



Universiteit  
Leiden  
The Netherlands

## Data-driven machine learning and optimization pipelines for real-world applications

Koch, M.

### Citation

Koch, M. (2020, September 1). *Data-driven machine learning and optimization pipelines for real-world applications*. Retrieved from <https://hdl.handle.net/1887/136270>

Version: Publisher's Version

License: [Licence agreement concerning inclusion of doctoral thesis in the Institutional Repository of the University of Leiden](#)

Downloaded from: <https://hdl.handle.net/1887/136270>

**Note:** To cite this publication please use the final published version (if applicable).

Cover Page



Universiteit Leiden



The handle <http://hdl.handle.net/1887/136270> holds various files of this Leiden University dissertation.

**Author:** Koch, M.

**Title:** Data-driven machine learning and optimization pipelines for real-world applications

**Issue Date:** 2020-09-01

# Data-Driven Machine Learning and Optimization Pipelines for Real-World Applications

**Proefschrift**

ter verkrijging van  
de graad van Doctor aan de Universiteit Leiden,  
op gezag van Rector Magnificus prof.mr. C.J.J.M. Stolker,  
volgens besluit van het College voor Promoties  
te verdedigen op dinsdag 1 september 2020  
klokke 10.00 uur

door

**Milan Koch**

geboren te Ostercappeln, Duitsland  
in 1990

## Promotiecommissie

Promotor:	Prof. Dr. T.H.W. Bäck	
Co-promotor:	Dr. H. Wang	
Overige leden:	Prof. Dr. A. Laat	(voorzitter)
	Prof. Dr. S. Mostaghim	(University of Magdeburg, DE)
	Prof. Dr. F. Duddeck	(Technical University of Munich, DE)
	Dr. A. Kononova	
	Prof. Dr. H. Hoos	
	Prof. Dr. M. Bonsangue	(secretaris)

Copyright © 2020 Milan Koch.

Figures and diagrams are generated using PGF/TIKZ, INKSCAPE, MATPLOTLIB and NETRON.

---

---

# Contents

## List of Symbols

<b>1</b>	<b>Introduction</b>	<b>1</b>
1.1	Background . . . . .	1
1.2	Objectives . . . . .	2
1.3	Outline of the Dissertation . . . . .	3
<b>2</b>	<b>Automated Damage Assessment</b>	<b>7</b>
2.1	Objectives . . . . .	7
2.2	Data . . . . .	10
<b>3</b>	<b>Methods</b>	<b>13</b>
3.1	Machine Learning . . . . .	13
3.2	Time Series Classification . . . . .	17
3.3	Decision Trees and Random Forests . . . . .	18
3.4	Feature Engineering . . . . .	20
3.5	Hyperparameter Optimization . . . . .	25
3.6	Neural Networks . . . . .	25
3.7	Evolutionary Algorithms and Genetic Programming . . . . .	28
3.8	Automated Machine Learning . . . . .	29
<b>4</b>	<b>Multivariate Time Series Classification</b>	<b>33</b>
4.1	Related Work . . . . .	33
4.2	Overview of Approaches for MTSC . . . . .	35
4.3	General Approach and Plain Hand-Crafted Pipeline . . . . .	35
4.4	Hand-Crafted Approach with Genetic Programming . . . . .	42
4.5	AutoML . . . . .	45
4.6	Neural Network Architectures . . . . .	47

<b>5 Solving Real-World Problems</b>	<b>51</b>
5.1 Case A: Predicting the Vehicle Crash Impact Point . . . . .	51
5.2 Case B: Predicting the Damaged Parts of a Vehicle Crash . . . . .	61
5.3 Case C: Automated Machine Learning for PD Patients . . . . .	71
5.4 Discussion of Boruta Algorithm . . . . .	86
 <b>6 Advanced and Optimized Pipelines</b>	 <b>87</b>
6.1 Data Sets . . . . .	87
6.2 Performance Evaluation Measure . . . . .	89
6.3 Experimental Settings . . . . .	91
6.4 Results and Analysis . . . . .	91
6.5 Conclusions and Outlook . . . . .	93
 <b>7 Data-Driven Services in the Car Industry</b>	 <b>97</b>
7.1 Introduction . . . . .	98
7.2 Related Work . . . . .	99
7.3 A Data-driven Service for Crash Damage Prediction . . . . .	100
7.4 Methodology towards Data-Driven Services . . . . .	101
7.5 Conclusions and Outlook . . . . .	111
 <b>8 Conclusions and Outlooks</b>	 <b>113</b>
8.1 Conclusions . . . . .	113
8.2 Outlook . . . . .	114
 <b>Appendix A Decision Tree of Case A</b>	 <b>117</b>
 <b>Bibliography</b>	 <b>119</b>
 <b>English Summary</b>	 <b>135</b>
 <b>Samenvatting</b>	 <b>137</b>
 <b>About the Author</b>	 <b>139</b>

---

## List of Symbols

$\mathbf{x}$	Data points
$N$	Number of data points
$\mathbf{X}$	Time series
$L$	Length of time series
$d$	Dimension of time series
$\Phi$	Features extracted from time series
$\mathcal{F}$	Feature function
$\mathcal{A}$	Machine learning algorithm
$\Theta$	Configuration space
$\mathcal{H}$	Algorithm configurator
$f$	Performance metric
$N_{cv}$	Number of folds in cross-validation
$Y$	True target label
$\hat{Y}$	Predicted target label
$\mathcal{C}$	Classes
$q$	Number of classes
$\mathbf{S}$	EEG measurement
$w$	Weight
<b>CDF</b>	Cumulative distribution function

## LIST OF SYMBOLS

---

$\alpha$	Significance level
$\Omega$	Parse tree
$\psi$	Genetic operator
$\zeta$	Arithmetic operator
$p(\cdot)$	Probability function
$\mathbb{E}$	Expectation
$\Pr(\cdot \cdot)$	Probability distribution
$f(\cdot)$	Function
$G$	Gini impurity
$E$	Entropy
$b$	Bias