A DATA-DRIVEN APPROACH TO REMOTE FAULT DIAGNOSIS OF HEAVY-DUTY MACHINES

Tomas Olsson

2015
A DATA-DRIVEN APPROACH TO REMOTE FAULT DIAGNOSIS OF HEAVY-DUTY MACHINES

Tomas Olsson

Akademisk avhandling som för avläggande av teknologie doktorsexamen i datavetenskap vid Akademin för innovation, design och teknik kommer att offentligen försvaras tisdagen den 10 november 2015, 14.00 i Omega, Mälardalens högskola, Västerås.

Fakultetsopponent: Docent Amy Loutfi, Örebro universitet

Copyright © Tomas Olsson, 2015
ISSN 1651-4238
Printed by Arkitektkopia, Västerås, Sweden
A DATA-DRIVEN APPROACH TO REMOTE FAULT DIAGNOSIS OF HEAVY-DUTY MACHINES

Tomas Olsson

Akademisk avhandling

som för avläggande av teknologie doktorsexamen i datavetenskap vid Akademin för innovation, design och teknik kommer att offentligen försvaras tisdagen den 10 november 2015, 14.00 i Omega, Mälardalens högskola, Västerås.

Fakultetsopponent: Docent Amy Loutfi, Örebro universitet

Akademin för innovation, design och teknik
Abstract

Heavy-duty machines are equipment constructed for working under rough conditions and their design is meant to withstand heavy workloads. However, the last decades technical development in cheap electronically components have lead to an increase of electrical systems in traditionally mainly mechanical systems of heavy-duty machines. As the complexity of these machines increases, so does the complexity of detecting and diagnosing machine faults. However, the addition of new electrical systems, such as on-board computational power and telematics, makes it possible to add new sensors that measure signals relevant for fault detection and diagnosis, and to process signals on-board or off-board the machines.

In this thesis, we address the diagnostic problem by investigating data-driven methods for remote diagnosis of heavy-duty machines, where a part of the analysis is performed on-board the machine (fault detection), while another part is performed off-board the machine (fault classification). We propose a diagnostic framework where we use a novel combination of methods for each step in the diagnosis. On-board the machine, we have used logistic regression as an anomaly detector to detect faults that will lead to a stream of individual cases classified as anomalous or not. Then, either on-board or off-board, we can use a probabilistic anomaly detector to identify whether the stream of cases is truly anomalous when we look at the stream of cases as a group. The anomalous group of cases is called a composite case. Thereafter, off-board the machine, each anomalous individual case is classified into a fault type using a case-based reasoning approach to fault diagnosis. In the final step, we fuse the individual classifications into a single aggregated classification for the composite case. In order to be able to assess the reliability of a diagnosis, we also propose a novel case-based approach to estimating the reliability of probabilistic predictions. It can, for instance, be used for assessing the confidence of the classification of a composite case given historical data of the predictive reliability.

ISSN 1651-4238
A Data-Driven Approach to Remote Fault Diagnosis of Heavy-duty Machines

Tomas Olsson

2015
Abstract

Heavy-duty machines are equipment constructed for working under rough conditions and their design is meant to withstand heavy workloads. However, the last decades technical development in cheap electronically components have lead to an increase of electrical systems in traditionally mainly mechanical systems of heavy-duty machines. As the complexity of these machines increases, so does the complexity of detecting and diagnosing machine faults. However, the addition of new electrical systems, such as on-board computational power and telematics, makes it possible to add new sensors that measure signals relevant for fault detection and diagnosis, and to process signals on-board or off-board the machines.

In this thesis, we address the diagnostic problem by investigating data-driven methods for remote diagnosis of heavy-duty machines, where a part of the analysis is performed on-board the machine (fault detection), while another part is performed off-board the machine (fault classification). We propose a diagnostic framework where we use a novel combination of methods for each step in the diagnosis. On-board the machine, we have used logistic regression as an anomaly detector to detect faults that will lead to a stream of individual cases classified as anomalous or not. Then, either on-board or off-board, we can use a probabilistic anomaly detector to identify whether the stream of cases is truly anomalous when we look at the stream of cases as a group. The anomalous group of cases is called a composite case. Thereafter, off-board the machine, each anomalous individual case is classified into a fault type using a case-based reasoning approach to fault diagnosis. In the final step, we fuse the individual classifications into a single aggregated classification for the composite case. In order to be able to assess the reliability of a diagnosis, we also propose a novel case-based approach to estimating the reliability of probabilistic predictions.
Abstract

Heavy-duty machines are equipment constructed for working under rough conditions and their design is meant to withstand heavy workloads. However, the last decades technical development in cheap electronically components have lead to an increase of electrical systems in traditionally mainly mechanical systems of heavy-duty machines. As the complexity of these machines increases, so does the complexity of detecting and diagnosing machine faults. However, the addition of new electrical systems, such as on-board computational power and telematics, makes it possible to add new sensors that measure signals relevant for fault detection and diagnosis, and to process signals on-board or off-board the machines.

In this thesis, we address the diagnostic problem by investigating data-driven methods for remote diagnosis of heavy-duty machines, where a part of the analysis is performed on-board the machine (fault detection), while another part is performed off-board the machine (fault classification). We propose a diagnostic framework where we use a novel combination of methods for each step in the diagnosis. On-board the machine, we have used logistic regression as an anomaly detector to detect faults that will lead to a stream of individual cases classified as anomalous or not. Then, either on-board or off-board, we can use a probabilistic anomaly detector to identify whether the stream of cases is truly anomalous when we look at the stream of cases as a group. The anomalous group of cases is called a composite case. Thereafter, off-board the machine, each anomalous individual case is classified into a fault type using a case-based reasoning approach to fault diagnosis. In the final step, we fuse the individual classifications into a single aggregated classification for the composite case. In order to be able to assess the reliability of a diagnosis, we also propose a novel case-based approach to estimating the reliability of probabilistic predictions. It can,
for instance, be used for assessing the confidence of the classification of a composite case given historical data of the predictive reliability.
Sammanfattning


I denna avhandling adresserar vi diagnostikproblemet genom att använda data-drivna metoder för att diagnostisera tunga maskiner. En del av analysen görs ombord på maskinen (feldetektering) och en del görs på distans i en centraliserad server (felklassificering). Vi presenterar ett diagnostiskt ramverk där vi använder en innovativ kombination av metoder i varje steg av diagnosticeringen. Ombord på maskinen har vi använt logistisk regression som avvikelsedetektor för att klassificera individuella mätningar (fall) som normala eller onormala (avvikande) för att upptäcka fel. Detta resulterar i en sekvens (ström) av individuella fall som är klassificerade som normala eller avvikande. Sedan använder vi, antingen ombord eller på servern, en probabilistisk avvikelsedetektor som kan upptäcka om strömmen med fall är avvikande när vi tittar på strömmen av fall som en grupp med fall istället för som individer. En avvikande grupp med fall kallas för ett sammansatt fall. På servern klassificeras sedan de sammansatta fallen genom att de individuella, avvikande fallen klassificeras med hjälp av case-based reasoning och som

To my wife Elisabeth and to my children Joel, Jonathan and Elsa
Acknowledgements

First I would like to thank my main supervisor Peter Funk and my co-supervisors Ning Xiong at Mälardalen University (MDH), Anders Holst at SICS Swedish ICT (SICS) and Marcus Bengtsson at MDH/Volvo Construction Equipment (Volvo CE). Without their effort, this thesis would not have been written. I would also like to thank Jonas Larsson and Elisabeth Källström at Volvo CE for their important help and support during the thesis work and last but not least, I would like to thank Björn Levin and Daniel Gillblad at SICS for supporting my studies towards my PhD.

Good colleges matter a lot. Therefore, I want to thank all ITS-EASY students involved in the ITS-EASY industrial research school at MDH and the ITS-EASY school staff for all support and interesting discussion during the years. Then, I am also thankful for the support and inspiration from my colleges at SICS in Kista and the encouragement from my colleges at SICS Västerås. Also a great thank to all other fellow PhD students and colleges at MDH, especially members of the AI group. Thank you all!

Finally, I would like to thank my wife Elisabeth, and my children Joel, Jonathan and Elsa for their patient while I was working on this thesis.

This research has been supported by Volvo Construction Equipment (Volvo CE), SICS Swedish ICT, and the Knowledge Foundation (KKS) through ITS-EASY, an Industrial Research School in Embedded Software and Systems, affiliated with the School of Innovation, Design and Engineering (IDT) at Mälardalen University.

Tomas Olsson
Västerås, September 2015
vii
Acknowledgements

First I would like to thank my main supervisor Peter Funk and my co-supervisors Ning Xiong at Mälardalen University (MDH), Anders Holst at SICS Swedish ICT (SICS) and Marcus Bengtsson at MDH/Volvo Construction Equipment (Volvo CE). Without their effort, this thesis would not have been written. I would also like to thank Jonas Larsson and Elisabeth Källström at Volvo CE for their important help and support during the thesis work and last but not least, I would like to thank Bjorn Levin and Daniel Gillblad at SICS for supporting my studies towards my PhD.

Good colleges matter a lot. Therefore, I want to thank all ITS-EASY students involved in the ITS-EASY industrial research school at MDH and the ITS-EASY school staff for all support and interesting discussion during the years. Then, I am also thankful for the support and inspiration from my colleges at SICS in Kista and the encouragement from my colleges at SICS Västerås. Also a great thank to all other fellow PhD students and colleges at MDH, especially members of the AI group. Thank you all!

Finally, I would like to thank my wife Elisabeth, and my children Joel, Jonathan and Elsa for their patient while I was working on this thesis.

This research has been supported by Volvo Construction Equipment (Volvo CE), SICS Swedish ICT, and the Knowledge Foundation (KKS) through ITS-EASY, an Industrial Research School in Embedded Software and Systems, affiliated with the School of Innovation, Design and Engineering (IDT) at Mälardalen University.

Tomas Olsson
Västerås, September 2015
Paper A

Fault Diagnosis of Heavy Duty Machines: Automatic Transmission Clutches

Tomas Olsson, Elisabeth Källström, Daniel Gillblad, Peter Funk, John Lindström, Lars Häkansson, Joakim Lundin, Magnus Svensson, Jonas Larsson.

In Proceedings of the ICCBR 2014 Workshops (CBRDM'14), Cork, Ireland, September 2014.

Paper B

Case-Based Reasoning for Explaining Probabilistic Machine Learning

Tomas Olsson, Daniel Gillblad, Peter Funk, Ning Xiong.

In International Journal of Computer Science & Information Technology (IJCSIT), 6(2), April 2014.

Paper C

Explaining Probabilistic Fault Diagnosis and Classification using Case-based Reasoning

Tomas Olsson, Daniel Gillblad, Peter Funk, and Ning Xiong.

In Case-Based Reasoning Research and Development. Proceeding of the 22d International Conference on Case-Based Reasoning (ICCBR-2014), Cork, Ireland, September 2014.

Paper D

A Probabilistic Approach to Aggregating Anomalies for Unsupervised Anomaly Detection with Industrial Applications

Tomas Olsson and Anders Holst.


The included articles have been reformatted to comply with the PhD thesis layout.
List of Publications

Papers Included in the PhD Thesis


Paper C  *Explaining Probabilistic Fault Diagnosis and Classification using Case-based Reasoning.* Tomas Olsson, Daniel Gillblad, Peter Funk, and Ning Xiong. In Case-Based Reasoning Research and Development. Proceeding of the 22d International Conference on Case-Based Reasoning (ICCBR-2014), Cork, Ireland, September 2014.


---

1The included articles have been reformatted to comply with the PhD thesis layout.
**Paper E: Fault Diagnosis via Fusion of Information from a Case Stream.**

Tomas Olsson and Ning Xiong and Elisabeth Källström and Anders Holst. In Case-Based Reasoning Research and Development. Proceeding of the 23h International Conference on Case-Based Reasoning (ICCBR-2015), Frankfurt, Germany, September 2015.

**Related Publications not Included in the PhD Thesis**


3. *Representation and similarity evaluation of symbolic time series in uncertain environments* N Xiong, T Olsson, P Funk In Workshop on Reasoning about Time in CBR at 22nd International Conference on Case-Based Reasoning (RATIC 2014) (2014)

**Licentiate Thesis**


**Other AI-related peer-reviewed publications**


---

2A licentiate degree is a Swedish graduate degree halfway between M.Sc. and Ph.D.


**Patent**


## Contents

I Thesis 1

1 Introduction 3
   1.1 Research Problem and Contributions . . . . . . . . . . . . 6
      1.1.1 Research Questions . . . . . . . . . . . . . . . . . . 7
      1.1.2 Research Contributions . . . . . . . . . . . . . . . . 8
   1.2 Thesis Outline . . . . . . . . . . . . . . . . . . . . . . . . 10

2 Background 11
   2.1 Fault Diagnosis . . . . . . . . . . . . . . . . . . . . . . . . 11
   2.2 Heavy-Duty Machines . . . . . . . . . . . . . . . . . . . . . 14
   2.3 Data Streams . . . . . . . . . . . . . . . . . . . . . . . . . . 14
   2.4 Data-Driven Methods . . . . . . . . . . . . . . . . . . . . . 16
      2.4.1 Case-Based Reasoning . . . . . . . . . . . . . . . . . . 18
      2.4.2 Probabilistic Machine Learning . . . . . . . . . . . . 21
      2.4.3 Anomaly Detection . . . . . . . . . . . . . . . . . . . . 22

3 Research Summary 25
   3.1 Diagnostic Framework . . . . . . . . . . . . . . . . . . . . 25
      3.1.1 Statistical Model for On-board Fault Detection of
            Individual Cases . . . . . . . . . . . . . . . . . . . . . . 26
      3.1.2 Probabilistic Anomaly Aggregation for Group
            Anomaly Detection . . . . . . . . . . . . . . . . . . . . . . 27
      3.1.3 Case-Based Diagnosis of Individual Anomalies . . . . 28
      3.1.4 Case Stream Fusion for Group Classification . . . . 28
      3.1.5 Reliability Estimation for Diagnosis
            Predictions and Method Selection . . . . . . . . . . . . . 29
   3.2 Prototype Implementation . . . . . . . . . . . . . . . . . . 30
3.3 Research Methodology and Research Process ............................................. 34

4 Abstracts of Included Papers 37
4.1 Paper A: Fault Diagnosis of Heavy Duty Machines: Automatic Transmission Clutches ................. 37
4.2 Paper B: Case-Based Reasoning for Explaining Probabilistic Machine Learning ......................... 38
4.3 Paper C: Explaining Probabilistic Fault Diagnosis and Classification using Case-based Reasoning ........ 39
4.4 Paper D: A Probabilistic Approach to Aggregating Anomalies for Unsupervised Anomaly Detection with Industrial Applications .................. 40
4.5 Paper E: Fault Diagnosis via Fusion of Information from a Case Stream .......................... 40

5 Related Work 43
5.1 CBR Systems ......................................................... 43
5.2 Hybrid Systems ..................................................... 45
5.3 Other Systems ......................................................... 47
5.4 Reliability of Diagnosis ............................................. 48

6 Conclusions and Future Work 51

Bibliography 55

Appendices 65

A The k-Nearest Neighbour Algorithm 66

B The kNN as an Approximation to Probabilistic Learning 68
8.5.3 Evaluation of the Estimation of the Prediction Error 107
8.5.4 Case-based Explanation Examples 109
8.6 Conclusions and Future Work 111
Bibliography 113

9 Paper C:
Explaining Probabilistic Fault Diagnosis and Classification using Case-based Reasoning 117
9.1 Introduction 119
9.2 Case-based Explanation 120
9.3 Preliminaries 122
  9.3.1 Similarity and True Metrics 122
  9.3.2 Statistical Metrics 123
  9.3.3 Logistic regression 123
9.4 The Case-based Explanation Framework 124
  9.4.1 Statistical Measures of Similarity for a Probabilistic Classifier 125
9.5 Explaining Fault Diagnosis 127
  9.5.1 Fitting Logistic Regression 129
  9.5.2 Estimating Local Accuracy with kNN 129
  9.5.3 Case-based Explanation Examples 131
  9.5.4 Analyzing the Local Accuracy 133
9.6 Conclusions and Future Work 134
Bibliography 135

10 Paper D:
A Probabilistic Approach to Aggregating Anomalies for Unsupervised Anomaly Detection with Industrial Applications 139
10.1 Introduction 141
10.2 Background 142
  10.2.1 Collective Anomalies 142
  10.2.2 Unsupervised Anomaly Detection 143
10.3 Aggregation of Anomalies 143
  10.3.1 A Naive Aggregation Scheme 143
  10.3.2 Probabilistic Aggregation 144
  10.3.3 Identifying Anomalous Groups 145
10.4 Evaluation 146
# List of Figures

1.1 An example of a wheel loader. ........................................ 4  
1.2 An on-board and off-board architecture for processing sig- 
    nals from a machine. ............................................. 5  
1.3 The thesis uses a combination of different methods with 
    applications to fault diagnosis of heavy-duty machines. . . 6  
2.1 The CBR management cycle [1]. ................................. 19  
3.1 The proposed fault diagnosis framework. ...................... 25  
3.2 The connection between the papers and the steps in the 
    diagnostic framework. ........................................... 27  
3.3 The part of the fault diagnosis framework implemented in 
    the prototype. .................................................. 30  
3.4 The architecture of the implemented prototype. ............ 31  
3.5 The prototype on-board visualisation. ......................... 32  
3.6 The prototype off-board anomaly management interface. . 33  
3.7 The prototype off-board diagnosis interface. ................ 33  
7.1 On-board and off-board fault diagnosis system. ............ 80  
8.1 Number of $k$ nearest neighbors (x-axis) versus the root 
    mean square error (y-axis) of the estimated prediction error.109  
9.1 The accuracy learning curve for classifying steel faults. . . 130  
9.2 MSE for the kNN algorithm using different distances and 
    various $k$. ....................................................... 132  
10.1 Histogram of anomaly scores of the artificial data. ........ 147
10.2 Outlier detection curves for different value of $p$ and varying non-anomalous standard deviation. Curves are close to zero for $p > 0.4$. .......................................................... 148
10.3 The zoomed tail of the histogram of anomaly scores of container move durations. The y-axis denotes number of moves and it is cut at about 20. ............................. 149
10.4 Histogram of anomaly scores per crane. The x-axis is cut at 50 sec and each crane has different y-axis scale. ......... 150
10.5 Histogram of anomalies scores for each crane. The x-axis shows from 20 up to 1000 sec. The y-axis has a uniform scale from 0 to 10. .................................................. 151
10.6 The outlier detection curve for cranes. The y-axis is number of identified anomalous cranes. The x-axis is the different values of $p$. .......................................................... 152
10.7 The sorted number of times each crane was identified as an outlier. .......................................................... 152
10.8 Outlier detection curve for anomalous road segments. ... 153
10.9 The sorted number of times (y-axis) that each road segment (x-axis) was identified as an outlier. .................. 154

11.1 The original on-board and off-board fault diagnosis system.163
11.2 The extended fault diagnosis framework. .................. 164
List of Tables

3.1 Papers in chronological order and the related step in the diagnostic framework. Parentheses indicate an indirect relation as shown in Fig. 3.2. .......................... 28

7.1 The Area Under the Curve (AUC) for the anomaly detection. .................................................. 83
7.2 The mean squared error (MSE) for clutch slip severity diagnosis. .................................................. 85

8.1 The building and location attributes with summary of the data followed by measuring period start and energy performance. Climate zone indicate location in Sweden: zone 1 is in the most northern part and 4 is in the most southern part of Sweden. ........................................ 106
8.2 Heating systems and how many households with the energy source. ........................................ 107
8.3 The regression weight for each feature. .................. 108
8.4 Household example 1 ........................................ 110
8.5 Household example 2 ........................................ 111
8.6 The three most similar cases to household example 2 .... 112

9.1 The four derived distances. .............................. 128
9.2 Attributes: 27 independent variables, and the last rows shows the 7 fault classes. ......................... 129
9.3 The mean squared error (MSE) for different distances (best is in bold font). ............................. 131
9.4 Fault 1: Low local accuracy 20% (2 of 10) and low class probability 33.4%. .......................... 132
9.5 Fault 2: Low local accuracy 10% (1 of 10) and high class probability 73.4%. .......................... 133

10.1 Performance (in %) for different values of the non-anomalous standard deviation: Proposed Algorithm (above) and baseline (below). The last column (*) shows the performance of the robustness test. ........................................ 148

11.1 The data sets with normal and fault cases with corresponding sizes. ..................................... 170
11.2 Accuracy of the information fusion approaches for detected faults. ...................................... 172
11.3 Metrics showing the performance of the PRAAG algorithm. 172
11.4 Total accuracy of the information fusion approaches for all classes. ............................. 173
I

Thesis
Chapter 1
Introduction

Heavy-duty machines are complex working machines that many times are run under rough conditions and are therefore designed to do hard work without breaking. Examples of heavy-duty machines are wheel loaders, rail mounted gantry cranes, and trucks. As with many other mechanical systems, more and more electronically systems and other features are added to the machines, and hence, the complexity of the machines increases. As the complexity increases, finding and diagnosing faults becomes more difficult. Thus, the time to identify a fault gets longer and so does the service time, while naturally, the uptime of the machines decreases.

In contrast, many machine operators and transportation companies have small economical margins which makes it important to minimise cost of repairs and services, while maximising the uptime of the machines. In addition, many manufacturers are moving from only selling machines and replacement parts to selling new functional products where the customers also can be guaranteed a level of availability.

As a step towards addressing the above problems, this thesis investigates data-driven approaches to automating early detection and diagnosis of faults of heavy-duty machines. By being able to continuously monitor machines in order to detect problems early and diagnose the problems, it is expected that less unplanned stops will be made and thus, the uptime will increase. Data-driven methods are methods that learn and generalise from data in contrast to methods where models are.
Chapter 1

Introduction

Heavy-duty machines are complex working machines that many times are run under rough conditions and are therefore designed to do hard work without breaking [2]. Examples of heavy-duty machines are wheel loaders (Figure 1.1), rail mounted gantry cranes, and trucks. As with many other mechanical systems, more and more electronically systems and other features are added to the machines, and hence, the complexity of the machines increases. As the complexity increases, finding and diagnosing faults becomes more difficult. Thus, the time to identify a fault gets longer and so does the service time, while naturally, the uptime of the machines decreases.

In contrast, many machine operators and transportation companies have small economical margins which makes it important to minimise cost of repairs and services, while maximising the uptime of the machines [3, 4]. In addition, many manufacturers are moving from only selling machines and replacement parts to selling new functional products where the customers also can be guaranteed a level of availability [3, 5].

As a step towards addressing the above problems, this thesis investigates data-driven approaches to automating early detection and diagnosis of faults of heavy-duty machines. By being able to continuously monitor machines in order to detect problems early and diagnose the problems, it is expected that less unplanned stops will be made and thus, the uptime will increase. Data-driven methods are methods that learn and generalise from data in contrast to methods where models are

\[\text{\footnotesize Image of wheel loader used with permission from Volvo Construction Equipment.}\]
built manually using human expertise. A common instance of an expert method is the first principle model were a mathematical model of a physical system is constructed beforehand [6]. Other types of expert methods are rule based expert system and business rules [7]. Nevertheless, although all applications in this thesis are for heavy-duty machines, the described approach can also be used for other types of machines.

The basic idea is to gather data from sensors on-board a fleet of machines that are used for training data-driven methods into detecting and diagnosing faults. However, unlike fault detection and diagnosis in industrial processes or power stations [8, 9] that can quite easily be equipped or connected to as much computer power as needed, most heavy-duty machines have limited internet connection and limited computational power. For instance, construction machines, like wheel loaders, are often used at remote places with bad or no Internet connection while trucks travel long distances where the Internet connections are of varying quality. Thus, for heavy-duty vehicles there is a trade-off between what can be processed on-board and off-board the machines. In cases, where there is good Internet connection and plenty of computational resources, this will of course not be an issue. However, the use of the existing but limited on-board processing leads to faster processing time compared to when all data should be sent off-board for processing. When the on-board analysis manages a part of the fault diagnosis, the off-board analysis will
Chapter 1. Introduction

Figure 1.1: An example of a wheel loader.

built manually using human expertise. A common instance of an ex-
pert method is the first principle model were a mathematical model of
a physical system is constructed beforehand [6]. Other types of expert
methods are rule based expert system and business rules [7]. Neverthe-
less, although all applications in this thesis are for heavy-duty machines,
the described approach can also be used for other types of machines.
The basic idea is to gather data from sensors on-board a fleet of ma-
chines that are used for training data-driven methods into detecting and
diagnosing faults. However, unlike fault detection and diagnosis in indus-
trial processes or power stations [8, 9] that can quite easily be equipped
or connected to as much computer power as needed, most heavy-duty
machines have limited internet connection and limited computational
power. For instance, construction machines, like wheel loaders, are of-
ten used at remote places with bad or no Internet connection while trucks
travel long distances where the Internet connections are of varying qual-
ity. Thus, for heavy-duty vehicles there is a trade-off between what can
be processed on-board and what can be processed off-board. In cases, where there
is good Internet connection and plenty of computational resources, this
will of course not be an issue. However, the use of the existing but limited
on-board processing leads to faster processing time compared to when all
data should be sent off-board for processing. When the on-board anal-
ysis manages a part of the fault diagnosis, the off-board analysis will
do not need as much computational power as if all data would be gathered
for off-board processing while at the same time the off-board processing
will have access to greater computational resources than the on-board
analysis. Thus, an architecture as shown in Fig. 1.2 with both on-board
and off-board processing is proposed in this thesis.

Figure 1.2: An on-board and off-board architecture for processing signals
from a machine.

In this thesis, we propose a framework for remote fault diagnosis
that can use data collected from many different but similar machines
for detecting and diagnosing faults where on-board processing is done to
filter out non-interesting events and off-board analysis is done to classify
faults or identify anomalies. For this purpose, this thesis uses a novel
combination of methods from case-based reasoning and probabilistic ma-
chine learning for fault diagnosis (Fig. 1.3). On-board the machine, we
use a supervised approach to probabilistic anomaly detection to detect
faulty cases, then off-board the machine we use unsupervised anomaly
detection to identify an anomalous stream of cases, and last, we use
case-based reasoning in combination with information fusion to make a
diagnosis.

In addition, any learning system is no better than its predictions and
for fault diagnosis it is essential to be able to justify the confidence in
a prediction. In probabilistic classification, probabilities are used for
assessing the confidence in the prediction. However, the confidence of
the prediction is dependent of the correctness of the used probability
model, and thus, if the model is not correct for all types of input, then
the prediction and the probability assessment are not reliable. So, there
is a need for improving the means for justifying predictions that are not
dependent on the used prediction method and that can be used as a
Figure 1.3: The thesis uses a combination of different methods with applications to fault diagnosis of heavy-duty machines.

second opinion on the reliability of a prediction. For this purpose, we propose a justification that uses a case-based reasoning approach that estimates the reliability directly from the most relevant previous cases.

1.1 Research Problem and Contributions

The problem addressed in this thesis is how to support remote fault diagnosis of heavy-duty machines. Fault diagnosis is a two-step process that consists of (1) detecting and (2) classifying the fault. Fault detection aims at detecting when a system does not perform as expected while fault classification is about determining the type (or class) and severity of a fault. Generally speaking, each step can be done manually or automatically, or a combination of both. The goal of this thesis is to support automatic diagnosis as far as possible.

We assume that there is a fleet of similar machines that are constantly monitored and that data can be collected from the machines at least once per day or when deemed necessary. However, it is also assumed that some data processing might be performed both on-board and off-board the machine so that the computation can be distributed in the system. In addition, it is assumed that there is an off-board system (back office)
that either can automatically analyse data from the machines or support manual decision-making by experts when needed.

The research questions and contributions in this thesis can be related to fault diagnosis and in the used data-driven methods: case-based reasoning (CBR), probabilistic machine learning and anomaly detection. The thesis includes five papers, referred to as Paper A, B, C, D, and E. Below, we list the high-level research topics with the corresponding papers that addresses them.

Fault diagnosis

- Fault Detection [Paper A,D,E]
- Fault Classification [Paper A,E]
- Reliability of Fault Diagnosis [Paper B,C]

Method improvements

- Case-Based Reasoning [Paper A,B,C,E]
- Probabilistic Machine Learning [Paper B,C]
- Anomaly Detection [Paper A,D,E]

In the following sections, we first list the addressed low-level research questions (RQ1-4) related to these areas. Then, we present the research contributions (RC1-5) of the thesis.

1.1.1 Research Questions

RQ1: How to detect faults online with limited computational resources? This research question assumes that only limited computational resources are available and thus, only relatively simple algorithms can be used. The problem is to filter out cases that are faulty from those that are not.

RQ2: How to assess the anomalousness of a group of cases? This research question is related to [RQ1] but considers assessing the faultiness of a group of cases as a collective, while [RQ1] addresses fault
detection of individual cases. This research question addresses the problem that if the fault detection for individual cases is error prone, we can increase the significance of the detection by looking at many related cases together.

**RQ3:** How to diagnose individual cases and groups of cases? This research question is related to [RQ1] and [RQ2], but computational resources are not limited as in an online setting. In addition, since in any realistic setting on-board a machine, cases are not generated one by one but as stream of cases. Thus, we don’t need to have only a single case as basis for a diagnosis but a group cases.

**RQ4:** How to identify when supervised learning fails to produce a reliable diagnostic model? In order for a decision to be made, there need to be some measure of confidence in a prediction from the diagnostic system. This research question addresses how the confidence in the reliability of a diagnosis can be measured.

### 1.1.2 Research Contributions

**RC1: A statistical on-board anomaly detection algorithm that can detect faults** The first thesis contribution presented in Paper A addresses the problem of detecting faults online using limited resources [RQ1]. In the proposed approach, a statistical model (Gaussian mixture model) is used to process the features in order to compute the likelihood that each feature value is normal. The output of the statistical model is then fed to a probabilistic classifier (logistic regression) that classifies a case as normal or anomalous. Detected anomalies are thereafter sent off-board for further analysis using case-based reasoning to predict the severity of a fault. Both the Gaussian mixture model and the logistic regression are assumed to be built off-board the machine, but being used on-board the machine. Thus, only limited computational resources are needed for the on-board processing.

**RC2: Group anomaly detection using probabilistic anomaly aggregation** The second contribution presented in Paper D addresses research question [RQ2]. The paper presents an approach to detecting anomalous groupings of individual anomalies from, for instance, the above on-board anomaly detector, to find likely faulty groups of cases.
Assuming that on-board anomaly detector might return a large number of false positives over time, it might be good to distinguish between randomly occurring anomalies and a significant increase in anomalies. Thus, this can, as presented in Paper E, be used to assess whether a vehicle generates a significantly unlikely number of anomalies in a stream of cases instead of only classifying each anomaly separately as in Paper A. This anomaly detection approach can be used both on-board and off-board.

**RC3: Fault classification of individual cases using a probabilistic case similarity** The third contribution in Paper A partially addresses [RQ3]. Paper A presents a CBR-based approach to diagnosis that uses the $k$-nearest neighbour algorithm to predict the severity of an anomalous case that is detected on-board. The same statistical model for normal cases as for the statistical on-board anomaly detector (described in research contribution [RC1]) was used, so that similarity between faults were measured as how similarly cases deviate from the normal cases with respect to the statistical model.

**RC4: Fault classification using information fusion over a stream of cases** The fourth research contribution is presented in Paper E also addresses [RQ3]. Paper E investigates approaches to fusing the classification of the individual cases from an anomalous group of cases. A sliding window of the most recent cases are individually classified as anomalous or not, and we can then detect whether an unlikely number of cases are anomalous as a collective compared to a much larger sliding window. Then, the individual anomalous cases in the most recent window are classified using the CBR classifier in [RC3], and thereafter, a final classification is inferred by fusing the results of the individual classifications.

**RC5: A CBR approach to reliability estimation of a probabilistic prediction** The last contribution in this thesis is to estimate the prediction performance of a probability model by using the local prediction performance given by the average performance of the previous most similar cases. Paper B and C propose a novel approach to comparing cases that use probability distributions to model cases and a statistical similarity metric to compare them so that cases similar with
Chapter 1. Introduction

respective to the probability model can be retrieved. Then, Paper B applies the approach to justifying the reliability of a linear regression algorithm by using the local mean absolute error given by the k-nearest neighbours. Paper C applies the same approach to probabilistic classification (logistic regression) of faults but uses the local accuracy given by the k-nearest neighbours. Using this approach, it is possible to present a set of previous cases together with an estimate of the expected prediction reliability as a decision support. Notice that this approach can be used regardless of whether a pure CBR approach is used or a probabilistic machine learning method as in the included papers. It is only required that the learning algorithm can be related to a similarity metric. Paper E proposes the use of this approach to diagnose the composite cases constructed from the individual anomalies from a stream of cases, but it could also be used for estimating the reliability of the fused classification given a case base of previous composite cases.

1.2 Thesis Outline

The outline of the thesis is as follows. Section 2 gives an overview of background information on the problems addressed and methods used in this thesis such as fault diagnosis, heavy-duty machines, data streams and data-driven methods. Section 3 gives an overview of the research conducted in this thesis. Section 4 lists and summarises each included paper. Section 5 presents related work. Finally, Sect. 6 ends with a set of conclusions and describes future work.
Chapter 2

Background

In this chapter, we describe specific problems related to remote fault diagnosis of heavy-duty machines. We also look at the relation to data streams. Last, we describe data-driven methods and anomaly detection.

2.1 Fault Diagnosis

In this thesis, we use fault diagnosis in its broadest sense of the term by including both fault detection and fault classification. Many times in related literature, fault diagnosis only includes the latter[10].

Notable, there are different definitions of what a fault is and its relation to the concept of a failure. In control engineering, fault diagnosis has been studied for many years and the working group SAFEPROCESS\(^1\) of the International Federation of Automatic Control (IFAC) uses the following definitions presented in [10, 7]. A fault [10, p. 710] is

“An unpermitted deviation of at least one characteristic property or parameter of the system from the acceptable / usual / standard condition” .

Further, a failure [10, p. 710] is defined as

“A permanent interruption of a system’s ability to perform a required function under specified operating conditions” [10].

\(^1\)http://tc.ifac-control.org/6/4
Similarly, the author of [8] uses the following definition of a fault:

"A fault can be defined as anomalous behaviour causing systems or processes to deviate unacceptably from their normal operating conditions or states." [8, p. 5]

In this case, the fault precedes the failure, so that a fault will result in a failure if not detected in time.

In contrast, in maintenance terminology (MT) (see the Swedish-European standard [11]), a fault [11, p. 11] is defined as a

"state of an item characterised by inability to perform a required function, excluding the ability during preventive maintenance or other planned actions, or due to lack of external resources."

which is followed by a comment

"NOTE A fault usually results from a failure, but in some circumstances it may be a pre-existing fault."

Moreover, a failure [11, p. 9] is defined as

"termination of the ability of an item to perform a required function”.

Yet, it continues with the following comments:

"NOTE 1 After failure the item has a fault, which may be complete or partial."

"NOTE 2 'Failure’ is an event, as distinguished from 'fault’, which is a state."

Thus, MT seems to have the opposite definition compared to control engineering in that the failure is the event that results in a fault that is a state, so the failure precedes the fault. However, the standard also defines a degraded state [11, p. 11] to be a

"state in which the ability to provide the required function is reduced, but within defined limits of acceptability”,

"NOTE A degraded state may be the result of faults at lower indenture level".
Then, the indenture level [11, p. 7] is defined as a

“level of sub-division within an item hierarchy”

“NOTE 1 Examples of indenture levels are: system, sub-system and component”.

So the contradiction between the two definitions is not complete, since control engineering considers the failure of the system while MT considers the failure of a specific component in the system. Thus, in this thesis, fault detection considers the task of detecting when a fault occurs at a lower component level before it results in a failure at a higher system level, that is, before the monitored component stops working as intended because of a failure of a subcomponent. Consequently, we also consider a fault to be a state that is a result of a failure. We only use anomalous behaviours as an indication of a fault and not as the definition of a fault. Fault states that do not result in an immediate malfunctioning component are also called *incipient faults* [12, 13].

There are two main approaches to fault diagnosis: model-based and data-driven approaches [14, 15]. Model-based approaches use detailed mathematical models of the physical system that can be used to detect when there is a fault. Data-driven or data-based methods use measurements from sensors attached to the machine to automatically learn how the machine normally operates. The drawback of the former is that it requires a lot of detailed knowledge of the system in order to be applied while the drawback of the latter is, depending of the used method, that the data-driven method can be more or less a black box in that it is not always easy to understand its inner working. However, if you lack data, the only choice is to use a model-based approach.

Traditionally in control engineering, the fault detection part is automated while fault classification has been done by other means [8]. Typically, the expected sensor outputs have been modelled and then the residuals (difference) of expected output compared to the actual output have been computed [6, 9, 8]. Then, if the residuals are too large we have identified a faulty part, but the type of the fault is not necessarily identified.

In this thesis, we consider fault diagnosis as a learning problem in that we assume that we have a set of labelled cases of faults and non-faulty data. Thus, we are not only interested in detecting faults but in automating the full diagnosis as far as possible. So, in this work, we only consider data-driven methods for fault detection and diagnosis.
2.2 Heavy-Duty Machines

This thesis is about analysing sensor data collected remotely from heavy-duty machines in order to detect and diagnose faults. However, the conducted research covers data collected from three different types of machines. The main application is for wheel loaders, but we have also used data from container cranes and long distant traveling heavy vehicles carrying cargo. These three types of vehicles have quite different characteristics. All three are used under rough conditions. However, the three vehicle types come with different problems with respect to data analysis. Container cranes are used together in a container stack so that there are about 20 or more identical cranes. Thus, it is easy to monitor in that we can assume similar data are generated from all cranes in a stack. In addition, container cranes are used under similar conditions with a quite limited number of possible working conditions. In contrast, cargo-carrying vehicles can travel different routes under varying circumstances, for instance, different drivers and varying road conditions. Thus, cargo-carrying vehicles have larger individual variations than container cranes. Yet, the cargo-carrying vehicles can be quite similar in size and working conditions, such as, using the same motorways. However, wheel loaders are used in much more diverse circumstances. They can be working in small or large constructions sites. They can work in cities or at remote sites far from any decent Internet connection. They can work high up in the mountains or deep down in a mine. These are issues to take into account when evaluating the conducted research. Nevertheless, regardless of type of vehicle, we assume that data from the vehicles can be collected at least once per day and that the data comes as streams of data to a remote back office that can further analyse the data.

2.3 Data Streams

Data-driven methods from machine learning and data mining have traditionally been used for analysing static data sets that are not updated very frequently [16, 17]. However, when we continuously monitor and analyse data that are generated on-board running machines we are not longer looking at approximately static data but at streaming data. Potentially, we could monitor each component of a machine such as the engine, transmission, breaks, axles, and even the driver. Each component
could also have a large number of sensors. All these sensors would generate never-ending series of measurements that would be impossible to store and then analyse sufficiently fast using traditional methods due to the limited computational resources, memory and communication bandwidth available on-board a machine. Thus, we can do some computation on-board the machine in order to lessen the amount of data that is then sent off-board for further processing and thereby make efficient use of the limited on-board computational resources and the limited bandwidth.

According to [18, 19, 20], learning from data streams differs from learning from databases in the following ways. A data stream learning algorithm should:

1. Use a small constant time to process a data element
2. Use a only a fixed amount of memory
3. Scan data only once when learning (not several times)
4. Be able to predict at anytime
5. Adjust to changes to the underlying data generating mechanism.
6. Build models equally good as when doing batch learning

In order to manage these requirements, various approaches are used [19, 20]. Approximation is used so that instead of giving an exact solution, only inexact solutions are computed with a guaranteed size of the error. Similarly, randomisation can be used such that the solution is correct with a predefined probability. A related approach is to sample the stream to build a subset of the data to represent the whole stream. Another approach is to represent the stream with a finite window with the most recent elements. Other approaches construct various forms of summaries of the data, such as, counting number of items, estimate mean values, estimate moments and other similar statistics [17].

One of the research areas within data stream mining investigates approaches to detecting concept drift that is used for detecting when a learning algorithm stops performing well and should restart learning [21]. Relearning is often done by first forgetting previous learned elements and then by learning from the most recent elements. However, this setup assumes that the correct classification is learned quite immediately after making a prediction, but this is not the case when doing diagnosis since the true label is not learned until quite long time afterwards. In addition,
the work in this thesis differs in that we do not have an algorithm that learns to classify directly from the stream. Instead, we assume that the machine is functioning in its normal mode and that any change in underlying data generation process is an indication of a fault and not an indication of the need to restart learning.

2.4 Data-Driven Methods

This chapter presents the data-driven methods used in this thesis. That the methods are data-driven means that we mainly use data to directly do predictions or by generalising to a model from the data. Thus, we do not explicitly model a system using a mathematical model. By a data-driven method, we also mean a method that learns from experience and improves its performance. Thus, we adopt the following definition of a learning program proposed by Tom Mitchell in his book on machine learning [22]:

**Definition 1.** A computer program is said to learn from experience $E$ with respect to some class of tasks $T$ and performance measure $P$, if its performance at tasks in $T$, as measured by $P$, improves with experience $E$.

This is a very generic notion of learning that can encompass many learning methods, even approaches that only constitutes of updating a database to improve query answering [22]. The experience is the data used by the method to learn, which we in this thesis call the cases or the case base. A task can then be to predict a class (classification) or to predict a real value (regression) for a new case, but also something more complex such as learning a utility function for selecting the best choice of actions, for instance, using reinforcement learning. In this thesis, we only look at classification and regression.

An important aspect of the performance measure is that it should measure the ability to generalise to new cases (new experience), since for old cases the true answer of a prediction is already known. Thus, the choice of performance measure is closely related to the kind of learning task. In case of classification, we usually count the fraction of correctly classified cases (accuracy) and for regression we usually compute the mean absolute error or the mean squared error between the predicted value and the true value. If the task is binary classification, that is,
we only have two classes (for instance, -1 or 1) to choose from, then we can also compute precision and recall as well as the area under the ROC curve (AUC) as performance measures. Precision is defined as the fraction of positive classifications that is classified correctly, while recall is defined as the fraction of all positive cases that is correctly classified as positive. AUC is a more generic measure that computes the area of the curve of varying the decision threshold for different fractions of true positive classification cases and different fractions of false negatively classifications [23].

The most common type of learning is supervised learning, where a method can learn from cases with known labels, that is, the true class or the true value of the cases are known. Another type of learning is unsupervised learning, which is when there are no labelled cases, but where new information is learned from the cases. Thus, unsupervised learning is often harder to evaluate than supervised learning, since the performance measure cannot be related to knowing the true class or value. A third type of learning is called semi-supervised learning where the labels of some of the cases are known, but not for most of the cases. This thesis uses both supervised and unsupervised methods. Classification and regression are supervised methods while anomaly detection can be both supervised and unsupervised.

In addition, learning can be done by generalising directly from a set of cases on demand or by first generalising into a model and then, make a prediction, or by a combination of both. The former is sometimes called lazy learning since most or a large part of the computation is delayed till when a prediction is done or instance-based learning, since the inference is done directly from the examples.

An important notice on the selection of learning algorithm; The no-free lunch theorem states that there is no universally best learning method for all situation but one method might work better than another in one application and vice versa for another [24]. The theorem implies that many times it is more important to select the right features than the right learning method, since the features must reflect the true dependency in relation to the learned tasks, otherwise no method will be able to learn the mapping from the features to the predicted value. In addition, it is better to start out with a simple method and then make it more complex when needed. So, in this thesis we use rather simple learning methods in most cases but use them in novel combinations. For instance, the simplest method to learning is the kNN algorithm, but
given sufficiently many training cases it is able to give reasonable performance for any application domain [25]. Thus, kNN is a reasonable baseline for comparing more complex learning methods, since if a more complex method is not better than kNN then there is probably something wrong with that method. However, since kNN compute predictions on demand from cases stored in memory, it can be rather slow and memory consuming. So, a model-based learning method can be faster and less memory consuming, and therefore preferred over kNN.

In the following sections, we give an overview of a set of data-driven methods used in the various papers included in the thesis. We start by describing case-based reasoning that makes inference directly from cases. Thereafter, we present probabilistic machine learning that uses probability theory for learning. Last, we describe anomaly detection where the learning task is classification into normal or anomalous cases but where the true classes might not be known beforehand.

### 2.4.1 Case-Based Reasoning

Case-based reasoning (CBR) is a AI approach that is inspired by cognitive theories developed by Roger C. Schank [26, 27] in the 70’s and 80’s. The idea is that human are often using analogical or similarity reasoning to solve new problems. That is, if a new situation resembles old situation then old solutions to old problems can be used to solve the new problem. For instance, when software developers solve new problems they often use a search engine to find out whether other people already solved the same or a similar problem, and then, adjust the found solutions to their own needs. CBR is about automating this type of problem solving process, where the inference is done directly from the cases without first generalising to a model that is common in other learning methods.

Nevertheless, CBR has much in common with other learning approaches in that CBR uses previous observations to generalise to new observations. However, while machine learning is focusing on automated means of doing this, it is common in CBR to manually specify and fine-tune a similarity metric and to manually create and select new cases, although much research has also focused on how to do that automatically. The latter, more automated version of CBR, is called knowledge-light CBR [28] while the former, more handcrafted CBR, is called knowledge-intensive CBR [28, 29], and of-course, there are approaches in between those two extremes. The benefit of knowledge-intensive CBR compared
to knowledge-light CBR is that a lot of knowledge can be coded into the system and thereby; less number of cases are needed to do inference. CBR is also referred to as a lazy learning or an instance-based learning approach in that the inference is done directly from the cases when a prediction is done.

CBR can best be described as a conceptual framework for reasoning or a knowledge management methodology that follows the CBR cycle shown in Fig. 2.1. Agnar Aamodt and Enric Plaza presented the CBR knowledge management cycle with the following four processes in their seminal paper from 1994 [1]:

1. RETRIEVE the most similar case or cases
2. REUSE the information and knowledge in that case to solve the problem
3. REVISE the proposed solution
4. RETAIN the parts of this experience likely to be useful for future problem solving

![Figure 2.1: The CBR management cycle](image)
Research issues in CBR are essentially related to the four steps in the CBR management cycle.

In the Retrieve step, given a new case problem, the set of cases most relevant or similar are retrieved. Two important research questions in this step are how the cases are represented and how to measure the similarity between cases. Cases are typically represented with two parts: a problem and a solution, and sometimes an outcome. The problem describes the setting in which the solution has been applied and the outcome is the result or the evaluation of applying the solution. However, how each part is represented varies, e.g., attribute-value pairs or something more complex. The similarity metric algorithm can be very simple or very advanced. Another research issue is algorithms for indexing cases for easy retrieval and for maintaining the case-base, which is closely related to the choice of similarity metric.

In the Reuse step, the retrieved solution(s) are applied to the new problem. For this step, an adaptation algorithm is needed that can adapt the retrieved solution(s) to the problem. Algorithms for adaptation are an important research question, since there is no generic approach to adaptation. However, most current CBR implementations leave it to the users to do the adaptation.

In the Revise step, the outcome of applying the adapted solution from the previous step is evaluated. If the outcome was not successful, a more elaborate process of finding a new solution is required. Often this means asking human expert to provide a solution.

In the Retain step, the new case is analysed and only the parts of the case that are deemed important for the previous steps are stored. Research questions in this step includes both issues from knowledge modelling and feature extraction and selection in machine learning.

Another research topic is to relate CBR to other inference approaches such as Fuzzy logic and other types of probabilistic reasoning [30, 31, 32]. One search direction is to formalise CBR in a theoretical framework [33, 34]. A related research direction is to broaden the scope of CBR to include a more generic case matching based on preferences or utilities, not only on the similarity between problems [35, 34].

A common way of implementing a CBR system is to define a similarity metric, retrieve the most similar historical cases, and then do some inference. A good similarity metric is therefore essential when implementing a CBR system. According to [36], a good similarity metric should give a measure of the usefulness of previous cases given a query.
of a new case. Then, three assumptions behind a good similarity metric are: (1) Similarity between a query and a case implies usefulness. (2) The similarity is based on a priori known facts. (3) As cases can be more or less useful for a query, similarity must provide a quantitative measurement. When representing cases as feature vectors, a typical similarity metric has the following three components [37]:

- **local similarity functions** that compute the similarity between the values of two cases for a single feature
- **feature weights** that indicate the importance of features
- an **amalgamation function** that combines the local similarity functions and weights into one global similarity value

The *global* similarity metric between two cases might then be defined as follows, using the summation as amalgamation function:

\[
s(x, y) = \sum_k w^k s_k(x^k, y^k)
\]

where \(s_k\) denotes the *local* similarity function between two case feature values and \(w^k\) denotes a importance weight for each feature \(k\).

The most commonly used and most basic CBR algorithm is the knowledge-light \(k\)-nearest neighbour algorithm (kNN) that can be used in the retrieve and reuse step. We use kNN as CBR approach in the current thesis and kNN is described in detail in Appendix A.

In this thesis, we mainly address research issues within the retrieval step. In Paper A, Paper B, and Paper C, we deal with the problem of defining a similarity metric and how to represent features. As a contribution in relating CBR to probabilistic reasoning and a sound theoretical foundation, we propose a principled approach to defining similarity metrics based on a statistical model.

### 2.4.2 Probabilistic Machine Learning

Probabilistic machine learning is an approach to machine learning that uses probability theory to learn from the cases [25]. In probabilistic machine learning, one usually makes a distinction between parametric and non-parametric approaches, where the difference is that a limited set of parameters are needed to fit the former approach, while the non-parametric uses fewer preselected parameters, but where the number of
parameters grows with the number of training cases. Thus, a parametric approach makes stronger modelling assumptions. For instance, the $k$-nearest neighbour classifier often used in CBR can be considered a non-parametric approximation to a fully probabilistic classifier (see Appendix B) while logistic regression is a parametric probabilistic classifier that makes more model assumptions with corresponding parameters [25]. If a parametric approach fits the data very well, the parametric model is preferred over a non-parametric approach. However, if the data do not follow any identifiable statistical model, then a non-parametric approach, such as CBR, might be better.


### 2.4.3 Anomaly Detection

Anomaly detection is a research field that investigates various ways of identifying cases that deviate substantially from what is considered normal [38]. Since, anomaly detection does not require that all possibly anomalous fault classes are known in advance to be useful, it has become a quite popular approach for fault detection. Instead, it is assumed that the non-anomalous cases are substantially more common than the anomalous cases, and thus, it is possible to distinguish the anomalous cases from the non-anomalous. Similar to machine learning, data-driven anomaly detection approaches can be classified into two major types: model-based and distance-based methods, where the latter corresponds to cased-based or instance-based methods in machine learning. A typically model-based approach is to fit data to a statistical model and then, if the likelihood of a data point is below a threshold, it is considered an anomaly. In contrast, distance-based methods do not necessary make any modelling assumptions, but use the distance to the neighbours as basis for deciding whether a case is anomalous.

In this thesis, we mainly use model-based anomaly detection, and anomaly detection is proposed both for on-board diagnosis and for off-board diagnosis in Paper A, Paper D and Paper E. Paper A and Paper
E use supervised anomaly detection on-board machines, while Paper D and Paper E present an unsupervised approach to anomaly detection of groups of cases for off-board analysis.
Chapter 3

Research Summary

This chapter presents a summary of the research conducted within this thesis. Section 3.1 describes the proposed diagnostic framework with its components. Section 3.2 presents a working prototype of parts of the framework. Section 3.3 describes the used research methodology and the conducted research process.

3.1 Diagnostic Framework

Figure 3.1 shows the proposed fault diagnosis framework. The framework consists of four steps that are divided into an on-board component and an off-board component.
Chapter 3

Research Summary

This chapter presents a summary of the research conducted within this thesis. Section 3.1 describes the proposed diagnostic framework with its components. Section 3.2 presents a working prototype of parts of the framework. Section 3.3 describes the used research methodology and the conducted research process.

3.1 Diagnostic Framework

Figure 3.1 shows the proposed diagnostic framework. The framework consists of four steps that are divided into an on-board component and an off-board component.

Figure 3.1: The proposed fault diagnosis framework.
The proposed approach in Fig. 3.1 is a generic framework that assumes a stream of data that will be classified as faulty or normal, with a severity estimation, a measure of prediction confidence and/or reliability and potentially a set of relevant cases. The first step (1) is to extract relevant features from the signals. Then, next step (2), a local anomaly detector assesses whether the features are anomalous, and thus, potentially faulty by assigning an anomaly score per case. In the third step (3), a stream anomaly detector analyses the stream of cases generated by the previous step (2) in order to detect a anomalous group of the most recent cases. The true anomaly will be detected as a change in the distribution of generated anomaly score or as a change in the anomaly frequency. The third step can also be performed on-board or the machine but then, the anomaly scores cannot be compared to other similar machines. The last step (4) is about analysing the output from the stream anomaly detection in order to find the fault types and estimate the confidence and reliability of the diagnosis. Observe that step 1 and 2 only works on individual cases, but step 2 produces a stream of data with anomaly scores per case that is then analysed in step 3.

The following sections give a summary of the work done in this thesis with respect to the last three steps of the framework. The feature extraction step is not a contribution of this thesis, but we assume that we can extract relevant features. Figure 3.2 shows the relationship between the papers and the proposed framework, while Table 3.1 shows the same relations in chronological order of when the papers were composed.

### 3.1.1 Statistical Model for On-board Fault Detection of Individual Cases

Paper A describes the problem of detecting faults on-board a wheel loader. The idea is to train a lightweight probabilistic classifier in form of logistic regression to distinguish between normal cases and faulty cases. It is assumed that we have a large number of normal cases but a small number of faulty cases. A novel approach to modelling the individual features was also presented. Each feature was modelled independent of the other features using a Gaussian mixture model (GMM) over the non-anomalous cases and then the log-likelihood of the feature-cluster pairs were used as new features for training the anomaly detector. We showed that the logistic regression-based anomaly detector was able to detect individual fault cases, but the GMM did not improve the performance
The proposed approach in Fig. 3.1 is a generic framework that assumes a stream of data that will be classified as faulty or normal, with a severity estimation, a measure of prediction confidence and/or reliability and potentially a set of relevant cases. The first step (1) is to extract relevant features from the signals. Then, next step (2), a local anomaly detector assesses whether the features are anomalous, and thus, potentially faulty by assigning an anomaly score per case. In the third step (3), a stream anomaly detector analyses the stream of cases generated by the previous step (2) in order to detect a anomalous group of the most recent cases. The true anomaly will be detected as a change in the distribution of generated anomaly score or as a change in the anomaly frequency. The third step can also be performed on-board or the machine but then, the anomaly scores cannot be compared to other similar machines. The last step (4) is about analysing the output from the stream anomaly detection in order to find the fault types and estimate the confidence and reliability of the diagnosis. Observe that step 1 and 2 only works on individual cases, but step 2 produces a stream of data with anomaly scores per case that is then analysed in step 3.

The following sections give a summary of the work done in this thesis with respect to the last three steps of the framework. The feature extraction step is not a contribution of this thesis, but we assume that we can extract relevant features. Figure 3.2 shows the relationship between the papers and the proposed framework, while Table 3.1 shows the same relations in chronological order of when the papers were composed.

### 3.1.1 Statistical Model for On-board Fault Detection of Individual Cases

Paper A describes the problem of detecting faults on-board a wheel loader. The idea is to train a lightweight probabilistic classifier in form of logistic regression to distinguish between normal cases and faulty cases. It is assumed that we have a large number of normal cases but a small number of faulty cases. A novel approach to modelling the individual features was also presented. Each feature was modelled independent of the other features using a Gaussian mixture model (GMM) over the non-anomalous cases and then the log-likelihood of the feature-cluster pairs were used as new features for training the anomaly detector. We showed that the logistic regression-based anomaly detector was able to detect individual fault cases, but the GMM did not improve the performance for detecting faults compared to using the original features.

### 3.1.2 Probabilistic Anomaly Aggregation for Group Anomaly Detection

The approach presented in Paper A detects individual fault cases. However, it will probably never have perfect detection, but always a fraction of false positives. This problem is addressed in Paper D and Paper E that describe an approach to detecting anomalous groups of cases using a probabilistic approach to anomaly aggregation. The approach does not assume a statistical model of the individual cases, but uses the ranking of cases from an individual case anomaly detector as only input. Any anomaly detector that produces a ranking as output can be used and thus, we are not forced to use the logistic regression approach from Paper A. Then, given a group of individually ranked anomalies, we can compute the probability of observing an equally or more extreme ranked value with respect to all ranked cases. If the probability is low, an anomalous group is detected. We have showed that the proposed group anomaly detector works better than a Gaussian baseline anomaly detec-
Table 3.1: Papers in chronological order and the related step in the diagnostic framework. Parentheses indicate an indirect relation as shown in Fig. 3.2.

<table>
<thead>
<tr>
<th>Paper</th>
<th>Diagnostic Step</th>
</tr>
</thead>
<tbody>
<tr>
<td>Paper B</td>
<td>(Fault Diagnosis Reliability)</td>
</tr>
<tr>
<td>Paper C</td>
<td>Fault Diagnosis Reliability</td>
</tr>
<tr>
<td>Paper A</td>
<td>Local Anomaly Detection, Fault Diagnosis</td>
</tr>
<tr>
<td>Paper D</td>
<td>(Stream Anomaly Detection)</td>
</tr>
<tr>
<td>Paper E</td>
<td>Stream Anomaly Detection, Fault Diagnosis</td>
</tr>
</tbody>
</table>

...tor in Paper D and that it can be applied to a stream of cases in Paper E.

### 3.1.3 Case-Based Diagnosis of Individual Anomalies

Given that we have detected a set of cases as anomalous and faulty, we would like to classify the individual cases or estimate their severity. Paper A presents a CBR based approach to fault diagnosis where a k-nearest neighbour algorithm was trained to predict the severity of cases. As similarity metric, we used the weighted Manhattan distance and as weights, we used the Maximal Information Coefficient that measures the mutual information between continuous variables. As in case of the anomaly detector, we used the same approach to modelling features using a GMM for each feature independent of the other features over the non-anomalous cases. Thus, cases were compared with respect to how they deviate from each other with respect to the statistical model of the non-anomalous cases (GMM). We showed that the CBR approach could learn to estimate the severity of cases using all the five features that were proposed in the paper, but as in case of the anomaly detector, the GMM did not improve the performance. In Paper E, we used the same setup to CBR for classification, but without the GMMs.

### 3.1.4 Case Stream Fusion for Group Classification

Paper E presents an approach that combines the work in Paper A and Paper D in that we consider a stream of cases instead of individual
cases. By using a large sliding window and a small sliding window with the most recent cases, we can detect when the small window is anomalous compared to the large window using a similar approach as in Paper D. After identifying an anomalous window (a group of cases forming a composite case), the individual cases are classified using a similar CBR approach as in Paper A. Then, we investigate three approaches to fusing the contribution of the classification of the anomalous cases into a single aggregated classification: simple and weighted majority voting and Dempster-Shafer theory based fusion. We showed that fusion is better than using only the classification of the last anomalous case as group classification, and that, weighted majority vote was slightly better than the others for one of three compared data sets.

3.1.5 Reliability Estimation for Diagnosis

Predictions and Method Selection

In the end of a diagnosis, we would like to have a reliability estimation of the predicted class or the severity estimation in order to make a decision of whether to trust the prediction. When using a CBR classification approach it is common to measure the confidence in a prediction as the fraction of nearest neighbours that have the majority class, while many machine learning approaches assign a probability distribution for the prediction. However, in the latter instance, there is no direct link to the cases that were used for training the machine learning approach, so the probability distribution is no better than the fit of the used probability model. Paper B addresses this problem by a CBR based approach for estimating the reliability of a linear regression model, while Paper C extends the approach to classification. The proposed approach (1) uses a statistical approach to similarity metric that directly connects cases to the used probability model, and (2) measures the reliability of the prediction using the local error. For regression, the local mean absolute error was used and for classification, the local accuracy was used. In both cases, a kNN algorithm was trained to estimate the local errors, and we showed that it was indeed possible to learn to estimate the local error in both domains.

In Paper E, the fusion results in a weight distribution over possible fault classes that can be interpreted as a probability distribution and therefore, reliability estimation could be computed using a case base of historical composite cases.
In addition, the proposed approach can also be used for selecting or combining the output of several probabilistic machine learning methods for improving the diagnostic performance. Notice also that the local error can also be used as reliability measure for a pure CBR approach where the similarity metric is already defined.

### 3.2 Prototype Implementation

For demonstration, we implemented a prototype of a part of the proposed framework. It consists of the components shown in Fig. 3.3 and is mainly based on the approach presented in Paper A. The prototype is implemented for detecting oil leakage and clutch slippage in the transmission by analysing data signals from the transmission during gear changes. Clutch slippage is when the time to execute a gear switch takes longer than expected and it is a symptom of lost oil pressure due to oil leakage. However, the focus in this thesis is on detecting and diagnosing oil leakage, not clutch slippage. What is lacking from the full framework is the off-board anomaly detection for streaming data. So, the gear changes are only analysed individually and the stream of gear changes at a collective level is not considered.

![Figure 3.3: The part of the fault diagnosis framework implemented in the prototype.](image)

The prototype was implemented as a client-server architecture shown in Fig. 3.4. The on-board client is pushing raw data signals from the gear changes of the transmission to an off-board diagnostic server.
In the prototype, the on-board client extracts features and computes the probability that a gear change is anomalous using logistic regression trained on all available cases in combination with the Gaussian mixture model trained on the non-anomalous cases. Then, the diagnostic server uses the kNN to make a diagnosis by classifying an anomaly as actually normal or as faulty (oil leakage) using a majority vote and by estimating the severity as the mean valve opening of the most similar cases. In addition, it can compute a confidence in the classification in form of the fraction of nearest neighbour cases with the predicted class. We did not implement the CBR based reliability justification using the local accuracy proposed in this thesis but it could easily have been added to the prototype.

The prototype GUI consists of three views that are used to demonstrate and interact with the prototype: (1) an on-board visualisation view, (2) an off-board anomaly management view and (3) an off-board case-based diagnosis view. Figure 3.5 shows a on-board visualisation of the machine. The upper part shows two meters. The left meter shows a measurement of clutch slippage where values close to 1.0 indicate an ongoing slippage. Similarly, the right shows a meter for indicating anomalies, where values close to 1 indicate detected anomalies. In lower part shows the last 100 gear changes and their probability of being anomalous with a smoothed curve.

Figure 3.6 shows the off-board view of the system. In the upper
part, we can see a list of detected anomalies (using a threshold where probability of being anomalous > 0.9) with timestamps. The lower part shows the raw data signals for each anomaly and it is also possible to show and compare signals from several anomalies. There is also a button that opens a new view for making case-based diagnosis for the anomalies selected from the list.

The case-based diagnosis view is shown in Fig. 3.7. The diagnosed cases are shown in the upper left and the ranked list of related cases for a selected diagnosed case is shown in the upper right. The ranking is ordered from the closest to the most distant case. The lower part shows the signals for the selected related cases so they can be compared. As can be seen, the diagnosed cases are shown with a predicted fault type (class), an estimated severity and a classification confidence.

The prototype was implemented using a combination of the SVALI data stream management system [39], Python\(^1\) (and the Python machine learning library scikit-learn [40] and IPython for command line execu-

\(^1\)http://www.python.org
part, we can see a list of detected anomalies (using a threshold where probability of being anomalous > 0.9) with timestamps. The lower part shows the raw data signals for each anomaly and it is also possible to show and compare signals from several anomalies. There is also a button that opens a new view for making case-based diagnosis for the anomalies selected from the list.

The case-based diagnosis view is shown in Fig. 3.7. The diagnosed cases are shown in the upper left and the ranked list of related cases for a selected diagnosed case is shown in the upper right. The ranking is ordered from the closest to the most distant case. The lower part shows the signals for the selected related cases so they can be compared. As can be seen, the diagnosed cases are shown with a predicted fault type (class), an estimated severity and a classification confidence.

The prototype was implemented using a combination of the SVALI data stream management system [39], Python (and the Python machine learning library scikit-learn [40] and IPython for command line execution.

Figure 3.6: The prototype off-board anomaly management interface.

Figure 3.7: The prototype off-board diagnosis interface.
tion [41]), Java\textsuperscript{2}, Google Web Toolkit\textsuperscript{3} and Google Charts JavaScript library\textsuperscript{4}. The SVALI (Stream VALIdator) is a data stream management system for handling continuous queries and is implemented on top of a object-oriented SQL relational database. SVALI is easily extended with external Python code. The Google Web Toolkit is a library for creating interactive Web Applications using an only Java-approach where Java code is translated into JavaScript code that is executed in a web browser. Google Charts is a JavaScript library for plotting various forms of curves in a browser. Scikit-learn is an advanced machine learning library in Python.

The on-board client is implemented using SVALI and Python. We used the Gaussian mixture model from the scikit-learn machine learning library. The diagnostic server was implemented using Java with a connector to SVALI. The GUI was implemented using Google Web Toolkit and Google Charts. Anomaly detector (logistic regression) and Gaussian mixture model was trained off-line using scikit-learn, but the attribute weights from logistic regression were than exported to and used in a logistic regression implemented in SVALI. The CBR diagnosis was also implemented using SVALI, which has a built in support for the kNN algorithm and enhanced case indexing using X+-tree \cite{42}. So, the cases were stored in SVALI and retrieved using X+-tree indexing. The on-board client and off-board diagnosis server and GUI client were implemented and run on a laptop computer attached to the CAN bus\textsuperscript{5} of a wheel loader.

### 3.3 Research Methodology and Research Process

In this thesis, we have used an exploratory research methodology to identify research problems and then an empirical research methodology to evaluate proposed solutions. For evaluation, we used a combination of proprietary data gathered from real machines and public data sets. In two instances we used artificial or simulated data. For evaluating the

\textsuperscript{2}https://java.com/en/
\textsuperscript{3}http://www.gwtproject.org
\textsuperscript{4}https://developers.google.com/chart/
\textsuperscript{5}The CAN (controller area network) bus is used by the on-board hardware for off-board communication.
The performance of the used algorithms, we followed the common scientific method in machine learning by dividing the data sets into a training set, a validation set (if relevant) and a test set, only reporting the result of the test set in order to ensure the independence between the data an algorithm learned from and the data that it was evaluated on. As evaluation metrics, we used the metrics that best suited the conducted learning task.

The conducted research started out from literature studies that identified that research in CBR lacked a theoretical understanding of the relation between the used similarity metrics and probability theory, which resulted in proposing a statistical definition of a similarity as presented in Paper A, B and C.

Further literature studies also identified that learning systems in general lack intuitive and easy to understand explanation facilities. This observation resulted in the development of the reliability estimation approach described in Paper B and C. In Paper A and Paper B, we used proprietary data set to test the algorithms, while in Paper C, we used a public data set.

In case of Paper A, we started by exploring possible ways of performing fault diagnosis efficiently by supporting both automatic and manual diagnosis. A requirement was that the system should be able to retrieve a set of most similar previous observed cases and thereby, a CBR-based approach was investigated.

The work presented in Paper D was the result of exploring the problem of detecting anomalous groups of cases, where we identified the need for approaches that did not assume a statistical model for the cases. In case of Paper D, the proposed approach was evaluated using an artificial data set with the ground truth available and two industrial data sets lacking ground truth.

When developing the prototype of the framework, we also identified the problem of diagnosing data streams, which resulted in the research presented in Paper E that builds on the approaches presented in Paper A and Paper D. The evaluation in Paper E used a combination of simulation of data streams and three public data sets.
Chapter 4

Abstracts of Included Papers

4.1 Paper A: Fault Diagnosis of Heavy Duty Machines: Automatic Transmission Clutches

Tomas Olsson, Elisabeth Källström, Daniel Gillblad, Peter Funk, John Lindström, Lars Håkansson, Joakim Lundin, Magnus Svensson, Jonas Larsson.

The paper was presented at the workshop on Synergies between CBR and Data Mining at the 22nd International Conference on Case-Based Reasoning (CBRDM’14). Tomas was the main author of the paper. He developed most of the algorithms and implemented the reported experiments using data that was collected by Volvo CE.

Abstract

This paper presents a generic approach to fault diagnosis of heavy-duty machines that combines signal processing, statistics, machine learning, and case-based reasoning for on-board and off-board analysis. The used methods complement each other in that the on-board methods are fast and lightweight, while case-based reasoning is used off-board. 


Chapter 4

Abstracts of Included Papers

4.1 Paper A:

Fault Diagnosis of Heavy Duty Machines: Automatic Transmission Clutches

Tomas Olsson, Elisabeth Källström, Daniel Gillblad, Peter Funk, John Lindström, Lars Håkansson, Joakim Lundin, Magnus Svensson, Jonas Larsson.

The paper was presented at the workshop on Synergies between CBR and Data Mining at the 22nd International Conference on Case-Based Reasoning (CBRDM’14). Tomas was the main author of the paper. He developed most of the algorithms and implemented the reported experiments using data that was collected by Volvo CE.

Abstract This paper presents a generic approach to fault diagnosis of heavy-duty machines that combines signal processing, statistics, machine learning, and case-based reasoning for on-board and off-board analysis. The used methods complement each other in that the on-board methods are fast and lightweight, while case-based reasoning is used off-board.
for fault diagnosis and for retrieving cases as support in manual decision making. Three major contributions are novel approaches to detecting clutch slippage, anomaly detection, and case-based diagnosis that is closely integrated with the anomaly detection model. As example application, the proposed approach has been applied to diagnosing the root cause of clutch slippage in automatic transmissions.

4.2 Paper B:

Case-Based Reasoning for Explaining Probabilistic Machine Learning

Tomas Olsson, Daniel Gillblad, Peter Funk, Ning Xiong.

The paper was published in the International Journal of Computer Science and Information Technology (IJCSIT). Tomas was the main contributor of the paper.

Abstract This paper describes a generic framework for explaining the prediction of probabilistic machine learning algorithms using cases. The framework consists of two components: a similarity metric between cases that is defined relative to a probability model and an novel case-based approach to justifying the probabilistic prediction by estimating the prediction error using case-based reasoning. As basis for deriving similarity metrics, we define similarity in terms of the principle of interchangeability that two cases are considered similar or identical if two probability distributions, derived from excluding either one or the other case in the case base, are identical. Lastly, we show the applicability of the proposed approach by deriving a metric for linear regression, and apply the proposed approach for explaining predictions of the energy performance of households.
4.3 Paper C:

Explaining Probabilistic Fault Diagnosis and Classification using Case-based Reasoning

Tomas Olsson, Daniel Gillblad, Peter Funk, and Ning Xiong.

The paper was presented at the 22nd International Conference on Case-Based Reasoning (ICCBR’14) as a poster. Tomas was the main contributor of the paper.

Abstract  This paper describes a generic framework for explaining the prediction of a probabilistic classifier using preceding cases. Within the framework, we derive similarity metrics that relate the similarity between two cases to a probability model and propose a novel case-based approach to justifying a classification using the local accuracy of the most similar cases as a confidence measure. As basis for deriving similarity metrics, we define similarity in terms of the principle of interchangeability that two cases are considered similar or identical if two probability distributions, derived from excluding either one or the other case in the case base, are identical. Thereafter, we evaluate the proposed approach to explaining the probabilistic classification of faults. We show that with the proposed approach, it is possible to find cases for which the used classifier accuracy is very low and uncertain, even though the predicted class has high probability.
4.4 Paper D:

A Probabilistic Approach to Aggregating Anomalies for Unsupervised Anomaly Detection with Industrial Applications

Tomas Olsson and Anders Holst.

The paper was presented at the 28th International FLAIRS Conference (FLAIRS-28). The idea in the paper is the combined contribution of the two authors. Tomas is the main author, and has designed and performed all experiments.

Abstract This paper presents a novel, unsupervised approach to detecting anomalies at the collective level. The method probabilistically aggregates the contribution of the individual anomalies to detecting significantly anomalous groups of cases. As only input, the proposed method uses a list of cases ranked according to its individual anomaly. Thus, any anomaly detection algorithm can be used for scoring point anomalies. The applicability of the proposed approach is shown by applying it to an artificial data set and to two industrial data sets detecting anomalously moving cranes (using model-based detection) and anomalous fuel consumption of road segments (using instance-based detection).

4.5 Paper E:

Fault Diagnosis via Fusion of Information from a Case Stream

Tomas Olsson and Ning Xiong and Elisabeth Källström and Anders Holst.

The paper was presented at 23rd International Conference on Case-Based Reasoning (ICCBR’15) as a poster. Tomas was the main contributor of the paper, and has designed and performed all experiments.
4.5 Paper E:
Fault Diagnosis via Fusion of Information from a Case Stream

Abstract  This paper presents a novel approach to fault diagnosis applied to a stream of cases. The approach uses a combination of case-based reasoning and information fusion to do classification. The approach consists of two steps. First, we perform local anomaly detection on-board a machine to identify anomalous individual cases. Then, we monitor the stream of anomalous cases using a stream anomaly detector based on a sliding window approach. When the stream anomaly detector identifies an anomalous window, the anomalous cases in the window are classified using a CBR classifier. Thereafter, the individual classifications are aggregated into a composite case with a single prediction using an information fusion method. We compare three information fusion approaches: simple majority vote, weighted majority vote and Dempster-Shafer fusion. As baseline for comparison, we use the classification of the last identified anomalous case in the window as the aggregated prediction.
Chapter 5

Related Work

The research conducted in this thesis spans related research in CBR, machine learning, fault diagnosis, anomaly detection and change detection. To the best of our knowledge, no other work combines all the different aspects that we have investigated in this thesis. We restrict the list of related work to fault diagnosis of machines using artificial intelligence approaches, with a focus on CBR, and then, we end with some related work on explaining diagnosis by computing a reliability measure for a prediction. The related work in this chapter is a complement to the related work presented in each included paper. A good overview on issues in fault diagnosis and detection using intelligent methods is presented in [43].

5.1 CBR Systems

Fault diagnosis has been an application area for CBR since the beginning of the CBR field. Early CBR-based systems for engineering diagnosis are [44], MOLTKE [45, 46] and the successor PATDAX [47]. An early system for automotive troubleshooting is CELIA [48]. INRECA is another project developing fault diagnosis application combining inductive decision trees and CBR [49, 50]. An onboard CBR-based fault diagnosis system for locomotives called ICARUS was presented in [51]. As input, on-board fault messages were used and historical fault log and repair data were used as cases. An extension to [51] was presented in [52] for using the on-board sensor signals in addition to fault messages.
Chapter 5

Related Work

The research conducted in this thesis spans related research in CBR, machine learning, fault diagnosis, anomaly detection and change detection. To the best of our knowledge, no other work combines all the different aspects that we have investigated in this thesis. We restrict the list of related work to fault diagnosis of machines using artificial intelligence approaches, with a focus on CBR, and then, we end with some related work on explaining diagnosis by computing a reliability measure for a prediction. The related work in this chapter is a complement to the related work presented in each included paper. A good overview on issues in fault diagnosis and detection using intelligent methods is presented in [43]

5.1 CBR Systems

Fault diagnosis has been an application area for CBR since the beginning of the CBR field. Early CBR-based systems for engineering diagnosis are [44], MOLTKE [45, 46] and the successor PATDAX [47]. An early system for automotive troubleshooting is CELIA [48]. INRECA is another project developing fault diagnosis application combining inductive decision trees and CBR [49, 50]. An off-board CBR-based fault diagnosis system for locomotives called ICARUS was presented in [51]. As input, on-board fault messages were used and historical fault log and repair data were used as cases. An extension to [51] was presented in [52] for using the on-board sensor signals in addition to fault messages.
However, no details on the integration between the new approach and the CBR system are reported.

A CBR approach to diagnosing the sound of faulty gearboxes is presented in [53, 54, 55]. The approach used frequency and time analysis in form of discrete wavelet transforms and fast Fourier transforms to extract features and thereafter, a CBR approach is used for fault classification given faulty and normal recordings of gearboxes.

There has been extensive research within CBR-based diagnosis of oil well drilling. The CREEK approach was presented in [29]. CREEK is a knowledge intensive CBR approach (KICBR) that makes use of explicit knowledge models in the reasoning process in form of a semantic network. By using an explicit knowledge model, CREEK is able to reason using only a limited number of cases. In [56], a KICBR system that extends the CREEK approach is described that combines model-based diagnosis (a causal model) with a similarity based prediction for improving the accuracy of the drilling problem diagnosis. DrillEdge is a related framework presented in [57] that helps detecting, identifying and list countermeasures for upcoming problems in oil well drilling. The drilling operation is monitored in real time and a graphical GUI shows the operator how historical cases were related to the current situation and the previously applied solutions in similar situations. In relation to the work in this thesis, we have not yet addressed the high level reasoning on how to present potential problems to the operators, but worked on the basic framework for detecting problem on-board a machine. We also have so far taken a more knowledge-light approach in that we use fully automatic means to reasoning. However, a more knowledge-based, high level-reasoning component might be a good complement to the proposed framework.

In [58], an approach to CBR for early classification of faults for time series data is described. It is assumed that there is already a model-based fault detection system that, similar to our local anomaly detector, can indicate when there is a fault. The CBR system thereafter classifies the kind of fault. The authors compare five similarity metrics: three dynamic time warping (DTW) metrics, the Euclidean distance and the Manhattan distance. They show that DTW is better in this application, since it is less sensitive to shifts in the time series. It is also claimed that the CBR approach is better than model-based and machine learning approaches that they tried before. This is an approach that could be investigated for the off-board diagnosis of our work. For instance, it would be possible to
retrieve more data from the machines than used in the initial on-board fault detection for more detailed and precise diagnosis.

A CBR-based diagnosis system was presented in [59]. From about 5 years of customer solution reports, the authors created almost 1000 cases. The reports were a mix of natural language text and different attributes. Experts evaluated the system by assigning a utility measure for the top retrieved cases using 12 different criteria. This showed that the similarity between cases and the new case was clearly correlated with the utility value.

5.2 Hybrid Systems

There are plenty of hybrid systems using combinations of CBR and other learning approaches.

In [60], a CBR system combined with artificial neural networks (ANN) using repair reports as cases is described. The application is to support a help desk to find relevant fixes to problems through a web interface. Two different ANNs were used, a supervised ANN trained for classification and an unsupervised ANN used for clustering. The ANNs were used for case retrieval and the performance was compared to two traditional kNN approaches. The ANNs were faster and more accurate than using the brute force search implemented for the kNN.

A diagnostic framework with a distributed multi-agent system was presented in [61, 62]. The framework consists of three different types of agents: a single-signal diagnostic agent, a multi-signal diagnostic agent and a vehicle diagnostic agent. The single-signal agent classifies segmented data from a single sensor signal as good, bad or unknown. The multi-signal agent uses data from several signals to do the same. For classification they used a set of learned fuzzy rules. Then, the output from the signal agents was collected into a vector that was used as input to the vehicle diagnostic agent. The vehicle diagnostic agent used CBR to propose a diagnosis. Related to our approach, this is very similar in that it is a distributed system, but the papers have a much broader scope in that more kinds of faults are investigated. We differ in that we have only considered faults in the transmission instead of the engine and that we use other kinds of methods. However, a similar approach for segmenting and extracting features should be relevant to use with our approach as well.
In [63], the authors present an hybrid system that combines CBR with a ART-Kohonen neural network (ART-KNN) for diagnosing an electric engine. The ART-KNN in principle groups the cases into clusters with similar cases that are then used for case retrieval and to diagnose new cases. This is a very direct integration of ART-KNN with CBR that is possible because of the use of a distance metric when computing the weights of the ART-KNN.

A CBR approach to fault diagnosis and repair of cars is described in [64, 65, 66]. The authors present a dynamic CBR approach (DCBR) that in contrast to static CBR approaches lets the cases be updated dynamically. In addition, the system uses both sensor data (quantitative data) and textual data (qualitative data) for fault troubleshooting. The proposed approach consists of two steps: (1) find previous cases with similar symptoms to a new problem (described textually), and (2) interactively with a user explore potential fault classes using diagnostic tests and finally select repairs. The approach combines several AI approaches: CBR, text mining and reinforcement learning. CBR and text mining is used for comparing text documents containing information on faults. The approach identifies faults using a industrial language learner that can extract useful features from semi-structured natural language texts. Specifically DCBR is used to retrieve previous symptoms similar to the current symptoms. The user interact with the system by giving it feedback on the outcome of using it and then, reinforcement learning is used for adapting the matching of (1) symptoms to fault classes, (2) fault classes to diagnostic tests, and (3) diagnostic tests to repairs. It differs from our approach so far that we do not use any text mining to extract information for fault diagnosis and troubleshooting, but only focusing on classifying faults directly from measured signals. Adding user interaction to our system could be an obvious extension to our work.

A hybrid diagnostic system for gas turbine trips was presented in [67]. Five diagnosis systems – four rule-based and one CBR-based – were used to detect five different root cause problems. Then a second CBR-based system was used for fusing the output of the five into a single prediction. Each root cause diagnosis system also returns a confidence value in its prediction that was used by the CBR-based fusion. The authors claim that the CBR fusion is better than only using the root cause with the highest confidence. However, the reported experiment indicates that only one additional case would be classified incorrectly by using the maximum confidence. In this thesis, we classify a composite case using
fusion of its individual cases using the prediction with maximum fused weight (similar to a confidence value) (Paper E). As future work, we proposed using a similar approach as a complement as in [67]. The drawback is that for a CBR-based fusion, we need a case base with composite cases.

### 5.3 Other Systems

An anomaly detection system for detecting faults in time series of locomotives is presented in [52]. The proposed system uses, similarly to our method, a sliding window approach for detecting anomalies over data streams in form of time series. The approach consists of two steps. First, for each recorded on-board signal, an anomaly is detected using a non-parametric statistical test (the Wilcoxon rank sum test) comparing rank distribution of a sample of 100 data points from normal operation conditions with the 20 most recent data points. Then, a generalised regression artificial neural network was trained to fuse the result of the statistical tests in order to derive a final prediction. There are several similarities with our approach, but we differ in using a simpler algorithm, logistic regression, for fusing the contribution of each on-board signal, and then, we use the proposed PRAAG algorithm for detecting anomalies over streaming data. However, it would be interesting to compare the Wilcoxon rank sum test with PRAAG in that both uses only the ranking of data points (cases) for assessing the significance of anomalies.

An approach called VEDAS for vehicle-health and driver-status monitoring based on a distributed data mining model is presented in [68]. They use lightweight algorithms to detect parameters out of bounds and unusual driving behaviours. In another related paper, the MineFleet distributed mining platform is presented [69]. The platform is developed for fast computation of correlations, inner-products, and Euclidean distance matrices in a resource-constrained environment from continuous data streams observed on-board a vehicle [70]. In addition, both VEDAS and MineFleet use an off-board server for further analysis of data and by analysing data on-board they substantially decrease the amount of data sent off-board. However, there are no off-board diagnostic tools described for automatically classifying faults.

In [71], the authors compare three approaches to anomaly detection of a city bus fleet in normal traffic that compute the significance of a
deviation (an anomaly score) for a point deviation in form of p-values, that is, the probability of observing a similarly or more extreme value given a null hypothesis, which is in this case, the training cases. The three approaches are the one class support vector machine (OCSVM), conformal anomaly detection (CAD), and the “most central pattern” (MCP) algorithm. Of these three approaches, CAD is the most interesting in relation to this thesis. CAD was presented in a recent PhD thesis [72]. CAD is an approach that is clearly compatible to the CBR approach in that a nearest neighbour approach is utilised for computing the p-value for one case (sample) in relation to the training cases. The p-value is computed as the fraction of all cases with an equal or larger non-conformity measure. The non-conformity measure is the sum of the distance to the $k$-nearest neighbours. Since p-values are defined so that when the null hypothesis is true, then the p-values should be uniform. For aggregating the p-values (point animal score), they compute the p-value (another p-value than before) for a moving window of the p-values for the most recent measurement points. There are similarities to our group anomaly detection approach (PRAAG) in that they use a form of ranking, but only for single cases, while we do it for a group of cases.

5.4 Reliability of Diagnosis

There are a limited number of papers addressing the problem of reliability of predictions for fault diagnosis with intelligent systems. However, there are a couple of related papers that investigate approaches for estimating the reliability of CBR predictions for diagnosis.

In [73], the authors describe an approach for computing the confidence of the best solution of a CBR system. The proposed approach has been applied among others to diagnosing turbine trips [67] and medical equipment [74]. The idea is to use several indicators of confidence from the CBR system such as the average similarity to the best cases and the number/percentage of correct solution among the retrieved cases. In total 12 indicators are presented. In case of discrete confidence values (“low”, “high”), they then train a C4.5 decision tree to predict the confidence for the solution of a new case using the indicator functions. In case of numerical confidence values, they use a fuzzy membership function to compute the values. For optimising the confidence computations, they use a genetic algorithm. In comparison with our approach in Paper B
and Paper C, they compute the reliability of the CBR prediction while we address probabilistic models. The local accuracy used for justifying a prediction in Paper C corresponds to their use of the number/percentage of correct solution among the retrieved cases. However, the idea of using some more indicators of confidence in the estimate of the reliability of a prediction could easily be incorporated in the estimation of the local accuracy used in Paper B so that the local accuracy of cases with more distant neighbours could be lower.

In [75], a CBR-based decision support system for medical treatment is presented. The system helps deciding whether a child should be admitted to the hospital after treatment or be sent home. The system gives an explanation to a recommendation that also includes a confidence in the recommendation. The confidence measure is computed from four indicators of confidence that resembles the indicators in the previous paper. However, the paper turns the problem into a binary classification of high and low confidence. Thus, a threshold is selected for each indicator using leave-one-out classification such that the number of correctly classified cases with high confidence is maximised and the number of incorrectly classified cases with high confidence is minimised [76]. Thereafter, the confidences of the indicators are aggregated to produce a single confidence metric. We differ in our work that we estimate the reliability of probabilistic predictions and not CBR predictions. However, similarly to our work, they also use the local accuracy as one of the confidence indicators (called neighbour accuracy) but it is not entirely clear how it is computed and used in detail. However, in a related paper on spam filtering [77], the authors show that the probabilities of the naive Bayes classifier is not suited for estimating the true accuracy for spam filtering. Thus, this supports our claim that there is a need for a complementing approach for estimating uncertainty in probabilistic machine learning as proposed in Paper B and Paper C. In addition, the approach of Paper C could be extended with selecting thresholds for considering when a confidence is high or low in similar way as in [76].
In this thesis, we have proposed a generic framework for remote diagnosing heavy-duty machines. We have proposed a combination of on-board and off-board processing to address the research questions listed in the introduction. In an implementation of the framework in this thesis, we use a combination of methods from CBR and probabilistic machine learning in order to detect and classify faults.

The framework consists of four steps, where we have proposed solutions for each step that, if put together, can detect and diagnose faults for any machine. The first step is feature extraction from the on-board sensor signals. The second step is to detect when the signals deviates from what is considered normal. For this purpose, we used logistic regression trained to distinguish between normal and anomalous cases. The anomaly detector in the second step results in a stream of cases with an assigned anomaly score. Then, in the third step, a stream anomaly detector is used for filtering out spurious anomalies. The stream anomaly detector uses a sliding window approach where the most recent anomaly scores are used to identify whether the most recent cases are indeed anomalous as a collective and not generated by pure chance. An anomalous group of cases in the sliding window then constitutes a composite case. Last, in the fourth step, a CBR approach is used for classifying the individual cases, which are then fused into an aggregated classification for the composite case. In addition, a reliability estimation of the aggregated classification is performed.
Chapter 6

Conclusions and Future Work

In this thesis, we have proposed a generic framework for remote diagnosing heavy-duty machines. We have proposed a combination of on-board and off-board processing to address the research questions listed in the introduction. In an implementation of the framework in this thesis, we use a combination of methods from CBR and probabilistic machine learning in order to detect and classify faults.

The framework consists of four steps, where we have proposed solutions for each step that, if put together, can detect and diagnose faults for any machine. The first step is feature extraction from the on-board sensor signals. The second step is to detect when the signals deviates from what is considered normal. For this purpose, we used logistic regression trained to distinguish between normal and anomalous cases. The anomaly detector in the second step results in a stream of cases with an assigned anomaly score. Then, in the third step, a stream anomaly detector is used for filtering out spurious anomalies. The stream anomaly detector uses a sliding window approach where the most recent anomaly scores are used to identify whether the most recent cases are indeed anomalous as a collective and not generated by pure chance. An anomalous group of cases in the sliding window then constitutes a composite case. Last, in the fourth step, a CBR approach is used for classifying the individual cases, which are then fused into an aggregated classification for the composite case. In addition, a reliability estimation of the ag-
Aggregated classification could be computed given a case base of historical composite cases.

Each step in the framework should be implemented using the best available method for detecting and diagnosing faults with respect to the type of sensor signals and the type of faults. For instance, logistic regression is a relatively simple method but requires a training set of labelled cases. It could be replaced with another, resource efficient unsupervised anomaly detection method such as a simple statistical model or we could replace it with an artificial neural network with a limited number of neurons. Next, the stream anomaly detector could be implemented using another similar approach for change detection or concept drift detection. For the fault classification, we could also use another type of classifier, assuming that it can provide a weight distribution over the confidence of each possible fault type to be used in the fusion into a single classification. For instance, logistic regression or a naive Bayesian classifier could be used. Yet, CBR is the natural choice when retrieval of cases is important for supporting manual investigations. However, by using the approach proposed in this thesis in Paper B and Paper C for retrieving cases similar with respect to a probability model, we can also support case retrieval for probabilistic methods. So, the simple kNN approach used in this thesis for classification could also easily be replaced by a probabilistic method.

As future work, we should compare the proposed approaches with other similar approaches in order to investigate the best method in each case. However, many alternative methods need more data than was available in this thesis. For instance, unsupervised methods for anomaly detection require a large number of cases, although not necessarily labelled, that can be costly to collect from a machine. In addition, we would like to have a large set of data collected from many machines and with many different kinds of faults. This also means that we need an approach that actually can manage large data sets; so one research direction would be to scale the proposed approaches to a big data platform.

Another potential future research direction is to actually investigate approaches to detecting new fault types. In the current approach, we assume that cases can be diagnosed but we do not deal with fault of types not previously discovered. A novelty detection (or an anomaly detection) method could be applied to identify new fault classes.

Further, the on-board diagnosis should be able to automatically manage as many fault types as possible and only leave the hard cases to the
off-board system. For instance, we could train a set of classifiers to detect and classify anomalies and only those anomalies that no classifier identifies as a fault are sent off-board for analysis. Also, ensemble learning could be applied both for on-board fault detection and for the off-board fault classification in order to improve the diagnosis accuracy.

We could also, as future work, improve the reliability of the diagnosis by also consider other confidence indicators than local accuracy as described in the chapter on related work. For instance, the similarity or distance to cases should also be taken into account in the reliability estimation, and the idea of identifying the predictions with high confidence by using an optimised decision threshold is also worth investigating.
Bibliography


Bibliography


Appendices
Chapter A. The $k$-Nearest Neighbour Algorithm

Appendix A

The $k$-Nearest Neighbour Algorithm

The $k$-nearest neighbour algorithm (kNN) is a very simple and intuitive algorithm for real valued predictions (regression) and classification [22, 25]. The idea is that a new value can be approximated by neighbouring values located close to the new value.

Let a case be represented as a pair $(\mathbf{x}, y)$ with vector of real valued features $\mathbf{x} = \{x^1, x^2, \ldots, x^m\}$ of size $m$ and a target value $y$, and assume that we have a distance $d(\mathbf{x}_i, \mathbf{x}_j)$ that measures the closeness between two cases $(\mathbf{x}_i, y_i)$ and $(\mathbf{x}_j, y_j)$. Then, the kNN algorithm is used as follows for predicting the real value $\hat{y}$ of a new case $(\mathbf{x}', \hat{y})$:

1. Find the set of $k$ closest previous cases with respect to distance metric $d$ (denoted by $\mathcal{K}$)

2. Estimate $\hat{y}$ using the mean of the previous cases:

$$\hat{y} = \frac{1}{k} \sum_{j \in \mathcal{K}} y_j$$  \hspace{1cm} (A.1)

In case of predicting a class instead of a real value, the second step above is replaced with a majority vote:

$$\hat{y} = \arg \max_{c \in \mathcal{C}} \sum_{j \in \mathcal{K}} I(y_j = c)$$  \hspace{1cm} (A.2)

where $\mathcal{C}$ denotes the set of classes and $I$ is the indenticator function that is 1 if the argument is true and zero otherwise.

The most commonly used distance is the Euclidean distance defined as follows:

$$d_e(\mathbf{x}_i, \mathbf{x}_j) = \sqrt{\sum_{l=1}^{m} (x^l_i - x^l_j)^2}$$

Another commonly used distance is the Manhattan distance:

$$d_m(\mathbf{x}_i, \mathbf{x}_j) = \sum_{l=1}^{m} |x^l_i - x^l_j|$$

Since the scaling of the features values will affect the importance of a feature in the distance metric, one usually performs some form of normalisation. For instance, in this thesis we normally scale each feature to have a mean of zero and a standard deviation of one, but other types of normalisation are also possible, such as dividing with the difference between max and min values of each feature. In addition, it is common to multiply each feature with a weight that reflects its relative importance for predicting the target value $\hat{y}$. In this thesis, we use the mutual information between each feature and the $y$ as a weight. We compute the mutual information using the maximum information coefficient that is described in a later section. Thus, in case of the Euclidean distance the final distance is the weighted Euclidean distance:

$$d_e(\mathbf{x}_i, \mathbf{x}_j) = \sqrt{\sum_{l=1}^{m} (\mathbf{w}_l(x^l_i - x^l_j))^2}$$

where $\mathbf{w}$ is a vector of weights that includes weight of importance and normalisation.

Regarding the relation to CBR where we instead of a distance use a similarity metric, it can be noted that similarity and distance are coupled concepts in that a similarity can easily be created from a distance and vice versa. A similarity measure $s$ is commonly measured as a real value $s \in [0, 1]$ where $s = 0$ means no similarity and $s = 1$ means identity, while a distance $d$ is a real value $d \in [0, 1]$ where $d = 0$ means identity and $d > 0$ means a degree of dissimilarity. So, by a suitable transformation, for instance, $s = \frac{1}{1+d}$ or $s = e^{-d^2}$, it is possible to turn any distance to a similarity measure. Therefore, in the current work, we do not always distinguish between a similarity and a distance but assume that a proper transformation can be found if necessary.
where $C$ denotes the set of classes and $I$ is the identicator function that is 1 if the argument is true and zero otherwise.

The most commonly used distance is the Euclidean distance defined as follows:

$$d_e(\vec{x}_i, \vec{x}_j) = \sqrt{\sum_l (x^l_i - x^l_j)^2}$$

Another commonly used distance is the Manhattan distance:

$$d_m(\vec{x}_i, \vec{x}_j) = \sum_l |x^l_i - x^l_j|$$

Since the scaling of the features values will affect the importance of a feature in the distance metric, one usually performs some form of normalisation. For instance, in this thesis we normally scale each feature to have a mean of zero and a standard deviation of one, but other types of normalisation are also possible, such as dividing with the difference between max and min values of each feature. In addition, it is common to multiply each feature with a weight that reflects its relative importance for predicting the target value $\hat{y}$. In this thesis, we use the mutual information between each feature and the $y$ as a weight. We compute the mutual information using the maximum information coefficient that is described in a later section. Thus, in case of the Euclidean distance the final distance is the weighted Euclidean distance:

$$d_e(\vec{x}_i, \vec{x}_j) = \sqrt{\sum_{l=1}^{K} \omega^l (x^l_i - x^l_j)^2}$$

where $\vec{\omega}$ is a vector of weights that includes weight of importance and normalisation.

Regarding the relation to CBR where we instead of a distance use a similarity metric, it can be noted that similarity and distance are coupled concepts in that a similarity can easily be created from a distance and vice versa. A similarity measure $s$ is commonly measured as a real value $s \in [0, 1]$ where $s = 0$ means no similarity and $s = 1$ means identity, while a distance $d$ is a real value $d \in [0, \infty)$ where $d = 0$ means identity and $d > 0$ means a degree of dissimilarity. So, by a suitable transformation, for instance, $s = \frac{1}{1+d}$ or $s = e^{-d^2}$, it is possible to turn any distance to a similarity measure. Therefore, in the current work, we do not always distinguish between a similarity and a distance but assume that a proper transformation can be found if necessary.
As noted in Sect. 2.4.2, we can consider kNN as an approximation of probabilistic learning. In short: $y$ is estimated using the most probable value of the local probability distribution $\hat{p}(y|x', k, D)$ estimated from the closest neighbours around $\bar{x}'$ [25]. For more details, let us take one step back and look at the full theoretical generative probability distribution for a case as follows:

$$p(\bar{x}, y) = p(y|x) p(\bar{x}) \tag{B.1}$$

where $p(\bar{x})$ denotes the (possibly unknown) generative probability distribution over the cases, that is, the probability of picking a $x$, and $p(y|x)$ is the probability of picking any possible $y$ given $\bar{x}$ as input. Then, in probabilistic terms, kNN can then be reformulated into these three steps to approximate the full generative distribution from above as follows:

1. First, $\mathcal{K}$ is a sample of cases drawn from the distribution of cases $p(\bar{x})$ close to $\bar{x}'$ as measured using the distance $d$

2. Then, a maximum likelihood estimation\(^1\) (MLE) is done that local-

\(^1\)Maximum likelihood estimation is an approach to parameter estimation that picks the parameters of probability striation that maximise the probability of the data.
The kNN as an Approximation to Probabilistic Learning

As noted in Sect. 2.4.2, we can consider kNN as an approximation of probabilistic learning. In short:

\[ y \] is estimated using the most probable value of the local probability distribution \( \hat{p}(y | \tilde{x}, k, D) \) estimated from the closest neighbours around \( \tilde{x}_0 \) [25]. For more details, let us take one step back and look at the full theoretical generative probability distribution for a case as follows:

\[ p(\tilde{x}, y) = p(y | \tilde{x}) p(\tilde{x}) \] (B.1)

where \( p(\tilde{x}) \) denotes the (possibly unknown) generative probability distribution over the cases, that is, the probability of picking a \( x \), and \( p(y | \tilde{x}) \) is the probability of picking any possible \( y \) given \( \tilde{x} \) as input. Then, in probabilistic terms, kNN can then be reformulated into these three steps to approximate the full generative distribution from above as follows:

1. First, \( K \) is a sample of cases drawn from the distribution of cases \( p(\tilde{x}) \) close to \( \tilde{x}_0 \) as measured using the distance \( d \).

2. Then, a maximum likelihood estimation (MLE) is done that locally approximate the posterior distribution denoted as \( \hat{p}(y | \tilde{x}, k, D) \).

3. Last, the most probable value is picked such that

\[ \hat{y} = \arg \max_y \hat{p}(y | \tilde{x}, k, D). \]

For classification, we can approximate \( \hat{p}(y = c | \tilde{x}') \approx \frac{1}{k} \sum_{j \in K} I(y_j = c) \) that corresponds to the MLE for the Bernoulli distribution in case of binary classification [25]. Equation A.2 is then equivalent to \( \arg \max_{c \in C} \hat{p}(y = c | \tilde{x}') \) since \( k \) is scaling the result for all classes and thus do not effect the maximum value. For real value prediction, \( y \) can be considered approximately normally distributed close to \( \tilde{x}' \) so that \( \hat{p}(y | \tilde{x}') \approx N(y | \mu, \sigma) \) where \( \mu = \frac{1}{k} \sum_{j \in K} y_j \) comes from using MLE for the normal distribution while \( \sigma \) is not necessary for computing the estimation. Equation A.1 is then also equivalent to \( \arg \max_y \hat{p}(y | \tilde{x}'). \)