Pattern Analysis on Sensory Data

Abhishek Choudhary (22684)
M.Tech, MobE

Ashhar Zaman (22881)
M.Tech, MobE

Allen C George (22563)
M.Tech, MobE

Aman Raj (22893)
M.Tech, MobE

Abstract

Navigating a vehicle is becoming progressively intricate due to the integration of various features, including communication and entertainment options. These additional functionalities, designed to enhance the driving experience, add to the overall workload for the driver. Assessing the level of workload imposed on the driver is a complex task, with existing methods often focusing more on controlled experimental settings, such as questionnaires, rather than real-world scenarios. Our focus is on investigating physiological data that can be non-intrusively measured in the future, offering a practical approach for real-world applications.

1 Introduction

In the advancement of mobility domain inferences like the mapping of driving patterns, behaviour on different routes,road types,conditions of vehicle use etc... are becoming important by the day.Use of physiological rather than conventional means like audio, video or questionnaires is relatively easier to model in terms of complexity and size of data.

In this study, we attempted to create a model classifier based on labelled data collected from participant and then use this for classification on unknown data for different classes of road segments and driving patterns of different participants based on which inferences were made.

2 Dataset

The data is collected by hciLab Group, Institute for Visualization and Interactive Systems, Unversity of Stuttgart, Germany and is used under OPen DAtabase License.The dataset comprises approximately **2.21 Lakh** pings across features like 'AccelX', 'AccelY', 'AccelZ', 'Lightning', 'Speed GPS', 'Bearing GPS','ECG', 'SCR', 'Temp', 'HR', 'HRV LF'.Apart from the labelled data there is also unlabelled data from many participants on the same route. Interactive point maps are plotted to visualize the coordinates on different road types segments using **folium** library of python. GIANNINI (2022)

3 Data Pre-processing

3.1 Excluding Irrelevant features

In the dataset, certain features are deemed irrelevant for training and analysis. Features like Unnamed: 0, Time_Videorating, Time_Light, Time_Accel, Time_GPS, Frame_Videorating, and Rating_Videorating have been excluded to streamline the training and analysis process, reducing complexity.

3.2 Distribution Statistics of Features

To gain insights into the dataset, distribution statistics of features were computed. Visual aids, in the form of plots, were generated to facilitate both visualization and processing. These plots serve as a valuable tool for comprehending the spread and patterns within each feature, enhancing the overall understanding of the dataset.

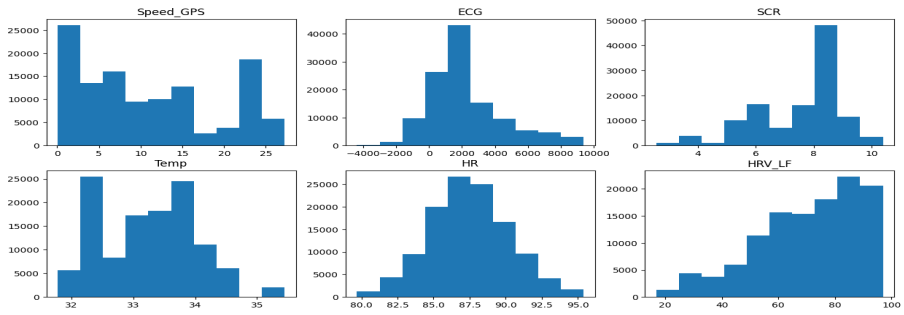


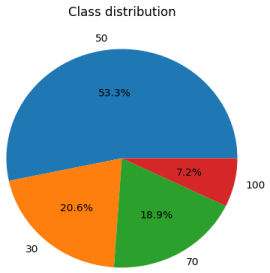
Figure 1: Distribution Statistics of Features

3.3 Refining Data by Outlier Removal

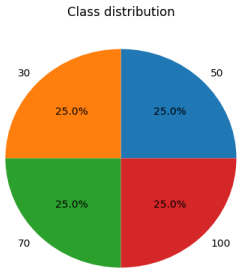
Box plots were utilized for visualizing the range of values across various features. Employing the Inter quartile Range (IQR) method, outliers were systematically identified and subsequently removed. Following this outlier removal process, another set of box plots was generated to visualize the impact of this refinement on the data distribution.

3.4 Stratified sampling

To comprehend the distribution of road types, a pie chart was created, revealing uneven class distribution within the 'target' column of the dataset. Recognizing the importance of balanced representation for effective training, stratified sampling was implemented. This methodology ensures proportional inclusion of each road type in the training dataset, thereby enhancing the model's ability to generalize across diverse road classes



(a) Before Stratified Sampling

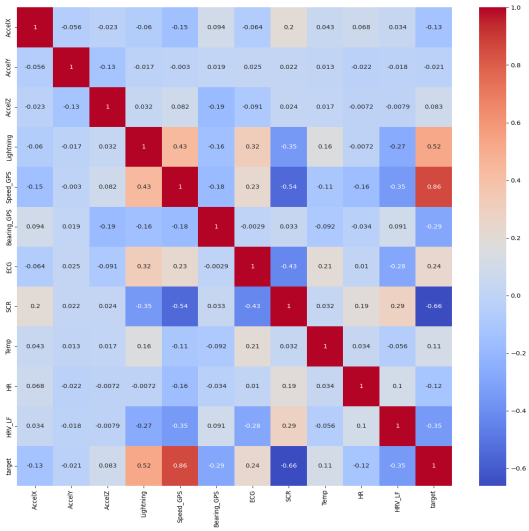


(b) After Stratified Sampling

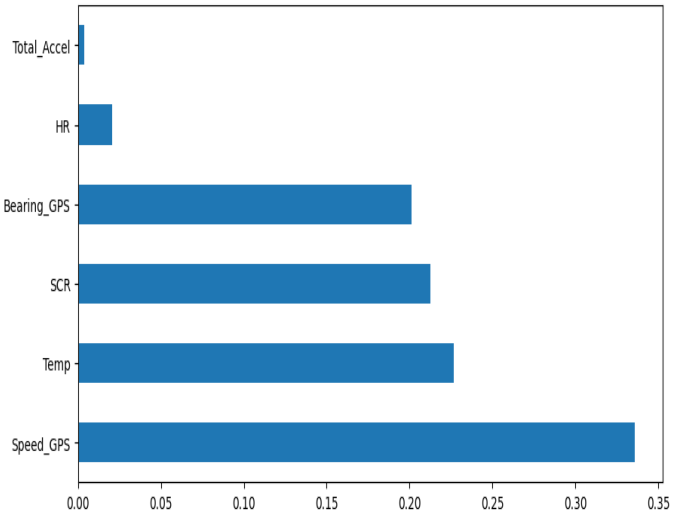
Figure 2: Pie chart of Class Distribution

3.5 Dimensionality Reduction

After assessing the correlation matrix and feature importance, 'lightning' and 'ECG' were identified as highly correlated with other features and taking some would not affect the model performance. Thus, these features were removed.



(a) Correlation Matrix



(b) Feature Importance

Figure 3: Feature Analysis

3.6 Scaling

Standard scaling was done on the data to scale the features in the standard normal distribution so that the model does not get dominated by high magnitude features.

3.7 Feature Transformation

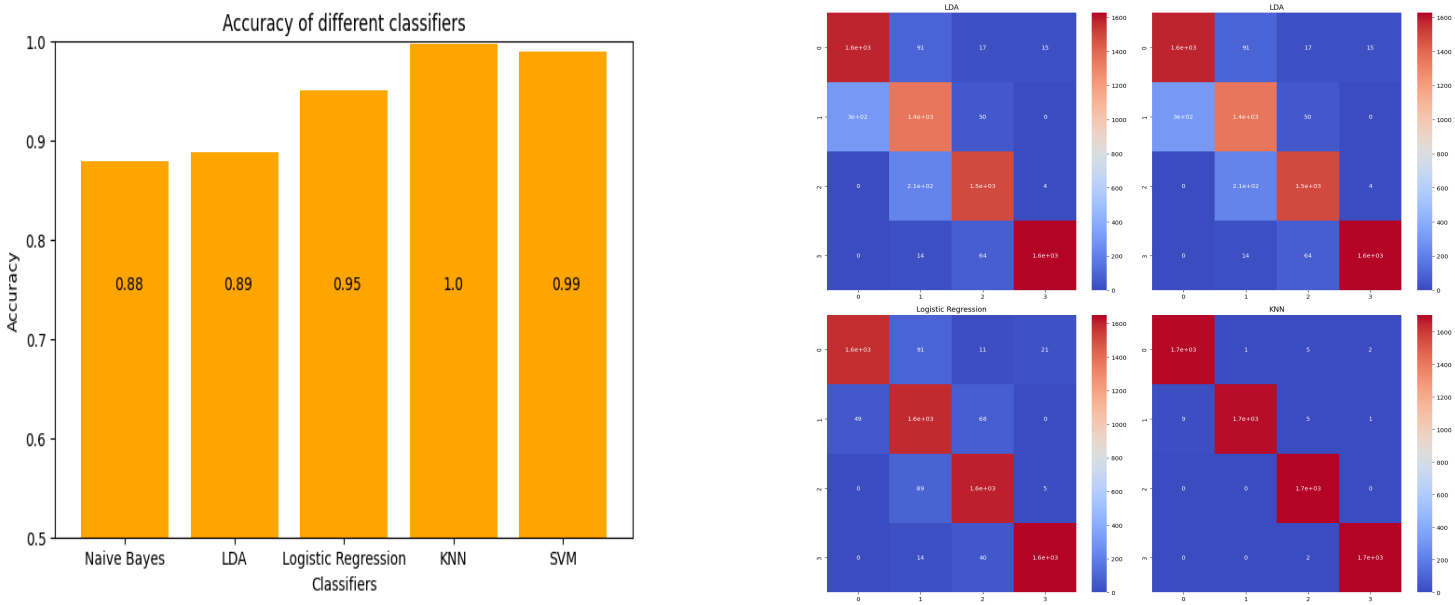
New feature such as total acceleration was added using the X, Y, Z acceleration given by: $a = a_x^2 + a_y^2 + a_z^2$ that was added to the features list instead of individual accelerations.

3.8 Data Test Train split

Test train split is done for the training of model with stratify sampling for the training and testing of the model for the calculation of accuracy. The Train data and test data was split in the ratio of 80:20 respectively.

4 Modelling

Following the scaling process we modelled for five different classifiers varying the parameters.Naieve Bayes is the simplest and easiest explainable model since it takes probabilistic approach so , it was modelled first.But it assumes a independence of features among them which is not true entirely.For kNN different number of nearest neighbours (3,5,7) were tested and based on accuracy '5' was choosen. Since non-linearity is there in data a non linear boundary classifier is taken. For this kernel for SVM is chosen to be 'rbf', although polynomial was tried with 2 and 3 degree 'rbf' is still marginally better by metric.LR with one vs all approach for classification was modelled.LDA classifier assumed Gaussian distribution for classes and does dimension reduction internally a to classify.



(a) Accuracy of Different Classifiers

(b) Confusion Matrix for Different Classifiers

Figure 4: Model Comparison

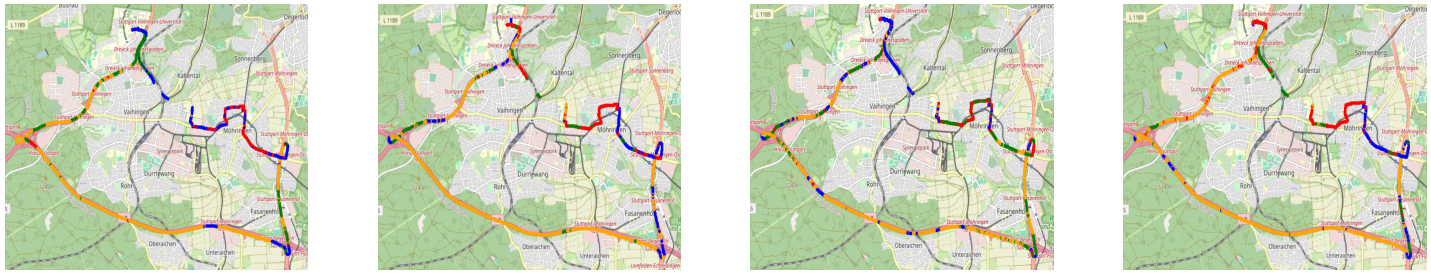
The trained models were tested on the test data that was stratified sampled earlier from splitting the original data. Since the data is more than three dimensional a visualization of decision boundary is not possible so, to pick a classifiers among the trained models we used accuracy measure and confusion matrix to get inference about mis-classified points. Clearly from model comparison plots the best classifier trained on the given data is **kNN** with an accuracy of **99.63** and only 25 mis-classified points out of 6,816 total test points.

5 Prediction of Unlabelled Data and Inference

After training the model, the model with highest accuracy was chosen i.e. K Nearest Neighbour for the prediction of other participants. The unlabelled data was pre processed before the actual prediction such as feature transformation and scaling. The predicted data was then plotted on interactive folium map and compared with the labelled data map which is found to be similar, the highways and freeways was showing higher speed than normal roads.The plots for different features overlayed on map data is then used fir further inference.

6 Inference Analysis

6.1 Road Class Prediction



(a) Participant 1 (b) Participant 2 (c) Participant 3 (d) Participant 4

Figure 5: Road Segments Map

The predicted road types and physiological feature patterns on these roads were plotted for each participant.Colour map for different classes was made such that:**Orange** >70kmph , **Green =** (50kmph,70kmph), **Blue =** (30kmph,50kmph), **Red** <30kmph.As seen in the map most of the participants on the freeway drive at higher speeds as evident from yellow class.Plot implies they were cruising at high speeds. Almost at all the turns/junction and round abouts there is a decrease in speed observered by different class colour at these points.**Stuttgart Vaihingen Universitat** and **Stuttgart Mohringen Ost** areas are places of high physiological stress and low speeds.

6.2 Heart rate and Heart Rate Variation

To find the high workload areas on the road heart rate > 140 and Heart rate Variation > 93 was located on the folium map. We found that congested areas and areas with more turns and exits i.e at low speeds are showing higher heart rate than freeways and highways.

(a) High heart rate regions (b) High heart rate variation regions

Figure 6: Heart rate and Heart rate variation

6.3 Altitude and Speed

We found regions with high altitudes by taking altitude GPS value ≥ 550 and regions with high speed, Speed GPS ≥ 30 and found that high speed is found at driving from high altitude to low altitude as seen in the graph.

(a) High Altitude regions (b) High speed regions

Figure 7: Altitude and Speed

Since no feature is there that directly represent Driver's Workload, we tried to make one by taking into account the effect of features like Heart Rate and Skin Conductance Response (SCR) which are dominant and directly relating and features like Body Temperature and Heart Rate Variation which inversely relate to driver's work load.Multiplying by different weight we made a feature **Driver Workload Index** ranging from 9.7-62.6(Higher value indicative of higher load).A heatmap of this feature overlayed on map data helps identifying possible points of interests of High Driver Workload.

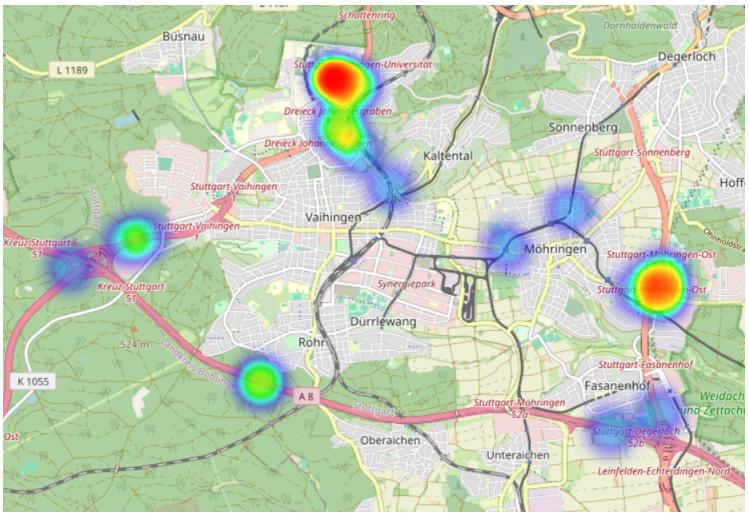


Figure 8: Driver Workload Index

7 Future Work

We can use these inferences for the workload adaptive systems such as ADAS (Advanced Driver Assistance Systems) that can use the workload on the driver and can work according to that for better safety of the driver.

References

GIANNINI, S. (2022). A dataset of real world driving to assess driver workload. In *Driving Dataset - HCILAB*. kaggle.

Stefan Schneegass, Bastian Pfleging, Nora Broy, Albrecht Schmidt, and Frederik Heinrich. 2013. A data set of real world driving to assess driver workload. In Proceedings of the 5th International Conference on Automotive User Interfaces and Interactive Vehicular Applications (AutomotiveUI '13). ACM, New York, NY, USA, 150-157. DOI=10.1145/2516540.2516561