BSDSAP:

Entwicklung eines KI-gesteuerten autonomen Roboters für die Navigation in einer häuslichen Umgebung

Projektbericht

eingereicht bei

Dr. Martin Prause

AKAD

von

Justin Stange-Heiduk

Hengstrücken 132

37520 Osterode am Harz

Telefon: 015233817587

Studiengang: Data Science

4. Fachsemester

Matrikelnummer: 8149363

Datum: 09.09.2025

[Abbildungsverzeichnis iii](#_Toc202942117)

[1. Einleitung 1](#_Toc202942118)

[1.1 Motivation und Problemstellung 1](#_Toc202942119)

[1.2 Zielsetzung der Arbeit 1](#_Toc202942120)

[1.3 Aufbau der Arbeit 1](#_Toc202942121)

[2. Methodisches Vorgehen 1](#_Toc202942122)

[2.1 CRISP-DM als Strukturmodell 1](#_Toc202942123)

[2.2 Überblick über verwandte Arbeiten (Related Work) 1](#_Toc202942124)

[2.2.1 Deep Learning für PdM (Zhong & Wang) 1](#_Toc202942125)

[2.2.2 Survival Analysis in der Instandhaltung (Carpentier et al.) 1](#_Toc202942126)

[2.2.3 Weitere Ansätze im Vergleich (z. B. GNN, Transformer, klassische ML) 1](#_Toc202942127)

[3. Business Understanding 1](#_Toc202942128)

[3.1 Zielsetzungen aus Sicht von OEMs und Werkstätten 1](#_Toc202942129)

[3.2 Herausforderungen: Datenverfügbarkeit, Blackbox-Modelle, wirtschaftlicher Nutzen 1](#_Toc202942130)

[3.3 Der SCANIA Component X Datensatz im Forschungskontext 1](#_Toc202942131)

[4 Data Understanding 1](#_Toc202942132)

[4.1 Datenquellen und Struktur (Operation, Repair, Spezifikation) 1](#_Toc202942133)

[4.2 Merkmalsübersicht: Sensorik, Histogramme, Zielgrößen 1](#_Toc202942134)

[4.3 Explorative Analyse und Besonderheiten (Imbalance, Zensierung) 1](#_Toc202942135)

[5 Data Preparation 1](#_Toc202942136)

[5.1 Feature Engineering und Preprocessing für Bi-LSTM 1](#_Toc202942137)

[5.2 Datensatzaufbereitung für XGBoost + AFT (inkl. Zensierung) 1](#_Toc202942138)

[5.3 Trainings-, Validierungs- und Testsplits 1](#_Toc202942139)

[5.4 Umgang mit fehlenden Werten und Normalisierung 1](#_Toc202942140)

[6 Modellierung 2](#_Toc202942141)

[6.1 Auswahl der Modellierungsstrategien 2](#_Toc202942142)

[6.1.1 Regression mit Bi-LSTM + Attention 2](#_Toc202942143)

[6.1.2 Survival Analysis mit XGBoost + Accelerated Failure Time 2](#_Toc202942144)

[6.2 Implementierung: Frameworks, Hardware, Laufzeitumgebung 2](#_Toc202942145)

[6.3 Hyperparameter, Trainingsstrategie und Modellarchitekturen 2](#_Toc202942146)

[7 Evaluation 2](#_Toc202942147)

[7.1 Regressionsmetriken (MAE, RMSE) 2](#_Toc202942148)

[7.2 Survival-Metriken (Concordance Index, ggf. Brier Score) 2](#_Toc202942149)

[7.3 Explainability (XAI) 2](#_Toc202942150)

[7.3.1 SHAP für XGBoost 2](#_Toc202942151)

[7.3.2 Attention-Analyse für Bi-LSTM 2](#_Toc202942152)

[7.4 Vergleich und Interpretation der Modelle 2](#_Toc202942153)

[7.5 Grenzen der gewählten Ansätze 2](#_Toc202942154)

[8 Fazit und Ausblick 2](#_Toc202942155)

[8.1 Zusammenfassung der Ergebnisse 2](#_Toc202942156)

[8.2 Bewertung der Modellansätze 2](#_Toc202942157)

[8.3 Mögliche Weiterentwicklungen (z. B. GNN, Multitask, Transfer Learning) 2](#_Toc202942158)

[Anhang A.1 iv](#_Toc202942159)

[Literaturverzeichnis v](#_Toc202942160)

[Eidesstattliche Versicherung vi](#_Toc202942161)

# Abbildungsverzeichnis

**Es konnten keine Einträge für ein Abbildungsverzeichnis gefunden werden.**

# Abkürzungsverzeichnis

Predictive Maintenance PdM

Remaining Useful Life RUL

# 1. Einleitung

## 1.1 Motivation und Problemstellung

## 1.2 Zielsetzung der Arbeit

## 1.3 Aufbau der Arbeit

# 2. Business Understanding

Die Phase des Business Understanding stellt im CRISP-DM-Modell den Ausgangspunkt für datengetriebene Projekte dar. Ziel ist es, die fachlichen Anforderungen und Rahmenbedingungen so zu erfassen, dass die anschließende Datenanalyse nicht isoliert, sondern in einem klaren geschäftlichen und anwendungsorientierten Kontext erfolgt. (Vgl. Chapman, 2000, S.14) Für den vorliegenden Anwendungsfall, die Untersuchung von Predictive-Maintenance-Ansätzen anhand von Sensordaten aus Nutzfahrzeugen, bedeutet dies, dass sowohl die übergeordneten Unternehmensziele als auch die spezifischen betrieblichen Herausforderungen betrachtet werden müssen.

Im weiteren Verlauf des Kapitels werden daher zunächst die relevanten Geschäftsziele im Kontext der vorausschauenden Wartung beschrieben. Darauf folgt eine Situationsanalyse, in der die verfügbaren Ressourcen sowie bestehende Restriktionen dargestellt werden. Anschließend werden die daraus abgeleiteten datenanalytischen Zielsetzungen präzisiert. Den Abschluss bildet ein Projektplan, der die methodische Vorgehensweise sowie den geplanten Einsatz von Werkzeugen und Verfahren aufzeigt.

## 2.1 Geschäftsziele im Kontext von Predictive Maintenance

Die zunehmende Digitalisierung und der Übergang zur Industrie 4.0 führen in der Automobil- und Nutzfahrzeugindustrie zu einem grundlegenden Wandel. Zentrale Treiber sind die Verfügbarkeit großer Datenmengen, der Einsatz neuer Sensortechnologien sowie die Integration von Machine-Learning-Methoden in industrielle Prozesse. Ziel ist es, Produktions- und Instandhaltungsprozesse so zu gestalten, dass sie nicht nur reaktiv, sondern zunehmend vorausschauend erfolgen können. Predictive Maintenance (PdM) nimmt dabei eine Schlüsselrolle ein. (Vgl. Zonta et al., 2020, S. 1)

PdM beschreibt einen Ansatz, bei dem historische und aktuelle Daten, domänenspezifisches Wissen sowie Verfahren der Statistik und künstlichen Intelligenz kombiniert werden, um drohende Ausfälle von Maschinen und Komponenten im Voraus vorherzusagen. Im Unterschied zu reaktiven oder präventiven Wartungsstrategien ermöglicht PdM eine zustandsorientierte und zeitnahe Planung von Instandhaltungsmaßnahmen. Damit verbunden sind wesentliche geschäftliche Zielsetzungen: die Reduktion von Stillstandszeiten, die Senkung von Wartungskosten sowie eine Steigerung von Produktivität und Qualität in komplexen Produktions- und Logistikumgebungen. (Vgl. Zonta et al., 2020, S. 2)

Eine zentrale Herausforderung für die Umsetzung von PdM ist die eingeschränkte Verfügbarkeit realer Datensätze. Herstellerunternehmen veröffentlichen in der Regel keine Betriebsdaten, da diese sensible Informationen zu Ausfallraten und Sensorausstattung enthalten. In der Forschung wird deshalb häufig auf synthetische oder simulierte Daten zurückgegriffen, die jedoch die Komplexität realer Systeme nicht vollständig abbilden können. (Vgl. Kharazian et al., 2025, S. 1)

Mit der Bereitstellung des SCANIA Component X-Datensatzes steht erstmals ein umfangreicher realweltlicher Datensatz für die Forschung zur Verfügung, der multivariate Zeitreihen umfasst und den schrittweisen Verschleiß einer anonymisierten Komponente dokumentiert. Dieser Datensatz ermöglicht Untersuchungen in verschiedenen Anwendungsfeldern von PdM, darunter Klassifikation, Prognose der Restlebensdauer, Anomalieerkennung und Überlebenszeitanalyse. Die Veröffentlichung im Rahmen der *Industrial Challenge 2024* verdeutlicht zudem den engen Bezug zwischen wissenschaftlicher Forschung und industrieller Praxis.

Damit ergibt sich für die vorliegende Arbeit ein zentrales Geschäftsziel: die Untersuchung und kritische Bewertung datengetriebener Ansätze des Predictive Maintenance, um deren Potenzial für die Optimierung von Wartungsstrategien, die effizientere Ressourcennutzung und die Verbesserung der Betriebssicherheit aufzuzeigen. (Vgl. Kharazian et al., 2025, S. 2)

## 2.2 Herausforderungen: Datenverfügbarkeit, Blackbox-Modelle, wirtschaftlicher Nutzen

Ein zentrales Hindernis für Predictive Maintenance liegt in der eingeschränkten Verfügbarkeit industrieller Sensordaten. Herstellerunternehmen veröffentlichen in der Regel keine Betriebsdaten, da diese sensible Informationen über Ausfallraten und eingesetzte Sensoren enthalten (Vgl. Kharazian et al., 2025, S. 1). Daher greifen viele Forschungsarbeiten auf synthetische Datensätze zurück, die die Komplexität realer Szenarien nur unzureichend widerspiegeln. Zusätzlich treten in industriellen Umgebungen häufig fehlerhafte oder verrauschte Messungen auf, etwa durch ungünstige Umgebungsbedingungen oder Sensorausfälle, was die Prognosequalität weiter beeinträchtigen kann (Vgl. Nunes et al., 2023, S. 1). Mit der Veröffentlichung des SCANIA Component X-Datensatzes liegt jedoch erstmals ein umfangreicher realweltlicher Datensatz mit zeitlicher Dimension vor, der neue Forschungsmöglichkeiten eröffnet (Vgl. Kharazian et al., 2025, S. 2).  
  
Die Leistungsfähigkeit moderner PdM-Ansätze basiert häufig auf komplexen Verfahren wie Deep Learning, Graph Neural Networks oder hybriden Prognosemodellen (ebd.). Diese Modelle sind zwar in der Lage, hochdimensionale Zeitreihendaten effektiv auszuwerten, zeichnen sich jedoch durch eine eingeschränkte Interpretierbarkeit aus. In industriellen Kontexten stellt diese „Black-Box“-Problematik eine erhebliche Hürde dar, da transparente Entscheidungsgrundlagen für die Akzeptanz neuer Systeme essenziell sind. Darüber hinaus zeigen viele Prognosemodelle eine begrenzte Generalisierbarkeit, da sie oft auf spezifische Komponenten zugeschnitten sind und die Wechselwirkungen zwischen verschiedenen Bauteilen nicht berücksichtigen (Vgl. Nunes et al., 2023, S. 1). Zonta et al. (Vgl. 2020, S. 1f ) betont in diesem Zusammenhang die zunehmende Komplexität der Industrie 4.0, die durch IoT, Big Data und Cyber-Physical Systems gekennzeichnet ist und weitere methodische Anforderungen an PdM-Lösungen stellt.  
  
Die wirtschaftliche Tragfähigkeit ist ein weiterer entscheidender Faktor für die Implementierung von PdM. Wartungskosten können je nach Branche zwischen 15 % und 60 % der gesamten Produktionskosten ausmachen. Effektive PdM-Systeme versprechen, ungeplante Produktionsstopps zu vermeiden, die Lebensdauer von Maschinen zu verlängern und die Ressourcennutzung zu optimieren. Damit verbunden sind potenziell signifikante Einsparungen, die jedoch den Aufwand für Datenmanagement, Infrastruktur und Modellimplementierung übersteigen müssen, um in der Praxis akzeptiert zu werden. Hier liegt eine wesentliche Herausforderung darin, den Nutzen präzise zu quantifizieren und in Relation zu den erforderlichen Investitionen zu setzen. (Vgl. Zonta et al., 2020, S. 2)

## 2.3 Datenanalytische Zielsetzungen

Aus den in Abschnitt 2.1 skizzierten Geschäftsanforderungen und der in Abschnitt 2.2 beschriebenen Ausgangslage ergeben sich konkrete datenanalytische Zielsetzungen für den Einsatz von Predictive Maintenance im industriellen Kontext. Predictive-Maintenance-Projekte folgen dabei einem wohldefinierten Ablauf von der Problemerfassung über die Datenerhebung bis zur entscheidungsunterstützenden Nutzung der Ergebnisse; sie sind eingebettet in die industrie-4.0-typische Infrastruktur aus IoT, Big Data, KI und Cyber-Physical Systems, die Datenverfügbarkeit und Automatisierung maßgeblich prägen. (Vgl. Achouch et al., 2022, S. 1f)

Erstes zentrales Ziel ist die Vorhersage des zukünftigen Zustands industrieller Komponenten in Form von Restlebensdauer-Prognosen oder Ausfallzeitpunkt-Schätzungen. Dies umfasst insbesondere die datengetriebene Prognostik auf Basis multivariater Sensordaten, einschließlich RUL-Bestimmung und Verfahren der Überlebenszeitanalyse, um präventive Eingriffe zeitlich zu planen und ungeplante Stillstände zu vermeiden. Der SCANIA-Component-X-Datensatz adressiert diese Zielsetzung explizit, da seine zeitlich aufgelösten Messreihen graduelle Degradationsprozesse abbilden und damit Aufgaben wie RUL-Prognose, Ausfallrisiko-Modellierung und Klassifikation imminenter Ausfälle auf realweltlicher Datenbasis ermöglichen.

Zweites Ziel ist die explizite Kopplung der modellbasierten Vorhersagen an betriebswirtschaftliche Optimationskriterien. In der Literatur werden Kostenminimierung sowie Verfügbarkeits- und Zuverlässigkeitsmaximierung als primäre Optimationsziele benannt; beide Ziele stehen teils in Zielkonflikten, weshalb eine wohldefinierte Zielsetzung und Abwägung für das jeweilige Anwendungsszenario erforderlich ist. Die datenanalytischen Ergebnisse sollen deshalb in eine Entscheidungsunterstützung überführt werden, die Wartungsplanung und Ressourceneinsatz unter Kosten- und Verfügbarkeitsgesichtspunkten koordiniert und in geeignete Systemarchitekturen eingebettet ist. (Vgl. Zhu et al., 2025, S. 2)

Zusammenfassend zielen die Analysen darauf, auf realweltlicher Zeitreihenbasis verlässliche und ökonomisch nutzbare Vorhersagen zu liefern. Der SCANIA-Datensatz stellt dafür eine geeignete Grundlage bereit, da er die für Prognostik und Ausfallrisiko-Modellierung relevanten Eigenschaften realer Flottenbetriebsdaten aufweist und damit eine über rein synthetische Daten hinausgehende Bewertung datengetriebener PdM-Methoden ermöglicht.

## 2.4 Überblick über verwandte Arbeiten (Related Work)

# 3. Data Understanding

## 3.1 Beschreibung des SCANIA-Datensatzes

## 3.2 Explorative Datenanalyse

## 3.3 Datenqualität und Herausforderungen

# 4. Data Preparation

### 4.1 Datenintegration und Vorverarbeitung 4.2 Umgang mit fehlenden Werten 4.3 Transformationen 4.4 Feature Engineering 4.4.1 Sliding Windows für Zeitreihen 4.4.2 Extraktion statistischer Merkmale (z. B. mit tsfresh)

### 4.4.3 Feature Selection 4.4.4 Bildung von Labels für Survival Analysis (Event, Time)

# 5. Modellierung

## 5.1 Auswahl der Modellierungsstrategien

### 6.1.1 Random Survival Forest

### 6.1.2 Survival Analysis mit XGBoost + Accelerated Failure Time

## 6.2 Modellarchitektur und Implementierung 6.3 Trainingskonfiguration und Hyperparameter

# 7. Evaluation

## 7.2 Survival-Metriken

## 7.4 Kostenvergleich

## 7.5 Grenzen der gewählten Ansätze

# 8. Deployment

## 8.1 Zielgruppe und Anwendungsszenario 8.2 Visualisierung und Nutzung der Vorhersagen 8.3 Beispielhafte Nutzung: Werkstatt, Flottenmanagement, OEM 8.4 Grenzen und offene Punkte für echte Produktion

# 9. Fazit und Ausblick

## 9.1 Zusammenfassung der Ergebnisse

## 9.2 Bewertung der Modellansätze

## 9.3 Mögliche Weiterentwicklungen (z. B. GNN, Multitask, Transfer Learning)

# Anhang A.1

# Literaturverzeichnis

# Eidesstattliche Versicherung

Ich versichere, dass ich das beiliegende Assignment selbstständig verfasst, keine anderen als die angegebenen Quellen und Hilfsmittel benutzt sowie alle wörtlich oder sinngemäß übernommenen Stellen in der Arbeit gekennzeichnet habe.