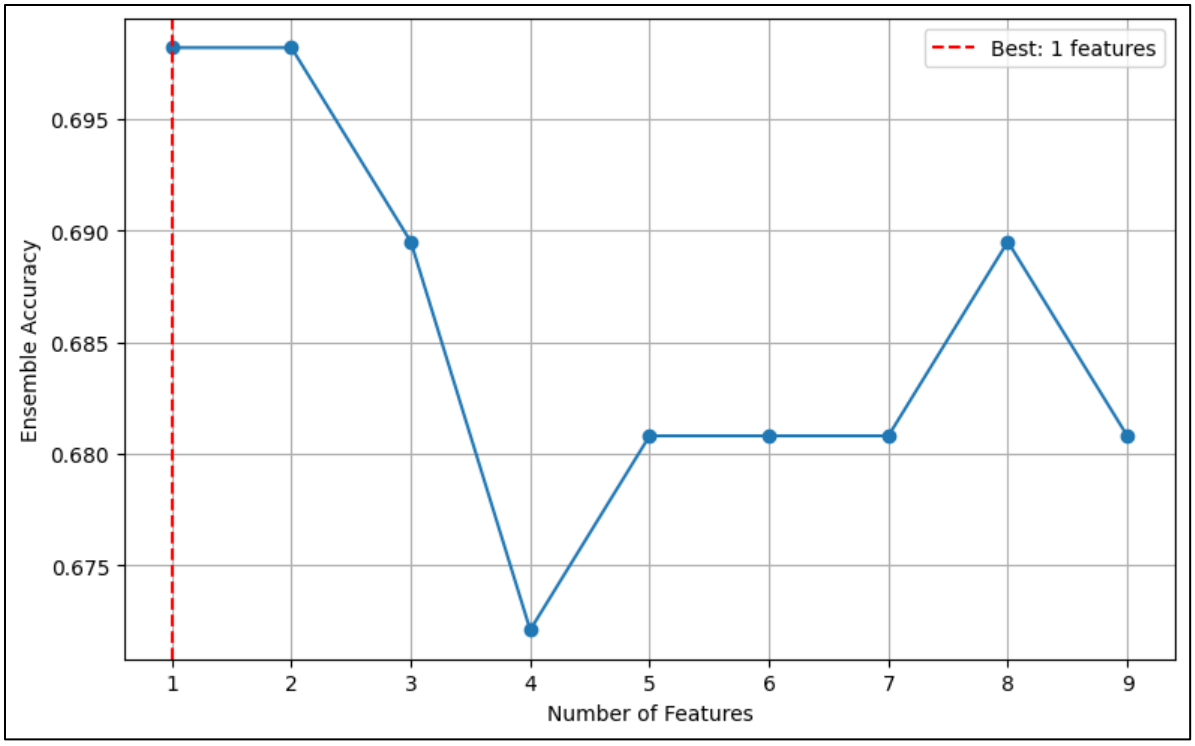

Figure 4.32: Chart showing ensemble model accuracy for aggressive braking prediction using different numbers of top features ranked by L1-Regularized Logistic Regression feature importance from the original 16 features

Table 4.11 presents the performance of various models in predicting aggressive braking using the top most feature ranked by L1-regularized Logistic Regression. Among the models, Logistic Regression, SVM, and the Ensemble classifier each achieved the highest accuracy of 0.70, with identical precision (0.35), recall (0.50), and F1-score (0.41).

Table 4.11: Evaluation metrics of different machine learning models for predicting aggressive braking using 1 feature selected from the original 16 based on L1-regularized Logistic Regression feature importance.

Model	Accuracy	Precision	Recall	F1 Score

Logistic Regression	0.70	0.35	0.50	0.41
Random Forest	0.57	0.48	0.47	0.46
XGBoost	0.52	0.42	0.42	0.41
SVM	0.70	0.35	0.50	0.41
Ensemble	0.70	0.35	0.50	0.41

Figure 4.33 provides a visual comparison of the performance metrics across all models used for predicting aggressive braking using the top 1 feature out of the 16 features.

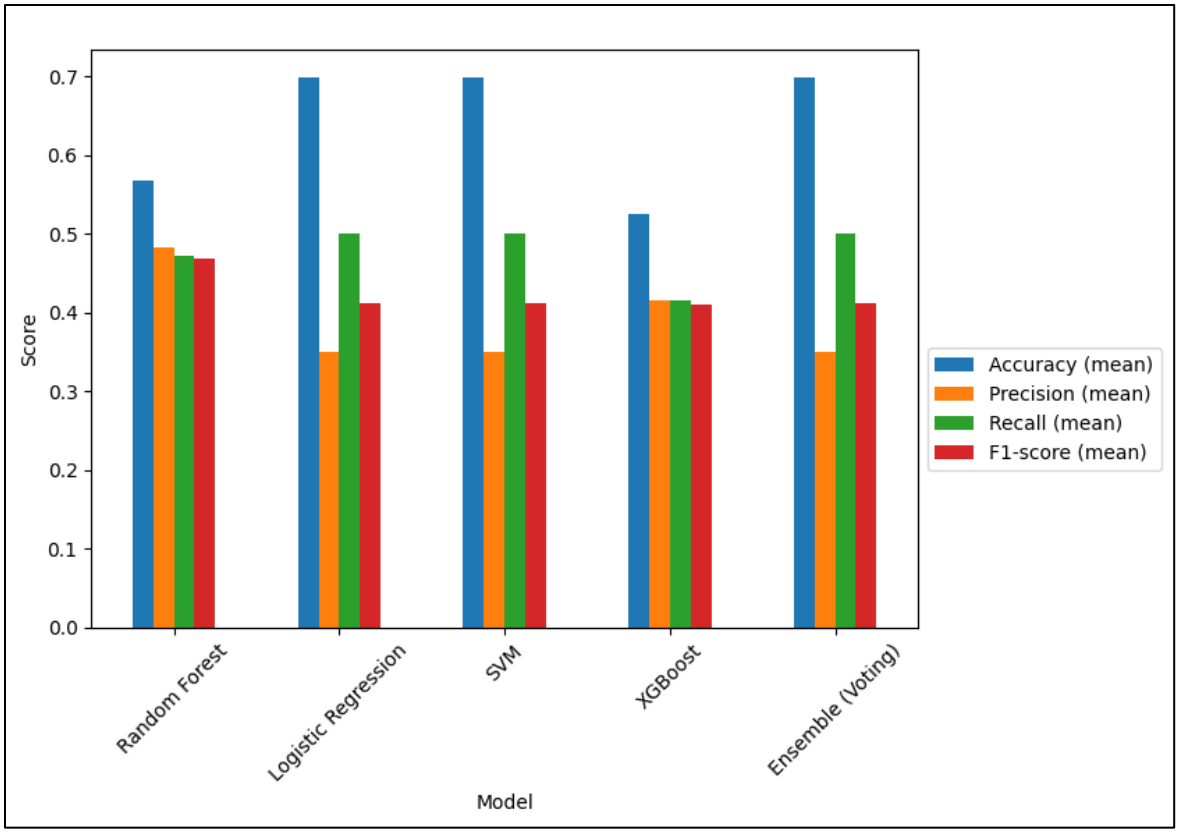

Figure 4.33: Comparison of performance metrics for the different machine learning models in predicting aggressive braking using the top one feature selected from the original 16 features based on the L1-regularized Logistic Regression feature importance method

Then, the second set of features, consisting of 10 variables, was used to predict aggressive Braking. This set included six difference-based EDA features, derived from the base and high values of phasic and tonic components, combined with the same four heart rate features. Like the previous model, logistic regression was employed to compute feature importance before model training, rather than directly using all features.

This time, the tonic max difference has emerged as the most important feature in predicting aggressive braking.

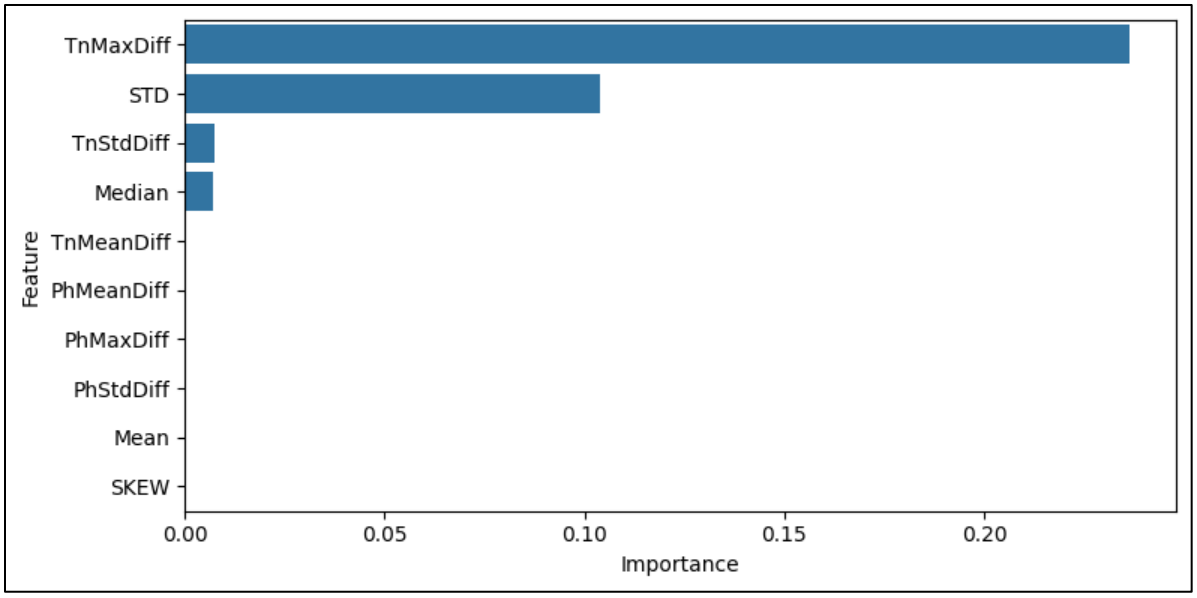

Figure 4.34: Feature importance chart for 10 features in aggressive braking prediction using L1-regularized Logistic Regression

Figure 4.35 shows that the ensemble model achieved its highest accuracy when the tonic max difference features were used. After this point, accuracy decreased as additional features were included.

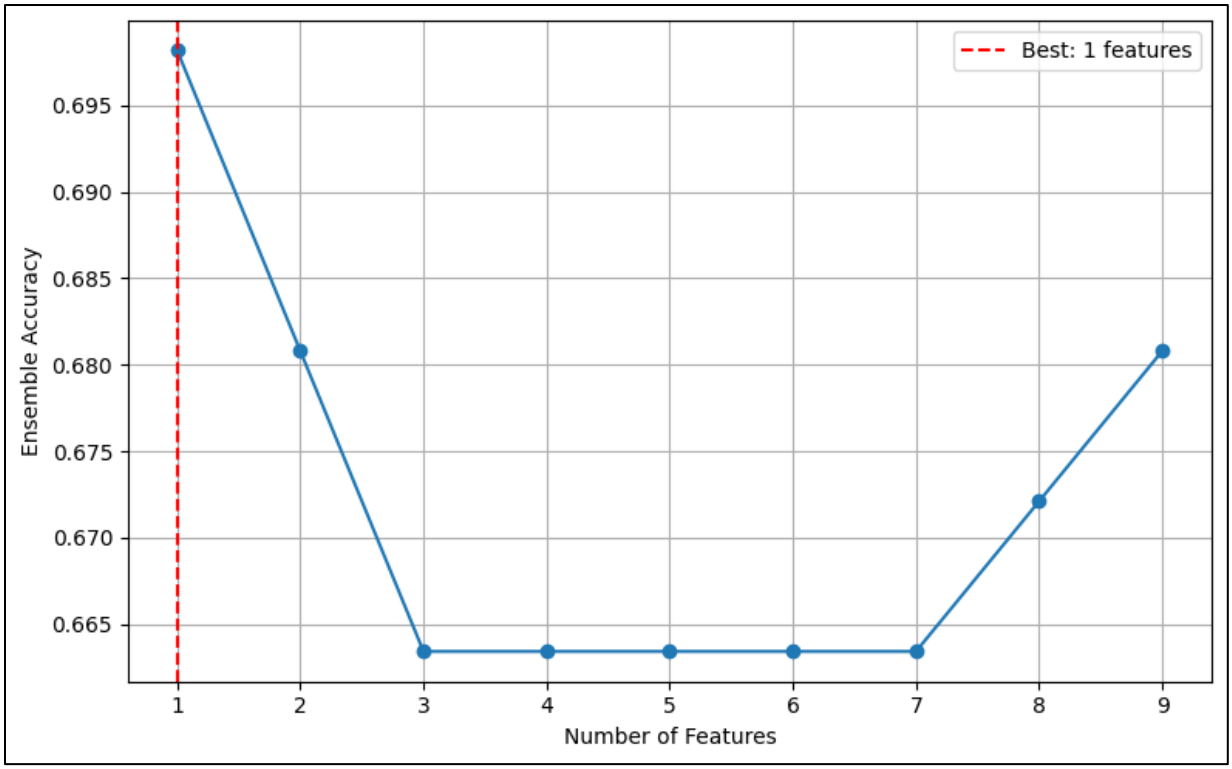

Figure 4.35: Chart showing ensemble model accuracy for aggressive braking prediction using different numbers of top features ranked by L1-regularized Logistic Regression feature importance from the 10 features.

Table 4.10 presents the performance of various models in predicting aggressive braking using the topmost feature ranked by Logistic Regression with L1 regularization. Among the models, Logistic Regression, SVM, and the Ensemble classifier each achieved the highest accuracy of 0.70, with identical recall (0.50), and F1-score (0.41). This suggests a consistent level of predictive performance across these models when using the full feature set.

Table 4.12: Evaluation metrics of different machine learning models for predicting aggressive braking using 1 feature selected from the 10 features based on L1-regularized Logistic Regression feature importance.

Model	Accuracy	Precision	Recall	F1 Score
Logistic Regression	0.70	0.34	0.50	0.41
Random Forest	0.56	0.46	0.46	0.46
XGBoost	0.53	0.43	0.45	0.44
SVM	0.70	0.35	0.50	0.41
Ensemble	0.70	0.34	0.50	0.41

Figure 4.36 provides a visual comparison of the performance metrics across all models used for predicting aggressive braking using the top 1 feature out of the 10 features.

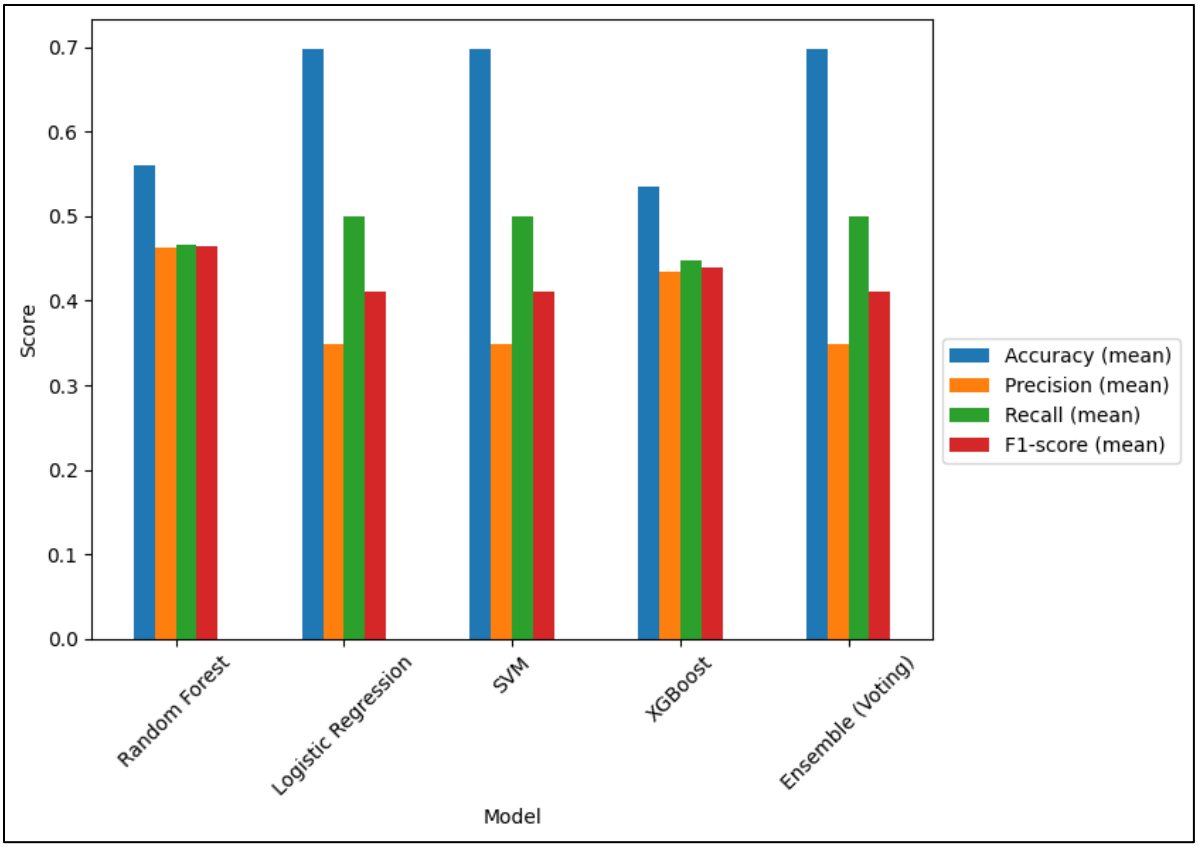

Figure 4.36: Comparison of performance metrics for different machine learning models in predicting aggressive braking using the top one feature selected from the 10 features based on L1-regularized Logistic Regression feature importance method

Chapter 5 : Discussion and Conclusion

The results indicate that SVM and ensemble classifiers were most effective for predicting emotional arousal. With just two selected features (tonic/phasic EDA and HR statistics), the SVM and a majority-vote ensemble achieved the highest accuracy (~0.64–0.65). Logistic regression and Random Forest were somewhat lower (~0.58–0.61). This aligns closely with the results reported by Nahid et al. (2024), who also used EDA signals collected in naturalistic driving settings. Their study found that neural networks could predict arousal with an accuracy of approximately 65% using a combination of tonic and phasic EDA features. Both studies highlight the potential of EDA features in driver state monitoring, while also acknowledging the inherent limitations of relying on a single physiological modality [49]. Roha et al. (2023) developed a machine learning framework to classify driver stress using biosignals—ECG, EDA, PPG, and respiration—collected during simulated driving. They used feature-level fusion and regularized deep kernel learning, followed by classification with SVM and Random Forest. The best results came from combining ECG and PPG signals using both SVM and RF, achieving an accuracy of (~0.93-0.97). EDA alone performed with around ~0.60 accuracy, but multimodal fusion consistently improved accuracy [56]. This also matches with the findings in the literature that SVMs often work better at physiological-arousal classification due to their ability to model complex decision boundaries. The voting ensemble uses the strengths of multiple algorithms to improve robustness, but at the cost of greater complexity. In practice, this means that a combined model gave more balanced predictions across precision, recall, and F1-score, suggesting that the ensemble model is valuable when maximizing overall accuracy is critical, but it must be weighed against interpretability and computational demand.

The aggressive acceleration task proved the most challenging. Here, XGBoost slightly outperformed all other models (achieving ~0.62 accuracy and F1 ≈0.60) using the six top features. The voting ensemble only reached ~0.55 accuracy in this case. In other words, a powerful boosting algorithm captured subtle patterns that simpler models missed, while the ensemble did not surpass

XGBoost alone. The lower accuracies suggest that aggressive acceleration may not induce strong, consistent physiological signals or that the positive-class examples were underrepresented, limiting the predictability. A notable point of comparison comes from Tsai et al. (2023), who used heart-rate variability (HRV) to detect aberrant driving behaviors such as speeding and hard braking in professional drivers. Their work achieved significantly higher accuracies (81–84%) by employing a deep learning approach (GRU networks) and incorporating temporal cardiac patterns. While their results are not directly comparable due to different target behaviors and physiological inputs, the contrast suggests that integrating HRV features and temporal modeling could substantially enhance model performance in future studies [18].

For aggressive braking, performance was stronger and was achieved with very few features. When using just the top one or two features (notably heart-rate mean and standard deviation), SVM, logistic regression, and the ensemble all reached high accuracy (~0.70). This indicates that aggressive braking events have a clearer physiological signature in our data, allowing even linear models with minimal inputs to predict well. In this task, the ensemble only matched the best individual models, meaning that once key predictors are identified, simpler models suffice. However, it is notable that precision remained low for braking across models (many false positives) while recall stayed around 0.5. This trade-off suggests the models tended to over-predict braking events to avoid missing true positives. In summary, braking could be detected with a minimal set of features and relatively simple classifiers, whereas acceleration required more complex modeling for only modest accuracy gains.

Across all tasks, our feature selection was very consistent. Both Random Forest importance and L1-regularized logistic regression repeatedly picked out similar features (e.g., tonic/phasic EDA max or std, HR median or skew) as top predictors. These features correspond to overall arousal level (tonic EDA) and sudden changes in mood (phasic EDA), as well as cardiac response (heart-rate statistics). Their recurrence suggests these physiological markers are robust indicators of driver arousal and stress. This agrees with prior work that says combining multiple autonomic

signals (EDA plus heart-rate) generally improves emotion detection. In our study, using both EDA and heart rate features likely provided complementary information about sympathetic and parasympathetic nervous system activity.

Several limitations and trade-offs affect these findings. Physiological signals are highly variable across individuals and contexts. Factors such as fitness level, hydration, time of day, or concurrent mental workload can change baseline EDA/HR, making it hard for a single model to fit everyone. In addition, our arousal labels came from self-reported mood surveys, which sometimes can be incorrect or biased. Participants chose from categories (calm, stressed, angry, etc.), which may not precisely match their instantaneous physiology. Self-report is known to suffer from recall and desirability bias, so our ground truth was imperfect. These issues likely set an upper limit on accuracy.

Finally, generalization remains a concern. Our sample size was relatively small (25 participants) and fairly homogeneous, as all participants were within a similar age range and from the same geographic location (Duluth, Minnesota), so the models may overfit these participants or specific driving conditions. While we used cross-validation, we could not test completely independent subjects or new driving environments. Therefore, the models' real-world performance outside this study is uncertain. Future work will need to validate and possibly recalibrate these models on larger, more diverse conditions to ensure robustness.

Future work

Expand the Dataset: Collecting more driving sessions from a broader range of participants (varying age, gender, driving experience, etc.) and environments can improve statistical power and model generalization.

Explore advanced model architecture: The feature importance process in a bigger dataset might become time-consuming and computationally expensive. In that case, Principal Component Analysis can be a more efficient way to reduce the number of features into a few principal components. Also, investigate time-series and deep-learning models that exploit temporal dynamics. For example, recurrent neural networks (LSTM, GRU) or convolutional LSTM networks. Because electrodermal activity (EDA), heart rate, and acceleration data are time-dependent, and their sequential patterns may provide deeper insights into the progression of emotional arousal or aggressive driving behaviors. Temporal models such as Long Short-Term Memory (LSTM), Gated Recurrent Units (GRU), or Convolutional LSTMs can capture such dependencies. These models could learn how patterns evolve over time, such as a sudden spike in phasic EDA followed by hard braking and may improve predictive accuracy.

Develop real-time, lightweight implementations: To deploy in-vehicles, models must run on low-power hardware with minimal latency. Future systems should use optimized classifiers (e.g. pruned decision trees, quantized neural networks, or embedded-friendly architectures so that driver arousal and aggression can be detected continuously in real time. Such implementations will enable closed-loop interventions (e.g., alerts or adaptive assistance) without heavy computation overhead.

References:

[1] “Aggressive Driving Infographic – How It Impacts Your Health & Driving.” [Online]. Available: https://flteensafedriver.org/aggressive-driving-impact/#:~:text=While%20aggressive%20driving%20is%20difficult,may%20have%20involved%20aggressive%20driving

[2] “Background On: Teen drivers.” [Online]. Available: https://www.iii.org/main-responsive/auto?page=1#:~:text=Motor%20vehicle%20crashes%20are%20the,such%20as%20speeding%20and%20tailgating

[3] “Minnesota Drivers’ Attitude and Behavior Traffic Safety Survey Report.” Minnesota Department of Public Safety, Sep. 02, 2022. [Online]. Available: https://www.lrl.mn.gov/docs/2023/other/231707.pdf#:~:text=Aggressive%20driving%208,Law%20enforcement%203

[4] “A Report on the State of Traffic Safety in Minnesota in 2023.” Minnesota Department of Public Safety, Jan. 2024. [Online]. Available: https://www.lrl.mn.gov/docs/2024/mandated/240203.pdf#:~:text=Careless%2C%20Negligent%2C%20or%20Erratic%20Driving,1.3

[5] V. Mishra et al., “Evaluating the Reproducibility of Physiological Stress Detection Models,” Proc. ACM Interact. Mob. Wearable Ubiquitous Technol., vol. 4, no. 4, pp. 1–29, Dec. 2020, doi: 10.1145/3432220.

[6] P. Zontone, A. Affanni, A. Piras, and R. Rinaldo, “Exploring Physiological Signal Responses to Traffic-Related Stress in Simulated Driving,” Sensors, vol. 22, no. 3, p. 939, Jan. 2022, doi: 10.3390/s22030939.

[7] S. Campanella, A. Altaleb, A. Belli, P. Pierleoni, and L. Palma, “A Method for Stress Detection Using Empatica E4 Bracelet and Machine-Learning Techniques,” Sensors, vol. 23, no. 7, p. 3565, Mar. 2023, doi: 10.3390/s23073565.

[8] K. Steinhauser et al., “Effects of emotions on driving behavior,” Transportation Research Part F: Traffic Psychology and Behaviour, vol. 59, pp. 150–163, Nov. 2018, doi: 10.1016/j.trf.2018.08.012.

[9] T. Aminosharieh Najafi, A. Affanni, R. Rinaldo, and P. Zontone, “Driver Attention Assessment Using Physiological Measures from EEG, ECG, and EDA Signals,” Sensors, vol. 23, no. 4, p. 2039, Feb.

2023, doi: 10.3390/s23042039.

[10] P. M. Litvak, J. S. Lerner, L. Z. Tiedens, and K. Shonk, “Fuel in the Fire: How Anger Impacts Judgment and Decision-Making,” in International Handbook of Anger, M. Potegal, G. Stemmler, and C. Spielberger, Eds., New York, NY: Springer New York, 2010, pp. 287–310. doi: 10.1007/978-0-387-89676-2_17.

[11] L. Montoro, S. Useche, F. Alonso, and B. Cendales, “Work Environment, Stress, and Driving Anger: A Structural Equation Model for Predicting Traffic Sanctions of Public Transport Drivers,” IJERPH, vol. 15, no. 3, p. 497, Mar. 2018, doi: 10.3390/ijerph15030497.

[12] L. Shamoa-Nir and M. Koslowsky, “Aggression on the Road as a Function of Stress, Coping Strategies and Driver Style,” PSYCH, vol. 01, no. 01, pp. 35–44, 2010, doi: 10.4236/psych.2010.11006.

[13] X. Zhang, R. Chang, X. Sui, and Y. Li, “Influences of Emotion on Driving Decisions at Different Risk Levels: An Eye Movement Study,” Front. Psychol., vol. 13, p. 788712, Feb. 2022, doi: 10.3389/fpsyg.2022.788712.

[14] K. Kyriakou et al., “Detecting Moments of Stress from Measurements of Wearable Physiological Sensors,” Sensors, vol. 19, no. 17, p. 3805, Sep. 2019, doi: 10.3390/s19173805.

[15] Y. Wang, W. Qu, Y. Ge, X. Sun, and K. Zhang, “Effect of personality traits on driving style: Psychometric adaption of the multidimensional driving style inventory in a Chinese sample,” PLoS ONE, vol. 13, no. 9, p. e0202126, Sep. 2018, doi: 10.1371/journal.pone.0202126.

[16] X. Liu et al., “The Influence of Personality and Demographic Characteristics on Aggressive Driving Behaviors in Eastern Chinese Drivers,” PRBM, vol. Volume 15, pp. 193–212, Jan. 2022, doi: 10.2147/PRBM.S323431.

[17] S. Hasaninasab, S. Jalili, and K. Mirani, “Evaluation Effects of Personality on the Young people Traffic Offense and Accident,” Iran J Sci Technol Trans Civ Eng, vol. 45, no. 1, pp. 491–503, Mar. 2021, doi: 10.1007/s40996-020-00399-5.

[18] C.-Y. Tsai et al., “Machine learning model for aberrant driving behaviour prediction using heart rate variability: a pilot study involving highway bus drivers,” International Journal of Occupational Safety and Ergonomics, vol. 29, no. 4, pp. 1429–1439, Oct. 2023, doi: 10.1080/10803548.2022.2135281.

[19] A. Mohammadzadeh Moghaddam, A. Sadeghi, M. Jalili Qazizadeh, H. Farhad, and M. Barakchi, “Investigating the relationship between driver’s ticket frequency and demographic, behavioral, and personal factors: Which drivers commit more offenses?,” Journal of Transportation Safety & Security, vol. 12, no. 2, pp. 225–244, Feb. 2020, doi: 10.1080/19439962.2018.1477894.

[20] V. C. Kummetha, U. Durrani, J. Mason, S. Concas, and A. Kondyli, “Driver Classification Using Self-reported, Psychophysiological, and Performance Metrics Within a Simulated Environment,” Data Sci. Transp., vol. 5, no. 2, p. 6, Aug. 2023, doi: 10.1007/s42421-023-00069-8.

[21] B. S. Shokri and H. R. Behnood, “Dangerous and Aggressive Driving: Detecting the Interrelationship by Data Mining,” Iran J Sci Technol Trans Civ Eng, vol. 46, no. 2, pp. 1709–1721, Apr. 2022, doi: 10.1007/s40996-021-00712-w.

[22] N. Habibifar and H. Salmanzadeh, “Relationship between driving styles and biological behavior of drivers in negative emotional state,” Transportation Research Part F: Traffic Psychology and Behaviour, vol. 85, pp. 245–258, Feb. 2022, doi: 10.1016/j.trf.2022.01.010.

[23] S. Milosevic, “Drivers’ fatigue studies,” Ergonomics, vol. 40, no. 3, pp. 381–389, Mar. 1997, doi: 10.1080/001401397188215.

[24] X. Zhang, X. Zhao, H. Du, and J. Rong, “A Study on the Effects of Fatigue Driving and Drunk Driving on Drivers’ Physical Characteristics,” Traffic Injury Prevention, vol. 15, no. 8, pp. 801–808, Nov. 2014, doi: 10.1080/15389588.2014.881996.

[25] Y. Zou, L. Ding, H. Zhang, T. Zhu, and L. Wu, “Vehicle Acceleration Prediction Based on Machine Learning Models and Driving Behavior Analysis,” Applied Sciences, vol. 12, no. 10, p. 5259, May 2022, doi: 10.3390/app12105259.

[26] P. Stephan, F. Wortmann, and K. Koch, “Understanding the Interactions Between Driving Behavior and Well-being in Daily Driving: Causal Analysis of a Field Study,” J Med Internet Res, vol. 24, no. 8, p. e36314, Aug. 2022, doi: 10.2196/36314.

[27] E. E. Miller and L. N. Boyle, “Driver Behavior in Road Tunnels: Association with Driver Stress and Performance,” Transportation Research Record, vol. 2518, no. 1, pp. 60–67, Jan. 2015, doi: 10.3141/2518-08.

[28] A. Lambert-Bélanger, S. Dubois, B. Weaver, N. Mullen, and M. Bédard, “Aggressive driving behaviour in young drivers (aged 16 through 25) involved in fatal crashes,” Journal of Safety Research, vol. 43, no. 5–6, pp. 333–338, Dec. 2012, doi: 10.1016/j.jsr.2012.10.011.

[29] McKenna, F. P., Waylen, A. E., Burkes, ME., Male and Female Drivers: How Different are They?. United Kingdom: AA Foundation for Road Safety Research. 1998.

[30] B. Simons-Morton, R. Chen, L. Abroms, and D. L. Haynie, “Latent Growth Curve Analyses of Peer and Parent Influences on Smoking Progression Among Early Adolescents.,” Health Psychology, vol. 23, no. 6, pp. 612–621, Nov. 2004, doi: 10.1037/0278-6133.23.6.612.

[31] A. P. Smith, “A UK survey of driving behaviour, fatigue, risk taking and road traffic accidents,” BMJ Open, vol. 6, no. 8, p. e011461, Aug. 2016, doi: 10.1136/bmjopen-2016-011461.

[32] A. Albraikan, B. Hafidh, and A. El Saddik, “iAware: A Real-Time Emotional Biofeedback System Based on Physiological Signals,” IEEE Access, vol. 6, pp. 78780–78789, 2018, doi: 10.1109/ACCESS.2018.2885279.

[33] P. Alinia et al., “Associations Between Physiological Signals Captured Using Wearable Sensors and Self-reported Outcomes Among Adults in Alcohol Use Disorder Recovery: Development and Usability Study,” JMIR Form Res, vol. 5, no. 7, p. e27891, Jul. 2021, doi: 10.2196/27891.

[34] M. Gjoreski, H. Gjoreski, M. Luštrek, and M. Gams, “Continuous stress detection using a wrist device: in laboratory and real life,” in Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct, Heidelberg Germany: ACM, Sep. 2016, pp. 1185–1193. doi: 10.1145/2968219.2968306.

[35] V. Mishra et al., “Continuous Detection of Physiological Stress with Commodity Hardware,” ACM Trans. Comput. Healthcare, vol. 1, no. 2, pp. 1–30, Apr. 2020, doi: 10.1145/3361562.

[36] V. Mishra et al., “The Case for a Commodity Hardware Solution for Stress Detection,” in Proceedings of the 2018 ACM International Joint Conference and 2018 International Symposium on Pervasive and Ubiquitous Computing and Wearable Computers, Singapore Singapore: ACM, Oct. 2018, pp. 1717–1728. doi: 10.1145/3267305.3267538.

[37] T. Vallès-Català, A. Pedret, D. Ribes, D. Medina, and M. Traveria, “Effects of Stress on Performance during Highly Demanding Tasks in Student Pilots,” The International Journal of Aerospace Psychology, vol. 31, no. 1, pp. 43–55, Jan. 2021, doi: 10.1080/24721840.2020.1841564.

[38] A. Horvers, N. Tombeng, T. Bosse, A. W. Lazonder, and I. Molenaar, “Detecting Emotions through Electrodermal Activity in Learning Contexts: A Systematic Review,” Sensors, vol. 21, no. 23, p. 7869, Nov. 2021, doi: 10.3390/s21237869.

[39] R. Sánchez-Reolid, M. T. López, and A. Fernández-Caballero, “Machine Learning for Stress Detection from Electrodermal Activity: A Scoping Review,” Nov. 02, 2020. doi: 10.20944/preprints202011.0043.v1.

[40] M. Maier, D. Elsner, C. Marouane, M. Zehnle, and C. Fuchs, “DeepFlow: Detecting Optimal User Experience From Physiological Data Using Deep Neural Networks,” in Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, Macao, China: International Joint Conferences on Artificial Intelligence Organization, Aug. 2019, pp. 1415–1421. doi:

10.24963/ijcai.2019/196.

[41] S. G. Holt et al., “Monitoring skin temperature at the wrist in hospitalised patients may assist in the detection of infection,” Internal Medicine Journal, vol. 50, no. 6, pp. 685–690, Jun. 2020, doi: 10.1111/imj.14748.

[42] P. Siirtola, H. Koskimäki, H. Mönttinen, and J. Röning, “Using Sleep Time Data from Wearable Sensors for Early Detection of Migraine Attacks,” Sensors, vol. 18, no. 5, p. 1374, Apr. 2018, doi: 10.3390/s18051374.

[43] A. Anzanpour et al., “Self-awareness in remote health monitoring systems using wearable electronics,” in Design, Automation & Test in Europe Conference & Exhibition (DATE), 2017, Lausanne, Switzerland: IEEE, Mar. 2017, pp. 1056–1061. doi: 10.23919/DATE.2017.7927146.

[44] B. Nakisa, M. N. Rastgoo, A. Rakotonirainy, F. Maire, and V. Chandran, “Long Short Term Memory Hyperparameter Optimization for a Neural Network Based Emotion Recognition Framework,” IEEE Access, vol. 6, pp. 49325–49338, 2018, doi: 10.1109/ACCESS.2018.2868361.

[45] E. Sheikhian, M. Ghoshuni, M. Azarnoosh, and M. M. Khalilzadeh, “Enhancing Arousal Level Detection in EEG Signals through Genetic Algorithm-based Feature Selection and Fast Bit Hopping,” Journal of Medical Signals & Sensors, vol. 14, no. 7, Jul. 2024, doi: 10.4103/jmss.jmss_65_23.

[46] L. Wan, Y. Yan, C. Zhang, C. Liu, T. Mao, and W. Wang, “Characteristics and identification of risky driving behaviors in expressway tunnels based on behavior spectrum,” International Journal of Transportation Science and Technology, vol. 16, pp. 5–17, Dec. 2024, doi: 10.1016/j.ijtst.2023.10.006.

[47] P. Qin et al., “Prediction of driving stress on high-altitude expressway using driving environment features: A naturalistic driving study in Tibet,” Traffic Injury Prevention, vol. 25, no. 3, pp. 414–424, Apr. 2024, doi: 10.1080/15389588.2024.2305420.

[48] Q. Meteier et al., “Classification of driver fatigue in conditionally automated driving using physiological signals and machine learning,” Transportation Research Interdisciplinary Perspectives, vol. 26, p. 101148, Jul. 2024, doi: 10.1016/j.trip.2024.101148.

[49] M. S. H. Nahid, T. Z. Tila, and T. S. Seecharan, “Detecting Emotional Arousal and Aggressive Driving Using Neural Networks: A Pilot Study Involving Young Drivers in Duluth,” Sensors, vol. 24, no. 22, p. 7109, Nov. 2024, doi: 10.3390/s24227109.

[50] F. Kamalov, H. Sulieman, A. Alzaatreh, M. Emarly, H. Chamlal, and M. Safaraliev, “Mathematical Methods in Feature Selection: A Review,” Mathematics, vol. 13, no. 6, p. 996, Mar. 2025, doi: 10.3390/math13060996.

[51] H. Liu, Y. Zhou, and C. Jiang, “Classifying metro drivers’ cognitive distractions during manual operations using machine learning and random forest-recursive feature elimination,” Sci Rep, vol. 15, no. 1, Mar. 2025, doi: 10.1038/s41598-025-92248-6.

[52] J. Li, H. Zhang, J. Zhao, X. Guo, W. Rihan, and G. Deng, “Embedded Feature Selection and Machine Learning Methods for Flash Flood Susceptibility-Mapping in the Mainstream Songhua River Basin, China,” Remote Sensing, vol. 14, no. 21, p. 5523, Nov. 2022, doi: 10.3390/rs14215523.

[53] S. Föll et al., “FLIRT: A feature generation toolkit for wearable data,” Computer Methods and Programs in Biomedicine, vol. 212, p. 106461, Nov. 2021, doi: 10.1016/j.cmpb.2021.106461.

[54] A. Greco, G. Valenza, A. Lanata, E. Scilingo, and L. Citi, “cvxEDA: a Convex Optimization Approach to Electrodermal Activity Processing,” IEEE Trans. Biomed. Eng., pp. 1–1, 2016, doi: 10.1109/TBME.2015.2474131.

[55] https://www.geotab.com/blog/what-is-g-force/

[56] V. S. Roha, N. Ganapathy, N. Spicher, S. Saha, and T. M. Deserno, “Assessment of Driver’s Stress using Multimodal Biosignals and Regularized Deep Kernel Learning,” in 2023 45th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), Sydney, Australia: IEEE, Jul. 2023, pp. 1–4. doi: 10.1109/EMBC40787.2023.10340564.

Appendix

A: Raw EDA Data and Separated EDA Data

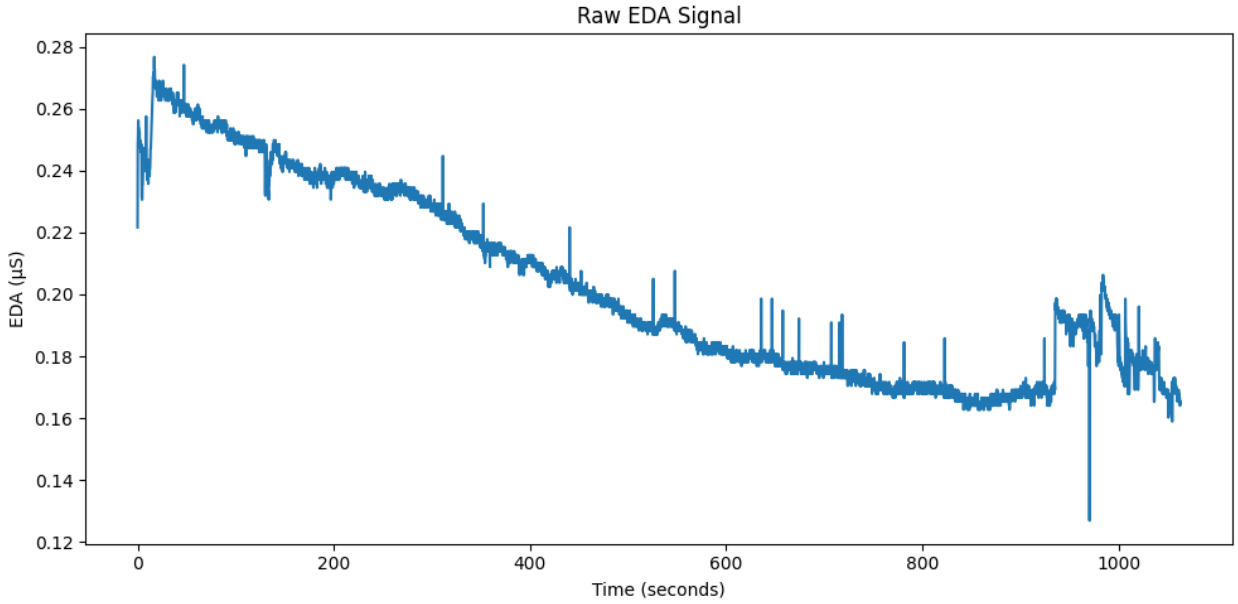

Figure 6.1: Raw EDA Signal

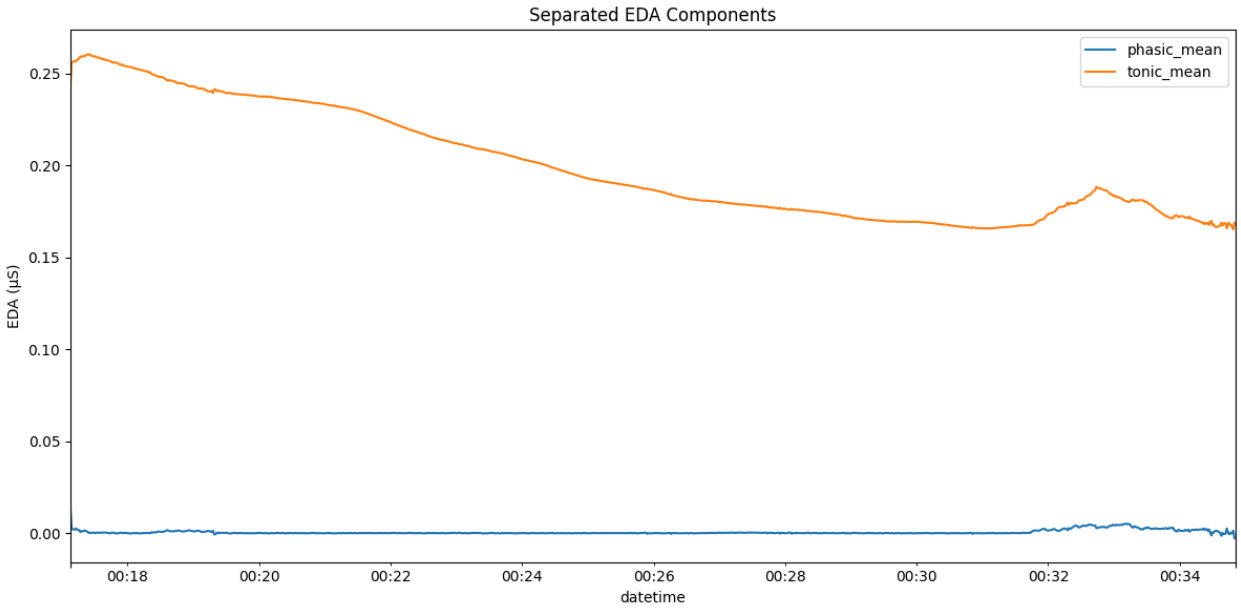

Figure 6.2: After cleaning and separating the raw EDA signal into phasic and Tonic components using FLIRT Toolkit

B. Code for Arousal/Aggressive Acceleration/Aggressive Braking Prediction Based on Random Forest Feature Importance with 16 Features

```
# Step 2: Import required libraries
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.preprocessing import StandardScaler
from sklearn.ensemble import RandomForestClassifier, VotingClassifier
from sklearn.linear_model import LogisticRegression
from sklearn.svm import SVC
from xgboost import XGBClassifier
```

```
from sklearn.model_selection import cross_val_score, StratifiedKFold
from sklearn.metrics import make_scorer, precision_score, recall_score, f1_score

# Step 3: Load your dataset
data = pd.read_csv('Final Dataset.csv')  # Adjust filename if needed

# Step 4: Define features and target
features = [
    'Base Phasic Mean', 'High Phasic Mean', 'Base Tonic Mean', 'High Tonic Mean',
    'Base Phasic std', 'High Phasic std', 'Base Tonic std', 'High Tonic std',
    'Base Phasic max', 'High Phasic max', 'Base Tonic max', 'High Tonic max',
    'Median', 'Mean', 'STD', 'SKEW'
]
X = data[features]
y = data['Arousal']

# Step 5: Scale features
scaler = StandardScaler()
X_scaled = scaler.fit_transform(X)

# Step 6: Train Random Forest and get feature importances
rf = RandomForestClassifier(random_state=42)
rf.fit(X_scaled, y)
importances = rf.feature_importances_

# Step 7: Rank features by importance
importance_df = pd.DataFrame({
    'Feature': features,
    'Importance': importances
}).sort_values(by='Importance', ascending=False)

# Step 7.1: Plot feature importances
```

```
plt.figure(figsize=(8, 4))
sns.barplot(x='Importance', y='Feature', data=importance_df)
plt.title("Feature importance chart for 16 initial features\n"
          "in arousal prediction using Random Forest",
          fontsize=12, loc='center')
plt.tight_layout()
plt.show()

# Step 8: Define individual models
models = {
    "Random Forest": RandomForestClassifier(random_state=42),
    "Logistic Regression": LogisticRegression(random_state=42),
    "SVM": SVC(probability=True, random_state=42),
    "XGBoost": XGBClassifier(random_state=42, use_label_encoder=False, eval_metric='logloss')
}

# Step 9: Define the Voting Classifier (Ensemble)
ensemble_model = VotingClassifier(estimators=[
    ('rf', models["Random Forest"]),
    ('lr', models["Logistic Regression"]),
    ('svm', models["SVM"]),
    ('xgb', models["XGBoost"])
], voting='hard')

# Step 10: Cross-validation setup
cv = StratifiedKFold(n_splits=5, shuffle=True, random_state=42)

# Step 11: Loop to find optimal number of features based on ensemble accuracy
accuracy_list = []
max_accuracy = 0
best_num_features = 0
patience = 8
```

```
no_improve_count = 0

for i in range(1, len(importance_df) + 1):
    current_features = importance_df['Feature'].iloc[:i].tolist()
    X_temp = data[current_features]
    X_temp_scaled = scaler.fit_transform(X_temp)

    acc = cross_val_score(ensemble_model, X_temp_scaled, y, cv=cv, scoring='accuracy').mean()
    accuracy_list.append(acc)

    if acc > max_accuracy:
        max_accuracy = acc
        best_num_features = i
        no_improve_count = 0
    else:
        no_improve_count += 1

    if no_improve_count >= patience:
        print(f"\nStopping early at {i} features. No improvement for {patience} steps.")
        break

# Step 12: Plot accuracy vs number of features
plt.figure(figsize=(8, 5))
plt.plot(range(1, len(accuracy_list) + 1), accuracy_list, marker='o')
plt.axvline(best_num_features, linestyle='--', color='red', label=f'Best: {best_num_features} features')
plt.title("Chart showing ensemble model accuracy for arousal prediction using different numbers\n"
          "of top features ranked by Random Forest feature importance from the original 16 features.")
plt.xlabel("Number of Features")
plt.ylabel("Ensemble Accuracy")
plt.legend()
plt.grid(True)
plt.tight_layout()
```

```
plt.show()

# Step 13: Final model evaluations using best features
final_features = importance_df['Feature'].iloc[:best_num_features].tolist()
X_selected = data[final_features]
X_selected_scaled = scaler.fit_transform(X_selected)

# Add Ensemble model for final evaluation
models["Ensemble (Voting)"] = ensemble_model

# Step 14: Evaluate models with cross-validation (include std deviation)
results = []
for name, model in models.items():
    acc_scores = cross_val_score(model, X_selected_scaled, y, cv=cv, scoring='accuracy')
    prec_scores = cross_val_score(model, X_selected_scaled, y, cv=cv, scoring=make_scorer(precision_score, average='macro'))
    rec_scores = cross_val_score(model, X_selected_scaled, y, cv=cv, scoring=make_scorer(recall_score, average='macro'))
    f1_scores = cross_val_score(model, X_selected_scaled, y, cv=cv, scoring=make_scorer(f1_score, average='macro'))

    results.append({
        'Model': name,
        'Accuracy (mean)': acc_scores.mean(), 'Accuracy (std)': acc_scores.std(),
        'Precision (mean)': prec_scores.mean(), 'Precision (std)': prec_scores.std(),
        'Recall (mean)': rec_scores.mean(), 'Recall (std)': rec_scores.std(),
        'F1-score (mean)': f1_scores.mean(), 'F1-score (std)': f1_scores.std()
    })

# Step 15: Display results
results_df = pd.DataFrame(results)
print(f"\nOptimal number of features: {best_num_features}")
```

```
print(f"Maximum Ensemble Accuracy: {max_accuracy:.4f}")
print("\nTop features selected:")
print(final_features)

print("\nModel Evaluation with 5-Fold Cross-Validation (Mean ± Std):\n")
print(results_df.to_string(index=False))

# Step 16: Plot comparison of all models (mean scores only)
results_plot = results_df[['Model', 'Accuracy (mean)', 'Precision (mean)', 'Recall (mean)', 'F1-score (mean)']].set_index('Model')
results_plot.plot(kind='bar', figsize=(10, 6))
plt.ylabel("Score")
plt.title(" Comparison of performance metrics for different machine learning models\n"
          "in predicting arousal using two features selected from the original 16 based\n"
          "on the Random Forest feature importance method", fontsize=12, loc='center')
plt.xticks(rotation=45)
plt.legend(loc='upper left', bbox_to_anchor=(1, 0.5))  # move legend out
plt.tight_layout(rect=[0, 0, 0.85, 1])  # adjust plot to fit legend
plt.show()
plt.tight_layout()
plt.show()
```

C. Code for Arousal/Aggressive Acceleration/Aggressive Braking Prediction Based on Random Forest Feature Importance with 10 Features

```
# Step 2: Import required libraries
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.preprocessing import StandardScaler
```

```
from sklearn.ensemble import RandomForestClassifier, VotingClassifier
from sklearn.linear_model import LogisticRegression
from sklearn.svm import SVC
from xgboost import XGBClassifier
from sklearn.model_selection import cross_val_score, StratifiedKFold
from sklearn.metrics import make_scorer, precision_score, recall_score, f1_score

# Step 3: Load your dataset
data = pd.read_csv('Final Dataset.csv')  # Adjust filename if needed

# Step 4: Define features and target
features = ['PhMeanDiff', 'TnMeanDiff', 'PhStdDiff', 'TnStdDiff', 'PhMaxDiff', 'TnMaxDiff', 'Median', 'Mean',
'STD', 'SKEW'
]
X = data[features]
y = data['Arousal']

# Step 5: Scale features
scaler = StandardScaler()
X_scaled = scaler.fit_transform(X)

# Step 6: Train Random Forest and get feature importances
rf = RandomForestClassifier(random_state=42)
rf.fit(X_scaled, y)
importances = rf.feature_importances_

# Step 7: Rank features by importance
importance_df = pd.DataFrame({
    'Feature': features,
    'Importance': importances
}).sort_values(by='Importance', ascending=False)
```

```
# Step 7.1: Plot feature importances
plt.figure(figsize=(8, 4))
sns.barplot(x='Importance', y='Feature', data=importance_df)
plt.title("Feature importance chart for 10 features\n"
          "in arousal prediction using Random Forest",
          fontsize=12, loc='center')
plt.tight_layout()
plt.show()

# Step 8: Define individual models
models = {
    "Random Forest": RandomForestClassifier(random_state=42),
    "Logistic Regression": LogisticRegression(random_state=42),
    "SVM": SVC(probability=True, random_state=42),
    "XGBoost": XGBClassifier(random_state=42, use_label_encoder=False, eval_metric='logloss')
}

# Step 9: Define the Voting Classifier (Ensemble)
ensemble_model = VotingClassifier(estimators=[
    ('rf', models["Random Forest"]),
    ('lr', models["Logistic Regression"]),
    ('svm', models["SVM"]),
    ('xgb', models["XGBoost"])
], voting='hard')

# Step 10: Cross-validation setup
cv = StratifiedKFold(n_splits=5, shuffle=True, random_state=42)

# Step 11: Loop to find optimal number of features based on ensemble accuracy
accuracy_list = []
max_accuracy = 0
best_num_features = 0
```

```
patience = 8
no_improve_count = 0

for i in range(1, len(importance_df) + 1):
    current_features = importance_df['Feature'].iloc[:i].tolist()
    X_temp = data[current_features]
    X_temp_scaled = scaler.fit_transform(X_temp)

    acc = cross_val_score(ensemble_model, X_temp_scaled, y, cv=cv, scoring='accuracy').mean()
    accuracy_list.append(acc)

    if acc > max_accuracy:
        max_accuracy = acc
        best_num_features = i
        no_improve_count = 0
    else:
        no_improve_count += 1

    if no_improve_count >= patience:
        print(f"\nStopping early at {i} features. No improvement for {patience} steps.")
        break

# Step 12: Plot accuracy vs number of features
plt.figure(figsize=(8, 5))
plt.plot(range(1, len(accuracy_list) + 1), accuracy_list, marker='o')
plt.axvline(best_num_features, linestyle='--', color='red', label=f'Best: {best_num_features} features')
plt.title("Chart showing ensemble model accuracy for arousal prediction using different numbers\n"
          "of top features ranked by Random Forest feature importance from the 10 features.")
plt.xlabel("Number of Features")
plt.ylabel("Ensemble Accuracy")
plt.legend()
plt.grid(True)
```

```
plt.tight_layout()
plt.show()

# Step 13: Final model evaluations using best features
final_features = importance_df['Feature'].iloc[:best_num_features].tolist()
X_selected = data[final_features]
X_selected_scaled = scaler.fit_transform(X_selected)

# Add Ensemble model for final evaluation
models["Ensemble (Voting)"] = ensemble_model

# Step 14: Evaluate models with cross-validation (include std deviation)
results = []
for name, model in models.items():
    acc_scores = cross_val_score(model, X_selected_scaled, y, cv=cv, scoring='accuracy')
    prec_scores = cross_val_score(model, X_selected_scaled, y, cv=cv, scoring=make_scorer(precision_score, average='macro'))
    rec_scores = cross_val_score(model, X_selected_scaled, y, cv=cv, scoring=make_scorer(recall_score, average='macro'))
    f1_scores = cross_val_score(model, X_selected_scaled, y, cv=cv, scoring=make_scorer(f1_score, average='macro'))

    results.append({
        'Model': name,
        'Accuracy (mean)': acc_scores.mean(), 'Accuracy (std)': acc_scores.std(),
        'Precision (mean)': prec_scores.mean(), 'Precision (std)': prec_scores.std(),
        'Recall (mean)': rec_scores.mean(), 'Recall (std)': rec_scores.std(),
        'F1-score (mean)': f1_scores.mean(), 'F1-score (std)': f1_scores.std()
    })

# Step 15: Display results
results_df = pd.DataFrame(results)
```

```
print(f"\nOptimal number of features: {best_num_features}")
print(f"Maximum Ensemble Accuracy: {max_accuracy:.4f}")
print("\nTop features selected:")
print(final_features)

print("\nModel Evaluation with 5-Fold Cross-Validation (Mean ± Std):\n")
print(results_df.to_string(index=False))

# Step 16: Plot comparison of all models (mean scores only)
results_plot = results_df[['Model', 'Accuracy (mean)', 'Precision (mean)', 'Recall (mean)', 'F1-score (mean)']].set_index('Model')
results_plot.plot(kind='bar', figsize=(10, 6))
plt.ylabel("Score")
plt.title(" Comparison of performance metrics for different machine learning models\n"
          "in predicting arousal using two features selected from the 10 features based\n"
          "on the Random Forest feature importance method", fontsize=12, loc='center')
plt.xticks(rotation=45)
plt.legend(loc='upper left', bbox_to_anchor=(1, 0.5))  # move legend out
plt.tight_layout(rect=[0, 0, 0.85, 1])  # adjust plot to fit legend
plt.show()
```

D. Code for Arousal/Aggressive Acceleration/Aggressive Braking Prediction Based on L1-regularized Logistic Regression Feature Importance with 16 Features

```
# Step 2: Import required libraries
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.preprocessing import StandardScaler
```

```
from sklearn.ensemble import VotingClassifier, RandomForestClassifier
from sklearn.linear_model import LogisticRegression, LassoCV

from sklearn.linear_model import Lasso
from sklearn.svm import SVC
from xgboost import XGBClassifier
from sklearn.model_selection import cross_val_score, StratifiedKFold
from  sklearn.metrics  import  make_scorer,  precision_score,  recall_score,
f1_score

# Step 3: Load your dataset
data = pd.read_csv('Final Dataset.csv')  # Adjust filename if needed

# Step 4: Define features and target
features = [
    'Base Phasic Mean', 'High Phasic Mean', 'Base Tonic Mean', 'High Tonic
Mean',
     'Base  Phasic  std',  'High  Phasic  std',  'Base  Tonic  std',  'High  Tonic
std',
     'Base  Phasic  max',  'High  Phasic  max',  'Base  Tonic  max',  'High  Tonic
max',
    'Median', 'Mean', 'STD', 'SKEW'
]
X = data[features]
y = data['Aggressive_Acc']

# Step 5: Scale features
scaler = StandardScaler()
X_scaled = scaler.fit_transform(X)

# Step 6: Train LassoCV and get feature importances
#lasso = LassoCV(cv=5, random_state=42).fit(X_scaled, y)
#importances = np.abs(lasso.coef_)  # Use absolute value of coefficients

# Step 6: Lasso Logistic Regression for feature selection
lasso      =      LogisticRegression(penalty='l1',      solver='liblinear',
random_state=42)
#lasso = Lasso(alpha=0.002)
lasso.fit(X_scaled, y)
```

```
# Step 7: Extract feature importance
lasso_importance = np.abs(lasso.coef_[0])
importance_df = pd.DataFrame({
    'Feature': features,
    'Importance': lasso_importance
}).sort_values(by='Importance', ascending=False)
# Step 8: Plot Lasso feature importance
plt.figure(figsize=(8, 4))
sns.barplot(x='Importance', y='Feature', data=importance_df)
plt.title("Feature importance chart for 16 initial features\n"
            "in aggressive acceleration prediction using L1-regularized
Logistic Regression",
          fontsize=12, loc='center')
plt.tight_layout()
plt.show()

# Step 7: Rank features by importance
#importance_df = pd.DataFrame({
#    'Feature': features,
#    'Importance': importances
#}).sort_values(by='Importance', ascending=False)

# Step 7.1: Plot feature importances
#plt.figure(figsize=(8, 4))
#sns.barplot(x='Importance', y='Feature', data=importance_df)
#plt.title('Feature Importance for Aggressive Acceleration (Lasso)')
#plt.tight_layout()
#plt.show()

# Step 8: Define individual models
models = {
    "Random Forest": RandomForestClassifier(random_state=42),
    "Logistic Regression": LogisticRegression(random_state=42),
    "SVM": SVC(probability=True, random_state=42),
      "XGBoost":  XGBClassifier(random_state=42,  use_label_encoder=False,
eval_metric='logloss')
}

# Step 9: Define the Voting Classifier (Ensemble)
ensemble_model = VotingClassifier(estimators=[
```

```
    ('rf', models["Random Forest"]),
    ('lr', models["Logistic Regression"]),
    ('svm', models["SVM"]),
    ('xgb', models["XGBoost"])
], voting='hard')

# Step 10: Cross-validation setup
cv = StratifiedKFold(n_splits=5, shuffle=True, random_state=42)

# Step 11: Loop to find optimal number of features based on ensemble accuracy
accuracy_list = []
max_accuracy = 0
best_num_features = 0
patience = 8
no_improve_count = 0

for i in range(1, len(importance_df) + 1):
    current_features = importance_df['Feature'].iloc[:i].tolist()
    X_temp = data[current_features]
    X_temp_scaled = scaler.fit_transform(X_temp)

      acc = cross_val_score(ensemble_model, X_temp_scaled, y, cv=cv,
scoring='accuracy').mean()
    accuracy_list.append(acc)

    if acc > max_accuracy:
        max_accuracy = acc
        best_num_features = i
        no_improve_count = 0
    else:
        no_improve_count += 1

    if no_improve_count >= patience:
          print(f"\nStopping early at {i} features. No improvement for
{patience} steps.")
        break

# Step 12: Plot accuracy vs number of features
plt.figure(figsize=(8, 5))
plt.plot(range(1, len(accuracy_list) + 1), accuracy_list, marker='o')
```

```
plt.axvline(best_num_features, linestyle='--', color='red', label=f'Best:
{best_num_features} features')
plt.title("Chart showing ensemble model accuracy for aggressive acceleration
prediction\n"
          "using different numbers of top features ranked by L1-regularized
Logistic Regression feature\n"
           "importance from the original 16 features.")
plt.xlabel("Number of Features")
plt.ylabel("Ensemble Accuracy")
plt.legend()
plt.grid(True)
plt.tight_layout()
plt.show()

# Step 13: Final model evaluations using best features
final_features                                                              =
importance_df['Feature'].iloc[:best_num_features].tolist()
X_selected = data[final_features]
X_selected_scaled = scaler.fit_transform(X_selected)

# Add Ensemble model for final evaluation
models["Ensemble (Voting)"] = ensemble_model

# Step 14: Evaluate models with cross-validation (include std deviation)
results = []
for name, model in models.items():
      acc_scores = cross_val_score(model, X_selected_scaled, y, cv=cv,
scoring='accuracy')
     prec_scores = cross_val_score(model, X_selected_scaled, y, cv=cv,
scoring=make_scorer(precision_score, average='macro'))
      rec_scores = cross_val_score(model, X_selected_scaled, y, cv=cv,
scoring=make_scorer(recall_score, average='macro'))
      f1_scores = cross_val_score(model, X_selected_scaled, y, cv=cv,
scoring=make_scorer(f1_score, average='macro'))

    results.append({
        'Model': name,
             'Accuracy (mean)': acc_scores.mean(), 'Accuracy (std)':
acc_scores.std(),
```

```
            'Precision (mean)': prec_scores.mean(), 'Precision (std)':
prec_scores.std(),
                'Recall (mean)': rec_scores.mean(), 'Recall (std)':
rec_scores.std(),
              'F1-score (mean)': f1_scores.mean(), 'F1-score (std)':
f1_scores.std()
    })

# Step 15: Display results
results_df = pd.DataFrame(results)
print(f"\nOptimal number of features: {best_num_features}")
print(f"Maximum Ensemble Accuracy: {max_accuracy:.4f}")
print("\nTop features selected:")
print(final_features)

print("\nModel Evaluation with 5-Fold Cross-Validation (Mean ± Std):\n")
print(results_df.to_string(index=False))

# Step 16: Plot comparison of all models (mean scores only)
results_plot = results_df[['Model', 'Accuracy (mean)', 'Precision (mean)',
'Recall (mean)', 'F1-score (mean)']].set_index('Model')
results_plot.plot(kind='bar', figsize=(10, 6))
plt.ylabel("Score")
plt.title(" Comparison of performance metrics for different machine learning
models\n"
         "in predicting aggressive acceleration using three feature selected
from the original 16 features based\n"
           "on the L1-regularized Logistic Regression feature importance
method", fontsize=12, loc='center')
plt.xticks(rotation=45)
plt.legend(loc='upper left', bbox_to_anchor=(1, 0.5))  # move legend out
plt.tight_layout(rect=[0, 0, 0.85, 1])  # adjust plot to fit legend
plt.show()
```

E. Code for Arousal/Aggressive Acceleration/Aggressive Braking Prediction Based on L1-regularized Logistic Regression Feature Importance with 10 Features

```
# Step 2: Import libraries
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.linear_model import LogisticRegression
from sklearn.ensemble import RandomForestClassifier
from sklearn.svm import SVC
from xgboost import XGBClassifier
from sklearn.preprocessing import StandardScaler
from sklearn.model_selection import train_test_split
from sklearn.metrics import accuracy_score, classification_report

# Step 3: Upload dataset
from google.colab import files
#uploaded = files.upload()

# Step 4: Load and prepare data
data = pd.read_csv('Final Dataset.csv')  # Adjust name if needed

features = ['PhMeanDiff', 'TnMeanDiff', 'PhStdDiff', 'TnStdDiff',
'PhMaxDiff', 'TnMaxDiff',
    'Median', 'Mean', 'STD', 'SKEW'
] #'
X = data[features]
y = data['Aggressive_Brk']

# Step 5: Train-test split & scaling
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3,
stratify=y, random_state=42)
scaler = StandardScaler()
X_train_scaled = scaler.fit_transform(X_train)
X_test_scaled = scaler.transform(X_test)

# Step 6: Lasso Logistic Regression for feature selection
```

```
lasso = LogisticRegression(penalty='l1', solver='liblinear',
random_state=42)
lasso.fit(X_train_scaled, y_train)

# Step 7: Extract feature importance
lasso_importance = np.abs(lasso.coef_[0])
feature_importance_df = pd.DataFrame({
    'Feature': features,
    'Importance': lasso_importance
}).sort_values(by='Importance', ascending=False)
# Step 8: Plot Lasso feature importance
plt.figure(figsize=(8, 4))
sns.barplot(x='Importance', y='Feature', data=feature_importance_df)
plt.title('Feature Importance using Lasso Regression')
plt.tight_layout()
plt.show()

print("Top  Features by Lasso Importance:")
print(feature_importance_df.head(4))
top_features = feature_importance_df['Feature'].head(2).tolist()
print("Top selected features:", top_features)

X_train_top5 = X_train_scaled[:, [features.index(f) for f in top_features]]
X_test_top5 = X_test_scaled[:, [features.index(f) for f in top_features]]

# Step 9: Train models
models = {
    "Random Forest": RandomForestClassifier(random_state=42),
    "Logistic Regression": LogisticRegression(random_state=42),
    "SVM": SVC(random_state=42),
      "XGBoost":  XGBClassifier(random_state=42,  use_label_encoder=False,
eval_metric='logloss')
}

results = []
for name, model in models.items():
    model.fit(X_train_top5, y_train)
    y_pred = model.predict(X_test_top5)
    report = classification_report(y_test, y_pred, output_dict=True)
    results.append({
```

```
        'Model': name,
        'Accuracy': accuracy_score(y_test, y_pred),
        'Precision (Macro)': report['macro avg']['precision'],
        'Recall (Macro)': report['macro avg']['recall'],
        'F1-score (Macro)': report['macro avg']['f1-score'],
        'Precision (1)': report['1']['precision'],
        'Recall (1)': report['1']['recall'],
        'F1-score (1)': report['1']['f1-score'],
        'Precision (Weighted)': report['weighted avg']['precision'],
        'Recall (Weighted)': report['weighted avg']['recall'],
        'F1-score (Weighted)': report['weighted avg']['f1-score'],
    })

# Step 10: Display results
results_df = pd.DataFrame(results)
print("\nModel Evaluation with Top 5 Features:\n")
print(results_df.to_string(index=False))
```

F. Survey

1. Name:

2. For matching the EDA and driving data please enter your trip date, start time (start driving), and end time (end driving). Example: 16th February, 9:00 am to 10:00 am:

3. Rate your mood before driving:
   - Happy
   - Stress
   - Calm
   - Fatigued
   - Angry
4. Rate your mood while driving:
   - Happy
   - Stress
   - Calm
   - Fatigued

- Angry

5. Rate your mood after driving:
- Happy
- Stress
- Calm
- Fatigued
- Angry

6. Self-reported sleep quality
Duration of sleep (h/day):

7. Sleep efficiency(%)
(To determine sleep efficiency, divide the entire time spent asleep by the total time spent in bed.):

8. How alert you were while driving on a scale of 1 to 9?
Extremely Alert: Score of 1
Alert: Score of 3
Neither Alert nor Sleepy: Score of 5
Sleepy (but no difficulty remaining awake): Score of 7
Extremely Sleepy (fighting sleep): Score of 9