Deep learning applications for condition monitoring of rotating systems

Jesse Miettinen
Deep learning applications for condition monitoring of rotating systems

Jesse Miettinen

A doctoral thesis completed for the degree of Doctor of Science (Technology) to be defended, with the permission of the Aalto University School of Engineering, at a public examination held at the lecture hall K216 of the school on 20th October 2023 at 12:00.

Aalto University
School of Engineering
Department of Mechanical Engineering
Mechatronics Research Group
**Supervising professors**
Assistant Professor Raine Viitala
Aalto University School of Engineering
Department of Mechanical Engineering
Finland

**Thesis advisors**
D.Sc. Tuomas Tiainen
Aalto University School of Engineering
Department of Mechanical Engineering
Finland

**Preliminary examiners**
Professor Robert Gao
Case Western Reserve University
Department of Mechanical and Aerospace Engineering
USA

Associate Professor Roel Pieters
Tampere University
Automation Technology and Mechanical Engineering
Finland

**Opponents**
Professor Robert Gao
Case Western Reserve University
Department of Mechanical and Aerospace Engineering
USA

---

Aalto University publication series
**DOCTORAL THESES** 148/2023

© 2023 Jesse Miettinen

ISSN 1799-4934 (printed)
ISSN 1799-4942 (pdf)

Unigrafia Oy
Helsinki 2023

Finland
Abstract

Vibration-based condition monitoring can provide crucial information of the performance and health state of rotating machines for maintenance scheduling. Condition-based maintenance schedules can increase the reliability and useful lifetime of rotating machines and reduce maintenance downtime. Deep learning models show great promise for automating vibration-based condition monitoring systems. These models can be optimised for accurate and precise fault detection, fault diagnosis and virtual sensing functions, for example. Automating these functions can, for instance, reduce the manual labour required from experts, decrease the time from a fault occurring to the corrective maintenance action, and enable the condition monitoring of large machine fleets.

To this end, the research improved the current deep learning-based condition monitoring models, and studied best practices for the application of these models for vibration analysis. A virtual sensor application based on long short-term memory (LSTM) and a training procedure for largely varied operating conditions were demonstrated. The virtual sensor application for a rotor system and the training procedure were evaluated with vibration data covering largely varied operating speeds and support stiffnesses. Fault diagnosis models based on one-dimensional convolutional neural networks (1D CNN) were optimised and compared with vibration data acquired with different sensors mounted on many locations on a drivetrain. The feature normalisation of 1D CNNs was investigated with vibration datasets acquired under laboratory and real operating conditions.

The results demonstrate that deep learning models for vibration-based condition monitoring can learn to function over a large span of operating conditions. Simultaneously, the results also suggest that the training dataset should contain vibration samples from the desired operating condition range. The results imply that torsional vibration can be an effective data source for the fault diagnosis of rotating machines. Furthermore, the results show that the type, mounting location, and number of vibration sensors influence the model performance significantly. In addition, the results related to the feature normalisation experiments on 1D CNN-based models indicate that the current algorithms can still be improved.

Overall, the findings in this research are relevant to the future research and development projects towards deep learning-based condition monitoring systems for rotating machines. Moreover, the developed applications and the findings related to the best optimisation practices can contribute to automatic condition monitoring systems providing timely information for maintenance planning. This in turn can reduce the downtime and maintenance costs of rotating machinery.

Keywords Condition Monitoring, Rotating Machines, Deep Learning

Author Jesse Miettinen

Name of the doctoral thesis Deep learning applications for condition monitoring of rotating systems

Publisher School of Engineering

Unit Department of Mechanical Engineering

Series Aalto University publication series DOCTORAL THESES 148/2023

Field of research Mechanical Engineering (ENG206Z)

Manuscript submitted 7 April 2023 Date of the defence 20 October 2023

Permission for public defence granted (date) 16 June 2023 Language English

Monograph Article thesis Essay thesis


ISSN (printed) 1799-4934 ISSN (pdf) 1799-4942

Location of publisher Helsinki Location of printing Helsinki Year 2023

Tekijä
Jesse Miettinen

Väitöskirjan nimi
Neuroverkkomenetelmät pyörivien koneiden kunnonvalvonta

Julkaisija
Insinööritieteiden korkeakoulu

Yksikkö
Konsteknikinan laitos

Sarja
Aalto University publication series DOCTORAL THESES 148/2023

Tutkimusala
Konsteknikka (ENG206Z)

Käsikirjoituksen pvm
07.04.2023

Väitöspäivä
20.10.2023

Väittelyluvan myöntämispäivä
16.06.2023

Kieli
Englanti

Monografia
Artikkeliväitöskirja
Esseeväitöskirja

Tiivistelmä

Tässä tutkimuksessa kehitettiin syväoppimisovelluksia värähtelyn analysointiin. Työ painottui sekä kunnonvalvonta soveltuvien syväoppimismallien kehittämiseen, että parhaimpien optimointimenetelmien löytämiseen. Tutkimuksessa kehitettiin syväoppimiseen perustuvaa virtuaalinen sensori ja optimointimenetelmä. Virtuaalisensoriosovelluksen ja optimointimenetelmän toimivuus validoidiin datalla, joka sisälsi laajalti vaihtelevilla pyöräismoneuksilla ja tuennanäökkyyksillä kerättyjä paperikoneetelantä värähtelynätteitä. Tutkimuksessa vertaillisikin myös 1D konvolutoineuroverkkoihin perustuvia viikadioagnostiikkaamalleja, jotka optimoitiin eri kohdista voimansiirtolinjaen kerättyllä datalla. 1D konvolutoineuroverkkomallien piirteiden normalisointia tutkittiin myös värähtelydatalla, jota oli kerätty laboratorio olokuvahtia, sekä oikeassa käytössä.

Tutkimuksen tulokset osoittivat, että syväoppimismallit voivat analysoida pyörivien koneiden värähtelyä huomioimien ympäristön ja käytön aiheuttamat värähtelyyn muutokset. Samalla tulokset myös osoittivat, että syväoppimismallin tarkkuus ja luotettavuus ovat korkeimmat värähtelyalueilla, joka vastaa koulutusdation sisältämää värähtelyä. Lisäksi tutkimustulokset osoittivat vääntöväärähtelyn olevan erinomainen mitattava suure pyörivien koneiden ja erityisesti hammaspysyrien viikadioagnostiikkaan varten. Tulokset indikoivat myös, että tarkemmalta konvolutoverkkojen piirteiden normalisoinnilla voidaan yhä parantaa viikadioagnostiikkaamalleja.

Tämän tutkimuksen löydökset ovat relevantteja syväoppimismalleihin perustuvien pyörivien koneiden kunnonvalvontajärjestelmien tutkimus- ja kehitysprojekteille. Tutkimuksessa kehitetyt sovellukset ja niiden optimointiin liittyvät parhaat käytännöt voivat parantaa nykyisiä automaattisia kunnonvalvontajärjestelmiä tarjoten tarkempia ja nopeita värähtelyanalysejä. Ajankohtaisella informaatiolla koneiden toiminnasta ja kunnosta huoltoonmuitelminen päivittäminen ja toteuttaminen helpottuvat.

Avainsanat
Kunnonvalvonta, Pyörivät Koneet, Syväoppiminen

ISBN (painettu)
978-952-64-1432-4

ISBN (pdf)
978-952-64-1433-1

ISSN (painettu)
1799-4934

ISSN (pdf)
1799-4942

Julkaisupäivity
Helsinki

Painopäivity
Helsinki

Vuosi
2023

Sivumäärä
164

urn
Preface

Before all else, I wish to underline the significance of the input of our research group to this work. Without the group, the studies included in this dissertation would have been much constrained. The test rigs and the measured data that were made available were invaluable. In addition, initiating the research concerning deep learning and rotating machines with zero experience now seems like an obvious choice, but in 2019 it was a bold move. Therefore, this work should be considered as an achievement of our research group, a group of people who I also wish to express gratitude in this preface.

To me, this dissertation is the unlikely culmination of a path I naively set up in 2018 after deciding to return to full-time studies. The plan was to learn programming, put all the data science and machine learning courses to the study plan, graduate as a Master in Mechanical Engineering, and then, make a living as an AI project salesman. This plan felt like quite far-fetched and difficult. The courses were challenging and dropping out dropped in my mind several times. Thanks to ”ML harkat” study group, all was going according to the plan for the first semester until Raine and Tuomas intercepted it. They offered a Master’s thesis position in an interesting project, with one condition: say ‘yes’ or at least ‘maybe’ for continuing to the Doctoral studies. The tactical ‘maybe’ turned into the unlikely ‘yes’ during the time working in the research group.

Along the road towards finishing this dissertation, there has been many ups and downs, as there are with any dissertation. At times, the despair of having the manuscripts in eternally long rejection cycles felt overwhelming. Fortunately, there were many people around for support during the moments of despair. First and foremost, the supervisor and the advisor of my Doctoral thesis, Prof. Raine Viitala and Dr. Tuomas Tiainen, respectively. Raine was the best possible supervising professor I can imagine. He never showed doubt towards my work, nor did he push to work on unnecessary stuff. Tuomas was also a tremendous advisor. His writing tips and comments on many of my manuscripts did not only improve my writing but also my thinking. Both, Raine and Tuomas were always available
for assisting in any way necessary and could often quickly understand a complex new topic and then share helpful thoughts.

Then there were many people who were able to provide crucial measured data for the studies in this work. Dr. Risto Viitala provided the rotor and bearing vibration dataset for the virtual sensor experiments. Sampo Haikon, Dr. Ivar Koene and Jouni Pekkarinen designed and built a thruster test rig and then Sampo Haikonen conducted measurements that were crucial for the gear fault diagnosis experiments. Joni Keski-Rahkonen and the condition monitoring team in Kongsberg Maritime Finland Oy provided large amounts of vibration data from real operating conditions for the study concerning fault diagnosis with whitened features. Also, I’d like to thank Dr. Andre Böhme for sharing his thoughts on gear faults during the lengthy revision process of the article concerning gear fault diagnosis.

I also wish to express gratitude towards the researchers in our group. Our small AI team: Aku Karhinen and Aleksanteri Hämäläinen were great additions to the group. It was fun to work with you! With Sampo Laine, who approached vibration analysis from the modelling perspective, we had many fruitful conversations. Overall, I enjoyed spending time with the research group, although the 10:30 am lunches were not designed for my rhythm.

Thank you Prof. Roel Pieters and Prof. Robert Gao for pre-examining this dissertation. It was a delightful surprise that the dissertation manuscript was accepted without corrections. I got to move on to other projects in a more-relaxed schedule during the summer of 2023!

It is also important to acknowledge that this dissertation would not have started nor finished without funding. I worked under many research projects funded by the Academy of Finland and Business Finland. The research projects were Reboot IoT, AI-ROT and GOOD. For the final year, the Deans funding was essential for finishing the thesis. In addition, I made full use of Triton, the Aalto high-performance computing cluster coordinated by the School of Science.

Finally, I would like to thank my family for all the support. Kirsi, my mom, has supported me throughout the almost 25 years of studying. I was never forced to do my homework, and I always had the freedom to handle studying related matters independently. Perhaps this is the reason it took almost 25 years to finish studying. A very important person during my Master's and Doctoral studies was Irina. Without her support, I would have likely burnt out, and for sure, enjoyed the previous 5 years way less.

Agia Galini, Greece, September 20, 2023
Contents

Preface 1

Contents 3

List of Publications 5

Author's Contribution 7

Abbreviations 9

Symbols 11

1. Introduction 13
   1.1 Research problem ............................. 15
   1.2 Objectives of the research ..................... 17
   1.3 Limitations of the research .................... 18
   1.4 Research methods ............................. 19
   1.5 Scientific contribution ....................... 20

2. State-of-the-art 21
   2.1 Rotor system condition monitoring ............. 21
       2.1.1 Vibration of rotating systems ............ 22
       2.1.2 Vibration data acquisition ............... 26
       2.1.3 Rotating system failures .................. 27
   2.2 Data-driven virtual sensors .................... 28
   2.3 Intelligent fault diagnosis models ............. 30
       2.3.1 Open source datasets ..................... 31
       2.3.2 Machine learning-based fault diagnosis models 32
       2.3.3 Deep learning-based fault diagnosis models 33
       2.3.4 Training paradigms ....................... 35

3. Materials and Methods 41
   3.1 LSTM-based vibration estimator from indirect measurements 41
       3.1.1 Rotor vibration data description ......... 43
## Contents

3.1.2 Training procedure for generalisation over many operating conditions .................................... 44  
3.2 Torsional and lateral vibration-based intelligent fault diagnosis ................................................. 45  
  3.2.1 Description of the test rig and the fault data ...... 47  
  3.2.2 Model architectures and optimisation .............. 49  
  3.2.3 Description of ablation studies ...................... 53  
3.3 Whitening of CNN-based fault diagnosis model features ......................................................... 55  
  3.3.1 Fault data description ................................. 56  
  3.3.2 Implicit and approximate whitening of 1D CNN features ................................................. 58

4. Results .................................................. 63  
  4.1 Estimation accuracy of LSTM-based virtual sensor ............................ 63  
  4.2 Torsional vibration data for CNN-based gear fault diagnosis ............................ 66  
  4.3 Fault diagnosis accuracy of models with network deconvolution ......................... 69

5. Discussion .............................................. 73  
  5.1 Scientific and practical impact .............. 76  
  5.2 Future research ................................. 79

6. Conclusions .......................................... 81

Publications ........................................... 97
This thesis consists of an overview and of the following publications which are referred to in the text by their Roman numerals.


Author’s Contribution

Publication I: “Bidirectional LSTM-Based Soft Sensor for Rotor Displacement Trajectory Estimation”

The author conceptualised the research together with Tuomas Tiainen, Kari Hiekkonen and Raine Viitala. The data was acquired by Risto Viitala. The author curated the data with the Tuomas Tianen. The author was responsible of the software, formal analysis of the results and writing the manuscript. The co-authors assisted in revising the manuscript.

Publication II: “Comparing torsional and lateral vibration data for deep learning-based drive train gear diagnosis”

The author conceptualised the research together with Joni Keski-Rahkonen. Ivar Koene and Sampo Haikonen were responsible of designing and constructing the test rig. Sampo Haikonen conducted the measurements and acquired the data with the test rig. The author implemented the software required for the experiments. The author was responsible of the formal analysis of the results. The author wrote the manuscript. The co-authors assisted in revising the manuscript.

Publication III: “Whitening CNN-Based Rotor System Fault Diagnosis Model Features”

The author conceptualised the research together with Riku-Pekka Nikula and Joni Keski-Rahkonen. The author was responsible of the curation of the data related to both datasets. Riku-Pekka Nikula and Joni Keski-Rahkonen assisted in the data curation related to the azimuth thrusted dataset. Riku-Pekka Nikula contributed to the data preprocessing algo-
Author's Contribution

Algorithm for the azimuth thruster dataset. The author designed the methodology and implemented the software. The author was responsible of the formal analysis of the results. The author wrote the manuscript, and was responsible for revising it. The co-authors assisted in revising the manuscript.
### Abbreviations

**1D CNN** 1-Dimensional Convolutional Neural Network  
**AE** Autoencoder  
**AI** Artificial Intelligence  
**BCE Loss** Binary Cross-entropy Loss  
**BN** Batch Normalisation  
**BPFI** Ball Pass Frequency of the Inner Race  
**BPFO** Ball Pass Frequency of the Outer Race  
**BSF** Ball Spin Frequency  
**CBM** Condition-based Maintenance  
**CE Loss** Cross-entropy Loss  
**Conv1D** 1-Dimensional Convolutional Layer  
**CPM** Center Point Movement  
**DBN** Deep Belief Network  
**DFT** Discrete Fourier Transform  
**DWT** Discrete Wavelet Transform  
**EMD** Empirical Mode Decomposition  
**FCL** Fully-connected Layer  
**FT** Fourier Transform  
**FTF** Fundamental Train Frequency  
**GAN** Generative Adversarial Network
Abbreviations

GMF  Gear Meshing Frequency
hp   Horsepower
IFD  Intelligent Fault Diagnosis
im2col  Image to Column Transform
k-NN K-Nearest Neighbours
LSTM Long Short-Term Memory
MAE  Mean Absolute Error
ML   Machine Learning
MLP  Multilayer Perceptron
MSE  Mean Squared Error
ND   Network Deconvolution
RDConv1D Dilated 1-Dimensional Convolutional Layer with Residual Connection
RMS  Root Mean Square
RNN  Recurrent Neural Network
RL   Reinforcement Learning
RPM  Rounds Per Minute
RUL  Remaining Useful Lifetime
SAE  Sparse Autoencoder
SDA  Stacked Denoising Autoencoder
SGD  Stochastic Gradient Descent
SRDCNN Stacked Residual Dilated Convolutional Neural Network
STFT Short-time Fourier Transform
SVM  Support Vector Machine
TE   Transmission Error
TICNN Convolution Neural Networks with Training Interference
VAE  Variational Autoencoder
WDCNN Deep Convolutional Neural Network with Wide First-layer Kernels
WT   Wavelet Transform
Symbols

\( \mathcal{D} \) Dataset

\( D_{ball} \) Bearing ball diameter

\( D_{pitch} \) Average diameter of the bearing inner and outer races

\( \mathcal{D}_{query} \) Query dataset

\( \mathcal{D}_S \) Source dataset

\( \mathcal{D}_{support} \) Support dataset

\( \mathcal{T}_T \) Target dataset

\( \beta \) Angle between ball and race groove

\( f_b \) Bearing ball frequency

\( f_c \) Critical speed

\( f_{ca} \) Bearing cage frequency

\( f_g \) Gear rotating speed

\( f_i \) Bearing inner ring frequency

\( f_{ir} \) Base frequency of a bearing inner ring

\( f_{or} \) Base frequency of a bearing outer ring

\( f_r \) Shaft speed

\( \mathcal{L} \) Error function

\( N_t \) Number of teeth in a gear

\( \mathcal{T}_S \) Source task

\( \mathcal{T}_T \) Target task
Symbols

\( z_i \)  Number of undulations in bearing inner ring
\( z_o \)  Number of undulations in bearing outer ring
\( \beta \)  Angle between the race groove and ball
\( \phi \)  Load angle from the radial plane
\( \theta \)  Parameters of neural network
\( \omega \)  Hyperparameters
\( \omega_i \)  Angular velocity of the bearing inner ring
1. Introduction

After the invention of fire, one of the most important innovations of mankind is the exploitation of rotating devices. Rotation is movement around an axis. Rotating systems were, most likely, first designed for vehicular purposes. For example, carts on wheels moved people and goods along the ancient roads of the copper age. Ever since, rotating mechanical systems have been present in the daily lives of most humans. Consequently, humanity has developed almost an endless number of elegant new rotating system designs. Today, the modern day car(t) transmits power from the engine to the wheels with powertrains. Furthermore, most green power plants, such as nuclear power plants, transform thermal energy of heated steam into electric energy by rotating a turbine and a generator. Moreover, numerous clocks transform the battery power to the precise movement of the hand using complex gear systems. All these crucial rotating systems are most often used in the modern world without us, the users, even acknowledging, unless they malfunction.

The present rotating systems transmitting mechanical power typically rely on components such as shafts, bearings, gears and couplings. The currently available design and manufacturing techniques enable the production of rotating systems with components that can bear high loads and rotating speeds. That is, cruise ships weighing over tens of millions of kilograms are moved through the water with thrusters. Simultaneously, steam turbines may rotate at speeds of thousands of revolutions per minute. Despite the sophisticated design, manufacturing and assembly methods available, the rotating components may still fail. Mechanical components will eventually wear after exhaustive use. Furthermore, hazardous environments such as salt water may corrode the components prematurely. In addition, numerous other failure modes related to rotating devices exist. Most often, all failures are undesirable, and sometimes, even catastrophic if safety of people or the environment is compromised.

Fortunately, failures of mechanical components can often be avoided by applying a suitable maintenance strategy. Many sophisticated maintenance strategies tailored for specific rotor systems and their requirements
exist. Regardless of the maintenance strategy, unless it is run-to-break (and then replace), the maintenance schedules of any rotating machine can be optimised dynamically based on the most recent observations of the machine condition [1]. With accurate and continuous condition monitoring, faults can be detected and diagnosed and an estimate of the remaining useful lifetime (RUL) can be made. With such information regarding any machine, timely maintenance actions can be scheduled and performed. With timely maintenance actions, the frequency of unexpected failures likely decreases, and thus, the safety and reliability of the rotating machines increase.

Numerous different condition monitoring systems for rotating machines have been developed. One of the distinguishing features between condition monitoring systems relates to the hardware used for data acquisition. The most common data acquisition systems used for condition monitoring typically include sensors for vibration, oil or temperature measurements [1]. Therefore, much research has been conducted related to these types of condition monitoring systems. For example, numerous studies have contributed to the development of vibration-based condition monitoring [2–8]. In addition, the lubricant oil condition may reveal a malfunctioning component [1, 9, 10]. Furthermore, a great majority of cars include an engine temperature indicator in the dashboard next to the gasoline level and speed indicators.

Although oil and thermographic analysis can reveal faults quite reliably, vibration analysis may provide some advantages for condition monitoring. First, vibration patterns may reveal very specific information regarding a faulty component, which enables accurate diagnosis [5, 11, 12]. Second, vibration patterns often change with little deterioration in the machine condition, which gives much advance notice for maintenance planning [1]. In contrast, temperature and oil analysis techniques might not give much information regarding the fault type or severity [9].

Another distinguishing feature between condition monitoring systems relates to the level of automation in the data analysis. It seems that many condition monitoring systems rely on both manual and automated data analysis. For example, some condition monitoring systems can automatically notify the operators if some measured value exceeds a predetermined threshold level. Such detected anomalies can then lead to further condition analysis by experts. The experts can often diagnose the condition accurately using numerous well-established signal processing tools. For example, Fourier transform (FT), short-time Fourier transform (STFT) and Wavelet transform (WT) can reveal frequent patterns in the raw vibration data [1, 5, 11]. However, manually inspecting every anomaly can require much expensive labour, which hinders the frequent analysis of large fleets. Therefore, increasing the level of automation in condition monitoring systems is desirable.
Vibration-based condition monitoring seems very promising for further automation of the condition monitoring of rotating machines. Rotating machine parts typically vibrate in motion [13–15]. Even in perfect condition, a rotor in motion always has some vibration energy. The vibration can be in the lateral, torsional and axial directions [15]. Lateral vibration is vibration that is perpendicularly directed toward the axis of rotation. Torsional vibration is the form of vibration that twists the rotating component around the axis of rotation. Axial vibrations displace the rotor along the axis. Faults often excite the rotating components, and induce a characteristic signature in the machine vibration [1]. These vibration signatures can be observed with sensors while monitoring the condition of a machine. Probably the most frequently used sensors for lateral vibration data acquisition are variants of accelerometers, since they can be easily mounted on many machines [16]. Torsional vibration data can be acquired with, for example, rotary encoders and torque transducers [1]. Rotary encoders can measure the angular displacement of the rotating component to which the encoder is attached. Torque transducers are often coupled between two rotating shafts, and they sample the torque transmitted through the transducer.

During recent decades, researchers have intensively developed artificial intelligence (AI) for automatic condition monitoring [4, 7, 8, 17]. AI and especially deep learning are attractive techniques for condition monitoring, since the modern data acquisition systems and computational resources allow the optimisation of models for numerous condition monitoring tasks. Moreover, recent work has shown that the powerful pattern recognition abilities of deep learning models are extremely promising for vibration data analysis [18, 19]. That is, deep learning models can recognise the distinct vibration signatures of various faults related to rotating components [20–22]. With the current data acquisition systems and the intensive research towards applied deep learning, it seems that fully automated and intelligent condition monitoring systems could be deployed in industry in the near future. However, few studies have yet demonstrated convincing evidence of the reliable performance of deep learning-based condition monitoring algorithms. Much work is still required to find the best deep learning models and best practices for data acquisition and preprocessing.

1.1 Research problem

The high reliability of the data acquisition system is critical in many applications. All sensors and related instruments should produce accurate signals for the designed period of time without failures. Designing a reliable data acquisition system can be difficult if many sensors are required or if the sensors are mounted in a hazardous environment. The likelihood of
Introduction

Sensor malfunction increases with the number of sensors in the system. In addition, mere vibration of the rotating machine can become a hazardous factor for the data acquisition system. Another challenge regarding the data acquisition system relates to the mounting of the sensors. Sometimes the hazardous environment or geometric constraints may completely prevent or at least hinder the mounting and replacing of sensors.

Numerous deep learning models for vibration-based condition monitoring are relatively agnostic towards the type of acquired vibration data. That is, the same model architectures can be optimised to recognise and distinguish faults from different vibration signals, such as lateral acceleration, angular acceleration or torque. Many of the latest deep learning-related fault diagnosis studies have employed lateral vibration signals acquired with accelerometers [17, 19, 23–31]. This paradigm understandably stems from the state of publicly available datasets, but fails to consider all the circumstances related to vibration and fault diagnosis. For a number of reasons, lateral vibration may not always be the optimal source for monitoring the rotating system condition. For example, an accelerometer attached to a component housing is likely to acquire data containing vibration patterns originating from excitations transmitted through the support structure of that housing. Such vibration patterns may confuse the condition monitoring algorithm. For example, deep learning-based algorithms may learn confusing relationships between redundant vibration patterns and faults. Furthermore, lateral vibration data can be quite sensitive to damping, since the vibration patterns excited by the rotating components are transmitted through the non-rotating parts to the sensor. Damping of the non-rotating parts may reduce the vibration amplitudes, and therefore, decrease the performance of the condition monitoring algorithm. Thus, the performance of deep learning-based condition monitoring applications could be improved by finding more reliable transmission paths for the vibration signatures.

Fault diagnosis has received much attention in the research space consisting of deep learning-based condition monitoring [17]. Many promising fault diagnosis techniques for time-series data employ 1-dimensional convolutional neural networks (1D CNN). Numerous studies have shown that 1D CNNs can learn and distinguish features from time-series data sampled with sensors, such as accelerometers [17, 21, 32]. Furthermore, the current 1D CNNs seem promising for the most recent transfer learning and meta-learning-related fault diagnosis research [33, 34]. However, many of these studies include batch normalisation (BN) as the undisputed and undiscussed technique for feature normalisation. Feature normalisation is essential for the gradient descent-based optimisation of deep neural networks. BN is a computationally effective feature normalisation technique, which often results in significantly higher test performances than corresponding models without BN [35, 36]. However, BN can be considered
as an incomplete normalisation method. BN was not designed to decorrelate the features between deep learning model layers, as discussed in the original publication [35]. Thus, the currently effective fault diagnosis models could still improve with more sophisticated feature normalisation techniques.

1.2 Objectives of the research

This research aims to improve and develop the condition monitoring systems with three different objectives related to deep learning. The objectives concern the data acquisition and data analysis aspects of condition monitoring. The first objective (Publication I) is to increase the reliability of data acquisition systems, and simultaneously, to provide an alternative technique for direct physical measurements. The second objective (Publication II) is to reveal the differences in 1D CNN-based fault diagnosis model performances between lateral and torsional vibration data. The third objective (Publication III) is to improve the performance of data analysis by studying the feature normalisation of the latest 1D CNN-based fault diagnosis models.

The first study in this dissertation focuses on recurrent neural networks (RNN) for virtual sensor applications. A virtual sensor, or synonymously soft sensor, is a computational model that can estimate a measurable quantity from indirect measurements. Such virtual sensors could be employed either as a reserve for faulty sensors or as an alternative for sensors that are expensive or difficult to mount. RNNs are advantageous techniques for virtual sensor applications, since they require little mathematical modelling of the underlying rotating system mechanics. With sufficient data from sensors, these deep learning models can learn to estimate the measurable quantity from secondary measurements. However, the numerous parameters of the RNNs are typically optimised using stochastic gradient descent (SGD). SGD may result in local optima that does not generalise over all the desired operating dimensions. For example, a poorly optimised RNN may not estimate some quantity of a rotor system correctly at some typical rotating speeds. Therefore, this study aims to develop an optimisation technique that results in satisfactory optima enabling generalisation over a large span of operating conditions.

The second study in this dissertation investigates the effectiveness of torsional vibration data for deep learning-based gear fault diagnosis. Torsional vibration patterns can be measured with sensors attached directly to the rotating components. Such sensors are, for example, rotary encoders and torque transducers. Torsional vibration may be transmitted more effectively through the rotating machine components than lateral vibration, since torsional vibration is not significantly affected by the damping of
bearings, bearing housings or gear boxes. Therefore, torsional vibration data may include patterns that are easier to diagnose. The aim of this study is to test this hypothesis by employing 1D CNN-based models in diagnosing faults from torsional and lateral vibration acquired from various locations along a drivetrain. 1D CNNs are adequate for validating pattern visibility, since they learn features autonomously, and thus the diagnosis accuracies are not biased by features designed by experts.

The third study in this dissertation concerns 1D CNN-based fault diagnosis models. 1D CNN-based models are promising computational models for automating the diagnosis functions of condition monitoring systems. These models can learn to compute effective features for fault recognition from relatively unprocessed sensor data. Consequently, numerous studies have proposed different architectures and training algorithms for 1D CNN-based fault diagnosis. A major share of these proposed models use batch normalisation. This study seeks to improve the feature normalisation technique of all 1D CNN-based fault diagnosis models. The improvement is targeted towards the normalisation of features.

1.3 Limitations of the research

Numerous previous studies have proposed effective deep learning models for condition monitoring. Therefore, this research concerns the investigation of general problems related to the application of deep learning models for condition monitoring. In Publication I, the employed model is a variant of Long Short-Term Memory (LSTM) [37]. In Publications II and III, the employed 1D CNNs were also implemented by closely following the originally proposed architectures [21, 38, 39].

The studies in this dissertation mostly concern laboratory measurements. Publication I exploits a dataset acquired from a test rig of a full-scale paper machine roll. Publication II examines data acquired from a down-scaled azimuth thruster with artificially produced gear faults. Publication III employs a published dataset acquired from another test rig with artificially induced bearing faults. In addition, Publication III uses an azimuth thruster dataset, which included healthy and faulty data acquired during real operating conditions.

This research focuses on studying problems related to the application of deep learning models for condition monitoring. To this end, all the publications employ supervised learning. That is, the model in Publication I is optimised by the mean squared error (MSE) between the measured sensor values and the estimated sensor values at given trajectories. Furthermore, the models in Publication II and III are optimised with cross-entropy loss for a pre-determined label space. Supervised learning may not be the optimal choice for fault diagnosis applications in the real world. Acquiring
sufficiently richly labelled fault datasets for supervised learning may require exhaustive effort and infeasible costs. However, supervised learning is a suitable paradigm for testing the hypotheses in these fault diagnosis publications.

1.4 Research methods

The research methodology in this work consists of

- Literature review on the state-of-the-art techniques related to machine learning-based condition monitoring
- Experimental validation and testing of implemented algorithms
- Statistical analysis of the experimental results

The literature review conducted for this compilation targets condition monitoring assisted with artificial intelligence, in general. The literature review associated with Publication I consists of studies reporting either rotor dynamics or virtual sensors developed with state-space approaches or machine learning (ML). The literature review in Publication II focuses on deep learning-based fault diagnosis techniques with an emphasis on gear faults and torsional vibration. Publication III reviews literature regarding deep learning-based fault diagnosis methods and the normalisation techniques for neural network features.

Each publication in this research is constructed around the experimental validation of deep learning models for the acquired datasets. Publication I evaluates a LSTM-based virtual sensor and the developed training technique for a paper machine roll in motion. Publication II compares 1D CNN-based fault diagnosis models on data acquired with a laboratory test rig. Publication III evaluates the employed feature normalisation technique with three different 1D CNN models on an open source laboratory dataset and on a real-world dataset acquired in real operating conditions.

A crucial cornerstone in each publication is statistical analysis of the results. Publication I evaluates the performance of the virtual sensor on the large test dataset with mean absolute error (MAE) in the time domain and frequency domain. Publication II examines the model performances by the means and standard deviations of accuracies in experiments that were repeated multiple times. Similarly, the fault diagnosis experiments were repeated in Publication III and the corresponding accuracies were evaluated by their means and standard deviations.
1.5 Scientific contribution

Publication I builds on three scientific contributions. First, the publication demonstrates a novel application for RNNs. A LSTM-based model estimates rotor displacement trajectories from bearing housing vibration. Second, the study demonstrates a training procedure that generalises the model in many operating conditions. Third, the work indicates that LSTM is superior to other previously proposed ML models for virtual sensor applications.

Publication II experimentally demonstrates that torsional vibration can be a more effective source for gear condition monitoring data than lateral vibration. A large majority of deep learning-based condition monitoring studies employ lateral vibration data, and neglect torsional vibration. Furthermore, most related studies experiment with data acquired with a few sensors from a simple test rig. Publication II employs a laboratory drive train designed to share the natural frequencies with a real maritime thruster. Moreover, the data is acquired with numerous torque transducers, rotary encoders and accelerometers. The results show that models can recognise gear faults more accurately from torsional vibration than from lateral vibration. The experiments consider the vibration phenomena in a much broader view than related previous research.

Publication III demonstrates a possible improvement for all CNN-based fault diagnosis models. The improvement relates to feature normalisation, which is crucial for deep learning models. This study investigates a feature normalisation technique that theoretically includes better normalisation functionalities than the commonly employed BN. This normalisation technique is designed to decorrelate the features in addition to the scaling and shifting which BN performs. The technique known as network deconvolution (ND) was adapted from a previous study and modified for 1D CNN layers. This technique has previously been examined with 2D CNNs and image data. Publication III extends the scope of the available scientific knowledge regarding the application of ND to 1D CNNs and time-series data. Furthermore, the performance of the technique is validated with ablation experiments on relevant fault diagnosis datasets.
This literature review consists of three sections related to the publications. Section 2.1 covers rotor dynamics, typical failure modes of rotor systems and common components of condition monitoring systems. Section 2.2 visits published literature concerning virtual sensors or similar applications with an emphasis on ML-based methods. Section 2.3 reviews the latest literature related to ML techniques for machine fault diagnosis applications.

2.1 Rotor system condition monitoring

Three common maintenance strategies are run-to-break, time-based preventive maintenance and condition-based maintenance (CBM) [1]. The run-to-break strategy is the most suitable for machines that do not result in significant economic losses due to failures. Typically such a maintenance strategy relies on cheap machines that can be quickly replaced by a contingency. However, this strategy is not optimal for circumstances where failures can result in catastrophic outcomes, or where the cost of lost production is higher than the cost of replacing or repairing the failed machine. For such circumstances, a time-based maintenance strategy has often been employed. This preventive maintenance strategy relies on pre-scheduled maintenance plans, which include frequent maintenance breaks for machine health inspections and component replacements. Unfortunately, frequent maintenance breaks can result in downtime and replacements of components with long remaining useful lifetimes. Fortunately, CBM can reduce wasted maintenance time and the number of discarded healthy components. The CBM strategy plans maintenance breaks based on recent assessments of the machine condition. Only components with little remaining useful lifetime are replaced according to this maintenance strategy.

CBM strategies for rotating machinery often employ vibration-based condition monitoring systems [1]. Vibration can be an effective source for
State-of-the-art condition monitoring data, since it often includes characteristic signatures related to the normal or faulty motion of the machine components. Furthermore, numerous powerful signal processing techniques can be deployed for vibration data analysis.

This section presents rotor vibration and typical means for vibration data acquisition in three sections. Section 2.1.1 describes the dynamic behaviour of rotating machines. Section 2.1.2 reviews some of the common techniques for vibration data acquisition. Section 2.1.3 discusses some of the common failure modes of rotating machines and the corresponding effects on the machine vibration.

2.1.1 Vibration of rotating systems

Vibration emerges in every rotating system in motion despite the health state of the system. The system components dictate the vibration signatures. That is, every component in the machine contributes to the machine vibration with characteristic frequencies. For example, deviations from the ideal roundness profile of the shaft or from the ideal material distribution of the shaft likely excite the system at frequencies proportional to the rotating speed [15]. Furthermore, bearings and gears affect the rotor system vibration at characteristic frequencies [1, 40–42]. In addition, the foundation can have a significant effect on the rotor vibration [43]. This section covers such vibration phenomena that can be considered as dynamics of a healthy system with mere manufacturing and assembly inaccuracies.

Rotor vibration

Every rotor has many natural frequencies depending on their properties, such as stiffness, mass and dimensions. Each natural frequency corresponds to a particular eigenmode of the rotor. Figure 2.1 shows some of the eigenmodes of rigid and flexible shafts. The top three on the left correspond to the eigenmodes of a rigid shaft. The top three on the right correspond to the eigenmodes of a flexible shaft. The lowest shaft twists around the axis of rotation, and corresponds to a torsional eigenmode.

Most often, the rotating system operating speeds should avoid the natural frequencies of the rotors. The rotating speeds that correspond to the rotor natural frequencies are the critical speeds $f_c$ of the rotor [44]. These critical speeds can magnify the rotor vibration excessively due to harmonic resonance. A speed range lower than the critical speed is often referred to as the subcritical speed range. Depending on the vibration sources, resonance can also occur at subcritical speeds.

Common vibration sources caused by inaccuracies in manufacturing and assembly include, for example, unbalance, misalignment and bending stiffness variation. These properties often increase the vibration amplitudes
at the lowest integer multiples of the rotating speed, i.e. the harmonic frequencies. Unbalance is the condition where the centre of mass of the rotor does not align with the axis of rotation [1, 15]. Unbalance typically excites the system at the rotating speed (1X), and therefore unbalance-related vibration can be observed at the harmonic frequencies of the rotor [1]. Misalignment is the condition where two rotors are coupled so that their axes of rotation do not align. Misalignment often increases vibration at the lowest even harmonic frequencies (2X) [1]. Bending stiffness variation changes the rotor deflection under a load, such as the gravity, depending on the rotor angle. Bending stiffness variation can excite the rotor twice per revolution (2X) [45]. Overall, rotating systems include many vibration sources that can excite the system at the harmonic frequencies (1X, 2X, 3X ...) of the rotating speed. Therefore, the subcritical rotating speeds of $1/2f_c$, $1/3f_c$, and so on should also be avoided.

**Bearing vibration**

Bearings are crucial for all rotating machines, since they couple rotors to the supporting foundation. Many different bearing types, such as oil film bearings [13], aerostatic bearings [46] and rolling element bearings [45], exist for different applications. The oil film bearing and the aerostatic bearing allow the rotor to slide freely in a bearing sleeve with some lubrication between the rotor and the bearing surfaces. The most common bearing is the rolling element bearing, which also allows relatively high rotating
Since rolling element bearings are common in the industry, and the majority of the experiments in this dissertation concern rolling element bearings, this section focuses on this type of bearing. Figure 2.2 shows a cross-section of a typical rolling element bearing. The rolling elements in this bearing are balls. Rolling element bearings consist of an inner race, an outer race and the rolling elements. The outer race is typically attached to the stationary foundation, and therefore does not rotate. The inner race is in turn fixed to the rotating shaft, and therefore it rotates with the same frequency. The balls can move between the inner and outer races. Furthermore, the balls are often placed in a cage that locks the relative distance between the elements.

The vibration of rolling element bearings has been thoroughly studied. Therefore, the vibration patterns of a healthy bearing with mere manufacturing or assembly inaccuracies are well-known. These inaccuracies tend to result to geometric deviations to the bearing. The geometric deviations, such as the oval shape of the inner ring, can excite the rotating system [45]. Equation 2.1 shows the frequency \( f_i \) corresponding to the geometric deviations of the inner ring \( i \). The number of undulations in the inner ring is \( z_i \), and \( \omega_i \) is the rotating speed of the rotor [14].

\[
f_i = z_i \cdot \omega_i \tag{2.1}
\]

Another specific vibration frequency related to bearing motion is the cage frequency \( f_{ca} \). The cage frequency corresponds to the speed the balls rotate around the axis of rotation of the bearing. Equation 2.2 shows the relation between the cage frequency \( f_{ca} \) and the bearing assembly. \( D_{ball} \) is the ball diameter, \( D_{pitch} \) is the average diameter of the inner and outer races and \( \beta \) is the angle between the ball and the race groove [14].

\[
f_{ca} = \frac{\omega_i}{2} \left( 1 - \frac{D_{ball} \cdot \cos \beta}{D_{pitch}} \right) \tag{2.2}
\]
The ball frequency related to the rotation of the balls follows Equation 2.3 [14].

\[ f_b = \frac{\omega_i}{2} \left( \frac{D_{\text{pitch}}}{D_{\text{ball}} \cdot \cos \beta} - 1 \right) \]  

(2.3)

Equation 2.4 shows the relationship between the inner ring base frequency \( f_{ir} \), the inner ring frequency \( f_i \) and the cage frequency \( f_{ca} \). Equation 2.5 shows the outer race base frequency \( f_{or} \), where \( z_o \) is the number of undulations on the outer race. The outer race is assumed to be fixed.

\[ f_{ir} = |f_i - f_{ca}| \]  

(2.4)

\[ f_{or} = f_{ca} \cdot z_o \]  

(2.5)

**Gear vibration**

Many gear types have been developed to transmit power between rotors in a drivetrain. Types of gear include, for example, the bevel gear, the worm gear and the planet gear [47]. Typically, gears can change the torque, the rotating speed and the direction of rotation between the rotors. A gear can consist of two or more wheels with teeth meshing together. In a gear consisting of two wheels, the smaller gear with fewer teeth is typically called the gear pinion and the larger gear with more teeth is often called the gear wheel. The gear pinion rotates faster than the gear wheel. Simultaneously, the smaller torque from the gear pinion is transmitted to a larger torque to the gear wheel. During torque transmission, the gears typically vibrate at the meshing frequency. Equation 2.6 shows the gear meshing frequency (GMF), which is proportional to the gear rotating speed \( f_g \) multiplied by the number of teeth in the gear \( N_t \) [48].

\[ GMF = f_g \cdot N_t \]  

(2.6)

Gear vibration is greatly affected by the transmission error (TE). TE can modulate the amplitude and frequency of the meshing vibration. Most common reasons for TE under constant load are related to the non-ideal geometry of the gear teeth. These deviations in the ideal teeth geometry can be manufactured intentionally. For example, some material is often removed from the tip of the teeth to reduce the impact between contacting teeth. Unintentionally produced geometry errors exist too. For example, imprecision in the gear-cutting process can result in systematic deviations in the teeth geometry, which can cause vibration at the harmonics of the gear rotating speed. The dynamic deviations to the teeth geometry can be caused by the deformation of the meshing teeth under load. Overall, transmission error is present in all healthy gears, and can increase the vibration amplitudes at the meshing frequency, and its harmonics [1].
Furthermore, transmission error can give rise to amplitudes at some frequencies near the meshing frequency. These frequencies are commonly known as sideband frequencies.

### 2.1.2 Vibration data acquisition

Vibration data can be acquired during condition monitoring. Vibration-based condition monitoring, however, can be practised intermittently or continuously. That is, the condition monitoring equipment can be permanently mounted on the machine or mounted for certain intervals [1]. Permanent condition monitoring systems are often mounted on the critical equipment that should be shut down immediately after a fault is detected. Intermittent condition monitoring systems are more suitable for machines that either do not often break unexpectedly or do not cause severe consequences if they malfunction.

Vibration data should be acquired from close proximity to the monitored component. A failure in a component often changes the vibration signature observed with vibration sensors. However, the vibration signal originating from the fault is often transmitted through other components, which likely affect the acquired signal [1]. A long transmission path can dissolve the fault signal to the background noise and other signals.

There are numerous different vibration sensors used for condition monitoring purposes. Most often, condition monitoring systems for rotating machines include sensors for lateral and torsional vibration. Lateral vibration sensors can acquire data regarding the displacement, velocity or acceleration of the machine. Typically, condition monitoring systems rely on accelerometers. Accelerometers can be easily mounted on the machines and can acquire vibration signals from a broad frequency range. Torsional vibration sensors can measure angle, angular velocity, angular acceleration and torque, for example. Common torsional vibration sensors for condition monitoring include rotary encoders and torque transducers.

Condition monitoring systems can acquire data with different sampling rates. Typically, the sampling events are triggered by an encoder pulse or a sampling clock. Encoder-triggered sampling enables data acquisition at corresponding angles. Such data can be useful for synchronous vibration analysis, such as roundness error analysis [49] and bearing stiffness variation analysis [43]. The sampling rate varies between vibration samples acquired at varied rotating speeds if encoder-triggered sampling is applied. Vibration data with a constant sampling rate can be acquired with a sample clock. Despite the sampling trigger, condition monitoring systems often require high sampling rates, such as thousands of hertz, since many vibration signals and their harmonics occur at high frequencies [1].
2.1.3 Rotating system failures

The components of rotating machines can fail for many reasons, such as inaccuracies in manufacturing or assembly, high loads and improper maintenance. Manufacturing inaccuracy can relate to unbalanced shafts that excite other parts of the system and cause premature failures. Similarly, misaligned gears and shafts due to inaccuracies in assembly can increase vibration in the system. Excessive loading of the rotating system may eventually cause fatigue problems. Failure of lubrication due to improper maintenance can increase friction between the moving components and cause wear on the surfaces of the components, for example. The severity of the faults can vary significantly. For example, small defects may not even be noticed due to their relatively small effect on the machine vibration. In contrast, some faults may develop significantly, and therefore affect the operation of the machine.

International Standard ISO 15243 describe rolling element bearing failure modes and their characteristics [50]. These modes include fatigue, wear, and fracture and cracking, for example. Fatigue can be initiated on the surfaces and under the surfaces of the rolling elements due to the loads between the rolling contacts. Figure 2.3 shows an example of a subsurface fatigue failure progressing. Wear results in the removal of material from the contact surfaces due to inadequate lubrication or friction, for example [50]. Cracks in the bearings can start to develop if the stresses in the components exceed their tensile strength [50]. Fractures are defined in this international standard as the "complete separation of a part of the component," and they can be caused by excessive crack propagation or impacts.

Typically, all these bearing failure modes affect the vibration patterns of the rotating machines. Equations 2.7-2.10 show some of the frequencies that are typically considered to be affected by bearing failures [1, 51]. These frequencies are the ball pass frequency of the outer race (BPFO), the ball pass frequency of the inner race (BPFI), the fundamental train frequency (FTF), i.e. the ball cage speed, and the ball spin frequency (BSF). In these equations, $n$ is the number of rolling elements in the bearing, $f_r$ is the shaft rotating frequency, $d$ is the rolling element diameter, $D$ is the bearing pitch diameter and $\phi$ is the load angle from the radial plane.

\[
BPFO = \frac{n f_r}{2} \left(1 - \frac{d}{D} \cos \phi\right) \quad (2.7)
\]
\[
BPFI = \frac{n f_r}{2} \left(1 + \frac{d}{D} \cos \phi\right) \quad (2.8)
\]
\[
FTF = \frac{f_r}{2} \left(1 - \frac{d}{D} \cos \phi\right) \quad (2.9)
\]
\[
BSF = \frac{f_r D}{2} \left(1 - \frac{d}{D} \cos \phi\right) \quad (2.10)
\]
Figure 2.3. Subsurface fatigue failure progressing on a bearing [50].

International Standard ISO 10825 defines the terminology related to failure modes of gears. These failure modes include, for example, wear, pitting and scuffing [52]. Wear in gears occurs similarly as in bearings, as the two contacting surfaces slide and remove material from each other. Pitting relates to the surface fatigue on the active tooth flanks. Pitting is characterised by small and shallow holes which can eventually progress to larger and deeper holes that deform the active tooth flank surfaces [52]. Scuffing can be caused by the failure of the lubricating film between the contacting teeth and result in visible bands in the teeth sliding direction [52]. All these gear failures can affect the vibration signatures of the rotating machine. Some of the frequencies that can be observed to detect gear faults are the meshing frequency and sidebands, and their harmonics [48].

2.2 Data-driven virtual sensors

Virtual sensors, synonymous to soft sensors, are computational models that can estimate a measurable quantity from other related information, such as secondary measurements. Virtual sensors can be divided into two branches: model-based and data-driven virtual sensors [53]. Model-based virtual sensors often rely on prior knowledge of the system dynamics. Well-known model-based virtual sensors include observers, variants of the Luenberger observer and variants of the Kalman filter [54]. Previously,
these techniques have been used, for example, to estimate the torsional vibration and rotating speed of a thruster [55], state of charge of batteries [56] and side-slip angle and transversal forces of a wheel [54].

Although these model-based virtual sensors can estimate the internal states of many systems, they can suffer from several disadvantages. Often, the non-linear interactions between the components of the system are modelled with simplified assumptions, such as linearisation. For example, a Kalman filter estimated the lateral displacement of a drill collar. The model was based on lumped masses connected by linear torsional springs [57]. Such idealisations of the system dynamics can result in a gap between the modelled dynamics and the real dynamics. Modelling complex and nonlinear system dynamics can be difficult, if high accuracy is desired. For example, modelling torque losses in gears accurately can require detailed information related to the friction between meshing teeth. Moreover, the accuracy of these models can be limited to a particular range of operating conditions.

Data-driven virtual sensors seem to offer promising alternatives for these model-based sensors. Data-driven virtual sensors require very little prior knowledge and no simplifying assumptions of the system dynamics. These algorithms are optimised with historical data acquired from the system [53].

Data-driven virtual sensors can be divided into traditional machine learning-based and deep learning-based virtual sensors. Previously investigated traditional machine learning-based virtual sensors have relied on algorithms such as multilayer perceptrons (MLP) [58], partial least squares [59] and support vector machines (SVM) [60]. To some extent, these models have also been experimented on with rotating systems. For example, SVMs were used to estimate the displacement of a rotor suspended by a hybrid magnetic bearing from the control coil currents [61, 62]. Unfortunately, these traditional machine learning-based virtual sensors suffer also from disadvantages. Traditional machine learning-based models can be limited in their capacity to learn complex functions from large data. Likely, these data-driven models are not very accurate over large ranges of operating conditions. For example, changes in the rotating speeds or other operating conditions can change the vibration levels significantly, and require more complex virtual sensor functions. Studies demonstrating the performance of these traditional data-driven virtual sensors over many operating conditions are very scarce.

Deep learning-based virtual sensors can provide a significant advantage over traditional machine learning-based virtual sensors because of their capability to learn very complex functions from large data. This advantage has been demonstrated by many studies in other application fields of deep learning. For example, deep learning models outperformed traditional machine learning in many benchmark tasks, such as natural language
processing [63], image recognition [64] and reinforcement learning [65]. Thus, deep learning models seem promising for virtual sensor applications that require accuracy over many operating conditions.

Deep learning-based virtual sensors have been developed for numerous applications in many industries [66]. The majority of the previous studies seem to concern the chemical industry. For example, deep learning models have been applied for monitoring of penicillin fermentation [67], hydrocracking [68] and polyethylene production [69]. However, only a minority of previous works seem to have considered rotating machines [66]. Some studies have proposed stacked autoencoders and deep belief networks to estimate the deformation of the air pre-heater of a power station boiler [70–72]. Nevertheless, very few virtual sensor-related works have considered the monitoring of rotor vibration. Moreover, many of the previous works concern data acquired with relatively low sampling rates, such as one sample per hour or per day. Virtual sensors for rotor vibration monitoring are required to compute the vibration signals acquired at high frequency, such as thousands of times per second.

2.3 Intelligent fault diagnosis models

Effective condition monitoring systems rely on timely and accurate fault diagnosis. Such requirements for fault diagnosis can be fulfilled with anomaly detection software and expert analysis. Unfortunately, condition monitoring experts can be expensive, and sometimes there can be an unfeasible lag between anomaly detection and fault diagnosis. Therefore, the automation of fault diagnosis is desired in many condition monitoring applications. Many recently developed automatic fault diagnosis systems employ AI, and hence, are often called intelligent fault diagnosis (IFD) models.

IFD models can be described in two steps. First, the models extract features from data, and then the models recognise the fault from the extracted features [18]. These fault diagnosis algorithms can be further divided into traditional ML-based and DL-based techniques [17]. These two categories are different in the way the features are extracted. ML-based fault diagnosis systems rely on manually designed features, such as peak value, mean frequency or energy entropy [17]. However, these manually tailored features can be very sensitive to variations in the vibration signatures caused by changes in the operating conditions or by the inaccuracies in manufacturing and assembly. DL-based fault diagnosis models may have an advantage over traditional ML, since they require no manual feature design. DL models can be optimised to extract features autonomously and recognise the faults from the features.

The remainder of this section reviews the published literature regard-
ing these data-driven models and the resources and techniques for their optimisation. The review consists of four sections. Section 2.3.1 presents some open source benchmark datasets often used in IFD model studies. Section 2.3.2 views some of the most significant publications related to traditional ML-based fault diagnosis models. Section 2.3.3 covers some of the best-performing DL-based fault diagnosis models. Finally, Section 2.3.4 reviews the latest techniques related to different training algorithms for DL-based condition monitoring models.

### 2.3.1 Open source datasets

ML and DL-based fault diagnosis models optimised with the supervised learning paradigm require extensive training datasets. For some condition monitoring applications, the acquisition of extensive datasets can be difficult. For example, algorithms for fault diagnosis often require datasets containing representative distributions of healthy and faulty samples under relevant operating conditions. Faulty data is often scarcely available, since the machines are often repaired after the fault has been noticed. Moreover, numerous failure modes exist, and the failures may require much time to develop.

A number of ways for acquiring representative training datasets exist. Likely, the most reliable datasets for fault diagnosis purposes can be gathered by continuous or intermittent monitoring of a fleet of machines over a long period of time. By collecting different faults occurring over time from machines with similar configurations, the fault diagnosis models can eventually become accurate [17]. Another promising technique for data acquisition includes simulations [1, 73, 74]. Simulations can generate large amounts of data from rich conditions in relatively short time. However, the difference between data generated with simulations and data acquired from real machines can hinder the deployment of models trained with simulated data. A very common approach in ML-based condition monitoring studies is to evaluate the models with a dataset acquired in a laboratory [20]. Although laboratory data may suffer from similar differences between data distributions as simulated data, it still can provide better grounds for research since the data includes real measurements.

Currently, a number of published fault datasets are available that are used for IFD research purposes. Table 2.1 lists some of these popular datasets. The CWRU dataset is the bearing fault dataset provided by the Case Western Reserve University [75]. This dataset includes vibration samples of one healthy and nine faulty bearings measured with two accelerometers on a test rig operated with rotating speeds between 1730 RPM and 1797 RPM. The Paderborn dataset was provided by Paderborn University [76]. This dataset includes vibration data from 32 bearings, of which 6 were healthy, 12 were artificially broken and 14 were broken.
in accelerated lifetime tests. The vibration data includes motor current and acceleration signals acquired under 900 RPM and 1500 RPM. The PRONOSTIA dataset including data from 17 bearings in run-to-failure tests was provided by the FEMTO-ST Institute [77]. The vibration samples include acceleration signals from two accelerometers. The vibration samples were acquired under 1500 RPM, 1650 RPM and 1800 RPM rotating speeds. The MFPT bearing fault dataset was provided by the Society For Machinery Failure Prevention Technology [78]. This dataset includes vibration data related to 23 different healthy and faulty bearing conditions. Three of the conditions relate to real machines, and the remaining 20 to a test rig operated at a 1500 RPM rotating speed. Finally, the PHM dataset includes vibration data of a gearbox in healthy condition and in various faulty conditions [79]. The data was measured with two accelerometers under many rotating speeds between 1800 RPM and 3000 RPM [79].

Table 2.1. Some of the commonly used fault datasets in intelligent fault diagnosis studies.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Measured quantity</th>
<th>Sensors</th>
<th>Rotating speeds (RPM)</th>
</tr>
</thead>
<tbody>
<tr>
<td>CWRU</td>
<td>Acceleration</td>
<td>2</td>
<td>1730 &amp; 1750 &amp; 1772 &amp; 1797</td>
</tr>
<tr>
<td>Paderborn</td>
<td>Motor current &amp; Acceleration</td>
<td>3</td>
<td>900 &amp; 1500</td>
</tr>
<tr>
<td>PRONOSTIA</td>
<td>Acceleration</td>
<td>2</td>
<td>1500 &amp; 1650 &amp; 1800</td>
</tr>
<tr>
<td>MFPT</td>
<td>Acceleration</td>
<td>1</td>
<td>1500</td>
</tr>
<tr>
<td>PHM</td>
<td>Acceleration</td>
<td>2</td>
<td>1800 &amp; 2100 &amp; 2400 &amp; 2700 &amp; 3000</td>
</tr>
</tbody>
</table>

2.3.2 Machine learning-based fault diagnosis models

ML-based models for fault diagnosis of rotating machines have been developed for decades [17]. In the literature, these ML-based methods are commonly described as consisting of a separate feature extraction technique and a traditional machine learning algorithm used for fault recognition [18]. The feature extraction techniques typically compute values from the vibration data in the time domain, the frequency domain or the time-frequency domain [18]. The time domain features can be, for example, root mean square (RMS), standard deviation and kurtosis [80]. The frequency domain features often rely on the Fourier transform. The frequency domain features can relate to the amplitudes at particular known fault frequencies, such as the BPFI, BPFO and GMF [81]. The time-frequency domain features typically rely on signal transforms such as the short-time Fourier transform, Hilbert-Huang transform [82] and wavelet transform [83]. A number of traditional machine learning models have been proposed for recognising the extracted features, such as support vector machines,
k-Nearest Neighbours (k-NN) and shallow neural networks [17, 18].

Numerous works have demonstrated the application of SVMs for the fault diagnosis of rotating machines [8]. For example, low speed bearing faults were diagnosed with SVM based on the statistical features of vibration data and acoustic emissions [84]. In addition, the performance of SVM for bearing fault diagnosis on the time-frequency domain features extracted with wavelet-based functions has been evaluated [85]. Bearing faults of an induction motor were also classified with SVM from the RMS, crest and kurtosis values extracted from the continuous wavelet transform [86]. Moreover, time domain features and fractal dimensions of the vibration signals in the CWRU bearing fault dataset were experimented on with SVM [87].

In addition to SVM, the literature covering other ML-based fault diagnosis algorithms is extensive. For example, k-NN, LDA and AdaBoost M1 were optimised and trained to recognise azimuth thruster anomalies from time and frequency domain features extracted from accelerometer signals [81]. A shallow neural network was optimised to recognise bearing faults of an induction motor from statistical features selected from the continuous wavelet transform of a vibration signal [86]. Furthermore, Naive Bayes and k-NN were used to experiment on the performance of features derived from the statistics, frequencies and autoregressive coefficients of motor current data for induction motor fault detection and diagnosis [88].

2.3.3 Deep learning-based fault diagnosis models

Recently, many studies have demonstrated that DL models can effectively recognise faults from vibration data [7, 17, 18, 20, 89]. This substantial attention towards DL-based fault diagnosis likely follows from the advantages these models provide compared to conventional ML-based fault diagnosis models. DL models require no preprocessed features because they can recognise the machine condition from the relatively raw representations of vibration data. That is, these models can process the vibration data in the time, frequency or time-frequency domains [17, 90].

DL models learn to extract and recognise features from the vibration data autonomously under simultaneous optimisation. Learning the features autonomously can save much manual effort in feature design and feature engineering. In addition, DL models can learn complex hierarchical features which can characterise large data distributions. This attribute, inherent to DL models, enables the optimisation of one DL-based model for many fault diagnosis tasks such as gear and bearing condition monitoring tasks. Furthermore, the powerful hierarchical features can reduce the sensitivity of the diagnosis algorithm to the variations in the operating conditions. However, in order to learn highly accurate functions over many operating conditions, DL models require training data from these operating
conditions.
Currently, many researchers are trying to circumvent the problems related to lack of fault data with techniques such as transfer learning and few-shot learning. Transfer learning and few-shot learning are covered briefly in Section 2.3.4. Much of the remaining basic theory is explained well in many reviews and books [7, 17, 91]. Therefore, the remaining section focuses on the most important DL models for fault diagnosis.

Based on the previous DL-based fault diagnosis studies, it seems that some DL models are more popular [17, 18]. These popular models include CNNs, autoencoders (AE), deep belief networks (DBN) and recurrent neural networks [7, 17, 20]. All of these model types enable the learning of features from the vibration data. However, the features learned by different model types are not similar in structure. The feature structure may bring some constraints to the application of the model. For example, CNNs can learn features from vibration data in the time domain while optimising AEs and DBNs for similar tasks can be difficult [90, 92].

CNNs can learn to extract and recognise features from vibration data in the time, frequency and time-frequency domains [90]. These models require very little data preprocessing. The most common CNN models for fault diagnosis are 1D CNNs and 2D CNNs. 1D CNNs typically process vibration data in the time or frequency domain. 2D CNNs often compute vibration data in the time-frequency domain, since the model design requires 2D representations of the data.

One of the publications proposing a 1D CNN-based fault diagnosis model with most impact introduced a model known as WDCNN. WDCNN includes wide first-layer kernels, which can extract low level features from the time or frequency domain representation of the vibration signal [21, 90]. Other models based on WDCNN have also been proposed, such as RNN-WDCNN and Convolution Neural Networks with Training Interference (TICNN). RNN-WDCNN consists of WDCNN and an adjacent LSTM path [93]. TICNN is an ensemble of WDCNNs trained with dropout [94]. Plenty of other 1D CNN-based models have also been proposed. For example, stacked residual dilated convolutional neural network (SRDCNN) was presented, and the model performance was evaluated on the CWRU dataset [32].

Currently, 2D CNNs seem to be very promising for vibration-based fault diagnosis. For example, pretrained AlexNet and ResNet were shown to be effective on the CWRU dataset when the vibration data was formulated to the time-frequency domain with a short-time Fourier transform [90]. In addition, AlexNet and ResNet were employed as a baseline for another model using 2D CNNs merged with a capsule network structure and an inception block [95]. This model and the baseline models also learned features from the STFT spectra of the vibration data from the CWRU and Paderborn datasets.
In addition to CNNs, autoencoder based models have been thoroughly studied. AEs are typically optimised layer-by-layer with the unsupervised learning paradigm. That is, the models learn to encode the data to lower dimensions and then to reconstruct the down-sized data to the original dimension. More encoding and reconstructing layers can be trained between the previous ones after the reconstruction error has decreased to a satisfactory level. One of the most popular related studies examined a stacked AE based model with bearing and gear fault data [19]. Moreover, other similar approaches include stacked denoising autoencoder (SDA) experiments for bearing diagnosis [96] and a stacked sparse autoencoder model (SAE) for induction motor diagnosis [97].

Deep belief networks have been employed in many related works. In these works, DBNs often process manually extracted features from vibration data [7]. One study even showed that the accuracy of DBN can be reduced significantly if raw vibration data is diagnosed [92]. Therefore, DBNs can be considered to be closely related to the conventional ML-based fault diagnosis. However, some proposed models have architectures which include DBN and other model structures. For example, a DBN with convolutional functions was proposed for bearing fault diagnosis [98]. Furthermore, a similar model was merged together with an autoencoder for experiments on locomotive bearing fault diagnosis [99].

Recurrent neural networks are DL models specifically designed for sequential computation. RNNs can process the vibration data in many different representations. Most importantly, RNNs can learn temporal features, such as vibration patterns, from the time domain data or the changes in some preprocessed features over time. Thus, there are many possible approaches for RNN-based fault diagnosis. For example, an autoregressive model with exogenous inputs (NARX) was proposed for DC motor bearing fault diagnosis [100]. The vibration and stator current data were preprocessed with discrete wavelet transform (DWT) and discriminative analysis. Another study diagnosed raw time-series data from multiple sensors on a wind turbine test rig with LSTM [101]. The final hidden state of the LSTM was passed to the classifier. Another study performed fault detection with an LSTM-based autoencoder [102]. The reconstruction error of the trained model was used to detect faults.

### 2.3.4 Training paradigms

Currently, there are a few different machine learning paradigms with different learning objectives. Although under discussion, these learning paradigms are supervised, unsupervised, reinforcement, transfer and meta learning. The paradigms are not all equally suitable for condition monitoring. The objective of reinforcement learning (RL) is to learn what to do. That is, RL is concerned with optimising interactions between
an agent and an environment [103]. Therefore, applications related to rotating system control seem to be a more meaningful domain for RL than applications related to rotating system monitoring. Thus, this section excludes RL and describes the other paradigms.

Supervised learning is one of the most-employed paradigms in the research related to condition monitoring [17]. The objective in supervised learning is to learn a function between two distributions. That is, the ML model $f_{\theta}(\cdot)$, parameterised by $\theta$ is optimised with labeled data. The dataset $\mathcal{D} = (x_1, y_1) ... (x_N, y_N)$ consists of $N$ samples of input data $x_n$ and corresponding target data $y_n$, where $n \in N$. In condition monitoring, the input samples $x_n$ can be, for example, multidimensional arrays, such as time-series data from many sensors. The target samples $y_n$ can be discrete classes for health states or some other regression target, such as time-series data from other sensors.

Equation 2.11 formulates the supervised learning objective more generally. The objective is to find parameters $\theta^*$, which minimise an error function $\mathcal{L}$ on a dataset $\mathcal{D}$. The parameters $\omega$ are the hyperparameters, which are often searched before the supervised model optimisation. These hyperparameters include, for example, the model architecture and training algorithm parameters. After successful supervised optimisation, the model $\hat{y}_n = f_{\theta}(x_n)$ should satisfy condition $y_n \approx \hat{y}_n$.

$$\theta^* = \arg\min_{\theta} \mathcal{L}(\mathcal{D}; \theta, \omega)$$ (2.11)

The supervised learning techniques employed for condition monitoring can be divided to classification and regression. A fault diagnosis model computing a probability distribution over all known health states from a vibration sample establishes a classifier for a classification task. For such classification purposes, the target function is often some variant of cross-entropy loss (CE Loss), which is a measure of the difference between the target distribution and the computed distribution. Equation 2.12 shows the formula for CE Loss, where $y_n^j \in \{0,1\}$ are the $K$ health states in the target distribution. The correct health state is encoded as 1. The estimated distribution also consists of $K$ probabilities denoted by $f_n^j \in [0,1]$. The CE Loss is often averaged over a batch of $N$ samples. Binary cross-entropy loss (BCE loss) is suitable for anomaly detection tasks where the target distribution includes only labels for healthy and not-healthy conditions. Equation 2.13 shows the BCE loss, where $y_n \in \{0,1\}$ is the target label, $f_n^\ast \in [0,1]$ is the probability of a fault, and $N$ denotes the number of samples in the batch.

$$CE \text{ Loss} = -\frac{1}{N} \sum_{n=1}^{N} \sum_{j=1}^{K} y_n^j \log(f_n^j)$$ (2.12)

36
\[ BCE\ loss = -\frac{1}{N} \sum_{n=1}^{N} (y_n \log(f^n) + (1 - y^n)(1 - \log(f^n))) \] (2.13)

Supervised learning for regression models includes different target functions. The regression models in condition monitoring applications can, for example, estimate a variable. The target function is often established to minimise the difference between the estimated value and the correct value of that variable. Suitable loss functions include, for example, mean squared error and mean absolute error. Equations 2.14 and 2.15 show the formulas for MSE and MAE, respectively. In the equations, \( y_{t,k} \) is the estimated value and \( y_{t,k}^{true} \) is the correct value in a multivariate time-series estimation task consisting of \( K \) variables and \( T \) time steps.

\[ \text{MSE} = \frac{1}{K} \frac{1}{T} \sum_{k=1}^{K} \sum_{t=1}^{T} (y_{t,k} - y_{t,k}^{true})^2 \] (2.14)

\[ \text{MAE} = \frac{1}{K} \frac{1}{T} \sum_{k=1}^{K} \sum_{t=1}^{T} |y_{t,k} - y_{t,k}^{true}| \] (2.15)

Equation 2.11 also describes the unsupervised learning objective. The difference between unsupervised and supervised learning relates to the formulation of the dataset. The dataset \( \mathcal{D} = [x_1...x_N] \) consists of unlabeled samples. The objective of unsupervised learning is to learn hidden structures in the data instead of a function that maps the input data to the correct target values. Such hidden structures can be learned, for example, with the approach used for optimising autoencoders. In this approach, the optimised function \( \hat{x}_n = f_\theta(x_n) \) satisfies the condition \( x_n \approx \hat{x}_n \).

Unsupervised learning models are especially popular in studies concerning fault diagnosis [89]. One of the common unsupervised learning model types is the autoencoder [7]. AEs require no labeled data, since they are optimised to compress the vibration data to lower dimensions, and then, to reconstruct the original signal. The target function can be MSE, for example. Other target functions for autoencoders reconstructing vibration data have also been developed, such as the maximum correntropy [104].

Interest in applying transfer learning to fault diagnosis has surged during the past years [17]. Transfer learning a promising approach, since conventional deep learning-based fault diagnosis models suffer from sparse data. That is, the deep models optimised with supervised learning require large and representative datasets, which are sometimes difficult to acquire. Most often, condition monitoring systems acquire much vibration data from healthy machine states. However, faults are often avoided and quickly repaired, and therefore, vibration data from the faulty machine states are scarcely available. Transfer learning techniques have been designed for such data sparsity problems.
Generally, the objective in transfer learning is no different from supervised or unsupervised learning. Transfer learning merely exploits learned knowledge from a previous domain to improve learning in a new domain. The improvement can be in terms of data efficiency or increased accuracy [105, 106]. Formally, a problem for transfer learning is often considered to consist of source data $D_S$, a source task $T_S$, target data $D_T$ and a target task $T_T$ [107]. The ML model $f_\theta(\cdot)$ is first trained for $T_S$ with $D_S$. The learned model is then either transferred to both, the target task $T_T$ and the target data $D_T$, or one or the other. This approach is different from unsupervised and supervised learning, since the data distribution can change between $D_S$ and $D_T$. Furthermore, tasks $T_S$ and $T_T$ can be significantly different.

There are a number of different transfer learning techniques, of which fine-tuning is sometimes considered as the most popular [106, 108]. Fine-tuning can be considered as a way to initialise the parameters $\theta$ of the model $f_\theta(\cdot)$. In supervised and unsupervised learning, the parameters are often initialised randomly and then optimised according to Equation 2.11. These optimised parameters that can extract features from the source dataset $D_S$, which are relevant for the source task $T_S$, can be useful in new problems. Typically, fine-tuning means that a model optimised with the source domain is adopted to the new domain with only some parameters randomly reinitialised. The adopted model can then be optimised again for the target task $T_T$ with target data $D_T$. Such retraining may reduce the risk of overfitting, even though the target data $D_T$ would be relatively small [106]. Therefore, transfer learning seems promising, especially for the data sparsity problems related to condition monitoring applications.

Meta-learning is another promising ML paradigm for fault diagnosis. Similarly to transfer learning, the currently developed meta-learning applications for fault diagnosis can be useful for sparse data problems [109]. Meta-learning focuses essentially on learning to learn. It is different from transfer learning, since the objective in meta-learning is to learn from multiple tasks. Formally, meta-learning can be described with Formula 2.16, where the aim is to minimise the expected loss $L$ over a distribution of tasks $p(T)$ by finding suitable hyperparameters $\omega$ [106].

$$\min_{\omega} \mathbb{E}_{T \sim p(T)} L(T; \omega) \quad (2.16)$$

Typically, the suitable hyperparameters $\omega$ are learned by episodic training. The training episodes can consist of $M$ source tasks $D_{sampled} = \{(D_{support}, D_{query})^{(i)}\}_{i=1}^M$ with data sampled from the dataset $D$. In each training episode, a model is trained with the support data $D_{support}$ and then validated with the query data $D_{query}$. The meta learning relates to the optimisation of the hyperparameters $\omega$ between the training episodes. Depending on the meta learning algorithm, the optimised hyperparame-
ters can relate to the initialisation parameters $\theta$ of the base-learner $f_\theta(\cdot)$, the gradient optimiser, or to some metrics in the output space, for example [110].

Few-shot learning is an application field closely related to meta learning. Few-shot learning can be described as a process where a model is optimised to recognise unseen samples based on very few examples. Commonly, few-shot problems are described with the N-way and K-shot notation. That is, the learning model $f_\theta$ is trained to recognise N classes with K training samples from each class. For example, the support data $D_{\text{support}}$ could include 3 bearing health conditions and 1 vibration example from each condition. The learning model then needs to classify 3 similar bearing health conditions from new unseen vibration samples in the query data $D_{\text{query}}$. In essence, few-shot learning algorithms are trained with episodes to recognise similar samples, and to distinguish different samples. Some of the best performing models for such few-shot learning are matching networks [111], prototypical networks [112], and relation networks [113]. Some promising results related to fault diagnosis of rotating machines have also been published [33, 110, 114].
3. Materials and Methods

This chapter describes the materials and methods employed in this dissertation. Overall, all the research in this dissertation concerns experimental research and statistical analysis of the results. Publication I presents a virtual sensor developed with the experimental dataset acquired in the laboratory. Publications II and III study fault diagnosis models with experimental data acquired in the laboratory and in the real maritime environment.

All the publications concern the condition monitoring of rotating machinery using deep learning algorithms. However, Publication I relates to a different condition monitoring domain to Publications II and III. Publication I presents a virtual sensor, or synonymously a soft sensor, for vibration estimation of a large rotor. A virtual sensor application can be useful for the condition monitoring of production processes, such as paper or metal production. Publications II and III present studies for the fault diagnosis of rotating machinery.

Despite the different applications, each publication studies some aspect of generalisation related to optimisation of deep learning models for rotating machinery applications. Publication I demonstrates the effect of scarce and sufficient training data distribution in the operating condition space. Publication II studies the performance of deep learning models on different input data combinations. That is, the classification accuracy of the models was compared between lateral and torsional vibration sensors. Publication III presents two case studies on the effect of activation normalisation on the diagnosis accuracy of faulty rotating machinery.

3.1 LSTM-based vibration estimator from indirect measurements

The objective in the research for Publication I was to develop a data-driven virtual sensor for a large flexible rotor. A data-driven virtual sensor is a computational model optimised with historical data to estimate a quantity that could be measured with a sensor. The data-driven virtual sensor in
this research was optimised to estimate the centre-point movement (CPM) of a full-scale paper machine roll. The dataset included real measured vibration data of the full sized paper machine roll in Figure 3.1. Section 3.1.1 briefly introduces the dataset. The data acquisition process and the dynamics of the rotor system are explained in more detail in the publication presenting the device for adjusting a horizontal foundation stiffness of a rotor [115].

One specific aim in the study in Publication I was to demonstrate the generalisability of RNNs over a broad range of operating conditions. Therefore, a training procedure was developed and experimented in the study. Section 3.1.2 presents this training procedure.

The virtual sensor was based on LSTM, since LSTM is one of the most effective RNNs for long sequences. Table 3.1 details the architecture of the model in this study. The model architecture established a computational algorithm similar to a state-space model. That is, the model computed the rotor displacement from the bearing reaction forces in a deterministic order one time step at a time. Thus, the input and output dimensions in Table 3.1 correspond to the sampled sensor values at each time instant. The LSTM architecture consisted of four layers on both bidirectional paths. Each cell held 81 variables in the memory. The fully connected layer combined both bidirectional paths and computed the output values for each time step.

The bidirectional path introduces some constraints to the real-time application of such virtual sensors. Bidirectional RNNs compute the input sequences in opposite directions. With time-series data, the other RNN path computes the data in the reverse direction. However, computing the input sequence backwards in time merely introduces a small time lag to the most recently estimated output value.
Table 3.1. LSTM-based virtual sensor architecture. (Publication I)

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Input dimension</td>
<td>6</td>
</tr>
<tr>
<td>Output dimension</td>
<td>2</td>
</tr>
<tr>
<td>Number of layers</td>
<td>4</td>
</tr>
<tr>
<td>Bidirectional</td>
<td>True</td>
</tr>
<tr>
<td>Cell size</td>
<td>81</td>
</tr>
<tr>
<td>Fully-connected dimension</td>
<td>2x81</td>
</tr>
</tbody>
</table>

### 3.1.1 Rotor vibration data description

The dataset for virtual sensor development was acquired on a laboratory test rig including a full-scale paper machine roll. Figure 3.1 shows this test rig. The dataset includes vibration data of the rotor and the bearing housings. The rotor vibration, expressed as the CPM of the middle cross-section, was measured with four laser displacement sensors. The CPM was computed from the four displacement signals with the four-point method [49]. The bearing housing forces were measured with six force sensors. Figure 3.2 shows three of the force sensors mounted on a bearing housing. A similar system was assembled on both, the tending side and drive side bearings. Figure 3.3 shows an example of measured rotor CPM and bearing reaction forces.

![Figure 3.2](image.png)

**Figure 3.2.** Two vertically directed force sensors were placed between the bearing housing and the cradle. One horizontally directed sensor was mounted between the cradle and the stiffness adjustment mechanism [115]. (Publication I)

In order to acquire data with a representative distribution of the rotor system dynamics, the rotating speed and horizontal foundation stiffness were varied during the data acquisition. For each rotating speed between 4 Hz and 18 Hz in 0.05 Hz steps, rotor vibration was measured for 100 full revolutions at all possible horizontal foundation stiffnesses. The horizontal foundation stiffness was controlled at both rotor ends with the device...
Materials and Methods

Figure 3.3. Example of drive-end bearing reaction forces (left) and rotor centre-point movement at the middle cross-section (right). (Publication I)

shown in Figure 3.2. The horizontal foundation stiffness could be varied in a range between 2.04 MN/m and 18.32 MN/m. Depending on the rotating speed, the measured stiffness range varied. At higher speeds, the lower stiffness range was avoided because of the close proximity to the critical speed of the rotor.

The complete dataset consists of over 5000 vibration samples, each including 100 revolutions of data. However, the sampling rate varied significantly between these samples. The sampling events in the data acquisition system were phase locked and triggered by an encoder 1024 times per revolution. Therefore, the sampling frequency in the dataset changed with the rotating speed of the rotor. Each vibration sample was resampled to a sampling rate of 2000 Hz. The dataset is available online [116].

3.1.2 Training procedure for generalisation over many operating conditions

Changes to the operating conditions can crucially affect the vibration of the rotor. For example, the vibration amplitudes of the test rig can grow in parallel with the rotating speed. Furthermore, as the horizontal vibration stiffness of the rig was varied, the horizontal natural frequencies changed as well [115]. These variables to the operating conditions can introduce complexity to the system dynamics, which increases the difficulty of training a data-driven virtual sensor.

A data-driven virtual sensor based on LSTM for many operating conditions can be optimised with supervised learning. However, LSTM requires many training samples to learn the system dynamics. Moreover, these training samples should contain vibration from much of the operating condition space. That is, the training samples should include vibration measured at 4 Hz, 18 Hz and many rotating speeds in between. In addition,
the training samples should cover the horizontal foundation stiffness range densely.

To optimise the LSTM to many operating conditions, the training samples were randomly divided in the operating condition space constituted by the rotating speed and horizontal foundation stiffness. Figure 3.4a shows the division into training and test data, where 25% of the vibration samples were randomly selected for training and 75% for testing. In the figure, each red and black point in the heatmap represents a vibration sample with 100 revolutions of data. The red vibration samples were assigned to the training dataset and the black vibration samples were included in the test dataset. The lighter points in the operating condition space were excluded from training and testing. With such a division, the model could learn the system dynamics over the complete operating condition space.

To validate the hypothesis regarding the necessity of training the model with many vibration samples densely distributed in the operating condition space, another training and test data division was designed. Figure 3.4b shows this control division. The training set in this division consisted of 50% of the vibration samples between 4 Hz and 9 Hz rotating speeds. The vibration samples acquired under higher rotating speeds were excluded for testing purposes.

In addition to random training and test data division in the operating condition space, another technique for data augmentation was used. This data augmentation technique is time window division. Figure 3.5 demonstrates how time window division increases the amount of training data. Each 100-revolution sample selected for training purposes was divided into shorter windows. These roughly 1.4 second-long and overlapping time windows were used as the training samples for the model. These time windows were used for model optimisation in random order to introduce more stochasticity into the training.

The LSTM-based virtual sensor was optimised with supervised learning. The target function during optimisation was MSE, as shown in Equation 2.14. More specifically, the MSE was calculated between the measured and estimated CPM signals, which all included 1.4 s of vibration. The test performance was evaluated in the time domain and the frequency domain. The estimated time domain signals were compared to the measured time domain signals with MAE, as shown in Equation 2.15. The frequency domain representations of the estimated signals and the measured signals were compared quantitatively.

### 3.2 Torsional and lateral vibration-based intelligent fault diagnosis

The aim of Publication II was to compare torsional vibration to the widely used lateral vibration for deep learning-based gear fault diagnosis. The
(a) Training vibration samples (red) and test vibration samples (black) divided randomly in the operating condition space. The training set included 25% and the test set included 75% of the samples.

(b) Data division procedure applied to the control experiments. 50% of the samples below a 9 Hz rotating speed were selected to the training set (red). All vibration samples over a 9 Hz rotating speed constituted the test set (black).

Figure 3.4. Data division procedure in the operating condition space divides the vibration samples randomly based on the rotating speed and the horizontal support stiffness. The red points are the training samples and the black points are the test samples including vibration data from 100 revolutions of the paper machine roll. Figure (a) shows the data divided randomly over all operating conditions. Figure (b) shows the control data division for the experiment where the training data was limited in the rotating speed dimension. (Publication I)
Materials and Methods

Figure 3.5. Data augmentation technique for increasing the number of samples in the training dataset. Shorter and overlapping windows can be extracted from the original vibration sample. In this example, overlapping time windows are extracted from a 10-second sample including two bearing reaction force signals. (Publication I)

study employed a laboratory drive train designed to share the natural frequencies with a maritime thruster. A bevel gear was designed and installed on the rig with multiple artificially produced gear faults. A thorough data acquisition system was mounted on the test rig for torsional and lateral vibration data acquisition. The test rig, the gear faults and the acquired data are briefly explained in Section 3.2.1. A more thorough description of the down-scaled thruster components and dynamics were presented in another study [117]. Three popular 1D CNN-based models were optimised in extensive ablation studies on the vibration data acquired with the different sensors. Section 3.2.2 details the model architectures and presents the optimisation algorithm. Section 3.2.3 describes the ablation studies conducted with the 1D CNN models on the different data combinations from the lateral and torsional vibration sensors.

3.2.1 Description of the test rig and the fault data

The test rig representing an azimuth thruster in terms of the natural frequencies of the system consisted of many components. Figure 3.6 shows the test rig in the configuration used for data acquisition in this study. The test rig included a driving motor and a load motor, multiple shafts and flywheels, bearings and two gears. The upper gear ratio was 3:1, and the lower gear ratio was 4:1. In addition to the components, Figure 3.6 shows the sensors of the data acquisition system and their correspond-
Materials and Methods

Figure 3.6. Test rig representing a down-scaled azimuth thruster and the vibration measurement instrumentation. The artificial faults were produced in the lower gear box shown in Figure 3.7 without the roof of the housing. (Publication II)

ing positions. The data acquisition system included four piezo-electric accelerometers, five rotary encoders and two torque transducers. The data was acquired simultaneously with all sensors at a 3 kHz sampling rate. The rotary encoders measured angular displacement signals. The angular displacement signals were differentiated twice to angular acceleration signals.

The artificial gear faults were produced in the lower gear box, as shown in Figure 3.7. The lower gear box contains a bevel gear with 15 teeth in the pinion and 60 teeth in the wheel. Figure 3.7 (a) shows the bevel gear. The artificial gear faults were produced by gluing thin metal sheets, such as the one in Figure 3.7 (b), to the pinion gear, as shown in Figure 3.7 (c). Three different thicknesses of metal sheets were glued onto the pinion gear teeth. In addition, one, two or three sheets with the same thickness were attached to the teeth to produce more gear faults. Table 3.2 lists the different health conditions in the dataset. For each condition, 150 seconds of data was simultaneously measured with all sensors at speeds from 500 RPM to 1500 RPM in 250 RPM steps. These rotating speeds indicate the speed of the drive shaft between the drive motor and the upper gear.

Table 3.2. Produced gear health conditions and related indices. (Publication II)

<table>
<thead>
<tr>
<th>Condition index</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sheet thickness (mm)</td>
<td>N/A</td>
<td>0.01</td>
<td>0.01</td>
<td>0.01</td>
<td>0.03</td>
<td>0.03</td>
<td>0.03</td>
<td>0.05</td>
<td>0.05</td>
<td>0.05</td>
</tr>
<tr>
<td>Num sheets</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
</tbody>
</table>
Materials and Methods

(a) The bevel gear in the lower gear box without the housing roof.

(b) One thin metal sheet used for the artificial gear faults.

(c) Thin metal sheet glued onto the gear pinion.

Figure 3.7. Thin metal sheets were glued on the teeth of the lower gear pinion in order to produce artificial gear faults. (Publication II)

3.2.2 Model architectures and optimisation

Three models were trained for gear fault diagnosis in this work. The models have been proposed and evaluated for fault diagnosis purposes previously. All three models are different 1D CNNs. Tables 3.3, 3.4 and 3.5 list the architectures of these models. Each of the model architectures include several modifications specific to this study.

Table 3.3 presents the architecture of Ince’s model. Ince’s model was originally proposed for detecting bearing faults from motor current signals [38]. In contrast to the previously proposed model, each of the three 1D convolutional layers (Conv1D) was followed by batch normalisation. Furthermore, the number of input channels were adjusted to each task. The number of input channels corresponded to the number of sensors used to acquire the vibration data in each task. In addition, the stride and padding of the first layer were adjusted accordingly to enable the processing of vibration samples of different lengths. Strides of 5, 10 or 20 paired with the padding of 18, 36 or 72 enabled the model to process vibration samples with lengths of 2048, 4096 or 8192, respectively. Finally, the fully connected layer (FCL) estimated 4 or 10 probabilities depending on the ablation study. The probabilities corresponded to the gear health states included in each ablation study.

Table 3.4 shows the details for the architecture of the second model in these fault diagnosis experiments. This Deep Convolutional Neural Network with Wide First-layer Kernels (WDCNN) was proposed for bearing fault diagnosis from lateral vibration. WDCNN was shown to recognise
Ince’s model architecture and the modifications used in this study. (Publication II)

<table>
<thead>
<tr>
<th>Layer</th>
<th>Kernel Size</th>
<th>Channels In</th>
<th>Filters</th>
<th>Stride</th>
<th>Padding</th>
</tr>
</thead>
<tbody>
<tr>
<td>Conv1D</td>
<td>9 × 1</td>
<td>1 to 11</td>
<td>60</td>
<td>5 / 10 / 20</td>
<td>18 / 36 / 72</td>
</tr>
<tr>
<td>Conv1D</td>
<td>9 × 1</td>
<td>60</td>
<td>40</td>
<td>5</td>
<td>4</td>
</tr>
<tr>
<td>Conv1D</td>
<td>9 × 1</td>
<td>40</td>
<td>40</td>
<td>9</td>
<td>3</td>
</tr>
<tr>
<td>FCL</td>
<td>1 × 1</td>
<td>400</td>
<td>4 / 10</td>
<td>N/A</td>
<td>N/A</td>
</tr>
</tbody>
</table>

Patterns related to the bearing conditions in the original study [21]. The architecture of the model includes five 1D-convolutional layers with batch normalisation and max pooling (Conv1D). The final two fully-connected layers compute the probabilities for each health state from the features extracted by the convolutional layers. Similarly to Ince’s model, the original architecture was modified for the purposes of the ablation studies. The input channels of the first layer were varied between 1 and 11 to match with the number of sensors used to acquire the vibration data in the experiments. In addition, different window sizes were experimented with. The stride of the first layer was adjusted to 16, 32 and 64 in the ablation studies with window sizes of 2048, 4096 and 8192, respectively. Finally, the last fully connected layer computed 4 or 10 probability values corresponding to the 4 or 10 health states included in the ablation studies.

WDCNN model architecture with the modifications used in this study. (Publication II)

<table>
<thead>
<tr>
<th>Layer</th>
<th>Kernel Size</th>
<th>Channels In</th>
<th>Filters</th>
<th>Stride</th>
<th>Padding</th>
</tr>
</thead>
<tbody>
<tr>
<td>Conv1D</td>
<td>64 × 1</td>
<td>1 to 11</td>
<td>16</td>
<td>16 / 32 / 64</td>
<td>24</td>
</tr>
<tr>
<td>Conv1D</td>
<td>3 × 1</td>
<td>16</td>
<td>32</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Conv1D</td>
<td>3 × 1</td>
<td>32</td>
<td>64</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Conv1D</td>
<td>3 × 1</td>
<td>64</td>
<td>64</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Conv1D</td>
<td>3 × 1</td>
<td>64</td>
<td>64</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>FCL</td>
<td>1 × 1</td>
<td>192</td>
<td>18</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>FCL</td>
<td>1 × 1</td>
<td>18</td>
<td>4 / 10</td>
<td>N/A</td>
<td>N/A</td>
</tr>
</tbody>
</table>

Table 3.5 shows the architecture and modifications of the third model adapted in this study. Similarly to WDCNN, a Stacked Residual Dilated Convolutional Neural Network (SRDCNN) was demonstrated to be an effective model for bearing fault diagnosis from lateral vibration data [32]. SRDCNN consists of five dilated convolutional layers with residual connection (RDConv1D) and two fully-connected layers. Each dilated con-
volutional layer with residual connection has three adjacent convolutional filters, which establish a gating mechanism. The gating mechanism functions similarly to the input gate mechanisms in LSTM. The activations of the first two 1D-convolutional filters are multiplied element-wise together. The activation function of the first filter is sigmoid, which scales the activation values to the range between 0 and 1. This filter can be considered as the control gate. The activation function of the second filter is tanh, which scales the activation values to the range between -1 and 1. This filter can be considered as the proposal gate. The activation values of the control gate dictate the share of the activations of the proposal gate passed forward through the layer. The third filter establishes a residual connection. That is, this filter merely scales the dimensions of the layer input values accordingly. The residual values are then summed to the controlled proposal values.

Similarly to Ince’s model and WDCNN, SRDCNN was also modified for the purposes of this study. First, the input channels were varied according to the number of sensors used to acquire the vibration data for the experiments. Second, the stride was adjusted to 2, 4 or 8 to match the different vibration sample lengths of 2048, 4096 and 8192, respectively. Third, the final fully connected layer computed 4 or 10 probabilities, which corresponded to the number of health states specific to the ablation studies.

Table 3.5. SRDCNN model architecture and the modifications used in this study. (Publication II)

<table>
<thead>
<tr>
<th>Layer</th>
<th>Kernel Size</th>
<th>Channels In</th>
<th>Filters</th>
<th>Stride</th>
<th>Padding</th>
<th>Dilation</th>
</tr>
</thead>
<tbody>
<tr>
<td>RDConv1D</td>
<td>64 × 1</td>
<td>1 to 11</td>
<td>32</td>
<td>2 / 4 / 8</td>
<td>31</td>
<td>1</td>
</tr>
<tr>
<td>RDConv1D</td>
<td>32 × 1</td>
<td>32</td>
<td>32</td>
<td>2</td>
<td>31</td>
<td>2</td>
</tr>
<tr>
<td>RDConv1D</td>
<td>16 × 1</td>
<td>32</td>
<td>64</td>
<td>2</td>
<td>30</td>
<td>4</td>
</tr>
<tr>
<td>RDConv1D</td>
<td>8 × 1</td>
<td>64</td>
<td>64</td>
<td>2</td>
<td>28</td>
<td>8</td>
</tr>
<tr>
<td>RDConv1D</td>
<td>4 × 1</td>
<td>64</td>
<td>64</td>
<td>2</td>
<td>24</td>
<td>16</td>
</tr>
<tr>
<td>FCL</td>
<td>1×1</td>
<td>4096</td>
<td>100</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>FCL</td>
<td>1×1</td>
<td>100</td>
<td>4 / 10</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
</tr>
</tbody>
</table>

This study consisted of numerous experiments. In each experiment, the models were trained to diagnose the vibration samples that included data from a specific set of sensors. Every experiment followed similar procedures for data preprocessing and model optimisation. The data preprocessing consisted of data division and data augmentation. The data division separated training, validation and test sets from the 150-second vibration samples. Specifically, the first 50% was assigned to the training set, the third quarter to the validation set and the last quarter to the...
Figure 3.8. Training, validation and test sets included, respectively, the first half, the third quarter and the last quarter of every 150-second vibration sample included in the experiment. This example demonstrates the data division for the experiments conducted with vibration data acquired with both torque transducers. (Publication II)

test set. A similar division was followed for every 150-second vibration sample corresponding to the experiment. That is, the training data included roughly the first 75 seconds measured at the five discrete rotating speeds between 500 RPM and 1500 RPM with all the different gear health states. Figure 3.8 shows this data division procedure applied for the 500 RPM vibration samples acquired with the healthy gear and the torque transducers.

The data augmentation included the extraction of overlapping windows from the divided vibration samples. Three different window sizes were used: 2048, 4096 and 8192 time steps. The overlap between adjacent time windows was 128 time steps. The windows consisted of 1 to 11 channels each corresponding to data acquired with a particular sensor. Table 3.6 lists the number of vibration samples in the augmented training, validation and test sets in every experiment. The number of augmented samples in these datasets depended on the window size and the number of health states included in the experiment. Each experiment in an ablation study included same number of samples. The ablation studies are explained in more detail in the following Section 3.2.3.

The model optimisation algorithm was similar in every experiment for all three models. The criterion for the error was based on the cross-entropy loss. The models were validated against the validation data set between every epoch. The maximum number of epochs in each experiment was set to 20. The model parameters corresponding to the lowest validation error were used for evaluating the trained models against the test set.
Table 3.6. The number of samples in the augmented training, validation and test sets per experiment in the ablation studies with different window sizes and gear health states. (Publication II)

<table>
<thead>
<tr>
<th>Ablation study</th>
<th>Window size</th>
<th>Conditions</th>
<th>Training samples</th>
<th>Validation samples</th>
<th>Test samples</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>2048</td>
<td>All</td>
<td>86572</td>
<td>42486</td>
<td>42486</td>
</tr>
<tr>
<td>B</td>
<td>4096</td>
<td>All</td>
<td>84972</td>
<td>40886</td>
<td>40886</td>
</tr>
<tr>
<td>C</td>
<td>8192</td>
<td>All</td>
<td>81772</td>
<td>37686</td>
<td>37686</td>
</tr>
<tr>
<td>D</td>
<td>2048</td>
<td>1-4</td>
<td>34632</td>
<td>16996</td>
<td>16996</td>
</tr>
</tbody>
</table>

3.2.3 Description of ablation studies

The experiments were organised in extensive ablation studies, A, B, C and D, with different window sizes and gear health states as Table 3.6 shows. The same experiments were conducted in each ablation study. The differences between the ablation studies relate to the length of the window, i.e., 2048, 4096 or 8192, and to the number of health states included in the experiments. Ablation study D was conducted with the least severe faulty conditions only, since they were considered more difficult to recognise. The differences between the experiments related to the sensors used to acquire the data. Each ablation study included 38 different experiments. Table 3.7 lists all the experiments and corresponding sensors. In the ablation studies, the experiments were grouped into one sensor experiments, two sensor experiments and experiments with many sensors.

Figure 3.9 summarises the ablation studies in this work. For each ablation study, all 38 experiments with vibration data from different sensors were conducted. In each experiment, all three models were optimised and evaluated 20 times. The experiments were repeated in order to reduce uncertainty of the model performances. The model performances formed the basis for the comparison between the experiments with different vibration data. The performance of the models was expressed in terms of the average accuracy of 20 trials in an experiment.

![Ablation study diagram](https://via.placeholder.com/150)

Figure 3.9. All three models were optimised and evaluated 20 times in each of the 38 experiments in each ablation study. The ablation studies included a total of 9120 optimisation and training runs. (Publication II)
Table 3.7. Description of the sensors used to acquire data for each experiment in the ablation studies. (Publication III)

<table>
<thead>
<tr>
<th>Sensor #</th>
<th>Sensors</th>
<th>Experiment name</th>
<th>Sensors</th>
<th>Experiment name</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Encoder 1 E1</td>
<td>Accelerometer 1 A1</td>
<td>Encoder 2 E2</td>
<td>Accelerometer 2 A2</td>
</tr>
<tr>
<td></td>
<td>Encoder 3 E3</td>
<td>Accelerometer 3 A3</td>
<td>Encoder 4 E4</td>
<td>Accelerometer 4 A4</td>
</tr>
<tr>
<td></td>
<td>Encoder 5 E5</td>
<td>Upper Torque Transducer T1</td>
<td>Lower torque Transducer T2</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Encoders 1,2 E12</td>
<td>Encoders 1,3 E13</td>
<td>Encoders 1,4 E14</td>
<td>Encoders 1,5 E15</td>
</tr>
<tr>
<td></td>
<td>Encoders 2,3 E23</td>
<td>Encoders 2,4 E24</td>
<td>Encoders 2,5 E25</td>
<td>Encoders 3,4 E34</td>
</tr>
<tr>
<td></td>
<td>Encoders 3,5 E35</td>
<td>Encoders 4,5 E45</td>
<td>Accelerometers 1,2 A12</td>
<td>Accelerometers 1,3 A13</td>
</tr>
<tr>
<td></td>
<td>Accelerometers 1,4 A14</td>
<td>Accelerometers 2,3 A23</td>
<td>Accelerometers 1,4 A14</td>
<td>Accelerometers 3,4 A34</td>
</tr>
<tr>
<td></td>
<td>Accelerometers 2,4 A24</td>
<td>Accelerometers 3,4 A34</td>
<td>Upper &amp; Lower torque transducers T all</td>
<td></td>
</tr>
<tr>
<td>&gt;2</td>
<td>Encoders 1,2,3,4,5 E all</td>
<td>Accelerometers 1,2,3 A123</td>
<td>Accelerometers 1,2,3,4 A all</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Encoders 1,2,3,4 &amp; Upper &amp; Lower torque transducers E all, T all</td>
<td>Accelerometers 1,2,3 &amp; Upper &amp; Lower torque transducers A123, T all</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Accelerometers 1,2,3,4 &amp; Upper &amp; Lower torque transducers A all, T all</td>
<td>Accelerometers 1,2,3 &amp; Encoders 1,2,3,4,5 A123, E all</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Accelerometers 1,2,3,4 &amp; Encoders 1,2,3,4,5 A all, E all</td>
<td>Accelerometers 1,2,3 &amp; Encoders 1,2,3,4,5 &amp; Upper &amp; Lower torque transducers A123, E all, T all</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Accelerometers 1,2,3,4 &amp; Encoders 1,2,3,4,5 &amp; Upper &amp; Lower torque transducers A all, E all, T all</td>
<td>Accelerometers 1,2,3 &amp; Encoders 1,2,3,4,5 &amp; Upper &amp; Lower torque transducers A123, E all, T all</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Accelerometers 1,2,3,4 &amp; Encoders 1,2,3,4,5 &amp; Upper &amp; Lower torque transducers A all, E all, T all</td>
<td>Accelerometers 1,2,3,4 &amp; Encoders 1,2,3,4,5 &amp; Upper &amp; Lower torque transducers A all, E all, T all</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
3.3 Whitening of CNN-based fault diagnosis model features

Publication III investigated a potential improvement for the features of all 1D CNN-based fault diagnosis models. The improvement concerned the normalisation of the CNN model features. The adapted feature normalisation technique was named network deconvolution in the original work [118]. The original work provided thorough experiments and arguments regarding the necessity of using ND for feature normalisation instead of batch normalisation. In the previous work, ND was integrated into MLP and 2D CNN models, which were then experimented with in image classification tasks with many related benchmark datasets. Different metrics related to convergence and accuracy showed that ND can be superior to batch normalisation. However, the original work did not provide any results regarding 1D CNN experiments on sequential data.

The research regarding Publication III adapted ND and applied it for 1D CNN models. The models were experimented on with two different datasets. The first dataset was the CWRU dataset, which is a well-established benchmark for fault diagnosis, as discussed in Section 2.3.1. The experiments related to this dataset include the conventional load domain shift approach adapted from many related studies [21, 94, 95]. Although this conventional approach has gained some critique [90], it can still be used for model comparison. The second dataset included vibration data collected from three similar azimuth thrusters between mid-2018 and the end of 2019. Gear and bearing faults occurred in these three thrusters during the period of observation. These datasets and the preprocessing are explained in more detail in Section 3.3.1.

Overall, the fault diagnosis experiments were conducted with supervised learning, and with İnce's model, WDCNN and SRDCNN. The architectures of these models were visited in Section 3.2.2. The models were applied as baselines in a similar configuration in this study. The batch normalisation functions in these models were changed to ND in order to compare the two normalisation techniques. Furthermore, in the experiments concerning the azimuth thruster fault diagnosis, the final fully-connected layers were modified to estimate the probability of a fault. That is, the output dimension of the models was changed to one value. These models were then trained for anomaly detection. The error function in these experiments was thus the binary cross entropy, as shown in Equation 2.13.

In both experiments with the CWRU data and the azimuth thruster data, the models were compared with the average accuracy and the standard deviation of the accuracy of 10 repeated trials. In each trial, the numbers of samples per health state were balanced.
3.3.1 Fault data description

The first case study in this research was based on the conventional load domain shift experiments with the CWRU bearing fault dataset. In the load domain shift experiments, a deep learning model is first trained with all data acquired under some motor load. The model is then tested with data acquired under other loads. Typically, the employed loads are reported in horsepower (hp), and the loads are 1 hp, 2 hp and 3 hp. Respectively, the vibration data related to these loads was also acquired under the following constant rotating speeds: 1772 RPM, 1750 RPM and 1730 RPM. The vibration data was acquired with two accelerometers. One was mounted on the drive-end and one on the fan-end of the drive motor, which is shown in Figure 3.10.

The task for the model in these experiments was to classify the bearing condition from the vibration data acquired with accelerometers. The dataset consisted of 10 different bearing conditions. The 9 faulty conditions were produced by electro-discharge machining dents to the ball (B), the inner ring (I) and the outer ring (O) of the bearings. For each component, dents of three different diameters were machined: 0.1778 mm ($L$), 0.3556 mm ($M$) and 0.5334 mm ($H$). This study used the data related to the drive-end bearing. The vibration samples related to these faults were divided into shorter windows, as demonstrated in Figure 3.5. The window size for these samples was 2048 time steps. The overlap between training samples was 2016 time steps. The test samples were not overlapped. Table 3.8 presents the complete augmented dataset for this study.

The second case study in this research concerned the fault detection of azimuth thrusters. An azimuth thruster is a device that marine vessels can use for propulsion or manoeuvre. Figure 3.11 shows an illustration of the azimuth thrusters analysed in this case study. This case study included...
Table 3.8. Number of samples per class and load domain in the augmented bearing fault dataset. (Publication II)

<table>
<thead>
<tr>
<th>Load</th>
<th>Split</th>
<th>H</th>
<th>B_L</th>
<th>B_M</th>
<th>B_H</th>
<th>I_L</th>
<th>I_M</th>
<th>I_H</th>
<th>O_L</th>
<th>O_M</th>
<th>O_H</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Test</td>
<td>59</td>
<td>59</td>
<td>59</td>
<td>59</td>
<td>59</td>
<td>59</td>
<td>59</td>
<td>59</td>
<td>59</td>
<td>59</td>
</tr>
<tr>
<td></td>
<td>Test</td>
<td>59</td>
<td>59</td>
<td>59</td>
<td>59</td>
<td>59</td>
<td>59</td>
<td>59</td>
<td>59</td>
<td>59</td>
<td>59</td>
</tr>
<tr>
<td></td>
<td>Test</td>
<td>59</td>
<td>59</td>
<td>59</td>
<td>59</td>
<td>59</td>
<td>59</td>
<td>59</td>
<td>59</td>
<td>59</td>
<td>59</td>
</tr>
</tbody>
</table>

Figure 3.11. Azimuth thruster with the vibration sensors, and faulty component locations indicated. The accelerometers were placed near the input shafts of the thrusters. The faults occurred in the gear and the bearing for the pinion gear. (Publication III)

vibration data of three such thrusters in healthy and faulty condition. During the data acquisition, faults occurred in these three thrusters in the gear and the bearing of the pinion shaft. The time the faults occurred in the bearings and gears did not completely overlap. The vibration data was acquired with an accelerometer mounted in a similar location on each thruster. Figure 3.11 shows the locations of the faulty components and the sensor.

The vibration data was acquired with encoder-triggered sampling 1024 times per revolution. Each vibration sample included four revolutions, and thus the length of the samples was 4096. During the data acquisition, the thrusters were operated at many different rotating speeds. Figure 3.12 shows the distribution of the rotating speeds in the data. Because of the varied rotating speeds and encoder triggered sampling, the sampling rate also varied in the dataset.

During the monitoring period of roughly a year and a half, the three
thrusters were healthy in the beginning but suffered similar failures of the pinion shaft bearing and the gear. These failures were detected and labeled by experts. In addition to labelling, the experts also assessed the severity of the faults. Since the failures occurred on different dates, and the thrusters were kept in operation for some time after the failures occurred, the severity of the faults varied in the dataset. Thus, the labels in this study merely state if the bearing and the gear is healthy or faulty. Table 3.9 lists the number of samples per thruster and the component health state. In the table, B refers to the bearing and G refers to the gear.

The lowest rows of the Table 3.9 show the number of augmented training, validation and test samples for each fault detection experiment. The augmentation consisted of time window division, similar to the bearing fault diagnosis experiments with the CWRU dataset. The windows included 2048 time steps, and the overlap between the training and validation time windows was 2016 time steps. The test time windows did not overlap. Training, validation and test time windows were extracted from different vibration samples. The sets included equal number of healthy and faulty samples. Two different experiment types were conducted with the dataset. The first experiment type consisted of optimising the models for bearing fault detection. The second experiment type concerned gear fault detection.

### 3.3.2 Implicit and approximate whitening of 1D CNN features

Deep learning models tend to converge to the optimum faster if the input data has been normalised. That is, with zero mean, unit variance and decorrelated data, the required number of gradient steps and training samples in a batch decreases. Currently, state-of-the-art normalisation techniques can not only normalise the input to a deep model, but also the input to every layer of the network [36]. By a large margin, the most
Table 3.9. Number of healthy and faulty vibration samples per thruster and component. The lowest row lists the number of augmented vibration samples in the training, validation and test sets. (Publication II)

<table>
<thead>
<tr>
<th>Thruster</th>
<th>Condition</th>
<th>B</th>
<th>G</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Healthy</td>
<td>263</td>
<td>283</td>
</tr>
<tr>
<td></td>
<td>Faulty</td>
<td>468</td>
<td>448</td>
</tr>
<tr>
<td>2</td>
<td>Healthy</td>
<td>108</td>
<td>200</td>
</tr>
<tr>
<td></td>
<td>Faulty</td>
<td>843</td>
<td>751</td>
</tr>
<tr>
<td>3</td>
<td>Healthy</td>
<td>125</td>
<td>216</td>
</tr>
<tr>
<td></td>
<td>Faulty</td>
<td>405</td>
<td>314</td>
</tr>
<tr>
<td>Total</td>
<td>Healthy</td>
<td>496</td>
<td>699</td>
</tr>
<tr>
<td></td>
<td>Faulty</td>
<td>1716</td>
<td>1513</td>
</tr>
<tr>
<td>Data splits</td>
<td>Train</td>
<td>48230</td>
<td>77740</td>
</tr>
<tr>
<td></td>
<td>Val</td>
<td>11960</td>
<td>19110</td>
</tr>
<tr>
<td></td>
<td>Test</td>
<td>368</td>
<td>592</td>
</tr>
</tbody>
</table>

common normalisation technique is batch normalisation. BN can shift and scale the inputs to a layer based on batch statistics. By consistently shifting and scaling the inputs to a layer during the training, BN stabilises the input distribution to each weight. This stabilisation is considered to mitigate the internal covariate shift problem [35]. However, BN does not attempt to decorrelate the activations, since computing the covariance matrix and the inverse square root of the covariance matrix is considered to be expensive [35]. Despite this compromise, BN is the widely used, mostly undiscussed and undisputed method of choice for many, if not all current CNN-based fault diagnosis models.

This study adapted network deconvolution to 1D CNN-based fault diagnosis models. ND is an implicit and approximative technique for layer input whitening. In essence, ND can be considered as a similar layer normalisation technique to BN, but with the capability to decorrelate the features: i.e., ND was designed to shift, scale and decorrelate inputs to each weight. More specifically, ND can whiten the input data to each weight by computing the inverse square root of the covariance matrix. The following paragraphs review the application of ND for vibration data in more detail.

In order to compute the covariance matrix of the input data to each layer, the vibration is first transformed with a similar transform to the commonly known image to column transform (im2col). This transform rearranges the data and the weights of the kernels in the convolutional layer into columns. Conventionally, the convolutional computations presented in Equation 3.1 are considered as kernels $k_{j,n}^l$, computing cross-correlations on some section of the vibration data $x_{n,i}^l$. The activation $y_{j,i}^l$ relates to the $i$-th
Materials and Methods

Input channel 1

Input channel 2

\[ X^{i=1}_{n=1} \quad X^{i=1}_{n=2} \]

\[ X^{i=2}_{n=1} \quad X^{i=2}_{n=2} \]

\[ k^{n=1} \quad k^{n=2} \]

**Figure 3.13.** Computations of one convolutional filter with two kernels on vibration data from two sensors represented in the matrix multiplication format. (Publication III)

Section of the input data to the layer \( l \) computed by the filter \( j \). The filters consist of \( N \) kernels with \( M \) weights. Each kernel \( k_{*, n}^j \) corresponds to an input channel \( n \). Considering the vibration data, these cross-correlations between \( k_{j,n}^l \) and \( x_{i,n}^{l,i} \) are computed on many sections along the data. On the first layer, the kernels can be considered to slide through the vibration data along the time dimension. Figure 3.13 demonstrates the im2col transform for the first layer. In the figure, the input data consists of two vibration samples, i.e., input channels, with \( 2 \times M \) time steps. The convolutional filter consists of two kernels with \( M \) weights. Both kernels \( k_n \) compute two cross-correlations on the corresponding vibration sections \( x_{n}^{i=*} \). The sections do not overlap. The adjacent vibration sections processed by the same kernel are then stacked into the columns of the transformed matrix \( X \). The transformed matrix \( X \) includes \( 2 \times M \) columns that correspond to the \( M \) weights of two kernels. Simultaneously, the weights in the kernels are stacked into one column vector. The output of this filter is the matrix vector multiplication \( Xw \). Additional vibration samples in a minibatch would increase the number of rows in \( X \), and additional channels would increase the number of columns.

\[ y^{l,i}_{j} = \sum_{n=0}^{N} k_{j,n}^l \ast x^{l,i}_{n} + b_{j,n} \]  

(3.1)

The covariance matrix of the input data to a layer is then computed on centred data, as Equation 3.2 shows. The mean vector \( \mu \) contains the means of the inputs to each weight. That is, \( \mu \) includes the column means of \( X \). \( N \) corresponds to the number of sections in each column in the transformed matrix \( X \).

\[ \text{Cov} = \frac{1}{N}(X - \mu)^T (X - \mu) \]  

(3.2)

ND whitens the inputs with the inverse square root of the covariance
matrix. The inverse square root is approximated iteratively. That is, the matrix is approximated with the coupled Newton-Schulz iteration. Before the iteration, two matrices are initialised: $Y_0 = \text{Cov} + \epsilon \cdot I$ and $Z_0 = I$. The matrices are updated every iteration with Equations 3.3 and 3.4. During the iteration, the matrices $Y_k$ and $Z_k$ converge approximately to the square root and inverse square root of the covariance matrices, as shown in Formulas 3.5 and 3.6.

$$Y_{k+1} = \frac{1}{2} Y_k (3I - Z_k Y_k) \quad (3.3)$$

$$Z_{k+1} = \frac{1}{2} (3I - Z_k Y_k) Z_k \quad (3.4)$$

$$Y_{k+1} \rightarrow \text{Cov}^{\frac{1}{2}} \quad (3.5)$$

$$Z_{k+1} \rightarrow \text{Cov}^{-\frac{1}{2}} \quad (3.6)$$

Equation 3.7 shows the whitening computations. The activation values $y^l_j$ of filter $j$ in layer $l$ are computed with centred input data $(X^l - \mu^l)$ and weights corrected with the deconvolution matrix $D \approx (\text{Cov} + \epsilon \cdot I)^{-\frac{1}{2}}$, where $\epsilon$ is a small constant added for numerical stability. Since the whitening correction is applied to the weights instead of directly to the input data, the whitening correction can be considered as implicit.

$$y^l_j = (X^l - \mu^l) \cdot (D^l \cdot w^l_j) + b^l_j = X^l \cdot D^l \cdot w^l_j + b^l_j - \mu^l \cdot D^l \cdot w^l_j \quad (3.7)$$
4. Results

This chapter reviews the most important results of this research. Section 4.1 details the findings regarding Publication I. Section 4.2 shows the results regarding Publication II. Section 4.3 views the research findings of Publication III.

4.1 Estimation accuracy of LSTM-based virtual sensor

The aim of Publication I was to train the rotor system dynamics over a large range of operating conditions for a data-driven virtual sensor. For this purpose, and considering the supervised learning approach, a specific strategy for training and test data division was designed. The training and test data were constructed by randomly dividing the vibration samples based on the rotating speed and the horizontal foundation stiffness of the system during data acquisition. The importance of such a uniform random division was evaluated by comparing the model accuracy to the accuracy of another model trained with data including limited rotating speeds. Figure 3.4 shows these two data divisions.

The virtual sensor was evaluated with 5-second vibration samples. One such sample was extracted from each 100-revolution-long test vibration sample measured at a specific rotating frequency and foundation stiffness. Figure 4.1 shows an example of one extracted 5-second vibration sample. The blue vibration signal was measured when the rotating speed was 10.5 Hz and the horizontal foundation stiffness was 18.32 MN/m. The measured time-series data is presented in the upper two subplots in the horizontal and vertical directions. The vibration signals estimated by the trained model are plotted in red in the same upper two subplots. The mean absolute error between these signals was 0.0064 mm. This error is close to the MAE over all 5-second test windows, which was 0.0063 mm. The related model was trained with the proposed training and test data division technique in Figure 3.4a. The lower two subplots in Figure 4.1 show the discrete Fourier transforms (DFT) of the corresponding time
domain representations in the upper subplots.

Since the horizontal foundation stiffness, which was adjusted in the research, affects the horizontal vibration most significantly, the remaining results section focuses on this direction. Figure 4.2 demonstrates the accuracy of the model estimates for the horizontal CPM. The plots in this figure include the DFTs of all corresponding measured and estimated signals. Signals corresponding to the different horizontal foundation stiffnesses are also included in the plots, even though the horizontal foundation stiffness is not included in the axes of the plots. Figure 4.2a presents the DFTs of the measured 5-second test windows. The corresponding DFTs of the vibration estimated by the model trained with randomly selected training set are shown in Figure 4.2b. The difference between these corresponding DFTs of the measured and estimated vibration is in Figure 4.2c. The difference between the DFTs is consistently less than 0.05 mm.

In relation to the results in Figure 4.2, virtual sensor optimisation was conducted with the limited training data, as shown in Figure 3.4b. The training data included randomly selected samples from a rotating speed range between 4 Hz and 9 Hz. Figure 4.3 shows three different plots related to the performance of the model trained with limited data. Figure 4.3a shows the DFTs of the measured 5-second CPM test windows, which consisted of vibration data acquired at rotating speeds higher than 9 Hz. Figure 4.3b shows the DFTs of the corresponding estimated CPM windows. The amplitudes of the DFTs of the estimated CPM test windows

Figure 4.1. The top two subplots show estimated (red) and measured (blue) CPM with MAE of 0.0064 mm. The two lower subplots show the corresponding discrete Fourier transforms of the estimated and measured CPM signals.
Results

(a) DFTs of the measured 5-second test windows in terms of the rotating speed during the data acquisition.

(b) DFTs of the estimated 5-second test windows corresponding to the measured samples in Figure 4.2a.

(c) Difference between the corresponding DFTs from 5-second test windows from Figure 4.2a and Figure 4.2b.

Figure 4.2. DFTs of the measured (a) and estimated (b) horizontal CPMs and their difference (c). (Publication I)
Results

(a) DFTs of the measured 5-second test windows in terms of the rotating speed during the data acquisition.

(b) DFTs of the estimated 5-second test windows corresponding to the measured samples in Figure 4.3a.

(c) Difference between the corresponding DFTs from 5-second test windows from Figure 4.2a and Figure 4.2b.

Figure 4.3. DFTs of the measured and estimated horizontal CPMs and their difference in the tests where the training data was limited to a range between 4 Hz and 9 Hz. (Publication I)

are consistently lower than their measured counterparts. Furthermore, the error grows significantly with the rotating speed. This relation can be observed from the error plot in Figure 4.3c.

4.2 Torsional vibration data for CNN-based gear fault diagnosis

The research in Publication II concerned the suitability of torsional vibration data for deep learning-based gear fault diagnosis purposes. The results relate to the different experiments with torsional and lateral vibration data from many different sensor sets. The torsional vibration data was acquired with five rotary encoders and two torque transducers. The lateral vibration data was acquired with four accelerometers. The results concerning the lateral vibration data can be considered as baseline results for the torsional vibration-related results.

Table 4.1 lists the average accuracies in the experiments concerning the ablation studies A and D, which yielded the most relevant findings. The
results of experiments in both studies were conducted with the same window size including 2048 time steps. The average accuracies related to ablation studies B and C with larger window sizes are presented in Publication III. The experiments in ablation study A included all the gear health conditions listed in Table 3.2. The experiments in ablation study D included only the least severe gear faults (conditions 2, 3 and 4) and the healthy state. In Table 4.1, the results are grouped based on the number of sensors used to acquire the data for the experiments.

The highest accuracies in the group with one sensor relate to experiments T1 with data from the upper torque transducer. In study A, the accuracy in T1 was 98.3% while the second-highest accuracy in the experiment E5 was 98.1%. Overall, the accuracies were close to 90.0% or more in these experiments in study A. The differences between accuracies of the experiments were more drastic in ablation study D with less severe faults. The accuracy in T1 was 99.9%, while most of the accuracies were significantly below 90.0%, and some even below 80.0%. In the experiment group with two sensors, the highest accuracy in ablation study A related to A13. This accuracy was 98.9% while the second-highest accuracy related to T all was 98.8%. The difference was not significant. In ablation study D the highest accuracy related to T all. Similar to the previous group, this accuracy of 98.7% was again significantly higher than the accuracies in most of the other experiments. In the group with many sensors, the highest accuracy in both studies relates to the same experiments in both ablation studies. The accuracy of A123, T all was 99.9% in study A and 100.0% in study D.

Several important observations can be made from the accuracies presented in Table 4.1. First, the lowest accuracy was always higher in the experiment group with more sensors. Second, the models were very sensitive to the location of the rotary encoders. The models more accurately diagnosed the data from the encoders near the faulty lower gear than the data from other encoders closer to the upper gear. In addition, the models more accurately diagnosed the data from the three accelerometers mounted on the lower gear box housing than the data from the fourth accelerometer on the bearing housing. In contrast, the models diagnosed more accurately the data from the upper torque transducer mounted further away from the lower gear box than the data from the lower torque transducer. Third, the experiments conducted with torque transducer data resulted in the highest accuracies in all experiment groups with one exception, where the difference was small. Fourth, the models were systematically less accurate in the experiments in ablation study D than in ablation study A, if the models diagnosed vibration samples without data from torque transducers. The experiments with torque transducer data sometimes resulted in higher accuracies in ablation study D.

Figure 4.4 shows the average accuracies and standard deviations per model over 20 trials in the six experiments in ablation study A.
### Table 4.1

Average accuracy in the experiments in ablation study A and D. The experiments and corresponding sensors are detailed in Table 3.7. The highest average accuracy of ablation study A in each group has been underlined. Similarly, the highest average accuracy in each group concerning ablation study D has been marked with bold font. (Publication III)

<table>
<thead>
<tr>
<th>Sensor #</th>
<th>Avg. Accuracy (%)</th>
<th>Avg. Accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Experiment A</td>
<td>D</td>
</tr>
<tr>
<td>E1</td>
<td>90.9</td>
<td>72.8</td>
</tr>
<tr>
<td>E2</td>
<td>89.4</td>
<td>74.4</td>
</tr>
<tr>
<td>E3</td>
<td>89.7</td>
<td>71.5</td>
</tr>
<tr>
<td>E4</td>
<td>93.4</td>
<td>82.1</td>
</tr>
<tr>
<td>E5</td>
<td>98.1</td>
<td>97.5</td>
</tr>
<tr>
<td></td>
<td>T2</td>
<td></td>
</tr>
<tr>
<td>E12</td>
<td>92.3</td>
<td>75.9</td>
</tr>
<tr>
<td>E14</td>
<td>91.4</td>
<td>77.7</td>
</tr>
<tr>
<td>E23</td>
<td>91.2</td>
<td>80.6</td>
</tr>
<tr>
<td>E25</td>
<td>93.1</td>
<td>74.3</td>
</tr>
<tr>
<td>E35</td>
<td>94.5</td>
<td>74.1</td>
</tr>
<tr>
<td>A12</td>
<td>98.5</td>
<td>92.3</td>
</tr>
<tr>
<td>A14</td>
<td>95.5</td>
<td>86.6</td>
</tr>
<tr>
<td>A24</td>
<td>95.0</td>
<td>85.8</td>
</tr>
<tr>
<td>T all</td>
<td>98.8</td>
<td><strong>98.7</strong></td>
</tr>
<tr>
<td></td>
<td>A all</td>
<td>99.0</td>
</tr>
<tr>
<td>A all, T all</td>
<td>99.8</td>
<td><strong>100.0</strong></td>
</tr>
<tr>
<td>A all, E all</td>
<td>94.2</td>
<td>88.6</td>
</tr>
<tr>
<td>A all, E all, T all</td>
<td>97.0</td>
<td>91.2</td>
</tr>
<tr>
<td>E all, T all</td>
<td>94.8</td>
<td>90.4</td>
</tr>
</tbody>
</table>
Figure 4.4. Average accuracy and standard deviation per model in the experiments with the three lowest accuracies (left) and the three highest accuracies (right) in ablation study A. The average accuracy of all the models in the experiment is on the right from the average accuracies of the models. (Publication II)

4.3 Fault diagnosis accuracy of models with network deconvolution

The research in Publication III pursued an improvement to the layer normalisation of 1D CNN-based fault diagnosis models. This results section presents the accuracies of the three 1D CNN models in the three different fault diagnosis tasks. All the models were tested in their conventional configuration, i.e., they include BN. Models with BN are referred to as the baseline models in this section. All the models were also tested with ND. This section presents the accuracy of each baseline model and model with ND in all the experiments.
Figure 4.5. Average and standard deviation of accuracy in repeated bearing fault diagnosis experiments with the CWRU dataset. The input data included vibration acquired with the drive-end sensor. Red bars show baseline model accuracy, and blue bars show modified model accuracy. (Publication III)

Figure 4.5 shows the model performances in the bearing fault diagnosis tests with load domain shift between the training and test data. These experiments were conducted with the data acquired with the drive-end accelerometer. On average, ND improved WDCNN and SRDCNN performance. WDCNN with ND was more than 5 percentage points more accurate on average than WDCNN with BN. Similarly, SRDCNN with ND was more than 8 percentage points more accurate than the corresponding baseline. However, the accuracy of Ince’s model with ND was lower and within one standard deviation from the average of the baseline.

Figure 4.6 shows the model performances in similar bearing fault tests, as in Figure 4.5. However, the input data in these experiments included vibration acquired with both the fan-end and the drive-end accelerometers. On average, all models with ND performed significantly better than the baseline models. The difference between the WDCNN models was over 11 percentage points. Between the SRDCNN models the difference was over 14 percentage points, and between Ince’s models the difference was over 2 percentage points. The models with ND achieved mostly significantly higher accuracy than the baseline models in the tests including training data sampled with a 3 hp load.

Figure 4.7 shows the accuracies in the azimuth thruster fault detection experiments conducted with vibration data acquired in real operating conditions. The figure shows the accuracy of the models in two experiments:
### Results

#### Figure 4.6. Average and standard deviation of accuracy in repeated bearing fault diagnosis experiments with the CWRU dataset. The input data included vibration acquired with the drive-end and fan-end sensors. Red bars show baseline model accuracy, and blue bars show modified model accuracy. (Publication III)

#### Figure 4.7. Average and standard deviation of accuracy in repeated azimuth thruster anomaly detection experiments. Red bars show baseline model accuracy, and blue bars show modified model accuracy. (Publication III)
bearing fault detection and gear fault detection. All the models detected the faults in both experiments with an accuracy of between 84\% and 91\%. The differences between corresponding models with ND and with BN were small, according to the average and the standard deviation of the accuracy.
5. Discussion

Deep learning models have been investigated for condition monitoring purposes for only relatively few years. One of the earliest significant works related to deep learning applications for condition monitoring was published in 2016 [38]. Currently, numerous works have focused on deep learning-based condition monitoring applications and have proposed sophisticated algorithms. Moreover, the most recent works have focused on solving the problem related to lack of data. Transfer learning and few-shot learning seem to be promising techniques for increasing the reliability of deep learning models especially in fault diagnosis tasks. However, other important practical problems regarding the application of deep learning for vibration-based condition monitoring seem to be relatively unaddressed. For example, vibration data of rotating machinery is greatly affected by the operating conditions, inaccuracies related to manufacturing and assembly, and the measurement instrumentation. Simultaneously, deep learning models are sensitive to the training data distributions; that is, they may not function reliably in condition monitoring applications when the vibration patterns change. These preconditions can limit the deployment of any sophisticated deep learning algorithm for condition monitoring tasks. To this end, techniques for training deep learning models for the vibration-based condition monitoring of rotating systems were studied. Publication I investigated the importance of the sufficient inclusion of all operating conditions in the training data. Publication II studied the suitability of torsional vibration data for gear fault diagnosis purposes, and compared torsional vibration to lateral vibration. Publication III studied 1D CNN layer normalisation techniques for improving the features of fault diagnosis models. The remainder of this chapter first discusses the results, and then two independent sections are dedicated to the impact and possible research paths in the future. Section 5.1 focuses on the scientific and practical impact of this research. Section 5.2 concerns possible future research paths.

The results regarding the training procedure proposed in Publication I indicate that deep learning models can learn system dynamics over a
large span of operating conditions. That is, the results shown in Figure 4.2 demonstrate the capability of the model to interpolate in the operating condition space. However, the results also demonstrate the limitations of such data-driven approaches. The trained virtual sensor could estimate the vibration accurately only near the operating conditions that were present in the training data. More specifically, the model could not estimate the vibration levels accurately at rotating speeds that were distinct from the rotating speeds in the training data. This limitation relates to the poor capability of the model to extrapolate in the operating condition space. Figure 4.3 shows that the estimates are more inaccurate when the rotating speed increases further away from the rotating speed range used in the training.

The results in Publication II yielded several key findings. For example, torsional vibration seems to be a source of data for deep learning-based gear fault diagnosis purposes with high potential. The experiments with data acquired with the torque transducers in particular resulted in high accuracies, which, most often, were the highest accuracies in the experiment groups in each ablation study. Furthermore, the models were not very sensitive to the longer transmission path if the data was acquired with the torque transducers. The lateral vibration data acquired with the accelerometers also resulted in high accuracies as well. However, the results showed that the models were more sensitive to the transmission path of the lateral vibration. The data acquired with the accelerometer which was mounted the furthest from the faulty gear, resulted in slightly lower average accuracies than the data acquired with the other accelerometers mounted on the gear box. Some experiments with the angular acceleration signals resulted in high accuracies. However, the models were very sensitive to the transmission path of the torsional vibration if the data was acquired with the rotary encoders.

Unfortunately, the reasons for the poor performance of some of the encoder experiments were not investigated. However, likely explanations were suggested. First, the accuracy of the models seemed to correlate with the rotating speed of the encoder. Higher rotating speeds resulted in lower accuracies on average. This decrease in the average accuracy in the experiments with data from faster rotating encoders could be caused by increased noise levels in the angular acceleration signals. The rotary encoders sampled the angular displacement of the shaft at 3 kHz. The angular displacement signals acquired with encoders at higher rotating speeds could include more initial high-frequency noise. The acquired angular displacement signals were then differentiated twice to angular acceleration. The differentiation may increase the noise in these signals, which hinders the performance of the models. Another likely explanation relates to the location of the poorly performing encoders. Encoders 1, 2 and 3 were mounted in close proximity to the upper gear, which may introduce
some confusing vibration patterns in the torsional vibration data. However, the patterns did not seem to affect the performance of the models in the experiments conducted with the upper torque transducer.

Another important finding relates to the performance of the 1D CNN-based models on vibration data acquired simultaneously with multiple sensors. Convolutional neural networks compute features that may be based on many adjacent vibration signals. In essence, they perform implicit sensor fusion. The results in this study showed advantages and disadvantages related to this capability. The models reached the best results with data acquired with all the accelerometers and torque transducers. In addition, the lowest accuracies between the experiment groups increased with the number of sensors. The disadvantage of this sensor fusion capability seems to be the capability to learn confusing features. The models were not accurate in most of the experiments conducted with the data acquired with one encoder. Therefore, the vibration patterns in some of the angular acceleration signals were likely confusing. Unfortunately, the models could not learn to neglect these confusing signals, even in the experiments where more clear signals were provided. For example, the average accuracy in experiment T all, with data from both torque transducers, in study D was 98.7%. In the same ablation study, the experiment E all, T all, with the same data combined with all encoder data, resulted in mere 90.4% accuracy.

Publication III shows promising results regarding the network deconvolution. The models with network deconvolution were compared to the baseline models in three different experiments conducted with the CWRU dataset and azimuth thruster dataset. In the two experiments with CWRU data, the models with ND outperformed the baseline models on average, with one exception. Ince’s model with ND resulted in a slightly lower average accuracy than Ince’s model with BN, as Figure 4.5 shows. However, these two accuracies were within one standard deviation from each other. The most significant increase in the average accuracy in these CWRU experiments relates to WDCNN. ND improved the average accuracy of WDCNN over 5 percentage points in the experiments with the drive-end data, and over 11 percentage points in the experiments with the drive-end and fan-end data, as shown in Figures 4.5 and 4.6, respectively.

The results related to the anomaly detection experiments with the azimuth thruster data did not show similar improvements to the accuracy of the models. The models with ND detected the gear faults and the bearing faults with similar accuracies with the models with BN, as shown in Figure 4.7. That is, the differences between the accuracies of the improved models were within the standard deviation of the baseline models, and thus, the differences were not statistically significant. Therefore, it can be concluded that both normalisation techniques were equally good.

The average accuracies in the experiments with the CWRU data can
be compared to other published results with the same data. The results regarding the CWRU load domain shifts differ slightly from the other published results. For example, the average accuracy of WDCNN in the experiment with load domain shift from 1 hp to 3 hp in Figure 4.5 was 99.9%. In contrast, the same experiment conducted in the original publication proposing WDCNN resulted in an accuracy of 91.0%, and in [90], the reported accuracy for the same experiment was 88.6%. The cause for such differences is difficult to analyse, since many works have not published their software. Furthermore, the reported accuracy can change significantly because of the differences between the used hyperparameters and training algorithm. Moreover, it may be possible that some works have not reported the average accuracy over a suitable number of experiments.

A problem regarding the azimuth thruster data was noticed during the research. The problem affects the interpretation of the results. Both the bearing faults and gear faults occurred in the azimuth thrusters close in time, and many vibration samples were influenced by both faults. This may hinder the reliability of the results. That is, it may be difficult to determine whether or not the models actually learned to detect faults in the bearing and in the gear. However, based on the results, it can be considered that the models learned to detect anomalies quite reliably, at least.

An important aspect to consider regarding the results in the experiments in this dissertation relates to the generalisability of the results. All experiments conducted with a laboratory test rig with specific instrumentation may not be repeatable outside the laboratory. For example, the trained models may function sufficiently with the specific test rig used to acquire the training data. However, even disassembling, and then reassembling the test rig may influence the vibration behaviour of the machine so that the trained deep learning model becomes unreliable. Furthermore, the mounting of the sensors and the used sensors may influence the results, especially in Publication II. For example, the orientation of the accelerometers was not optimised to acquire the best possible vibration patterns from the lower gear. Furthermore, if the sensors were changed to other torque transducers, encoders or accelerometers, the results could be different. Fortunately, most of the experiments in the publications in this dissertation resulted in the satisfying performance of the developed models. Therefore, the results can be considered to be appropriate and relevant considering the research aims of each publication.

5.1 Scientific and practical impact

The studies in this dissertation addressed general challenges related to the application of deep learning for vibration-based condition monitoring. Overall, the findings concerning the advantages and disadvantages of
Discussion

Deep learning applications for vibration analysis are relevant for condition monitoring of rotating systems. This section discusses the scientific and practical implications of the results.

Publication I demonstrated a novel virtual sensor application based on LSTM. Similar studies where an RNN is optimised to estimate the signals measured with a sensor from another sensor are rare. More specifically, there are not many studies demonstrating similar results where an RNN estimates time-series signals measured in the kHz regime. Some studies use LSTM in a sequence-to-value configuration where the model estimates one value from a longer vibration sample. In addition, RNNs have been used in the encoder-decoder configuration for sequence-to-sequence tasks. However, in this approach the input signals are not computed from the output signals in a deterministic order, and thus, the approach should not be compared with the experiments in Publication I. The LSTM-based virtual sensor in Publication I estimated time-series signals in a deterministic order in sequence-to-sequence configuration without the encoder-decoder architecture. Moreover, the virtual sensor learned the system dynamics between the sensors of the test rig and could estimate the lateral displacement of the rotor one time step at a time. The deterministic time-series estimation enables the virtual sensor to compute the desired quantity continuously. In contrast, the encoder-decoder configurations and the sequence-to-value configuration are not as suitable for continuous estimation.

Publication I also experimentally demonstrated the limitations of deep learning models when the vibration patterns change due to the operating conditions. The poor extrapolation capability of the data-driven virtual sensor indicates that the accuracy of deep learning models outside the operating conditions in the training data is limited. Fortunately, the LSTM-based model was accurate with largely distributed training data spanning all operating conditions. Generalisation over the whole operating condition space is a promising result considering future deep learning applications. The deep learning models seem to be able to learn complex features that can change appropriately based on the operating conditions. However, all vibration-related deep learning studies should take the limitation into account in the future. A model that has been trained and tested with vibration data from a narrow range of operating speeds has likely overfitted, despite the test performance of the model.

The practical implications of the results regarding the virtual sensor application are promising. Such a data-driven virtual sensor could be used to estimate signals that are difficult to measure. For example, the displacement of the centre-point of a paper machine roll was accurately estimated in the study. Installing the sophisticated multi-probe roundness instruments for the centre-point measurements can be difficult because of geometric constraints and high costs. Moreover, the data-driven virtual
sensors are not limited to this application. This approach could be used widely in industrial rotor systems as an alternative method for direct measurement and model-based state estimation. In addition, data-driven virtual sensors could be used as back-up sensors during primary sensor malfunctions.

Publication II concerns torsional vibration, which has been largely neglected in related works. Most of the related studies have used lateral vibration data acquired with accelerometers. The results showed that torsional vibration is an effective source for gear fault diagnosis data. In addition, the 1D CNN models were shown to be accurate despite the very versatile operating speeds and drive-train assembly. Similar experimental validation of deep learning-based fault diagnosis algorithms over a vast range of operating conditions are rare. Furthermore, many fault diagnosis studies have employed simple test rigs with one or very few sensors. Finally, the sensor fusion capabilities can be considered scientifically and practically relevant. Combining lateral and torsional vibration in the input data resulted in highly accurate models. Likely, the models learned better features by combining lateral and torsional vibration. Therefore, in future studies and applications of CNNs, the data acquisition system should include sensors for both lateral and torsional vibration. Including both types of sensors likely increases the fault diagnosis accuracy.

Publication III addressed the feature normalisation techniques between 1D CNN layers. Feature normalisation is commonly used in CNN-based fault diagnosis models. Most studies use batch normalisation, which is crucial for the optimisation and performance of the model. However, few works have discussed the feature normalisation of fault diagnosis models. The feature normalisation technique experimented with in Publication III theoretically provided a better optimisation landscape for the 1D CNN models. Experimentally, the results seemed promising. Many results showed that models with network deconvolution were more accurate than models with batch normalisation. Therefore, the results in Publication III align with previous results regarding network deconvolution [118]. That is, network deconvolution was demonstrated to perform well in the experiments with 2D CNNs and image data in the original article proposing the technique. Considering the previously published results, it is also important to underline the contribution of Publication III. Publication III reports additional results regarding the performance of network deconvolution with 1D CNNs and time-series data. Finally, the practical impact of the results in Publication III can be considered to be significant. Although, batch normalisation is a good feature normalisation technique and network deconvolution was adapted only slightly modified for 1D CNNs, this feature normalisation technique can be included in any CNN-based fault diagnosis model.
5.2 Future research

Deep learning-based condition monitoring seems to be an actively studied field. The number of studies published seems to grow year-on-year [17]. Moreover, deep learning-related research frequently provides new breakthrough techniques. These breakthrough techniques can often be studied and applied in condition monitoring. This section briefly discusses several problems and possible solutions related to deep learning-based condition monitoring.

Likely, the most difficult problem related to deep learning-based condition monitoring is the lack of data. Insufficient data in particular hinders the optimisation of these algorithms for fault diagnosis tasks. Vibration data including different faults is difficult to acquire. Moreover, the data is heavily affected by the operating conditions of the machine and the transmission path between the fault and the sensor. Furthermore, faults may occur at many locations at different severity. Therefore, a reliable deep learning-based fault diagnosis model would require a significant amount of training data for supervised learning. Thus, optimising deep learning-based models for fault diagnosis with the traditional supervised learning methods seems challenging.

Fortunately, there are many promising techniques that may solve this data sparsity problem. A number of works have shown that transfer learning provides some improvements to the performance of deep learning-based fault diagnosis models [90]. However, many reported results still leave room for improvement. Another promising technique is few-shot learning. Few-shot learning models can learn new classes from very few samples. Learning to diagnose faults from one example seems very promising. However, the limitations of these models regarding variations to vibration patterns due to operating conditions and realisations of faults could be investigated more. Currently, the work towards these problems has been started in the Mechatronics research group in Aalto University, and the author is involved [33].

In addition to algorithms related to the model architecture and optimisation, more data could be artificially generated. Data can be artificially generated with simulations and generative deep learning models. Currently, rotor dynamics are well-known and sophisticated simulations of rotating machines can be designed. With such simulations, vibration data could be generated without breaking any components. However, simulation and reality tend to deviate. That is, simulations often include ideal dynamics that may result in differences between the simulated vibration and the measured vibration. Therefore, more work related to the optimisation of deep learning models with simulated data is likely necessary. Currently, related work is ongoing in the Mechatronics research group in Aalto University. A promising technique concerning the randomisation
Discussion

of excitations and model parameters, such as masses and stiffnesses in simulations has been studied, and the author is involved [119]. In addition to simulations, generative models have been studied intensively during the recent years. Many studies proposing different generative adversarial networks (GAN) have been published. Additionally, state-of-the-art generative models based on variational autoencoders (VAE) can generate images from text prompts. For example, the image on the cover of this dissertation was generated with a model based on VAE from a prompt: "ship fleet condition monitoring, artificial intelligence, futuristic systems". Thus, there is likely room for more studies related to vibration data generation with deep generative models.
6. Conclusions

In this research, deep learning-based applications for condition monitoring purposes were developed and investigated. Limitations and advantages related to the application of deep learning for vibration-based monitoring were both discovered and quantified. Suggestions for the sources of vibration data and recommendations for the acquisition of data were given considering the reliability of the optimised models in both data-driven virtual sensor and fault diagnosis functions.

Constant monitoring of rotor systems decreases the likelihood of surprising malfunctions and product deficits by allowing timely maintenance actions. Constant monitoring can be difficult to arrange for a number of reasons. Occasionally the machine geometry, processed materials or frequent sensor malfunctions can prevent constant monitoring. Virtual sensors can enable the monitoring of a machine when primary measurements with a physical sensor is prevented. Virtual sensors based on deep learning can be optimised to estimate the primary measurements from certain secondary measurements with historical data. In this work, a LSTM-based virtual sensor and a suitable optimisation procedure were developed using experimental data.

Gear failures can excite torsional and lateral vibration in rotating systems. Analysing torsional vibration to detect and diagnose gear faults can provide an advantage over lateral vibration analysis. The vibration patterns excited by the gear faults are always transmitted through the components to the sensors. Measured lateral vibration is often affected by damping and dynamics of the non-rotating parts, such as bearings, gears, and bearing and gear housings, and other foundations. Such interference can hinder the performance of deep learning-based fault diagnosis models. To this end, torsional vibration data acquired with torque transducers and rotary encoders was compared to lateral vibration data acquired with piezoelectric accelerometers in extensive ablation experiments concerning gear fault diagnosis tasks. In addition, the sensor fusion capabilities of 1D CNNs were demonstrated to be effective and limitations related to confusing features were discussed.
Most CNN-based fault diagnosis models rely on batch normalisation for the centring and scaling of the features. During optimisation, normalisation of the features is often crucial for feasible computation times and satisfying test performance. However, batch normalisation, which is the most widely used feature normalisation technique, includes a compromise in the algorithm because of computational constraints. Theoretically, whitening the features between the layers, which includes decorrelation of the activations, provides a better optimisation landscape because of the sparser representation of the features between the layers. However, few studies concerning fault diagnosis have considered whitening instead of batch normalisation. For this purpose, network deconvolution, an approximative and implicit feature whitening technique, was adapted for the layer normalisation of 1D CNN-based fault diagnosis models. Models with network deconvolution were compared to similar models with batch normalisation on a benchmark dataset including vibration samples of faulty bearings, and on a dataset including vibration samples of healthy and faulty azimuth thrusters in real operating conditions. The experiments revealed benefits in increased accuracy and limitations in additional hyperparameter tuning of the network deconvolution algorithm.

To summarise, deep learning models can be optimised for many condition monitoring tasks. With sufficient data, the developed CNN-based and LSTM-based algorithms can be deployed for fault diagnosis and virtual sensor tasks, respectively. The experimental results in this dissertation are relevant for any development project towards deep learning-based condition monitoring systems from the design stage of the data acquisition to the evaluation stage of the optimised model.


[114] Chuanjiang Li, Shaobo Li, Ansi Zhang, Qiang He, Zihao Liao, and Jianjun Hu. “Meta-learning for few-shot bearing fault diagnosis under complex working conditions”. In: Neurocomputing 439 (2021), pp. 197–211. ISSN: 0925-2312. DOI: https://doi.org/10.1016/j.neucom.2021.01.099.


Deep learning applications for condition monitoring of rotating systems

Jesse Miettinen