Predicting comfort in autonomous driving from vibration measurements using machine learning models

KATE ASARAR
Predicting comfort in autonomous driving from vibration measurements using machine learning models

Kate Asarar

Master of Science in Engineering
Master program in Vehicle Engineering
KTH Royal Institute of Technology

Supervisor at BMW: Georg Burkhard
Supervisor at KTH: Jenny Jerrelind
Examiner at KTH: Jenny Jerrelind

Date of presentation: September 18, 2020

Swedish title: Komfortprediktering i autonoma fordon från vibrationsmätningar med hjälp av maskininlärningsmodeller
Acknowledgement

It is my genuine pleasure to express my deepest gratitude and thanks to my supervisor and guide Jenny Jerrelind, Associate Professor in Vehicle Dynamics at KTH Aeronautical and Vehicle Engineering. Her dedication and keen interest in the work of her students and her overwhelming support in all aspects of this work has inspired and motivated me to a very great extent to accomplish this task.

I thank profusely my supervisor at BMW Georg Burkhard for giving me the opportunity to be a part of the innovation process at a leading OEM.

It is my privilege to thank my partner Savo for his constant encouragement and most importantly for his patience and for believing in me even when I was full of doubt.

I’m deeply grateful to the person without whom neither this work nor I would have existed, my lovely mother Assia Chakkor. The stubbornness that I got from her has been my greatest asset in finishing this work and for that and much more I will be forever be thankful.

- Kate Asarar
Sammanfattning


De maskininlärningsmodeller som undersöktes i detta examensarbete har visat stor potential för att upptäcka komplexa mönster som kopplar känslor och tankar till mekaniska variabler. Modellerna kunde förutsäga rätt komfortnivå med upp till 50% precision när 6 eller 7 nivåer av komfort användes. Vid uppdelning i hög mot låg komfort, dvs att kunna förutsäga en av två komfortnivåer, kunde modellerna uppnå en precision på upp till 75.4%.
Abstract

Highly automated driving is approaching reality at a high speed. BMW is planning to put its first autonomous driving vehicle on the road already by 2021. The path to realising this new technology is however, full of challenges. Not only the transverse and longitudinal dynamic vehicle motion play an important role in experienced comfort but also the requirements and expectations of the occupants regarding the vertical dynamic vibration behaviour. Especially during long trips on the motorway where the so far active driver becomes the chauffeured passenger, who reads, works or sleeps in his newly gained time. These new use-cases create new requirements for the future design of driving comfort which are yet to be fully discovered.

This work was carried out at the BMW headquarters and had the aim to use different machine learning models to investigate and identify patterns between the subjective comfort values reported by participants in a study, on a comfort scale of 1-7 and the mechanical vibrations that they experienced, measured in m/s². The data was collected in a previous independent study and statistical methods were used to insure the quality of the data. A comparison of the ISO 2631-1 comfort ratings and the study’s findings is done to understand the need for a more sophisticated model to predict comfort in autonomous driving. The work continued by investigating different dimensionality reduction methods and their influence on the performance of the models. The process used to build, optimise and validate neural networks and other models is included in the method chapter and the results are presented. The work ends with a discussion of both the prediction results and the models re-usability.

The machine learning models investigated in this thesis have shown great potential for detecting complex pattern that link feelings and thoughts to mechanical variables. The models were able to predict the correct level of comfort with up to 50% precision when trying to predict 6 or 7 levels of comfort. When divided into high versus low discomfort, i.e. predicting one of two comfort levels, the models were able to achieve a precision of up to 75.4%.

Excluded from this thesis is the study of differences in attentive vs inattentive state when being driven in a autonomous driving vehicle. It became clear shortly before the start of this work, that the experiment that yielded the data used for it failed to find a statistically significant difference between the two states.
# Contents

1 Introduction  
1.1 Aim .................................................. 1  
1.2 Outline .................................. 1  

2 Theoretical background  
2.1 Autonomous driving vehicles  
2.2 Passenger comfort in road vehicles  
  2.2.1 What is comfort?  
  2.2.2 Comfort in ADV's  
  2.2.3 Measuring comfort  
  2.2.4 Measuring vibrations  
2.3 Artificial neural networks  
  2.3.1 Biological neurons  
  2.3.2 Artificial neuron  
  2.3.3 Previous use of machine learning to objectify comfort  
2.4 Constructing a neural network  
  2.4.1 Feedforward neural network  
  2.4.2 Activation functions  
  2.4.3 Weights  
  2.4.4 Training a neural network using backpropagation and gradient descend  
2.5 Other machine learning models  
  2.5.1 K-nearest neighbors  
  2.5.2 Support vector machine  
  2.5.3 Random forest  
  2.5.4 One-vs-Rest  
  2.5.5 One-vs-One  
  2.5.6 Feature extraction  
2.6 Model evaluation  

3 Data collection study  
3.1 Study objectives and method  
3.2 Simulator  
3.3 Vibration and comfort measurements
Chapter 1

Introduction

Autonomous driving technology stand to bring along great improvements to our lives and industries. One of the biggest benefits is thought to be gained through increased road safety as each year human error is found to be the leading cause in 60% of the traffic incidents that kill approximately 1.25 million humans [1]. Another increasingly important benefit is the ability of autonomous driving to reduce vehicle emissions. Driving strategies in autonomous driving mode can be compared to adaptive cruise control systems (ACC) which have been shown to reduce fuel consumption by 5-7% [2]. Furthermore, decrease in overall number of traffic accidents would lead to less traffic blockage, resulting in lowering the amount of time vehicles are driving at lower velocities or standing still, which is the least efficient operating point of most combustion engines. Lastly, autonomous driving technology is predicted to have personal and societal benefits such as extending the freedom of easy and available transportation for otherwise neglected minorities in society, such as the elderly and people with disabilities [3].

Global vehicle manufacturers have already realised the potential this technology has and are preparing to release several levels of autonomous driving vehicles (ADV:s) between the years 2021-2030 [4]. However, as the technology is moving from being a question of feasibility into a reality, a new set of questions is proposed regarding the technology acceptance and usage. While acceptance remains a very complex and difficult part to predict, various studies are pointing out the way to gain it.

Technology acceptance has been proven to be based on two fundamental aspects, perceived ease of use and perceived usefulness [5]. For perceived ease of use the interface between machine and human should be made as simple and clear as possible. For increasing perceived usefulness the driving mode should enable the passenger to make as much use of their time as possible during the drive. In a study on the German public's view on autonomous driving [6] it was found that relaxing, surfing the internet, as well as reading, working were some of the most likely activities to be carried out while in autonomous driving mode. Furthermore, the results of another study [7] showed that the most positively viewed aspects of ADV:s are comfort, safety, reliability and flexibility. As comfort will play an even bigger role when driving becomes autonomous, an investigation of whether the state of art comfort measuring methods, as well as the expectations put on comfort are precise enough is needed.
During any driving mode, passengers are exposed to two main disturbances, leading to increased discomfort, namely road and load disturbances. Their effect however, varies depending on whether the passenger is attentive or not. While actively driving, load disturbances can be predicted and the passenger is able to take measures to minimise the undesirable outcome on themselves, by for example leaning in to the curve or tensing their muscles. In autonomous driving mode however, the task of avoiding load disturbances is taken over by the vehicle’s computers. This is achieved through a combination of adapting acceleration profile to the desired values [8], as well as keeping the magnitude and frequency within certain limits [9]. While load disturbances can be minimised using a smooth driving profile, road disturbances are difficult to avoid. A sophisticated algorithm could steer the vehicle away from deep irregularities in the road, but most of them are unavoidable. It is therefore important to look into the effects of autonomous driving on passenger comfort and evaluate if current implementations will fill the need of the future.

1.1 Aim

This work aims at suggesting a system for comfort prediction given vehicle vibrations measurements. The system shall consist of a data transformation pipeline where relevant physical properties are calculated from recorded vibration signals, followed by a predictive machine learning model. Furthermore, this work aims at investigating the extent to which different machine learning models are able to learn and find patterns between measured vibrations and their comfort labels. Lastly, data wrangling methods which can increase the prediction accuracy shall be further investigated.

The work shall include the following stages. First, a deep literature research is performed to gain a deep understanding of comfort and machine learning applications. Second, the collected data is to be cleaned and transformed to a format suitable for machine learning applications. Third, a method to evaluate the performance of the machine learning models is constructed based on the intended real life application of the predictive system. Finally, the machine learning models are constructed, optimised and evaluated. Lastly, a thesis is written describing the detailed methodology and results of the work which is performed at BMW in Germany.

1.2 Outline

The outline of this report is as follow. Firstly, in Chapter 2 theoretical background to the machine learning, neural networks and comfort measurements as well as their evaluation method is given. Chapter 3 describes the comfort evaluation study that has been done prior to the work presented in this report and the resulting dataset which is then used in this work. Chapter 4 describes the methodology of setting-up, training and evaluating different machine learning models as well as the different assessments used to measure their accuracy. In Chapter 5 the results of the work are presented and discussed. In Chapter 6 the work and its outcome is summarised and an outlook on future work is given.
Chapter 2

Theoretical background

In this chapter central concepts and terms in the technologies discussed in this thesis are defined and explained. Starting with introducing the concept of autonomous driving, followed by the importance of technology acceptance, moving on to an elaboration on comfort and how it is measured and ending with an introduction to machine learning and an overview of the state of the art usage of machine learning in the field of measuring comfort.

2.1 Autonomous driving vehicles

An autonomous driving vehicle is defined as a vehicle which can drive itself with no need of instructions from a human driver [10]. The term is used however, also for vehicles that need a level of assistance from the driver as well and are therefore not autonomous by definition. The Society of automotive engineers (SAE) introduced for this reason an indexing table for categorising vehicles depending on their ability to drive independently of human control. With the use of norm SAE J3016 [11], autonomous vehicles can be divided into six different levels of automation. Ranging from no driver assistant systems at all to fully autonomous self-driving vehicles. A summary of the different levels is given in Table 2.1. So far (2020) very few vehicles have been able to support level 3 automation. Most vehicle manufacturers are however set on reaching it within the upcoming year or two [12]. High levels of autonomous driving services are predicted to result in significant changes to day-to-day vehicle usage. For the sake of keeping up with the changes in customer behaviour, this shift has to be predicted and preparations have to be made in order to ensure the success of future vehicle models.
Table 2.1: Summary of driving automation levels for on-road vehicles according to SAE J3016 [11].

<table>
<thead>
<tr>
<th>SAE level</th>
<th>SAE name</th>
<th>Definition</th>
<th>Execution of steering and acceleration/deceleration</th>
<th>Monitoring of driving environment</th>
<th>Fallback level</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>No automation</td>
<td>Human driver is fully operating the vehicle, even when intervention systems (e.g. ABS(^a) or ESP(^b)) are active</td>
<td>Driver</td>
<td>Driver</td>
<td>Driver</td>
</tr>
<tr>
<td>1</td>
<td>Driver assistance</td>
<td>One driver assistance system supports the driver during steering or acceleration/deceleration (e.g. ACC(^c))</td>
<td>Driver &amp; System</td>
<td>Driver and Driver</td>
<td>Driver</td>
</tr>
<tr>
<td>2</td>
<td>Partial automation</td>
<td>One or multiple driver assistance systems support the driver during steering and acceleration/deceleration</td>
<td>System</td>
<td>Driver</td>
<td>Driver</td>
</tr>
<tr>
<td>3</td>
<td>Conditional automation</td>
<td>Autonomous driving with the requirement, that the human driver intervenes when requested by the system</td>
<td>System</td>
<td>System</td>
<td>Driver</td>
</tr>
<tr>
<td>4</td>
<td>High automation</td>
<td>Autonomous driving without the expectation of an intervention on the human drivers part</td>
<td>System</td>
<td>System</td>
<td>System</td>
</tr>
<tr>
<td>5</td>
<td>Full automation</td>
<td>Full-time performance of all driving tasks by an automated driving system under all external conditions that can be carried out by a human driver</td>
<td>System</td>
<td>System</td>
<td>System</td>
</tr>
</tbody>
</table>

\(^a\)ABS:Anti-lock braking system  
\(^b\)ESP:Electronic stability control  
\(^c\)ACC:Adaptive cruise control
2.2 Passenger comfort in road vehicles

Road disturbances are transported from the road to the passenger through the vehicle body in the form of vibrations. These vibrations are then felt by the passenger and could lead to a feeling of discomfort. To classify what level of vibration is comfortable, one should start with understanding what comfort is and how we experience it. In this chapter comfort research is explored and comfort is defined using the latest agreed upon definition in normal driving as well as autonomous driving contest. Furthermore, the common measuring procedure of comfort is explained as well as the steps taken to require the data used for this thesis.

2.2.1 What is comfort?

The definition of comfort has been changed in research throughout the years. The Cambridge dictionary offers the lexical definition ”the pleasant and satisfying feeling of being physically or mentally free from pain and suffering, or something that provides this feeling” [13]. The definition hints to the relationship between comfort and discomfort, i.e. explaining that they are opposites on a continuous scale, and insinuating that the only way to experience comfort is to eliminate discomfort. Herzberg et al. use the same definition and refer to comfort as ”the absence of discomfort, [...] a state of no awareness at all of a feeling” [14]. Other researchers argue that the relation between the two is experienced in different dimensions, and that there could be factors present that give rise to both comfort and discomfort simultaneously. Zhang et al. present a notion that comfort and discomfort are two different entities [15] i.e. discomfort is the result of physiological and biomechanical factors, whilst comfort is the product of aesthetics. It is therefore possible for both comfort and discomfort to simultaneously exist.

A great example of such a situation is put forward by Bubb et al. by means of the race vehicle seat argument [16]. Bubb et al. argues that the seat of a fast vehicle is perceived as uncomfortable due to the unfavourable but performance elevating seat design and hard connection to the stiff chassis. At the same time it is rated as comfortable as its looks help lowering the passengers expectations of comfort compared to a normal seat. A vehicle seat would therefore not fall on the curve connecting discomfort and comfort in the traditional sense see Figure 2.1b, but somewhere where both discomfort and comfort is relatively high. He also presented the comfort pyramid in vehicle comfort design, meant to show the fundamental aspects of comfort evaluation in vehicles see Figure 2.1a.
2.2.2 Comfort in ADV:s

In autonomous driving level 3 to 5, the driver will be free to direct their focus on any task they choose. For less than level 5 however, the passenger will still be required to take back control of the vehicle if needed, making facing the road the only possible seating position and the number of possible in-vehicle activities is therefore restricted. Activities such as working, eating and entertainment however, are still possible. Figure 2.2 shows a number of activities the former driver can engage in while traveling in an autonomous driving vehicle as per the results of the questionnaire from [6]. With driver acceptance rooting in perception of technology usage, it is therefore a good starting point to explore what limitations today’s vehicle impose for carrying out previously mentioned activities. Revisiting the Bubbs comfort pyramid illustrated in Figure 2.1 the main limiting factor would be bad lightning, vibrations and too high levels of noise. Lightning is a factor better improved separately, however noise and vibration are often coupled and a reduction in vibration will lead to a reduction in noise levels. A correlation between the influence of vibration and the ability to perform tasks such as reading is studied in [17, 18]. The results show that difficulty in reading is increased with increased vibration acceleration magnitude. Autonomous driving will not only influence the perceived comfort of vibrations, it will also change their type. Research looking for the desired driving style in autonomous driving found that minimising lateral acceleration, as well as adapting a more informative speed change profile is more positively accepted and viewed as safer [8, 19]. The optimised driving style also happens to lower load transfer as it’s dependant on lateral acceleration. The only vibration source left then is that coming from the road and is therefore a central theme for comfort in autonomous driving.
2.2.3 Measuring comfort

The methodology of measuring comfort is divided into different steps. This is done when developing industry vibration evaluation standards as well as when conducting company research. Firstly, the study participants answer questionnaires asking them to rate their experienced comfort from a ride on an interval scale. The comfort measured using such methods is subjective as it depends on the participants' personal interpretation. The study is then followed by the use of linear regression to objectify the obtained comfort values. When using the linear regression model, the comfort values would be objective as they are independent of who is experiencing the vibrations. In following these two steps, external validity is increased by means of drivers assessing comfort in normal usage scenarios. Internal validity is then further added by the use of objective comfort evaluation with proved relation between vibration and effect on perceived comfort. These steps are introduced in detail in the upcoming paragraphs.

Subjective comfort assessment

Subjective comfort assessment is done through experiments, where subjective ratings are gathered through a number of statements that are given relative to a predefined interval scale. Examples of comfort evaluation questionnaire that are often used for measuring comfort are the:

- SAE recommended practice for subjective rating scale for evaluation of noise and ride comfort characteristics related to motor vehicle tyres [20]
- Porter’s 7 scale comfort rating scale [21]

The participants are either few appointed experts in the field, or a large group of people with varying experience levels. The set-up can differ between real road
driving and simulator studies and each environment enforces different challenges. Demands for repeatability of the experiment however, make the driving simulator a more favourable mean as it allows for greater control of background variables.

The experimental plan of action often follows a common pattern. It starts with having the participants take a personal questionnaire to later assess if any external influences might explain misfitting data. The second step is to prepare the participants for the trial, this is often done through a presentation of the setup and questionnaire. After the introduction participants are led to take place in the driving simulator where they experience short sequences of vibration and give a rating for the experienced discomfort value after the ride is over. A final questionnaire can be added to make sure that all opinions are expressed.

Researchers and vehicle developers are faced with several issues when attempting to rate comfort subjectively. Some of the factors to be considered are not directly coupled to the ride yet may influence the test drivers perceived level of comfort. Test conditions such as humidity [22] and fatigue [15] can effect the passengers significantly. Predetermined opinions on comfort for example due to the aesthetics of the set up [23] can also function as an undesired anchor. Furthermore, different understanding of vocabulary used for rating between individuals can lead to misinterpretations. The advantage is however, as previously mentioned, stronger external validity level.

**Objective comfort assessment**

There are several main standardized methods for predicting comfort and hazards of body vibrations and shock. Two of which are the ISO 2631-1 standard, used primarily in Europe [24] and the British standard BS 6841 used in the united kingdom [25]. The difference between them is seen mainly in frequency weightings, averaging procedures and assessment methods [26]. In this thesis the focus will be on the ISO standard as preferred method.

### 2.2.4 Measuring vibrations

Vibration measurement is meant to record the body’s movement in reaction to a specific excitation. There are different vibration sensors that can be used for the task: accelerometers (piezoelectric), velocity sensor, proximity probes etc. Which one to choose depends on the relevant frequency range, the accepted accuracy of measurement and the facility of use for the purpose. In the next chapters the required frequency range is explained, and the use of these measurements to objectify comfort is elaborated on. In the method chapter, a review of the remaining two criteria are given as well as that of the experiment set-up and equipment description.

**Acceleration measurement location**

The measured vibrations’ magnitude, direction and frequency, vary depending on the measurement location. This is largely due to the changes in material stiffness along the travel way of the vibrations to each location. As the transmission function between two locations on the body and vehicle is usually difficult to define or calculate. It is critical that the measuring location is chosen at an area where
there is physical contact between the vehicle and passenger in order to eliminate the influence of variation in passengers’ anthropometric measurements.

In a driving scenario, the contact interfaces are defined in similar fashion for both the ISO and BS standard see Figure 2.3. The vibrations at each location is measured in a local coordinate system to correctly reconstruct the body’s movements. In Figure 2.3 three interfaces are presented: backrest, denoted with $b$ in the local coordinate system, ischial tuberosities- seat marked with $s$, and feet-ground marked with $f$. In autonomous driving the hand transmitted vibrations can be neglected as the drivers no longer holds the driving wheel.

![Figure 2.3: Suggested vehicle-to-body vibration measurement locations in ISO 2631-1 and BS 6841 standards](image)

**Objective comfort scaling**

Road vibrations can be felt by passengers in a vehicle on a wide frequency band extending from less than 1 Hz to more than 300 Hz [24]. Higher frequencies cause unacceptable noise while driving. Vehicle designers therefore make sure to limit the occurrence of vibrations above 300 Hz. For whole body vibrations the ISO and BS standards consider reducing vibrations with frequencies within the bandwidth 0.5 to 80 Hz sufficient to obtain a comfortable ride [27]. The resulting severity of discomfort due to exposure to vibrations varies depending on the frequencies present. Discomfort varies between motion sickness at 0 to 0.5 Hz, to increased overall discomfort between 4 and 8 Hz to disappear again close to the limit of 80 Hz. Frequency however, is not the only vibration component contributing to discomfort. The duration ($T$), amplitude $\hat{a}$ and direction of disturbance ($x,y,z$) given in the local coordinate system, also have a significant effect on experienced discomfort. To give an overall discomfort value, vibrations are considered alongside the frequencies and direction they appear at as well as the exposure time. Hence three different
values are calculated, weighted root mean square (r.m.s), weighted root mean quad (r.m.q.), and vibration dose value (VDV) using ISO and BS standard, respectively.

The root mean square is the value used by the ISO standard when only steady state vibrations are present and is defined as:

\[
\text{r.m.s} = \left[ \frac{1}{T} \int_{t=0}^{t=T} a^2_w(t) \, dt \right]^{1/2}.
\]

The weighted root mean quad is used to evaluate vibration signals in vehicles where also high magnitude accelerations with short duration i.e. shocks are present, r.m.q is defined as:

\[
\text{r.m.q} = \left[ \frac{1}{T} \int_{t=0}^{t=T} a^4_w(t) \, dt \right]^{1/4},
\]

where \( T \) is the measuring duration period, \( a_w \) the instantaneous frequency weighted acceleration. The VDV value is obtained by not averaging r.m.q with the measurement duration:

\[
\text{VDV} = \left[ \int_{t=0}^{t=T} a^4_w(t) \, dt \right]^{1/4}.
\]

The difference between the last two is that while r.m.q. is a good measurement for the overall severity of the vibration, it is still an average value meaning that it does not increase or decrease with the changing measurement duration. VDV on the other hand accumulates rather than averages the measured vibrations. Making it especially preferable to use in vehicle vibration evaluation as it incorporate the fact that vibrations tend to be more uncomfortable the longer the duration of exposure.

The crest factor e.g. the ratio between the highest peak in a signal and its r.m.s value is used to guide the use of r.m.s or VDV values. The maximum crest value for using r.m.s, given in the ISO standard, is nine [24] and for the British standard sex [25]. VDV should be used when the crest values is higher than the limit as to factor in the effect of the chock waves.

A mapping of vibration accelerations and their resulting comfort values is given in Table 2.2. The amplitudes are frequency weighted, as to better present the effect of certain frequencies on discomfort. The weighting graphs used by ISO and BS are shown in Figure 2.4.
The weighted vibrations are then given a discomfort value according to Table 2.2

<table>
<thead>
<tr>
<th>r.m.s weighted acceleration value</th>
<th>Degree of discomfort</th>
</tr>
</thead>
<tbody>
<tr>
<td>Less than 0.315 m/s²</td>
<td>Not uncomfortable</td>
</tr>
<tr>
<td>0.315 – 0.63 m/s²</td>
<td>A little uncomfortable</td>
</tr>
<tr>
<td>0.5 – 1.0 m/s²</td>
<td>Fairly uncomfortable</td>
</tr>
<tr>
<td>0.8 – 1.6 m/s²</td>
<td>Uncomfortable</td>
</tr>
<tr>
<td>1.25 – 2.5 m/s²</td>
<td>Very uncomfortable</td>
</tr>
<tr>
<td>Greater than 2 m/s²</td>
<td>Extremely uncomfortable</td>
</tr>
</tbody>
</table>

2.3 Artificial neural networks

Artificial Neural Networks (ANN) are inspired by the way the brain works. For every input via our senses, be it a picture, sound, or taste, a group of specific neurons in the brain are triggered. The neuron is therefore in simple terms an information processing cell. Every neuron is trained to give out an impulse only if its individual activation criteria are met. The neuron activation was discovered by Wiesel and Hubel during an experiment where they monitored a cats brain activity [28]. During their experiments the cat was directed to look at a line of light against a sheet of paper. The discovery was made when the line was put at a certain angle and position, and at the same time the transmitters monitoring the part of the animals brain activity started to pick up electric signals. These were later proven to be the signal output of specialized neurons for light detection corresponding to the specific part of the cats retinas which the light fell on.

Much the same way, ANNs use activation functions in order to couple input data with output data. Unlike decision trees, the connections are created with no need for pre-decided classification criteria between the input and output. They are namely
created arbitrary to fit as much of the data set as possible. To have a nonlinear system, several layers containing a finite number of neurons are put between the input and output, see Figure 2.5. The so called hidden layers in the middle are essential for creating fast adapting and robust systems. It comes however, on the cost to transparency, computational feasibility and sometimes overfitting. These problems can be solved using additional algorithms which are discussed later in Chapter 3.

![Neural Network Diagram](image)

Figure 2.5: Layout of a neural network with an input and output layer with 2 and 1 value, respectively, as well as 4 hidden layers.

### 2.3.1 Biological neurons

A neuron in a brain consists of many different parts working together to pass on information see Figure 2.6. The dendrites of the neuron include a range of both chemical and electrical synapses through which information is transferred into the neuron nucleus. The electrical synapses transfer the electrical signal coming from the presynaptic side directly to the postsynaptic side, i.e. the nucleus. Electrical signals create therefore direct, strong and unadjustable connection the transmitter and receiver of the signal which allows for shortening reaction time. The chemical synapses however, have a constructed synaptic cleft separating the presynaptic and the postsynaptic sides. In order for the information to flow in this case, the electrical signal is converted to a chemical one with the help of so-called neurotransmitters. These neurotransmitters cross the cleft and thereby transfer the information to the nucleus after which they are rapidly degraded to make room for new ones. In spite
of the more complex functioning of chemical information transfer, due to the rapid
decay of the neurotransmitters and their versatile roles i.e. they can either simulate
the postsynaptic or slow down such simulation, the process is still done rapidly and
with more precision.

![Composition of a brain neuron](image-url)

Figure 2.6: Composition of a brain neuron [29].

When the received level of electrical signal simulation in the nucleus exceeds a
threshold value, the cell nucleus of the neuron sends out an electrical pulse which
is then transferred on to other neurons. This is done by the means of the axon, an
extension of the nucleus cell. The axon leads eventually either to other dendrites or
to other kind of cells, thereby controlling them [29].

## 2.3.2 Artificial neuron

An artificial neuron consists of sex elements and follows the same procedure as the
brain neuron. The information is introduced to the neuron via the input $x_{1-n}$ which
is then multiplied with its respective weight $\omega_{ij}$. The weight is used to strengthen
or weaken the connection between two nodes and is therefore applied between every
two connected neurons. The input and weight product is then added to the bias
$b$ by means of a transfer function see Figure 2.7. The activation function is fed
the output of the transfer function and depending on this values relation to the
activation threshold an outgoing value is sent out to other neurons. Instead of
electrical signals, the artificial neuron uses integers as inputs and outputs.
2.3.3 Previous use of machine learning to objectify comfort

Machine learning and neural networks specifically have been introduced as a powerful classifying method for comfort in vehicles in a handful of studies. In [30] a linear regression method as well as a neural network were used to predict comfort values depending on different inputs.

The traditional modeling approach of comfort uses linear regression such as random forest [31] as was done in [32]. This method takes in data, which is then filtered according to preset rules and attributes until sufficient classification is reached. Random forest offers transparency in terms of the algorithms classification procedure, as it is namely defined by the code. The extraction of fitting classifying features
in the data requires however a lot of preparatory work and is time consuming [33]. It is suited for cases where the connections between subjects and their classes are necessary to know in detail. This is however not the case for comfort objectifying.

Deep neural networks with one or more hidden layers lack the transparency of the random forest algorithm. They enable however for detection of new connections previously unknown as well as for the construction of complex modelling systems and non-linear models with no need for previously determined classification rules.

In [30] one way analysis of variance (ANOVA) was compared to neural networks in term of seat comfort prediction. It was shown that artificial neural networks can achieve lower error values as well as derive a higher statistically significant difference between different chairs. In [34] Stammer explored the use of artificial neural networks for the objectification of comfort in passenger vehicles. Moser et al. [35] used dynamic time warping to study the effect of time sequences on comfort, this can be done using neural networks as well and is included in the objectives of this thesis.

2.4 Constructing a neural network

The universal approximation theorem states that a feedforward neural network with a single hidden layer containing a finite number of neurons will approximate any given function to an arbitrary accuracy [36]. While being a very powerful tool in theory, in practice, they can be computationally hard to train. Several algorithms however, can be used to increase the learning speed of the network [37]. In this part of the thesis, the constructed neural network topology is explained as well as a few well fitting activation and learning algorithms.

2.4.1 Feedforward neural network

A Feedforward network consists of an input layer, an arbitrary number of hidden layers and an output layer. Let $\chi$ be an arbitrary data set and let $Y$ be its (related) label set consisting of two classes $-1, 1$. The data set that is provided by $X$ and their respective labels in $Y$ are then divided into two subsets, i.e. a training and a validation data set. The ratio between the data points in each of the sets depends on the total number of data [38]. Training the NN refers to the initial procedure where the weights of each neuron are adjusted with each epoch (learning iterations). The goal is to have the weight value that will activate each neurons activation function only when needed. This is done repeatedly until the network has been trained using all available training data points.

2.4.2 Activation functions

The main purpose of an activation function is outputting a value between two boundaries, usually 0 and 1. Any arbitrary function that has an upper and lower limit for increasingly high and low values, respectively can be used. Different activation functions can be used at every layer, to influence computational and learning time as well as the format of the output. There are however, a set of activation functions which are especially suited for certain machine learning problems such as the
Softmax function when learning and predicting on multi-class datasets. It is used to compute the probability distribution of the output belonging to one of several classes. The output is a value between 0 and 1, where 1 refers to 100% chance and 0 to 0% chance. It is therefore mostly chosen as the activation layer of the output layer and defined as:

\[
\sigma(y)_i = \frac{e^{y_i}}{\sum_{j=1}^{K} e^{y_j}}. \tag{2.6}
\]

In this work a rectified linear unit (Relu) activation function is used for the hidden layers. Its output varies between 0 for all negative inputs in the range \([-\infty, 0]\) and 1 for any input bigger than zero. For the final layer the Softmax function is used which gives out the probability distribution over the output classes.

2.4.3 Weights

Setting the appropriate initial hyperparameters when constructing a neural network is crucial for fast learning. In simple artificial neural networks, these parameters are the weights and number of hidden layers. The rule of thumb for hidden layers is to keep adding them until the classification error stagnates. The weights however are best initialised with random values that differ from zero and in \([39]\) it is shown that weights drawn from a Gaussian distribution with a standard deviation of 0.2 lead to faster learning.

2.4.4 Training a neural network using backpropagation and gradient descend

Backpropagation \([40]\) is the preferred learning method for training artificial neural networks due to its accuracy and versatility. To train a network with \(\vec{x}\) as input and \(\vec{y}\) as output, a square cost function is defined as:

\[
L = \sum_{i=1}^{n} (\hat{y}_i - y(x_i))^2, \tag{2.7}
\]

where \(\hat{y}\) is the artificial network output. At the beginning, the cost function \(L\) will produce relatively large values. However, after a number of epochs the loss function drops and the accuracy increases. The cost function is reduced by changing the weights of the neurons according to their contribution to the error. This is done by calculating the derivative of the cost function for every neuron from the output layer back to the input, see Figure 2.8.
2.5 Other machine learning models

Deep neural networks are a versatile and powerful tool. But they require good, clean large quantities of data. This was not ensured for the available data for this work. Efforts are made to improve on the data which will be further explained in Chapter 4. Furthermore, since there are very few studies on the subject of classifying long vibration signals, one cannot exclude other methods to be superior to DNNs. Hence, a conclusive investigation would include testing different machine learning models to see if they reach a higher level of success. The following sections give an introduction to the classifiers investigated in this work. Unlike DNNs, the rest of the classifier models operate best with small input dimensions, a long acceleration measurement is therefore not appropriate to use. The suggested solution is to extract features from the data, see Section 2.5.6, lower the features’ dimensions either by means of transformation or picking out feature that cover the most variance and using the result as the input. First a quick introduction of the classifiers is given, followed by the theory of feature extraction in vibration signals.

2.5.1 K-nearest neighbors

K-nearest neighbor [41] is built on the assumption that similar objects cluster together in the feature space depending on feature similarity. For supervised learning it works by defining a number of neighbouring data points \( (k) \) that should be used to conclude the class of the one data point. Afterwards, the method of calculating the distance between each point is decided. For this work Euclidean distance is used. Running the algorithm would therefore group the signals into clusters, if the feature space is correctly chosen then the clusters should reflect different comfort levels. Figure 2.9 offers a visual representation of the clustering method in two feature dimension and using three neighbouring data points as \( k \).
2.5.2 Support vector machine

Support vector machine (SVM) models \[42\] aim to find a hyperplane in an N-dimensional space of features that best separates different classes of data. It does so by maximising the margin between the closest points and the hyperplane, so called support vectors. Figure 2.10 shows the intended usage of both the hyperplane and maximum margin to separate data points taken from two different classes.

![Figure 2.10: Example of an SVM classifier \[43\].](image)

2.5.3 Random forest

A random forest classifier \[31\] combines the outcome of several decision trees. Each tree is specialized in predicting classes based on a subset of features, see Figure 2.11. After running the algorithm, the predictions of each tree are summarised and the most popular prediction becomes the classifier’s prediction. This method is more
robust when dealing with uneven instances in each class but depends on a high
correlation within the classes to work well.

![Random Forest Simplified](image)

Figure 2.11: Example of an random forest classifier logic [44].

2.5.4 One-vs-Rest

The strategy behind a One-vs-Rest [45] classifier is to train one classifier per class
to predict the probability of a point belonging to that class and not in any of the
other classes, see Figure 2.12. It compares one class at a time with all the other
existing classes. When the algorithm is run, the class of the classifier which gives
the highest probability for the point to be in their class is chosen. For this work, the
One-vs-Rest was used with a Support Vector Classifier estimator with linear kernel
from the Python library SKlearn [46].

![One-vs-Rest Classifier](image)

Figure 2.12: Example of a one-vs-rest classifier set-up [47].

2.5.5 One-vs-One

One-vs-One [48] works in a similar ways to One-vs-Rest but instead of comparing
one class to all other classes, it compares as the name suggests only two classes at a
time, see Figure 2.13. Due to it comparing only two classes at a time, the number of
data points being compared is smaller than when using One-vs-Rest which can lead
to faster training. The same estimator was used for One-vs-One as for One-vs-Rest
for this work, i.e. a Support Vector Classifier estimator with linear kernel from the Python library SKlearn [46].

![Figure 2.13: Example of a One-vs-One classifier set-up [47].](image)

### 2.5.6 Feature extraction

Feature extraction was done for this work using TSFresh [49]. TSfresh is a library in python that can be used to extract hundreds of characteristics from time series. The library also contains functions which pick out the feature with the highest relevance for the a classification problem. This can also be done by picking out feature with the highest explained variance and using those as input for the classifier.

### 2.6 Model evaluation

Deep neural networks can be evaluated in many different ways depending on the problem at hand. For this work, the classifier needs to identify important parameters in the input and couple it to the output. The goal is to find a pattern in the data and to have as high a success rate in predicting comfort as possible. It is therefore more important to be often correct than seldom wrong. The models are therefore seen as better performing based on their accuracy. However, a high number of false negative prediction could mean that some error occurred while learning or that the data was not correctly divided between validation and training. Based on these arguments each model is evaluated with regards to its precision equation (2.8), recall equation (2.9) and F-score equation (2.10) value when run on a validation data set. To calculate the precision as well as other model matrices, the confusion matrix is used. A confusion matrix (see Figure 2.14) is a way of presenting the performance of a neural network. It is a two dimensional table where one dimension shows the predicted results of the neural network and the second the correct label.
Figure 2.14: Example of a confusion matrix for binary classification

<table>
<thead>
<tr>
<th>Predicted labels</th>
<th>Positive</th>
<th>Negative</th>
</tr>
</thead>
<tbody>
<tr>
<td>Positive</td>
<td>TP</td>
<td>FP</td>
</tr>
<tr>
<td>Negative</td>
<td>FN</td>
<td>TN</td>
</tr>
</tbody>
</table>

With TP being the number of true positive predictions for a certain class, FP the false positives, FN the false negatives and TN the true negative predictions. From this table several qualitative quantities can be calculated, for example:

- **Precision**: Used to evaluate how well the model is in avoiding labeling the wrong classes as belonging to a certain class.

  \[
  Precision = \frac{TP}{TP + FP} \quad (2.8)
  \]

- **Recall**: Measures how prone the model is to predict false negatives and thereby miss-classifying important classes.

  \[
  Recall = \frac{TP}{TP + FN} \quad (2.9)
  \]

- **F-score**: The harmonic mean of precision and recall.

  \[
  F\text{-score} = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (2.10)
  \]
Chapter 3

Data collection study

The work of this thesis builds on the work of a previous study where a driving simulator was used to measure comfort in different autonomous driving scenarios. In this chapter the objectives of the study are presented, then the set-up of the experiment and the used equipment is given, followed by a description of used vibrations and finally the results of the study.

3.1 Study objectives and method

The study \[50\] aimed to clarify the influence of different vibration directions and patterns on comfort during both attentive and inattentive driving. The participants were subjected to different sets of vibrations and their comfort levels after each set of vibrations was noted. To ensure a realistic feeling during the ride, vibrations were gathered while driving on real roads and later manipulated to movements in different directions inside a driving simulator, see Table 3.1. In total five different road segments were used to gather vibration data. The speed was kept constant at 70 km/h using cruise control. The measuring duration was kept to exactly 30 seconds. The five different street segments were chosen due to the naturally resulting vibrations and vehicle movements along the relevant directions, see Table 3.1.

From these five segments, 14 different vibration sequences were created for the simulator input. Where five were the measured vibrations with no alteration, 6 had increased amplitudes of the first road vibrations while keeping the frequency distribution unaffected. The last three used a ramp function to increase the vibration amplitude of one dynamic phenomena at a time, while minimising vibrations in all other directions to white noise.

Table 3.1: Road segments and their main resulting vehicle dynamic response

<table>
<thead>
<tr>
<th>Segment</th>
<th>Mainly occurring vertical dynamic phenomena</th>
<th>Intensity</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Pitch, roll and heave</td>
<td>Very high</td>
</tr>
<tr>
<td>2</td>
<td>Pitch and heave</td>
<td>High</td>
</tr>
<tr>
<td>3</td>
<td>Roll and heave</td>
<td>High</td>
</tr>
<tr>
<td>4</td>
<td>None (smooth road)</td>
<td>None</td>
</tr>
<tr>
<td>5</td>
<td>Pitch, roll and heave</td>
<td>High</td>
</tr>
</tbody>
</table>
3.2 Simulator

A complete simulation must serve all of the participant’s sensory channels. Meaning that a the participant must have visual, tactile as well as auditory perception to perceive the ride as realistic. The visual input was supplied using screens covering the available field of vision from the driver seat. Auditory perception is supplied by means of a video recording taken during the primary vibration measurements. Tactile perception for vibrations is ensured via the use of a dynamic simulator. Based on these criteria a dynamic high fidelity simulator with an original replica of the vehicle is used for an increased realism.

3.3 Vibration and comfort measurements

The participants of the data gathering study consisted of 79 BMW employees of which 60 males and 19 females with an average age of 35.7 years and a standard deviation of 12.6 years \[50\]. Each participant sat through the 14 road segments rides with no pauses except for a few seconds of smooth road rides between each segments. Fog was used to indicate the end of each segment and disappeared at the beginning of the next. The participants were asked to give a comfort value between 0 and 6 after each ride and could stop the ride if they felt too ill. The experiment was repeated in a randomised matter for inattentive driving scenario during autonomous driving. In a random pattern, the participants were asked to shift their attention from attentive to inattentive by asking them to fill out a survey during the ride or ride attentively by again asking them to focus on the road. During each ride, bodily vibration measurements were recorded. Acceleration measurements were taken at four different places; 3 on the participants bodies (head, chest and thighs) and one on the vehicle see Figure \[3.1\]. Specifications of the sensors are given in Table \[3.2\].
Figure 3.1: Measuring positions

Table 3.2: Sensor data

<table>
<thead>
<tr>
<th>Position</th>
<th>Sensor</th>
<th>DoF</th>
</tr>
</thead>
<tbody>
<tr>
<td>Head</td>
<td>Dytran 7556A1</td>
<td>6</td>
</tr>
<tr>
<td>Chest</td>
<td>PCB 3713E1125G</td>
<td>3</td>
</tr>
<tr>
<td>Thighs</td>
<td>PCB 3713E1125G</td>
<td>3</td>
</tr>
<tr>
<td>Seat</td>
<td>PCB 356A16</td>
<td>3</td>
</tr>
<tr>
<td>vehicle</td>
<td>Dytran 7556A1</td>
<td>6</td>
</tr>
</tbody>
</table>

Results from the study

The results of the study are 79 sets of data for 14 roads in two different awareness states recorded at four different position. Some measurements were excluded from the database due to various reasons such as participant being too distracted or feeling ill, faulty equipment, etc.
Chapter 4

Method

The method used for creating the results presented in this work can be divided into three stages. First the data was preprocessed and manipulated in different ways to optimise learning and prediction. The second part is the creation of the fitting model for the problem and the given data and the third part being the evaluation of the model as well as it’s optimisation.

In this chapter, the three parts are presented in the order mentioned above. A closer look at the data and its analysis as well as a comparison to the ISO standard is given. The problems faced when dealing with the dataset are explained and the used solutions are presented. Furthermore, different model topologies are presented in combination with the different data inputs. Finally the models are evaluated using fitting measures and metrics.

4.1 Analysis performed on the data

To ensure the usability of the data, statistical analysis were performed to investigate the correlation of the acceleration values and the discomfort ratings. This was done firstly by taking a closer look at the average accelerations per rating in the frequency spectrum using Fast Fourier Transform (FFT) \[51\]. The spread of both the ratings and average accelerations were plotted using box-plots to better understand the connection between the accelerations and the ratings. These plots were primarily used to determine which direction of measurement to use as input as using all three directions would lead to complexities in the learning of the models due to dimensionality.

4.2 Comparison to the ISO 2631-1

For comparison of the comfort classification of the ISO standard and the machine learning models, the comfort measurements are used as a reference base. The ISO 2631-1 comfort scale however, is ambiguous due to overlapping r.m.s intervals for different comfort levels. This makes a direct comparison difficult. As a workaround, two different comparisons were made, one using the highest values (see Table 4.1) and one using the lowest (see Table 4.2). In both tables, the number of comfort levels is kept to the original six levels. In Table 4.1 the highest ranges of original
table are used as the cut off for the different intervals and in Table 4.2 the lowest ones are used.

Table 4.1: Higher value ISO-2631-1

<table>
<thead>
<tr>
<th>r.m.s weighted acceleration value</th>
<th>Degree of discomfort</th>
<th>Numerical value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Less than 0.315 m/s²</td>
<td>Not uncomfortable</td>
<td>1</td>
</tr>
<tr>
<td>0.315 – 0.63 m/s²</td>
<td>A little uncomfortable</td>
<td>2</td>
</tr>
<tr>
<td>0.63 – 1 m/s²</td>
<td>Fairly uncomfortable</td>
<td>3</td>
</tr>
<tr>
<td>1 – 1.6 m/s²</td>
<td>Uncomfortable</td>
<td>4</td>
</tr>
<tr>
<td>1.6 – 2.5 m/s²</td>
<td>Very uncomfortable</td>
<td>5</td>
</tr>
<tr>
<td>Greater than 2.5 m/s²</td>
<td>Extremely uncomfortable</td>
<td>6</td>
</tr>
</tbody>
</table>

Table 4.2: Higher value ISO-2631-1

<table>
<thead>
<tr>
<th>r.m.s weighted acceleration value</th>
<th>Degree of discomfort</th>
<th>Numerical value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Less than 0.315 m/s²</td>
<td>Not uncomfortable</td>
<td>1</td>
</tr>
<tr>
<td>0.315 – 0.5 m/s²</td>
<td>A little uncomfortable</td>
<td>2</td>
</tr>
<tr>
<td>0.5 – 0.8 m/s²</td>
<td>Fairly uncomfortable</td>
<td>3</td>
</tr>
<tr>
<td>0.8 – 1.25 m/s²</td>
<td>Uncomfortable</td>
<td>4</td>
</tr>
<tr>
<td>1.25 – 2 m/s²</td>
<td>Very uncomfortable</td>
<td>5</td>
</tr>
<tr>
<td>Greater than 2 m/s²</td>
<td>Extremely uncomfortable</td>
<td>6</td>
</tr>
</tbody>
</table>

4.3 Preprocessing the data

The goal of this step is to facilitate the learning of the models and classifiers by finding a general way of creating a dataset that covers the variances of the original data while reducing it’s dimensions. Furthermore, statistical analysis was included in the preprocessing step to get a better understanding of the behaviour of the data. These are explained first, followed by feature extraction and methods used to speed up the learning process.

4.3.1 The dataset

Only a portion of the dataset acquired by the study was usable in this work as big insecurities remained in the measurement quality of three road segments. The final count landed on 716 different labeled measurement. Each measurement is cut to be 26 seconds (original 30 seconds) long with a sampling frequency of 1000 Hz. This concludes that each measurement produces an array of 26000 measured accelerations with one respective label. With the different comfort levels as classes, the distribution of labels in each class for every road segment is shown in Figure 4.1. The x-axis represents the different road segments used in this thesis, the y-axis represents the comfort rating on the scale 0 to 6 and the z-axis represents the number of times the road received each particular comfort rating.
Figure 4.1: Distribution of ratings for each road segment

Figure 4.2 represents the standard deviation of the comfort rating of each road segment as well as the mean comfort given for that road. The x-axis shows the road number and the y-axis represents the comfort rating.

Figure 4.2: Mean and standard deviation for each road segment

4.3.2 Reducing dimensionality

Dimensionality in time series comes from the recording length rather than the number of sampled variables. Longer sampling times however, do not always lead to better classifying results. For deep neural networks and other machine learning methods, the smaller the input dimension is, the faster the learning process will be. Due to the limiting computational capabilities of the hardware available for this work, there is a need to reduce the dimension of the dataset, from 26000 sample points to something more manageable. The dimension reduction has to be done in a way however, that preserves the information contained in the original data.

For time signals, the dimension reduction can be done in two different ways. By extracting physical features from the signal such as the mean value, frequency range, signal energy etc. or by creating new data describing the original data however with no physical units, so called meta data. Meta data can also be extracted from the physical features, further lowering the input dimension.
For this work the TsFresh library \[49\] was used to extract physical features from the vibration time series. Furthermore, the method of principle component analysis was used to create meta data of both the original signals as well as the TsFresh extracted features. These inputs were then used to train different models and their outcome was compared.

**Principle component analysis**

In order to reduce dimensionality while keeping as much of the data variance intact, the used method has to remove or replace linearly correlated variables. With that, any variables that are related and do not contain unique information about the signal are dropped. A principle component analysis (PCA) uses orthogonal transformation along the projection with the biggest variance in the data. The results of the transformation is the data along a new coordinate system where the first coordinate, so called first principal component, has the greatest variance in the data, the second the second biggest variance and so on \[52\]. The variance can also be summed and the final explained variation of all the produced principal components can be calculated. Thanks to this, the dimension of the input can be reduced significantly by choosing a total explained variation less than 100%.

The first Principle component is given by:

$$PC_1 = \arg\max_{\text{PC}} \left\{ \frac{PC^T \ast X^T \ast X \ast PC}{PC^T \ast PC} \right\} ,$$

(4.1)

where \( PC \) is an arbitrary vector used for projecting the data along a new coordinate system and \( X \) is the data matrix with a zero empirical mean along its columns. The rest of the components are given by:

$$PC_k = \arg\max_{\text{PC}} \left\{ \frac{PC^T \ast \tilde{X}_k^T \ast \tilde{X}_k \ast PC}{PC^T \ast PC} \right\} ,$$

(4.2)

where \( \tilde{X}_k^T \) is given by subtracting the \( k \)th components from \( X \):

$$\tilde{X}_k = X - \sum_{s=1}^{k-1} X \ast PC_s \ast PC_s^T$$

(4.3)

For this work, the function sklearn.decomposition. PCA from the library Sklearn \[46\] was used up to a total explained variation of 98%.

**4.3.3 Unbalanced classes in the data**

Uneven number of data instances in different classes poses a series of problems in machine learning. As the model learns, it will start to favour the class that increases its probability of making a correct prediction, leading to biased models. The methods for dealing with uneven classes in machine learning can be divided into two categories, one concerning further processing of the input data and the other dealing with commands and functions added to the model itself.
Manipulating the data

In the case managing the data to limit the effects of uneven classes there are three methods that can be used:

- **Upsampling**: where the dataset is increased by duplicating existing data until all classes are represented equally.

- **Down-sampling**: where the dataset volume is decreased to include only as many from each class as there is data points in the smallest one.

- **Augmenting the data**: e.g. the same concept as upsampling but instead of just duplicating the data, noise and other affects are added as to avoid creating a reversed biased model.

For this work, the up- and down-scaling methods are being used. The number of data points the different methods resulted in are shown in Figure 4.3. The x-axis represents the comfort levels and the y-axis represents the total number of data points corresponding to each comfort level in the dataset. As can be seen in the figure, using the up- and down-sampling methods results in an even number of data points per class while the original data had a very high number of data points corresponding to level 4 of comfort and very few corresponding to levels 0, 1 and 6. Furthermore, when using the down-sampling method the least number of classes were found in the class corresponding to the comfort level of 0. As this would result in a very small data base, this class was excluded and the next biggest class was used as reference. This resulted in six total classes for the outcome when using the down-sampling method.

![Figure 4.3: Number of data points in each class with different sampling methods](image)

Another way of handling this problem which was also motivated by the small size of the data set is creating a binary class distribution, where the data is divided into two classes, i.e. low and high comfort where low comfort includes all values from 0-3 and high the values from 4 to 6. This resulted in a perfectly even class distribution with each of the low and high discomfort classes having 358 data points.
As mentioned in the introduction of this thesis, comfort is a very complex feeling and many things can effect humans perceiving of it. Adding noise to the signals may have had unpredictable effects on the occupants of the vehicle which are impossible to predict. Since there is no way of eliminating the uncertainty that such augmentations would have an effect on the occupants ratings, it’s investigation is left out.

4.3.4 optimising the model

Model optimisation is a very important detail when it comes to getting the model to perform as well as possible. The idea behind it is to vary the different model settings such as number of layers, nodes and optimising functions until the optimum combination is found.

The biggest trap that a model can fall into while learning is overfitting. Which is when a model fits every point of the training data so well that it looses the ability to generalise to unseen data points. Overfitting often happens with the data has a lot of white noise present, the model then adapts to it as being as important an input as the main variables. This results in poor generalisation for new data that differs from the training data mostly in white noise rather than important features. Overfitting and biased learning can be combated however, using functions that arbitrary lower the confidence in the weights of certain nodes. Two methods commonly used methods are:

- Dropout: which is a function assigning a probability to not use the connection to a random set of nodes in each learning iteration. This leads to a more rigid model with less bias and overfitting

- Batch normalisation: normalises the activation for each layer so that no nodes activation gets too high or too low. Also allows each different layer to train more independently, thereby speeding up the learning process.

Both methods are used when creating the deep neural networks models. To keep the scope of the thesis focused. The dropout rate as well as the normalisation is kept the same for the different variations of model architecture.

4.4 Finding the best model

In order to find the most accurate model, several architectures of well-known machine learning classifiers were investigated. KNN, SVM, random forest, One-vs-Rest and One-vs-One were all tested with different numbers of classes and sampling methods. A summary of the different tested configurations is given in Table 4.3

<table>
<thead>
<tr>
<th>Artifacts</th>
<th>Range</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of classes</td>
<td>2; 6; 7</td>
</tr>
<tr>
<td>Resampling technique</td>
<td>Up; Down; Original</td>
</tr>
<tr>
<td>Dimension reduction method</td>
<td>PCA; orginal</td>
</tr>
</tbody>
</table>
Furthermore, since a stronger interest was put on the use of DNN:s for the classification given their versatility, several variations were investigated. However, due to the irregular and unpredictable relationship between the DNN parameters and success of classification more parameters had to be investigated in comparison with other models, see Table 4.4. The number of nodes in the table were used for the first layer of the neural network and the following hidden layers were created by multiplying it with 1.5 for the next layer, 2 for the following and 1 for the last hidden layer, depending on the number of layers used.

Table 4.4: Tested DNN models artifacts and input variations

<table>
<thead>
<tr>
<th>Artifact</th>
<th>Range</th>
</tr>
</thead>
<tbody>
<tr>
<td>Layers</td>
<td>1 - 4</td>
</tr>
<tr>
<td>Nodes</td>
<td>200; 500; 800; 1000; 2000</td>
</tr>
<tr>
<td>Number of classes</td>
<td>2; 6; 7</td>
</tr>
<tr>
<td>Resampling technique</td>
<td>Up; Down; Original</td>
</tr>
<tr>
<td>Dimension reduction method</td>
<td>PCA; Original</td>
</tr>
</tbody>
</table>

4.4.1 Creating the results

In order to test the effect of the chosen models artifacts (number of nodes and layers) as well as the preprocessing method of the input data on the effectiveness of the classification. A grid evaluation method was developed for creating, training and evaluating the models. A model was created, trained, tested and verified for each combination of the parameters.

The evaluation is done through splitting the data into two parts. Each with the same distribution of classes to avoid biased learning. The neural network is trained on the training set and uses the testing set to test the prediction accuracy after each epoch. The difference between predicted and actual output is then used to further optimise the model. The training and testing sets can be merged, shuffled and split again to obtained a low bias model, a method called K-fold [53]. Simultaneously the model is tested on the testing dataset in order to keep track of it’s performance. The models in this work are not validated with unseen data as the dataset was too small and all data points are needed during training to reach a high performance. The validation of the models will be done in future work using newly recorded data from a second experimental study.
Chapter 5

Results

In this chapter the result of the vibration data analysis as well as the comparison between the ISO-2631-1 predicted comfort levels and the ones measured in the study are presented. Furthermore, the most successful machine learning models’ accuracy are introduced.

5.1 Data analysis

The acceleration measurement from different locations in the body are transferred into frequency domain using FFT. The results are plotted and studied to explore any visible correlation between different roads’ average acceleration amplitudes and their corresponding comfort levels. In Figure 5.1, the averaged measurements for each comfort level recorded by the chest sensor from road number 2 are plotted within the most active range of 1-20 Hz. The right plot in Figure 5.1 magnifies the interesting range of 4-8 Hz as mentioned in section 2.2.4. Based on data plots and correlation calculations between average maximum acceleration amplitudes and comfort, it was decided to use the measurements in the z-direction (vertical axis) from the chest sensor, as these show higher correlation.

Figure 5.1: Averaged chest acceleration for road 2 grouped by comfort level.
Figure 5.1 shows acceleration values of up to 10 m/s$^2$ within a frequency range of 1 to 5 Hz. A clear correlation can be seen between acceleration amplitudes and levels of discomfort, where increasing acceleration values result in higher discomfort.

The left plot in Figure 5.2 shows the average of the 1000 maximum acceleration values of each passenger per road segment. Average maximum accelerations vary from values, as low as 0.45 m/s$^2$ for road segment 7 to values of up to 2.1 for road segment 2. Acceleration spreads show relatively small values, varying for different road segments from less than 0.2 m/s$^2$ absolute difference for road segment 7 up to 0.6 m/s$^2$ for road segment 2.

![Accelerations value distribution for each road.](image1)

![Value distribution of ratings for each road.](image2)

Figure 5.2: Value distribution of the average maximum acceleration and their ratings.

It can be clearly seen in this boxplot that the experiment is able to reproduce similar vibration accelerations for most passengers as the spread of acceleration values is relatively small. Variations in the real-life roads used for the original measurements are possibly the main reason behind the different spreads, whereas larger spreads are the results of larger local variations in the road surface.

The right plot in Figure 5.2 depicts the distribution of comfort values for each road segment. The spread of ratings is large for most roads ranging from 3 levels of rating to covering the whole rating scale of 6. Average comfort rating varies between 2 for road segments 4 and 7 and 5 for road segments 1 and 11.

The large spread indicates that the subjective feeling of comfort is very different for each individual passenger. Particularly difficult to assess are road segments 3, 7, 8 and 9. Road segments with higher ratings generally exhibit higher average maximum acceleration values when comparing the results to the plots in Figure 5.2.

Variables unaccounted for could be the reason behind the high variation in acceleration measurements. Confounder variables such as the participants weight, height and body composition, etc, are predicted to have influenced the transmitted vibrations from the vehicle. Since these variables were not recorded, it was not possible to estimate their effect on the accelerations. The varying weight of the passengers in particular, could be an explanation to the distribution seen in the left plot in Figure 5.2.
The quality of the data can be further discussed in regards of how well it was able to capture the set of variables that influence comfort. This can be seen in the average and variance of comfort values for each road see right plot in Figure 5.2 and Figure 4.2. Smaller variances can be seen as indicators of agreement within the participants on the comfort level of the road, meaning that the variables that influence comfort as suggested in Table 3.1 are indeed the main cause of certain levels of discomfort. However, for the data used in this work the variances were relatively large, with an average standard deviation of 1.38. Due to the standard deviation being fairly similar for all 11 roads, its high value is most probably not an indicator that the roads were difficult to assess but rather that there were underlying differences in the way participants perceived vibrations which were not accounted for.

In summary, while the correlation between higher acceleration values and higher discomfort ratings is visible, the spread of ratings per individual road segment hints at inconclusiveness in the data.

5.2 Comparison with ISO 2631-1

The comparison between the predicted discomfort levels using the ISO 2631-1 standard and original values observed in the study is given in Figure 5.3. The first graph shows the re-scaled predicted comfort levels, using the lower boundary table shown in Table 4.2, the second graph depicts the predicted values calculated with the higher boundary graph shown in Table 4.1. Comfort values from the study are plotted in the last graph for comparison. The predicted comfort rating values obtained by the ISO-2631-1 are re-scaled to match the range used in the study. The standard does not predict discomfort levels above 2 when using the higher range comparison and 1 when using the lower ranges.
As shown in the first graph, the ratings vary between 4 and 3 using the higher boundary and between 3 and 5 when using the lower boundary, whilst study values range from 0 to 6. The original data exhibits strong fluctuations in comfort ratings whilst the ISO 2631-1 predicted data is relatively steady. The small spread of values in the upper graph is a result of larger r.m.s acceleration value bin sizes used for each level of comfort. This results in a less sensitive comfort prediction. Using the lower boundary table leads to increased sensitivity while maintaining the steady prediction output making it a better approach to predict comfort ratings. The comparison between the ISO 2631-1 standard and the study participants comfort levels show that it is sufficiently accurate to use the standard for predicting discomfort trends, e.g. if its rising or falling, on a broad and general level. The overlapping intervals for the different levels and the way they are interpreted can however, have a great effect on the outcome of the investigation and.

5.3 Machine learning models

In this section, the results of the machine learning classifiers are presented. Different variations of deep neural networks were investigated, totaling 280 different models reaching varying degrees of success. Following, plots of the models with the highest
precision scores for each examined number of classes and their corresponding dimension reduction method are given. The results for predicting seven levels of comfort are summarised in Table 5.1 and for two levels of comfort in Table 5.2.

In order to investigate the performance of the models based on each class, the confusion matrix of the best performing models are plotted in the following. The plots combine the predicted labels (x-axis) given by the model and the correct label (y-axis) given by the dataset where the correctly predicted data points per class are given by the matrix’ diagonal, see Section 2.6. As precision per class results vary independently of using PCA dimensional reduced data, original data plots are used exclusively for the comparison.

Table 5.1: Summary of best performing DNN with seven comfort levels.

<table>
<thead>
<tr>
<th>Input type</th>
<th>Sampling method</th>
<th>Number of layers</th>
<th>Number of nodes</th>
<th>Precision</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ts-Fresh</td>
<td>Up</td>
<td>4</td>
<td>200</td>
<td>49.5%</td>
</tr>
<tr>
<td>Ts-Fresh</td>
<td>Down</td>
<td>1</td>
<td>500</td>
<td>41.9%</td>
</tr>
<tr>
<td>Ts-Fresh</td>
<td>None</td>
<td>1</td>
<td>2000</td>
<td>24.7%</td>
</tr>
<tr>
<td>Ts-Fresh + PCA</td>
<td>Up</td>
<td>2</td>
<td>1000</td>
<td>51.2%</td>
</tr>
<tr>
<td>Ts-Fresh + PCA</td>
<td>Down</td>
<td>2</td>
<td>200</td>
<td>42.9%</td>
</tr>
<tr>
<td>Ts-Fresh + PCA</td>
<td>None</td>
<td>1</td>
<td>500</td>
<td>26.4%</td>
</tr>
</tbody>
</table>

The sampling method with the highest precision used to handle class imbalance is the up-sampling method with 51.2%, followed by down-sampling with 42.9%. The lowest performance is achieved by leaving the class distribution unchanged, resulting in a precision score of 26.4%. The differences in precision score using the investigated sampling methods are further explained with figures of each method’s confusion matrix in the continuation of this chapter. The number of nodes, resulting in the best performance vary between 200 to 2000, covering the complete investigated range. Furthermore, 50% of the models were able to achieve highest performance using a single layer followed by 2 layers. No clear correlation could be observed between number of layers or nodes and precision. Using PCA to decrease the input dimension lead to a slight increase in precision, approximately 1.5% higher precision on average in comparison to the same model topology with original data size and features. The reason can be found in the noise reduction capabilities of the PCA data transformation, as disturbances are dropped when calculating the new set of features leading to increased precision.

Figure 5.4 shows the confusion matrix of the best performing neural network trained on up-sampled class distribution data. The best predictions are made for the classes 0, 1 and 6 with prediction scores between 60% and 100%, while classes 3 and 4 have the lowest scores with 19% and 26%, respectively. The generally high precision of the up-sampling method is a result of the increased precision for the most duplicated classes, e.g. 0, 1 and 6. Consequently, the improved performance for said classes is achieved by increased similarity in the data, resulting as a consequence of duplicating data points in the originally smaller classes. The model is trained more often on the same data and is biased to adjust its weights to better match said classes. A more even distribution of classes could have resulted in less data duplication and therefore lower bias towards small classes. On the other hand,
originally larger classes, e.g. 2-5, continue to contain a large amount of variance that the model is unable to properly learn, leading to low class-precision values.

Figure 5.4: Confusion matrix for original data and up-sampled class distribution

Figure 5.5 shows the confusion matrix of the best performing neural network trained on down-sampled class distribution data. The performance is optimised for classes 1, 4 and 6 where the model is able to correctly predict at least 50% of the data points belonging to these classes, showing values of 67%, 62% and 50%, respectively. The lowest performance is observed for classes 2, 3 and 5 with 42%, 23% and 8% precision, respectively. As the down-sampling method represent a dataset with an even class distribution, the model is not able to maintain a bias towards any particular class as a results of its size. The differences in scores are therefore argued to be the result of large feature variance in input data for certain classes rather than the sampling method itself. The generally lower precision when compared to up-sampling is a consequence of the lack of priory mentioned pure duplication in the data for the up-sampling method, as well as the smaller size of the dataset. Major parts of the data were excluded using down-sampling.
Figure 5.5: Confusion matrix for original data and down-sampled class distribution

Figure 5.6 shows the confusion matrix of the best performing neural network trained on data with original class distribution. The lowest precision score is obtained in class 0 and the highest score was obtained for class 6 with precision scores of 0% and 62% respectively. Furthermore, classes 1, 3 and 4 are predicted with a precision higher than 20%, while class 2 and 5 achieved precision scores of 10% and 6%, respectively. The original distribution of the data was strongly uneven, as class 4 was 17 times bigger than class 1 with a difference of 158 data points. The low number of data points explains the 0% precision for class 0 but does not account for the relatively high precision of classes 6 and 1 as they are both smaller than class 4 which only achieved a precision of 28%. These differences in precision scores are therefore a result of feature input variance within the classes and are unrelated to the sampling method.

Figure 5.6: Confusion matrix for original data and class distribution

A final conclusion can consequently be made regarding the easiest and most
complicated classes to predict. Classes with constantly high precision scores are 1
and 6, while classes 2, 3 and 5 have the lowest precision scores, independently
of the used sampling method. This can be directly linked to the measurement
scale used in the study and the participants ability to objectify their perceived ride
comfort. The results indicate that the participants are better able to classify which
physical movements of the body result in the most and least amount of discomfort.
Furthermore, they are less able to precisely distinguish between the middle stages
of the discomfort spectrum. Label 0 and 4 are exceptions of the given conclusion.
The low precision for label 0, with the exception of the influence of class bias in the
up-sampling method, is due to its low representation in the dataset as it is scarcely
used as comfort level. This level would correspond to a static vehicle ride which was
never the case as all rides had some level of vibrations. Label 4 falls in the middle of
the comfort scale and could have been selected when the participant is uncertain as
to how to classify the ride. This would explain why it is the most popular class but
is equally probable to be classified as class 3 or 5 especially when using the original
data distribution.

Table 5.2 shows the best DNN performance on binary labeled data. Transform-
ing the data labels into binary classes with low = 0 and high = 1 discomfort resulted
in equally distributed classes. The usage of sampling methods to handle uneven class
distribution was therefore not required and only the original distribution is used in
both model testing and training.

<table>
<thead>
<tr>
<th>Input type</th>
<th>Sampling method</th>
<th>Number of layers</th>
<th>Number of nodes</th>
<th>Precision</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ts-Fresh</td>
<td>None</td>
<td>2</td>
<td>800</td>
<td>73.7%</td>
</tr>
<tr>
<td>Ts-Fresh +</td>
<td>None</td>
<td>2</td>
<td>1000</td>
<td>75.4%</td>
</tr>
<tr>
<td>PCA</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The highest performing topology using original feature input reached a precision
score of 73.7%, while reducing the feature input size using PCA led to a slightly
enhanced score of 75.4%. This follows the trend seen earlier with the PCA feature
input enabling the model to achieve a few percentage higher precision score and can
be attributed to the noise reduction in means of dropping negligible features. These
results are significantly higher in comparison to the multi-label classification results.
Due to the use case at hand, predicting comfort in autonomous driving, it is more
valuable to achieve maximum precision scores when predicting uncomfortable rides,
making the binary classification the most useful method for predicting comfort.

The confusion matrix plots of the best performing models using binary classes are
given in Figures 5.7 and 5.8 to better understand the models ability to differentiate
between classes. The x-axis shows the predicted label and the y-axis shows the true
class label, as given by the dataset. The class normalised confusion matrix is also
plotted next to the value count confusion matrix for easier performance comparison
between classes.

Figure 5.7 shows the classification results of the best performing model trained
with original feature input. Class 0 was predicted with a 76% precision, slightly
higher than class 1 with 71% precision score. The model was better able to predict the lower discomfort class, indicating that the more comfortable rides are easier to classify than rides that lead to higher discomfort. The difference is however not large enough to draw decisive conclusions of what class is easiest to classify. Furthermore, the percentage of falsely classified uncomfortable rides is 29%, this is a considerably high recall score and should be further improved in the future.

Figure 5.7: Confusion matrix for original input data and binary class distribution

Figure 5.8 shows the classification results of the best performing model trained with PCA reduced feature input. Class 0 was predicted with a 80% precision, significantly higher than class 1 with 71% precision score. While this model has the highest overall precision score, it does not perform equally as well for both classes. Other models were able to achieve a more even precision score for both classes but had a lower individual as well as overall precision and are therefore not included in this report. Furthermore, the trend of better prediction being made for the lower discomfort class is visible here as well, further solidifying the theory that the more comfortable rides are easier to classify than rides that resulting in higher discomfort. The reduction in feature space seems to have a positive effect only on the low discomfort class, leaving the higher discomfort classification precision unchanged in comparison to the original input.
Comparing both class distributions e.g. 7-classes ranging from 0 to 6 and binary classes with only high and low discomfort levels shows a clear advantage in terms of increased precision with the binary discomfort scale. However, the dataset used in this study was not collected using binary class labels. There is therefore an unknown level of uncertainty as to how well the divide of classes into high and low would fit that given by the study participants. On the other hand, binary discomfort comes with a loss of information in terms of precise discomfort level. For introductory application in the field of autonomous driving, this loss can be accepted, however, with increasing customer demand on comfort, multi-class distribution prediction methods have to be introduced. As priory stated, multi-class discomfort modeling needs further improvement before being used in real world driving scenarios.

5.4 Other investigated models

Different standard classification models were investigated alongside deep neural network to find the overall best performing model. The five trained and tested models are KNN, SVM, random forest, One-vs-Rest as well as One-vs-One, see chapter 4. For a precise comparison between the classifiers and the deep neural network, the models were trained on the same datasets in different variations, using the two different sampling methods as well as original class distribution. The input was further varied by training and testing with both PCA reduced feature input and original feature input. In the following section the results of the best performing models in each variation of data are presented. To assess the models performance as effectively as possible the precision scores as well as recall and F-score are plotted for each combination of input, see Section 2.6. The relevant metrics are plotted in bar plots with the x-axis denoting the classifier type and the y-axis denoting the average achieved metrics for all data classes.

In this section the three studied classifiers random forest, One-vs-Rest and One-vs-One are referred to as combination classifiers. Combination classifiers are setup of several different classifiers which are trained on a subset of features (random forest)
or a class in the data (One-vs-Rest and One-vs-One). When predicting labels for the test data, these classifiers output the class belonging to the classifier which is able to match the input data to its past training data with the highest probability, see Chapter 4. Hence, combination classifiers are less sensitive to outliers in comparison to KNN and SVM which try to find a global optimal solution for all points in the data rather than focusing on a subset of it at a time. In the discussion of the following results the classifiers will be assessed in comparison to each other as well as in groups of combination versus non-combination classifier groups. In the discussion of the different results the precision score is exclusively used to compare the performance of the classifiers. Recall and F-score values are discussed when the values differentiate significantly between the classifiers.

Figure 5.9 shows the three metrics results for classifiers trained with up-sampling 7-class distribution output and original dimension Ts-Fresh features input. The plotted metrics show a high classification precision (over 50%) for the combination classifiers, e.i. random forest, One-vs-Rest and One-vs-One, with One-vs-One outperforming the other classifiers reaching a precision score of 59.8%, followed by One-vs-Rest and random forest with 57.8% and 52.2%, respectively. The classifiers, KNN and SVM performed in comparatively poorly on this dataset reaching a precision scores of only 38.2% and 30.9%, respectively.

The difference in performance between the combination classifiers and KNN and SVM classifiers in is traced back to the increased number of outliers. This is a natural consequence of the up-sampling method, as with added duplicated data the total number of outliers is increased. As seen earlier with the performance of the neural network models on different classes, the most difficult classes to classify are the ones in the mid-range of the discomfort spectrum. This is further visible in Figures 4.1 and 4.2, as most comfort ratings are clustered in the mid range and each road has a high standard deviation of at least 1. In conclusion, the duplication of the data points in smaller classes consequently lead to an increase in outliers, resulting in a decreased performance for outlier-sensitive classifiers SVM and KNN. Combination classifiers on the other hand were better able to handle the added outliers and were better able to take advantage of the increase in data points given by the up-sampling method.
In Figure 5.10, the three evaluation metrics of the five investigated models are plotted for PCA reduced dimension Ts-Fresh features input and up-sampled 7-class distribution output. The plotted metrics show a relatively high classification precision specifically for the combined classifiers as expected when using the up-sampling method. The best performing classifier is instead the random forest classifier with 59.1%. Furthermore, using PCA reduced input yielded an improved precision for both KNN and SVM reaching 39.2% and 34.9%, respectively. The results for the One-vs-Rest and One-vs-One show however a decrease in all three metrics to 47.2% and 52.5%, respectively.

The recall and F-score remain high in both Figure 5.9 as well as Figure 5.10, indicating the lack of bias in the models for the biggest classes. For the SVM classifier in Figure 5.9, the difference between recall and precision is the biggest which is also reflected in a lower F-score. This is seen as an indication that the model is prone to miss-label classes from least popular classes as belonging to the most popular classes, i.e. the model is biased towards the bigger classes. Furthermore, it is worth mentioning that the One-vs-One classifier achieved a 10% performance increase in comparison to the best performing deep neural network while maintaining a high recall value, indicating the models ability to distinguish between the different classes to a very high degree.

Using PCA reduced input shows an improved performance for all of KNN, SVM and random forest classifiers. There is however a decrease in performance for One-vs-Rest and One-vs-One when using PCA reduced input. The reason behind this drop in performance is not clear and not expected as both classifiers use a support vector classifier estimator with linear kernel, which is the same method used for the SVM. The difference in performance seems to be dependent on the topology of the One-vs-One and One-vs-Rest. Figure 5.11 shows the three metrics for the five classifiers with original dimension Ts-Fresh features input and an down-sampled 6-class distribution output. The plotted metrics show a relatively low classification precision for all classifiers as the down-sampling method combined with the high
number of classes renders relatively few data points for training per class. The highest precision was achieved using random forest and reached 39.2% followed by SVM and KNN with a precision score of 36.9% and 37.6%, respectively. The lowest precision score of 36.4% is achieved by One-vs-One and One-vs-Rest.

Figure 5.11: Classifier metrics with original input dimension and down-sampling class distribution

Figure 5.12 shows the three evaluation metrics of the five investigated models for PCA reduced dimension Ts-Fresh features input and down-sampled 6-class distribution output. The plotted metrics show relatively high classification precision with no clear advantage for the combined classifiers random forest, One-vs-Rest and One-vs-One with precision scores of 35.1.8%, 36.3% and 34.8%, respectively. SVM exhibits a higher precision score of 39.1% followed by KNN at 36.3%. Similar to original input performance, all classifiers performed equally well.

F-score and recall are comparably high for all classifier, with the SVM showing a lower recall value similar to the case with the up-sampled dataset. This is seen as an indication of the SVM classifier’s inability to reduce false negatives when trained on original feature data. With even distribution of data over classes, there is no advantage to be found when using a combination classifier (random forest, One-vs-Rest and One-vs-One), compared to other types of classifiers which is evident in the performance results.
The dimension reduction with PCA yield better precision scores in comparison to original input dimension only for SVM classifier, allowing an increase of 2.2%. Furthermore the recall score is higher with reduced dimension, indicating that using PCA in combination with this classifier on an even class-distributed dataset is a successful combination.

The usage of the down-sampling method eliminated the risk of duplicating outliers, this is visible through the relative increase in performance for SVM to match that of combination classifiers. Due to the decreased dataset size however, all models performance is decreased, with only SVM performing better on this set that on the up-sampled set. PCA dimension reduction did not yield better results for any of the classifiers with the down-sampling method, indicating the importance of sampling method and dataset size in comparison to feature dimension. Figure 5.13 shows the three evaluation metrics of the five investigated models for original Ts-Fresh features dimension input and original-sampled 7-class distribution output. The plotted metrics show low precision scores for all classifiers with the best performance reaching 25.89% for the random classifier followed by SVM and One-vs-Rest 25.11% and 23.62%, respectively. The two lowest performances were seen for classifiers One-vs-One and KNN with with precision values reaching 23.26% and 22.67%, respectively. While F-score and recall are comparatively similar for all classifiers, SVM exhibits a significant reduction in recall reaching a score of 14.79%.
The major reason for the reduced performance value is found in the strongly uneven class distribution in combination with the small dataset size. This results in a smaller group of training data points for several classes, making their prediction especially difficult. The findings are consequent with the ones seen for deep neural networks, see Figures 5.5 and 5.6, leading to the conclusion that original class distribution is the least favorable when seeking high precision classification. Figure 5.14 shows the three evaluation metrics of the five investigated models for PCA dimension reduced Ts-Fresh features input and original 7-class distribution output. With this set of input, the best performing classifier, i.e. One-vs-Rest was able to reach a precision score of 28.3%, followed by SVM and One-vs-One with precision scores of 24.8% and 23.4%, respectively. KNN and random forest show the lowest performances with precision scores of 22.25% and 20.43%, respectively.

Using PCA reduced dimension resulted in no significant change regarding the performance of KNN, SVM and One-vs-One, while reducing the ability of random forest to correctly classify the signals and slightly improving the results achieved by One-vs-Rest. As the influence of PCA on the One-vs-One and One-vs-Rest should have been relatively comparable, it is not deductible why a change occurs. When
looking closer at the influence of PCA reduction on the performance of Random forest a positive trend can be observed using up-sampling and a negative one using down-sampling. This leads to difficulties in predicting the influence of input dimensions in combination with different sampling methods.

The overall performance on original class distribution is significantly lower in comparison to up- and down-sampling. This trend is compliant with results obtained via deep neural networks on uneven class distribution in combination with PCA reduced input dimension, as priory shown.

Maintaining the original data class distribution preserves the relatively high total number of data points. The uneven distribution however results in a higher bias for classes with many data points. As these classes are few in this work the classifiers’ overall precision score is lowered significantly by poor learning of the classes with fewer data points. Comparing the results to a random probability of picking a class from the 7 classes (14.3%), the question can be raised whether said models are significantly superior in predicting the proper outcome. In Figure 5.15 the three evaluation metrics of the five investigated models are plotted for original dimension Ts-Fresh features input and original binary class distribution output. The best performing classifier is the random forest classifier, reaching a precision score of 75.4% followed by both KNN and SVM with precision scores of 73.2% and 73.1%, respectively. The least performing classifiers are the One-vs-One and One-vs-Rest with precision scores of 73.39% and 73.35%, respectively.

![Figure 5.15: Classifier metrics with original input dimension and original class distribution](image)

As One-vs-Rest and One-vs-One train inner classifiers on each class in the data, they are expected to reach a very similar precision score at the end of training. This becomes apparent when analysing the results of both classifiers, reaching comparatively similar results. As previously shown via deep neural networks, higher precision scores are mainly achieved for the low discomfort class, while lower precision scores were found for the high discomfort class, see Figures 5.7 and 5.8.

It can be argued that the same distribution in performance per class is the reason for relatively low performance values of One-vs-One and One-vs-Rest, as these classifiers combine the prediction accuracy of each sub-classifier. Random forest on the other hand, is able to achieve a higher precision as it uses features rather than classes to train on. The performance of KNN and SVM is consistent
with that of the previously presented neural network results. Both show a low level of bias indicated by a high recall score. In Figure 5.16 the three evaluation metrics of the five investigated models are plotted for PCA reduced dimension Ts-Fresh features input and original binary class distribution output. The best performing classifier is the random forest classifier, reaching a precision score of 75.97% followed by both KNN and SVM with precision scores of 73.2% and 73.1%, respectively. The least performing classifiers are the One-vs-One and One-vs-Rest with precision scores reaching 73.41% and 73.32%, respectively.

No significant improvement can be seen with the combination of PCA reduced input and original input features and dimension. Similar trends regarding performance differences between classifiers as observed for classifiers trained on original input, can be observed in the present case.

The performance on binary class distribution was similarly high for both the investigated deep neural networks and different classifiers. The same argument put forward regarding the application of the binary class distribution as mentioned previously apply here as well. The data used in this thesis was not meant to be further transformed to a binary class distribution. New data has to be gathered on this scale for improved external validity.
Chapter 6

Summary and Outlook

The present work investigates the possibility of deriving a machine learning model that connects vibrations omitted by a passenger’s body during a drive in autonomous driving mode and the comfort level that said passenger experiences. It aimed at doing so by first investigating the data and applying different methods for reducing dimensions, followed by handling uneven class distributions with means of up- and down-sampling. The dataset was additionally divided into high and low discomfort classes. Said data wrangling was followed by constructing and investigating both deep neural network models as well as regression models to see if any specific machine learning approach is better suited for the prediction task.

Priorly, experimentally gathered, 6 DoF simulator data on passenger comfort was used as training and testing datasets for deep neural networks (DNN), KNN, SVM, random forest, as well as One-vs-One and One-vs-Rest classifiers. The data consisted of acceleration measurements from different body parts with corresponding labels for discomfort ranging between 0 to 6. The Python library Ts-fresh was used for feature extraction from the data and said features were used throughout this work. PCA feature selection method was applied for reducing input data dimension and compared to original feature data. All models were investigated using a binary and 7-class distribution. Uneven class distribution handling using up- and down-sampling was performed on 7-class distributed data, exclusively.

All investigated models showed clear ability to learn and find patterns in the data with a significantly higher accuracy than that seen when randomly guessing. With the exception of the performance of random forest, One-vs-One and One-vs-Rest on up-sampled class distribution, no clear advantage could be found when using DNN compared to other investigated classifiers. Random forest showed a consistent comparatively high precision score for different sampling methods, where other models’ performance fluctuated based on input data. KNN and SVM’s performance saw a decrease with up-sampling as the number of outliers in the data increased, while One-vs-One and One-vs-Rest were simultaneously reaching high precision scores of up to 57.8% and 59.8%, respectively. These high scores were achieved greatly due to the increased dataset size. In combination with down-sampling and original class distribution the performance of all classifiers and DNN:s decreased as the dataset size got smaller. The up-sampling method showed best performance with values reaching up to 60%, followed by down-sampling reaching 43% while leaving the distribution unchanged enabled reaching a precision score of 28.3%. Using PCA lead
in most cases to up to 2% increase in precision and faster learning time with the exception of the influence of up- and down-sampling on One-vs-Rest classifier.

Binary class distribution showed a consistently higher classification precision in comparison to 7-class distribution reaching a precision score of 75.4% using deep neural networks. This increase in performance was greatly due to the increased of the data points per class ratio. This is accompanied with a loss in granularity. Within the 7-class distribution, classes 1 and 6 achieved the highest precision for most DNN:s while classes 2, 3 and 5 achieved the lowest. This was taken as an indication that discomfort of rides which fall on the far ends of the spectrum were easier to quantify compared to rides with an average discomfort level. Using binary class distribution, low comfort was predicted with higher precision in comparison to high comfort, further indicating that certain vibrations are easier to predict than others.

In conclusion, it can be said that with the use of up-sampling methods in combination with machine learning algorithms, passenger discomfort can be predicted to a high degree of precision. To make any decisive conclusions however, a larger testing sample of data should be used and confounder variables should be taken into account. New data should be gathered on a binary discomfort scale to ensure the fitness of the methods suggested in this work.

Suggested future work includes using collaborative filtering methods to predict the drivers individual comfort level. For this, more data has to be gathered. It is therefore also suggested to include investigating the usage of vibration measurements applications on smartphones for specific roads, followed by a questionnaire to expand the data set with both vibration measurements and labels. Furthermore, reinforcement learning can be used to improve on the model precision and lower the influence of outliers through means of continuous learning with bigger datasets.

This work used different methods for handling uneven class distribution, including dividing the data into high and low comfort classes. This resulted in high classification precision but lead to decreased external validity of the model as a whole. Next steps should include the a repeat of the data gathering study using different discomfort value scales, including a binary value scale.
Bibliography


[29] David Kriesel. A Brief Introduction to Neural Networks, Figure 2.3. 2005.


52


