

Informationsextraktion aus Gebäudeenergiesystemen

Information Extraction from Building Energy Systems

Von der Fakultät für Maschinenwesen der Rheinisch-Westfälischen Technischen Hochschule Aachen zur Erlangung des akademischen Grades eines Doktors der Ingenieurwissenschaften genehmigte Dissertation

vorgelegt von

Florian Stinner

Berichter: Univ.-Prof. Dr.-Ing. Dirk Müller

Prof. Dr.-Ing. habil. Jörn Plönnigs

Tag der mündlichen Prüfung: 12.12.2025

Diese Dissertation ist auf den Internetseiten der Universitätsbibliothek online verfügbar.

Florian Stinner

Information Extraction from Building Energy Systems

Kurzfassung

Nichtwohngebäude sind ein großer Emittent von CO₂-Emissionen. Die Verbesserung der energetischen Nutzung erfolgt im Allgemeinen durch eine personalintensive Analyse des Betriebes, wobei moderne Regelungs- und Fehleranalysemethoden helfen können.

Für diese Methoden existiert in Automationssystemen in der Regel kein direkt maschinell auswertbares passendes und erforderliches Metadatenmodell. Die Metadaten von Datenströmen sind jedoch in je Betreiber individuellen Anlagenkennzeichnungsschlüsseln codiert, weshalb die automatische Übersetzung nur umständlich möglich ist. Ergänzend zu den Zeichenketten der Schlüssel und zusätzlichen Beschreibungen liefern Zeitreihen wertvolle Informationen. Bei der Übersetzung der Metadaten können Algorithmen auf Basis von künstlicher Intelligenz helfen.

In dieser Dissertation wird anhand eines Informationsextraktionsprozesses gezeigt, wie aus Rohdaten (Zeichenketten und Zeitreihen) Informationen über die Metadaten extrahiert werden können. Ziel des Prozesses ist es, acht verschiedene Anwendungen in Energiesystemen mit den erforderlichen Metadaten zu unterstützen. Zuerst werden die Metadaten aus den Rohdaten extrahiert und darauf folgend für die jeweiligen Anwendungen geclustert, um z.B. die Daten von einem Raum zusammenzufassen. Es werden verschiedene Qualitätskriterien der Informationen ausgewertet, die auch eine Bewertung der anwendungsrelevanten Metadaten umfasst.

Die Metadatenextraktion erzielte anwendungsbezogen eine Informationsqualität (F1 Wert) von bis zu 100 % und skalierbar einsetzbar von bis zu 94 %. Das auf der Metadatenextraktion folgende Clustern von Datenströmen zu Anwendungen konnte bis zu 100 % Genauigkeit erreichen, so dass teilweise die erforderlichen Datenströme für Anwendungen wie Raumautomationen vollständig zugeordnet werden konnten. Wenn die Metadatenextraktion und das Clustern als gesamter Informationsextraktionsprozess betrachtet werden, so erreichte jede Anwendung mit mindestens einem Verfahren eine Gesamtbewertung zwischen 88 % und 98 % richtig zugeordneten Metadaten. Die Analyse ergab, dass die Komplexität, wie die Anzahl der Eingangsdaten und der jedem Datenstrom zugeordneten Metadaten, nur in einem geringen Maße die Ergebnisse der Algorithmen beeinflusst.

Insgesamt konnte somit ein Informationsextraktionsprozess zur Unterstützung von Energieeinsparanwendungen etabliert werden. Der Großteil der algorithmischen Ergebnisse benötigt so wenig zusätzliche Eingabewerte von Nutzern, so dass diese in reale Prozessen in Energiesystemen integriert werden könnten, jedoch können diese zusätzlichen Eingabewerte durch die Kombination von Informationen weiter gesenkt werden. Diese Arbeit liefert einen Einblick, worauf in zukünftiger Entwicklung bei Algorithmen und Integration dieser in Unternehmensabläufe geachtet werden kann.

Abstract

Non-residential buildings are a major source of CO₂ emissions. Improvements in energy use are generally achieved through labor-intensive analysis of operations, with modern control and fault analysis methods providing support.

Automation systems generally do not have a directly machine-readable metadata model that is suitable and necessary for these methods. However, the metadata of data streams is encoded in individual plant identification keys for each operator, which makes automatic translation difficult. In addition to the character strings of the keys and additional descriptions, time series provide valuable information. Artificial intelligence-based algorithms can help with the translation of metadata.

This dissertation uses an information extraction process to show how information about metadata can be extracted from raw data (character strings and time series). The aim of the process is to support eight different applications in energy systems with the necessary metadata. First, the metadata is extracted from the raw data and then clustered for the respective applications, e.g., to aggregate the data from a room. Various information quality criteria are evaluated, which also includes an assessment of the application-relevant metadata.

Metadata extraction achieved an application-specific information quality (F1 value) of up to 100 % and was scalable for use up to 94 %. The clustering of data streams for applications following metadata extraction achieved up to 100 % accuracy, meaning that in some cases the data streams required for applications such as room automation could be completely assigned. When metadata extraction and clustering are considered as an overall information extraction process, each application achieved an overall rating of between 88 % and 98 % correctly assigned metadata using at least one method. The analysis showed that complexity, such as the number of input data and the metadata assigned to each data stream, only had a minor influence on the results of the algorithms.

Overall, an information extraction process was established to support energy-saving applications. The majority of the algorithmic results require so few additional input values from users that they could be integrated into real processes in energy systems. However, these additional input values can be further reduced by combining information. This work provides insight into what can be considered in future development of algorithms and their integration into business processes.

Inhaltsverzeichnis

Nomenklatur	IX
Abbildungsverzeichnis	XIII
Tabellenverzeichnis	XVII
1 Motivation	1
2 Stand der Wissenschaft und Technik	4
2.1 Informationsextraktion in Gebäudeautomationssystemen	4
2.1.1 Definition von genutzten Datenquellen und Algorithmen	5
2.1.2 Extraktion von Metadaten	6
2.2 Metadaten im Gebäudeautomationssystem	15
2.2.1 Industry Foundation Classes (IFC)	15
2.2.2 Virtual Buildings Information System (VBIS)	16
2.2.3 Brick Schema	16
2.2.4 Project Haystack	17
2.2.5 BUDO Schema	17
2.2.6 Weitere Schemata/Bewertung	18
2.3 Algorithmen	18
2.3.1 Klassische Merkmalsextraktion	18
2.3.2 Fortschrittliche Merkmalsextraktion	19
2.3.3 Klassifizierungsalgorithmen	21
2.3.4 Ähnlichkeit von Zeichenketten	21
2.4 Zusammenfassung Forschungsbedarf	23
3 Methodik	25
3.1 Informationsextraktion	26
3.1.1 Daten und Information	26
3.1.2 Datentypen	26
3.1.3 Datenqualität	27
3.1.4 Informationsqualität	29
3.1.5 Informationsextraktionsprozess	36
3.2 Auswahl eines Metadatenschema	39
3.2.1 Auswahlprozess Metadatenschema	39

3.2.2	Ergebnisse Bewertung Metadatenchema	40
3.2.3	Fazit Metadatenchema	41
3.3	Metadatenextraktion auf Basis von Zeichenketten	42
3.3.1	Klassifizierungsalgorithmen für Zeichenketten	42
3.3.2	Merkmalsextraktionsverfahren Zeichenketten	43
3.3.3	Blockweise Merkmalsextraktion	43
3.3.4	Gesamtheitliche Merkmalsextraktion	46
3.3.5	Komposita und Wortdatenbanken	47
3.3.6	Korrektur fehlerhafter Schreibweisen und Abkürzungen	49
3.3.7	Normalisierung Synonyme	51
3.3.8	Algorithmus zur Verwendung von Ähnlichkeitswerten	51
3.4	Zeitreihenklassifizierung	52
3.4.1	Datenvorprozessierung	52
3.4.2	Nutzung der Merkmalsextraktion	53
3.4.3	Einflüsse von Störgrößen, Nutzungsmustern	53
3.4.4	Klassifizierer	54
3.5	Datensatz	54
3.5.1	Mortar Datensatz	55
3.5.2	OOM4ABDO Datensatz	55
3.5.3	Vergleich von Datensätzen	56
3.5.4	Übersetzung für Ground Truth	59
3.6	Anwendungen in Energiesystemen	61
3.6.1	Fehler- und Anomalieerkennung	61
3.6.2	Regelung	64
3.6.3	Energiesystemanalyse	67
3.6.4	Zusammenfassung Anwendungen	69
4	Ergebnisse	69
4.1	Metadatenextraktion auf Basis von Zeichenketten	70
4.1.1	Auswertung Datensätze	71
4.1.2	Anwendungsbezogene Bewertung	86
4.1.3	Zusammenfassung von zeichenkettenbasierten Verfahren	86
4.2	Zeitreihenklassifizierung	87
4.2.1	Ergebnisse Mortar Datensatz	87
4.2.2	Ergebnisse Einflüsse von Nutzungsmustern	91
4.2.3	Ergebnisse OOM4ABDO Datensatz	91
4.3	Anwendungen in Gebäudeenergiesystemen	92
4.3.1	Fehler- und Anomalieerkennung	93
4.3.2	Regelung	96

4.3.3	Energiesystemanalyse	96
4.3.4	Zusammenfassung Anwendungen	99
4.4	Gesamtbewertung Verfahren	99
4.4.1	Bewertung Datenqualität	101
4.4.2	Bewertung Genauigkeit	101
4.4.3	Bewertung Vollständigkeit	101
4.4.4	Bewertung Ausgeglichenheit	102
4.4.5	Bewertung Konformität	102
4.4.6	Bewertung Zuverlässigkeit	103
4.4.7	Bewertung Generalisierbarkeit	104
4.4.8	Bewertung Operationalisierbarkeit	104
4.4.9	Bewertung Effektivität	104
4.4.10	Bewertung Gesamtverfahren von Rohdaten zu Anwendung	105
5	Diskussion	105
5.1	Einordnung Informationsqualitätskriterien	107
5.2	Einordnung Metaextraktionsalgorithmen	108
5.3	Einordnung Anwendungen	110
5.4	Einordnung Gesamtverfahren	112
6	Fazit und Ausblick	115
	Literaturverzeichnis	117
A	Informationsqualität	173
A.1	Learning Ability Index (LAI)	173
A.2	Herleitung Effektivität	174
B	Metadatenextraktionsverfahren in Energiesystemen	175
C	Sprachverarbeitung	175
C.1	Wortfilter	175
C.2	Korpora Statistiken	177
C.3	Synonyme	177
C.4	Komposita	178
C.4.1	Wortdatenbank	179
C.4.2	Tuggener	180
C.4.3	Repodiac	180
C.4.4	Stinner	181
C.5	Ergebnisse Komposita Trennung	182
C.5.1	Korrektur fehlerhaft geschriebener Wörter	183

C.5.2	Ergebnisse Korrektur fehlerhaft geschriebener Wörter	185
D	Merkmalsextraktion von Zeitreihen	185
D.1	TSFast	187
E	Statistiken	195
E.1	Statistische Merkmale Datensatz	195
E.2	Bewertung von statistischen Merkmalen	197
E.3	Ergebnis von statistischen Merkmalen als prädiktive Vorhersage von Ergebnissen . . .	198
E.4	Statistiken Anwendungen	200
E.4.1	RLT Anlagen	200
E.5	Anwendungen	213
E.5.1	Raumlufttechnische Anlage	214
E.5.2	Wärmeübertrager	218
E.5.3	Außentemperatursensor	218
E.5.4	PID Regelung der Raumtemperatur	219
E.5.5	Modellprädiktive Regelung	220
E.5.6	Raummonitoring	221
E.5.7	Energiefluss/Topologie	222
E.5.8	Quartiersmonitoring	222
E.6	Metadaten-Kennzeichnung	222
E.6.1	Gunning Fog Index	223
E.6.2	Automated readability index	223
E.6.3	Semantische Komplexität	223
E.6.4	Semantische Bewertung	224
E.6.5	Informationsentropie	225
E.7	Zusätzliche Ergebnisse	226
E.7.1	Vergleich Aikido mit RoBERTa	226
F	Verwendete Umgebung	227
F.1	Verwendete Umgebung Zeitreihenklassifizierung	228

Nomenklatur

Formelzeichen

Symbol	Bedeutung
$\cos(\alpha)$	Cosinus-Distanz
$L_{a,b}$	Damerau-Levenshtein Distanz
av	Durchschnitt (engl. average)
Bes	Datenstrombeschreibung
FN	Falsch Negativ
FP	Falsch Positiv
GBK	Generalisierbarkeit
Gen	Genauigkeit
ID	Identifikator
IQ	Informationsqualität
J	Jaro-Winkler-Distanz
KD	Kennzeichen-zu-Datenstrom-Verhältnis
KDD	Kennzeichen-zu-Datenstrom-Diversitäts-Verhältnis
KNN	künstliches neuronales Netz
KW	Kennzeichen-zu-Wort-Verhältnis
KWD	Kennzeichen-zu-Wort-Diversitäts-Verhältnis
LAI	Learning Ability Index
Max	Maximum
Min	Minimum
NaN	Ungültige Werte (engl. Not a Number)
nom	nominell
Name	Datenstromname
Rel	Relevanz
Rob	Robustheit
Sen	Sensitivität
SSC	Ähnlichkeitswerte (engl. Similiarity Score)
Vol	Volumenstrom

Fortsetzung auf der nächsten Seite

Formelzeichen

Symbol	Bedeutung
WD	Wort-zu-Datenstrom-Verhältnis
WDD	Wort-zu-Datenstrom-Diversitäts-Verhältnis
WN	Wahr Negativ
WP	Wahr Positiv
ZD	Anzahl-Zeichen-zu-Datenstrom-Verhältnis
ZK	Anzahl-Zeichen-zu-Kennzeichen-Verhältnis
ZW	Anzahl-Zeichen-zu-Wort-Verhältnis

Indizes und Abkürzungen

Symbol	Bedeutung
AKS	Anlagenkennzeichnungsschlüssel
BERT	Bidirectional Encoder Representations from Transformers
BHKW	Blockheizkraftwerke
BIM	Building Information Modelling
BoW	Bag of Words
BUDO	Buildings Unified Data Point Naming Schema for Operation Management
CatBoost	categorical boosting
CNN	Convolutional Neural Network
CRF	Conditional Random Field
DFIG	Data Fusion Information Group
DTW	Dynamic Time Warping
ECE	erwarteter Kalibrierungsfehler (engl. Expected Calibration Error), siehe Formel 3.7
F1	F-Maß mit $\alpha = 0$, siehe Formel 3.4
GAS	Gebäudeautomationssystem
GES	Gebäudeenergiesystem
HLK	Heizung, Lüftung, Klimatechnik
ID	Identifikator
IE	Informationsextraktion
IFC	Industry Foundation Classes
IKT	Informations- und Kommunikationstechnik
IQ	Informationsqualität
ISE	Institut für Solare Energiesysteme
JDL	Joint Directors of Laboratories
KI	Künstlicher Intelligenz
KNN	künstliches neuronales Netz
LCS	Longest Common Subsequence
LCSS	Longest Common Substring
LLM	Large Language Model
LSPV	längste Strecke aufeinanderfolgender Werte über Null (engl. longest stretch of positive values)
LSTM	Long short-term memory
MCC	Matthews correlation coefficient, siehe Formel 3.5
ME	Merkmalsextraktion
MIPV	Mittelwert des Index der positiven Werte (engl. mean of indices of positive values)

Symbol	Bedeutung
ML	Machine Learning
MLP	Multi-Layer Perceptron
MPC	modellprädiktive Regelung (engl. Model Predictive Control)
MPV	Mittelwert der positiven Werte (engl. mean of positive values)
MVTS	Multivariate Time Series Transformer
NLP	Natural Language Processing
NWG	Nichtwohngebäude
OOM4ABDO	Projekt: Objektorientiertes Monitoring als Grundlage für einen effizienteren Betrieb sowie kostengünstige Bestandsoptimierung durch Anwendung von Machine Learning-Techniken, BMWK FKZ 03SBE006A
OSA	Optimal String Alignment
PID	verstärkend (P), integrierend (I) und differenzierend (D) (engl. Proportional-integral-derivative)
PPV	Anteil positiver Werte (engl. proportion of positive values)
R&ID	Rohrleitungs- und Instrumentierungsdiagramme
RF	Random Forest Klassifizierer
RLT	raumluftechnisch, wie in „raumluftechnischen Anlagen“ (RLT-Anlagen)
RoBERTa	Robustly Optimized BERT
Rocket	RandOm Convolutional KErnel Transform
SdT	Stand der Technik
SVM	Support Vector Machines
TGA	technischen Gebäudeausrüstung
THG	Treibhausgas
TSFEL	Time Series Feature Extraction Library
TSFresh	Time Series FeatuRe Extraction on basis of Scalable Hypothesis tests
TVOC	Gesamtsumme der Emissionskonzentration (engl. Total volatile organic compounds)
UCR	University of California
UFZ	Helmholtz-Zentrum für Umweltforschung
VAV	Volumenstromregler (engl. variable air volume)
VBIS	Virtual Buildings Information System
XGBoost	eXtreme Gradient Boosting

Abbildungsverzeichnis

1.1	Ablauf der Dissertation anhand der verschiedenen Ebenen eines Informationsextraktionsprozesses. Es werden die verschiedenen Schritte in der Arbeit mit den jeweiligen Kapiteln aufgeführt. Insgesamt wird der Prozess von Daten über Informationen hin zum Wissen verdeutlicht.	3
2.1	Übersicht über die verwendeten Verfahren zur Metadatenextraktion. Mit grünen Haken versehene Verfahren existieren in einem weitreichenden Verfahren, benötigen aber weitere Untersuchungen. Mit gelben Strich versehene Verfahren existieren nur vereinzelt.	7
2.2	Struktur und Aufbau des BUDO Schemas	18
2.3	Das Long short-term memory mit seinen vier Gattern (Vergessengatter, Eingangsgatter, Zellenaktualisierung, Ausgangsgatter).	19
2.4	Aufbau der Merkmalsextraktion durch MultiRocket: Die Zeitreihen werden mithilfe einer Convolution Schicht in verschiedene Convolutional Kerne unterteilt (im Beispiel: A, B, C). Aus jedem der MultiRocket Convolutional Kerne werden vier Merkmale extrahiert: Anteil positiver Werte (PPV), Mittelwert der positiven Werte (MPV), Mittelwert des Index der positiven Werte (MIPV), längste Strecke aufeinanderfolgender Werte über Null (LSPV). Für das Beispiel B ist der PPV $\frac{3}{5}$, da drei von fünf Werten positiv sind, MPV ist $\frac{60+20+100}{3} = 60$, MIPV beträgt $\frac{0+1+2}{3} = 1$ und LSPV $0 \rightarrow 2 = 3$. Diese Merkmale können dann einem linearen Klassifizierer übergeben werden.	20
2.5	Die vier verwendeten phonetischen Verfahren aufgezeigt am Beispielwort „Temperature“.	23
3.1	Klassische Informationsqualitätskriterien der Klassifikation auf Basis des wahren und vorhergesagten Zustands.	31
3.2	Data Fusion Information Group (DFIG)-Modell zur Datenfusion im Kontext der Metadatenextraktion aus Gebäudeautomationssystem (GAS)-Daten	37
3.3	Ergebnisse der Untersuchung der unterschiedlichen Metadatenschemata (Brick Schema, VBIS). Einzelne bedeutet, dass jede im Datensatz vorkommende Vokabel nur einmal gezählt wurde. Gesamt bedeutet, dass die Vokabel mit ihrem Vorkommen gewichtet wurde.	41
3.4	Ein Vergleich der Zuweisungen von Kennzeichen zu Instanzen bei der blockweisen und gesamtheitlichen Merkmalsextraktion.	44

3.5	Abbildung der jeweiligen Merkmalssätze („Basis“ und „Alle“) mit Merkmalsextraktionsarten („Blockweise“ und „Gesamtheitlich“) zusammen mit den verwendeten Algorithmen („CatBoost“ und „Long short-term memory (LSTM)“). LSTM kann aufgrund seiner Eigenschaften, die auch die Reihenfolge der Blöcke des Gesamtnamens erfordert, nur „Gesamtheitlich“ durchgeführt werden.	44
3.6	Abbildungen der blockweisen und gesamtheitlichen Merkmalsextraktion. Hellgelb sind alle Merkmale vom Merkmalssatz „Basis“ und Dunkelgelb alle Merkmale vom Merkmalssatz „Alle“ markiert.	45
3.7	Ablauf der Detektion von Repodiac. Vom letzten Zeichen ausgehend wird jeweils überprüft, ob sich am Ende eines gefundenen Wortes ein neues Wort befindet.	50
3.8	Vergleich der Signal Typen von Mortar (blau) und OOM4ABDO Datensatz (rot). In roter Schrift sind die Klassen markiert, die gesondert untersucht wurden.	57
3.9	Vorgehen zur Erstellung der Ground Truth. Es wird von Anfangsblock und Endblock ausgehend jede Blockkombination auf eine Übereinstimmung in der Übersetzungstabelle überprüft. Schon übersetzte Blockkombinationen werden im weiteren nicht mehr weiter betrachtet.	60
3.10	Grundsätzlicher Aufbau einer PID Regelung. Für den in dieser Arbeit betrachteten Anwendungsfall sind die Eingangssignale (Messwert und Sollwert) relevant.	65
3.11	Übersicht über alle verwendeten Verfahren im Bereich „Anwendungen“. Meta bedeutet die Bewertung des Metadatenextraktionsverfahrens, Cluster ZK das Clustern auf Basis von Zeichenketten und Cluster ZR das Clustern auf Basis von Zeitreihen.	70
4.1	Abfolge von Untersuchungen bei der Merkmalsextraktion aus Zeichenketten von der Auswahl der zu untersuchenden Informationsqualitäten (1) bis hin zur anwendungsbezogenen Bewertung (7)	72
4.2	Vergleich des allgemeinen F1 Wertes, MCC und ECE der untersuchten Merkmalsextraktionsverfahren (ME-Verfahren, links) und Merkmalssätze (rechts) bei Training und Testen auf gleichen (blau) und unterschiedlichen Datensätzen (rot).	74
4.3	Vergleich der Merkmalswichtigkeit beim Merkmalssatz „Alle“ beim Merkmalsextraktionsverfahren „Blockweise“. Die in rot markierten Merkmale sind Merkmale, die nur in diesem Merkmalssatz enthalten sind.	76
4.4	Vergleich der Merkmalswichtigkeit beim Merkmalssatz „Alle“ bei „Gesamtheitlich“. Die in rot markierten Merkmale sind Merkmale, die nur in diesem Merkmalssatz enthalten sind.	77
4.5	Vergleich des allgemeinen F1 Wertes, MCC und ECE von Merkmalsextraktionsverfahren „Blockweise“ und „Gesamtheitlich“ bei Variation des Trainingsanteils und Training und Testen auf gleichen und unterschiedlichen Datensätzen.	79

4.6	Vergleich des Medians des allgemeinen F1 Wertes, MCC und ECE von Merkmalsextraktionsverfahren „Blockweise“ bei Variation des minimalen Kennzeichenauftritts und Training und Testen auf gleichen und unterschiedlichen Datensätzen.	81
4.7	Vergleich des Medians des allgemeinen F1 Wertes, MCC und ECE von Merkmalsextraktionsverfahren „Blockweise“ bei Variation der minimalen Blocklänge und Training und Testen auf gleichen und unterschiedlichen Datensätzen. Bei 0 werden alle Blöcke verwendet.	81
4.8	Vergleich des Median des allgemeinen F1 Wertes, MCC und ECE von Merkmalsextraktionsverfahren „Blockweise“ bei Variation des minimalen Blockauftretens und Training und Testen auf gleichen und unterschiedlichen Datensätzen. Minimum Blocklänge=0 bedeutet, dass alle Blöcke verwendet werden.	82
4.9	Vergleich des allgemeinen F1 Wertes, MCC und ECE von Merkmalsextraktionsverfahren „Gesamtheitlich“ bei Normalisierung und keiner Normalisierung und Training und Testen auf gleichen und unterschiedlichen Datensätzen.	84
4.10	Ergebnisse des allgemeinen F1 Wertes und ECE bei Merkmalsextraktionsverfahren „Blockweise“ und „Gesamtheitlich“ bei Training und Testen auf gleichen und unterschiedlichen Datensätzen bei schemaübergreifender Betrachtung.	85
4.11	Ergebnisse des F1 Wertes der Zeichenketten -Metadatenextraktion „Gesamtheitlich“ für die Bereitstellung anwendungstypischer Metadaten bei Training und Testen auf gleichen und unterschiedlichen Datensätzen.	86
4.12	Eine Übersicht über die Untersuchungen und untersuchten Eigenschaften, die im Rahmen von Zeitreihenanalysen auf den zwei Datensätzen durchgeführt werden. . .	88
4.13	Vergleich des allgemeinen F1 Wertes, MCC, ECE und des F1 Wertes bei Multilabel -Betrachtung der untersuchten Kombinationen von Merkmalsextrahierern und Klassifizierern bei Metadatenextraktion durch Zeitreihen bei Anwendung auf dem Mortar Datensatz.	89
4.14	Vergleich des F1 Wertes der verschiedenen Anwendungen bei Berücksichtigung der Datenstromtypen bei den Kombinationen MiniRocket mit Ridge und MultiRocket mit CatBoost auf dem Mortar Datensatz.	90
4.15	Vergleich unterschiedlicher Informationsqualitätskriterien (IQ-Wert) im gesamten Mortar Datensatz auf Basis von klassischer Merkmalsextraktion. Werte mit „Multi“ sind bei Berücksichtigung einer Multilabel-Zuordnung.	91
4.16	Vergleich unterschiedlicher Informationsqualitätskriterien (IQ-Wert) im gesamten OOM4-ABDO Datensatz. Werte mit „Multi“ sind bei Berücksichtigung einer Multilabel-Zuordnung.	92
4.17	Vergleich des F1 Wertes der verschiedenen Anwendungen bei Berücksichtigung der Datenstromtypen beim OOM4ABDO Datensatz.	93
4.18	Ergebnisse bei der Erkennung von zusammengehörigen Datenströmen bei RLT Anlagen im Testdatensatz	94

4.19	Verteilung der Limits des Maximums der wahren Datenströme („Max Wahr Limit“) bzw. Minimum der falschen Datenströme („Min Falsch Limit“) bei allen Datenströmen des Trainings- und Testdatensatzes von RLT Anlagen mit den c_i aus Tabelle 4.1.	95
4.20	Ergebnisse bei der Erkennung von zusammengehörigen Datenströmen bei Wärmeübertragern im Testdatensatz bei Anwendung der c_i aus Tabelle E.4.	95
4.21	Ergebnisse bei der Erkennung von zusammengehörigen Datenströmen bei Raummonitoring im Testdatensatz bei Anwendung der c_i aus Tabelle E.4.	97
4.22	Vergleich des F1 Wertes, MCC und ECE bei untersuchten Kombinationen von Merkmalsextrahierern und Klassifizierern bei unterschiedlichen Mediumskombinationen (Total, Wasser-Wasser, Luft-Luft) bei der Anwendung „ Energiefluss “.	98
4.23	Vergleich des F1 Wertes, MCC und ECE der untersuchten Kombinationen von Merkmalsextrahierern und Klassifizierern bei unterschiedlichen Datentypskombinationen (Nur Vorlauftemperatur, Vor- und Rücklauftemperatur und inklusive Volumenstrom) bei der Anwendung „ Quartier “.	100
4.24	Vergleich unterschiedlicher IQ Werte bei Anwendung von Zeitreihenklassifizierung, Zeichenkettenklassifizierung. Dies ist eine kompakte Zusammenfassung bisheriger Ergebnisse innerhalb des IE-Prozesses mit der zusätzlichen Darstellung der neu berechneten Effektivität.	102
4.25	Vergleich unterschiedlicher Cohens Kappa IQ-Werte. Es wird das Ratio (Berücksichtigung der unterschiedlichen Grundgesamtheiten) und der eigentliche Cohens Kappa dargestellt. Diese werden entweder mit Berücksichtigung einzelner Merkmale oder aller Merkmale berechnet.	103
4.26	Vergleich unterschiedlicher IQ Werte bei unterschiedlichen Anwendungen bei Anwendung von Zeitreihenklassifizierung, Metadatenextraktion und der Zuordnung innerhalb der Anwendung. Dies ist eine kompakte Zusammenfassung bisheriger Ergebnisse innerhalb des IE-Prozesses.	105
4.27	Vergleich unterschiedlicher IQ Werte bei unterschiedlichen Anwendungen bei Anwendung von Zeitreihenklassifizierung, Metadatenextraktion und der Zuordnung innerhalb der Anwendung. Dies ist die kennzahltechnische Abbildung der gesamten IE-Kette	106
5.1	Einordnung Gesamtverfahren für das Einsparpotential bei menschlichem Einsatz und Einsatzpotential der Anwendungen im Gesamtverfahren. Regelungsanwendungen sind aufgrund ihrer komplexen Zusammenhänge nur eingeschränkt anwendbar.	113

Tabellenverzeichnis

2.1	Beispiele für unterschiedliche Blöcke, die das Kennzeichen „Temperatur“ enthalten. Sie enthalten Abkürzungen, falsch geschriebene Wörter, aber auch Wörter, die nur im Kontext das Kennzeichen „Temperatur“ enthalten.	5
2.2	Statistische Merkmale der in den Ansätzen zur Informationsextraktion verwendeten Datensätze. NaN kennzeichnet Werte, die aus der Veröffentlichung nicht ersichtlich sind. Werte mit ~ sind aus der Quelle geschätzt. Gebäudeübergreifend bedeutet, dass mit Gebäude(n) A trainiert und mit Gebäude(n) B getestet wurde (Transferlernen). Die folgenden Abkürzungen wurden verwendet: ZR sind die Zeitreihen, ZK Zeichenketten, Top Topologieerkennung, Komb immer die Ergebnisse des kombinierten Verfahrens. Gen ist die Genauigkeit, Klass die Ergebnisse für die Klassifizierung, Clust die Werte für ein Clustering für eine Anwendung. Die Werte sind teilweise aus den Grafiken der Aufsätze entnommen.	8
2.2	Statistische Merkmale der in den Ansätzen zur Informationsextraktion verwendeten Datensätze. NaN kennzeichnet Werte, die aus der Veröffentlichung nicht ersichtlich sind. Werte mit ~ sind aus der Quelle geschätzt. Gebäudeübergreifend bedeutet, dass mit Gebäude(n) A trainiert und mit Gebäude(n) B getestet wurde (Transferlernen). Die folgenden Abkürzungen wurden verwendet: ZR sind die Zeitreihen, ZK Zeichenketten, Top Topologieerkennung, Komb immer die Ergebnisse des kombinierten Verfahrens. Gen ist die Genauigkeit, Klass die Ergebnisse für die Klassifizierung, Clust die Werte für ein Clustering für eine Anwendung. Die Werte sind teilweise aus den Grafiken der Aufsätze entnommen.	9
3.1	Bewertung von Cohens Kappa anhand seines Wertes.	33
3.2	Statistische Eigenschaften des Mortar-Datensatzes.	55
3.3	Formeln der statistischen Merkmale der Datensätze.	58
3.4	Statistische Merkmale der Datensätze. Für kursiv markierte Merkmale sind Übersetzungen der Blöcke notwendig (höherer Personalaufwand).	58
3.5	Beispiel für die Definition des Datensatzes in der Anwendung „Energiefluss“. Das originale Paar mit der ID 1 wird mit einem anderen Paar gemixt, wobei „von“ und „zu“ jeweils mit dem anderen Paar ausgetauscht werden.	68
4.1	Anteile c_i der einzelnen Distanzen für die Ähnlichkeitsberechnung von SSC_{total} bei raumluftechnischen Anlagen.	94

1 Motivation

Der Klimawandel ist die größte ökonomische und gesellschaftliche Herausforderung des 21. Jahrhunderts [336]. Um die Treibhausgasemissionen nach dem Pariser Klimaabkommen für das 1,5 °C zu senken, muss die anthropogene Treibhausgas (THG)-Emissionsbilanz bis 2050 ausgeglichen sein (Netto-Null) [89, 359, 377]. Die Nutzung von Gebäuden ist für 40 % der deutschen THG-Emissionen verantwortlich, wobei auf Nichtwohngebäude (NWG) rund die Hälfte (22 %) dieser Emissionen entfällt [471]. Da während der durchschnittlich 50 Jahre langen Nutzungsdauer [199] von NWG 74,6 % der Lebenszyklus-THG-Emissionen entfallen [377], ist neben der Sanierung auch die Optimierung des Betriebs erforderlich [4]. Moderne NWG sind mit einem Gebäudeautomations-system (GAS) ausgestattet [459], dass neben der Hauptfunktion, der Regelung der Gebäudetechnik, auch Energie einsparen soll [337]. Aber in der Praxis sind bis zu 90 % der Flächen von NWG ineffizient betrieben [459], weshalb durch einen nutzerangepassten Betrieb ein Einsparpotential von 15 bis 45 % der verwendeten Energie besteht [162]. Automatische Algorithmen können dazu beitragen, die Barrieren für einen effizienten Betrieb zu überwinden und zur Erreichung der Klimaziele beizutragen [415].

Für automatische Algorithmen bestehen verschiedene Optionen im Gebäude: die Optimierung der Regelung [81, 148], die Fehlerbehebung in technischen Anlagen und Sensoren [27, 130, 239, 273, 283, 398, 494], die Identifikation von suboptimal versorgten Gebäudeteilen [53, 160, 163, 318] und die Nutzung von modernen den Energieeinsatz optimierenden Regelungsmethoden wie der modellprädiktiven Regelung (MPC - engl. Model Predictive Control) [114, 436]. Für eine automatische Ausführung benötigen all diese Verfahren ein gutes Metainformationsmodell [67, 114, 366], das beispielsweise die Metadaten „Sensor“, „Sollwert“ oder „Parameter“ von Regelungen umfasst. Diese Informationen können zwar in „Building Information Modelling (BIM)“-Modellen für Neubauten und Bestandsgebäude [402, 456] abgelegt werden [151, 196, 440], aber in realen BIM-Modellen werden keine bis wenige für den Betrieb wichtige Metainformationen wie Sensoren abgelegt [259, 324, 375]. Daher ist es erforderlich, eine alternative Quelle für diese Informationen zu nutzen.

Die meisten entscheidenden Metainformationen werden in der Praxis in für jeden Betreiber individuell aufgebauten [66, 430] Anlagenkennzeichnungsschlüsseln (AKS) für Datenströme codiert [430], die für Bestandsgebäude eine personalaufwändige [367, 397], individuelle Übersetzung der AKS-Systematik erforderlich macht [45, 46, 431, 432, 433]. Dennoch basieren AKS-Systeme in der Regel auf einer grundlegenden Grammatik und Wortliste, sodass sich der Einsatz von Künstlicher Intelligenz (KI) in Form von Natural Language Processing (NLP) anbietet [245].

In der Forschung wurden bereits zahlreiche Ansätze zur Übersetzung von Zeitreihen und Daten-

strombenennungen in nutzbare Metadaten auf Basis künstlicher Intelligenz entwickelt [465], wovon die meisten nur eine der Datenquellen verwenden [39, 465] oder nur einzelne Anlagentypen oder Metadatenklassen [39, 465]. Des Weiteren wurde die Funktionsweise technischer Anlagen in Bezug auf deren Verbindung untereinander, auch als Relationsinterferenz bzw. Topologie bezeichnet, nur unzureichend untersucht [465], da sie sich nur auf einzelne Anlagentypen in vor allem luftbasierten Systemen konzentrieren [39, 465].

Die vorliegende Arbeit verfolgt das Ziel, einen Informationsextraktionsprozess (IE-Prozess) zu entwerfen und auszuführen, der aus Rohdaten aus Energiesystemen von Gebäuden und Quartieren Informationen für die Nutzung von Anwendungen zur Optimierung des Energieeinsatzes generiert (vgl. Abbildung 1.1). In diesem Prozess sollen erstmals [39, 465] die drei Verfahren (Metadatenextraktion auf Basis von Zeitreihen und Zeichenketten sowie Topologieerkennung) gemeinsam angewendet werden. Ziel des IE-Prozesses ist eine Bewertung der unterschiedlicher Ebenen anhand ausgewählter Informationsqualitäts (IQ)-Metriken, um eine Steuerung und Optimierung des Prozesses zu ermöglichen, beispielsweise durch eine Normalisierung der Datensätze. Insgesamt sollen aus Rohdaten (Zeitreihen und Zeichenketten) Informationen (Metadaten) gewonnen werden, die dann zur Wissensgenerierung (Anwendungen) genutzt werden können. In jeder Ebene soll untersucht werden, inwiefern die gewünschten Anwendungen durch den Prozess unterstützt werden.

Initial wird in Kapitel 2 der Stand der Technik und Wissenschaft im Kontext eines IE-Prozesses vorgestellt, einschließlich der Informationsextraktion in Gebäudeautomationssystemen (Kapitel 2.1), der Verwendung von Metadatenschemata in der Gebäudeautomation (Kapitel 2.2), passenden KI-Algorithmen (Kapitel 2.3) und abschließend eine Zusammenfassung des erforderlichen Forschungsbedarfs (Kapitel 2.4).

Die Methodik in Kapitel 3 umfasst die Beschreibung der Ebenen des Informationsextraktionsprozesses in der Gebäudeautomation inklusive zugehörigen Definitionen (Kapitel 3.1), die Auswahl eines einheitlichen Metadatenschemas (Kapitel 3.2), die Algorithmen zur Metadatenextraktion aus Zeichenketten und Zeitreihen (Kapitel 3.3 und 3.4) und der verwendete Datensatz in Kapitel 3.5. Schlussendlich werden in Kapitel 3.6 die acht untersuchten Anwendungen aus den Gebieten „Fehler- und Anomalieerkennung“ (Kapitel 3.6.1), „Regelung“ (Kapitel 3.6.2) und „Energiesystemanalyse“ (Kapitel 3.6.3) vorgestellt.

Den Ergebnissen (Kapitel 4) der der Metadatenextraktion auf Basis von Zeichenketten und Zeitreihen (Kapitel 4.1 und 4.2) folgen die Ergebnisse der Zuordnung von Datenströmen zu Anwendungen (4.3). Diese Ergebnisse werden in Kapitel 4.4 einer Gesamtbewertung unterzogen, die in Kapitel 5 erörtert wird und in Kapitel 6 in den Gesamtkontext eingeordnet wird.

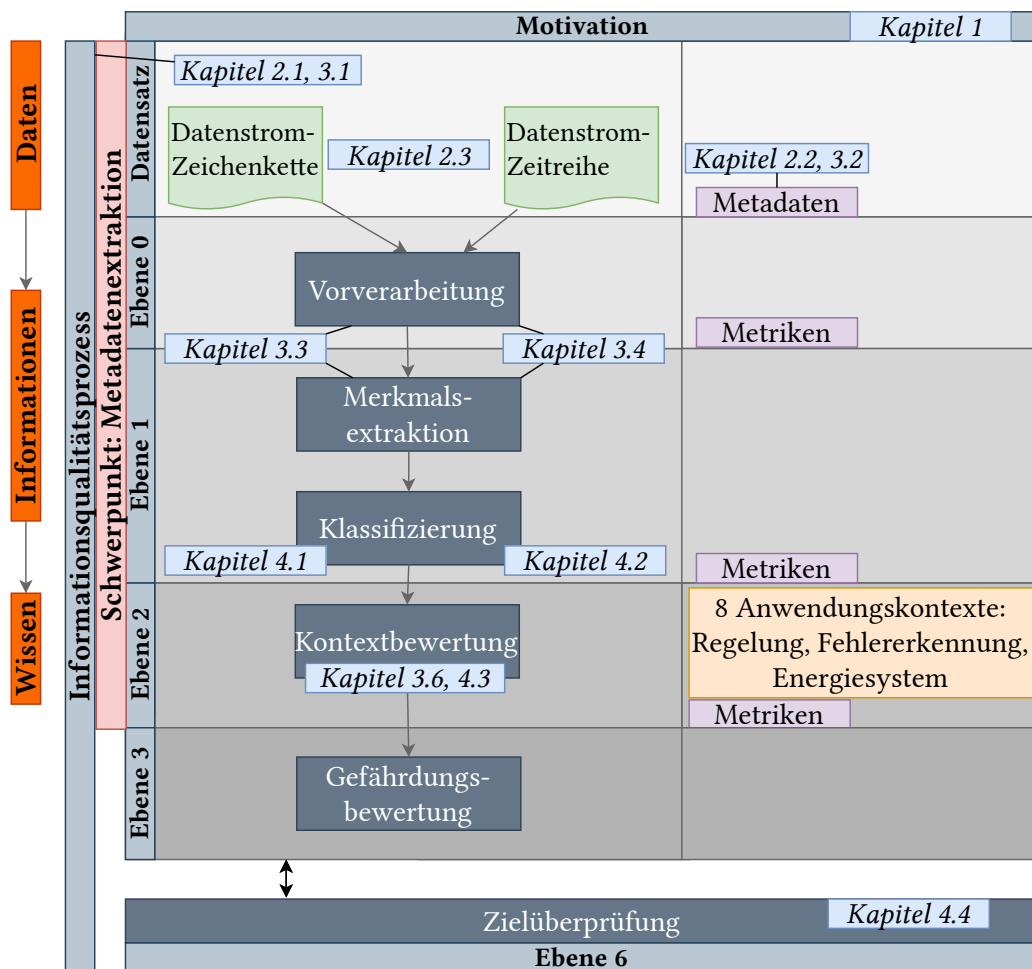


Abbildung 1.1: Ablauf der Dissertation anhand der verschiedenen Ebenen eines Informationsextraktionsprozesses. Es werden die verschiedenen Schritte in der Arbeit mit den jeweiligen Kapiteln aufgeführt. Insgesamt wird der Prozess von Daten über Informationen hin zum Wissen verdeutlicht.

2 Stand der Wissenschaft und Technik

In diesem Kapitel wird der aktuelle Stand der Wissenschaft und Technik im Rahmen dieser Dissertation näher erläutert. Zuerst erfolgt in Kapitel 2.1 eine Übersicht über Verfahren zur Extraktion von Metadaten in Gebäudeautomationssystemen. Daraufhin werden in Kapitel 2.2 verschiedene Verfahren zur Erstellung von Metadaten-Modellen im Gebäudebereich vorgestellt. Die im Rahmen dieser Arbeit untersuchten Algorithmen und Verfahren zur Merkmalsextraktion werden in Kapitel 2.3 einer näheren Betrachtung unterzogen.

2.1 Informationsextraktion in Gebäudeautomationssystemen

Die Qualität der Informationsextraktion in GAS hängt maßgeblich von den verfügbaren Datenquellen und deren Qualität ab [361]. Von den für GAS verfügbaren Datenquellen sind die folgenden besonders für die Informationsextraktion relevant:

1. Zeitreihendaten der GAS-Datenströme,
2. Namen der GAS-Datenströme,
3. Beschreibung der GAS-Datenströme,
4. Funktionsbeschreibung des GAS,
5. Rohrleitungs- und Instrumentierungsdiagramme (R&ID) des GES,
6. Funktionsliste des GES.

Bei der Auswahl geeigneter Datenquellen für einen universellen Informationsextraktionsansatz müssen jedoch mehrere Einschränkungen berücksichtigt werden. Datenstromlisten ohne eindeutige maschinenlesbare Zuordnung [140], die im Gegensatz zu industriellen Systemen fehlende Standardisierung von R&ID in der Gebäudeautomation [435, 469], eine fehlende Standardisierung und geringe Menge nutzbarer Funktionsbeschreibungen [469] und die benötigte aufwändige Aufbereitung von Funktionslisten [24] machen diese Datenquellen nur mit großem Personalaufwand nutzbar, weshalb in dieser Arbeit darauf verzichtet wird.

Als praktikable Grundlage für die Informationsextraktion aus GAS verbleiben somit die Zeitreihendaten sowie die Namen und Beschreibungen der Datenströme.

Tabelle 2.1: Beispiele für unterschiedliche Blöcke, die das Kennzeichen „Temperatur“ enthalten. Sie enthalten Abkürzungen, falsch geschriebene Wörter, aber auch Wörter, die nur im Kontext das Kennzeichen „Temperatur“ enthalten.

Block	Bedeutung
HT	Hochtemperatur
TM	Temperaturfühler
Tmp	Temperatur
Raumtempeatur	Raumtemperatur
STB	Sicherheitstemperaturbegrenzer
zu warm	Grenztemperatur überschritten

2.1.1 Definition von genutzten Datenquellen und Algorithmen

In diesem Kapitel werden die grundlegenden Begriffe der verwendeten Datenquellen und Algorithmen definiert. Die Namen und Beschreibungen der GAS-Datenströme können zusammenfassend als Zeichenketten bezeichnet werden. Ein Datenstrom wird von Stinner u. a. [432] definiert als einen „Informationsträger, der kontinuierlich Informationen über einen Zustand liefert“.

Eine Zeichenkette besteht aus Blöcken, die durch Sonderzeichen, Ziffern und Änderung von Groß- und Kleinschreibung getrennt werden. Der Beispielsname „4120-Room-Set-Temperature-RTs“ besteht aus den Blöcken „4120“, „Room“, „Set“, „Temperature“ und „RTs“. Die zugehörige Beschreibung „Room temperature setpoint“ beinhaltet die Blöcke „Room“, „temperature“ und „setpoint“. Auch die physikalische Einheit „degrees-celsius“ kann in zwei Blöcke unterteilt werden: „degrees“ und „celsius“. Ein Block kann ohne klare Trennung auch mehrere semantische Einheiten enthalten (z.B. „ROOMSETTEMP“). Es kann für diese Blöcke auch eine Segmentierung vorgenommen werden, die aber in Vorversuchen keine sinnvollen Ergebnisse lieferten [433].

Jedem Block können ein oder mehrere Kennzeichen oder Tags zugewiesen werden. So kann dem Block „Temperature“ naheliegend das Kennzeichen „Temperatur“ zugeordnet werden. Nun gibt es beispielsweise für das Kennzeichen „Temperatur“ viele verschiedene Übersetzungen. In dem in dieser Arbeit verwendeten Datensatz von Zeichenketten existieren bei ca. 36.000 unterschiedlichen Datenströmen, in denen das Kennzeichen „Temperatur“ verwendet wurde, 540 verschiedene Formen für „Temperatur“. Tabelle 2.1 zeigt unterschiedliche Beispiele für die Herausforderung bei der Verwendung von Zeichenketten nur für das Kennzeichen „Temperatur“. Eine Vielzahl an Kennzeichen wird als Metadaten bezeichnet. Die Bewertung von Metadaten auf ihre Nützlichkeit in Gebäudeanwendungen wird von Fierro u. a. [139] als semantische Suffizienz bezeichnet. Dieser Begriff wird auch hier verwendet. Eine Zeitreihe definiert sich als die semantische gleiche, zeitlich aufeinander folgende Aufzeichnung eines Datenstroms aus einer gleichen Datenquelle, z.B. einem Temperatursensor.

Im Machine Learning werden im Kontext der Informationsextraktion in GAS-Systemen unterschiedliche Begriffe verwendet. Ein Merkmal (engl. feature) ist eine für jede Instanz (Datenstrom) ein-

heitlich gewonnene Eigenschaft. Bei Zeitreihendaten aus der Gebäudetechnik können dies beispielsweise das Tagesmaximum einer Raumtemperatur, die Schwankungsbreite eines Ventilstellsignals oder die Korrelation zwischen Außentemperatur und Heizkreisvorlauftemperatur sein. Für Zeichenketten aus GAS können Merkmale etwa die Präsenz bestimmter Abkürzungen wie „T“ oder „Fnkt“ oder die Position von Schlüsselblöcken im Datenstromnamen darstellen. Ein Klassifizierer verarbeitet diese gebäudetechnischen Merkmale und ordnet den Datenstrom einer spezifischen Klasse zu, etwa als „Raumtemperatursensor“, „Ventilstellsignal“ oder „Betriebsmeldung einer Lüftungsanlage“. Solche Zuordnungen ermöglichen die automatische Interpretation von Daten aus unterschiedlichen Gebäudetypen und technischen Anlagen.

Unüberwachtes Verfahren hingegen können gebäudetechnische Datenströme anhand von Merkmalen in unterschiedliche Cluster unterteilen. Sie können beispielsweise ähnlich verlaufende Temperatursensoren in einem Gebäudetrakt gruppieren, hydraulisch verbundene Anlagenteile identifizieren oder Heizkreise mit vergleichbarem Regelverhalten clustern. Diese Gruppierung erfolgt ausschließlich anhand von Ähnlichkeitsmerkmalen der Datenströme, ohne dass eine manuelle Vorkategorisierung der Datenströme nötig ist.

2.1.2 Extraktion von Metadaten

Fierro [134] definiert vier Herausforderungen, die er als zentral für den skalierbaren Einsatz von Anwendungen in Gebäuden ansieht: eine einheitliche Beschreibung der Umgebung, eine flexible Umgebung, die Variabilität und Integration der Metadatenbeschreibung und die skalierbare Anwendbarkeit der Metadaten. Als zentrale Herausforderung sieht er die Heterogenität der Gebäudeumgebungen. Diese Heterogenität spiegelt sich auch in den Metadaten und aufgezeichneten Daten der in dieser Dissertation verwendeten Datensätzen wider und erschwerte auch in Vorversuchen entscheidend die automatisierte Übersetzbarkeit in ein standardisiertes Metadaten-Modell.

Diese Heterogenität spiegelt sich auch in den verwendeten Datensätzen wider. Die Varianz der Datenstromtypen, der Standort, die Anlagentechnik und Gebäudeanzahlen der unterschiedlichen in diesem Kapitel untersuchten Ansätzen aus der Literatur erschweren den direkten quantitativen Vergleich. Tabelle 2.2 gibt einen Überblick über die verwendeten Datensätze und Ergebnisse der im Folgenden vorgestellten Ansätze, die in Tabelle B.1 im Anhang einer genaueren Übersicht unterzogen werden. Die Ergebnisse hängen sehr stark vom verwendeten Datensatz und der Auswertung ab, weshalb die Ergebnisse im wesentlichen nur qualitativ verglichen werden können. Insgesamt kann aber beobachtet werden, dass je mehr Datenstromtypen oder Geräteklassen verwendet werden, die Ergebnisse schlechter werden, jedoch hatte die Anzahl der Datenströme nur einen geringen Einfluss auf die Ergebnisse. Es wurden aber nur in den meisten Fällen keine praxisrelevante Menge an Datenstromtypen (>100) mit einer geringen Anzahl Gebäuden (<10) verwendet.

In Abbildung 2.1 wird eine Übersicht über die untersuchten Verfahren gegeben, die in der Metadatenextraktion und Topologieerkennung zeitreihen- und R&ID-basierte Verfahren umfassen, aber

auch zusätzlich alternative und zeichenkettenbasierte Verfahren bei der Metadatenextraktion.

Die Klassifizierung ist zentral für einen gelungenen IE-Prozess in GAS-Systemen, weshalb im Folgenden die bestehenden Ansätze für Klassifizierung von GAS-Zeitreihendaten (Kapitel 2.1.2.1) und -Zeichenketten (Kapitel 2.1.2.2), aber auch alternative sowie kombinierte Ansätze (Kapitel 2.1.2.3) erörtert werden. Abschließend wird als zentrale Anwendung die Topologieerkennung in Kapitel 2.1.2.4 untersucht und abschließend die gesamten Erkenntnisse und Forschungslücken in Kapitel 2.1.2.5 aufgezeigt.

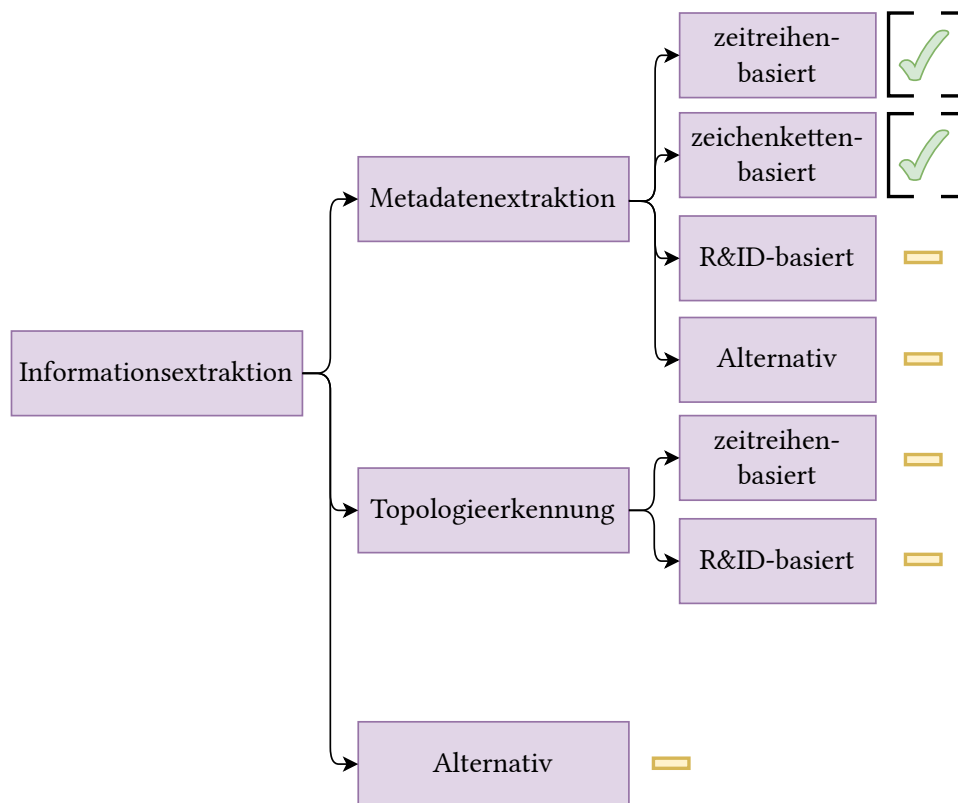


Abbildung 2.1: Übersicht über die verwendeten Verfahren zur Metadatenextraktion. Mit grünen Haken versehene Verfahren existieren in einem weitreichenden Verfahren, benötigen aber weitere Untersuchungen. Mit gelben Strich versehene Verfahren existieren nur vereinzelt.

Tabelle 2.2: Statistische Merkmale der in den Ansätzen zur Informationsextraktion verwendeten Datensätze. NaN kennzeichnet Werte, die aus der Veröffentlichung nicht ersichtlich sind. Werte mit ~ sind aus der Quelle geschätzt. Gebäudeübergreifend bedeutet, dass mit Gebäude(n) A trainiert und mit Gebäude(n) B getestet wurde (Transferlernen). Die folgenden Abkürzungen wurden verwendet: ZR sind die Zeitreihen, ZK Zeichenketten, Top Topologieerkennung. Komb immer die Ergebnisse des kombinierten Verfahrens. Gen ist die Genauigkeit, Klass die Ergebnisse für die Klassifizierung, Clust die Werte für ein Clustering für eine Anwendung. Die Werte sind teilweise aus den Grafiken der Aufsätze entnommen.

Quelle	Daten- quellen	Datenstrom- typen	Daten- ströme	Geräte- klassen	Geräte	Gebäude	Gebäude- übergreifend	Ergebnisse
[17]	ZK	NaN	68.000	NaN	NaN	50	Nein	Gen 91 %
[34]	ZK, ZR	~367	21.416	NaN	NaN	4	Ja	Gen komb < 99 %, nur ZR 63 % nur ZK 95 %
[45]	ZK	NaN	16.000	NaN	NaN	2+60	Ja	Gen 50 bis 70 %
[56]	ZK	467	38.710	NaN	NaN	72	Ja	F1 93 bis 100 %
[78]	ZR, Top	6	~680	2 (RLT, VAV)	114	2	Nein	Gen bis zu 94 %
[146]	ZR	1	3.224	1 (Licht)	365	1	Nein	Gen < 100 %
[147]	ZR	22	5.140	NaN	NaN	1	Nein	Gen < 73 %
[152]	ZR	12	~7.200	1 (RLT)	614	421	Ja	Gen < 78 %
[168]	ZR	~900	3.300	38	NaN	13	Ja	Gen < 88 %
[189]	ZR	3	900	1 (Räume)	1	1	Nein	Gen < 96 %
[193]	ZR	21	2.500	NaN	NaN	3	Ja	Gen/F1 ca. 89 %
[194]	ZK	21	2.500	NaN	NaN	3	Ja	Klass F1 41,6 bis 100 %, Clust F1 22 bis 96 %
[191]	Top	NaN	13.270	3 (RLT, VAV, Räume)	3.284	5	Ja	Gen/F1 ~ 89 %, Top < 94 %
[215]	ZK	69	NaN	NaN	NaN	2	Ja	F1 < 68 %
[244]	Top	~17	NaN	3 (RLT, VAV, Räume)	237	1	Nein	Gen ~80 %, Top < 88 %
[245]	ZK	119	6.551	NaN	NaN	5	Ja	Gen 83 bis 99 %, F1 25 bis 31 %
[266]	ZK	~9	>20000	2	NaN	9	Ja	Klass F1 90 bis 98 %, Clust Gen 80 bis 90 %

Fortsetzung auf nächster Seite

Tabelle 2.2: Statistische Merkmale der in den Ansätzen zur Informationsextraktion verwendeten Datensätze. NaN kennzeichnet Werte, die aus der Veröffentlichung nicht ersichtlich sind. Werte mit ~ sind aus der Quelle geschätzt. Gebäudeübergreifend bedeutet, dass mit Gebäude(n) A trainiert und mit Gebäude(n) B getestet wurde (Transferlernen). Die folgenden Abkürzungen wurden verwendet: ZR sind die Zeitreihen, ZK Zeichenketten, Top Topologieerkennung. Komb immer die Ergebnisse des kombinierten Verfahrens. Gen ist die Genauigkeit, Klass die Ergebnisse für die Klassifizierung, Clust die Werte für ein Clustering für eine Anwendung. Die Werte sind teilweise aus den Grafiken der Aufsätze entnommen.

Quelle	Daten- quellen	Datenstrom- typen	Daten- ströme	Geräte- klassen	Geräte	Gebäude	Gebäude- übergreifend	Ergebnisse
[275]	Top	6	NaN	3 (RLT, VAV, Räume)	NaN	6	Ja	Klass ca. 97 %, Top 90 %
[286]	ZK	NaN	4.900	NaN	NaN	3	Ja	Gen 89 %
[298]	ZK, ZR	18	11.000	NaN	NaN	5	Ja	Gen 85 %
[299]	ZK	NaN	220	NaN	NaN	2	Ja	NaN
[316]	ZR	9	5.610	NaN	NaN	76	Ja	Gen 93 bis 98 %
[321]	ZK, ZR	NaN	NaN	NaN	NaN	2	Nein	F1 ~90 %
[347]	ZR, Top	NaN	15.000	3 (RLT, VAV, Räume)	290	1	Nein	Gen 80 %
[378]	ZR	5	4.860	NaN	NaN	150	Nein	F1 77 %
[381]	ZR, Top	~9	36.940	3 (RLT, VAV, Räume)	1.281	9	Ja	Klass F1 bis 98 %, Clust Gen bis 91 %
[467]	ZK	812	1.061.810	NaN	NaN	182	Ja	Gen 71 bis 85 %
[475]	ZK	NaN	7.200	NaN	NaN	4	Ja	F1 82 %
[497]	ZK	62	NaN	NaN	NaN	2	Ja	F1 87 %

2.1.2.1 Zeitreihenbasierte Ansätze

Bei den zeitreihenbasierten Ansätzen dominieren einfache merkmalsextrahierende Methoden. Merkmalsextrahierende Methoden verwenden statistische Merkmale wie Maximum, Median, aber auch komplexere Merkmale wie Entropien oder Fast Fourier Transformation.

Chen u. a. [78] verwendeten 23 Merkmale aus den Zeitreihen gemeinsam mit einem Entscheidungsbaumansatz zur Klassifizierung von Datenströmen, jedoch nur auf der begrenzten Anzahl von sechs Datenstromtypen und zwei Geräteklassen, was die Aussage für die skalierbare Nutzbarkeit einschränkt. Für die Klassifizierung der Zeitreihen hat Holmegaard u. Kjaergaard [189] 27 va-

lierte Datenströme mit nur 3 Datenstromtypen verwendet und damit weitere 903 Datenströme klassifiziert. Fütterer u. a. [147] verwendeten 13 verschiedene Algorithmen im überwachten Lernen, um mit 8 Merkmalen 22 Typen von Datenströmen zu klassifizieren. In Reeve u. a. [381] wurden im größten Datensatz mit ~36.000 Datenströmen Variable Volumenstromgeräte (engl. Variable Air Volume: VAV) und RLT-Geräte und ihre Datenstromtypen auf der Basis von Support Vector Machines (Support Vector Machines (SVM)) erkannt. Fortschrittliche Klassifizierer finden immer stärker Anwendung. Rana u. a. [378] verwendeten XGBoost mit 11 teilweise zeitlich überlappenden einfachen Merkmalen und deren statistische Eigenschaften (Minimum, Maximum, Varianz, Standardabweichung), die vier verschiedene Klassen voneinander unterschieden.

Mertens u. Wilde [316] nutzen TSFresh mit sieben Klassifizierern (bester: XGBoost), um Merkmale aus Zeitreihen zu extrahieren und wählen dann mit statistischen Methoden sinnvolle Merkmale für den GAS-Nutzungsfall aus, wobei nur ein geringer Einfluss der Saisonalität festgestellt wurde. Es wurde trotz Verwendung des Mortar-Datensatzes nur eines der sieben verfügbaren Jahre genutzt, was die Vergleichbarkeit einschränkt.

Aus den bisherigen zeitreihenbasierten Ansätzen ergibt sich zum einen, dass kein eindeutig bester Merkmalsatz für die Klassifikation von GAS-Daten existiert. Zum anderen wurden meist baumbasierte Verfahren verwendet, die im Allgemeinen gute und sichere Ergebnisse lieferten.

2.1.2.2 Zeichenkettenbasierte Ansätze

Die Nutzung von Bezeichnungen und Metadaten in textlicher Form zeigt eine größere methodische Vielfalt. So werden verschiedene Merkmalsätze der Texte herangezogen, aber auch deutlich unterschiedliche Klassifizierer (Tag-Extraktion). Zusätzlich wurden Ähnlichkeitsverfahren verwendet, die als entscheidender Faktor identifiziert wurde.

Schumann u. a. [405] bewertete die reine Ähnlichkeitsanalyse von Bezeichnungen als nicht direkt nutzbar für die Datenstromtyp-Identifikation, sondern als sinnvoll für die Senkung der benötigten Eingaben zur Zuordnung von Klassen. Lin u. a. [286] untersuchte mit einem Conditional Random Field (CRF)-Ansatz auf Basis einer Cosinus-Ähnlichkeitsvariante den Einfluss erhöhter menschlicher Arbeitsleistung auf die Metadatenextraktionsqualität. Der Grenznutzen war bei mehr gelabelten Blöcken stark abnehmend, wirkte jedoch stark auf den erforderlichen Arbeitsaufwand bei der späteren Modellanpassung für noch nicht klassifizierte Blöcke. Die Überprüfung des Personalaufwands scheint vielversprechend, um den Algorithmus effizienter auszuführen.

Der Ansatz von Koh u. a. [245] besteht aus zwei Schichten, einer Schicht mit CRF für die Tag-Extraktion und einer Multi-Layer Perceptron (MLP)-Schicht für den Aufbau von Tag-Sets mit Bag of Words (BoW). In diesem Anwendungsfall führt dies zu einer repräsentativen Verteilung der enthaltenen Informationsstücke und hat deshalb bei strukturierten Datenströmen Nachteile, da die Grammatik der Datenbezeichnungen oft wertvolle Positionsinformationen für Tags enthält.

Im NLP-Bereich haben sich LSTM-Netzwerke bewährt. Char-LSTM-CRF mit einem siamesischen Netz auf Basis eines umfangreichen Wörterbuchs ist die Grundlage für einen Algorithmus von Jiao u. a. [215]. Darauf aufbauend wird ein CNN mit einem MLP gekoppelt, um die Ähnlichkeit zwischen den Datenströmen zu berechnen.

Die Zufallsauswahl war die beste Auswahl in einem Vergleich von Bhattacharya u. a. [45] für die skalierbare Extraktion von Metadaten mithilfe von aktiven Lernen auf der Basis von Regular Expression. Hier wurden die Solltemperaturen mit den Zeitreihendaten der klassifizierten Temperaturmessungen verglichen. Räume mit nicht zufriedenstellenden Temperaturen wurden ausgegeben. Die Zufallsauswahl ist in einer Untersuchung von Stinner u. a. [433] nur schwer zu überbieten, weshalb diese eine einfache, aber gute Auswahl darstellt.

Hong u. a. [194] verwenden k-mer als Merkmalsextraktor und SVM für die Klassifizierung. Almas-hor u. a. [17] verwenden Auto Encoder um die Metadaten in kombinierte Merkmale zu übersetzen, die dann einer Klassifizierungsschicht übergeben werden. Als Input werden alle Daten von über 100 Gebäuden verwendet, weshalb das Gebäude von den gleichen Gebäuden lernt, mit denen es getestet wird. Außerdem werden die Metadaten des Mortar-Datensatzes verwendet, die im Gegensatz zu realen Datensätzen von realen Fehlern bereinigt wurden und konsistent überall die gleichen Begriffe für die gleiche Klasse verwenden.

Auch auf Transformern basierende Modelle wurden in den letzten Jahren stärker in den Fokus gerückt. Both u. a. [56] verwendeten eine Bidirectional Encoder Representations from Transformers (BERT)-Variante mit einem „Schalenmodell“ zur ebenenweisen Identifikation der Klassen der Datenströme.

Der umfangreichste Ansatz war von Waterworth u. a. [467] mit einer gebäudeübergreifenden Klassifizierung in 182 Gebäuden. Der zweistufige Prozess kombiniert eine eigene tokenbasierter Merkmalsextraktion und nutzt ein mit speziellen GAS-Begriffen vortrainiertes Robustly Optimized BERT (RoBERTa)-Modell [107, 291] als Eingabe für einen speziellen Tagset-Klassifizierer. Der Vorteil ist, dass somit die Merkmale sinnvoller und generalisierbar extrahiert werden können. Kritisch anzumerken ist, dass die Aufteilung in Trainings- (80 %) und Testdaten (20 %) nicht explizit berücksichtigte, ob diese das gleiche Vokabular verwenden könnten. Dies ist ein Faktor, der die Ergebnisse stark beeinflussen kann, wenn Gebäudeautomatisierer ähnliche Strukturen in verschiedenen Gebäuden implementieren.

In der darauf aufbauenden Arbeit beschränken Waterworth u. a. [468] die verwendeten Gebäude auf drei und verwenden einen klassischen Bi-LSTM-CRF-Ansatz [197] mit schwacher Supervision. Der tatsächliche Nutzen und Aufwand hängt aber stark von der benutzerfreundlichen Anwenderumgebung ab [358], das in dieser Arbeit nicht fokussiert wird. Nach ihrer Literaturrecherche [468] soll ein typischer Ansatz für GAS-Daten mindestens 10 % der Gebäudeautomationsdaten als Trainingsdaten aufweisen.

Bei zeichenkettenbasierten Verfahren variierten die verwendeten Ansätze deutlich stärker als bei

den zeitreihenbasierten. Insbesondere Ähnlichkeitswerte von Zeichenketten als unüberwachte Methode war beliebt. Transformer-basierte Verfahren wie BERT lieferten auch gute Ergebnisse, aber wurden nicht mit dem vollständigen Merkmalsraum auf großen Datenmengen angewendet. Diese haben auch den Nachteil, dass diese eine sehr lange Trainingszeit und entsprechendes Material benötigen. Hier könnten wiederum baumbasierte Verfahren mit GAS-spezifischen Merkmalen eine leichtgewichtigere Alternative darstellen.

2.1.2.3 Alternative und kombinierte Ansätze

Ein zwar unterschwelliger, aber arbeitsintensiver Ansatz, der auf der Erfassung von Informationen durch Benutzeraktionen (Ablesen Thermostate, Lichtschalter) basiert, wurde von Fürst u. a. [146] entwickelt. Für GES wurde R&ID nur in zwei Fällen zur Klassifizierung von Datenstromtypen verwendet: Zhang [501] nur für die Erkennung von Klassen einzelner Symbole, während Stinner u. a. [435] auch die Topologie der Systeme erkannte, aber eine fehlende Verbindung von Symbolen zu den Zeitreihen mokierte.

In einem Beitrag von Gao u. Bergés [152] wurden sechs verschiedene Zeitreihen- und Datenstromkennzeichnungsansätze aus der Literatur [278, 153, 192, 47, 34, 244] mit unterschiedlichen Algorithmen (bester: Random Forest Klassifizierer (RF)) an einem großen Datensatz von RLT-Geräten (12 Datenstromtypen in 614 RLT-Geräten in 421 Gebäuden an 35 Standorten in verschiedenen Klimazonen) verglichen. Die Trainingsstrategie hatte kaum Auswirkungen auf die Ergebnisse.

Hamiz Ahmed [168] verwendete ein bidirektionales LSTM [197] mit GloVe [353] für die Metadatenextraktion und einen Random Forest Klassifizierer basierend auf statistischen Merkmalen für die Zeitreihenklassifikation. Die Vorhersage zwischen verschiedenen Gebäuden war hier ein großes Problem, was auch im Rahmen dieser Arbeit untersucht werden sollte.

Hong u. a. [193] verwendeten verschiedene Methoden zur Klassifizierung von Datenströmen mittels Transfer Learning. Sie verwenden SVM, Logistische Regression (LR) und RF mit 11 Merkmalen für die erste Klassifizierung. Für das übergreifende Transferlernen verwenden sie unüberwachte Gaussian Mixture Models (GMM).

Balaji u. a. [34] verwendeten eine Kombination aus hierarchischem Clustering und RF zur Klassifizierung. Balaji u. a. [34] wiesen darauf hin, dass mehrere Instanzen für einen Tag (z. B. „T“, „temp“ und „temperature“ für den Tag „temperature“) schwer zu erlernen sind und hierarchisches Clustering hierbei helfen könnte. Zunächst werden die Datenstrombezeichnungen anhand von 6 Merkmalen und 24 Merkmalen der Zeitreihe gruppiert. Auf dieser Basis wenden sie hierarchisches Clustering an, um die Datenstromtypen voneinander zu unterscheiden und die Anzahl der Cluster entsprechend zu reduzieren. Ein gleichartiges Vorgehen mit anderen Clustering-Algorithmus (k-mers) wurde auch von Mishra u. a. [321] durchgeführt. Da sich alle vier hier verwendeten Gebäude auf einem Campus befinden, ist davon auszugehen, dass die Kennzeichnungen sehr ähnlich sind.

Die Analyse ergab, dass nur wenige kombinierte Verfahren verwendet wurden und hier auch wiederum baumbasierte Verfahren dominierten. Wenn kombinierte Verfahren verwendet wurden, dann wurden nur vergleichsweise kleine Datensätze untersucht.

2.1.2.4 Topologieerkennung als spezieller Anwendungsbereich

Die Topologieerkennung, die Identifikation von Verbindungen zwischen technischen Anlagen, bildet einen spezialisierten Anwendungsbereich. In der Gebäudetechnik als „relationelle Interferenz“ oder „Beziehung“ bezeichnet, entspricht sie in mathematischen bzw. informationstechnischen Netzwerken dem Konzept der Topologie [75, 90, 95, 235].

Die Verbindungen zwischen RLT-Geräten, VAV-Geräten und Räumen wurden auf Basis von aktiven Sollwerten von Koh u. a. [244] untersucht. Die Signale wurden auf Korrelationen untersucht, wobei als Zeitreihentransformation Fast Fourier Transform (FFT) mit euklidischem Abstand, DTW, Mittelwert, Varianz und Anzahl der Änderungen und als Klassifikator acht klassische Machine Learning (ML)-Ansätze verwendet wurden.

Pritoni u. a. [364] verwenden die Zonentemperaturen, die Zulufttemperatur und den Volumenstrom, um die Anschlüsse von RLT-Geräten, VAV-Geräten und Zonen im Gebäude zu identifizieren. Sie waren auch auf die Verbindungen innerhalb der RLT- und VAV-Geräte beschränkt.

Hong [191] verwendet Markov-Ereignismodelle, da Hidden-Markov-Modelle (HMM) die zeitlichen Zusammenhänge in der Topologie nicht ausreichend abdecken. Anhand der Übergangszustände kann bestimmt werden, welche Zeitreihen miteinander korreliert sind. HMM boten in Vortests keine guten Ergebnisse, die im Rahmen der Erstellung der Studie von Stinner u. a. [434] getätigt wurden, weshalb diese nicht berücksichtigt wurden.

Von Li u. a. [275] wurde die Topologie auf der Grundlage der Kurzzeit-Fourier-Transformation (engl. short-time Fourier transform: STFT) unter Verwendung eines überwachten CNN bestimmt. Im Postprocessing wurden unmögliche Verbindungen (z.B. zwei RLT-Geräte zu einem VAV) auf Basis von Graphenschnitten aussortiert. Dies erfordert besonderes Wissen über die Struktur der Verbindungen, die nur sehr aufwändig bei einer mediumsübergreifenden Topologieerkennung durchgeführt werden kann.

Stinner u. a. [432] nutzten CNN-basierte Algorithmen zur Klassifizierung von Verbindungen von Wärmeübertragern und Wärmepumpen, wobei sie simulierte und reale Daten als Eingabe nutzten. Die simulativen Daten schnitten in diesem Fall besser ab. Nachfolgende Ergebnisse des Autors mit anderen Ergebnissen und Vorverarbeitungsfunktionen zeigten aber, dass dies nur bei einzelnen Geräten sichtbar wird.

Chen u. a. [78] kombinierte alleinig die Metadatenklassifizierung und Topologieerkennung in einem Verfahren, in dem er Zeitreihen zur Klassifizierung nutzte und dann die begrenzten semantischen Informationen zur Topologieerkennung. Dabei wird nur ein begrenzter Datensatz verwendet (nur 6 Arten von Datenströmen von 114 Geräten aus zwei Gerätetypen).

Die Erkennung von Verbindungen durch eine aktive Steuerung der Systeme muss zu Zeiten erfolgen, in denen die Gebäude nicht genutzt werden [244], weshalb Monitoringdaten nur eingeschränkt nutzbar sind. Für eine gebäudeübergreifende Analyse müssten aufwändig mehrere Gebäude aktiv gesteuert werden, was ein eigenes Sicherheitsmanagement erfordert. Daher wird in dieser Arbeit kein aktiver Ansatz verfolgt.

Gemeinsam bei allen untersuchten Ansätzen war, dass nur sehr wenige Klassen und Beziehungen untersucht wurden. Auch wurden nur fast ausschließlich einzelne Anwendungen der Topologie (z.B. RLT-Anlagen) in einem Medium (nur Luft) betrachtet. Hier besteht noch ein großer Forschungsbedarf, um diesen Ansatz skalierbar nutzbar zu machen. Beispielsweise ist nicht bei jedem Temperatursensor bekannt, ob dieser Lufttemperaturen misst.

2.1.2.5 Erkenntnisse und Forschungslücken

Die Analyse des aktuellen Stands der Forschung zusammen mit den Ergebnissen in Tabelle 2.2 zeigt klare Tendenzen auf.

- Bei der Zeitreihenerkennung dominieren Entscheidungsbäume mit einfachen Merkmalsextrahierern, während kaum komplexe künstliches neuronales Netz (KNN) verwendet wurden.
- Bei wenigen Klassen wurden bei Zeitreihen Ergebnisse im F1 Wert und der Genauigkeit oberhalb von 85 % erzielt.
- Bei mehr als 20 Klassen wurden die Ergebnisse im Trend schlechter als 70 %, was eine hohe Abhängigkeit der Ergebnisse von der Anzahl der Klassen aufzeigt.
- Der Trend in der Zeichenkettenerkennung geht hin zu Transformer basierten Large Language Models (LLM), die jedoch nicht immer bessere Ergebnisse liefern.
- Die Ergebnisse für die zeichenkettenbasierten Algorithmen erreichten auch F1-Werte von 95 bis 100 %, wobei 100 % nur bei Nutzung von wenigen Klassen erreicht wurden.
- Die meisten Ansätze waren insgesamt nur auf wenige Klassen konzentriert und selten wurde schemaübergreifend gearbeitet, was jeweils tendenziell bessere Ergebnisse zur Folge hat.
- Die Topologieerkennung erreichte Genauigkeitswerte von bis zu 94 %, allerdings beschränkten sich diese Ansätze nur auf eine Art der Verbindung von Geräten (z.B. Lüftungsgerät zum Raum) mit einem Transportmedium (Luft). Für die skalierbare, automatisierte Anwendung müssten auch diese Werte mediumsübergreifend höher liegen.
- Die Eignung extrahierter Metadaten für konkrete Anwendungsfälle wurde bisher kaum evaluiert. Und wenn diese evaluiert wurden, so beschränkten sich die Anwendungsfälle meist auf einzelne Anlagen und Maßnahmen. Es wurde nur selten mit konkreten Daten aus Gebäuden analysiert, welche Metadaten für eine große Bandbreite an Anwendungsfällen erforderlich sind und dies auch auf einem großen Datensatz evaluiert.

- Die systematische Bewertung der Informationsqualität im Extraktionsprozess fehlt. Es wurden zwar teilweise Bewertungen der Qualität vorgenommen, aber nicht systematisch untersucht, welche Auswirkungen einzelne Qualitäten auf den Prozess haben.

Die vorliegende Arbeit soll diese Lücke schließen, indem sie neue Ansätze für die Metadatenextraktion und Topologieerkennung innerhalb eines Informationsextraktionsprozesses entwickelt und evaluiert, die mit größeren Klassenanzahlen umgehen können und eine zuverlässige gebäude-, schema- und mediumsübergreifende Anwendbarkeit gewährleisten. Die Bewertung unterschiedlicher Informationsqualitäten in verschiedenen Anwendungsfällen zeigt ein vollständigeres Bild auf, welche Hürden im Informationsextraktionsprozess in GAS aktuell existieren. Durch die Kombination der Analyse von zeitreihenbasierten und zeichenkettenbasierten Verfahren soll eine robuste und umfassende Lösung für die Metadatenextraktion in heterogenen Gebäuden geschaffen werden.

2.2 Metadaten im Gebäudeautomationssystem

In der Praxis erfolgt die Definition der Metadaten in Gebäudeautomationssystemen in erster Linie über ihren Namen und die Beschreibung. Infolgedessen sind die Metadaten der Anlagentechnik oder GAS [373] gegenwärtig häufig entweder unstrukturiert oder in einer Form verfasst, die eine automatisierte Auswertung nicht zulässt [46, 237] auch wegen fehlerhaften Übergaben von Daten [510].

Es gibt viele „Standards“ für den Aufbau von Metadaten im Gebäude, aber keiner hat sich eindeutig durchgesetzt [46], was durch die fehlenden auswertbaren Metamodelle die bedeutsame Portabilität von Anwendungen einschränkt [25, 40, 80, 201, 373].

Nachfolgend erfolgt eine Auswahl von fünf Metadaten schemata, welche als potenzielle Kandidaten für eine durchgängige Kennzeichnung von Metadaten im Gebäude in Frage kommen.

- **Industry Foundation Classes (IFC)** als Vertreter des BIM,
- **Virtual Buildings Information System (VBIS)** als code-basiertes Verfahren analog zu AKS und mit Integration in BIM Systemen,
- **Brick Schema** als Ontologie-Ansatz für die Gebäudetechnik,
- **Project Haystack** als tag-basiertes Verfahren,
- **BUDO Schema** als AKS mit Verbindungen zu Ontologien.

2.2.1 Industry Foundation Classes (IFC)

Der IFC Standard [112] Modell 4 etabliert sich zunehmend als offener Austauschstandard im Bereich von BIM. IFC kann durch standardisierte Konstruktionsdaten [69] neben dem Bauprozess die

Optimierung des Energiesystems unterstützen [375]. Eine Umfrage unter Architektenkammern in Deutschland zeigte jedoch, dass lediglich 18 % der Befragten BIM nutzen [383]. Dies ist ein Hinweis darauf, dass insbesondere kleine Bauunternehmen Schwierigkeiten bei der Umstellung auf BIM haben [101, 159], wobei die internationale Durchdringung von IFC und BIM länder- und regionenweise variiert [77, 220, 300]. IFC wird in erster Linie für den Bauprozess [420] und das Gebäudemodell [141] verwendet, wobei es in der Regel lediglich als Produktinformationsmodell verstanden wird [196, 230]. Die Möglichkeit Informationen über den Betrieb zu integrieren [68, 94, 158, 307] werden nicht genutzt [259, 324] und somit nur unzureichend für den Lebenszyklus vorbereitet [196]. Beim IFC-Modell fehlen aktuell insbesondere speziellere Eigenschaften, insbesondere für das Facility Management [413, 474], weshalb die Informationstiefe der auf dem IFC-Standard basierenden Modelle für den Betrieb schon in der Theorie nicht besonders hoch ist [280].

2.2.2 Virtual Buildings Information System (VBIS)

Das Virtual Buildings Information System (VBIS) [452] stellt ein System zur Bezeichnung von Bauteilen in Gebäudesystemen und BIM-Systemen dar. Die exakte Bezeichnung eines Bauteils kann in bis zu vier Stufen aufgestellt werden: Beispiel Mechanik (ME) → RLT-Anlage (engl. Air Handling Unit: „AHU“) → Wärmerückgewinnung (engl. Heat Recovery: „HR“) → modular („Mo“). Die einzelnen Auswahlmöglichkeiten ergeben in ihrer Gesamtheit einen Code von ME-AHU-HR-Mo, um alle Eigenschaften abzubilden. Insgesamt stehen 31 unterschiedliche Bereiche wie „Mechanik“ zur Verfügung, die verschiedene Aspekte eines Gebäudes abbilden.

2.2.3 Brick Schema

Das Brick Schema [32] ermöglicht die Erstellung eines Gebäude-Energiesystems mit seinen Systemen auf Basis von ontologiebasierten Formalismen [32, 33] und wurde als Antwort auf die fehlende Strukturierung von Gebäudebetriebsdaten entwickelt [46]. Von Fierro u. a. [136] wurde das Brick Schema zu „Brick +“ erweitert, wodurch eine Darstellung deutlich umfangreicherer semantischer Informationen ermöglicht wird. Das Brick Schema umfasst eine Vielzahl von Beziehungen, welche gemäß dem Selbstverständnis der Entwickler eine präzise Beschreibung technischer Anlagen, Systeme, Standorte, Messgrößen, Medien und Datenstromtypen sowie deren Verbindungen ermöglichen. Das Brick Schema wird mithilfe eines Konsortiums unterstützt, das Forschung und Industrie verbindet [63].

Brick weist insbesondere Defizite hinsichtlich der Elemente der klassischen Gebäudeautomation (z.B. Steuerungsgeräte oder Regelungseinstellungen) oder Räumen oder Zonen auf [109, 322], die durch eigene anreichernde Ontologien von den Autoren behoben wurden bzw. durch Anbindung anderer Ontologien wie RealEstateCore [125] gelöst wurden. Diese Eigenschaften sind jedoch von entscheidender Relevanz für die Integration in ein Informationssystem, welches die Informationen für die Anwendungen über den gesamten Lebenszyklus bereitzustellen hat.

2.2.4 Project Haystack

Das Projekt Haystack definiert 230 Tags [369], die zur Darstellung eines semantischen Modells verwendet werden können. Ein Tag stellt dabei eine Verknüpfung zwischen technischen Systemen dar (Referenzen). Das bereits vorgestellte Brick Schema übersteigt das Project Haystack in der Expressivität und Vollständigkeit deutlich [372], weshalb auf eine weitere Betrachtung verzichtet wird.

2.2.5 BUDO Schema

Im Jahr 2018 wurde das Buildings Unified Data Point Naming Schema for Operation Management (BUDO) Schema auf Basis des Benennungsschemas des Fraunhofer Institut für Solare Energiesysteme (ISE) [382] veröffentlicht [430]. Das BUDO Schema wurde im Rahmen des Projektes „Objektorientiertes Monitoring als Grundlage für einen effizienteren Betrieb sowie kostengünstige Bestandsoptimierung durch Anwendung von Machine Learning-Techniken“ (kurz: OOM4ABDO, BMWK FKZ 03SBE006A) entwickelt und gegenüber dem Standard des Fraunhofer ISE deutlich erweitert. Die Grundstruktur des BUDO Schemas ist in Abbildung 2.2 dargestellt. Im Rahmen einer Untersuchung von 125 internationalen Normen und Standards in der Gebäudetechnik konnte festgestellt werden, dass die Struktur der Datenstromnamen in Abhängigkeit vom Betreiber(-typ) variiert.

Um den genannten Unterschieden Rechnung zu tragen, wurden am Anfang des BUDO-Schemas auf Basis der vorherigen Untersuchung 63 individuelle Kategorien definiert. Danach stehen dem Nutzer drei feste Systemkategorien zur optionalen Auswahl näherer Beschreibungen zur Verfügung. Beispielhaft sei hier die Kategorie „Kessel“ mit der näheren Beschreibung „Brennwertkessel“ genannt. Im Anschluss besteht die Möglichkeit, eine optionale Bezeichnung zu vergeben (beispielsweise eine Nummerierung). Anschließend können das Medium (z.B. Wasser), Signaltyp, Funktionstyp und weitere freie Kategorien wie die physikalische Einheit definiert werden.

Jede Eigenschaft, die im BUDO Schema definiert wird, kann auch in das Brick Schema übersetzt werden [432]. Fast alle in der Untersuchung genutzten Normen und Standards definieren eine hierarchische Struktur, die Zuweisungen zu Anlagen wie RLT-Anlagen oder Räumen explizit vorsieht. Diese Zuweisungen können in der hierarchischen Struktur des BUDO Schemas direkt abgebildet werden. Das BUDO Schema ist so flexibel, dass auch eine Unterscheidung zwischen versorgten und versorgenden Anlagen und Medien näher definiert werden kann. Die Verbindungen zwischen technischen Geräten, wie beim Brick Schema möglich, sind für mehr als 1-zu-n-Beziehungen im BUDO Schema nur schwierig abzubilden, sind aber grundsätzlich möglich und erhöhen die Länge des Datenstromnamens erheblich. Jedoch waren kompliziertere Beziehungen in Beispielgebäuden des Brick Schemas nicht häufig anzutreffen.

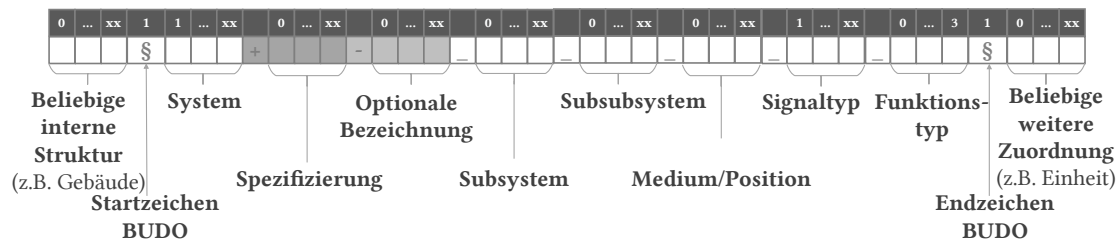


Abbildung 2.2: Struktur und Aufbau des BUDO Schemas

2.2.6 Weitere Schemata/Bewertung

Im Rahmen dieser Arbeit wird auf eine detaillierte Analyse weiterer Ontologien und Beschreibungsarten für Metadaten verzichtet. In diesem Kontext sind die folgenden Beschreibungsarten von Relevanz: Google Digital Buildings [42], SAREF [103] mit seiner Erweiterung für Gebäude SAREF4BLDG [362], sowie RealEstateCore [125]. Jede dieser Beschreibungsarten weist spezifische Stärken und Schwächen auf, wobei keine der genannten die Informationstiefe von Brick Schema, Projekt Haystack oder VBIS erreicht oder weit verbreitet ist [137].

Eine aktuelle Metadatenauswertung auf Basis der in vielen Gebäuden eingesetzten Anwendungen fehlt derzeit [297]. Daher wird zunächst geprüft, welches Metadatenschema für die in dieser Arbeit betrachteten Anwendungen am besten geeignet ist (siehe Kapitel 3.2).

2.3 Algorithmen

Im Rahmen der Dissertation wird ein IE-Prozess entworfen, der zwei Datenquellen beinhaltet zur Metadatenextraktion: Zeitreihen und Zeichenketten, die einer Merkmalsextraktion und Klassifizierung unterzogen werden.

Die Merkmalsextraktion besteht aus statistischen (Kapitel 2.3.1) und fortschrittlichen (Kapitel 2.3.2) Merkmalen aus der Zeitreihe. Fortschrittliche Merkmalsextraktion kann Algorithmen aus dem Bereich der neuronalen Netze umfassen. Algorithmen für die Klassifikation werden in Kapitel 2.3.3 vorgestellt, die auch für Zeichenketten gleichermaßen verwendet werden.

Um Datenstromnamen zu klassifizieren oder zusammengehörige Datenströme anhand des Namens zu identifizieren, können Textdistanzen bzw. -ähnlichkeiten verwendet werden, die in Kapitel 2.3.4 vorgestellt werden.

2.3.1 Klassische Merkmalsextraktion

Die klassische Merkmalsextraktion kann sehr variantenreich sein, weshalb zahllose Ansätze existieren. Für diese Arbeit werden drei Ansätze und Pakete gesondert betrachtet, die eine größere Verbreitung in der Nutzung besitzen [449], die im Folgenden vorgestellt werden:

- **Time Series Feature Extraction on basis of Scalable Hypothesis tests (TSFresh)** Christ u. a. [88] kombiniert 63 verschiedene Methoden zur Charakterisierung von Zeitreihen mit insgesamt 794 extrahierten Merkmalen. Es können automatisiert nur relevante Merkmale ausgewählt werden.
- **Time Series Feature Extraction Library (TSFEL)** [35] extrahiert statistische, spektrale, temporale und energetische Merkmale und kann die Zeitreihen einer Datenvorbereitung (Filterung, Normalisierung) unterziehen.
- **Multivariate Time Series Transformer (MVTs)** [11] kann laut Aussage der Autoren eigentlich Merkmale aus multivariaten Zeitreihen extrahieren, aber nutzt keine multivariaten Merkmale wie die Kovarianz zweier Zeitreihen wie Temperaturmessung oder Sollwert. Er kann nur aus multivariaten Zeitreihen univariate Merkmale extrahieren.

2.3.2 Fortschrittliche Merkmalsextraktion

In diesem Kapitel werden fortschrittliche Verfahren zur Merkmalsextraktion beschrieben, die in neuronalen Netzen verwendet werden. Zunächst werden auf neuronalen Netzen basierende Long short-term memories (LSTM) und darauf aufbauend die auf Convolutional Neural Networks (CNN) basierende Algorithmenfamilie Rocket vorgestellt. LSTM und CNN können auch bei Zeitreihendaten [133, 222, 223, 224] aus der Gebäudeautomation [432] erfolgreich eingesetzt werden.

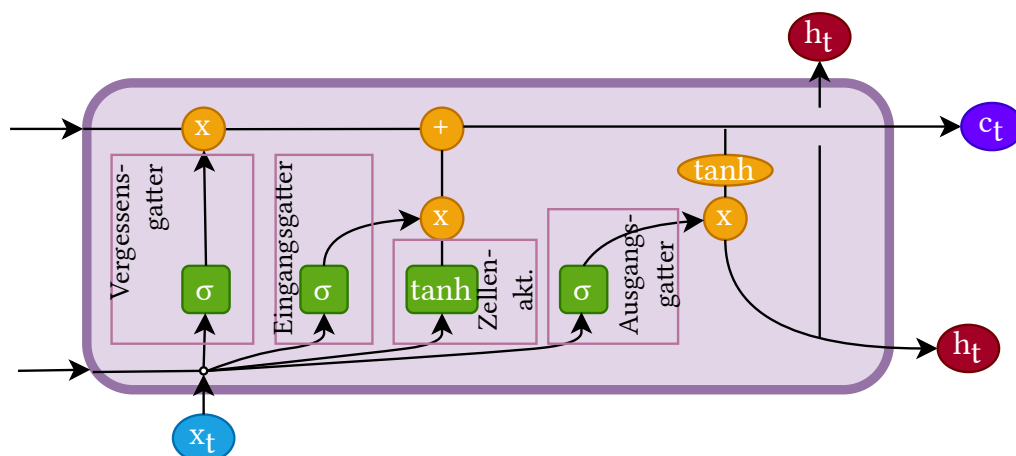


Abbildung 2.3: Das Long short-term memory mit seinen vier Gattern (Vergessensgatter, Eingangsgatter, Zellenaktualisierung, Ausgangsgatter).

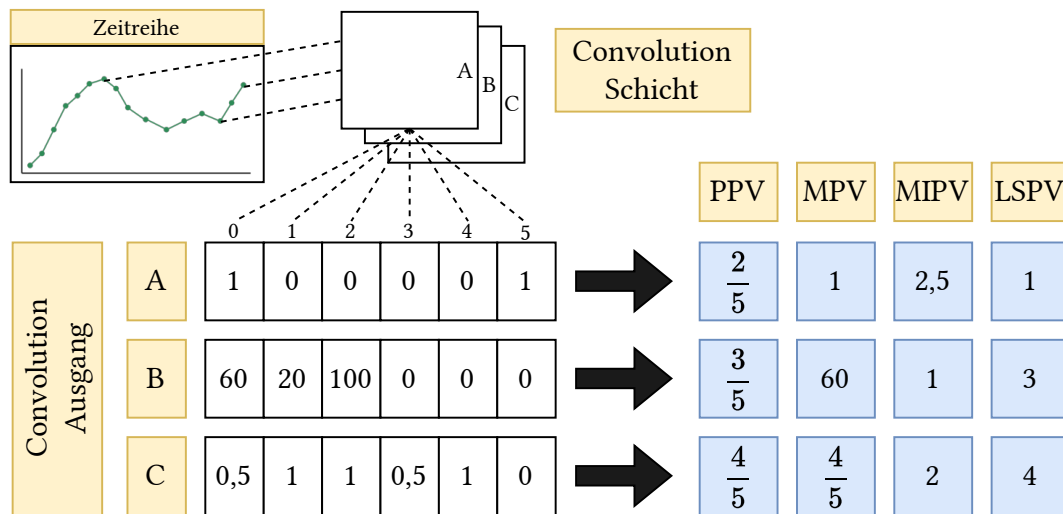


Abbildung 2.4: Aufbau der Merkmalsextraktion durch MultiRocket: Die Zeitreihen werden mithilfe einer Convolution Schicht in verschiedene Convolutional Kerne unterteilt (im Beispiel: A, B, C). Aus jedem der MultiRocket Convolutional Kerne werden vier Merkmale extrahiert: Anteil positiver Werte (PPV), Mittelwert der positiven Werte (MPV), Mittelwert des Index der positiven Werte (MIPV), längste Strecke aufeinanderfolgender Werte über Null (LSPV). Für das Beispiel B ist der PPV $\frac{3}{5}$, da drei von fünf Werten positiv sind, MPV ist $\frac{60+20+100}{3} = 60$, MIPV beträgt $\frac{0+1+2}{3} = 1$ und LSPV $0 \rightarrow 2 = 3$. Diese Merkmale können dann einem linearen Klassifizierer übergeben werden.

- **Long short-term memories (LSTM)** [186] basieren auf neuronalen Netzen, die grundsätzlich Eingabe-, Ausgabe- und versteckten Schichten besitzen. Im LSTM sind diese Schichten als Gatter umgesetzt: Vergessensgatter, Eingangsgatter, Zellenaktualisierung und Ausgangsgatter (siehe Abbildung 2.3). Mithilfe dieser Gatter kann ein LSTM eine Sequenz (z.B. Zeichenketten) gut erfassen, weshalb LSTM zusammen mit CRF-Schichten für die Übersetzung von Zeichenketten verwendet werden.
- **Rocket** [105] ist eine schnelle, energieeffiziente Alternative zur klassischen Merkmalsextraktion auf Basis von am University of California (UCR)-Datensatz vortrainierter einschichtiger CNN [13, 253, 492]. Das CNN wird aufgeschnitten und aus den Gewichten des CNN werden Zufallsfaltungskerne mit ihrer Länge, Dilatation und Auffüllung ausgewählt [106]. Zufallsfaltungskerne besitzen ein hohes Potenzial, die Merkmale von Zeitreihen zu erfassen und eine entsprechende Zeitreihe mit weniger Merkmalen darzustellen [105]. Um die Abhängigkeit von einem Kern zu verringern, werden 10.000 Kerne verwendet. Als Merkmale werden in jedem Kern Global Max Pooling und dem Anteil positiver Werte (PPV) berechnet. Durch die begrenzte Anzahl an Merkmalen können lineare Klassifikatoren verwendet werden, die bei klassischen ML-Zeitreihenaufgaben (Klassifikation, Regression, Vorhersage) in der Regel versagen, da sie selten lineare Zusammenhänge aufweisen [208].

- **MiniRocket** [106] wurde entwickelt, um die Ausführungszeit von Rocket um den Faktor 10 bis 100 zu reduzieren. Dies wurde durch die alleinige Nutzung des PPV und der Fixierung der Rocket-Parameter erreicht.
- **MultiRocket** [439] erweiterte MiniRocket um drei weitere Merkmale [296]: längste Strecke aufeinanderfolgender Werte über Null (LSPV), der Mittelwert der positiven Werte (MPV) und der Mittelwert des Index der positiven Werte (MIPV). In Abbildung 2.4 ist ein Beispiel eines MultiRocket-Kerns mit seinen vier Merkmalen dargestellt. MultiRocket konnte die Ergebnisse auf dem UCR-Datensatz von MiniRocket übertreffen. Im Rahmen dieser Arbeit wurde MultiRocket in die Bibliothek sktime mit den entsprechenden Schnittstellen integriert [294].

2.3.3 Klassifizierungsalgorithmen

Für die Einteilung von Datenströmen in Klassen können verschiedene Klassifizierungsalgorithmen wie Random Forest, XGBoost, CatBoost und als linearer Algorithmus der Ridge Klassifizierer verwendet werden, die im Folgenden vorgestellt werden.

- **Random Forests (RF)** sind eine Unterart der Entscheidungsbäume [61]. Sie sind leicht interpretierbar, haben eine kurze Laufzeit bei insgesamt guten Ergebnissen für Zeitreihen, was sich insbesondere für die Klassifikation von Energiedaten bestätigt [36].
- **eXtreme Gradient Boosting (XGBoost)** [79] verwendet exakte Greedy-Algorithmen für die Auswahl von Entscheidungsbäumen und nutzt Regulierungsterme in der Zielfunktion für die Vermeidung von Overfitting.
- **CatBoost** [74, 113, 370] (für „categorical boost“) kann kategoriale Merkmale direkt verarbeiten und nutzt „oblivious trees“ zur Vermeidung von Overfitting. CatBoost profitiert kaum von der genauen Einstellung der Hyperparameter und kann somit robust ohne Hyperparameter Tuning eingesetzt werden [170].
- **Ridge Classifier** ist ein lineares Klassifizierungsmodell, das einen Regularisierungsterm (Ridge-Regularisierung) verwendet, um Overfitting zu vermeiden [187, 188]. Der Ridge Klassifizierer ist der empfohlene Klassifizierer für die Rocket-Merkmalsextraktion [105], weshalb er hier als Vertreter der einfachen linearen Klassifizierung verwendet wird.

2.3.4 Ähnlichkeit von Zeichenketten

In dieser Arbeit werden elf verschiedene Algorithmen zur Ähnlichkeitsbewertung von Zeichenketten verwendet. Die Ähnlichkeit von Zeichenketten wird beispielsweise zur Erkennung von Plagiaten [267], zur Normalisierung von Personendaten [458] oder zur Bewertung der semantischen Ähnlichkeit von Texten [207] verwendet. In den meisten Fällen wird die Ähnlichkeit als normalisierte Form der umgekehrten Distanz von Zeichenketten zwischen 0 (keine Ähnlichkeit) und 1 (absolute Ähnlichkeit) definiert [331].

Nach Naumann [331] können die Ähnlichkeiten von Zeichenketten in fünf Kategorien eingeteilt werden: editierbasiert, token-basiert, phonetisch, domänen-abhängig und hybride Verfahren. Für die ersten drei Kategorien werden unterschiedliche Verfahren im Folgenden vorgestellt. Die detaillierte Beschreibungen der Jaro-Winkler-, Cosinus- und Damerau-Levenshtein-Distanz wurden weitgehend bereits in Stinner u. a. [429] behandelt.

2.3.4.1 Editierbasierte Verfahren

Editierbasierte Verfahren kalkulieren die benötigten Operationen für die Umwandlung von Zeichenkette a in die Zeichenkette b :

- **Damerau-Levenshtein-Distanz** [102, 268] ist eine Erweiterung der Hamming Distanz [169] für ungleichlange Zeichenketten, der die gewichtete Summe der benötigten Operationen (Ersetzung, Einfügung, Löschung, Vertauschung) für die Umwandlung einer Zeichenkette a in die Zeichenkette b anzeigt.
- **Optimal String Alignment (OSA)-Distanz** [182] wird gleich berechnet wie die Damerau-Levenshtein-Distanz, nur dass für jedes Zeichen nur eine Operation erlaubt ist.
- **Longest Common Subsequence (LCS)** [41] ist die längste gemeinsame Abfolge von Zeichen zwischen zwei Zeichenketten.
- **Longest Common Substring (LCSS)** [470] ist die längste gemeinsame Abfolge von Zeichen zwischen zwei Zeichenketten ohne Änderung von Zeichen. Bei den Zeichenketten „Temperatur“ und „Temporär“ wäre der LCS „Tempr“ und der LCSS „Temp“.
- **„Prefix“** ist die Überschneidung von zwei Zeichenketten jeweils von Beginn der Zeichenkette [84]. Der Prefix liefert für die tendenziell hierarchisch aufgebauten GAS-Datenstromnamen die größte gemeinsame Hierarchie.

2.3.4.2 Token-basierte Verfahren

Token-basierte Verfahren berechnen die Ähnlichkeit anhand von Häufigkeit von Zeichen /Token:

- Die **Jaro-Winkler-Distanz** [92, 143, 231] ist eine modifizierte Form der Jaro Distanz und arbeitet mit einem maximalen Übereinstimmungsbereich zweier Zeichenketten [267]. Sie überbewertet den Wortanfang [331], was bei hierarchisch strukturierten Datenstromnamen der Gebäudeautomation von Vorteil ist.
- Die **Cosinus-Distanz** misst, wie nahe sich zwei Zeichenketten-Vektoren a und b in $\mathbb{R}^n$ sind, basierend auf dem Cosinus des Winkels zwischen a und b auf Basis der Häufigkeit von Zeichenketten [82, 441]. Der Ochiai-Koeffizient ist eine mathematische Vereinfachung zur Berechnung der Cosinus-Ähnlichkeit und liefert äquivalente Ergebnisse [509] und wird deshalb in dieser Arbeit verwendet und als Cosinus-Distanz bezeichnet.

2.3.4.3 Phonetische Verfahren

Phonetische Verfahren basieren weitestgehend auf der Ähnlichkeit der Aussprache zweier Zeichenketten), wofür vier relevante deutsche Verfahren existieren, die im weiteren und in Abbildung 2.5 näher ausgeführt werden:

- **Kölner Phonetik** [360] ordnet analog zum im angelsächsischen Raum verbreiteten Soundex [190, 242] jedem der 26 Buchstaben eine Ziffer von 0 bis 9 zu.
- **Michael** [317] hat eine andere phonetische Darstellung entwickelt, die sich stärker an der Aussprache orientiert. So werden die Buchstaben „ch“ (wie in „Christopher“) und „k“ in den gemeinsamen Buchstaben „k“ umgewandelt.
- Die **erweiterte Kölner Phonetik** von Haase [167] ist eine für den Einsatz bei Namensdatenbanken optimierte Variante [393] der ursprünglichen Kölner Phonetik.
- **Reth u. Schek** [387] entwickelten ein anderes, auf Buchstaben basierendes System.

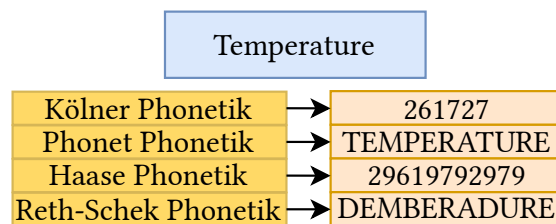


Abbildung 2.5: Die vier verwendeten phonetischen Verfahren aufgezeigt am Beispielwort „Temperature“.

2.4 Zusammenfassung Forschungsbedarf

Für die automatische Extraktion von Metadaten in Gebäudeautomationssystemen wurden zahlreiche Methoden entwickelt. Bisherige Ansätze betrachten überwiegend entweder Zeitreihen oder Zeichenketten von Datenströmen, während hier beide verwendet werden. Die bisherigen Ansätze beziehen sich hauptsächlich auf eine Sprache (Englisch), während in dieser Arbeit mit einem multilingualen Datensatz (Deutsch, Englisch) gearbeitet wird. Die deutsche Sprache weist Besonderheiten wie Komposita (z.B. „Temperatursensor“ statt „temperature sensor“) auf, die in dieser Arbeit näher betrachtet werden. Auch die zugrundeliegenden Metadaten schemata wurden nur selten einem großskaligen Praxistest der enthaltenen Metadaten unterzogen. Bisherige Ansätze beschränkten sich bis auf wenige Ausnahmen auf wenige Gebäude, während in dieser Arbeit Daten aus sechs Datensätzen mit insgesamt 250 Gebäuden untersucht werden, wovon 150 Gebäude im Rahmen des Dissertationsprojektes OOM4ABDO mit einer eigenen Monitoringlösung aufgenommen wurden. Dies führte zu einem bisher nur selten erreichte Bandbreite der Metadaten. Es ist

zu untersuchen, ob mithilfe eines größeren Datensatzes über mehrere Betreiber hinweg qualitativ sinnvolle Ergebnisse bei der Metadatenextraktion und der Anwendungszusammenführung erzielt werden.

Es gibt nur wenige Analysen von zusammenhängenden Gebäudeenergiesystemen und nur wenige echte Skalierungstests, die sich auf die Verbindungen zwischen RLT, VAV und Räumen konzentrieren. Auf der thermischen Seite gibt es jedoch mehr Verbindungsarten (z. B. Energieumwandlung zur Energieverteilung) und verwendete Medien (z.B. Wasser als Übertragungsmedium). Diese wurden bisher in der Wissenschaft nicht berücksichtigt. Auch eine Übertragung dieser Methoden auf Gebäude und Quartiere steht noch aus. Diese Lücke soll mit dieser Arbeit geschlossen werden. Es soll untersucht werden, inwiefern ein mediumsunabhängiger, skalierbarer Ansatz der Topologieerkennung auf Basis von Zeitreihen entwickelt werden kann.

In den bisherigen Beiträgen wurde die Informationsqualität (IQ) im Rahmen eines Informationsextraktionsprozesses (IE-Prozess) nicht berücksichtigt. Wenn übergeordnete Ebenen betrachtet wurde, dann wurden tendenziell die Daten-Plattformen oder einzelne Methoden betrachtet [337]. Im Rahmen dieser Arbeit wird ein IE-Prozess für Metadaten von Energiesystemen in Gebäuden und Quartieren entworfen, der insbesondere die IQ-Kriterien und deren Auswirkungen untersucht. Die automatische Übertragung der Metadatenerkennung auf Anwendungen ist erst ansatzweise erfolgt, wobei der Schwerpunkt auf der Erkennung von Fehlern in RLT-Geräten liegt. Es gibt jedoch eine Vielzahl von Anwendungen, die zwar einzeln untersucht wurden, um Verbesserungen im Betrieb der technischen Systeme zu erreichen, aber nicht daraufhin überprüft wurden, ob die Informationsextraktion aus den Metadaten die für die Nutzung erforderlichen Informationen liefern kann. Diese Lücke soll in Teilen in dieser Arbeit geschlossen werden. Für diesen Zweck werden acht Anwendungen betrachtet, die jeweils die Energieeffizienz in Gebäuden und Quartieren entscheidend beeinflussen können. Der Untersuchungsgegenstand ist, inwiefern eine anwendungsorientierte Metadatenextraktion und Zusammenführung von Datenströmen zu Anwendungen möglich ist. Im IE-Prozess werden verschiedene IQ-Kriterien auf ihre Anwendbarkeit in GAS untersucht. Sind verschiedene Kriterien sinnvoll anzuwenden oder überschneiden sich inhaltlich mehrere Kriterien?

Methodisch wurden bei Zeitreihen viele Ansätze mit Varianten von Entscheidungsbäumen mit klassischen Merkmalen wie Minimum und Maximum untersucht. Neuartige Methoden auf Basis von neuronalen Netzen waren tendenziell unterrepräsentiert. Im Bereich der Zeichenketten wurden zwar auch neuartigere Methoden wie auf Transformern basierenden BERT-Varianten genutzt, aber selten wurden sprachspezifische und domänenspezifische Merkmale für die Klassifizierung genutzt. Zeichenketten-Distanzen waren in einigen Fällen erfolgreich für das Clustering, aber nicht als Eingang für die Klassifizierung. Es wurde meistens nur eine Distanz verwendet und nicht mehrere auf ihre Nutzbarkeit untersucht.

Die gesamte in dieser Dissertation zu untersuchende Fragestellung lautet wie folgt: „Wie kann für multilinguale und heterogene Gebäudeautomationssysteme ein skalierbarer, anwendungsorien-

tierter Informationsextraktionsprozesses zur automatischen Metadatenextraktion aus Zeitreihen- und Zeichenkettendaten entwickelt werden, wobei konkrete Anwendungen zur Energieeinsparung betrachtet werden?“.

Aus dieser Fragestellung ergeben sich die folgenden Teilaufgaben:

- Verarbeitung gleichermaßen von Zeitreihen- und Zeichenkettendaten zur Metadatenextraktion,
- Daten aus multilingualen und heterogenen Gebäudeautomationssystemen gleichermaßen skalierbar zu verarbeiten,
- Unterstützung verschiedener Anwendungen zur Energieeinsparung, inklusive einer mediumsunabhängigen Topologieerkennung,
- Bewertung eines Informationsextraktionsprozesses für Daten von Gebäudeautomationssystemen.

3 Methodik

In diesem Kapitel wird der methodische Rahmen zur automatisierten Informationsextraktion aus Gebäudeautomationssystemen vorgestellt. Die automatisierte Optimierung des Energieeinsatzes in Gebäudeautomationssystemen erfordert einen skalierbaren Prozess, um aus heterogenen Rohdaten verwertbare Informationen zu extrahieren. Die Herausforderung liegt in der Heterogenität und mangelnden Strukturierung der Daten, die eine direkte Anwendung energieoptimierender Algorithmen erschweren. Der hier entwickelte Ansatz folgt einem systematischen Prozess, der von der Auswahl eines geeigneten Metadatenschemas über die Extraktion von Informationen aus Zeichenketten und Zeitreihen bis hin zur praktischen Anwendung in Energiesystemen reicht.

Die Methodik gliedert sich in sechs aufeinander aufbauende Schritte:

- Kapitel 3.1: der Informationsextraktionsprozess und seine Besonderheiten in der Gebäudeautomation,
- Kapitel 3.2: Auswahl eines Metadatenschemas,
- Kapitel 3.3: Verfahren zur Metadatenextraktion auf Basis von Zeichenketten der Datenströme inklusive Vorverarbeitung der Blöcke,
- Kapitel 3.4: Verfahren zur Metadatenextraktion auf Basis von Zeitreihenerkennung,
- Kapitel 3.5: der untersuchte Datensatz,

- Kapitel 3.6: acht Anwendungen zur Optimierung des Energieeinsatzes in Energiesystemen aus den drei Bereichen („Fehler- und Anomalieerkennung“, „Regelung“ und „Energiesystemanalyse“)

3.1 Informationsextraktion

Für die Anwendung von modernen Anwendungen im Gebäude fehlt für die skalierbare Anwendung in weiten Teilen ein adäquates Wissen über Informations- und Kommunikationstechnik (IKT) Technologien [444] und auch das passend ausgebildete an die sich steigenden Anforderungen anpassendes Personal [10] zur Implementierung der Strategien. Das Personal für die Ausführung von komplexen KI-Algorithmen in Verbindung von technischen Anlagen fehlt massiv [274], insbesondere in der Energietechnik [6, 7]. Deshalb bietet der in dieser Arbeit definierte automatisierte Informationsextraktionsprozesse Potential den benötigten Personalaufwand zu senken, wobei jedoch den erhöhten Anforderungen an das benötigte Personal Sorge getragen werden muss.

Im Bereich der Informationsextraktion in Energiesystemen werden verschiedene Begriffe mit unterschiedlichen Interpretationen verwendet [228]. Die im Rahmen dieser Arbeit verwendeten Begriffe „Daten“ und „Information“ werden in Kapitel 3.1.1 erklärt und in Kapitel 3.1.2 passende GAS-Datentypen definiert.

Die Begriffe Datenqualität und Informationsqualität und deren Bewertung werden in den Kapiteln 3.1.3 und 3.1.4 näher betrachtet und dieses Kapitel wird mit einer Gesamtbetrachtung des Informationsextraktionsprozesses in Kapitel 3.1.5 abgeschlossen.

3.1.1 Daten und Information

Die Begriffe Daten und Information werden häufig äquivalent oder ähnlich verwendet [512]. Nach Zins [512] sind Daten, Informationen und Wissen sequentiell geordnet. Daten sind hierbei das „Rohmaterial“, das durch eine Verarbeitungsfunktion zu Informationen werden. Und Informationen sind wiederum das „Rohmaterial“, das zu Wissen verarbeitet werden kann.

Die jeweilige Qualität des Rohmaterials (Daten, Informationen) bestimmt die Qualität des Produkts der Verarbeitungsfunktion [156, 479]. Die Verarbeitungsfunktion von Daten zu Informationen wird als Informationsextraktionsprozess (IE-Prozess) bezeichnet, der in Kapitel 3.1.5 näher beschrieben wird.

3.1.2 Datentypen

Im Folgenden werden die Begriffe für Datentypen definiert, die in dieser Arbeit verwendet werden.

- Eine **Observation** ist eine einzelne Beschreibung des Zustands (z.B. Temperatur) eines Systems (z.B. RLT-Anlage) aus einer Datenquelle (z.B. Temperatursensor) zu einem Zeitpunkt.

- Ein **Datenstrom** ist ein Strom von Observationen aus einer Datenquelle (z.B. Sensor, Verarbeitungsfunktion oder Parameter), der semantisch von anderen Datenströmen getrennt ist.
- Ein **virtueller Datenstrom** ist ein kombinierter Datenstrom als Ergebnis einer Verarbeitungsfunktion aus mehr als einem Datenstrom [110, 493].
- Ein **simulierter Datenstrom** ist ein (Zwischen-)Ergebnis eines Simulationsmodells, das ein Black-Box-Modell, Grey-Box-Modell oder White-Box-Modell sein kann [279, 292, 408, 428, 496].

3.1.3 Datenqualität

Die Datenqualität (DQ) wird als „Gesamtheit der Ausprägungen von Qualitäts[dimensionen] eines Datenbestandes [zur Zweckeignung]“ [157] definiert. Die Zweckeignung entspricht der Fähigkeit festgelegte Anwendungen zu unterstützen (engl. „fit for use“).

Die Datenqualität wird im Allgemeinen in verschiedenen Dimensionen gemessen [127]. So definiert die Data Management Association (DAMA) [23] die folgenden sechs primäre Dimensionen für die Datenqualität:

- **Vollständigkeit (Completeness)**: Der Anteil der gespeicherten Daten im Vergleich zum Potenzial von „100 % vollständig“.
- **Singularität/Einzigartigkeit (Uniqueness)**: Keine Instanz ist doppelt im Datensatz enthalten.
- **Einheitlichkeit (Consistency)**: Das Fehlen von Unterschieden, wenn zwei oder mehr Darstellungen einer Sache mit einer Definition verglichen werden.
- **Aktualität (Timeliness)**: Der Grad, in dem die Daten die Realität zum gewünschten Zeitpunkt darstellen.
- **Gültigkeit (Validity)**: Daten sind gültig, wenn sie mit der Syntax (Format, Typ, Bereich) ihrer Definition übereinstimmen.
- **Genauigkeit (Accuracy)**: Das Ausmaß, in dem Daten das beschriebene Objekt oder Ereignis in der „realen Welt“ korrekt beschrieben wird.

Diese sind in der Literatur in einer ähnlichen Weise definiert [37, 38, 338, 417, 418, 454].

Weiterführende Aspekte wie Aktualität, Gültigkeit und Genauigkeit werden in der vorliegenden Arbeit aufgrund des Anwendungsfalls in der Gebäudeautomation nicht im Detail betrachtet. Die Verzögerungen einer Temperatur-Messung, die entscheidend für die Bewertung der Aktualität sind, wird durch einen komplexen Prozess von Signalverarbeitungen (Spannung, Messaufnahme etc.) definiert, die nicht Bestandteil dieser Dissertation sind. Die Gültigkeit einer Messung könnte nur durch eine ausreichende Anzahl redundante Messungen unabhängiger Messgeräte überprüft werden, die in GAS nicht üblich sind. Die Gültigkeit eines Datenformats kann durch ein einheitliches

Metadatenschema garantiert werden. Die Genauigkeit wird durch die Qualität der Messung bestimmt [70, 272, 313, 448]. Keine der ausgeschlossenen Eigenschaften beeinflusst den IE-Prozess in einem berücksichtigungswerten Maße bzw. kann nicht mit angemessenen Mitteln berücksichtigt werden. Im Allgemeinen wird von einer menschlichen Fehlerquote bei der Datenbeschriftung von technischen Klassen (Sensordaten bzw. Fehlererkennung) von 1 bis 15 % ausgegangen [166, 241, 264, 293, 349]. Diese Fehlerquote hängt insbesondere von der Expertise und dem Stresslevel der menschlichen Klassifizierer ab [166, 264]. Im Folgenden wird diese 15 % Fehlerquoten-Grenze auch als Ziel-Grenze für den maschinellen Algorithmus verwendet.

Die drei ausgewählten Eigenschaften werden im folgenden näher erläutert:

- **Vollständigkeit (Completeness):** Eine hohe Datenqualität ist essenziell für skalierbare Energieanalysen [172]. Beim Erfassen von Sensordaten in der Gebäudeautomation können auf dem Weg von der Feldebene über Datenlogger zur Speicherung Verbindungsabbrüche, Softwarefehler oder Stromausfälle zu Datenlücken führen [31, 329]. Auch die Frequenz der Datenerfassung spielt eine Rolle: Während elektrische Messwerte häufig erfasst werden müssen, genügen für langsamer reagierende thermische Systeme oft geringere Frequenzen [198, 432, 466]. Dabei führt eine höhere Messfrequenz zu einem stark erhöhten Datenvolumen [399], was wiederum ökologische Auswirkungen hat [59, 150, 206, 209, 216, 227, 251, 326].
- **Singularität/Einzigartigkeit (Uniqueness):** In der Gebäudeautomation werden oft identische Sensoren (z.B. für Außentemperatur) mehrfach oder sogar gebäudeübergreifend eingesetzt. Zudem können Datenströme dupliziert oder umbenannt werden, wodurch Mehrdeutigkeiten entstehen [71, 99]. Es ist wichtig sicherzustellen, dass nur ein eindeutiger, korrekter Datenstrom genutzt wird und Kopien erkannt werden.
- **Einheitlichkeit (Consistency):** Eine klare und einheitliche Beschreibung der Daten ist entscheidend. Jeder Datenstrom muss einer einzigen, eindeutigen Datenklasse zugeordnet sein. Die Konsistenz und Eindeutigkeit in der Klassifizierung der Datenströme muss daher im Datensatz überprüft werden.

Für die Analysen werden auf Basis dieser Eigenschaften folgende Maßnahmen getroffen, um die Datenqualität sicherzustellen:

- Für Zeitreihenanalysen in Kapitel 4.2 keine fehlerhaften Sensordaten,
- Ausschluss von Datenströmen mit Datenlücken am Jahresanfang/-ende,
- Ausschluss von Datenströmen ohne Änderungen im Jahresverlauf,
- Für Anwendungen in Kapitel 4.3 nur gut beschreibbare Daten,
- Ausschließliche Annotation eindeutiger Metadaten; Mehrdeutigkeiten (z.B. „MW“) werden kontextabhängig aufgelöst,

- Singularität wird durch ein systematisches Auswahlverfahren, sichergestellt, bei dem Kopien der Datenstromdaten gemeinsam in Trainings- und Testdaten eingeteilt werden.
- Die Einheitlichkeit basiert auf einem einheitlichem Metadatenschema, das in Kapitel 3.2 ausgewählt wird.

3.1.4 Informationsqualität

Der Begriff Informationsqualität (IQ) wird sehr ähnlich wie der Begriff Datenqualität verwendet, wobei die Bedeutung stark vom Datenmodell und dessen Anwendung abhängt.

Die Definition der IEC International Electrotechnical Commission [203] beschreibt Informationsqualität als „das Ausmaß, in dem die Eigenschaften von Daten erklärte und implizierte Anforderungen erfüllen, wenn sie unter bestimmten Bedingungen verwendet werden.“ und liefert „ein allgemeines Datenqualitätsmodell für Daten, die in einem strukturierten Format innerhalb eines Computersystems aufbewahrt werden.“ Batini u. Scannapieco [37] gibt einen Überblick über die unterschiedlichen Definitionen von Modelleigenschaften durch verschiedene Autoren. Diese weichen teilweise stark von der Definition von IEC International Electrotechnical Commission [203] ab, jedoch überlagern sich auch Eigenschaften der Datenqualität mit der Informationsqualität.

Kenett u. Shmueli [233] grenzen die Informationsqualität ab als das Potenzial eines Datensatzes zur Erreichung eines bestimmten Ziel mit Hilfe einer bestimmten empirischen Analysemethode zu erreichen [93]. Auch Miller [319] definiert den Nutzen für das Ziel als zentral für die Informationsqualität.

Auf Basis dieser Eigenschaften wird die folgende Definition für Informationsqualität verwendet: „Informationsqualität ist die Fähigkeit von prozessierten Daten definierte Anforderungen zur Wissensextraktion zu erfüllen“.

Die Informationsqualität ist wie die Datenqualität mehrdimensional [9] und deren Definition und Abgrenzung ist je nach Quelle unterschiedlich [37, 124, 203, 262, 495]. Kenett u. Shmueli [232] definiert wiederum acht eigene Dimensionen, wovon aber zwei für die reale Implementierung von Informationsqualität entscheidend sind: Generalisierbarkeit und Operationalisierung. Die Generalisierbarkeit bedeutet, dass eine Informationsart nicht nur in Datensatz A, sondern auch in Datensatz B auffindbar ist und der IE-Prozess übertragbar ist. Die Operationalisierbarkeit bedeutet, dass die Definition der Dimensionen der Informationsqualität die Entscheidungsqualität bestimmt, die auf Basis dieser Informationen getroffen wird [156]. Der automatisierte Informationsqualitätsprozess soll immer ein Gleichgewicht zwischen Vollständigkeit und Genauigkeit herstellen [9, 37]. Jede Kennzahl für die Informationsqualitätsdimensionen ist biasbehaftet [363], weshalb mehrere Kennzahlen für eine gesamte Betrachtung der Informationsqualität verwendet werden sollten.

Daher ist es notwendig, zunächst auszuwählen, welche Informationsqualitätsmerkmale im Prozess verwendet werden sollen. Es ist auch wichtig, dass jedes Merkmal quantitativ abgebildet wer-

den kann [179]. Die Auswahl der Informationsqualitätsmerkmale soll im Folgenden genauer beleuchtet werden.

Im Kontext der Analyse von Gebäude(automations)daten sind auf Basis der Anforderungen an die Anwendungen in der Energietechnik die folgenden Dimensionen der Informationsqualität relevant:

- **Genauigkeit:** Exakte Zuordnung von Datenstromtypen verhindert fehlerhafte Modellierung und Alarmer in GAS.
- **Vollständigkeit:** Erhöht die Nutzbarkeit von extrahierten Metadaten und verringert die benötigten Nutzereingaben.
- **Ausgeglichenheit:** Wenn Metadaten-Klassen unterschiedliche Anzahl an Instanzen aufweisen, verbessert ein ausgeglichener Algorithmus die Nutzbarkeit bei seltenen Klassen.
- **Konformität:** Bei Nutzung unterschiedlicher Algorithmen zur Metadatenextraktion sind konforme Ergebnisse für die automatische Nutzung dieser erforderlich.
- **Zuverlässigkeit:** Wenn ein Algorithmus eine gewisse Klasse mit einer gewissen Wahrscheinlichkeit vorhersagt, garantiert ein zuverlässiger Algorithmus, dass diese Wahrscheinlichkeit im Schnitt auch in Realität auftritt.
- **Generalisierbarkeit:** Wenn ein Algorithmus auf einem Datensatz A trainiert wird, stellt die Generalisierbarkeit sicher, dass er auch auf Datensatz B sinnvolle Ergebnisse liefert.
- **Operationalisierbarkeit:** Bei Nutzung von Anwendungen zur Energieeinsparung ist es wichtig die benötigten operativen Datenströme für diese Anwendungen zu extrahieren.
- **Effektivität:** Der Ressourcenverbrauch sollte den Ergebnissen eines Algorithmus angemessen sein, also effektiv.

Diese Dimensionen werden im Folgenden erläutert.

3.1.4.1 Genauigkeit

Die Genauigkeit in der Informationsqualitätsforschung wird auch häufig als Fehlerfreiheit oder Richtigkeit bezeichnet [392]. Die Genauigkeit ist bei einem Algorithmus die Fähigkeit aus einem Datensatz von Instanzen die Metadaten so zu extrahieren, dass die Instanzen möglichst oft richtig kategorisiert werden.

Die Abbildung 3.1 stellt verschiedene ML-Scores vor. Die „Genauigkeit“ als ML-Score (engl. Accuracy) wird wie in Gleichung 3.1 dargestellt als Vergleich zwischen der Summe der wahren positiven und wahren negativen Ergebnisse im Vergleich zur Grundgesamtheit.

$$Gen = \frac{WP + WN}{WP + FP + WN + FN} \quad (3.1)$$

Die Bewertung der Genauigkeit hängt von der Verteilung der Kategorien innerhalb der Grundgesamtheit ab, weshalb die „ausgeglichene Genauigkeit“ hier eine Alternative sein kann.

	Wahrer Zustand positiv	Wahrer Zustand negativ	F1 Wert = $2 \times \frac{\text{Relevanz} \times \text{Sensitivität}}{\text{Relevanz} + \text{Sensitivität}}$
Vorhergesagter Zustand positiv	Wahr positiv (WP)	Falsch positiv (FP)	Relevanz = $\frac{WP}{WP+FP}$
Vorhergesagter Zustand negativ	Falsch negativ (FN)	Wahr negativ (WN)	Neg. Vorh. Wert = $\frac{WN}{FN+WN}$
	Sensitivität = $\frac{WP}{WP+FN}$	Spezifität = $\frac{WN}{FP+WN}$	Genauigkeit = $\frac{WP+WN}{WP+FP+FN+WN}$

Abbildung 3.1: Klassische Informationsqualitätskriterien der Klassifikation auf Basis des wahren und vorhergesagten Zustands.

3.1.4.2 Vollständigkeit

Für die Vollständigkeit ist entscheidend, dass auch alle in einer Klasse positiven Instanzen eines Datensatzes korrekt erkannt werden. Für die Bewertung dieser Vollständigkeit wird eine Bewertung auf der Grundlage von Relevanz (engl. Precision, Formel 3.2) und Sensitivität (engl. Recall, Formel 3.3) vorgenommen. Der Relevanz-Wert gibt an, wie viele der als positiv vorhergesagten Objekte tatsächlich positiv sind, der Sensitivität-Wert gibt an, wie viele der tatsächlich positiven Objekte auch als positiv vorhergesagt wurden.

$$\text{Rel} = \frac{WP}{WP + FP} \quad (3.2)$$

$$\text{Sen} = \frac{WP}{WP + FN} \quad (3.3)$$

Der F1-Wert, der in Formel 3.4 ist das harmonische Mittel aus Relevanz und Sensitivität. Allgemeiner ausgedrückt kann der F_beta-Wert ein gewünschtes Verhältnis von Relevanz und Sensitivität darstellen. Er ist biasbehaftet anhand der Anzahl der Instanzen jeder verwendeten Klasse. Das bedeutet, wenn eine Metadatenklasse wie eine Sensorart dominiert, dass seltene Klassen wie spezielle Sollwerte schlechter erkannt werden.

$$F1 = 2 \cdot \frac{\text{Rel} \cdot \text{Sen}}{\text{Rel} + \text{Sen}} \quad (3.4)$$

3.1.4.3 Ausgeglichenheit

Insbesondere für unausgeglichene Datensätze kann der Matthew-Korrelationskoeffizient (MCC) [310] eine Alternative sein, der in Gleichung 3.5 definiert wird. Der Wert kann zwischen -1 (kein richtiger Wert), 0 (Zufallsverteilung) und +1 (vollständige Übereinstimmung) liegen. Chicco u. Jurman [83] empfehlen ihn vor allem als Vergleichswert für binäre Klassifikationen. Hier können die absoluten Werte der Genauigkeit oder des F1-Wertes aufgrund der genannten Schwächen nur einen schlechten Hinweis auf die Qualität eines Algorithmus geben. Hier ist der Wert des MCC besser zu interpretieren und berücksichtigt die Verteilung der Grundgesamtheit.

$$\text{MCC} = \frac{WP \cdot WN - FP \cdot FN}{\sqrt{(WP + FP) \cdot (WP + FN) \cdot (WN + FP) \cdot (WN + FN)}} \quad (3.5)$$

3.1.4.4 Konformität

Für die Konformität von zwei ML-Algorithmen mit unterschiedlichen extrahierten Merkmalen und ordinalen Metadaten wird im Allgemeinen ein statistisches Maß für die Interrater-Reliabilität verwendet [455]. Cohens Kappa [91] ist hierfür ein im ML beliebtes Leistungsmaß [455] und ist in Formel 3.6 definiert.

$$\kappa = \frac{p_a - p_e}{1 - p_e} \quad (3.6)$$

p_a ist hierbei die tatsächliche Verteilung der Bewertungen und p_e die erwartete Verteilung der Bewertungen bei komplett zufälliger Auswahl der Bewertungen. Bei mehr als zwei Kategorien der Metadaten (z.B. Temperatur, Feuchtigkeit und CO₂-Wert) werden diese Kategorien in binäre Variablen („wahr“, „falsch“) umgewandelt [394]. Es können bei einem Interrater entweder alle Merkmale betrachtet werden oder alle Merkmale, die in einer Instanz vorkommen. Beispielsweise hat der Interrater A die Klassen 1 und 20 vorhergesagt und Interrater B die Klasse 1. Bei in unserem Beispiel 20 im Datensatz enthaltenen Klassen, können bei der Bewertung nur die Klassen 1 und 20 einzeln oder alle 20 Klassen bewertet werden. Diese Unterteilung findet auch hier statt.

Falls die Grundgesamtheit bei zwei Interratern unterschiedlich ist, wird das Verhältnis der Interrater mit dem Cohens Kappa der Grundgesamtheiten verglichen (Cohens Kappa Ratio). Cohens Kappa kann einen Wert von -1 (total konträre Einschätzungen) über 0 (Zufall) bis +1 (absolute Übereinstimmung) annehmen. Für die Bewertung von Cohens Kappa kann die folgende Tabelle von Landis u. Koch [258] verwendet werden, wobei die Bewertung „Schlecht“ nach Altman [18] schon bei einem Kappa unter 0,20 verwendet werden kann:

Tabelle 3.1: Bewertung von Cohens Kappa anhand seines Wertes.

Kappa Statistik	Stärke der Übereinstimmung
< 0,00	Schlecht (orig. poor)
0,00 - 0,20	Geringfügig (orig. slight)
0,21 - 0,40	Angemessen (orig. fair)
0,41 - 0,60	Mäßig (orig. moderate)
0,61 - 0,80	Erheblich (orig. substantial)
0,81 - 1,00	Fast perfekt (orig. poor)

3.1.4.5 Zuverlässigkeit

Nutzer müssen sich auf die gewonnenen Informationen aus einem System zuverlässig verlassen können und diesem vertrauen [211]. In der Kategorie Zuverlässigkeit wird überprüft, ob ein Algorithmus seine internen Wahrscheinlichkeiten für Klassen (z.B. zu 80 % Klasse Temperatursensor) zuverlässig vorhersagen kann. Somit bietet dieses eine Möglichkeit zu überprüfen, ob die Wahrscheinlichkeitswerte eines Algorithmus direkt als Wahrscheinlichkeit zu einer Klasse zugeordnet werden können

Zur Bewertung der Vertrauenswürdigkeit der Algorithmen und ihrer Anwendbarkeit im Kontext der Gebäudeautomation werden Kalibrierungskurven (auch Zuverlässigkeitsdiagramme) [256] eingeführt. Diese betrachten die Algorithmen und zeigen, ob die Algorithmen die Wahrscheinlichkeit einer Klasse korrekt vorhergesagt haben. Dazu werden die vorhergesagten Wahrscheinlichkeiten für die richtige Klasse mit dem tatsächlichen Anteil der richtigen Klasse verglichen. Für den Fall einer binären Klassifikation wird der erwartete Kalibrierungsfehler (engl. Expected Calibration Error, ECE) verwendet (siehe Gleichung 3.7). ECE berechnet die durchschnittliche absolute Differenz zwischen der Genauigkeit im Vergleich zur vorhergesagten Wahrscheinlichkeit (Vertrauen). Für kritische Anwendungen kann der maximale Kalibrierungsfehler (engl. Maximum Calibration Error, MCE) verwendet werden (siehe Gleichung 3.8). Dieser zeigt die Wahrscheinlichkeitsbereiche an, in denen eine hohe Abweichung zwischen Genauigkeit und Vertrauen besteht.

$$ECE = \sum_{m=1}^M \frac{|B_m|}{n} |\text{Gen}(B_m) - \text{Vertr}(B_m)| \quad (3.7)$$

Die Wahrscheinlichkeit wird in Behälter (engl. Bins) unterteilt (in unserem Fall 1 % bzw. 5 % große Bins) und für jedes Bin wird analysiert, wie groß der Unterschied ist.

$$MCE = \max_{m \in \{1, \dots, M\}} |\text{Gen}(B_m) - \text{Vertr}(B_m)| \quad (3.8)$$

Für die automatische Extraktion der Informationen muss ein Konfidenzniveau erreicht werden, so dass dem Algorithmus ohne weitere Nutzereingabe vertraut wird. Der *MCE* wird in Erfahrung aus GAS meist in den unteren Bereichen (<80 % Score) ermittelt [431], weshalb dieser in der Bewertung

der automatisierten Zuverlässigkeit nur eine sehr untergeordnete Rolle spielt.

Wichtig für die automatisierte Bewertung sind Scores »90 %, da diese die ungefähre menschliche Fehlerquote von 1 % bis 15 % widerspiegeln [166, 241, 264, 293, 349].

3.1.4.6 Generalisierbarkeit

Die Generalisierbarkeit der Algorithmen wird getestet, indem sie auf den Daten eines Gebäudekonglomerats trainiert und auf einem anderen Gebäudekonglomerat getestet werden. Die durchschnittliche Generalisierbarkeit GBK_{av} wird als normalisierte Summe der jeweiligen Informationsqualitäten betrachtet und die maximale Generalisierbarkeit GBK_{max} als maximaler Wert über alle Gebäudepaare.

$$GBK_{av,j} = \sum_{i=1}^M \frac{IQ_j}{M} \quad (3.9)$$

Hierbei werden die folgenden Variablen verwendet:

n : Anzahl an Informationsqualitätsdimensionen

i : Index für das jeweilige Gebäude A (Training) auf Gebäude B (Testen) M : Anzahl von Gebäudepaaren j : Index der Informationsqualitätsdimension

$$GBK_{max,j} = \max_{i=1}^M IQ_j \quad (3.10)$$

3.1.4.7 Operationalisierbarkeit

Die „Operationalisierbarkeit“ deckt ab, inwieweit die ermittelten Qualitätsmetriken unmittelbar zur Verbesserung der Nutzbarkeit von Anwendungen beitragen können. Dies ist insbesondere bei der Bewertung von Metadatenextraktions-Algorithmen entscheidend, ob diese eine Anwendung mit Metadaten unterstützen können. Deswegen wird ein anwendungsgewichteter ML-Score eingeführt, wobei die folgenden Symbole verwendet werden:

i : Index für den Datenstromtyp

j : Index für den Anwendungstyp

k : Index für Algorithmus

c_j : Korrekturfaktor für Anwendungstyp j

m_f : Anzahl Unteranwendungen

m_d : Anzahl Datenstromtypen

$n_{i,j}$: Anzahl von benötigten Datenströmen von Typ i benötigt bei Anwendungstyp j

$r_{i,k}$: Ergebnis von Algorithmus k zur Klassifizierung bei Datenstromtyp i

r_k : Gesamtergebnis von Algorithmus k zur Klassifizierung für die Anwendung

$$r_k = \frac{\sum_i^{m_d} \sum_j^{m_f} r_{i,k} \cdot n_{i,j} \cdot c_j}{\sum_i^{m_d} \sum_j^{m_f} n_{i,j} \cdot c_j} \quad (3.11)$$

3.1.4.8 Effektivität

Die Messung der Effektivität eines Informationsextraktionsprozesses ist schwierig und variiert je nach Szenario. Blasch u. a. [52] verwenden die Formel 3.12 zur Bestimmung der Effektivität:

$$\text{Effektivität} = \text{Informationsgewinn} \cdot \text{Fähigkeit} \cdot \text{Robustheit} \quad (3.12)$$

Nach der Herleitung in Kapitel A.2 kann die Effektivität wie folgt definiert werden.

$$\text{Effektivität} = X \cdot \frac{(I_{\text{post}} - I_{\text{base}})^2}{I_{\text{base}}} \cdot \frac{1 - 3 \cdot \sigma_I}{I_{\text{av}}} \quad (3.13)$$

Bei dieser Definition werden die folgenden Bestandteile und Definitionen verwendet:

- I_{base} ist hierbei ein Informationsstand bei einem Basisalgorithmus,
- I_{post} nach Durchführung des IE-Algorithmus,
- I_{av} : mittleres IQ-Abschneiden des Algorithmus,
- σ_I die Standardabweichung der Algorithmen,
- Der Korrekturfaktor X wird auf 1 gesetzt, da dieser real nur durch menschliche Experimente gewonnen werden kann.

Insgesamt kann die Effektivität nicht nur ein einzelnes Informationsqualitätskriterium absolut bewerten, sondern auch in Relation zu seiner Basis, was einen wertvollen neuen Einblick in den IE-Prozess liefern kann.

3.1.4.9 Zusammenfassung

Zusammenfassend bestimmt die Informationsqualität die Fähigkeit eines Informationsextraktionsprozesses, aus Rohdaten konsistente, vollständige und verlässliche Informationen zu generieren. Die hier definierten Dimensionen, von der Genauigkeit über Vollständigkeit, Ausgeglichenheit, Konformität und Zuverlässigkeit bis hin zu Generalisierbarkeit, Operationalisierbarkeit und Effektivität, bilden die Grundlage, um den Erfolg und die Übertragbarkeit der im IE-Prozess eingesetzten Methoden zu evaluieren. Als zentrale IQ-Kriterien im Bezug zur Analyse und Regelung der

Gebäudeautomation werden die Vollständigkeit (F1-Score), die Ausgeglichenheit (MCC) und die Zuverlässigkeit (ECE) betrachtet. Diese bieten ein gutes Bild über die Fähigkeit eines Algorithmus sinnvolle Beiträge für Anwendungen zu liefern. Die weiteren Qualitätsmerkmale werden am Ende einer Prüfung unterzogen, inwiefern diese den Prozess weiterhin unterstützen können.

3.1.5 Informationsextraktionsprozess

Der Informationsextraktionsprozess ermöglicht es Rohdaten in Informationen zu überführen, mit denen dann Entscheidungen umgesetzt werden können. Zwar stammen die klassischen Konzepte aus dem militärischen Bereich, der viele Datenquellen zusammenführen muss, aber sie können auch auf die Gebäudeautomation übertragen werden. Es ist entscheidend, dass im ersten Schritt ein Datenqualitätsprozess durchgeführt wird [302], um den manuellen Aufwand zu reduzieren [30], und der ein Bestandteil eines Risikomanagements ist [30].

Es existieren verschiedene Ansätze für einen IE-Prozess [73, 315], die im folgenden aufgelistet werden:

- Joint Directors of Laboratories (Joint Directors of Laboratories (JDL))-Architektur [472]
- Data Fusion Information Group (DFIG)-Modell [48]
- Cross Industry Standard Process for Data Mining (CRISP-DM) [414]
- Dasarathy-Modell [73]
- Durrant-Whyte [73]
- ETL-Prozesse (Extract-Transform-Load)
- Fuzzy-Logik-basierte Informationsfusion [119]

Die meisten Modelle außer dem JDL- und DFIG-Modell sind sehr spezialisiert auf Text- bzw. Datenanalyse großer heterogener Datenmengen, zu unspezifisch auch für organisationale Anforderungen oder zu komplex und ressourcenintensiv für die spezifischen Anforderungen der Gebäudeautomation. Für diese Arbeit wird deshalb das DFIG-Modell verwendet, das die aus dem US-Verteidigungsministerium stammende JDL-Architektur [472] um eine Nutzerbetrachtung [48] ergänzt und mit Eigenschaften aus dem Observe-Orient-Decide-Act (OODA) Prozess [49, 248] angereichert wird. Das DFIG-Modell besteht aus sieben Ebenen, die in Abbildung 3.2 abgebildet sind und im Folgenden erläutert werden.

3.1.5.1 Ebene 0: Vor-Objekt-Bewertung

Hier werden die Vorverarbeitung und ihre Herausforderungen diskutiert. An die Rohdaten wird auch eine Anforderung an das Datenformat gestellt, die in Kapitel 3.2 näher beleuchtet wird. Der

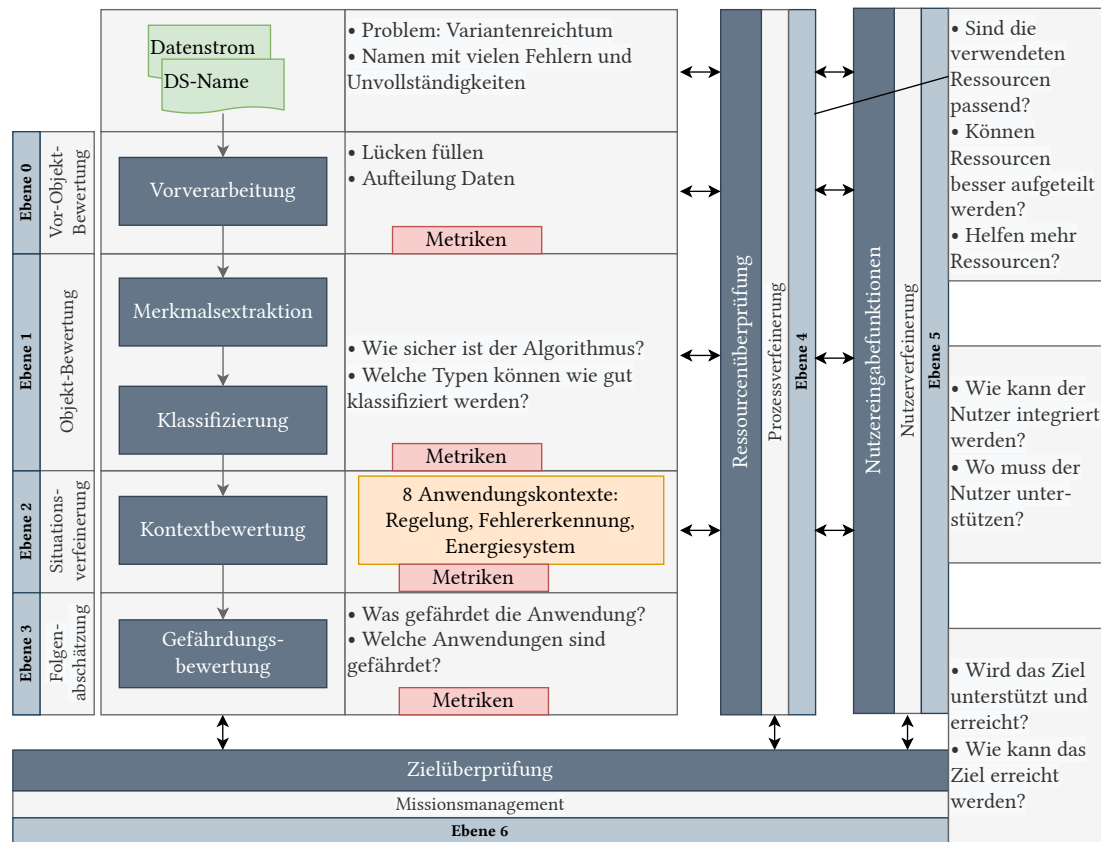


Abbildung 3.2: DFIG-Modell zur Datenfusion im Kontext der Metadatenextraktion aus GAS-Daten

in dieser Arbeit verwendete Datensatz selbst wird in Kapitel 3.5 beschrieben. Zentral für die Vorverarbeitung ist die initiale Datenqualität (Kapitel 3.1.3), die durch Vorverarbeitungsfunktionen erhöht werden kann. Die eigentlichen Vorverarbeitungsfunktionen werden in den Kapiteln 3.3.2 für einen Algorithmus zur Metadatenextraktion aus Zeichenketten von Datenströmen und in Kapitel 3.4.1 für einen Algorithmus zur Zeitreihenklassifizierung aus den Datenströmen selbst. In diesen Kapiteln wird jeweils ausgeführt, wie die Datenqualität der Rohdaten und vorprozessierten Daten nach den Datenqualitätskriterien bewertet werden kann.

3.1.5.2 Ebene 1: Objekt-Bewertung

In Ebene 1 wird das Objekt selbst identifiziert bzw. seine Objektbeschreibung verfeinert [315]. Die in dieser Arbeit behandelten Merkmalsextraktions- und Klassifizierungsalgorithmen wurden in Kapitel 2.3 beschrieben. Diese Ebene ist die Basis für den in dieser Arbeit entwickelten IE-Prozess. In ihr werden grundsätzliche Metadaten (z.B. „Temperatur“ oder „Alarm“) zu den Datenströmen hinzugefügt. Die IQ-Kriterien aus Kapitel 3.1.4 werden für die Bewertung der Performanz der Algorithmen herangezogen. Es wird überprüft, welche Algorithmen welche Metadaten hinzufügen und

wie auch die Übereinstimmung (Konformität) dieser Algorithmen ist. Die Informationszuweisungen, die aus der Objekt-Bewertung heraus vorgenommen werden, werden der Kontextbewertung als Eingang übergeben.

3.1.5.3 Ebene 2: Kontextbewertung

In der Kontextbewertung werden die klassifizierten Daten aus Ebene 1 in ihren jeweiligen Anwendungskontext (Situationsverfeinerung bzw. Kontextbewertung) [247, 315] überführt. So werden die acht Anwendungen aus drei Bereichen „Fehler- und Anomalieerkennung“, „GES-Regelung“ sowie der „Energiesystemanalyse“ näher betrachtet. Bei der Topologieerkennung wird z.B. die Verbindung zwischen Temperatursensoren erkannt. In dieser Ebene werden abermals die IQ-Kriterien herangeführt und die Kontextbewertungen auf Basis dieser Kriterien vorgenommen. Nach dieser Ebene ist die eigentliche algorithmische Arbeit vollständig durchgeführt. Die weiteren Ebenen sind vor allem Management-Ebenen, die vor allem den gesamten Prozess hin zur Durchführung der Anwendung unterstützen sollen.

3.1.5.4 Ebene 3: Gefährdungsbewertung

In der Gefährdungsbewertung wird überprüft, welche Algorithmen welche Ergebnisse liefern und welche Teilbereiche kritisch für die gewünschten Anwendungen sind. Es wird somit überprüft, für welche Anwendungen die Algorithmen automatisiert ohne Nutzereingabe verwendet werden können. Es werden aber auch kritische Ergebnisse in Teilbereichen (z.B. einzelnen Klassen der Objekt-Bewertung oder einzelnen Kontextbewertungen) betrachtet. Darauf aufbauend werden entsprechende Maßnahmen zur Senkung des Risikos zur automatischen Nutzung der Informationen aus den Algorithmen vorgeschlagen und durchgeführt. Es muss sichergestellt werden, dass die technischen Systeme nicht in Betriebszustände geraten, die aufgrund der durch den Datenfusionsprozess generierten Informationen zu einem Ausfall führen können [315].

3.1.5.5 Ebene 4: Ressourcenüberprüfung

Die Ebene 4 steuert den gesamten Prozess. Sie trifft Entscheidungen, um den Prozess effizienter zu gestalten [315] beispielsweise durch Fehleridentifikation und -behebungen [437]. Es muss im Zuge des IE-Prozesses überprüft werden, welche Ressourcen im Prozess welche Informationen bzw. Steigerungen der Informationsqualität leisten können. Ein Augenmerk muss vor allem auf den effizienten Einsatz von menschlichen Nutzereingabefunktionen geachtet werden. So müssen vor allem die Teilaspekte des IE-Prozesses auf Effizienz hin untersucht werden, die viele Nutzereingaben erfordern und wenig generalisierbar sind.

3.1.5.6 Ebene 5: Nutzereingabefunktionen

In dieser Ebene wird überprüft, wie mit zusätzlichen Nutzereingaben die Informationsqualität in den unterschiedlichen Prozesselementen verbessert werden können. Bei der Benutzerverfeinerung steuert ein Experte den gesamten Datenfusionsprozess, wobei er auf Basis von Übersichtsdarstellungen agiert [437] und die Ergebnisse von Algorithmen auf ihre Generalisierbarkeit [332, 437] und Plausibilität [50] überprüft werden. Es muss kritisch überprüft werden, ob zusätzliche Nutzereingaben einen ressourcengerechten Mehrwert in Form von Steigerungen der Informationsqualität erzeugen, wofür Reeve u. a. [381] einen Ansatz zur Berechnung des Arbeitsaufwandes für die manuelle Umbenennung von Datenströmen entwickelt hat. In dieser Arbeit soll deshalb eine anwendungsorientierte Bewertung erfolgen, um diesen Arbeitsaufwand möglichst zu minimieren.

3.1.5.7 Ebene 6: Zielüberprüfung

Es muss jeweils bei allen Ebenen der Informationsqualität überprüft werden, ob die in Ebene 2 definierten Anwendungen auch in Realität durch den IE-Prozess ausreichend unterstützt werden können. In dieser Ebene wird untersucht welche Informationsqualität in Ebene 2 erreicht wird und wie diese mit angemessenen Ressourcen verbessert werden könnte. Die Ressourcen sind im Bereich der Datenanalyse insbesondere in der Gebäudetechnik knapp, weshalb auf Automation der Prozesse gesetzt werden sollte. Aber auch die Margen sind in der Gebäudeautomation deutlich geringer, weshalb hier im Vergleich z.B. zu der Chemieindustrie nur wenig Ressourcen für die Optimierung von einzelnen Systemen aufgewendet werden kann. Das Missionsmanagement überprüft auch, ob angemessene Ressourcen verwendet werden im Kontext von sozialen, ökonomischen und politischen Randbedingungen [48, 51].

3.2 Auswahl eines Metadatenschema

Für die Bewertung und Auswahl des Metadatenschemas wird der in Kapitel 3.5 näher beschriebene Datensatz dieser Arbeit herangezogen. Die Übersetzungen der einzelnen Datenstromnamen werden als Grundlage genommen, um zu überprüfen, welches Metadatenschema wie viele der einzelnen Eigenschaften abbilden kann. Dabei werden zwei Aspekte analysiert:

- **Vokabelübersetzung:** Überprüfung, ob die Schemata die Vokabeln des Datensatzes vollständig oder eingeschränkt abbilden können.
- **Abdeckungsgrad:** Analyse, wie viel Prozent der Vokabeln (gewichtet und ungewichtet nach Häufigkeit im Datensatz) von den Schemata abgedeckt werden.

3.2.1 Auswahlprozess Metadatenschema

Für die Analyse der Metadatenschemata wurden folgende Versionen ausgewählt:

- **BUDO Schema** (Version vom 31.03.2023): Ein einheitliches Format mit umfangreicher Unterstützung für Kategorien und Vokabeln.
- **Brick Schema** (Version 1.2.1): Die am weitesten verbreitete Ontologie der Gebäudetechnik mit einem umfangreichen Vokabular.
- **VBIS** (Version vom 11.11.2022): Ein umfassendes, in der Bauindustrie insbesondere von Australien etabliertes Schema.

Es wird zum einen untersucht, inwiefern die Vokabeln ineinander übersetzbar sind. Zum anderen wird der Datensatz aus dem Projekt OOM4ABDO untersucht und genauer betrachtet welche Vokabeln in dem Testdatensatz abgebildet werden.

Die Anzahl der Vokabeln variiert erheblich zwischen den Schemata:

- **BUDO Schema:** 3.671 Vokabeln
- **Brick Schema:** 1.395 Vokabeln
- **VBIS:** 4.902 Vokabeln, wobei nach einer manuellen Überprüfung nur 2.061 Vokabeln einzigartig sind (durch Entfernen von Duplikaten).

Bei jeder einzelnen Vokabel des OOM4ABDO Datensatzes wurde manuell überprüft, inwiefern diese Vokabel im BUDO Schema, dem VBIS Schema und dem Brick Schema abgebildet werden kann. Es werden hierbei zwei Analysen abgebildet. Zum einen wird untersucht, ob das jeweilige Metadatenchema jede Vokabel übersetzen kann. Hierbei ist es unerheblich wie häufig eine Vokabel im gesamten Datensatz auftaucht. So wird die Vokabel „Temperatur“ alleine nur einmal aufgeführt. Zum anderen wird untersucht wie viele Prozent der Vokabeln des gesamten Datensatzes das Metadatenchema abbilden kann. Hierbei wird die Vokabel „Temperatur“ mit ihrer Anzahl der Vorkommen gewichtet.

Bei der Einordnung der Abbildungsfähigkeit werden vier Kategorien verwendet:

- **„vorhanden“:** 1-zu-1-Übersetzung möglich
- **„eingeschränkt vorhanden“:** Teil der Bedeutung wiedergegeben (Bsp. „relative Feuchtigkeit“ → „Feuchtigkeit“)
- **„sehr eingeschränkt vorhanden“:** Nur Grundkonzept möglich (Bsp. „Speicherprogrammierbare Steuerung“ → „Regelung“)
- **„nicht vorhanden“:** Keine Übersetzung möglich

3.2.2 Ergebnisse Bewertung Metadatenchema

In Abbildung 3.3 sind die Ergebnisse der Untersuchung der Metadatenchemata auf Basis der OOM4ABDO Daten aufgeführt. Der Datensatz wird durch das Brick Schema bei der einzelnen Zählung der Vokabeln nur bei 19,6 % der Vokabeln vollständig abgebildet, durch das VBIS nur bei

11,5 % der Vokabeln. Die Gewichtung durch das Vorkommen steigert die Bewertung beim Brick Schema auf 34,6 %. Bei VBIS verändert es sich sogar ins negative, so dass nur 11,2 % abgebildet werden. Dies liegt vor allem in fehlenden Sensoren und Regelungsbeschreibungen in VBIS begründet. Die nicht vorhandenen Vokabeln sind beim Brick Schema 65,2 % bei Einzelbewertung bzw. 46,1 % bei Vorkommensbetrachtung. VBIS bildet 77,8 % bzw. 72,9 % der Vokabeln nicht ab.

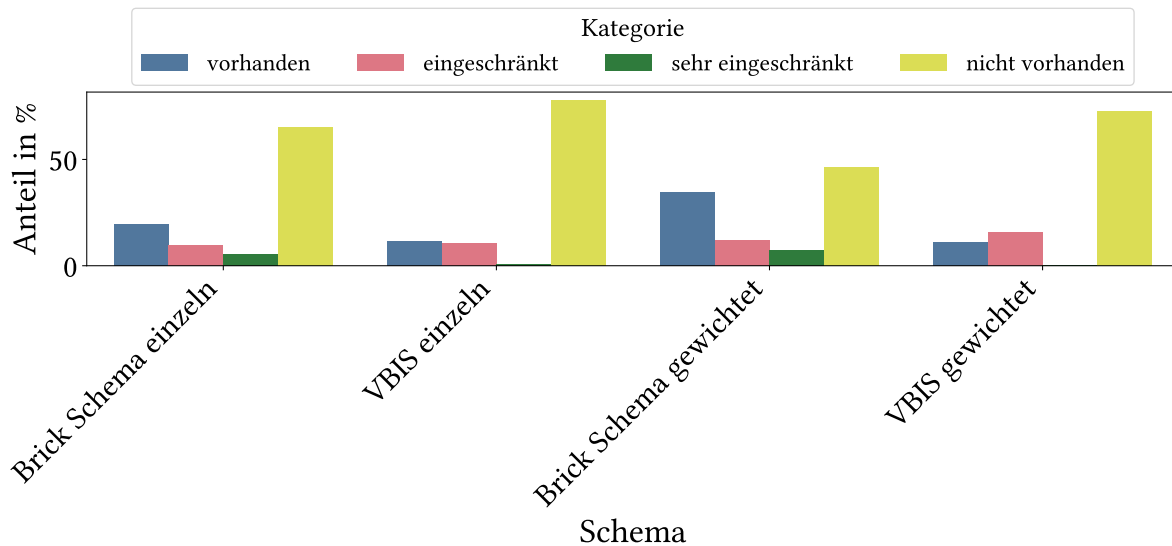


Abbildung 3.3: Ergebnisse der Untersuchung der unterschiedlichen Metadatenschemata (Brick Schema, VBIS). Einzeln bedeutet, dass jede im Datensatz vorkommende Vokabel nur einmal gezählt wurde. Gesamt bedeutet, dass die Vokabel mit ihrem Vorkommen gewichtet wurde.

3.2.3 Fazit Metadatenschema

Die Auswertung zeigt deutlich, dass Brick Schema und noch deutlicher VBIS die Anforderungen des OOM4ABDO Datensatzes nur unzureichend abbilden. Das BUDO Schema wurde auf die Anforderungen des Datensatzes angepasst und mit den im Datensatz enthaltenen Merkmalen weiterentwickelt, weshalb von diesem alle Anforderungen aus dem Datensatz zu 100 % abgebildet werden. Da hierdurch kein fairer Vergleich möglich ist, werden deshalb die Ergebnisse hier nicht näher beschrieben.

Die Ergebnisse zeigen aber deutlich, dass das Brick Schema und noch deutlicher VBIS mit mehr Vokabeln erweitert werden müssen, um die Informationsqualität der GAS-Realität abbilden zu können. Insbesondere im Bereich der Automation selbst, also beispielsweise der Nutzung von speziellen Sollwerten oder Geräten für die Regelung, fehlt es dem Brick Schema und VBIS an Expressivität, womit die Literatur [109, 322] bestätigt werden konnte. Bei Analyse der Modelle, die für das Brick Schema bereit gestellt werden, zeigt sich, dass fast nur Zuweisungen der Gebäude, Geräte und Signaltypen genutzt werden. Diese können alle vom BUDO Schema abdeckt werden. Die Verbindung

„feeds“, die anzeigt, dass ein technisches Gerät von einem anderen versorgt wird, kann im BUDO Schema bei einer Beziehung, die über 1-zu-n hinaus geht, nur umständlich umgesetzt werden.

3.3 Metadatenextraktion auf Basis von Zeichenketten

In diesem Kapitel wird die Extraktion von Metadaten aus Zeichenketten (Namen, Beschreibungen und Einheiten) vorgestellt. Zunächst werden zwei Klassifizierungsverfahren und anschließend zwei Merkmalsextraktionsverfahren. Bei den Verfahren zur Merkmalsextraktion werden unterschiedliche Merkmalssätze vorgestellt.

Datensätze von Zeichenketten der Gebäudeautomation haben verschiedene Herausforderungen:

1. Variantenreiche Datensätze
2. Unterschiedliche Sprachen
3. Deutsche Komposita
4. Fehlerhaft geschriebene Wörter und Abkürzungen
5. Synonyme
6. Hoher Personalaufwand der manuellen Übersetzung

Um diesen Herausforderungen zu begegnen, werden im Anschluss an den allgemeinen Algorithmus mehrere Vorverarbeitungsverfahren vorgestellt, die vor allem die Normalisierung des Datensatzes zum Ziel haben (Komposita, Korrektur fehlerhaft geschriebener Wörter und Normalisierung von Synonymen). Am Ende wird für das Clustering von Zeichenketten der Datenströme zu Anwendungen ein Algorithmus zur Verwendung von Ähnlichkeitswerten vorgestellt.

3.3.1 Klassifizierungsalgorithmen für Zeichenketten

Für die Übersetzung von Zeichenketten der Namen und Beschreibungen der Datenströme werden verschiedene Algorithmen mit verschiedenen Ansätzen zur Merkmalsextraktion verwendet. Als Algorithmen werden der CatBoost Algorithmus und ein konventioneller Algorithmus auf Basis von LSTM verwendet, da transformerbasierte LLM in Vorversuchen keine guten Ergebnisse zeigten (siehe Kapitel E.7.1).

3.3.1.1 CatBoost-Algorithmus

Der CatBoost Algorithmus wird als Vertreter der Entscheidungsbäume in diesem Vergleich verwendet. Insbesondere die Kategorisierung der einzelnen Merkmale und direkte Betrachtung von Merkmalskombinationen machen diesen Algorithmus für die Anwendung in Gebäudeautomationsdaten passend. Tendenziell werden Gebäudeautomationsdaten immer wieder mit gleichem Aufbau

benannt mit höchstens leichten Variationen. Deshalb liefern diese Daten kategorisierte Merkmale, die der CatBoost Algorithmus gut verwerten kann. Außerdem sind einzelne Merkmale nur im Zusammenhang zu verstehen. So kann z.B. der Block „MW“ die Übersetzungen Messwert, Mittelwert oder Megawatt beinhalten, das erst im Gesamtkontext z.B. bei „in MW“ (Megawatt) klar wird.

3.3.1.2 LSTM-basierter Algorithmus

Im Rahmen dieser Arbeit wurde ein konventioneller Algorithmus entwickelt, der den aktuellen Stand der KNN abbildet. Konventionelle Algorithmen wie LSTM wurden in der Literatur für die Merkmalsextraktion von Zeitreihen und Zeichenketten der Datenströme verwendet und lieferten meist die besten Ergebnisse [215, 465]. Deshalb wird ein Algorithmus auf Basis von bidirektionalen LSTM Algorithmen verwendet, der alle Blöcke der Zeichenkette als Input erhält („Gesamtheitlich“). Ein LSTM Modell kann nur sequentiell mit allen enthaltenen Blöcken ausgeführt werden, so dass eine einzelne Betrachtung der Blöcke („Blockweise“) nur umständlich zu erreichen ist.

3.3.2 Merkmalsextraktionsverfahren Zeichenketten

Im Folgenden werden zwei Merkmalsextraktionsverfahren vorgestellt, die eine Zeichenkette auf unterschiedliche Arten und Weisen in Merkmale überführen. Das erste Verfahren betrachtet eine Zeichenkette blockweise, während das zweite Verfahren die Zeichenkette gesamtheitlich betrachtet. Die Abbildung 3.4 zeigt den Unterschied in der Zuordnung von Kennzeichen auf, der zusammen mit den extrahierten Merkmalen in den kommenden Abschnitten näher erläutert wird. In Abbildung 3.5 finden sich eine Übersicht über die Kombinationen von Merkmalssätzen, Merkmalsextraktionsverfahren und Algorithmen, die im Rahmen dieser Arbeit verwendet werden.

3.3.3 Blockweise Merkmalsextraktion

In dieser Merkmalsextraktionsmethode wird ein Datenstromname in seine einzelnen Blöcke unterteilt, denen eine oder mehrere Kennzeichnungen zugewiesen werden (z.B. „Temperatur“). In den Ergebnissen wird er als „Blockweise“ bezeichnet.

Bei dem Datenstromnamen „4120-Room-Set-Temperature-RTs“ werden in die Blöcke „4120“, „Room“, „Set“, „Temperature“, „RTs“ unterteilt und ihnen Zuweisungen von Kennzeichen gegeben (Abbildung 3.4).

Meistens enthalten Zahlen wie der erste Block „4120“ nur eine ID, aber in diesem Fall gibt es auch eine gleichzeitige Beschreibung als Kennzeichen „Gebäude“. Die folgenden Blöcke „Room“ („Raum“), „Set“ („Sollwert“) und „Temperature“ („Temperatur“) sind klar eins zu eins kennzeichnbar. Der Block „RTs“ stellt eine Besonderheit dar, da er die englischen Begriffe „Room“, „Temperature“ und „Setpoint“ vereint. Zusätzlich kann ihm durch Fachwissen die Zuweisung des Mediums „Luft“ mit der Eigenschaft „Raum/Innen“ zugewiesen werden, was der Luftart „Raumluft“

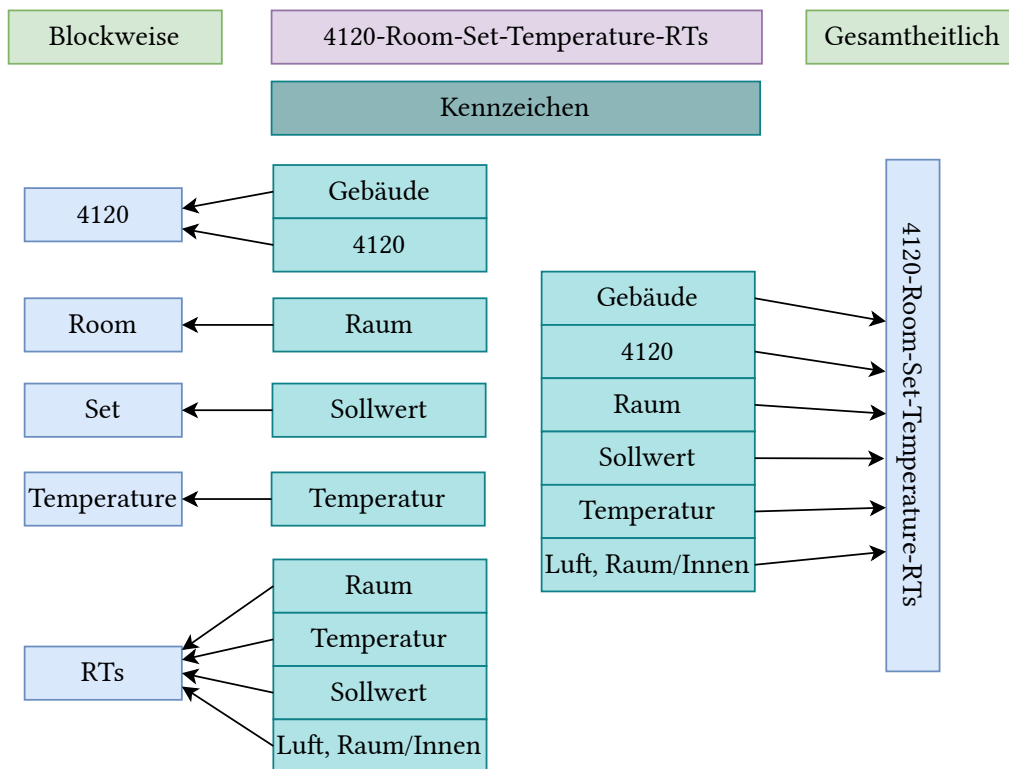


Abbildung 3.4: Ein Vergleich der Zuweisungen von Kennzeichen zu Instanzen bei der blockweisen und gesamtheitlichen Merkmalsextraktion.

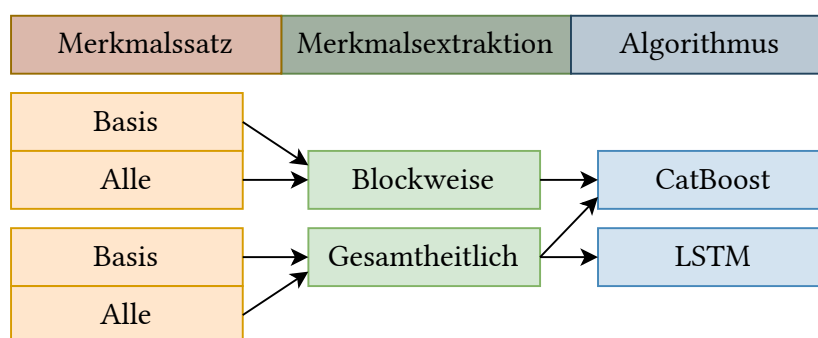
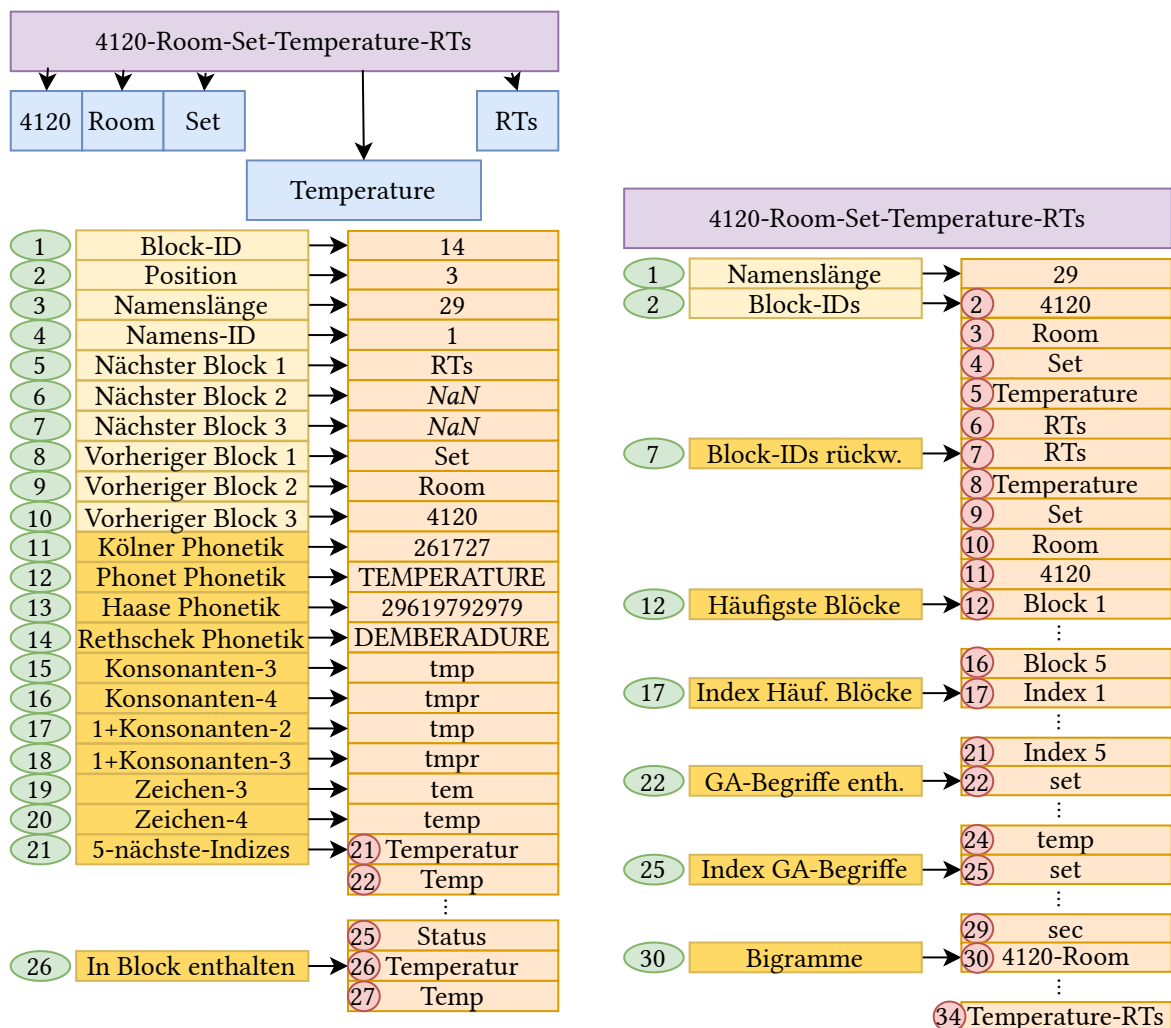


Abbildung 3.5: Abbildung der jeweiligen Merkmalsätze („Basis“ und „Alle“) mit Merkmalsextraktionsarten („Blockweise“ und „Gesamtheitlich“) zusammen mit den verwendeten Algorithmen („CatBoost“ und „LSTM“). LSTM kann aufgrund seiner Eigenschaften, die auch die Reihenfolge der Blöcke des Gesamtnamens erfordert, nur „Gesamtheitlich“ durchgeführt werden.

nach DIN EN 16798-3 [111] entspricht, obwohl die Eigenschaft „Luft“ nicht im Begriff auftaucht. Diese höhere Informationsdichte stellt eine größere Herausforderung für den Algorithmus dar.

Als Eingang für den Algorithmus wird jeder Block in einen von zwei Merkmalsätzen überführt: „Basis“ und „Alle“. In Abbildung 3.6a ist eine grafische Übersicht der Merkmalsätze anhand eines Beispielnamens dargestellt. Die einzelnen Merkmalsätze werden im Folgenden beschrieben.



(a) Vereinfachte Darstellung der **blockweisen** Merkmalsextraktion bei Block „Temperature“.

(b) Vereinfachte Darstellung der **gesamtheitlichen** Merkmalsextraktion beim Beispielnamen.

Abbildung 3.6: Abbildungen der blockweisen und gesamtheitlichen Merkmalsextraktion. Hellgelb sind alle Merkmale vom Merkmalsatz „Basis“ und Dunkelgelb alle Merkmale vom Merkmalsatz „Alle“ markiert.

3.3.3.1 Blockweiser Merkmalssatz „Basis“

Für den Merkmalssatz „Basis“ werden pro Block folgende Merkmale extrahiert:

- Block-ID
- nächster, übernächster und überübernächster Block
- vorheriger, vorvorheriger, vorvorvorheriger Block
- Position im Datenstromnamen
- Länge des Datenstromnamens

Dies ermöglicht die Einordnung eines Blocks in seinen umgebenden Gesamtzusammenhang. Ein großer Nachteil ist seine Abhängigkeit von der Position des Blocks, der aber durch die Berücksichtigung weiterer umgebender Blöcke aufgefangen wird.

3.3.3.2 Blockweiser Merkmalssatz „Alle“

Der Merkmalssatz „Alle“ erweitert „Basis“ um folgende Merkmale aus der Phonetik und auf Fachwissen basierte GAS-orientierte Merkmale:

- **Phonetische Merkmale:** um u.a. multilingualer phonetischer Schreibweise zu begegnen („Comfort“ statt „Komfort“, siehe Kapitel 2.3.4.3),
- **Konsonanten:** Erste drei bzw. vier Konsonanten und erster Buchstabe der Zeichenkette mit den ersten zwei bzw. drei Konsonanten,
- **Buchstaben:** Erste drei bzw. vier Buchstaben,
- **Berücksichtigung GAS-Wörter:**
 - **Damerau-Levenshtein Distanz** zu fünf ähnlichsten GAS-typischen Wörtern,
 - **Enthaltene Wörter:** Die enthaltenen GAS-typischen Wörtern selbst mit binärem Encoding („enthalten“, „nicht enthalten“).

Als Implementierung der Phonetik-Algorithmen wird die Abydos Bibliothek in Python genutzt [288]. Die Betrachtung von GAS-Zeichenketten zeigt, dass Konsonanten aus GAS-Begriffen häufig als Abkürzung verwendet werden, wie bei „Funktion“ (→ „Fnk“ bzw. → „Fnkt“). Auf Basis einer unabhängiger Wörterdatenbank aus OOM4ABDO (siehe Kapitel C.5 im Anhang) wurden 208 häufig vorkommende GAS-typische Wörter extrahiert.

3.3.4 Gesamtheitliche Merkmalsextraktion

In dieser Arbeit wurde eine neue Merkmalsextraktionsart entwickelt, die statt sequentieller Block-Verarbeitung alle Zuweisungen direkt dem Datenstromnamen zugeweißt und es werden alle redundanten Zuweisungen entfernt. In unserem vorherigen Beispielsnamen „4120-Room-Set-Temperature-RTs“ würden die Zuweisungen „Gebäude“, „4120“, „Raum“, „Sollwert“, „Temperatur“,

„Medium, Luft, Raum/Innen“ direkt dem Datenstromnamen „4120-Room-Set-Temperature-RTs“ zugewiesen. Als Eingang für den Algorithmus werden die in Kapitel 3.3.4 beschriebenen Merkmale verwendet.

Alle Kennzeichen eines Datenstromnamens werden vom Algorithmus als Klassen ausgegeben, was für einen Klassifizierer eine schwerere Herausforderung darstellt. Diese Art der Merkmalsextraktion hat den Vorteil, dass hierbei deutlich weniger Eingaben vom Nutzer erforderlich sind. Im Anhang in Abbildung 3.6b ist eine grafische Übersicht über die einzelnen Merkmalssätze („Basis“ und „Alle“) bei einem Beispielnamen dargestellt, die im Folgenden beschrieben werden.

3.3.4.1 Gesamtheitlicher Merkmalssatz „Basis“

Der Merkmalssatz „Basis“ enthält die folgenden Merkmale:

- Länge des Namens
- Block-IDs in Reihenfolge des Vorkommens (Bsp. „4120-Room-Set-Temperature-RTs“: „4120“ → „Block-ID-0“, „Room“ → „Block-ID-1“)

3.3.4.2 Gesamtheitlicher Merkmalssatz „Alle“

Beim Merkmalssatz „Alle“ werden neben den Merkmalssatz „Basis“ die folgenden Merkmale ergänzt:

- Block-IDs in rückwärtiger Reihenfolge (Im Bsp.: „RTs“ → „Block-ID-0“), da oftmals am Ende eines Namens Eigenschaften des Signaltyps eingetragen,
- Meist genutzte Blöcke,
- Ähnlichkeit (Damerau-Levenshtein Distanz) zu den meist genutzten Blöcken,
- Typische Wörter der Gebäudeautomation enthalten mit binärem Encoding („enthalten“, „nicht enthalten“),
- Ähnlichkeit der Blöcke (Damerau-Levenshtein Distanz) zu üblichen Wörtern in der Gebäudeautomation,
- Bi-gramme der Blöcke des Namens (analog zu „vorheriger Block“ bzw. „nächster Block“ aus dem blockweisen Merkmalen).

3.3.5 Komposita und Wortdatenbanken

Ein großes Problem bei der Anwendung automatischer Algorithmen für die Erkennung von Metadaten in Gebäuden ist die Sprachabhängigkeit vom Trainingsdatensatz. In der deutschen Sprache werden häufig zusammengesetzte Wörter verwendet (Komposita). Im Englischen wird zum

Beispiel die Vorlauftemperatur als zwei Wörter geschrieben: „flow temperature“. Die Komposita können theoretisch beliebig viele Wortbestandteile enthalten, deren Anzahl in der Realität jedoch limitiert ist. So kommt das laut Duden längste Wort auf maximal fünf Wortbestandteile: Aufmerksamkeitsdefizit-Hyperaktivitätsstörung (Aufmerksamkeit|Defizit|Hyper|Aktivität|Störung) [205].

Im Folgenden wird ein Ansatz zur Trennung von Komposita auf Basis verschiedener bestehender Ansätze und eigener Wortdatenbank erstellt.

3.3.5.1 Wortdatenbank

Der Datensatz aus der Gebäudeautomation enthält eine variantenreiche Anzahl an Wörtern. Um herauszufinden, ob ein gültiges Wort vorliegt, können Wortdatenbanken herangeführt werden:

- **Normen-Datenbank Perinorm** [44]: Ein selbstentwickelter Webcrawler und PDFExtraktor auf Basis von Selenium Webdrive [406] und pdfplumber [419] extrahierte Wörter aus 229 Normen mithilfe des Suchbegriffs „Gebäudeautomation“.
- **Free German Dictionary** [404] ist nach eigenen Angaben der größte deutsche Wortkorpus mit über 2 Millionen verschiedenen deutsche Wörtern.
- **GermaNet** der Abteilung „Allgemeine Sprachwissenschaft und Computerlinguistik“ der Universität Tübingen mit ungefähr 121.700 deutschen Komposita [180].
- **Digitale Wörterbuch der deutschen Sprache** [120] bietet einen Wortkorpus von 584.000 Einträgen.
- **Wortschatzportal der Universität Leipzig** [371] mit 49 öffentlich verfügbaren 49 Wortkorpora mit maximal 1 Millionen Sätzen pro Korpus.
- **Datenbank aus AKS-Standards**: Wortextraktion aus 121 Standards, Vorschlägen und Normen mit 3.800 neuen Wörtern zur Beschreibung von Datenstromnamen aus dem Aufbau des BUDO-Schemas.
- **Datenbank aus GAS-Dokumenten**: Wortextraktion aus 4.924 Dokumenten aus der Recherche zum BUDO-Schema mit 23.000 nicht in den vorherigen Datenbanken enthaltenen Wörtern.

Aus allen Wortkorpora wird ein Gesamtkorpora gebildet, der durch eine Vorverarbeitung auf signifikante Wörter reduziert wird. Die verwendeten Filter sind im Anhang in Kapitel C.1 dargestellt.

3.3.5.2 Ansätze zur Kompositazerlegung

Es existieren nur wenige Ansätze zur Kompositazerlegung. In dieser Dissertation werden drei Ansätze (Tuggener, repodioc und Stinner) miteinander verglichen und im Folgenden vorgestellt. Eine ausführlichere Beschreibung der Algorithmen findet sich im Anhang in den Kapiteln C.4.2 bis C.4.4.

3.3.5.2.1 Ansatz von Tuggener Der ursprüngliche Ansatz von Tuggener [447] basiert auf der Bewertung von Zeichenketten, ob diese am Anfang, in der Mitte oder am Ende von Wörtern auftauchen. Wenn die Wahrscheinlichkeit für eine Trennung an einer Stelle hoch ist, dann wird eine Trennung ausgegeben. Wichtig für einen funktionierenden Ansatz ist, dass er nicht mit zu vielen Komposita als Eingangswerte trainiert wird, weshalb vorwiegend mit technischen Wörtern genutzt werden. Technisch wurde der ursprüngliche Ansatz von Niklaus [335] erweitert, der auch hier verwendet wurde. In dieser Dissertation wurde der grundlegende Algorithmus technisch um eine lokale Datenbank auf Basis von SQLite [184] erweitert, um Speicherprobleme zu beheben.

Zusätzlich wurde in dieser Arbeit alternativ eine umfangreiche Nachbearbeitung eingeführt, die die Reihenfolge der Wahrscheinlichkeiten des ursprünglichen Ansatzes von Tuggener verwendet, aber für alle Varianten eine mögliche Trennung überprüft. Die genaue Nacharbeitung wird im Ansatz von Stinner in Kapitel 3.3.5.2.3 vorgestellt.

Ein Problem vom Ansatz von Tuggener ist, dass er immer nur an einer Stelle im aktuell verwendeten Datenblock trennen kann. Dies ist insbesondere problematisch, wenn ein Komposita aus mehr als zwei Bestandteilen besteht.

3.3.5.2.2 Ansatz von Repodiac Repodiac [385] entwickelte einen Ansatz auf Basis von performanten [287] Aho-Corasack-Zustandsautomaten [12] mit dem Free German Dictionary [404] als Lexikon. Der Ansatz von Repodiac besitzt eine Nachbearbeitung, die Prefixe (Bsp. „Vor-“, „Durch-“) und (Plural-)Suffixe (Bsp. „-ung“, „-en“) gesondert betrachtet. Auch die Änderung von Vokalen zu Umlauten (z.B. Vogel → Vögel) bei der Pluralform finden Beachtung. Der Algorithmus überprüft vom Ende des Wortes ausgehend, ob es sich um ein Wort handelt (siehe Abbildung 3.7) und prüft für nicht zuordenbare Artefakte, ob diese einem passenden Wort zugewiesen werden können

3.3.5.2.3 Ansatz von Stinner Als Basis für den hier in der Dissertation verwendeten Algorithmus wird der vorher vorgestellte Ansatz von Repodiac verwendet. Bestandteil des eigenen Algorithmus ist ein Postprocessing, in dem jeder gefundene Datenblock auf ein gültiges Wort mithilfe einer Wortdatenbank überprüft wird. Hierbei werden auch Umlautverschiebungen (z.B. Vogel → Vögel), den Ersatz von Umlauten (ä → ae) oder Mittelfixe („zu“ in „einzusehen“) berücksichtigt. Der Algorithmus arbeitet rekursiv und nutzt nach drei Iterationen den Algorithmus von Tuggener. Vorversuche zeigen eine Genauigkeit von 98,2 % im Vergleich zu 82,9 % beim besten bestehenden Ansatz (Repodiac).

3.3.6 Korrektur fehlerhafter Schreibweisen und Abkürzungen

Es gibt neun Fälle von fehlerhaften Schreibweisen und Abkürzungen in allen Datensätzen:

1. Abkürzung von Teilwörtern in Reihenfolge (Bürogebäude → Bürogeb)

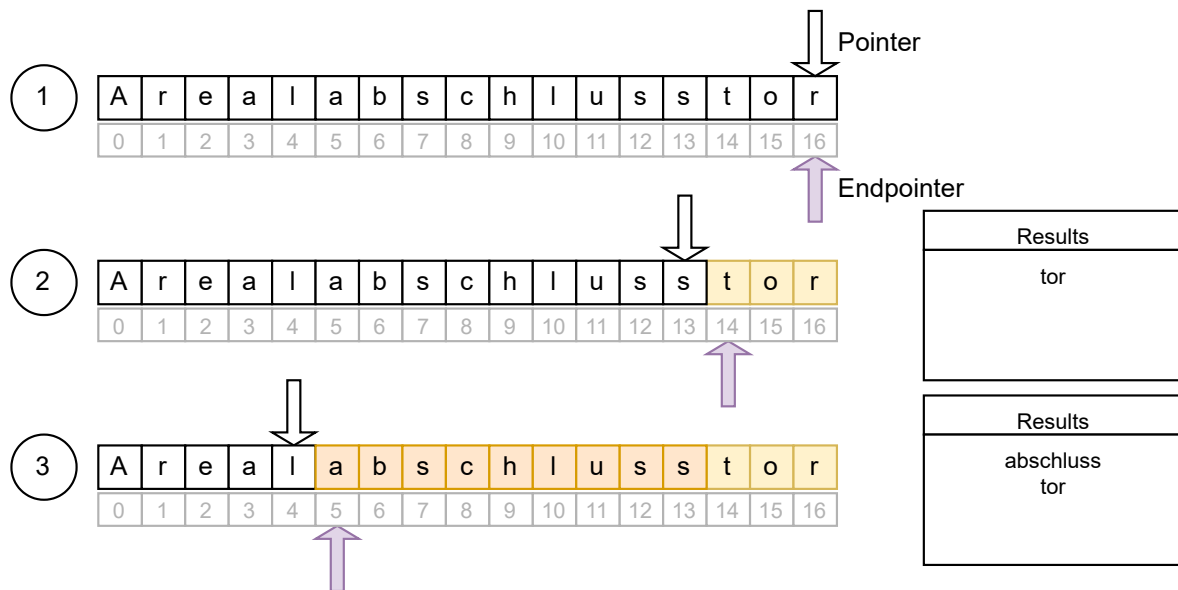


Abbildung 3.7: Ablauf der Detektion von Repodiak. Vom letzten Zeichen ausgehend wird jeweils überprüft, ob sich am Ende eines gefundenen Wortes ein neues Wort befindet.

2. Abkürzung von Teilwörtern mit Erstbuchstabe und (beliebigen) Konsonanten (Regelungs-**funktion** → Regelungs**fkt**)
3. Vertauschung von Zeichen (Temperaturleitkoeffizient → Temperaturleitkoeffiz**ei**nt)
4. Weglassen von Zeichen (Strombelastbar**ke**it → Strombelastbak**ei**t)
5. Fehlerhaftes Zeichen (Techn**ik** → Techn**nlk**)
6. Zusätzliche Zeichen (Wartezeit → Wartze**it**)
7. Weglassen von wiederkehrenden Zeichen (Sonnen**en**ergie → Sonnen**en**ergie)
8. Zu viele wiederkehrende Zeichen (Jalousieklapp**e** → Jalousieklapp**pp**e)
9. Rechtschreibfehler (die aber auch in den anderen Fehlern enthalten sind, jedoch schwieriger zu erkennen sind) (Chemikalien → Chem**ie**kalien)

Fehlerhaft geschriebene Wörter oder Abkürzungen können auch fehlerhafte Trennungen von Komposita zur Folge haben. So kann das Wort „Größenord“ für den Menschen direkt als „Größenordnung“ erkannt werden, ein Algorithmus würde jedoch die beiden Wörter „Größe“ und „nord“ erkennen. Um dem Problem der fehlerhaft geschriebenen Wörter oder Abkürzungen zu begegnen wurde ein Algorithmus zur Trennung dieser auf Basis des Algorithmus von Stinner (siehe Kapitel 3.3.5.2.3) entwickelt, der in Kapitel C.5.1 ausführlicher beschrieben wird.

Jedes Wort, was dieser nicht trennen kann und länger als 10 Zeichen ist, wird als Kandidat für ein fehlerhaft geschriebenes Wort erachtet. Dann werden die verschiedenen Fälle von fehlerhaften Schreibweisen durchgegangen und jeweils ein Handling durchgeführt. Gesetz dem Fall, dass auch

kein Wort gefunden wird, so wird das Wort nicht getrennt. Es kann auch ein Fachwort, Eigenname oder fremdsprachiges Wort sein, das sich noch nicht im Wörterbuch befinden.

Die Ergebnisse im Anhang (Kapitel C.5.2) zeigen, dass 43,7 % der fehlerhaft geschriebenen Wörter und Abkürzungen korrigiert werden können bei einer Fehlerquote von 11,9 %. Somit hat dieser Algorithmus ein Potential die Normalisierung des Zeichenketten-Datensatzes zu unterstützen.

3.3.7 Normalisierung Synonyme

Es gibt in der Gebäudeautomation Begriffe, die äquivalent genutzt werden können. So sind die Terme „Anlage“, „System“ und „Einrichtung“ weitestgehend miteinander austauschbar oder können ohne Informationsverlust weggelassen werden (Bsp. „Heizkreislaufsystem“ → „Heizkreislauf“). Auch die Terme „Steuerung“, „Regelung“, „Betrieb“ und „Funktion“ bedeuten in den meisten Fällen semantisch exakt das Gleiche. In Tabelle C.3 im Anhang sind verschiedene semantisch äquivalente Wort-Kombinationen in der Gebäudeautomation abgebildet.

3.3.8 Algorithmus zur Verwendung von Ähnlichkeitswerten

Der Algorithmus zur Nutzung der Ähnlichkeitswerte (engl. Similiarity Score: SSC) wurde zuerst von Stinner u. a. [429] genauer vorgestellt. Die methodischen Grundlagen für Ähnlichkeitswerte bei Zeichenketten wurden in Kapitel 2.3.4 näher erläutert.

Die Eingangsbasis ist für diesen Algorithmus immer nur der Datenstromname (Abk.: Name) bzw. die Datenstrombeschreibung (Abk.: Bes), die häufig aus Klartext besteht. Da der Datenstromname am Ende häufig Informationen über den Funktionstyp (z.B. binärer Input) und den Datenstromtyp (z.B. Temperaturmessung) enthält, aber keine über Hierarchie, werden in der Vorverarbeitung der Berechnung der Ähnlichkeitswerte je nach Vorkommen von Funktionstypen die letzten zwei bzw. vier Blöcke entfernt. Außerdem werden alle Sonderzeichen durch ein Leerzeichen ersetzt.

Allgemein wird eine Bewertung der Ähnlichkeit aus mehreren Ähnlichkeitswerten ermittelt. Jeder SSC untersucht einen anderen Teil eines Datenstromnamens und eine optimale Kombination verschiedener SSC führt zu stabileren Ergebnissen. In Formel 3.14 wird die Formel allgemein definiert mit den folgenden SSC als Unter-SSC:

- Jaro-Winkler-Distanz
- Cosinus-Distanz
- Damerau-Levenshtein-Distanz
- Optimal String Alignment (OSA)
- Longest Common Subsequence (LCS)
- Longest Common Substring (LCSS)

- Prefix

$$SSC_{\text{total}} = \sum_{i=0}^n c_{i,\text{Name}} \cdot SSC_{i,\text{Name}} + c_{i,\text{Bes}} \cdot SSC_{i,\text{Bes}} \quad (3.14)$$

Wobei i der Index für den SSC und c_i ein Korrekturfaktor zur Bewertung ist. Für jede Anwendung kann c_i neu gesetzt werden. Es ist jedoch wichtig, dass das c_i anhand eines Datensatzes ermittelt (trainiert) und anhand eines anderen Datensatzes getestet wird. Die Gewichtung der Beschreibung $c_{i,\text{Bes}}$ wird hierbei im Gegensatz zu Stinner u. a. [429] auf 0 festgesetzt, da die Beschreibung keine gesicherten Informationen über die Hierarchie enthält.

Für die Berechnung der Distanzen wird hauptsächlich das Python-Paket RapidFuzz [26] verwendet, da die Bibliothek in C++ geschrieben ist und einen großen Geschwindigkeitsvorteil gegenüber anderen vergleichbaren Bibliotheken besitzt [282]. Die Cosinus-Distanz und den Longest Common Substring (LCSS) sind nicht in RapidFuzz enthalten, weshalb hierfür das Paket TextDistance [282] verwendet wird.

3.4 Zeitreihenklassifizierung

Die Zeitreihenklassifizierung erfolgt auf einem Ausschnitt des OOM4ABDO Datensatzes und des Mortar Datensatzes mit Berücksichtigung der folgenden Prozessschritte:

- **Vorprozessierung:** Kapitel 3.4.1,
- **Merkmalsextraktion:** Kapitel 3.4.2,
- **Einflüsse von Störgrößen, Nutzungsmustern:** Kapitel 3.4.3,
- **Klassifizierung:** Kapitel 3.4.4.

3.4.1 Datenvorprozessierung

In der Datenvorprozessierung werden verschiedene Schritte vorgenommen. Jede Zeitreihe wird in ein Jahr mit 52 Wochen unterteilt, dass immer beim ersten Montag des Jahres beginnt. Die Auflösung wird auf 15-minütig normalisiert. Die Daten aus der Gebäudeautomation wurden mit „change of value“ aufgezeichnet, wobei bei binären Variablen (z.B. an/aus) nicht zwischen fehlerhafter Aufzeichnung und einer nicht erfolgten Änderung unterschieden werden kann.

Diese Ambivalenz von „Lücken“ in der Datenaufzeichnung wird in dieser Arbeit durch Nutzung einer vorwärts gerichteten Füllung der Datenlücken gelöst. Da es sehr viele binäre oder mehrstufige ganzzahlige Variablen gibt, ist eine Interpolation keine sinnvolle Füllung, da sonst binäre Variablen $x \in \{0, 1\}$ auch nicht sinnvolle Werte von 0,5 annehmen können. Zusätzlich wird ein Resampling durchgeführt, wobei die Eingangsdaten gemischt werden, damit einzelne Klassen nicht sequentiell in den Algorithmus übergeben werden.

3.4.2 Nutzung der Merkmalsextraktion

Aus der auf CNN-basierenden Rocket Familie werden „MiniRocket“ und „MultiRocket“ mit ihren jeweiligen Standardparametern ausgewählt, da diese nach Stinner u. a. [431] auf Gebäudeautomationsdaten die besten Ergebnisse erzeugten.

Die drei Merkmalsextrahierer TSFresh [88], TSFEL [35] und MVTs [11] (siehe Kapitel 2.3.1) werden gemeinsam in dieser Arbeit im Hinblick auf die Zielerreichung bewertet. Eine klassische Merkmalsextraktion hat den Vorteil, dass z.B. die durch einen baumbasierten Klassifizierer extrahierten Regeln auch menschenverständlich übersetzt werden können.

Bei genauer Untersuchung der Pakete ist auffällig, dass keines real multivariate Merkmale extrahiert. MVTs akzeptiert zwar multivariate Zeitreihen als Eingang, aber extrahiert nur jeweils auf den univariaten Zeitreihen Merkmale. Dies ist eine Lücke, die im Rahmen dieser Arbeit geschlossen werden soll.

Auf eine genaue Auflistung aller neu entwickelten Merkmale wird an dieser Stelle verzichtet und auf den Anhang in Kapitel D verwiesen. Die existierenden Merkmalsextrahierer haben Probleme große Datenmengen angemessen schnell zu berechnen [449]. Für den gesamten Mortar oder OOM4ABDO Datensatz würde eine Anwendung mehrere Wochen in Anspruch nehmen. Deshalb wurden die Extrahierer in ein Tool überführt, dass alle Merkmale mit dreidimensionalen Numpy-Arrays [175] als Eingang berechnet. Numpy hat den Vorteil, dass alle grundlegenden Funktionen in C entwickelt sind, das eine sehr maschinennahe Ausführung der Algorithmen garantiert. Der eigene entwickelte Merkmalsextrahierer wird im folgenden als „TSFast“ bezeichnet. Er erweitert die existierenden Ansätze um 122 weitere Merkmale, die die 43 Merkmale von TSFEL, 75 Merkmale von TSFresh und 35 Merkmalen von MVTs erweitert. Diese Merkmale befinden sich u.a. im Bereich der Entropien und Fraktale. Tabelle D.1 im Anhang gibt einen Überblick über alle implementierten Merkmale.

3.4.3 Einflüsse von Störgrößen, Nutzungsmustern

Die Klimazone hat einen Einfluss auf die Ergebnisse der Metadatenextraktion anhand von Zeitreihen [152]. Da in dem hier vorgestellten Datensatz vom Projekt OOM4ABDO jedoch nur eine Klimazone betrachtet wird, hat dies für diesen Datensatz keinen Einfluss. Für den Mortar-Datensatz könnte dies schon einen Einfluss haben, da der Datensatz Gebäude in der gesamten USA enthält.

ML-Algorithmen können zu besseren Ergebnissen führen, wenn die Merkmalsextraktionen sich auf unterschiedliche Metabereiche von Zeitreihen konzentrieren [154]. Es können besondere Tage und Ereignisse wie Wochentage (Montag), Wochenende, Feiertag oder Ferien, aber auch die Jahreszeit das Betriebsverhalten beeinflussen, das sich auch in den Zeitreihen widerspiegeln sollte.

Es wird die These in dieser Arbeit überprüft, ob tatsächlich saisonale bzw. wöchentliche Unterschiede in den Merkmalen bessere Ergebnisse erzeugen. Für eine Überprüfung einer eventuellen

zeitlichen Abhängigkeit werden die 15-minütig gesampleten Zeitreihen in wöchentliche Zeitreihenabschnitte unterteilt, die jeweils durch eine Überlappung von zwei Tagen in unterschiedliche Zeitreihen unterteilt werden. In jeder dieser Zeitreihe werden die entsprechenden Merkmale extrahiert. Für die monatliche Auswertung werden die Zeitreihen in vier-Wochen-Abschnitte mit jeweils einer Woche Überlappung unterteilt.

3.4.4 Klassifizierer

Für die Klassifizierung der Daten werden die in Kapitel 2.3.3 vorgestellten Klassifizierer verwendet. Da die Durchführung eines der verwendeten Deep Learning Ansätze von Stinner u. a. [432] auf einem Ausschnitt des Mortar-Datensatzes geschätzt mehrere Wochen dauern würde, wird in dieser Arbeit auf klassische Deep Learning Ansätze zur Klassifizierung verzichtet. Diese bieten ebenfalls laut Benchmarks bzw. Ergebnissen aus dem deutlich kleineren Datensatz von Stinner u. a. [432] keine Vorteile gegenüber klassischer Merkmalsextraktion oder den Rocket Merkmalsextrahierern.

Für die Rocket Familie wird als standardmäßiger Klassifizierer der Ridge Klassifizierer als schneller linearer Klassifizierer verwendet [105]. Ein großer Nachteil vom puren Ridge Klassifizierer ist, dass er keine Wahrscheinlichkeiten der Klassen berechnet und deshalb auch nicht ausgeben kann. Er berechnet immer nur die Ähnlichkeit zu den Klassen bei den vorliegenden Merkmalen und wählt die höchste Ähnlichkeit aus. Die Wahrscheinlichkeiten der Klassenzuordnung sind aber für die Nutzung des Expected Calibration Error (ECE) als Informationsqualitätsmerkmal der „Zuverlässigkeit“ entscheidend. Deshalb wird im Rahmen dieser Arbeit ein eigens entwickelter Ansatz verwendet, in dem Ridge Klassifizierer ein Kalibrierungsklassifizierer auf Basis eines in diesem Fall anwendbaren [352] isotonischer Baumklassifizierers [214] nachgeschaltet, der aus der Ausgabe des Ridge Klassifizierers Wahrscheinlichkeiten erzeugt. Im Folgenden wird der Ridge Klassifizierer immer als Kombination des Ridge Klassifizierers und des isotonischen Baumklassifizierers betrachtet und als Ridge Klassifizierer bezeichnet.

Insgesamt hat die Klassifikation auf Basis von Zeitreihen den Vorteil, dass sie personalärmer umgesetzt werden kann, da für ein neues AKS kein neues Anlernen erforderlich ist. Sie hat aber den Nachteil, dass bei ihr weniger Klassen definiert werden können, da eine MultiLabel-Strategie (z.B. „Temperatur“ am „Vorlauf“ eines „Wärmeübertragers“) nicht direkt erfolgreich umgesetzt werden kann. Dies könnte durch ein stufenweises Verfahren (Temperatur → Vorlauf → Wärmeübertrager) wie es Both u. a. [56] entwickelten, umgangen werden. Diese Kopplung würde aber mit stufenweiser, exponentieller Fehleranfälligkeit eine Komplexität in das Verfahren bringen, weshalb diese nicht verwendet wird.

3.5 Datensatz

Datensätze aus Gebäuden sind leider nur selten frei verfügbar [22] und deren Qualität hängt von der Zielsetzung der Erstellung ab [104]. In dieser Arbeit werden zwei unterschiedliche Datensätze

Tabelle 3.2: Statistische Eigenschaften des Mortar-Datensatzes.

Jahr	Anzahl der Datenstromtypen	Anzahl der Datenströme	Datenmenge in GB
2011	33	3.892	1,01
2012	36	4.800	1,24
2013	36	4.697	1,22
2014	39	5.224	1,35
2015	39	5.193	1,35
2016	41	6.515	1,69
2017	42	6.389	1,66
Summe	42	10.665	9,52

verwendet: der Mortar Datensatz [138], der die Entwicklung von skalierbaren Analysefunktionen für Gebäude(automations)daten fördern soll und der während des Projektes OOM4ABDO (BMWK Förderkennzeichen 03SBE0006A) entstandene und aufbereitete Datensatz, der alle Monitoring-Daten der Projektteilnehmer umfasst. Sie werden im Folgenden vorgestellt und auf Basis einer statistischen Analyse bewertet.

3.5.1 Mortar Datensatz

In der Dokumentation der Mortar-Datenbank [138] wird angegeben, dass die Datenbank Zeitreihendaten für 107 Gebäude und damit 26.000 Datenströme in den Jahren 2011 bis 2017 in 15-minütiger Auflösung enthält, wobei nicht jeder Datenstrom über den gesamten Zeitraum (2011-2017) aufgezeichnet wird.

Über die API vom 30.09.2023 der Mortar-Datenbank waren 10.665 Datenströme verfügbar, was nur 41 % der theoretisch verfügbaren Datenströme entspricht. In Tabelle 3.2 ist ein Überblick über die Daten nach Jahr sortiert aufgelistet. Laut Aussage von Gabe Fierro sind die Datenströme bei seinem Übergang seiner Arbeitsstelle verloren gegangen, da er auf die Gesamtdatenbank keinen Zugriff mehr hat [135].

3.5.2 OOM4ABDO Datensatz

Im Projekt OOM4ABDO (ursprünglich: object-oriented monitoring for augmented building and district operation) wurden fünf verschiedene Datenquellen aus sieben Liegenschaften bzw. Quartieren und einem einzelnen Nicht-Wohngebäude in ganz Deutschland genutzt. Es umfasst Daten aus acht Städten mit insgesamt 121.800 Datenströmen. Nicht jeder Datenstrom wurde über die gesamte Projektlaufzeit aufgezeichnet. Die Datenquellen werden im Folgenden vorgestellt.

- **E.ON ERC Hauptgebäude:** ist ein multifunktionales universitäres Gebäude, das mit verschiedenen Energiewandlern versorgt wird (Wärmepumpe mit Geothermiefeld, BHKW und Gas-

Brennwertkessel), die über verschiedene Verteilungssysteme in die unterschiedlichen Räume gebracht werden (Deckeninduktionsgeräte, Quellaufblüftung, Betonkernaktivierung, Fassadenblüftungsgerte, RLT-Anlagen) und 12.790 Datenströme liefert.

- **Stadt Aachen (Logger 1 & 2):** es wurden zwei Logger installiert, die 41.466 Datenströme von 51 Gebäuden bzw. 45.432 Datenströme von 71 Gebäuden aufzeichnen. Diese Gebäude umfassen Kindergärten, Schulen, Verwaltungsgebäude, Museen, Feuerwehren und Sportstätten.
- **Werksviertel Mitte München:** direkt am Ostbahnhof in liegend wird es mit Wärme und Kälte durch Blockheizkraftwerke (BHKW), Gaskessel, Wärmepumpe und Kältemaschine versorgt. Insgesamt liefert das ehemalige Gelände der Firma Pfanni aus fünf Gebäuden 16.420 Datenströme.
- **Helmholtz-Zentrum für Umweltforschung (UFZ) Leipzig:** Das Helmholtz-Zentrum für Umweltforschung (UFZ) besteht aus fünf Standorten: Leipzig, Halle, Magdeburg, Bad Lauchstädt und Falkenberg und liefert aus 17 Gebäuden 13.789 Datenströme.

3.5.3 Vergleich von Datensätzen

Im Folgenden wird ein Vergleich der Datensätze auf Basis der Metadaten der Zeitreihen und der Unterschiede in den statistischen Merkmalen der Zeichenketten auch im Vergleich zur Literatur beschrieben.

3.5.3.1 Vergleich der Zeitreihendatensätze

Es wurde verglichen, wie viele Datenströme gleichzeitig Namen und Zeitreihen in einer ausreichenden Qualität besitzen (mindestens ein vollständiges Jahr aufgezeichnet). Für insgesamt 4.350 Datenströme konnte im OOM4ABDO Datensatz kein Typ ermittelt werden (Anteil 4,7 %). Dies kann auf eine fehlende Beschreibung und eine nicht aussagekräftige Bezeichnung zurückzuführen sein. Deshalb wurden diese nicht in die Auswertung der Zeitreihen einbezogen.

Abbildung 3.8 zeigt die statistischen Unterschiede zwischen dem OOM4ABDO und Mortar Datensatz und die Besonderheiten sind im Folgenden aufgelistet:

- Die schwer zu klassifizierenden Betriebsmeldungen machen im OOM4ABDO Datensatz 25,3 % aus, aber beim Mortar Datensatz nur 0,2 % (Faktor 126).
- Der Messwert der Temperatur ist im Mortar Datensatz um 82 % häufiger vorhanden (36,8 % vs. 20,2 %), während Befehle um 148 % häufiger auftreten (23,3 % vs. 9,4 %).
- Gemessene Volumenströme kommen um 403 % häufiger im Mortar Datensatz vor (9,9 % vs. 2,4 %), obwohl diese in GAS tendenziell seltener aufzufinden sind.

- Andere für Anwendungen in GAS wichtige Klassen wie Energie- oder CO₂-Messwerte sowie Temperatur-Sollwerte tauchen dagegen im Mortar-Datensatz seltener auf.
- Messwerte der Feuchtigkeit fehlen komplett im Mortar-Datensatz, obwohl diese für die Be- und Entfeuchtung notwendig sind.
- Insgesamt werden rund 5,5 % der Klassen der Datenströme des OOM4ABDO Datensatzes im Mortar Datensatz nicht abgebildet.

Die hier vorliegende Auswertung der Datensätze zeigt, dass der OOM4ABDO Datensatz variantenreicher ist und einen höheren Anteil an schwer zu klassifizierenden Klassen enthält als der Mortar Datensatz, was vermuten lässt, dass dessen Klassifizierung eine größere Herausforderung ist. Die höhere Anzahl von Temperatursensoren und Volumenströmen kann in der hohen RLT-Abhängigkeit der US-amerikanischen Heizung, Lüftung, Klimatechnik (HLK) liegen. In reinen Luftsystemen werden an mehr Stellen im Gebäude Temperaturen und Volumenströme gemessen, u.a. durch die höhere Standardisierung der RLT-Systeme.

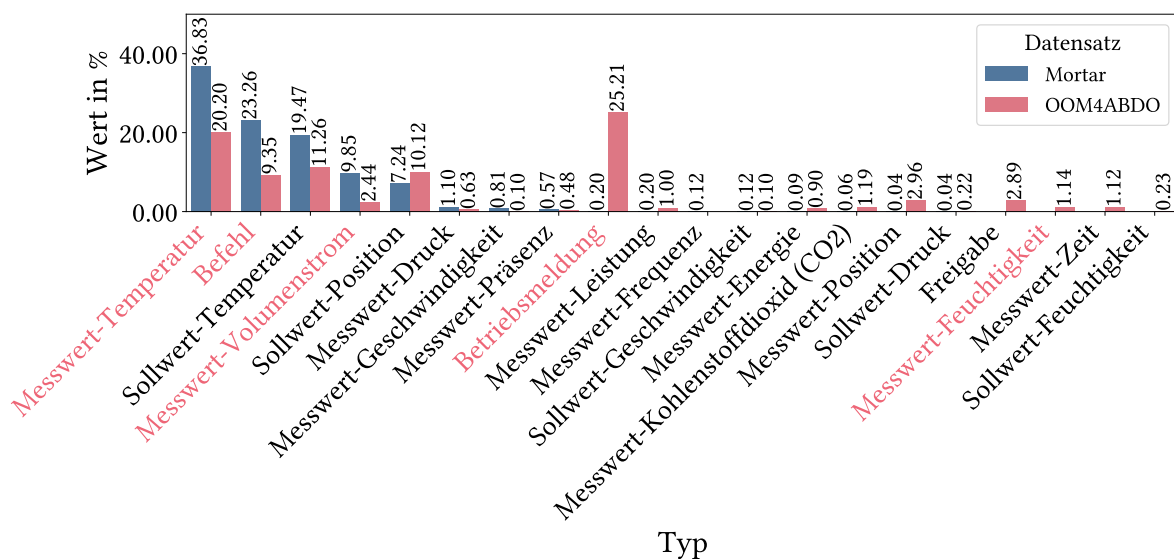


Abbildung 3.8: Vergleich der Signal Typen von Mortar (blau) und OOM4ABDO Datensatz (rot). In roter Schrift sind die Klassen markiert, die gesondert untersucht wurden.

3.5.3.2 Statistischer Vergleich

Zur Bewertung der Komplexität werden in diesem Abschnitt statistische Merkmale der Datensätze verglichen (siehe Kapitel Kapitel E.1 im Anhang). Es wurden insgesamt 59 statistische Merkmale berechnet, die in ihrer Gesamtheit in Tabelle E.2 im Anhang abgebildet sind. Im Folgenden werden die Berechnungen von ausgewählten statistischen Merkmalen beschrieben.

Tabelle 3.3: Formeln der statistischen Merkmale der Datensätze.

Statistisches Merkmal	Formel
Kennzeichen-zu-Wort-Verhältnis	$KW_i = \frac{ s_{k,i} }{ s_{w,i} }$
Kennzeichen-zu-Datenstrom-Verhältnis	$KD_i = \frac{ s_{k,i} }{ n_{ds,i} }$
Wort-zu-Datenstrom-Verhältnis	$WD_i = \frac{ s_{w,i} }{ n_{ds,i} }$
Anzahl-Zeichen-zu-Wort-Verhältnis	$ZW_i = \frac{ n_{z,i} }{ s_{w,i} }$
Anzahl-Zeichen-zu-Datenstrom-Verhältnis	$ZD_i = \frac{ n_{z,i} }{ n_{ds,i} }$

Tabelle 3.4: Statistische Merkmale der Datensätze. Für kursiv markierte Merkmale sind Übersetzungen der Blöcke notwendig (höherer Personalaufwand).

Statistisches Merkmal	ERC	Aachen_1	Aachen_2	Leipzig	München
$\frac{ n_{w,i} }{ n_{ds,i} }$	0,03	0,01	0,01	0,01	0,00
KW_i	1,66	1,80	1,26	0,69	0,98
KD_i	27,30	21,63	22,57	13,23	15,74
WD_i	16,45	12,00	17,86	19,14	16,03
ZW_i	3,99	4,12	3,78	3,39	3,63
ZD_i	65,64	49,39	67,42	64,85	58,13

3.5.3.2.1 Statistische Merkmale Als statistische Merkmale der Datensätze können die durchschnittliche Anzahl von Blöcken pro Datenstrom, Kennzeichnungen pro Datenstrom und Blöcken pro Kennzeichnung herangeführt werden. Die Formelbestandteile der statistischen Merkmale sind wie folgt definiert:

- $|n_{ds,i}|$: Anzahl der Datenströme n_{ds} in Datensatz i
- $|n_{w,i}|$: Anzahl der einzigartigen Wörter n_w in Datensatz i
- $|n_{z,i}|$: Anzahl der Zeichen s_w in Datensatz i
- $|s_{w,i}|$: Summe der Wörter s_w in Datensatz i
- $|s_{k,i}|$: Summe der Kennzeichen s_k in Datensatz i

In Tabelle 3.3 sind die entsprechenden Formeln für die Berechnung der einzelnen Merkmale aufgeführt.

3.5.3.2.2 Auswertung statistische Merkmale In Tabelle 3.4 sind die statistischen Merkmale der OOM4ABDO-Datensätze abgebildet. Es ist sichtbar, dass sich die proportionale Anzahl der einzigartigen Wörter ($|n_{w,i}|$) in allen Datensätzen stark unterscheiden, was sich auch im Kennzeichen-zu-Wort-Verhältnis (KW_i) widerspiegelt.

- Das Kennzeichen-zu-Datenstrom-Verhältnis (KD_i) kann doppelt so hoch sein (13,2 bis 27,3), was die Informationsbreite stark beeinflusst.

- Das Anzahl-Zeichen-zu-Wort-Verhältnis ZW_i unterscheidet sich im Verhältnis zu den anderen Kennwerten nicht sehr stark in den Datensätzen (3,4 bis 4,1).
- Die durchschnittlichen Zeichen pro Wort oder Block liegen zwischen 3,39 bis 4,12 (ZW_i), aber insgesamt gibt es durchschnittlich 12 bis 19,14 Wörter oder Blöcke pro Datenstrom (WD_i). In dem von Chen u. a. [78] verwendeten Datensatz haben dagegen nur 3 % der Datenstromnamen 6 oder mehr Blöcke.
- Die OOM4ABDO Datensätze enthalten im Durchschnitt 49,4 bis 65,6 Zeichen (ZD_i). Der Datensatz von Waterworth u. a. [467] enthält mit im Schnitt 28,7 Zeichen pro Datenstrom weniger Zeichen.

Der Vergleich mit anderen Datensätzen konnte nur auf statistischen Merkmalen durchgeführt werden, die auch in den entsprechenden Publikationen enthalten waren. Hierzu wurden aber jeweils meistens keine Aussagen getroffen oder nur einzelne Merkmale vorgestellt. Einzelne Merkmale können aber nur ein begrenztes Bild des Variantenreichtums eines Datensatzes zeigen. So könnte ein volltextbasiertes System auch eine höhere Anzahl an verwendeten Zeichen aufweisen, was aber eine einfachere Klassifizierungsaufgabe wäre. Auch eine konsistente Benennung kann die Aufgabe erleichtern, obwohl viele Blöcke und Zeichen verwendet wurden.

Es kann nicht vollständig geklärt werden, ob alle Datensätze aus der Literatur diese entsprechenden Statistiken aufweisen. Trotzdem kann ausgesagt werden, dass der hier verwendete Datensatz deutlich variantenreicher ist als die in der Literatur auswertbar verwendeten.

3.5.4 Übersetzung für Ground Truth

Im maschinellen Lernen werden mit Ground Truth die Basisdaten bezeichnet, für die bekannt ist welche Ergebnisse sie liefern müssen [213]. Für diese Arbeit wurden Bezeichnungen der Datenströme in allen OOM4ABDO-Datensätzen in eine Ground Truth übersetzt. Blöcke werden an einer bestimmten Stelle voneinander getrennt und im Gesamtzusammenhang übersetzt, wenn eine der folgenden Eigenschaften in diesem Zeichen der Zeichenkette auftreten:

- Wechsel von Kleinschreibung zu Großschreibung
- Wechsel von numerischem Zeichen zu alphabetischem Zeichen und andersherum
- Verwendung eines Sonderzeichens wie z.B. Leerzeichen, Punkt, Bindestrich oder Unterstrich

Bei Blöcken, die mehr als vier Zeichen lang sind und nicht nur aus Großschreibung bestehen (Beispiel: TMPTR), wurde angenommen, dass diese natürliche Wörter sind und deshalb in allen Datensätzen gleichermaßen verwendet werden, was trotz Fehlern ein guter Kompromiss ist.

Die Trennung und Übersetzung der jeweiligen Wörter und Blöcke wurde durch einen eigens entwickelten Algorithmus durchgeführt, der in Abbildung 3.9 dargestellt ist. Nach der Trennung wird anfangend von dem ersten und letzten Block (Beispiel: Raumbdiengerät, Sollwertschiebung, Pre,

Comf) jede Blockkombination auf schon übersetzte Blockkombinationen überprüft. Wenn diese Blockkombination noch nicht in der Übersetzungstabelle auftaucht, wird mit der Blockkombination von dem zweiten Block bis zum letzten Block (Beispiel: Sollwertschiebung, Pre, Comf) in einem zweiten Schritt fortgeschritten. Wenn eine Blockkombination gefunden wird, dann wird mit dem vorherigen Block (Beispiel: Raumbediengerät) als nächstes Ende der Blockkombination fortgefahren. Wenn alle Blöcke übersetzt sind, werden diese mit ihren Übersetzungen ausgegeben. Insgesamt konnten 9.300 Übersetzungen von Blöcken bzw. Blockkombinationen ermittelt werden.

In dieser Dissertation musste für jeden Block entschieden werden, ob dieser übersetzt werden soll. In der Praxis würden pragmatisch nur die für die jeweilige Organisation und die gewünschten Anwendungen sinnvollen Blöcke übersetzt.

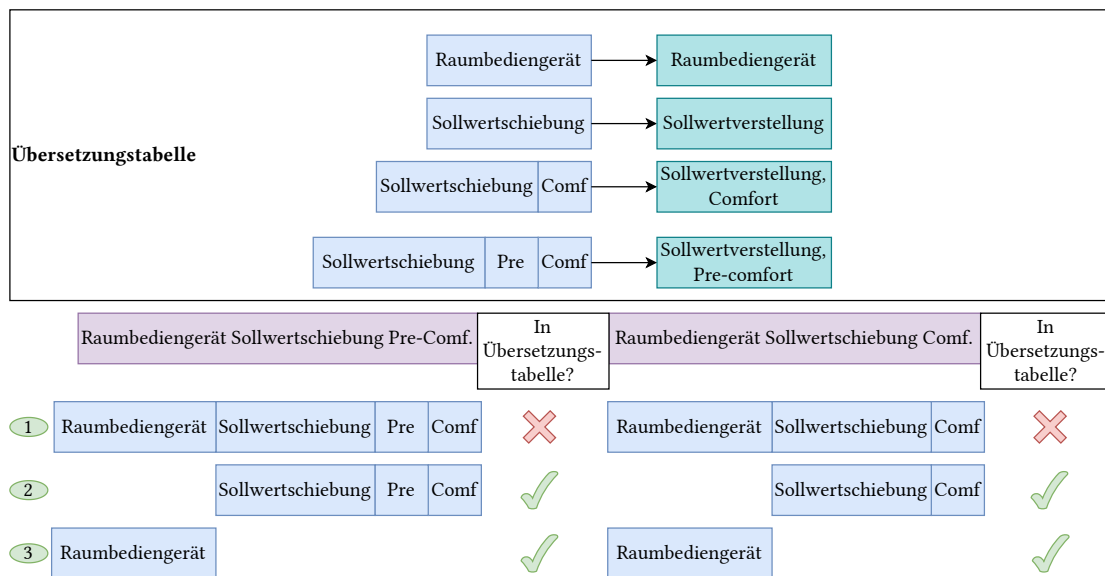


Abbildung 3.9: Vorgehen zur Erstellung der Ground Truth. Es wird von Anfangsblock und Endblock ausgehend jede Blockkombination auf eine Übereinstimmung in der Übersetzungstabelle überprüft. Schon übersetzte Blockkombinationen werden im weiteren nicht mehr weiter betrachtet.

Insgesamt sind im Datensatz 2,7 Mio. Kennzeichen enthalten in 2,1 Mio. Wörtern bzw. Blöcken. Für jedes Kennzeichen und jeden Block wurde überprüft, ob er von einer automatisierten Übersetzung auf Basis des vorher beschriebenen Algorithmus korrekt zugeordnet wird.

Auch wurde anhand des BUDO Schemas überprüft, ob sinnvolle Kombinationen von Kennzeichen in einem Label zugewiesen wurden. So könnte automatisiert gleichzeitig das Medium Luft und Wasser zugewiesen werden. Dies könnte jedoch auch korrekt sein, wenn in einem Luftsystem (z.B. RLT-Anlage) eine Wassertemperatur gemessen wird. Für jeden der insgesamt 1.700 Konflikte wurde entschieden, wie dieser zu behandeln ist. Dies konnte eine Änderung der verwendeten Kategorien, aber auch eine Zusammenführung von Kennzeichen umfassen. In dem Fall des gleichen Mediums

könnte Luft als übergeordnetes System-Medium und Wasser als Mess-Medium unterschieden werden.

Für die Zeitreihendaten können die Daten des OOM4ABDO-Projekts und der Mortar-Datenbank analysiert werden, während die Zeichenketten der Mortar-Datenbank durch die Normalisierung nicht verwendet werden können, um realitätsnahe Datensätze zu verwenden. Deshalb werden für die Verwendung der Zeichenketten nur die OOM4ABDO-Datensätze verwendet.

3.6 Anwendungen in Energiesystemen

Im Folgenden werden verschiedene Anwendungen von Optimierungen in Energiesystemen vorgestellt, deren Implementierung von strukturierten Gebäudeautomationsdaten profitieren. Für viele Anwendungen fehlt aber skalierbares Anwendungswissen [196] und es werden unterschiedliche Datenstromarten benötigt [57, 357], die schwer zu ermitteln sind wegen fehlender Standardisierung [178].

In dieser Arbeit soll ein Ansatz entwickelt werden, der die Anwendungen mithilfe eines IE-Prozesses mit informativen Metadaten aus GAS versorgt.

Die zentralen Anwendungen sind in dieser Arbeit:

- Fehler- und Anomalieerkennung (Kapitel 3.6.1),
- Regelung (Kapitel 3.6.2),
- Energiesystemanalyse (Kapitel 3.6.3).

Es wird für jede Anwendung herausgearbeitet, welche Datenstromtypen für diese erforderlich sind (semantische Suffizienz) (vollständige Statistiken: Kapitel E.4)) und in welchem Vertrauensbereich sich die Ergebnisse der Anwendungen befinden. Als Algorithmen zur Zuordnung von Datenströmen zu anwendungsbasierten Einheiten wie Regelkreis, Raum oder Energiezähler, werden neben Ähnlichkeitswerten von Datenstromnamen auch überwachtes Lernen auf Basis von Zeitreihen verwendet.

3.6.1 Fehler- und Anomalieerkennung

Die Fehler- und Anomalieerkennung (engl. Fault Detection and Diagnosis: FDD) in energetischen Anlagen im Gebäude und Quartier ist ein Thema, dem sich aufgrund von Fortschritten in KI-Modellen dem sich die Forschungswelt immer stärker zuwendet [183, 204, 240, 273, 415, 504]. Der skalierbare Einsatz von Gebäudeautomationsdaten zur Detektion von Fehlern im Gebäudebetrieb ist herausfordernd [311].

Roth u. a. [398] und Burak Gunay u. a. [65] geben einen Überblick über typische Fehler in Gebäudeautomationssystemen und ihren Einfluss auf den Energieverbrauch. Demnach sind Pumpen,

RLT-Anlagen, Licht und nicht bedarfsgerechte eingeschaltete HLK entscheidende Fehler für den Energieverbrauch. Deshalb werden die Heizung und Lüftung in diesem Anwendungsgebiet besonders betrachtet.

Mithilfe eines guten semantischen Modells, das oftmals bei Datensätzen fehlt [229], können Fehler in Energiesystemen eventuell automatisch erkannt und (assistiert) beseitigt werden [283, 284, 365, 376]. Für die Fehlererkennung fehlen wie bei der Metadatenextraktion oftmals Datensätze [333], aber deren Anzahl hat sich etwas verbessert [72]. Deshalb werden für die folgenden Ansätze eigene Datensätze gebildet, die entsprechend ausgewertet werden:

- Kapitel 3.6.1.1: Fehler- und Anomalieerkennung in raumluftechnische Anlagen als Beispiel für technische Anlagen,
- Kapitel 3.6.1.2: Wärmeübertrager als Schnittstelle im System,
- Kapitel 3.6.1.3: Außentempertursensoren als zentraler Sensor für die Gebäuderegelung.

3.6.1.1 Fehler- und Anomalieerkennung bei raumluftechnischen Anlagen

Raumluftechnische Anlagen sind insbesondere in Nichtwohngebäuden häufig anzutreffen [16]. Sie werden zur Konditionierung der Raumluf verwendet (heizen, kühlen, Frischluftzufuhr, Be- bzw. Entfeuchtung und Filterung der Luft).

Die Präzision von Algorithmen zur Fehlererkennung kann bei regelbasierten Systemen bis zu 94 % betragen [263]. Ebrahimifakhar u. a. [121] und Crowe u. a. [96] zeigen die meist auftretenden Fehlertypen in RLT-Anlagen in einem Datensatz von 1.360 Gebäuden auf. Ein Fehler im Sensor selbst wird nur selten überprüft, obwohl hier ein eventuell hohes Potential bei der Einsparung von Energie liegt [499].

Li u. a. [273] haben in einem Review verschiedene Fehler- und Anomaliedetektionsalgorithmen für raumluftechnische Anlagen und Kältemaschinen auf Basis ihrer verwendeten Datenstromtypen untersucht. Von 100 Publikationen konnten nur 50 für raumluftechnische Anlagen genutzt werden, wobei diese 129 verschiedene Fehlerbilder enthielten. Die verwendeten Sensoren und Leistungsmerkmale werden im Anhang in Kapitel E.4.1 untersucht und in spezifische Fehlergruppen eingeteilt. Insgesamt konnten 119 verschiedene Datenstromtypen identifiziert werden, wobei die meisten nur ein (52) oder zweimal (20) vorkommen.

Die quantitative Bewertung der Fehlererkennung bei RLT-Anlagen wird analog zu der Bewertung der Operationalisierbarkeit (Kapitel 3.1.4.7) vorgenommen. Im Folgenden wird der Datensatz und das Vorgehen bei diesem Anwendungsfall vorgestellt.

3.6.1.1.1 Datensatz für raumluftechnische Anlagen Im Datensatz von OOM4ABDO wurden 68 RLT-Anlagen mit insgesamt 375 Datenströmen identifiziert. Von den 129 in den wissenschaftlichen Aufsätzen untersuchten Fehler in RLT-Anlagen konnten anhand des in dieser Arbeit

verwendeten Datensatzes nur für 26, also knapp 20 %, der Fehler alle benötigten Datenstromtypen gefunden werden. In keiner RLT-Anlage konnten alle Fehlerbilder untersucht werden, da keine Anlage alle hierfür nötigen Datenströme enthält (maximal 17 Fehler).

Interessanterweise existieren nur wenige Datensätze, die jeweils die gleichen Fehler abdecken können. Für eine skalierbare Anwendung von z.B. Zeitreihenanalyse wäre eine Mindestbasis an Datenstromtypen erforderlich, die in allen RLT-Anlagen vorhanden ist oder eine deutlich größere Anzahl an RLT-Anlagen. Diese geringe Anzahl an nutzbaren Fehlern zeigt, dass eine Beachtung der vorhandenen Datenströme in der Forschung für die Praxis eine nutzbare Fehleranalyse liefern würde.

3.6.1.1.2 Vorgehen bei raumluftechnischen Anlagen In diesem Anwendungsfall wird überprüft, ob zum einen die Datenstromtypen der einzelnen Datenströme automatisiert erkannt werden können. Zum anderen wird untersucht, ob Datenströme anhand des Datenstromnamens auf Basis einer Ähnlichkeitsbewertung (Kapitel 3.3.8) zu einer RLT-Anlage geclustert werden können. Auf Basis einer Parameterstudie werden passende c_i für die Anwendung ermittelt.

Um eine Eingrenzung von der Anzahl der erforderlich zu erkennenden Datenströme vorzunehmen, wird die maximal verfügbare Anzahl an Datenströmen in einer RLT-Anlage herangeführt, die 15 beträgt. Von dieser ausgehend wird eine höhere Zahl ausgewählt (hier: 20), die eine obere Grenze der zu erkennenden Datenströme symbolisieren soll. Zusätzlich wird die exakte Zahl n an Datenströmen der jeweiligen RLT-Anlage verwendet. Jede Distanz SSC_i wird mittels aller 375 Datenstromnamen berechnet. Hieraus werden dann für die jeweiligen Grenzen die passenden Top-Werte ermittelt:

- Oberste Grenze: Top 20,
- Mittलगrenze: Top 15,
- Unterste Grenze: Top n .

Diese Grenzauswahl wird auch bei jedem der folgenden Ansätze mit einer Ähnlichkeitsbewertung vorgenommen.

3.6.1.2 Fehlererkennung bei Wärmeübertragern

Wärmeübertrager sind die Schnittstelle zwischen verschiedenen Wärme- bzw. Kälteströmen im Energiesystem. Für jeden Mediumswechsel (z.B. von Wasser zu Luft) und um hydraulische Systeme zu entkoppeln muss ein Wärmeübertrager eingesetzt werden. Die Wärmeverteilung weist bei etwa 78 % der Gebäude in einer Stichprobe von Mills [320] Mängel auf. Die genaue Wärmeverluste im Gebäude sind schwierig zu ermitteln, aber betragen annäherungsweise 6 % der eingesetzten Wärme [374]. Aus diesen Gründen muss die Nutzung der Energieströme und ihrer Temperaturniveaus untersucht werden.

Für die Informationsextraktion für die Fehlererkennung in Wärmeübertragern wurden 53 Wärmeübertrager mit 856 zugehörigen Datenströmen im OOM4ABDO Datensatz identifiziert. Wichtig für die Analyse von Fehlern in Wärmeübertragern sind die Temperaturen wie Vor- und Rücklauf und der Betrieb der Wärmeübertrager. Die Datenströme umfassen 202 Temperaturmessungen und 89 Regelsignale, die meistens Sollwerte darstellen. Der Rest sind allgemeine Alarmer, Betriebsmeldungen oder Positionen.

Da die maximale Anzahl der überprüfbaren Datenströme pro Wärmeübertrager mit 42 größer ist als bei den RLT-Anlagen werden die „Top 42“ und „Top 60“ der Datenströme betrachtet.

3.6.1.3 Fehlererkennung bei Außentemperatursensoren

Die Außentemperatur ist zentral für die Vorhersage der verwendeten Energie innerhalb eines Gebäudes [27, 87, 108, 174, 476]. Der Zusammenhang ist auch für die Regelung entscheidend, da die Raumtemperaturen meist anhand der Außentemperatur angepasst werden [87, 476]. Da es sich bei der gemessenen Außentemperatur um eine zentrale Größe in der Gebäudeautomation handelt, wird der Außentemperatursensor als Beispiel für die Fehlererkennung von einzelnen Sensoren herangeführt. In Stinner u. a. [431] wurde die Fehlererkennung von Außentemperatursensoren mithilfe von Zeitreihen vorgestellt, wobei ein F1 Wert von bis zu 92,3 % erzielt wurde. In dieser Arbeit wird anhand von Zeitreihen und Zeichenketten näher untersucht, ob die Außentemperaturmessung von anderen Datenstromtypen unterschieden werden kann.

3.6.2 Regelung

Die Optimierung bestehender Regelungen oder die Implementierung von neuartigen Regelungen wie einer modellprädiktiven Regelung erfordert aktuell ein hohes Systemwissen und somit hohen personellen Aufwand. Die Regelungsentwicklung in der Gebäudeautomation sollte über den Lebenszyklus eines Gebäudes dokumentiert und zwischen den Beteiligten ausgetauscht werden [443], was jedoch auch aufgrund fehlender Standardisierung bzw. Komplexität nicht der Fall ist. Um die immer komplexer werdenden Regelungssysteme stabil und optimiert zu steuern, müssen diese durch neue Lösungen permanent verbessert werden. Nichtsdestotrotz werden in der Gebäudeautomation nur rudimentäre innovative Steuerungsagentenfunktionen eingesetzt [265].

In dem Anwendungsbereich der Regelung wird gezeigt, ob Rohdaten aus dem Gebäude so umgewandelt und auch Zusammenhänge zwischen den Daten hergestellt werden können, dass die Implementierung von Anwendungen in der Regelung unterstützt werden kann. In Kapitel 3.6.2.1 wird das Beispiel der PID (engl. proportional-integral-derivative) Regelung als kleinste Regeleinheit im Gebäudeautomationssystem herangezogen. Darauf folgend wird in Kapitel 3.6.2.2 die modellprädiktive Regelung und deren Anforderungen vorgestellt.

3.6.2.1 Regelung der Raumtemperatur

Die organisatorisch kleinste Einheit für die Regelung im GAS sind die verstärkend (P), integrierend (I) und differenzierend (D) (engl. Proportional-integral-derivative) (PID)-Regelkreise von Raumtemperaturen mit ihren Sollwerten und Messwerten (siehe Abbildung 3.10). Um die Qualität eines in der Regel schlecht eingestellten [148, 149, 340, 502] PID Regler zu überprüfen, müssen die Sollwerte und die Messwerte zusammen betrachtet werden [2, 148, 149, 340]. Hindernisse für deren Optimierung sind meist Budget-Gründe bzw. zu wenig eingeräumte Zeit für die Erstellung der Gebäudeautomation. Hierbei könnten automatisierte Regelloptimierungswerkzeuge [149, 427] helfen, die jedoch den Regelkreis als semantische Einheit benötigen.

Ein Regelkreis könnte automatisiert über temporale Zusammenhänge (z.B. Einschalten morgens) erkannt werden. Für die Erzeugung von temporalen Zusammenhängen sind Versuchsdaten erforderlich, die zum einen nicht skalierbar durchgeführt werden können, aber auch detailliertes Wissen über die Regelkreise erfordern [364] und deren Nutzen ist fraglich [161]. Sie können jedoch zur Qualitätsüberprüfung genutzt werden [45, 161, 427], die nicht Teil dieser Arbeit ist.

Für die automatisierte Auswertung von PID-Regelungen müssen die entsprechenden Sollwerte mit den zugehörigen Messwerten gepaart werden. Hierfür sind insbesondere die Datenstrombeschriftungen eine nützliche Quelle. In Stinner u. a. [429] ist die genaue Einstellung der Parameter für die Sollwerterkennung beschrieben.

Da es pro Sollwert nur einen zuzuordnenden Messwert gibt, werden die „Top 5“, „Top 3“ und „Top 1“ bestimmt. Insgesamt hat der verwendete Datensatz 691 Temperatur-Sollwerte bzw. -Sollwertanpassungen und 183 Temperatursensoren.

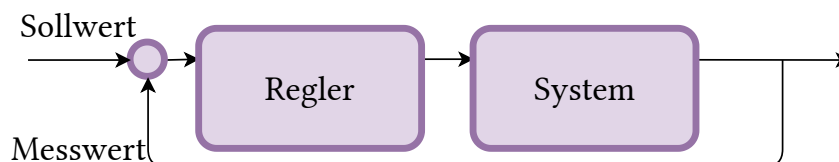


Abbildung 3.10: Grundsätzlicher Aufbau einer PID Regelung. Für den in dieser Arbeit betrachteten Anwendungsfall sind die Eingangssignale (Messwert und Sollwert) relevant.

3.6.2.2 Modellprädiktive Regelung

Modellprädiktive Regelung (MPC) ist ein Trendthema für Gebäudeenergiesysteme, das sich immer weiter beschleunigt [65]. Bei MPC wird ein Modell verwendet, um zukünftige Zustände eines Gebäudes vorherzusagen und dann auf diese zukünftigen Zustände a priori zu reagieren. Energieeinsparungen von 15 bis 50 % [114] können mithilfe von MPC simulativ ermittelt werden. Eine Reduzierung der Komplexität von technischen Anlagen [444] und große Totzeiten wie sie in Gebäuden

vorherrschen [100] bietet ein höheres Potential für die Anwendung von höheren Regelungsalgorithmen wie MPC, aber in der Realität werden diese aufgrund fehlender Schnittstellen nur selten eingesetzt [69, 265, 350].

Drgoňa u. a. [114] definieren sechs Hürden für die Anwendung von MPC in der Realität. In dieser Arbeit wird eine der Hürden, die Plug-and-Play-Implementierung, näher untersucht. Gebäudemodelle sind ein entscheidender Bestandteil der Verwendung von modellprädiktiver Steuerung in Gebäudeautomationssystemen [368] und deren Robustheit ist ein großes Problem in der Realisierung der MPC [225]. Ziel ist es, festzustellen, ob die verwendeten Algorithmen zur Informationsextraktion in der Lage sind, die für die Modellierung wesentlichen Datenstromtypen automatisch zu erkennen.

Welche Datenströme für die Bewertung wesentlich sind, hängt von der Definition der MPC-Zielfunktion ab [306]. Zhan u. Chong [496] definieren für die Bewertung fünf Level der Messgranularität für die folgenden sechs Datenkategorien:

- Energieverbrauch
- Innenraumbedingungen
- Systemzustand
- Umhüllungszustand
- Interne Störgrößen
- Außenbedingungen

Nur die drei Kategorien Energieverbrauch, Innenraumbedingungen und Außenbedingungen sind überhaupt durch die Daten aus der Gebäudeautomation abdeckbar. Die Bewertung des Modells hängt von der Zielgröße (Energieverbrauch) und zentralen Randbedingungen (Komfort, Außenbedingungen) ab. Die anwendungsbezogene Bewertung eines Algorithmus richtet sich daher danach, ob er in der Lage ist, die zentralen Datenströme für Energieverbrauch (elektrische Energie, Wärmeenergie, Gas- und Wasserzähler) und Innenraumbedingungen (Raumtemperatur, CO₂- und TVOC-Sensoren (Gesamtsumme der Emissionskonzentration, engl. Total volatile organic compounds)) und Außenbedingungen (Außentemperatur, solare Einstrahlung) zu erkennen.

Bei der Einführung von modellprädiktiven Regelungen in Gebäuden ist es nur durch ein sehr komplexes Verfahren möglich die notwendigen Datenströme direkt einem Gebäude zuzuweisen. Meistens sind die Eigenschaften von Außenbedingungen, Raumluftqualität und energetischem System nicht semantisch miteinander verknüpft. Deshalb wird auf die direkte Zuweisung zu einem Gebäude im Rahmen dieser Arbeit verzichtet. Die Anwendung „Modellprädiktive Regelung“ wird im Rahmen der Bewertung der Detektion von Datenstromtypen bei Zeichenketten (siehe Kapitel 4.1.2) und bei Zeitreihen (siehe Kapitel 4.2) betrachtet.

3.6.3 Energiesystemanalyse

Die Analyse eines Energiesystems als Ganzes ist sinnvoll, da der Energiefluss im Energiesystem betrachtet werden kann und Ineffizienzen direkt im Betrieb erkannt werden können. Insbesondere Energie-Dashboards sind für Facility Manager nützlich, um Systemfehler zu erkennen oder Energiemaßnahmen zu planen [442], wobei der Einfluss auf den Nutzer gering ist [442].

Im Rahmen dieser Arbeit wird in Kapitel 3.6.3.1 das Raummonitoring als kleinste Einheit eines Quartiersenergiesystems betrachtet. Dann folgt in Kapitel 3.6.3.2 eine Energieflussanalyse, also eine Analyse welche Temperaturen auch eine stoff- oder energieströmliche Verbindung haben. Auf Basis der Energieflussanalyse kann in Kapitel 3.6.3.3 ein gesamtes Quartiersmonitoring mit allen Verbindungen ermittelt werden.

3.6.3.1 Raummonitoring

Die Menschen in den Industrieländern verbringen mehr als 90 % ihrer Zeit in Räumen in Gebäuden [144], weshalb der Energieverbrauch in den Räumen gesondert betrachtet werden sollte. Die Sensoren und Aktoren für die Raumversorgung zeigen schlussendlich an, wie die Energie für die Menschen im Gebäude verwendet wird. Neben der Qualitätssicherung des Komforts [250, 400] hilft das Raummonitoring auch die Raumluftbehandlung nur auf die Gebäude- und Raumbeleuchtungszeit zu beschränken und unterstützt somit eine effiziente Energie- und Luftversorgung eines Gebäudes [400]. Wenn sich mehrere Nutzer in einem Raum aufhalten, kommt es häufig zu Konflikten in Bezug auf den Komfort [386], wobei die Temperatur und Raumluftqualität als entscheidend angesehen werden [343].

Das Energieeinsparpotential aufgrund von nutzerzentrierten HLK-Steuerungen und Verhaltensänderungen kann je nach Gebäudeausstattung und Klimazone auf 3 bis 45 % geschätzt werden [344, 500]. Durch ein gezieltes Raummonitoring können die Einsparpotentiale und Regelungen auf Basis von Nutzervorlieben realisiert werden.

Im Fall des Raumes ist die Herausforderung bei einem automatisierten Vorgehen alle zu einem Raum zusammengehörige Datenstromdaten semantisch zu clustern, was jedoch durch die Variation der Sensortypen verhindert wird [348]. Für die Umsetzung wird deshalb die Qualität des automatischen Mappings auf die benötigten Sensoren für die Raumkonditionen (Temperatur-, Feuchtigkeits-, CO₂- und TVOC-Sensoren) analysiert.

Insgesamt konnten im OOM4ABDO Datensatz 186 Räume mit insgesamt 1.156 zugehörigen Datenströmen identifiziert werden. Der Datensatz umfasst auch weitere Räume, aber im Rahmen der Arbeit wurde pragmatisch entschieden nur eine für die Analyse ausreichende, aber begrenzte Anzahl an Räumen zu betrachten. Datenströme, die in Räumen enthalten sein können, sind vor allem Temperaturen und zugehörige Sollwerte bzw. Sollwerte bei unterschiedlichen Qualitätsmerkmalen (z.B. Komfort). Aber auch Feuchtigkeits- bzw. Raumluftqualitätssensoren und -sollwerte oder die

Gesamtkonzentration von flüchtigen organischen Verbindungen (TVOC) sind in den Raumdaten enthalten.

Auf Basis der maximalen Anzahl von 14 Datenströmen in einem Raum werden die „Top 14“, „Top 25“ und „Top n“ bestimmt.

3.6.3.2 Energieflussanalyse/Topologieerkennung

Energieflussanalysen u.a. mithilfe von Sankey-Diagrammen bieten die Möglichkeit die energetische Nutzung nachzuvollziehen und Fehler in Energiesystemen zu erkennen. Insbesondere können untypische Nutzer erkannt werden und bei der Optimierung der Energienutzung die Zeit auf die größeren Verbraucher konzentriert werden.

Energieflussdarstellungen in RLT-Anlagen [3] und industriellen Prozesse [424] auf Landesebene konnten schon in Sankey-Diagrammen realisiert werden, wobei eine manuelle Verknüpfung erforderlich war. Soweit dem Autor bekannt ist, gibt es keinen automatisierten Ansatz zur Detektion von Energieflüssen in Energiesystemen.

Die automatisierte Berechnung von Wärmeströmen ist fehleranfällig, da zwei Temperaturen und ein zugehöriger Volumenstrom miteinander in Beziehung gesetzt werden müssen. Daher wird in dieser Arbeit ein Vergleich zwischen der Anwendung der Topologieerkennung von Wärme- und Temperaturflüssen vorgenommen. Temperaturflüsse berücksichtigen nur die Temperaturänderung innerhalb der Gebäudeenergiesysteme von einem hohen zu einem niedrigeren Temperaturniveau.

Es werden aus dem OOM4ABDO Datensatz 948 Temperatursensoren ausgewählt und 855 verschiedene Verbindungen identifiziert. Im Gegensatz zu vorherigen Ansätzen wird in dieser Arbeit ein mediumsunabhängiger Ansatz verfolgt. Es wird im zentralen Ansatz nicht unterschieden, ob die Energieübertragung durch Wasser-Wasser-, Wasser-Luft-, Luft-Wasser- oder Luft-Luft-Verbindungen stattfindet. Das Medium wird nur betrachtet, um zu erkennen, welche Informationsqualität die Detektion der einzelnen Übertragungsarten besitzen.

Im Datensatz wird jedem „richtigem“ Paar ein „falsches“ Paar aus einem anderen Datensatz zugewiesen, der, wie in Tabelle 3.5 abgebildet ist, jeweils die Datenströme „von“ und „zu“ des „richtigen“ Pairs ersetzt. Dieses Vorgehen führt zu 855 verbundenen und 1710 nicht verbundenen Datenströme, die analog zur Zeitreihenklassifikation in Kapitel 3.4 bewertet werden.

Tabelle 3.5: Beispiel für die Definition des Datensatzes in der Anwendung „Energiefluss“. Das originale Paar mit der ID 1 wird mit einem anderen Paar gemixt, wobei „von“ und „zu“ jeweils mit dem anderen Paar ausgetauscht werden.

ID	von	zu	Verbunden
1	bacnet115-005-H-HGrp1-TFl	bacnet115-005-H-HGrp1-TRt	Ja
2	bacnet115-005-H-HGrp1-TFl	bacnet387-H062-AEMWB03_T-RL	Nein
3	bacnet387-H062-AEMWB03_T-VL	bacnet115-005-H-HGrp1-TRt	Nein

3.6.3.3 Quartiersmonitoring

Als größte energetische Versorgungseinheit in dieser Arbeit wird das Quartiersmonitoring betrachtet. Betreiber von Quartiersenergielösungen haben wenig Zeit, ihren Betrieb zu optimieren. Automatisierte Auswertungen von Gebäude- und Quartiersdaten helfen daher, Energieeinsparpotenziale zu identifizieren [5, 431]. So konnten Stinner u. a. [431] zeigen, dass fehlende Abschaltungen der Energieversorgung in Nichtnutzungszeiten aus Zeitreihendaten erkannt werden können. Auch für das Quartiersmonitoring ist ein suffizientes semantisches Modell entscheidend [176], das aktuell in der Regel fehlt.

Für das Beispiel eines Quartiersmonitoring wird der Datensatz des Werksviertels herangezogen. In diesem sind 138 Wärmeübertrager aus der zentralen Versorgung, aber auch einzelnen Gebäuden, mit insgesamt 803 Datenströmen enthalten.

Im ersten Schritt werden die Datenströme zu Wärmeübertragern geclustert und dann werden die Temperaturen mithilfe von überwachtem Lernen einer Verbindungsüberprüfung unterzogen, wie sie im vorherigen Kapitel 3.6.3.2 dargestellt ist. Im Gegensatz zu dem Ansatz aus der Energieflussanalyse werden nur Wasser-Wasser-Mediumsübertragungen verwendet. Insgesamt können im Datensatz vom Werksviertel 1.011 variantenreiche Verbindungen zwischen 108 (Fern-)Wärme- und 31 (Fern-)Kälteübertragern definiert werden.

3.6.4 Zusammenfassung Anwendungen

Es wurden acht den Energieeinsatz optimierende Anwendungen aus drei Bereichen ausgeführt, wobei zusammenfassend folgende Verfahren für die Ergebnisproduktion verwendet werden, die auch in Abbildung 3.11 zu sehen sind:

- zwei Anwendungen nur allgemein in den Ergebnissen: „MPC“ und „Fehlererkennung bei Außentemperatursensoren“,
- vier auf Basis von Ähnlichkeitsanalysen der Zeichenketten: „RLT-Anlagen“, „Fehlererkennung bei Wärmeübertrager“, „Regelung der Raumtemperatur“, „Raummonitoring“,
- Energieflussanalyse/Topologieerkennung mithilfe einer Zeitreihenanalyse und
- das Quartiersmonitoring durch Kombination von Ähnlichkeits- und Zeitreihenanalyse.

Insgesamt müssen zwei (Energiefluss) bis 55 („RLT-Anlagen“) Kennzeichentypen detektiert werden, wobei diese im OOM4ABDO-Datensatz 24 („nach Wärmeübertrager“) bis 43.719 (Temperatur) mal enthalten sind.

Fehlererkennung	Fehlererkennung RLT	Meta	Cluster ZK	
	Fehlererkennung Wärmeübertrager	Meta	Cluster ZK	
	Fehlererkennung Außentemperatur	Meta		
Regelung	Regelung Raumtemperatur	Meta	Cluster ZK	
	Modellprädiktive Regelung	Meta		
Energiesystem	Raummonitoring	Meta	Cluster ZK	
	Energiefluss/Topologie	Meta	Cluster ZR	
	Quartiersmonitoring	Meta	Cluster ZK	Cluster ZR

Abbildung 3.11: Übersicht über alle verwendeten Verfahren im Bereich „Anwendungen“. Meta bedeutet die Bewertung des Metadatenextraktionsverfahrens, Cluster ZK das Clustern auf Basis von Zeichenketten und Cluster ZR das Clustern auf Basis von Zeitreihen.

4 Ergebnisse

Nachdem im vorherigen Kapitel die entwickelte Methodik aufgezeigt wurde, die anhand eines Informationsextraktionsprozesses die benötigten Metadaten für potentiell energieeinsparenden Anwendungen extrahiert. In diesem Abschnitt werden die Ergebnisse der Methodik anhand von Informationsqualitätskriterien vorgestellt:

- Kapitel 4.1: Ergebnisse der Metadatenextraktion anhand von Zeichenketten,
- Kapitel 4.2: Ergebnisse der Metadatenextraktion anhand von Zeitreihen,
- Kapitel 4.3: Ergebnisse der Zusammenführung von Datenströmen zu Anwendungen (z.B. raumluftechnischen Anlagen),
- Kapitel 4.4: Bewertung des gesamten IE-Prozesses anhand aller vorgestellten Informationsqualitätskriterien

4.1 Metadatenextraktion auf Basis von Zeichenketten

In diesem Kapitel werden die Ergebnisse der Metadatenextraktion anhand der Datenstromnamen und -beschreibungen vorgestellt. Hierfür wird nur der OOM4ABDO Datensatz mit seinen fünf Unterdatensätzen verwendet. Im Anhang in Kapitel C wird die Metadatenvorverarbeitung genauer beschrieben.

In Abbildung 4.1 ist der Ablauf der Untersuchung der Merkmalsextraktion aus Zeichenketten von der Auswahl der zu untersuchenden Informationsqualitäten (1) bis hin zur anwendungsbezogenen Auswertung (7):

1. **Auswahl Informationsqualitäten:** Kapitel 4.1.1 beschreibt welche Informationsqualitäten in der folgenden Auswertung verwendet werden. Die nicht in dieser Analyse verwendeten Informationsqualitäten werden in die Gesamtbewertung (Kapitel 4.4) einbezogen.
2. **Vergleich Merkmalsextraktion und Algorithmus:** Kapitel 4.1.1.1 stellt die Ergebnisse der unterschiedlichen Merkmalsextraktionsverfahren und Algorithmen aus den Kapiteln 3.3.1 und 3.3.2 vor.
3. **Vergleich Merkmalssatz:** Die Ergebnisse der einzelnen Merkmalssätze werden in Kapitel 4.1.1.2.1 vorgestellt.
4. **Vergleich Trainingsanteil:** Ziel aller Verfahren ist es mit möglichst wenig Aufwand die Metadaten zu extrahieren. Der Aufwand ist linear abhängig von der Menge der benötigten Trainingsdaten. Welche Ergebnisse bei welchem Trainingsanteil erzielt werden können, wird in Kapitel 4.1.1.3 dargestellt.
5. **Eingrenzung Nutzung Merkmale:** In der Theorie wird mit aussagekräftigeren und weniger diversen Merkmalen ein besseres Ergebnis erzielt. Auch die Nutzung von mehr natürlicher Sprache sollte in der Theorie die Ergebnisse verbessern. Ob diese Annahmen auch in der Realität zutreffen, wird in Kapitel 4.1.1.4 demonstriert.
6. **Gebäudeübergreifendes Training:** In der Praxis ist es unrealistisch, dass jedes Gebäude einzeln in allen Bestandteilen benannt wird. Deshalb wird in Kapitel 4.1.1.5 aufgezeigt, inwieweit ein gebäude- oder schemaübergreifendes Training möglich ist. Diese Modelle könnten bei ausreichenden Ergebnissen auch in der Realität skalierbar eingesetzt werden.
7. **Anwendungsbezogene Bewertung:** Eine anwendungsbezogene Bewertung der Merkmalsextraktion wird in Kapitel 4.1.2 vorgenommen. Diese zeigt auf, wie gut die Metadaten, die für die ausgewählten Anwendungen erforderlich sind, von den Algorithmen erkannt werden.

4.1.1 Auswertung Datensätze

Im Folgenden wird die Auswertung der Datensätze inklusive Nutzbarkeit der IQ-Kriterien näher eingeordnet.

Durch überwiegend negative Zuweisung von Kennzeichen ist die Auswertungsmatrix spärlich belegt, was sich für die Nutzung der Genauigkeit als IQ-Kriterium durch Werte oberhalb von 99,9 % als schwierig erweist. Deshalb liefert die Genauigkeit bei der Zeichenkettenerkennung keine wertvollen Erkenntnisse und wird dieses IQ-Kriterium in der Auswertung nicht berücksichtigt.

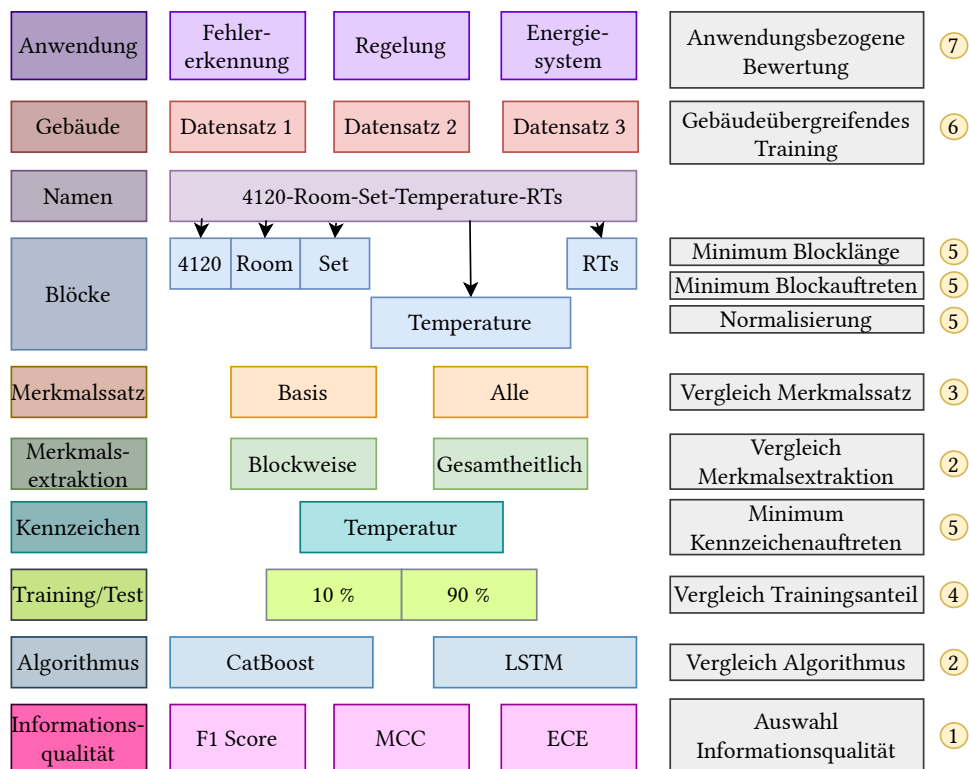


Abbildung 4.1: Abfolge von Untersuchungen bei der Merkmalsextraktion aus Zeichenketten von der Auswahl der zu untersuchenden Informationsqualitäten (1) bis hin zur anwendungsbezogenen Bewertung (7)

Die Konformität und Effektivität können nur bei zwei unterschiedlichen Algorithmen bzw. übergreifend über unterschiedliche Auswertungen sinnvoll verwertet werden. Die Generalisierbarkeit wird grundsätzlich in allen Unterpunkten auch für einzelne Datensätze untersucht („unterschiedliche Datensätze“), aber nicht gesondert ausgewiesen. Das volle Potential der Generalisierbarkeit wird aber erst bei der Gesamtuntersuchung in Kapitel 4.1.1.5 durch die Einbeziehung mehrerer Trainingsdatensätze gesondert betrachtet. Die Operationalisierbarkeit wird in Kapitel 4.1.2 detaillierter untersucht.

Auf Basis dieser Einschränkungen werden für die allgemeine Auswertung der Datensätze nur die folgenden Informationsqualitätskriterien herangezogen:

- Vollständigkeit (F1 Wert)
- Ausgeglichenheit (MCC)
- Zuverlässigkeit (ECE)

Die anderen IQ-Kriterien werden entweder gesondert in diesem Kapitel oder am Ende der Auswertung durch eine Gesamtbewertung in Kapitel 4.4 näher analysiert. In den vorherigen Parameterstudien werden diese jedoch nicht näher betrachtet, um diese Auswertung leichter verständlich zu halten. Die anderen IQ-Kriterien sind für eine Gesamtbewertung des Verfahrens sinnvoller einzusetzen, da sie ein Gesamtvermögen der Algorithmen zeigen. Dieser Abschnitt untersucht, ob durch die Verwendung von bestimmten Verfahren eine Verbesserung der IQ erzielt werden kann. Im digitalen Dokument existieren Verlinkungen auf die jeweiligen Verfahren.

4.1.1.1 Merkmalsextraktionsverfahren

Der F1 Wert und MCC der untersuchten Merkmalsextraktionsverfahren (ME-Verfahren) unterscheidet sich maßgeblich wie in Abbildung 4.2 links zu sehen ist. Während F1 Wert und MCC bei den untersuchten Verfahren bei Training und Testen auf dem gleichen Datensatz bei der Methode „Blockweise“ die besten Ergebnisse mit bis zu 98,4 % F1 Wert und MCC liefern, zeigt das Training und Testen auf unterschiedlichen Datensätzen, dass tendenziell das Verfahren „Gesamtheitlich“ die besseren Ergebnisse zeigt.

„Gesamtheitlich LSTM“ als Vertreter eines konventionellen Algorithmus zeigt um 10-20 % schlechtere Ergebnisse als die beiden anderen Methoden. Für den ECE ist unabhängig vom Training und Testen in allen Fällen die Methode „Blockweise“ mit Werten von 0,0 bis 0,2 % die beste Option (Vergleich „Gesamtheitlich“: 0,05 bis 0,25 %). „Gesamtheitlich LSTM“ ist abermals sehr weit entfernt von den Ergebnissen der anderen beiden Methoden, weshalb dieser nicht weiter betrachtet wird.

Insgesamt wird jedoch gezeigt, dass die entwickelten Verfahren „Blockweise“ und „Gesamtheitlich“ hohe Ergebnisse zeigen können und nur noch wenige Metadaten manuell geändert werden müssten. Bei „Blockweise“ muss jedem einzelnen Block ein Kennzeichen wie Temperatur

hinzugefügt werden, was tendenziell personalintensiver ist, als einem gesamten Label die benötigten Metadaten zuzuweisen, wie es bei „Gesamtheitlich“ geschieht. Auch können redundante Zuweisungen vermieden werden, wie sie z.B. durch die Definition einer Temperatur im Label mit AKS („4120-Room-Set-Temperature-RTs“: „Temperature“, „RTs“) und Klartext in der Beschreibung („Room temperature setpoint“: „temperature“) auftreten könnten. In dem Beispiel müsste das Kennzeichen „Temperatur“ bei „Blockweise“ drei mal zugewiesen werden, während bei „Gesamtheitlich“ nur eine Zuweisung erforderlich ist. Die Unterteilung in welchem Block welches Kennzeichen des Labels definiert ist, erfordert ein deutlich genaueres und für den Anwender anstrengenderes Vorgehen als die zentrale Zuweisung zum Label.

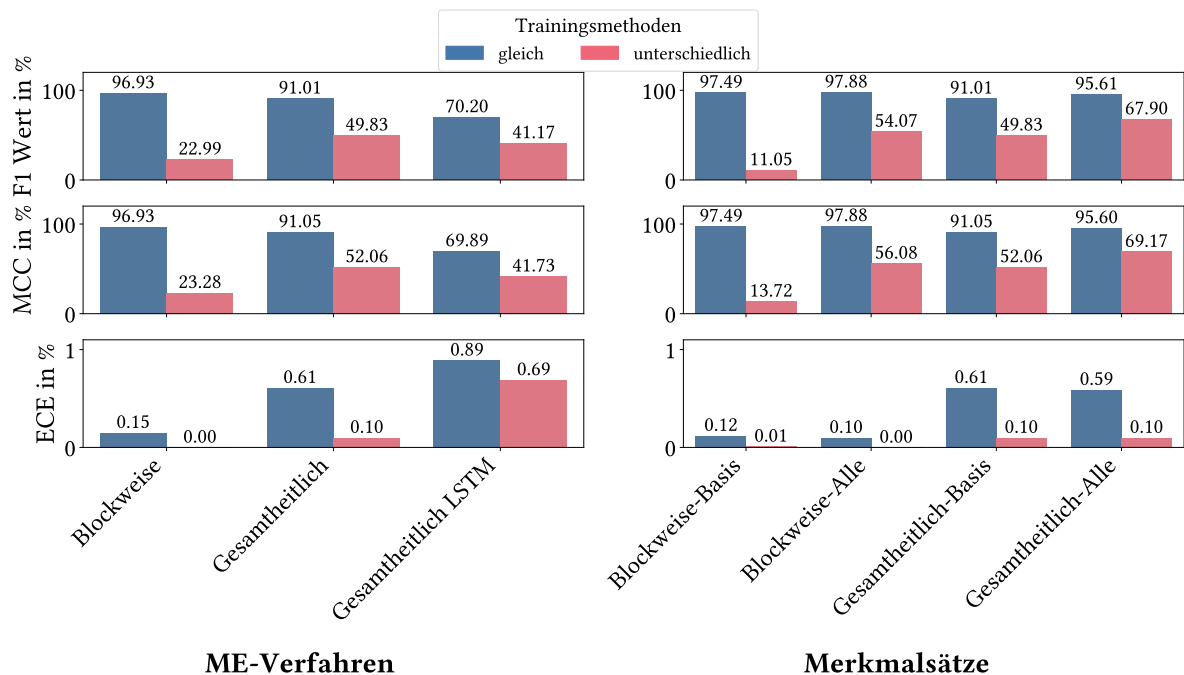


Abbildung 4.2: Vergleich des allgemeinen F1 Wertes, MCC und ECE der untersuchten Merkmalsextraktionsverfahren (ME-Verfahren, links) und Merkmalsätze (rechts) bei Training und Testen auf gleichen (blau) und unterschiedlichen Datensätzen (rot).

4.1.1.2 Merkmalssatz

Im Folgenden werden die Merkmalssätze der Merkmalsextraktionsverfahren „Blockweise“ und „Gesamtheitlich“ genauer untersucht. Es wird untersucht, ob eine höhere Anzahl an Merkmale zu besseren Ergebnissen führt und welche Merkmale in der Realität einen Einfluss besitzen.

4.1.1.2.1 Merkmalssatz bei blockweiser Merkmalsextraktion Die Merkmalssätze unterscheiden sich in der Anzahl der Merkmale. Während „Basis“ die geringste Anzahl aufweist, zeigt „Alle“ jedes Merkmal, das in dieser Arbeit untersucht wird.

Die Ergebnisse in Abbildung 4.2 rechts zeigen kein klares Bild, ob mehr Merkmale eine höhere IQ erzeugen. Der F1 Wert zeigt bei Training und Testen auf unterschiedlichen Datensätzen beim Datensatz „Alle“ mit 54,1 % mit einem durchschnittlichen Unterschied zu „Basis“ von 43,0 % die besten Werte. Beim Training und Testen auf gleichen Datensätzen zeigt sich dieses Verhalten auch, aber der Unterschied ist mit 0,9 % deutlich kleiner. Beim MCC zeigt sich das grundlegend gleiche Verhalten wie beim F1 Wert.

Der ECE profitiert bei gleichen Datensätzen nicht unbedingt von mehr Merkmalen, sondern erreicht seinen besten Wert beim „Basis“-Merkmalssatz. Alle Werte sind jedoch auch bei allen Merkmalssätzen in einem Bereich, wo ein Unterschied in der Realität keine Auswirkungen hat (0,004 bis 0,006 %). Bei unterschiedlichen Datensätzen zeigt sich jedoch, dass mehr Merkmale einen niedrigeren ECE liefern, da der Merkmalssatz „Alle“ die deutlich besten Werte zeigt mit 0,09 % (im Vergleich „Basis“: 0,12 %).

Da die Ergebnisse einen signifikanten Unterschied hin zum Besseren beim Merkmalssatz „Alle“ im Vergleich zum „Basis“-Merkmalssatz aufzeigen, ist es sinnvoll in der Praxis diesen Merkmalssatz zu verwenden. Die Ergebnisse zeigen, dass die Verwendung GAS-spezifischer Merkmale sinnvoll ist.

Falls es nur auf eine schnelle und effiziente Berechnung ankommt und nur auf dem gleichen Datensatz trainiert und getestet wird, kann auch der „Basis“-Merkmalssatz verwendet werden. Da aber die Merkmalsextraktion bei Zeichenketten in der Regel nur einmal durchgeführt wird, ist es fraglich, ob in der Praxis die Berechnungszeit relevant ist.

In Abbildung 4.3 sind die Merkmalswichtigkeiten (engl. „feature importance“) bei Nutzung des Merkmalssatz „Alle“ aufgeführt. Die Summe der Merkmalswichtigkeiten ergibt jeweils 100 %. Es ist sichtbar, dass die Umgebung („Nächster Block“ und „Vorheriger Block“) eines Blockes mit im Median 13,3 bzw. 13,5 % Merkmalswichtigkeit einen hohen Einfluss besitzt. In der Spitze können aber „Ähnliche Blöcke“ auf Basis der Ähnlichkeiten von in der Gebäudeautomation üblichen Wörtern und die Nutzung der Buchstabenkombinationen (inklusive Berücksichtigung der Konsonanten) mit 17,8 % bzw. 18,9 % einen höheren Einfluss erzielen. Dies ist abhängig vom verwendeten AKS. Wenn dieses viele Klartext-Beschreibungen nutzt, dann können ähnliche Blöcke auch einen höheren Einfluss erreichen. Auch die Phonetik ist im Median mit 13,1 % nahe an der Wichtigkeit der höchsten „alten“ Merkmale platziert. Ob bestimmte in der Gebäudeautomation üblichen Wörter tatsächlich im Block enthalten sind, spielt eine untergeordnete Rolle.

Es ist auch möglich mithilfe eines SHAP Values [305] die Wichtigkeit eines Merkmals auf das Modell selbst und nicht wie hier des Modellergebnisses zu bestimmen. Diese Überprüfung ist jedoch mit der in dieser Arbeit genutzten Computerumgebung aufgrund von Kapazitätsgründen trotz 128 GB Arbeitsspeicher nicht möglich.

4.1.1.2.2 Merkmalssatz bei gesamtheitlicher Merkmalsextraktion Bei „Gesamtheitlich“ werden auch wiederum zwei Merkmalssätze überprüft: „Basis“ und „Alle“, deren Ergebnisse in Abbildung 4.2 rechts aufgezeigt werden.

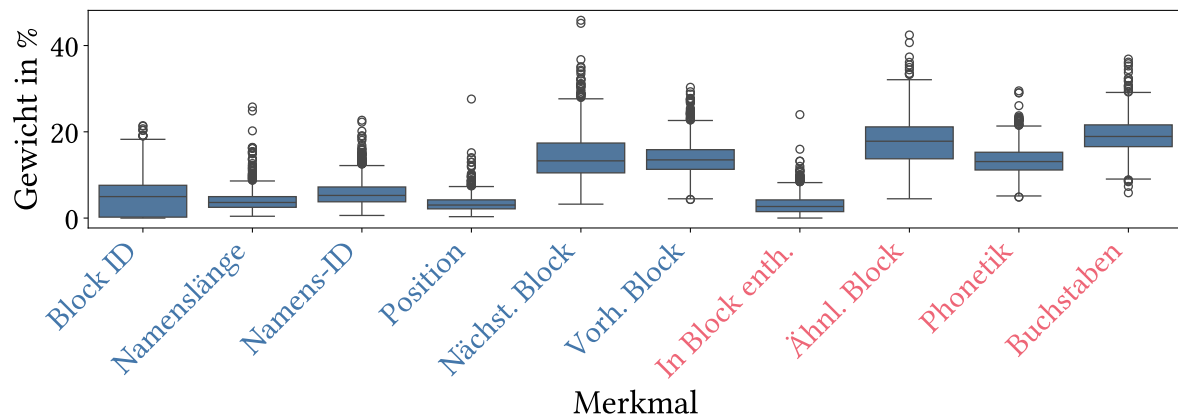


Abbildung 4.3: Vergleich der Merkmalswichtigkeit beim Merkmalssatz „Alle“ beim Merkmalsextraktionsverfahren „Blockweise“. Die in rot markierten Merkmale sind Merkmale, die nur in diesem Merkmalssatz enthalten sind.

In den Ergebnissen zum F1 Wert und MCC ist sichtbar, dass in allen Fällen der Merkmalssatz „Alle“ die besseren Ergebnisse liefert. Der F1 Wert steigt im Median von 91 auf 95,6 % bei gleichem Datensatz und von 49,8 auf 67,9 % bei unterschiedlichem Datensatz. Wie auch bei der Merkmalsextraktion „Blockweise“ wurden somit signifikante Steigerungen der Ergebnisse der Generalisierbarkeit durch die Einführung von sinnvollen GAS-spezifischen Merkmalen erreicht. Der ECE ist konträr zu diesen Ergebnissen nicht signifikant unterschiedlich bei unterschiedlichen Merkmalssätzen.

Abbildung 4.4 zeigt die Merkmalswichtigkeiten des Merkmalssatzes „Alle“. Es ist sichtbar, dass die Bigrame der Blöcke mit einem Median von 41,1 % eine wichtige Rolle bei der Unterscheidung der Klassen spielen. Die Block ID, die Block ID rückwärts, die Ähnlichkeit zu den meist genutzten Blöcken und die Ähnlichkeit zu typischen GAS-Begriffen besitzen mit 11,2 bis 13,5 % eine ähnliche Wichtigkeit. Wie bei „Blockweise“ ist das Merkmal, ob GAS-Begriffe im Namen enthalten sind, von untergeordneter Wichtigkeit, im Gegensatz zu seiner Ähnlichkeit. Aus Ressourceneffizienz-Gründen könnte dieses Merkmal ausgelassen werden. Aber in einzelnen Durchläufen erreichte dieses Merkmal auch Werte von bis zu 45 % und die Berechnung des Merkmals ist sehr schnell, weshalb eine Auslassung nicht als sinnvoll erachtet wird.

4.1.1.2.3 Bewertung Merkmalssätze Die Auswertung der unterschiedlichen Merkmalssätze der verschiedenen Merkmalsextraktionsverfahren zeigt, dass von den GAS-spezifischen Merkmalen vor allem die Generalisierbarkeit profitiert (also Training auf Datensatz A und Testen auf Datensatz B). Somit konnte die These der Verbesserung durch erweiterte Merkmalsextraktion bestätigt werden. Bei gleichem Datensatz profitierte vor allem die Merkmalsextraktion „Gesamtheitlich“ von den spezifischen Merkmalen. Die Ähnlichkeit zu GAS-typischen Blöcken war in beiden Verfahren erfolgreich, wurde aber im Summe durch die Ergebnisse der umgebenden Blöcken übertroffen.

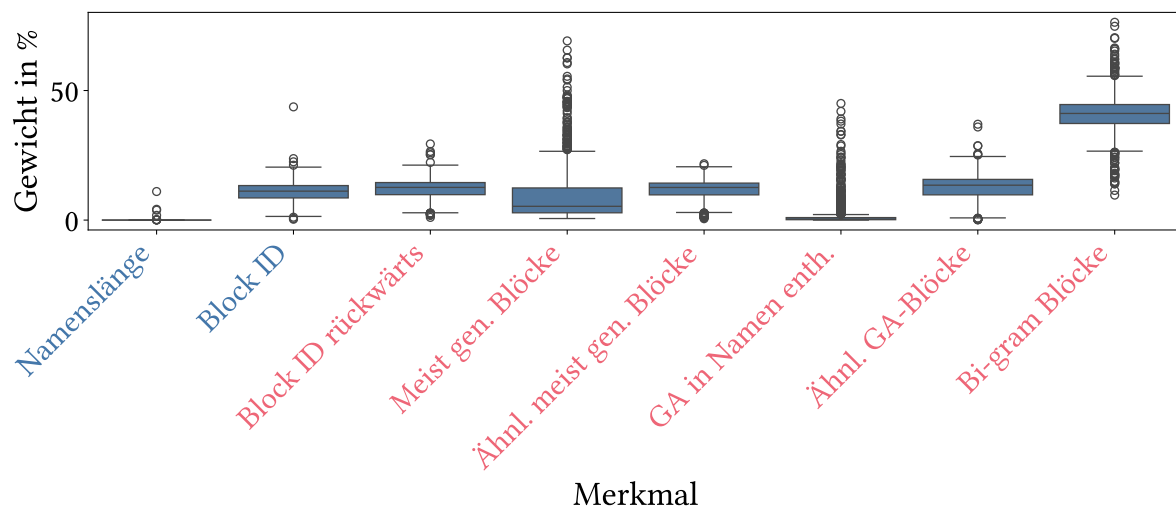


Abbildung 4.4: Vergleich der Merkmalswichtigkeit beim Merkmalssatz „Alle“ bei „Gesamtheitlich“. Die in rot markierten Merkmale sind Merkmale, die nur in diesem Merkmalssatz enthalten sind.

4.1.1.3 Unterscheidung Trainingsanteil

Bei einer Steigerung des Trainingsanteils ist zu erwarten, dass die Informationsqualität tendenziell steigt. In diesem Abschnitt wird überprüft, ob diese Annahme auch in Realität zu beobachten ist.

4.1.1.3.1 Trainingsanteil Blockweise Im Folgenden werden in Abbildung 4.5 die Ergebnisse bei Unterscheidung des Trainingsanteils eines Datensatzes bei der Zuordnung von Merkmalsextraktionsverfahren „Blockweise“ aufgeführt.

Der F1 Wert steigt bei Training und Testen auf dem gleichen Datensatz mit variablen Anteilen des Datensatzes mit höherem Trainingsanteil an. Der Bereich, in dem die Steigerung des F1 Wertes signifikant sinkt, liegt zwischen 4 und 6 % Trainingsanteil. Ab 12 % Trainingsanteil werden nur noch sehr geringere Steigerungen erzielt. Ein gegenteiliges Bild zeigt das Training und Testen auf unterschiedlichen Datensätzen. Der F1 Wert stagniert von Beginn an und schwankt um einen Wert von ungefähr 25 %. Das bedeutet, dass in diesem Fall beim Transferlernen der grundlegende Algorithmus nicht von einer Erhöhung des Trainingsanteils profitiert. Die Ergebnisse streuen je nach Paarung der Datensätze sehr stark von annähernd 0 % bis über 60 % F1 Wert. Das zeigt an, dass die Ergebnisse stark vom Datensatz abhängig sind.

Der MCC als Informationsqualitätskriterium zeigt ein annähernd gleiches Verhalten wie der F1 Wert. Er liegt nur minimal unterhalb der Werte des F1 Wertes, das für eine gute Ausgeglichenheit der Klassifizierungsergebnisse spricht.

Der allgemeine ECE sinkt tendenziell bei Nutzung von einem Anteil eines Datensatzes als Training und dem anderen Anteil des gleichen Datensatzes als Test mit Erhöhung des Trainingsanteils. Es ist

erkennbar, dass ab ca. 10 % Trainingsanteil der ECE nur noch minimal gesenkt wird und tendenziell nicht mehr sinkt. Die größten Veränderungen werden zwischen 0,5 % und 3 % Trainingsanteil erreicht, was auch zu erwarten ist. Bei Nutzung von einem Anteil eines Datensatzes für das Training und eines anderen Datensatzes für das Testen zeigt sich ein tendenziell anderes Bild: der allgemeine ECE stagniert bzw. sinkt nicht signifikant im Trend bei Erhöhung des Trainingsanteils und selbst zwischen 0,5 % und 3 % Trainingsanteils sinkt der ECE nicht signifikant. Auffällig ist auch, dass sich der ECE bei unterschiedlichen Datensätzen um einen Faktor 30 gegenüber dem ECE bei gleichem Datensatz unterscheidet ($\sim 0,15$ % gegenüber 0,005 %)

In der Praxis ist der Trainingsanteil sehr wichtig für die Akzeptanz eines Verfahrens. Es ist davon auszugehen, dass weniger Trainingsdaten, jedoch mehr benötigte konkrete Auswahl bei der Anwendung, von einem Anwender stärker akzeptiert wird. Die Kosten für die Umbenennung von existierenden Gebäuden ist annähernd linear abhängig von den benötigten Eingaben von Nutzern, wobei das für die jeweiligen Eingaben benötigte Vorwissen mit dem entsprechenden Gehalt einfließt. Je mehr Gebäude das gleiche AKS nutzen, desto geringer werden die benötigten Eingaben für neue Gebäude. Deshalb bestimmt der Anteil der Trainingsdaten annähernd linear die Kosten für die Umbenennung.

Für die Praxis sinnvoll ist bei wenigen Gebäuden von einem Trainingsanteil von 3 % bis 5 % auszugehen, da danach der Nutzen nur noch gering steigt. Der sinnvolle Anteil ist jedoch stark abhängig von den verwendeten Datensätzen und Algorithmen. Falls es sich um ein AKS handelt, dass in sehr vielen Gebäuden im Einsatz ist, dann lohnt es sich je nach Intensität der Nutzung den Trainingsanteil deutlich zu erhöhen, um bessere Ergebnisse zu erzielen. In der Literatur wurde ein minimaler Trainingsanteil von 10 % empfohlen [468], der aber bei dem hier vorliegenden Ansatz nicht unbedingt benötigt wird.

4.1.1.3.2 Trainingsanteil Gesamtheitlich Die Ergebnisse beim Merkmalsextraktionsverfahren „Gesamtheitlich“ sind im folgenden Unterkapitel näher beschrieben. Es wird hauptsächlich auf den Verlauf des Medians eingegangen, der in Abbildung 4.5 abgebildet ist.

Der F1 Wert steigt bei Training und Testen auf dem gleichen Datensatz ähnlich wie bei „Blockweise“ mit steigendem Trainingsanteil an. In diesem Fall sind jedoch die Steigerungen deutlicher sichtbar als bei „Blockweise“, da hier bei 0,5 % Trainingsanteil nur ein F1 Wert von 58,9 % erzielt wird, während „Blockweise“ schon 79,6 % erreicht. Auch wird bei 20 % Trainingsanteil ein geringerer F1 Wert von 94,5 % erreicht (im Vergleich „Blockweise“: 98,1 %). Die Abweichung der Ergebnisse sinkt mit steigendem Trainingsanteil, d.h. es werden mit höherer Wahrscheinlichkeit hohe Ergebnisse erzielt.

Das Training und Testen auf unterschiedlichen Datensätzen zeigt im Verlauf ein ähnliches stagnierendes Ergebnis wie bei „Blockweise“, aber auf einem fast doppelt so hohem Niveau (45-50 % F1 Wert zu 25 % bei „Blockweise“).

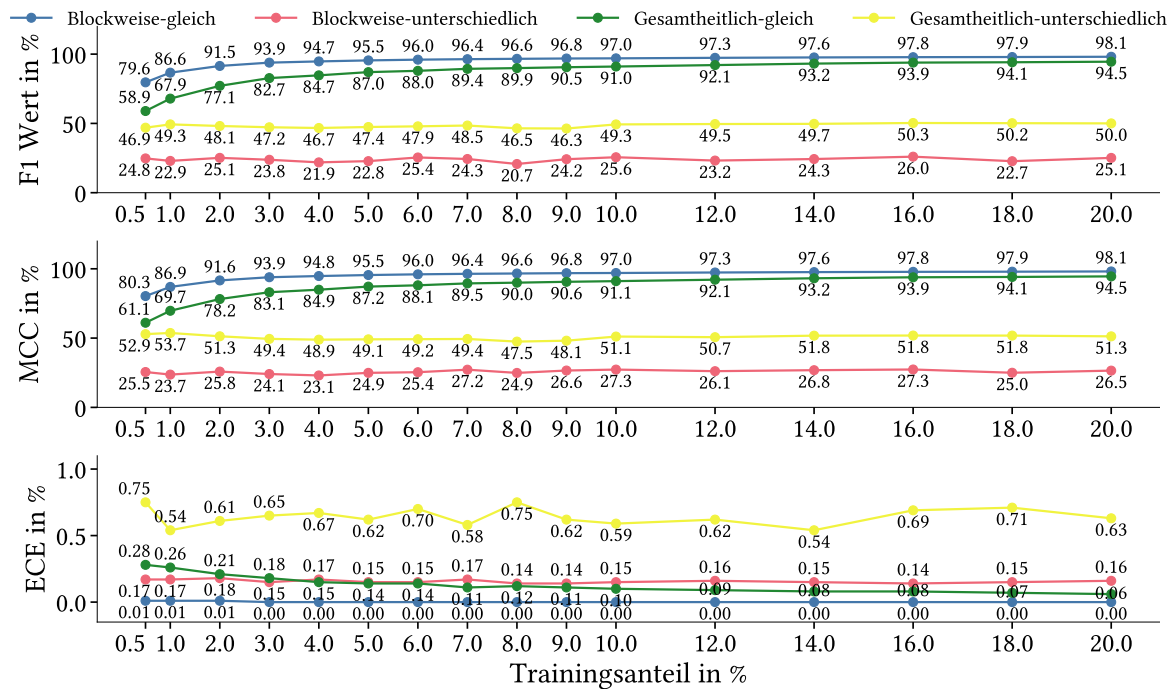


Abbildung 4.5: Vergleich des allgemeinen F1 Wertes, MCC und ECE von Merkmalsextraktionsverfahren „Blockweise“ und „Gesamtheitlich“ bei Variation des Trainingsanteils und Training und Testen auf gleichen und unterschiedlichen Datensätzen.

Der MCC zeigt qualitativ und annähernd quantitativ die gleichen Ergebnisse wie der F1 Wert. Der MCC ist tendenziell leicht größer wie beim F1 Wert, was wiederum für eine gute Ausgeglichenheit der Ergebnisse spricht.

Auch beim ECE werden die vorherigen Ergebnisse weitestgehend bestätigt. Er sinkt bei Training und Testen auf dem gleichen Datensatz mit steigendem Trainingsanteil von 0,2 % auf 0,08 %. Auch hier sinkt die Varianz mit steigendem Trainingsanteil. Bei Training und Testen auf unterschiedlichen Datensätzen wird ein vom Trainingsanteil annähernd konstantes Niveau von ca. 0,6 % erreicht.

Bei diesem Verfahren ist ein höherer minimaler Trainingsanteil als bei „Blockweise“ erforderlich. Hier wäre ein Trainingsanteil von 6 bis 10 % zu empfehlen, da in diesem Bereich die Steigerung der Ergebnisse nicht mehr stark ist.

4.1.1.4 Eingrenzung Nutzung Merkmale

In diesem Unterkapitel werden die einzelnen Merkmale näher untersucht, die eine Eingrenzung des Informationsspektrums ermöglichen. Hierzu wurde das Kennzeichenaufreten, die Blocklänge und das Blockauftreten berücksichtigt und im Folgenden dargestellt. Zusätzlich wird auch eine Normalisierung auf Basis von Komposita, Synonymen und der Korrektur falsch geschriebener Wörter durchgeführt. Im Anhang wird in Kapitel C.5 die Auswertungen zur Anwendung der Trennung

von Komposita näher beschrieben. Auf eine genauere Beschreibung der Ergebnisse wird an dieser Stelle verzichtet. Tendenziell sollte sich laut Literatur zeigen, dass eine Senkung der Eingangsgröße bzw. Klassen die Ergebnisse verbessert.

4.1.1.4.1 Minimum Kennzeichenauftreten Die Einschränkung auf ein minimales Auftreten der Kennzeichen lieferte wie in Abbildung 4.6 ausgeführt keine signifikanten Änderungen beim F1 Wert und MCC bei Anwendung auf unterschiedlichen Datensätzen und nur leichte Steigerungen von F1 und MCC von 96,9 % auf 97,4 % bei Anwendung auf dem gleichen Datensatz.

Der ECE dagegen steigt bei Einschränkung der Kennzeichen bei Anwendung auf gleichem und unterschiedlichem Datensatz. Es wäre aber eine Senkung zu erwarten. Bei der Einschränkung der Kennzeichen zeigt sich somit, dass die Senkung beim ECE keinen Vorteil bietet.

Bei einem minimalen Kennzeichenauftreten von 100 werden nur bis zu 4,1 % der Kennzeichen entfernt. Dafür wird die Kennzeichenvariation um bis zu 78,4 % reduziert. Das bedeutet, dass die Informationsbreite gesenkt wird, aber die Informationen insgesamt im Hauptteil erhalten bleiben. Es ist zu sagen, dass die Einschränkung auf ein Minimum Kennzeichenauftreten nur mit einem voll gelabelten Datensatz möglich ist und somit die Anwendung auf dem gleichen Datensatz fraglich ist. Da bei Anwendung auf unterschiedlichen Datensätzen bei F1 Wert und MCC nur geringe Änderungen zustande kamen und beim ECE sogar schlechtere Ergebnisse, ist von einer praktischen Anwendung abzusehen.

4.1.1.4.2 Minimum Blocklänge Die Variation der minimalen Blocklänge soll zeigen, ob mehr natürliche Sprache und weniger Abkürzungen zu einer Verbesserung der Ergebnisse führt, die in Abbildung 4.7 ausgeführt sind. Der ECE konnte sich nur minimal verbessern. Hierbei war es unerheblich, ob auf gleichem Datensatz oder unterschiedlichem Datensatz trainiert wurde. Der F1 Wert und MCC zeigten keinerlei Verbesserung unabhängig von Trainings- und Testart. Es zeigt sich somit, dass die Einschränkung nur auf tendenziell natürliche Sprache keine signifikante Verbesserung der Ergebnisse hervorruft.

4.1.1.4.3 Minimum Blockauftreten Während bei der Modifikation der minimalen Blocklänge die Qualität des Eingangs eine zentrale Rolle spielt, so ist bei der Modifikation des Blockauftretens der Variantenreichtum des Eingangs der KI das Ziel. Zu erwarten wäre, dass eine Begrenzung des Variantenreichtums des Eingangs eine Verbesserung der Ergebnisse in den IQ-Kriterien liefert. Die Ergebnisse der Versuchsreihe sind in Abbildung 4.8 zu sehen.

Im F1 Wert und MCC zeigt sich bei Erhöhung des minimalen Blockauftretens eine Senkung der Werte. Das heißt es zeigt sich das genau gegenteilig erwartbare Verhalten. Beim ECE zeigen sich keine signifikanten Veränderungen. Die Einschränkung anhand des minimalen Blockauftretens ist folglich nicht sinnvoll.

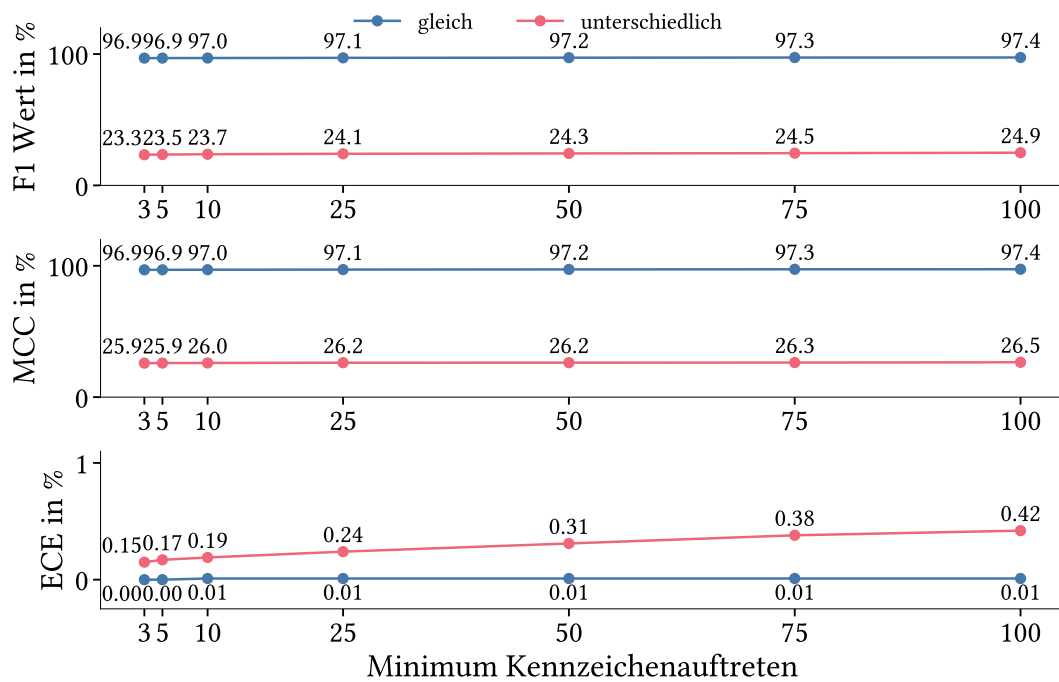


Abbildung 4.6: Vergleich des Medians des allgemeinen F1 Wertes, MCC und ECE von Merkmals-extraktionsverfahren „Blockweise“ bei **Variation des minimalen Kennzeichenauf-tretens** und Training und Testen auf gleichen und unterschiedlichen Datensätzen.

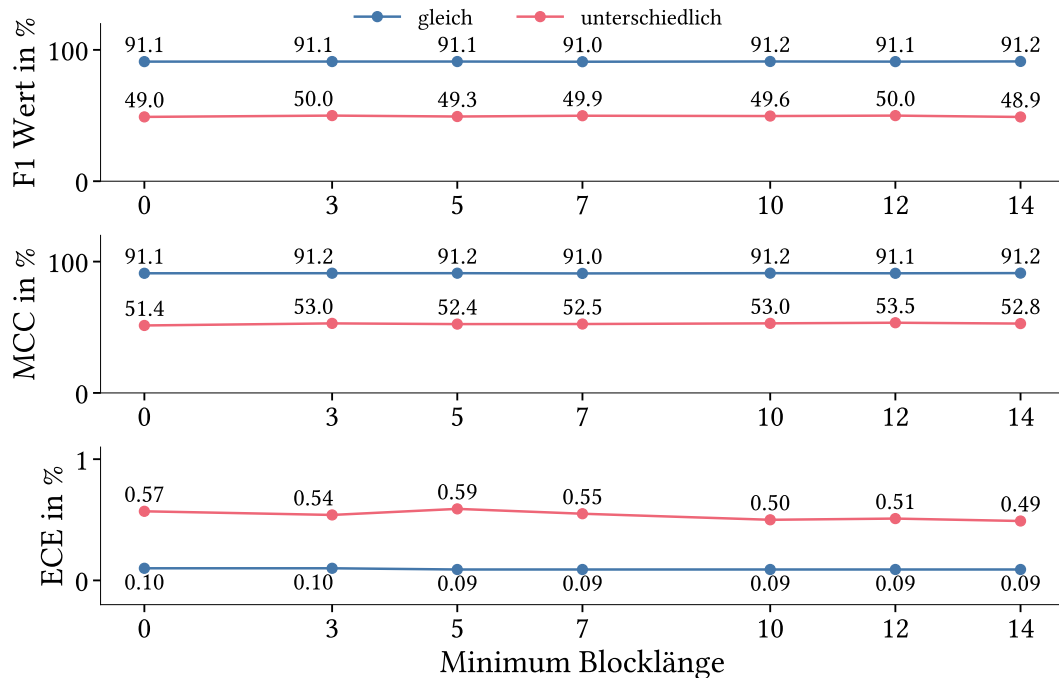


Abbildung 4.7: Vergleich des Medians des allgemeinen F1 Wertes, MCC und ECE von Merkmals-extraktionsverfahren „Blockweise“ bei **Variation der minimalen Blocklänge** und Training und Testen auf gleichen und unterschiedlichen Datensätzen. Bei 0 werden alle Blöcke verwendet.

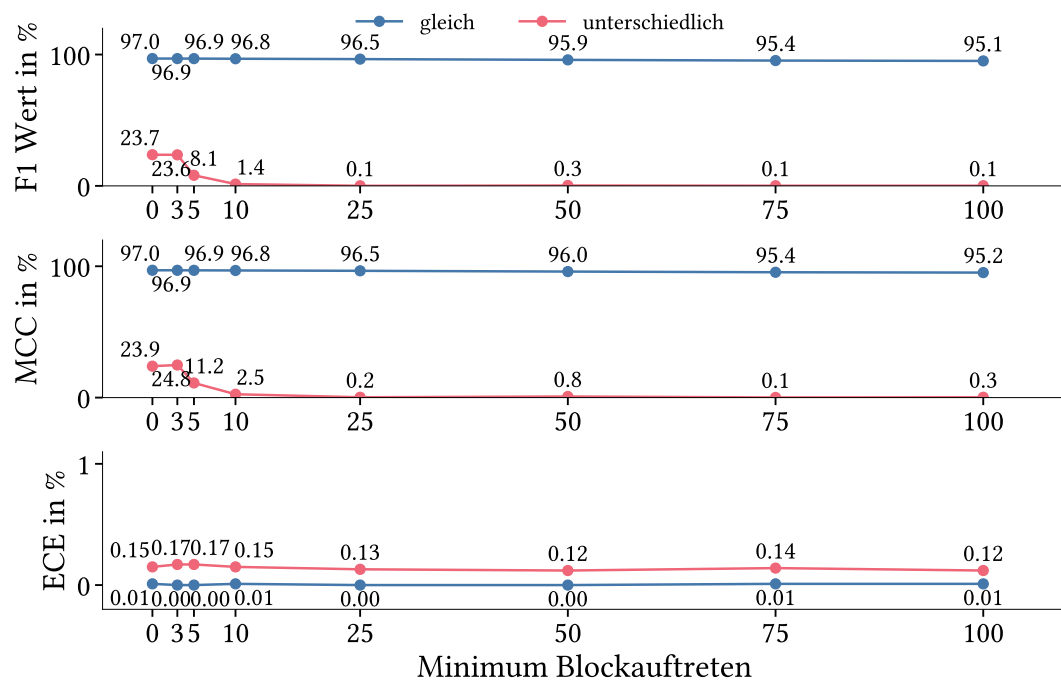


Abbildung 4.8: Vergleich des Median des allgemeinen F1 Wertes, MCC und ECE von Merkmals-extraktionsverfahren „Blockweise“ bei **Variation des minimalen Blockauftretens** und Training und Testen auf gleichen und unterschiedlichen Datensätzen. Minimum Blocklänge=0 bedeutet, dass alle Blöcke verwendet werden.

4.1.1.4.4 Normalisierung auf Basis von Komposita, Synonymen und Korrektur falsch geschriebener Wörter Im Folgenden werden die Ergebnisse der Normalisierung auf Basis der Trennung von Komposita, des Austauschen von Synonymen und der automatisierten Korrektur falsch geschriebener Wörter dargestellt. In Abbildung 4.9 sind die Ergebnisse aufgeteilt in Training und Testen auf dem gleichen und unterschiedlichen Datensätzen dargestellt. Es zeigt sich, dass nur beim ECE ein leichter Vorteil der Normalisierung besteht. Im MCC und F1 Wert sind die Ergebnisse für die Normalisierung bei unterschiedlichen Datensätzen besser, aber bei Nutzung auf gleichen Datensätzen wiederum schlechter. Das bedeutet, dass nur ein sehr minimaler Vorteil besteht, der aufgrund des hohen Personalaufwands in keiner Relation steht und deshalb nicht durchgeführt werden sollte. Der Personalaufwand ist für jeden Normalisierungsschritt erforderlich, also der Trennung von Komposita, Nutzung von Synonymen und Korrektur falsch geschriebener Wörter, wobei Synonyme hauptsächlich initial hohe Kosten verursachen.

Eine mögliche Erklärung für die nicht funktionierende Normalisierung von Wörtern ist die entgegen der Theorie eventuell erhöhte Entropie, die teilweise durch die Normalisierung von Wörtern entsteht. So können bei der Trennung von Komposita nicht nur die Wortbestandteile getrennt werden, sondern auch der Zusammenhang der Teile verloren gehen. So kann eine Wärmepumpe in die Bestandteile „Wärme“ und „Pumpe“ unterteilt werden. Eine Wärmepumpe hat zwar mit Wärme zu tun und sie enthält auch in wasserbasierten Systemen eine Pumpe, aber die Zuweisung der Kennzeichen „Wärme“ und „Pumpe“ wäre in diesem Kontext falsch und wird entsprechend im IQ-Kriterium bewertet. Falsch geschriebene Wörter tauchen zwar bei einzigartigen Blöcken häufiger auf, aber die Gesamtmenge der Vorkommen ist dagegen vernachlässigbar gering. Für eine technisch sinnvolle Einbindung der Normalisierung wäre ein entsprechender Algorithmus erforderlich, der nur semantisch sinnvolle Normalisierungen realisiert und deren Entwicklung wäre wirtschaftlich höchst fragwürdig.

4.1.1.5 Gebäudeübergreifendes Training

In diesem Unterkapitel wird das gebäudeübergreifende Training und Testen bei unterschiedlichen Bezeichnungssystemen näher überprüft. Insgesamt gibt es fünf verschiedene Datensätze. Jeder dieser Datensätze wird als Testdatensatz herangezogen und die vier anderen Datensätze zusammen als Trainingsdatensatz. Somit soll untersucht werden, ob der Algorithmus in der Lage ist, bei einer Datensatzanzahl $n \rightarrow \infty$ ein stabiles Ergebnis zu erzielen. Für das Training und Testen wurden aufgrund des Limits des Arbeitsspeichers nur 20 % des jeweiligen Datensatzes randomisiert verwendet.

4.1.1.5.1 Blockweises gebäudeübergreifendes Training Die Ergebnisse des Verfahren „Blockweise“ mit dem Merkmalssatz „Alle“ werden in Abbildung 4.10 gezeigt. Der MCC erreicht qualitativ die gleichen Ergebnisse wie der F1 Wert, weshalb auf eine Abbildung verzichtet wird.

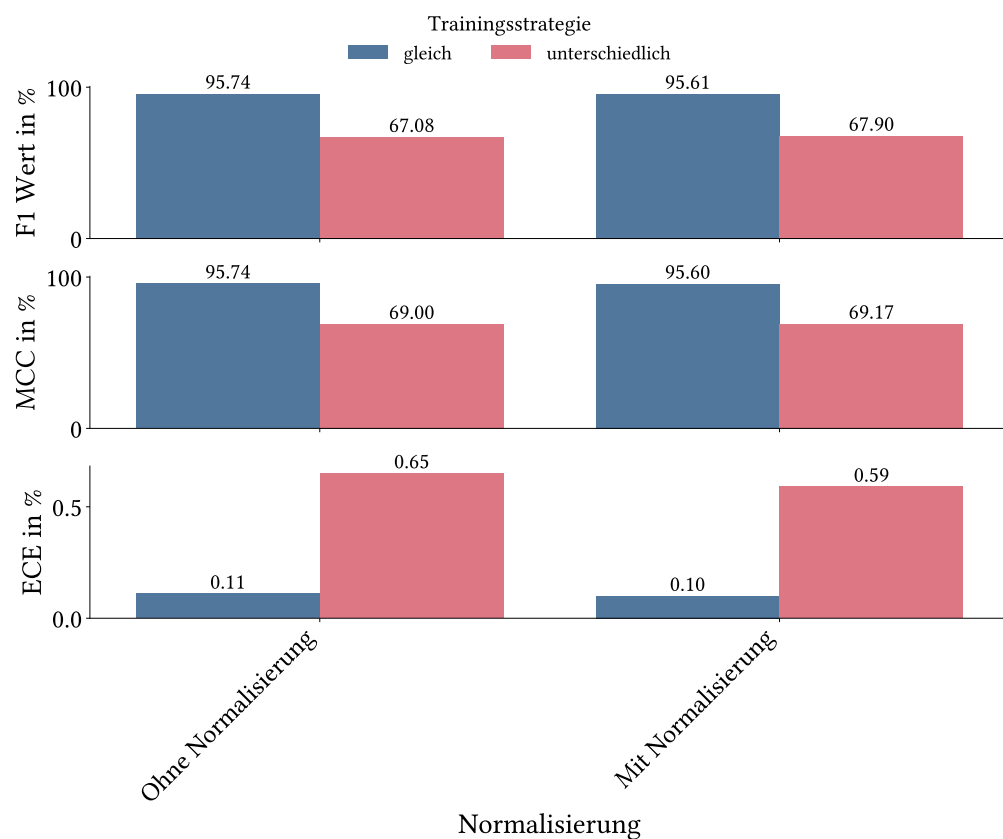


Abbildung 4.9: Vergleich des allgemeinen F1 Wertes, MCC und ECE von Merkmalsextraktionsverfahren „Gesamtheitlich“ bei **Normalisierung und keiner Normalisierung** und Training und Testen auf gleichen und unterschiedlichen Datensätzen.

Der F1 Wert weicht im Maximum nur 1 bzw. 2 % innerhalb der Testdatensätze ab und erreicht einen Spitzenwert von 83,7 %. Das bedeutet, dass ein relativ konstantes Ergebnis mit Hilfe von allen anderen Datensätzen als Trainingsdatensätzen erzielt werden konnte.

Der ECE ist mit maximal 0,017 % in einem Bereich, der die realen Ergebnisse nicht beeinträchtigt. Das bedeutet, dass die Vorhersage des Klassifizierers als korrekt angesehen werden kann.

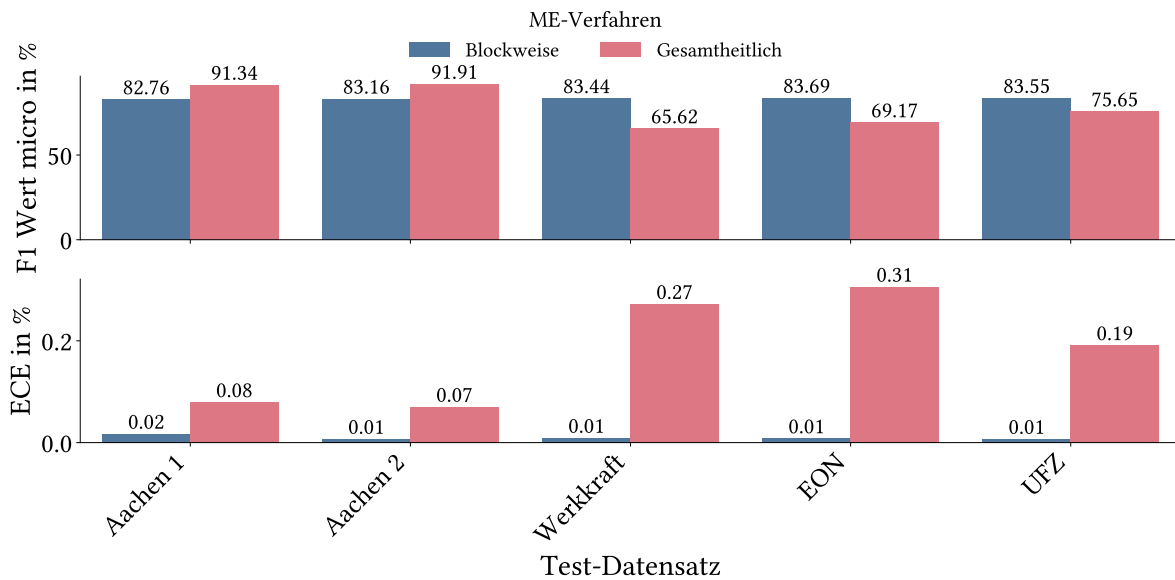


Abbildung 4.10: Ergebnisse des allgemeinen F1 Wertes und ECE bei Merkmalsextraktionsverfahren „Blockweise“ und „Gesamtheitlich“ bei Training und Testen auf gleichen und unterschiedlichen Datensätzen bei **schemaübergreifender** Betrachtung.

4.1.1.5.2 Gesamtheitliches gebäudeübergreifendes Training Im Folgenden werden die Ergebnisse beim Testen auf einem Datensatz und Training auf allen anderen Datensätzen des Merkmalsatzes „Gesamtheitlich“ dargestellt.

Die Ergebnisse beim F1 Wert und ECE (Abbildung 4.10) zeigen auf, dass die Datensätze der Stadt Aachen sehr gut erkannt werden, aber die anderen deutlich schlechter erkannt werden. Bei der Stadt Aachen wird in beiden Datensätzen in der Regel das umfangreiche AKS von der Firma Siemens verwendet, in dem die einzelnen Blöcke gut voneinander zu unterscheiden sind (Beispiel „079'H'HGrp1006'Hcrv'TOa“), was die Ergebnisse erheblich beeinflusst haben könnte.

Die Differenz zum Verfahren „Blockweise“ sind bis zu 18 % beim F1 Wert. Bei diesem Verfahren wurde auch kontinuierlich eine ähnliche Qualität in den Ergebnissen erzielt, während diese hier sehr schwankend ist und stärker vom Datensatz abhängig ist. Insgesamt existieren auch Datenstromtypen, die signifikant besser erkannt werden, wie mit 97 % die wichtigen Temperaturen gegenüber 90 % bei „Blockweise“.

4.1.2 Anwendungsbezogene Bewertung

In diesem Abschnitt wird die Bewertung des Metadatenextraktionsverfahrens anhand von den vorher definierten Anwendungen vorgenommen. Es wird das Merkmalsextraktionsverfahren „Gesamtheitlich“ mit dem Merkmalssatz „Alle“ verwendet. Die Abbildung 4.11 zeigt die Ergebnisse bei Training und Testen auf dem gleichen und unterschiedlichen Datensatz. Im Folgenden werden die Ergebnisse bei gleichem Datensatz vorgestellt. Einzig die Anwendung „RLT-Anlagen“ liegen mit einem F1-Wert von 92,5 % außerhalb der anderen Anwendungen (95,4 bis 98,5 %). Interessant ist hier die hohe Erkennung des Energieflusses, der nur die Messungen der Temperatur enthält. Dies ist insbesondere bei Betrachtung der Ergebnisse bei Test und Training auf unterschiedlichen Datensätzen, wo der Energiefluss, Wärmeübertrager und Außentemperatur mit 72,7 bis 80,3 % F1 Wert höhere Werte hat als die anderen Anwendungen mit maximal 49,1 % F1 Wert (Quartiersmonitoring). Das zeigt auch, dass trotz einer variantenreichen Benennung von Temperaturen, diese sinnvoll erkannt werden können, im Beispiel von Außentemperaturen jedoch nicht explizit die Eigenschaft „Frisch/Außen“. Die Eigenschaft „Temperatur“ kann oftmals auch über die Nutzung der physikalischen Einheit explizit gewonnen werden, die Eigenschaft „Frisch/Außen“ nur über die variantenreiche Benennung.

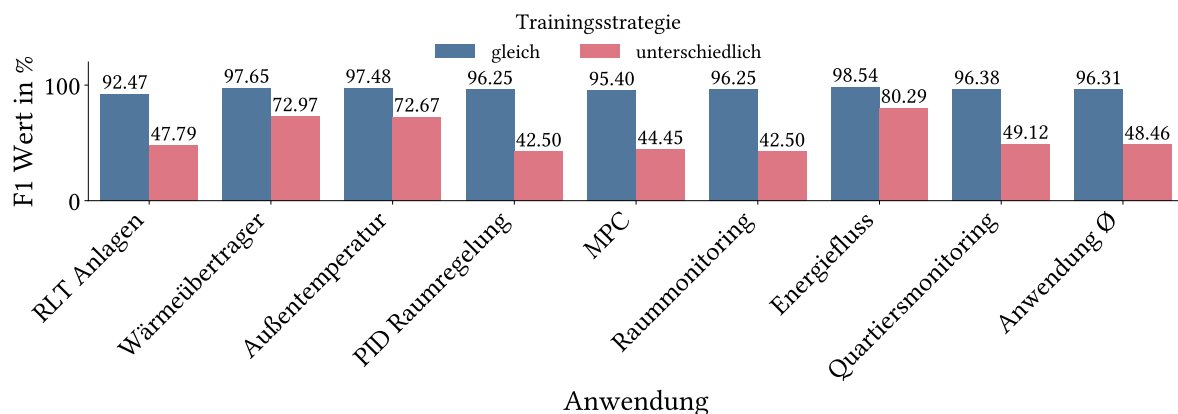


Abbildung 4.11: Ergebnisse des F1 Wertes der **Zeichenketten**-Metadatenextraktion „Gesamtheitlich“ für die **Bereitstellung anwendungstypischer Metadaten** bei Training und Testen auf gleichen und unterschiedlichen Datensätzen.

4.1.3 Zusammenfassung von zeichenkettenbasierten Verfahren

Es konnten insgesamt drei verschiedene Verfahren zur Extraktion von Metadaten aus Zeichenketten überprüft werden. Je nach Verfahren waren konstant gute Ergebnisse bis zu 98 % zu beobachten. Für die Praxis bedeutet dies, dass nur wenige extrahierte Metadaten von einer Person manuell geändert werden müssen.

Es ist insgesamt sichtbar, dass keine der Methoden zur Normalisierung der Datensätzen einem

dem Aufwand entsprechenden Nutzen hat. Hierbei scheint die Einbringung von mehr Entropie entscheidend gewesen zu sein.

Es konnte gezeigt werden, welche Merkmale sinnvoll eingebracht werden sollten. Die in dieser Arbeit neu eingebrachten Merkmale waren eine sehr sinnvolle Ergänzung zu Basis-Merkmalen, die sich eher am Stand der Technik orientieren. Hierbei wurde ein großer Datensatz von GAS-spezifischen Wörtern aufgebaut und genutzt.

Das gebäudeübergreifende Training und Testen war grundsätzlich erfolgreich. Dies ist insbesondere wichtig für den skalierbaren Einsatz der Methode. Wenn von einem Betreiber das gleiche AKS verwendet wird, dann kann das gebäudeübergreifende Verfahren den Personaleinsatz massiv senken. Es müssen dann nur noch nicht im Datensatz enthaltene Blöcke wie besondere technische Anlagen oder Synonyme übersetzt werden.

Für die Anwendungen konnte gezeigt werden, dass die Ergebnisse je nach Anwendung nochmal deutlich über den allgemeinen Ergebnissen liegen. Insbesondere die für die Untersuchung des sinnvollen Einsatzes der Energie erforderlichen Temperaturen konnten exorbitant gut erkannt werden. Somit könnten fast automatische Überprüfungen von Außentemperatur- oder Raumtemperaturen, aber auch des Energieflusses und somit der Topologie eines Gebäudes erfolgen.

4.2 Zeitreihenklassifizierung

In diesem Kapitel werden die Ergebnisse der Klassifizierung der Metadaten anhand der aufgenommenen Zeitreihendaten analysiert. Es wird überprüft welches Merkmalsextraktionsverfahren am geeignetsten ist und wie die Zeitreihenanalyse verwendet werden kann. Im Gegensatz zum zeichenkettenbasierten Ansatz wird hier ein Unilabel-Verfahren verwendet. Das bedeutet, dass z.B. für einen „Temperatur-Messwert“ beide Kennzeichen „Temperatur“ und „Messwert“ übereinstimmen müssen.

Abbildung 4.12 zeigt eine Übersicht über die bei Zeitreihen vorgenommenen Analysen, deren Ergebnisse im Folgenden vorgestellt werden.

4.2.1 Ergebnisse Mortar Datensatz

Im Folgenden werden die Ergebnisse auf den Zeitreihendaten des Mortar-Datensatzes dargestellt. Diese Ergebnisse gelten als Basis bei einem bereinigten Datensatz. Das bedeutet, dass die tendenziell maximal möglichen Ergebnisse erzielt werden.

4.2.1.1 Vergleich Kombinationen Merkmalsextrahierer und Algorithmen

In Stinner u. a. [431] wurde ermittelt, dass die Merkmalsextrahierer „MiniRocket“ und „MultiRocket“ die besten Ergebnisse für die Klassifizierung von Gebäudeautomationsdaten liefern. In diesem Ka-

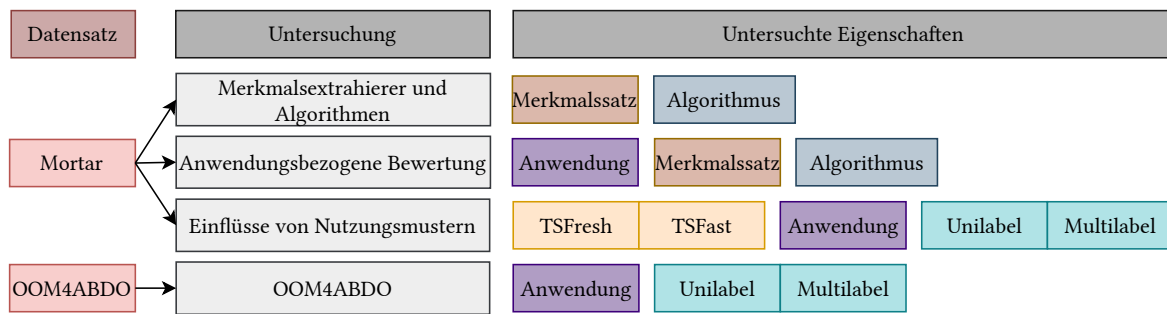


Abbildung 4.12: Eine Übersicht über die Untersuchungen und untersuchten Eigenschaften, die im Rahmen von Zeitreihenanalysen auf den zwei Datensätzen durchgeführt werden.

pitel werden die passenden Kombinationen von Merkmalsextrahierer und Klassifizierer ermittelt.

In Abbildung 4.13 wird der F1 Wert, MCC, ECE der Kombinationen bei Unilabel-Bewertung und F1 Wert bei Multilabel-Betrachtung dargestellt. Es ist klar ersichtlich, dass die Kombination „MiniRocket-Ridge“ mit 90,4 % F1 Wert die besten Ergebnisse liefert. Darauf folgt „MultiRocket-CatBoost“ mit 89,5 % F1-Wert. Interessanterweise erreicht die Kombination „MultiRocket-Ridge“ nur einen F1-Wert von 63,6 %. Der Grund dieser Diskrepanz bleibt offen, da CatBoost bei genauerer Betrachtung alle Merkmale von MultiRocket bei der Merkmalswichtigkeit relativ gleichmäßig zwischen 19 und 36 % erreicht. Das bedeutet, dass egal welches Merkmal von Ridge genutzt werden würde, dieses einen positiven Beitrag zum Ergebnis beitragen würde und diese sehr ähnlich bewertet werden würden. Der Isolation Forest hat keine Möglichkeit die Merkmalswichtigkeit auszugeben. Diese mangelnde Möglichkeit der Überprüfung der Ergebnisse zeigt allgemeine Probleme von Machine Learning Algorithmen auf Basis von KNN auf, es kann entweder gar nicht oder nur umständlich überprüft werden, warum Ergebnisse aufgrund des KNN-Algorithmus Besonderheiten aufweisen. Diese Probleme steigern sich mit der Komplexität des Algorithmus.

Der MCC zeigt qualitativ ähnliche Ergebnisse, wobei als Maximum „MiniRocket-Ridge“ nur 89 % MCC erreicht. Die benötigte Rechenzeit für MultiRocket ist das 6-fache von MiniRocket, für Ridge aufgrund des Isolation Forest etwa das 1,5-fache wie bei CatBoost. Es ist jedoch zu beachten, dass 1 % Verbesserung bei ungefähr 90 % Ergebnis eine Senkung der Deckungslücke (100 %-Ergebnis) um 10 % bewirkt.

Der ECE erreicht im Minimum 1,8 %. Dieser Wert wird nicht in der Kombination „MiniRocket-Ridge“ (2,2 %) erreicht, sondern „MiniRocket-CatBoost“. Da auch „MultiRocket-CatBoost“ mit 1,9 % einen guten Wert erzielt, kann gesagt werden, dass CatBoost als Algorithmus die besten ECE erreicht. Das Unilabel-Verfahren führt dazu, dass der ECE tendenziell höher als beim zeichenkettenbasierten Verfahren ausfällt aufgrund der Einzelbewertung der real existierenden Klassen pro Datenstrom.

Wenn die Unilabel-Bewertung auf eine Multilabel-Bewertung projiziert wird, wird jeder Bestandteil der Unilabel in ein eigenständiges Multilabel umgewandelt. „supply air temperature“ wird in

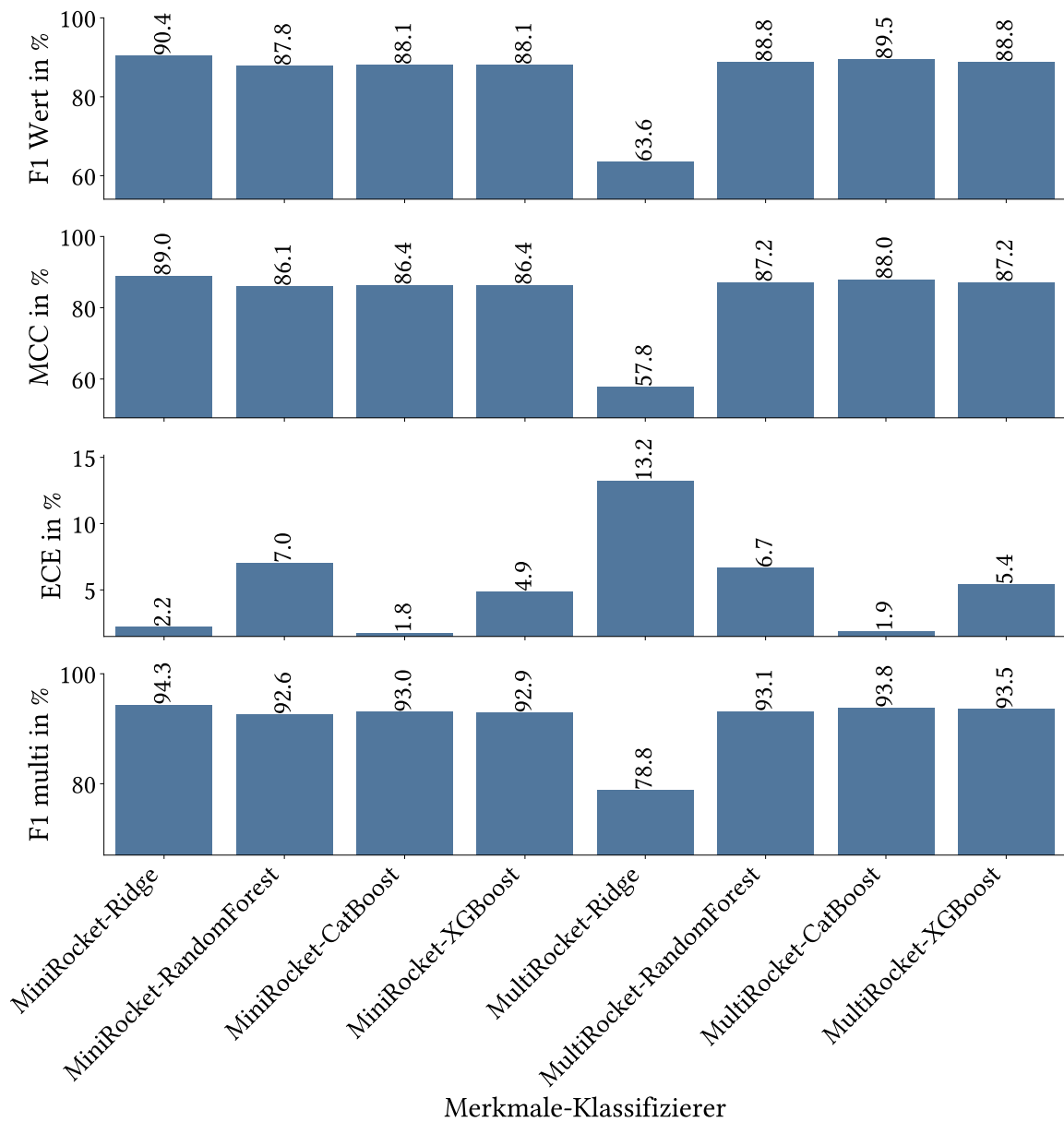


Abbildung 4.13: Vergleich des allgemeinen F1 Wertes, MCC, ECE und des F1 Wertes bei **Multi-label**-Betrachtung der untersuchten **Kombinationen** von Merkmalsextrahierern und Klassifizierern bei Metadatenextraktion durch **Zeitreihen** bei Anwendung auf dem **Mortar** Datensatz.

seine Bestandteile „supply“, „air“ und „temperature“ unterteilt. Die Ergebnisse des F1 Wertes zeigen, dass bei Multilabel-Betrachtung ein höherer F1-Wert mit 94,4 % erreicht wird. Dies ist darauf zurückzuführen, dass z.B. „Befehle“ oder einzelne Luft-Temperatur-Sollwerte nicht gut voneinander unterscheidbar sind.

4.2.1.2 Anwendungsbezogene Bewertung

Wichtig für die Bewertung der Fähigkeit der Algorithmen zur Nutzung in Metadaten-Frameworks ist die Kompetenz die für die jeweilige Anwendung erforderlichen Datenstromtypen zu erkennen.

Bei genauer Betrachtung der Ergebnisse der einzelnen Anwendungen in Abbildung 4.14 ist sichtbar, dass einzelne Datenstromtypen wie Außentemperatursensoren sehr gut erkannt werden (bis 99,2 % F1 Wert) und auch Raumregelungen, sowie Raummonitoring zeigen mit bis zu 98,0 % F1 Wert einen hohen Wert. Jedoch werden die Datenströme für Wärmeübertrager, Energiefluss und Quartiersmonitoring mit 82,7 % bis 92,0 % F1 Wert mit einem geringeren Wert erkannt. Bei einer ungefähr angenommenen maximalen Fehlerrate von Menschen von 15 % liegen Wärmeübertrager und Quartiersmonitoring unterhalb dieser maximalen Fehlerquote. Diese drei Anwendungen sind semantisch eng miteinander verknüpft, da es sich in jedem Fall um einen Energiestrom handelt.

Auch Raumregelung, Raummonitoring und auch Außentemperatursensoren sind sehr eng verknüpft, da sich die benötigten Eigenschaften überschneiden. So werden in jedem Fall Lufttemperaturen genutzt und die Temperatur-Regelungsparameter von Raumregelung und Raummonitoring überschneiden sich.

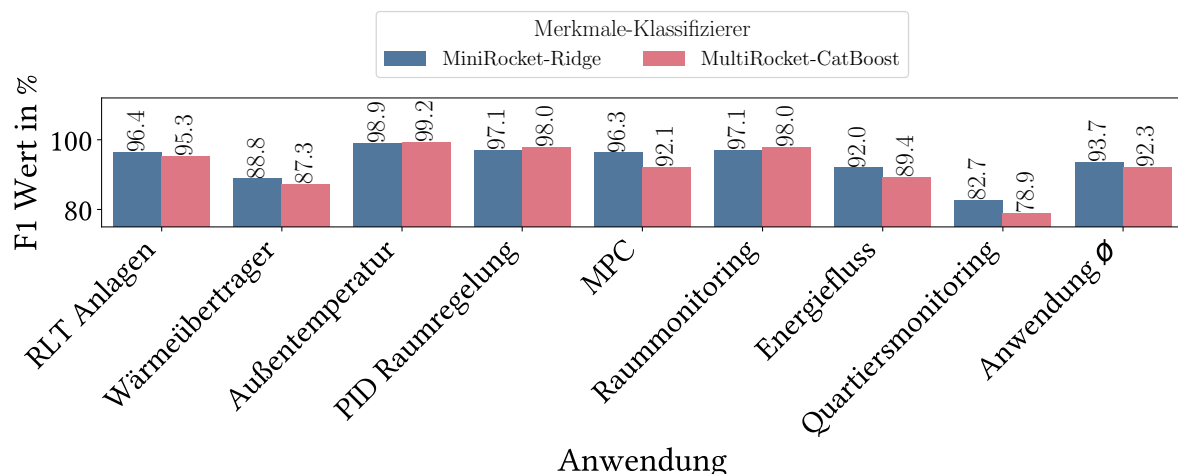


Abbildung 4.14: Vergleich des F1 Wertes der verschiedenen **Anwendungen** bei Berücksichtigung der Datenstromtypen bei den Kombinationen MiniRocket mit Ridge und MultiRocket mit CatBoost auf dem **Mortar** Datensatz.

4.2.2 Ergebnisse Einflüsse von Nutzungsmustern

Bei klassischen Merkmalsextraktionsverfahren wurde überprüft, ob mehr Merkmale, aber auch die Nutzung von unterschiedlichen zeitlichen Auflösungen (jährlich, monatlich, wöchentlich) einen Unterschied ausmacht. In Abbildung 4.15 sind die Ergebnisse des F1 Werts, des ECE, des F1 Wertes bei Multilabel-Projektion und bei Anwendungen dargestellt. Es zeigt sich, dass die zeitliche Auflösung keine große Rolle spielt, da die Ergebnisse im Bereich von 0,5 % F1 Wert schwanken. Einzig der ECE mit einem Abstand von maximal 0,8 % ist bei Betrachtung von wöchentlichen Zeitabständen signifikant besser. Ein Datensatz aus allen temporalen Merkmalen zusammen erreichte sogar leicht schlechtere Ergebnisse mit einem F1-Wert von 88,2 % (nicht abgebildet).

Die Nutzung der Rocket Familie lieferte jedoch in allen IQ-Kriterien außer dem ECE leicht bessere Werte. Das bedeutet, dass wenn nicht die Nutzung von klassischen Merkmalen für die Verständlichkeit eines Algorithmus wichtig ist, sollte die Rocket Familie verwendet werden.

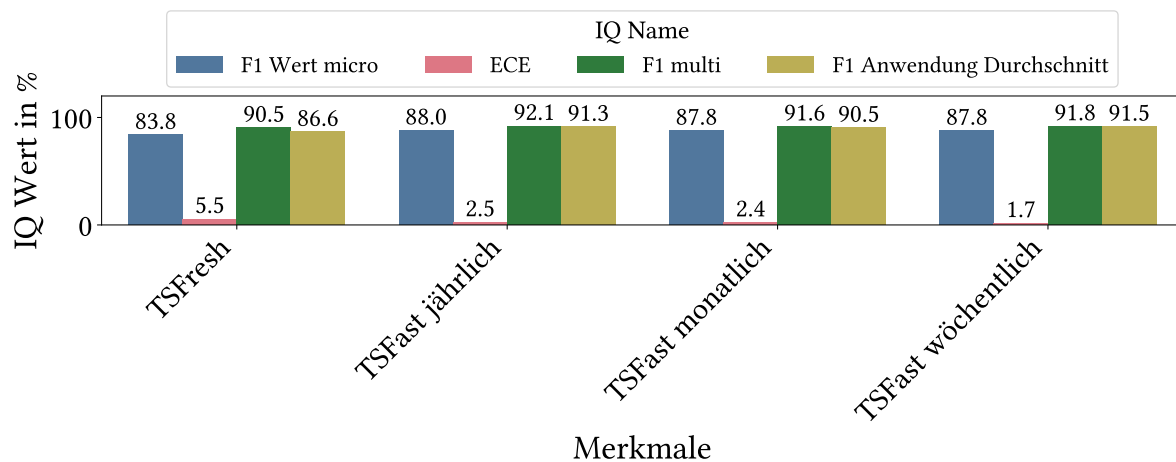


Abbildung 4.15: Vergleich unterschiedlicher Informationsqualitätskriterien (IQ-Wert) im gesamten **Mortar** Datensatz auf Basis von klassischer Merkmalsextraktion. Werte mit „Multi“ sind bei Berücksichtigung einer Multilabel-Zuordnung.

4.2.3 Ergebnisse OOM4ABDO Datensatz

Im Folgenden werden die Ergebnisse des innerhalb des OOM4ABDO Projektes aufgenommenen Datensatzes präsentiert. Abbildung 4.16 zeigt die Ergebnisse von MiniRocket mit CatBoost Klassifizierer. Insgesamt werden unilabel ein F1 Wert von 76,0 % und multilabel ein F1 Wert von 80,9 % erreicht. Diese Werte weichen fast 15 % von den vorher vorgestellten Ergebnissen des Mortar Datensatzes ab.

Der MCC weicht jeweils leicht (0,4 bis 2,7 %) vom F1 Wert ab. Die Anwendungen erreichen jedoch einen F1 Wert von durchschnittlich 89,8 % F1 Wert. Dieser weicht nur 4 % vom durchschnittlichen

F1 Wert des Mortar Datensatzes ab. Das gibt einen Hinweis darauf, dass die Diskrepanz in der Anzahl der anwendungsrelevanten Datenströmen und nicht-anwendungsrelevanten Datenströmen beim

OOM4ABDO Datensatz größer ist. Das bedeutet, dass in einem GAS tendenziell mehr Datenstromtypen aufgenommen werden als für die Anwendung relevant sind. Und dieses Verhältnis signalisiert im Zusammenhang der Ergebnisse, dass der Mortar Datensatz gleichförmiger aufgebaut ist, als der OOM4ABDO Datensatz, wie auch in Kapitel 3.5.3 dargestellt wurde. Dies ist eine einfachere Aufgabe für die Datenklassifizierung.

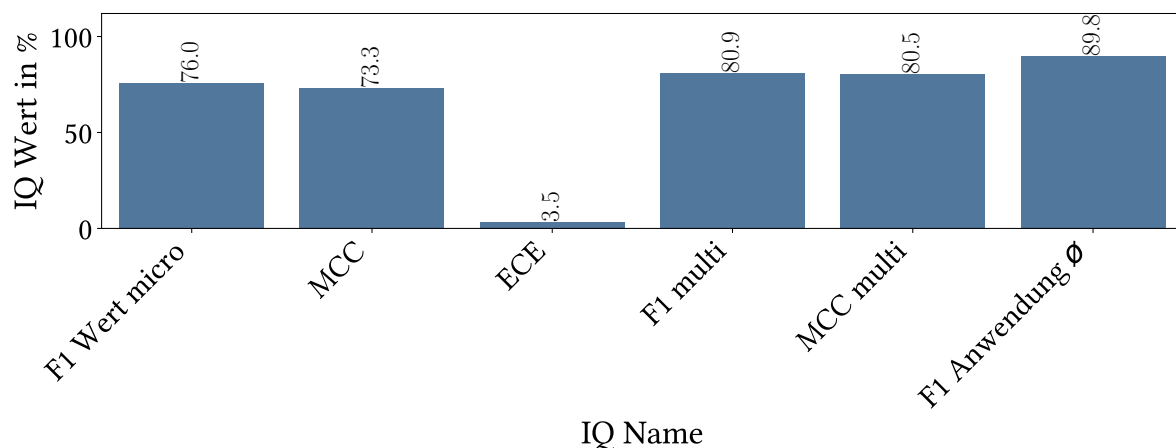


Abbildung 4.16: Vergleich unterschiedlicher Informationsqualitätskriterien (IQ-Wert) im gesamten **OOM4ABDO** Datensatz. Werte mit „Multi“ sind bei Berücksichtigung einer Multi-label-Zuordnung.

In Abbildung 4.17 sind die Ergebnisse anhand der einzelnen Anwendungen aufgeführt. Es ist sichtbar, dass Anwendungen auf einzelnen Datenstromtypen (Außentemperatur, Energiefluss) mit F1 Werten von 92,3 % stärker abschneiden. Dagegen sind vom Datenstromtyp her komplexere Anwendungen wie MPC, Raumregelung und Quartiersmonitoring mit 88,2 bis 88,3 % am schwächsten. Aber es gibt auch jeweilige Anwendungen, die im Mittelfeld mit 89,3 bis 89,8 % abschneiden, obwohl deren Komplexität augenscheinlich nicht stärker unterschiedlich ist, als beispielsweise MPC.

4.3 Anwendungen in Gebäudeenergiesystemen

In diesem Kapitel werden die Ergebnisse der Anwendungen auf Basis der vorherigen Metadatenextraktion beschrieben. Diese umfassen die Fehler- und Anomalieerkennung in Kapitel 4.3.1, die Regelung in Kapitel 4.3.2 und die Energiesystemanalyse in Kapitel 4.3.3.

In den meisten Fällen werden Ähnlichkeitswerte der Zeichenketten für die Anwendung verwendet. Es wurden hierbei sieben verschiedene Ähnlichkeitswerte verwendet, wobei Phonetik nicht verwendet wurde. Die einzelnen Algorithmen zur Berechnung der Ähnlichkeitswerte ist in Kapitel 2.3.4 beschrieben.

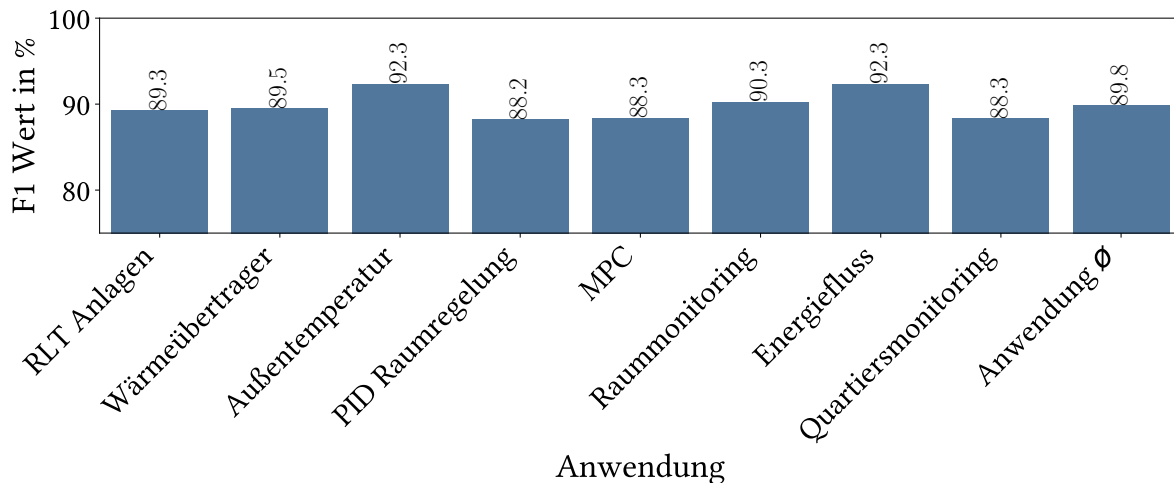


Abbildung 4.17: Vergleich des F1 Wertes der verschiedenen **Anwendungen** bei Berücksichtigung der Datenstromtypen beim **OOM4ABDO** Datensatz.

4.3.1 Fehler- und Anomalieerkennung

In der Fehler- und Anomalieerkennung wird zuerst die „Fehler- und Anomalieerkennung bei RLT-Anlagen“ vorgestellt. Die weiteren Anwendungsfälle sind die „Fehlererkennung bei Wärmeübertragern“ und „Außentemperatursensoren“ und deren Resultate werden im Folgenden vorgestellt.

4.3.1.1 Fehler- und Anomalieerkennung bei raumluftechnischen Anlagen

Die Ergebnisse der Parameterstudie sind in Tabelle 4.1 dargestellt. Die Hauptanteile sind der Prefix mit 43,6 % und der Longest Common Substring (LCSS) mit 41,8 %. Interessanterweise wird die Longest Common Subsequence (LCS) nicht mit einbezogen. Die LCS überspringt auch Buchstaben in einer Sequenz, weshalb der LCSS konsequenter die Hierarchie des Datenstromnamens berücksichtigt und deshalb auch einen großen Anteil beträgt. Die Editierdistanzen „Damerau-Levenshtein“ und „Optimal String Alignment (OSA)“ werden nicht verwendet, da diese auch die Hierarchie nicht berücksichtigen. Eine AHU1 wird von der AHU2 nur durch einen Buchstaben unterschieden, was bei den Editierdistanzen nur einen Unterschied von 1 ausmacht. Bei LCSS oder Prefix wird die gesamte Zeichenkette an dieser Stelle unterbrochen. Die Jaro-Winkler-Distanz überbewertet den Anfang einer Zeichenkette [331], die auch den hierarchischen Teil eines Datenstromnamens enthält. Deshalb wird diese auch in der Konsequenz mit 10,8 % berücksichtigt. Die Cosinus-Distanz erkennt die Ähnlichkeit von Zeichenketten anhand ihrer Vektoren und der Stellung dieser zueinander. Deshalb können einzelne Buchstaben auch einen hohen Unterschied ausmachen. Die Cosinus-Distanz fließt mit einem c_i von 3,8 % in die Bewertung ein.

Bei 20 einzuschränkenden Datenströmen („Top 20“), d.h. einer Auswahl von 20 Datenströmen, die dem Nutzer zur Auswahl für eine RLT-Anlage präsentiert werden, erreichen die Testergebnisse (Abbildung 4.18) eine Sensitivität von 100 %. Die Test-Ergebnisse bei Einschränkung der Datenströme

Distanz	c_i
Jaro-Winkler-Distanz	0.108
Cosinus-Distanz	0.038
Damerau-Levenshtein-Distanz	0.0
Optimal String Alignment (OSA)	0.0
Longest Common Subsequence (LCS)	0.0
Longest Common Substring (LCSS)	0.418
Prefix	0.436

Tabelle 4.1: Anteile c_i der einzelnen Distanzen für die Ähnlichkeitsberechnung von SSC_{total} bei raumluftechnischen Anlagen.

auf 15 („Top 15“) bzw. auf die richtige Anzahl an zusammengehörigen Datenströmen („Top n“) liegen 1,04 % bzw. 2,33 % unterhalb der Ergebnisse der Trainingsergebnisse (nicht abgebildet: 98,32 % bzw. 89,42 %).

Das bedeutet, wenn man einem Nutzer aus 20 Datenströmen auswählen lässt, dass in unserem Datensatz in jedem Fall die richtigen Datenströme für die RLT-Anlage enthalten sind. Eine vorherige Auswahl der Anzahl der enthaltenen Datenströme („Top n“) erscheint nicht sinnvoll zu sein. Da die Auswahl aus 20 oder 15 Datenströmen keinen großen Unterschied in der Praxis macht, ist eine Auswahl aus 20 Datenströmen zu empfehlen.

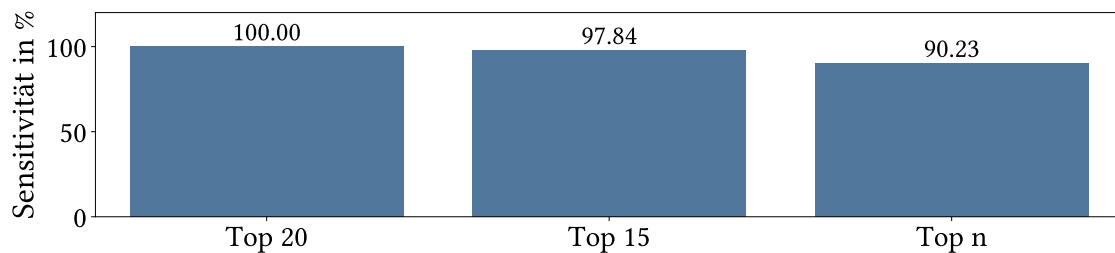


Abbildung 4.18: Ergebnisse bei der Erkennung von zusammengehörigen Datenströmen bei **RLT Anlagen** im **Testdatensatz**.

Eine Alternative um die Datenströme einer RLT Anlage zu ermitteln wäre die Implementierung von festen Grenzen von SSC_{total} bis dahin eine Zuweisung zu einer RLT-Anlage erfolgen kann. Wenn es allgemein gültige feste Grenzen gäbe, könnten diese die menschliche Arbeit zur direkten Zuweisung aus den Top 20 bzw Top 15 oder Bestimmung von n deutlich reduzieren.

In Abbildung 4.19 sind die jeweiligen Verteilungen und kumulierten Anteile der Grenzen der Ähnlichkeiten des maximal noch der zugehörigen RLT-Anlage zugewiesenen Datenstroms („Max Wahr Limit“) und der minimal nicht der RLT-Anlage zugewiesenen Datenstroms („Min Falsch Limit“) abgebildet. Die Verteilung zeigt eindeutig, dass keine eindeutige Grenze feststellbar ist. Der kumulierte Anteil von „Min Falsch Limit“ bei $lim < 0,35$, was die ungefähre Grenze für „Max Wahr Limit“ wäre, beträgt über 50 %. Das bedeutet, dass die Grenzsetzung von 0,35 auch die Grenze in der Hälfte aller Datenströme überschreiten würde. Da die Anzahl der falschen Datenströme in jedem Fall

größer ist, als die Anzahl der wahren Datenströme, müssten deutlich mehr Datenströme manuell aussortiert werden, als beim Verfahren der manuellen Exploration der festen Anzahl der Top 20 Datenströme. Deshalb ist eine Anwendung von festen Grenzen nicht sinnvoll für die Praxis.

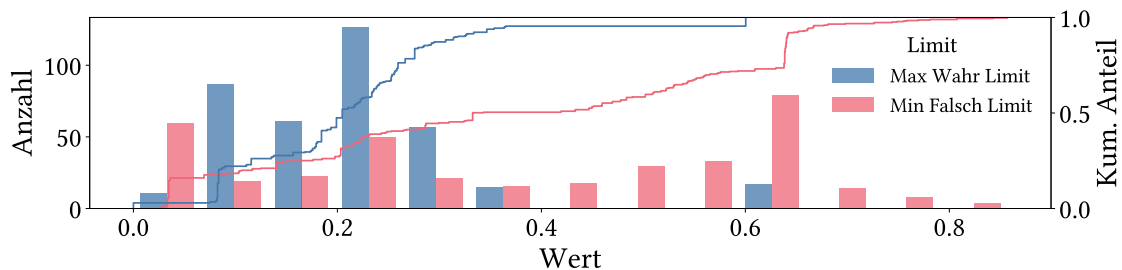


Abbildung 4.19: Verteilung der Limits des Maximums der wahren Datenströme („Max Wahr Limit“) bzw. Minimum der falschen Datenströme („Min Falsch Limit“) bei allen Datenströmen des Trainings- und Testdatensatzes von RLT Anlagen mit den c_i aus Tabelle 4.1.

4.3.1.2 Fehlererkennung bei Wärmeübertragern

In einem ersten Schritt werden alle Datenströme (auch z.B. Alarmer oder Betriebsmeldungen) von Wärmeübertragern betrachtet. Es werden zuerst die gleichen c_i der RLT Anlagen aus Kapitel 4.3.1.1 verwendet. Die Ergebnisse der Sensitivität sind in Abbildung 4.20 zu sehen. Es ist zu erkennen, dass die Ergebnisse mit einer Sensitivität von 99,3 % fast die Ergebnisse der RLT Anlagen erreichen (100 %).

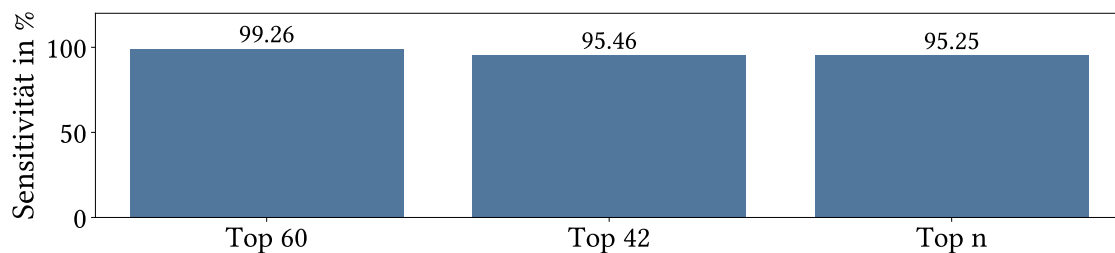


Abbildung 4.20: Ergebnisse bei der Erkennung von zusammengehörigen Datenströmen bei **Wärmeübertragern** im Testdatensatz bei Anwendung der c_i aus Tabelle E.4.

Die Blöcke für alle Datenstromtypen sind sehr inhomogen in ihrer Bezeichnung. Deshalb werden in einem zweiten Schritt nur die Temperaturen eines Wärmeübertragers zusammen betrachtet. Dies verringert sogar die Sensitivität bei „Top 60“ um geringfügige 0,36 % (ohne Abbildung). Somit spielt die Inhomogenität der Datenstromtypen bei dieser Anwendung nur eine geringe Rolle.

4.3.1.3 Fehlererkennung bei Außentempertursensoren

Da bei Außentempertursensoren nur eine Klassifizierung der Eigenschaften „Außen“, „Temperatur“ und („Sensor“ bzw. „Messwert“) stattfindet, wird diese nur in der Klassifizierung in den Kapiteln 4.1 und 4.2 untersucht.

4.3.2 Regelung

Im Folgenden werden als Beispiele für die Regelung in Energiesystemen die Ergebnisse der Anwendungen „Regelung der Raumtemperatur“ und die „modellprädiktive Regelung“ betrachtet.

4.3.2.1 Regelung der Raumtemperatur

Bei Verwendung der c_i aus Stinner u. a. [429] (Tabelle E.5) sind die Ergebnisse im Trainingsdatensatz bei zusätzlicher Nutzung der Datenstrombeschreibung mit 40 % Sensitivität schlechter als bei purer Nutzung des Datenstromnamens (100 %).

Die hohen Werte bei Nutzung des Datenstromnamens von 100 % Sensitivität werden ebenso in allen Fällen im Testdatensatz erreicht. Diese Ergebnisse sind unabhängig von der Nutzung verschiedener c_i Kombinationen der vorherigen Anwendungsfälle.

Die Ergebnisse zeigen, dass die gewählten c_i von Stinner u. a. [429] sinnvoll sind, jedoch eine Konzentration auf die Nutzung des Datenstromnamens erfolgen sollte. Die Datenstromnamen enthalten die Hierarchie des Datenstroms, die Beschreibungen nicht. Die Hierarchie ist aber in jedem untersuchten Datensatz auch bei Räumen im Allgemeinen zu beobachten. Deshalb weisen die Datenstromnamen eine höhere Informationsdichte für das Clustern der Raumtemperaturen auf.

Die Ergebnisse deuten darauf hin, dass die automatische Zuweisung von Temperaturen und Sollwerten mit geringem Personalaufwand erreicht werden kann. Danach kann die Regelqualität evaluiert werden und Fehler im System erkannt werden (z.B. festgesetzte Ventile, zu gering dimensionierte Pumpen etc.).

4.3.2.2 Modellprädiktive Regelung

Bei der „modellprädiktiven Regelung“ wird auf eine Gruppierung der Datenströme verzichtet. Die Ergebnisse der Detektion der benötigten Eigenschaften der Metadaten wird in den Kapiteln 4.1 und 4.2 näher erläutert. Die benötigten Eigenschaften sind im Anhang in Kapitel E.5.5 beschrieben.

4.3.3 Energiesystemanalyse

Im Folgenden werden die Ergebnisse der Energiesystemanalyse vorgestellt. Angefangen wird mit der kleinsten Einheit eines Quartiersenergiesystems, dem „Raummonitoring“. Darauf aufbauend

wird die „Energieflussanalyse“ und dann das „Quartiersmonitoring“ vorgestellt, bei dem Methoden aus dem Raummonitoring sowie der Energieflussanalyse kombiniert angewendet werden.

4.3.3.1 Raummonitoring

Beim Raummonitoring können, wie in Abbildung 4.21 dargestellt, in den Top 25 die Datenströme zu den einzelnen Räumen mit einer Sensitivität von bis zu 99,2 % zugeordnet werden. Interessanterweise sind die Ergebnisse mit 97,9 % bei den Top 14 nicht deutlich geringer als bei den Top 25. Ein Raummonitoring kann mit diesen Ergebnissen in wenigen Schritten direkt eingerichtet werden und kann beispielsweise auch eine automatisierte personalisierte Regelung von Räumen in Apps ermöglichen.

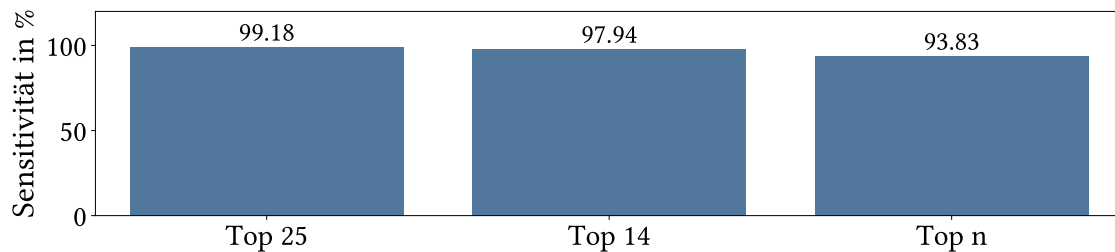


Abbildung 4.21: Ergebnisse bei der Erkennung von zusammengehörigen Datenströmen bei **Raummonitoring** im Testdatensatz bei Anwendung der c_i aus Tabelle E.4.

4.3.3.2 Energieflussanalyse/Topologieerkennung

In Abbildung 4.22 ist das Ergebnis des F1 Wertes, MCC und ECE beim gesamten Datensatz („Total“), aller Wasser-Wasser-Verbindungen und aller Luft-Luft-Verbindungen dargestellt. Der F1 Wert erreicht beim gesamten Datensatz einen Wert von maximal 98,0 %, bei Luft-Luft-Verbindungen erreichen diese sogar bis zu 99,1 % F1 Wert. Die Wasser-Wasser-Verbindungen, die auch im Quartiersmonitoring eingesetzt werden, erreichen jedoch nur Werte von bis zu 95,6 %. Der Ridge Klassifizierer erreicht minimal 91,0 % F1 Wert und ist in jedem Fall schlechter als CatBoost Kombinationen.

Der MCC zeigt, dass die Ergebnisse eine gewisse Abhängigkeit von der Verteilung der Klassen existiert („Verbunden“ und „nicht verbunden“). Der MCC für den gesamten Datensatz erreicht einen leicht nach unten abweichenden Wert von 95,7 % und die Wasser-Wasser-Verbindungen erreichen nur einen MCC von 90,7 %. Der Ridge Klassifizierer ist wieder in jedem Fall schlechter als CatBoost Kombinationen.

Der ECE erreicht bei CatBoost wie bei der Zeitreihenerkennung aus Kapitel 4.2 die niedrigsten Werte. Bei einem Wert von 0,83 % bei allen Temperaturverbindungen erreicht MultiRocket-CatBoost die besten Ergebnisse. Analog zum F1 Wert und MCC ist der Ridge Klassifizierer in jedem Fall

schlechter. Es ist auch sichtbar, dass die Wasser-Wasser-Verbindungserkennung schlechter funktioniert wie die Luft-Luft (minimal 1,74 % gegenüber 0,72 % ECE).

Insgesamt ist die Kombination MultiRocket-CatBoost bei Berücksichtigung aller relevanten Informationsqualitäten die beste Lösung für die Energieflussanalyse. Für die Praxis könnte somit eine unabhängige Verbindungserkennung nur anhand von Temperaturmessdaten genutzt werden, die die Energiesystemanalyse und automatisierte Erstellung von Sankey-Diagrammen, aber auch die automatisierte Abbildung topologischer Zusammenhänge in ontologischen Modellen wie Brick Schema, erleichtern könnte. Die Eigenschaft „Temperatur“ kann vergleichsweise einfach erkannt werden und somit den Aufwand für das Modellieren signifikant senken.

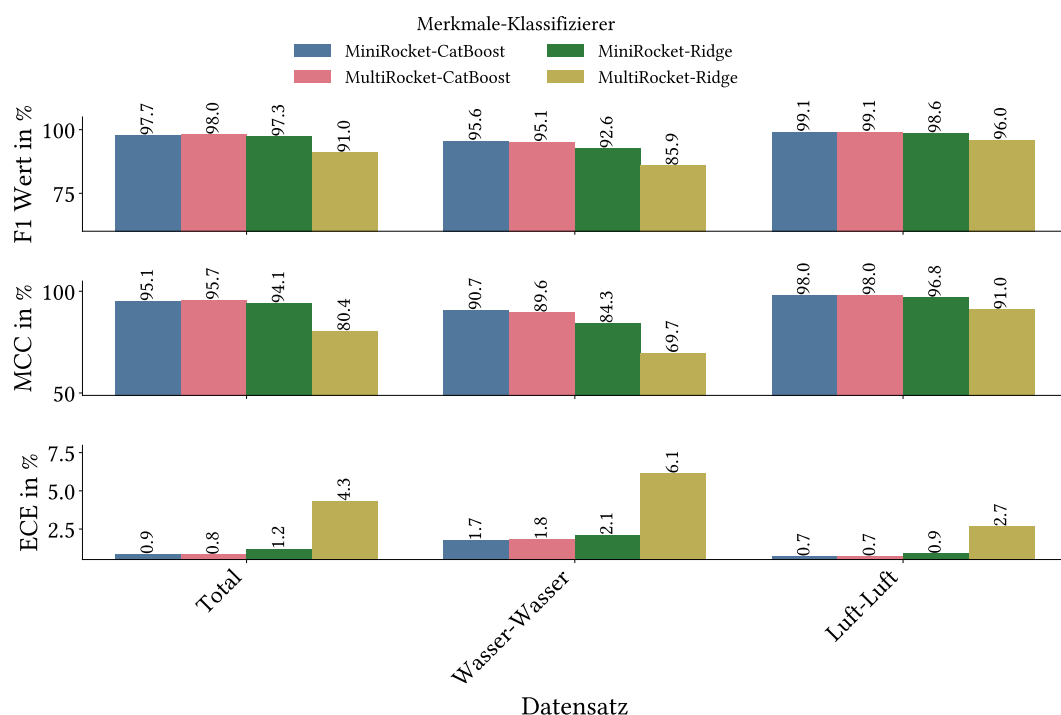


Abbildung 4.22: Vergleich des F1 Wertes, MCC und ECE bei untersuchten Kombinationen von Merkmalsextrahierern und Klassifizierern bei unterschiedlichen Mediumskombinationen (Total, Wasser-Wasser, Luft-Luft) bei der Anwendung „Energiefluss“.

4.3.3.3 Quartiersmonitoring

Im ersten Schritt werden anhand der Datenstromnamen alle Datenströme zu Wärmeübertragern geclustert. In einem zweiten Schritt wird anhand der Zeitreihe überprüft, ob Wärmeübertrager hydraulisch miteinander verbunden sind.

Wie beim Raummonitoring wird nur der zusammengehörige Beginn des Datenstromnamens betrachtet. In allen Auswahlmöglichkeiten wird eine Sensitivität von 99,1 % erreicht (ohne Abbil-

dung). Das bedeutet, dass immer die gleiche Anzahl an korrekten Sensoren miteinander verbunden wird.

Bei alleinigem Einbeziehen der Vorlauftemperaturen („Nur VL T“) werden, wie in der Abbildung 4.23 dargestellt, F1 und MCC Werte von 100 % und ein ECE von minimal 0,07 % erreicht. Dies bedeutet, dass die Vorlauftemperatur alleine ausreicht, um Verbindungen zu 100 % zu erkennen. Interessanterweise werden die Ergebnisse schlechter je mehr Sensoren einbezogen werden („Alle T(emperaturen“): F1 99,7 %, MCC 99,3 %, ECE 0,28 %). In Voruntersuchungen war das Ergebnis besser je mehr Sensoren in den Datensatz integriert wurden [128].

Dies kann darin begründet sein, dass im Fall der Vorlauftemperatur nur einer von zwei Sensoren beim Test der fehlerhaften Verbindungen ausgetauscht wird. Im Fall aller Temperaturen und Volumenströme kann der fehlerhafte Sensor in einem der sechs möglichen Datenströme auftreten, so dass dies für einen Algorithmus schwieriger zu erkennen ist.

Insgesamt erreicht im Schnitt die Kombination MultiRocket-CatBoost die besten Ergebnisse und sollte auch verwendet werden. Es ist in allen Abbildungen sichtbar, dass die Kombination MultiRocket-Ridge in keinem Fall ein in der Realität brauchbares Ergebnis erzielt.

Somit kann gezeigt werden, dass die Vorlauftemperaturen als alleiniger Indikator ausreicht um hydraulische Verbindungen innerhalb eines Quartiersenergiemonitorings zu erkennen. MiniRocket-Ridge oder MultiRocket-CatBoost können für diesen Anwendungsfall empfohlen werden. Das Clustern der Datenströme anhand des Namens war erfolgreich, aber es ist trotzdem minimaler Zeitaufwand des Personals erforderlich. Für große Liegenschaften und Quartiere könnte dieser Ansatz trotzdem eine große Erleichterung in der Strukturierung und Analyse unzähliger Energiezähler sein.

4.3.4 Zusammenfassung Anwendungen

Es konnten bei für unsere Untersuchungen maximale Auswahlmöglichkeiten für Datenströme in jeder Anwendung sinnvolle Ergebnisse mit bis zu 100 % Sensivität erzielt werden. Auch die Zeitreihenerkennung von hydraulischen Verbindungen zwischen technischen Geräten (Topologieerkennung) funktionierte mit einem F1-Wert von 95,7 % in einem sinnvollen Bereich. Es ist trotzdem sichtbar, dass in allen Fällen für eine spätere praktische Anwendung der Ergebnisse der Algorithmen in jedem Fall minimaler menschlicher Einsatz erforderlich bzw. sinnvoll ist.

4.4 Gesamtbewertung Verfahren

Für die Gesamtbewertung des Verfahrens wird zuerst die Datenqualität in Kapitel 4.4.1 betrachtet. Darauf aufbauend werden übersichtsweise die einzelnen IQ-Kriterien in den Kapiteln 4.4.2 bis 4.4.9 dargestellt. Für die Gesamtbewertung des Verfahrens wird in Kapitel 4.4.10 das vollständige

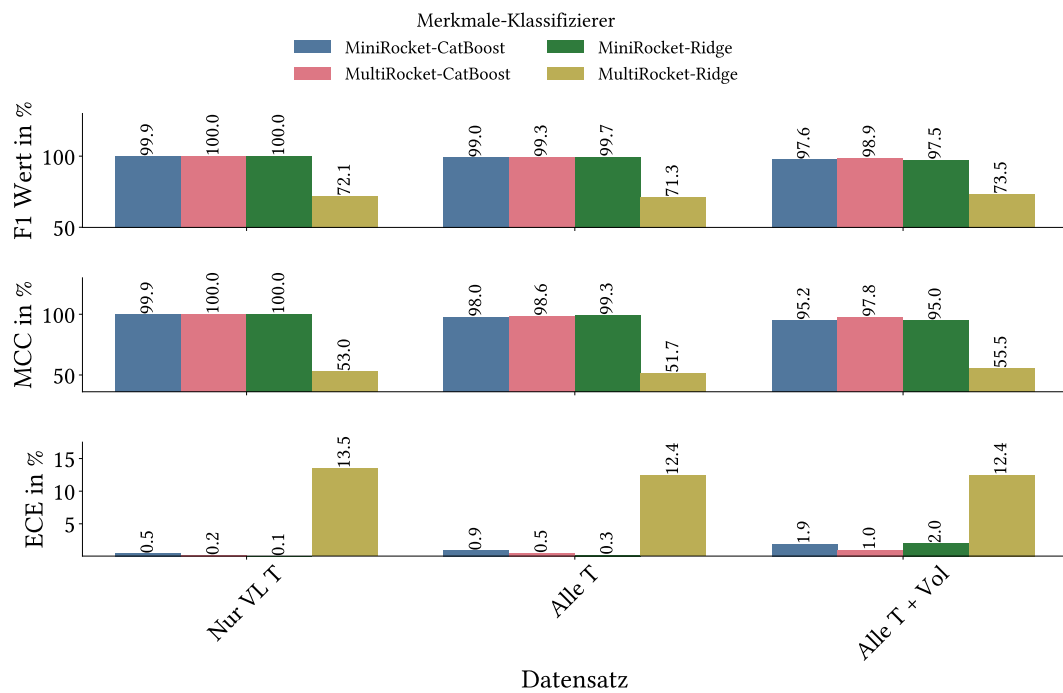


Abbildung 4.23: Vergleich des F1 Wertes, MCC und ECE der untersuchten Kombinationen von Merkmalsextrahierern und Klassifizierern bei unterschiedlichen Datentypskombinationen (Nur Vorlauftemperatur, Vor- und Rücklauftemperatur und inklusive Volumenstrom) bei der Anwendung „Quartier“.

Informationsextraktionsverfahren von der automatisierten Erstellung eines Metadatenmodells bis hin zur Nutzung in den Anwendungen betrachtet.

4.4.1 Bewertung Datenqualität

Die Datenqualität aller in dieser Arbeit verwendeten Datensätze wurde ausreichend überprüft. Es wurden insgesamt drei Datensätze für die Metadatenextraktion genutzt: ein zeichenkettenbasierter und zwei zeitreihenbasierte Datensätze. Für jede der acht Anwendungen in Energiesystemen wurde ein eigener Datensatz erstellt und auf seine ausreichende Datenqualität hin untersucht.

So wurde auf eine Vollständigkeit der Datensätze durch eine ausreichende Auflösung der Datenaufzeichnung und Handhabung von Lücken geachtet. Es wurde die Singularität/Einzigartigkeit aller Datensätze beachtet bzw. diese entsprechend in Trainings- und Testdatensätze unterteilt. Es wurde mit dem BUDO Schema während der Dissertation ein passendes Metadatenschema für Energiemonitoring- und Gebäudeautomationsdaten entwickelt und auf einem über 140 Gebäude fassenden Datensatz angewendet. Diese Benennung bietet die Grundlage für die daraufhin erfolgten Analyse der Daten.

4.4.2 Bewertung Genauigkeit

Die Genauigkeit bot beim zeichenkettenbasierten Verfahren keinerlei Informationsgewinn, da sie immer in einem Bereich von über 99 % liegt. Bei einem unilabel-basierten zeitreihenbasierten Verfahren zeigte die Genauigkeit keinerlei Vorteile gegenüber der Vollständigkeit (F1 Wert), Ausgeglichenheit (Matthews-Korrelationskoeffizient (MCC)) oder Zuverlässigkeit (ECE). Die biasbehaftete Genauigkeit bewegte sich entweder im Bereich der Relevanz oder Sensitivität, die Bestandteil des F1 Wertes sind. Deshalb ist es nicht empfehlenswert die Genauigkeit als zentrales IQ-Kriterium zu verwenden.

4.4.3 Bewertung Vollständigkeit

Der F1 Wert als zentrales IQ-Bewertungskriterium aus der Literatur konnte auch in dieser Arbeit weitestgehend bestätigt werden. Er erreichte bei Zeichenketten einen Wert von bis zu 96,9 %, aber bei Zeitreihen auf dem OOM4ABDO Datensatz nur 80,9 %. Jedoch zeigen die Ergebnisse beim Mortar Datensatz, dass auch F1 Werte von 94,3 % theoretisch erreicht werden können.

Die Diskrepanz liegt vor allem an der unterschiedlichen Zusammensetzung der Datensätze, die auch in Abbildung 4.26 für Anwendungen bestätigt wird, wo die Ergebnisse bei der Zeitreihenerkennung je nach Anwendung nicht weit voneinander entfernt sind.

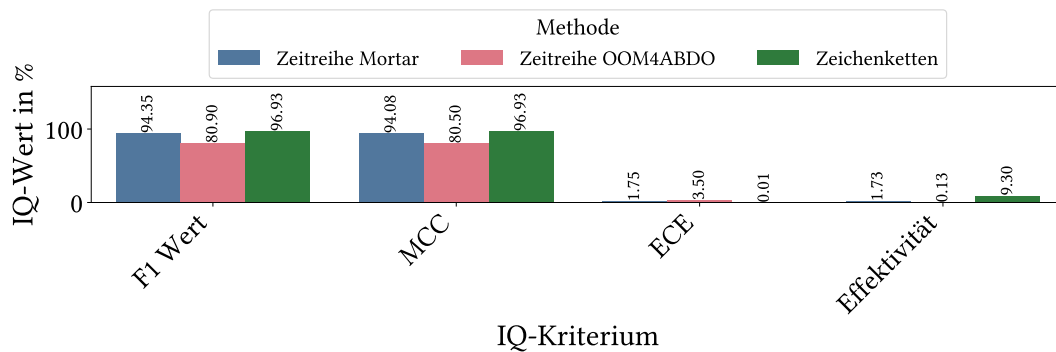


Abbildung 4.24: Vergleich unterschiedlicher IQ Werte bei Anwendung von Zeitreihenklassifizierung, Zeichenkettenklassifizierung. Dies ist eine kompakte Zusammenfassung bisheriger Ergebnisse innerhalb des IE-Prozesses mit der zusätzlichen Darstellung der neu berechneten Effektivität.

4.4.4 Bewertung Ausgeglichenheit

Der Matthews Correlation Coefficient (MCC) liefert bei einem nicht-binären Klassifizierer, wie er hier hauptsächlich angewendet wurde, eine gute Ergänzung zum F1 Wert. Insbesondere der Unterschied zwischen MCC und F1 Wert zeigt an, inwieweit die Ergebnisse einer Verzerrung unterliegen. Je weniger Klassen vorliegen, desto relevanter wird der MCC im IE-Prozess. Die Ergebnisse des MCC in Abbildung 4.24 spiegelt bei vielen Klassen in einer Multilabel-Klassifizierung weitestgehend die Ergebnisse des F1 Wertes wider. Die aus der Literatur empfohlene Anwendung bei binären Klassen konnte sich z.B. bei der Energiefluss-Anwendung zeigen.

4.4.5 Bewertung Konformität

Für die bisher noch nicht bewertete IQ-Dimension der Konformität wird der Cohens Kappa des besten Schätzers der Zeitreihenklassifizierung mit dem besten Schätzer der Metadatenextraktion verglichen. Es werden nur die Datenströme betrachtet, die sowohl in der Datenqualität der Zeitreihe, aber auch der Metadaten eine ausreichende Qualität aufweisen. Wie in Kapitel 3.1.4.4 beschrieben, wird Cohens Kappa mit allen existierenden Kennzeichen und auch nur Paaren von in den realen Daten und der Vorhersage pro Datenstrom enthaltenen Kennzeichen berechnet.

Abbildung 4.25 zeigt, dass der Ratio sich im Bereich 89 bis 97 % bewegt. Das bedeutet, dass das mögliche Potential der Algorithmen auf eine korrekte Vorhersage gut genutzt wird. Der einzelne Cohens Kappa fällt nach Tabelle 3.1 mit 53,4 % nur „mäßig“ aus, was an der geringeren Übereinstimmung der Grundgesamtheit liegt. Diese geringere Übereinstimmung liegt an den unterschiedlichen Methoden zur Nutzung der Metadaten. Während bei Nutzung der Zeichenketten ein Multilabel Verfahren verwendet wurde und mehrere Klassen parallel existieren können, ist für die Nutzung von Zeitreihen diese Unterteilung aufgrund der Nutzung vom Unilabel Verfahren nicht möglich. Es muss sich für eine dominierende Klasse entschieden werden. Der Cohens Kappa von

91,7 % bei Betrachtung aller Klassen zeigt aber auch, dass insgesamt die nicht vorhandenen Klassen „fast perfekt“ von beiden Interratern getroffen werden.

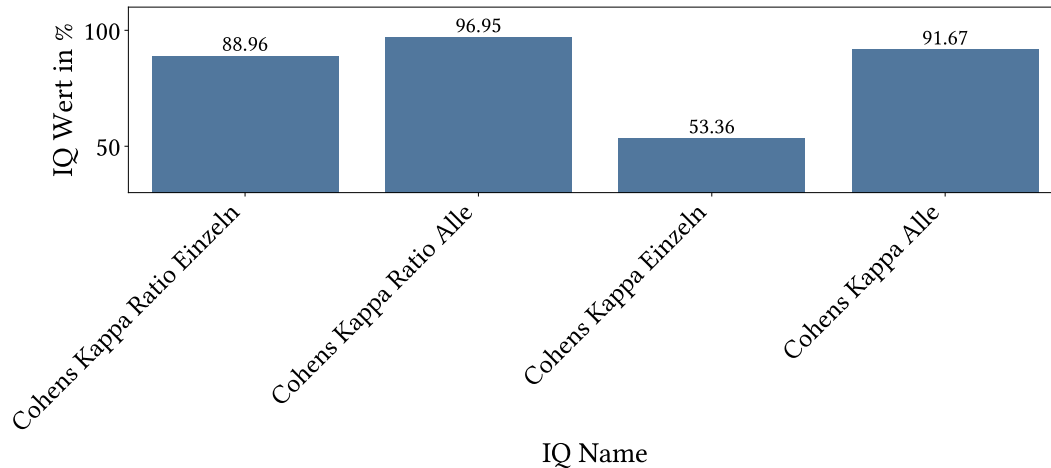


Abbildung 4.25: Vergleich unterschiedlicher Cohens Kappa IQ-Werte. Es wird das Ratio (Berücksichtigung der unterschiedlichen Grundgesamtheiten) und der eigentliche Cohens Kappa dargestellt. Diese werden entweder mit Berücksichtigung einzelner Merkmale oder aller Merkmale berechnet.

4.4.6 Bewertung Zuverlässigkeit

Der Expected Calibration Error (ECE) liefert einen guten Indikator, ob auf die interne Vorhersagewahrscheinlichkeit eines Klassifizierers zuverlässig vertraut werden kann. Er erreicht nach Abbildung 4.24 Werte von 0,01 bis 3,5 %, die jeweils auch in der Realität genutzt werden könnten. Bei 3,5 % ist je nach Nutzung schon eine Grenze erreicht.

Wenn man sich die Ergebnisse der Vorhersagewahrscheinlichkeit im Bereich >95 % und >99 % bei der Zeitreihenerkennung genauer anschaut, ist sichtbar, dass einzelne Klassen wie binäre Signale nicht auftreten. Einzelne Klassen wie Außentemperatursensoren können zu ca. 49 % mit >99 % Vorhersagewahrscheinlichkeit bzw. zu ca. 89 % mit >95 % vorhergesagt werden. Insgesamt erreichen viele Klassen, die die Eigenschaft „Temperatur“ besitzen eine Vorhersagewahrscheinlichkeit von über 95 %. Diese Werte ermöglichen auch eine skalierbare Anwendung insbesondere bei temperaturorientierten Anwendungen.

Da der Maximum Calibration Error (MCE) tendenziell in Instanzen auftritt, die sich unterhalb von 50 % Vorhersagewahrscheinlichkeit liegen, ist der MCE für eine Bewertung der Zuverlässigkeit nicht geeignet.

4.4.7 Bewertung Generalisierbarkeit

Bei der Generalisierbarkeit muss zwischen dem zeichenkettenbasierten Verfahren und dem zeitreihenbasierten Verfahren unterschieden werden.

Das zeitreihenbasierte Verfahren ist immer generalisiert anwendbar, jedoch muss betrachtet werden, ob für die Anwendung die erforderlichen Metadaten extrahiert werden können. So werden Signale, die Werte von 0 bis 1 oder 0 bis 100 aufweisen (z.B. Betriebsmeldungen oder Klappenstellungen) von jedem zeitreihenbasierten Verfahren nicht gut unterschieden. Insbesondere, wenn die Signale nicht gesetzt werden, wie Heizfreigaben im Sommer, kann die Zeitreihe nicht genutzt werden. Jedoch wurden z.B. Temperaturen wie die Außentemperatur sehr gut erkannt und können auch in einem skalierbaren Maßstab eingesetzt werden.

Das zeichenkettenbasierte Verfahren konnte skalierbar einen F1 Wert und MCC von bis zu 93 % erreichen, aber im Durchschnitt nur Werte von 82 bis 84 %. Dies reicht noch nicht für einen vollkommen skalierbaren Einsatz, aber insbesondere durch den Einsatz von LLM könnte der Wert noch stark steigen und ist nicht mehr weit von einem realen Einsatz im Rahmen der menschlichen Fehlerquote (1 bis 15 %) entfernt.

4.4.8 Bewertung Operationalisierbarkeit

Bei der Operationalisierbarkeit wird die Fähigkeit eines Verfahrens bewertet, wie Anwendungen mithilfe der Informationextraktion unterstützt werden können.

Abbildung 4.26 zeigt die Ergebnisse anhand der Zeitreihenerkennung des Mortar-Datensatzes, des OOM4ABDO Datensatzes und Nutzung der Metadatenextraktion anhand der Metadaten des OOM4ABDO Datensatzes. Zusätzlich sind die Ergebnisse des Clusters innerhalb der Anwendungen aufgeführt. Die Ergebnisse des Mortar-Datensatzes werden als höchstmögliche Erkennung als Vergleichsmaßstab aufgeführt.

Es ist sichtbar, dass das Clustern in den Anwendungen fast überall 100 % IQ-Wert erreicht. Es ist auch sichtbar, dass in allen Fällen die Metadatenextraktion auf Basis der Zeichenketten gegenüber der Zeitreihenerkennung die besseren Ergebnisse erzielt auf dem OOM4ABDO Datensatz.

4.4.9 Bewertung Effektivität

Die Ergebnisse der Effektivität (Methode siehe Kapitel 3.1.4.8) wurden bis jetzt noch nicht beschrieben. Die Berechnung der Effektivität besteht aus zwei Teilen: den Vergleich mit Basis-Ergebnissen und die Varianz der erzeugten Ergebnisse bei mehreren Durchläufen. Die Effektivität wird in diesem Beispiel anhand des F1 Wertes der Algorithmen berechnet und in % ausgedrückt. Abbildung 4.24 zeigt auch die Darstellung der Effektivität (Eff).

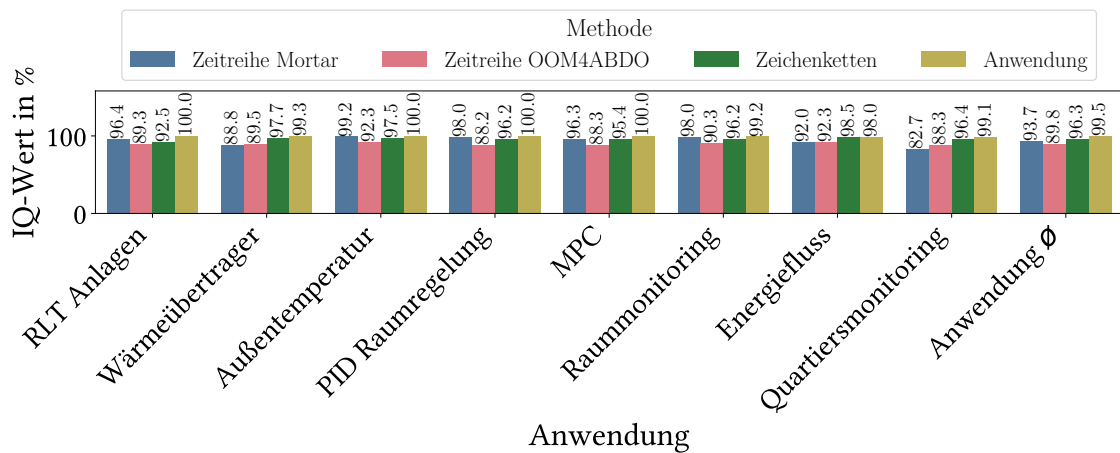


Abbildung 4.26: Vergleich unterschiedlicher IQ Werte bei unterschiedlichen Anwendungen bei Anwendung von Zeitreihenklassifizierung, Metadatenextraktion und der Zuordnung innerhalb der Anwendung. Dies ist eine kompakte Zusammenfassung bisheriger Ergebnisse innerhalb des IE-Prozesses.

Interessanterweise ist die Effektivität beim Zeichenketten Algorithmus (9,3 %) um einen Faktor 5 besser als bei Anwendung des Mortar Datensatzes (1,73 %) und dieser wiederum um den Faktor 13 größer als der bei Anwendung des OOM4ABDO Datensatzes. Das liegt darin begründet, dass das IQ-Kriterium der Effektivität stärker auf Verbesserungen gegenüber einer Basis reagiert, da der vergleichende Faktor des Informationsgewinns gegenüber einer Basis IG_{norm} höher ist als die Berücksichtigung seiner Abweichungen bei mehreren Durchläufen.

4.4.10 Bewertung Gesamtverfahren von Rohdaten zu Anwendung

Bei der Bewertung des Gesamtverfahrens werden die Ergebnisse der Metadatenextraktion und der Zeitreihenklassifizierung mit den Ergebnissen der Anwendungserkennung kombiniert.

Abbildung 4.27 zeigt die Ergebnisse der gesamten IE-Kette, indem die Ergebnisse der Anwendung mit dem jeweiligen Klassifizierungsansatz in der IE-Kette multipliziert werden. Die gesamte Kette darf nur eine Fehlerrate von maximal 15 % besitzen (abgeschätztes Maximum der menschlichen Fehlerquote im technischen Bereich [166, 241, 264, 293, 349]).

Jede Anwendung erreicht eine Fehlerquote von sogar maximal 7,5 % insbesondere bei der Metadatenextraktion durch Zeichenketten, die in jedem Fall die Erkennung durch die Zeitreihen im OOM4ABDO Datensatz überragt. Die Zeitreihenerkennung erreicht dagegen teilweise knapp keine 10 % Fehlerquote.

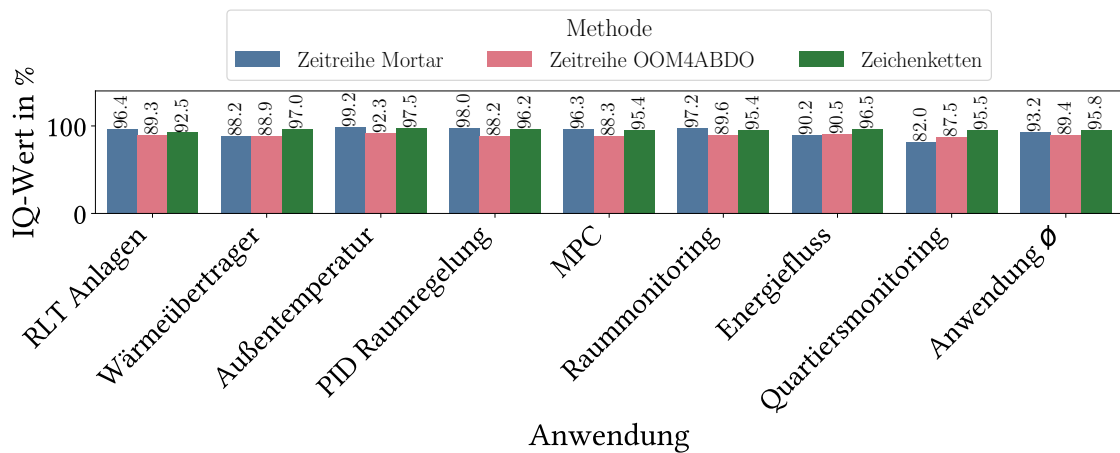


Abbildung 4.27: Vergleich unterschiedlicher IQ Werte bei unterschiedlichen Anwendungen bei Anwendung von Zeitreihenklassifizierung, Metadatenextraktion und der Zuordnung innerhalb der Anwendung. Dies ist die kennzahltechnische Abbildung der **gesamten IE-Kette**.

5 Diskussion

Diese Dissertation zeigt eine Methode auf, wie anhand eines Informationsextraktionsprozesses (IE-Prozess) die Metadaten von Datenströmen aus Energiesystemen extrahiert werden können. Bei diesem Prozess werden Zeichenketten und Zeitreihen als Datenquellen genutzt, so dass acht verschiedene Anwendungen anschließend durchgeführt werden könnten. Es wurden für die Nutzung der Zeichenketten und Zeitreihen jeweils Algorithmen entwickelt bzw. ausgewählt, die bisher noch nicht für Metadatenextraktion in Energiesystemen genutzt wurden. Die Methoden wurden für die Zeichenketten auf insgesamt fünf unterschiedlichen Datensätzen aus dem Projekt OOM4ABDO basierend auf vier verschiedenen Systematiken für Anlagenkennzeichnungsschlüssel (AKS) getestet. Diese fünf Datensätze bildeten die Daten aus der Gebäudeautomation und Energiemonitoring von knapp 150 Gebäuden ab. Für die Zeitreihe wurden zwei Datensätze ausgewählt (Mortar und OOM4ABDO), die jeweils über 100 Gebäude abbilden. Somit wurde auf einer großen Datenbasis anhand von Informationsqualitätskriterien bewiesen, inwiefern der Informationsextraktionsprozess in der Lage ist, die benötigten Metadaten für die untersuchten Anwendungen zu erzielen. Es wurden acht Anwendungen aus den Bereichen Regelung, Fehler- und Anomalieerkennung und Energiesystemanalyse betrachtet. Für jede der Anwendungen wurde ein eigener Datensatz aus dem OOM4ABDO Datensatz erstellt. Jede Anwendung erhielt ein Verfahren auf Basis der Zeichenketten oder Zeitreihen, um eine Verbindung zu einzelnen Einheiten wie einem Raum herzustellen.

Es wurden acht verschiedene Informationsqualitätskriterien (IQ-Kriterien) betrachtet, wovon zwei Kriterien gesondert für diese Dissertation formell entwickelt wurden (Operationalisierbarkeit und Effektivität).

5.1 Einordnung Informationsqualitätskriterien

Es gibt bei den Informationsqualitätskriterien mehrere Ziele, die betrachtet werden müssen. In Kapitel 4.4 wurde ein Überblick über die quantitativen Ergebnisse der IQ-Kriterien des IE-Prozesses gegeben.

Es sollte insgesamt eine Fehlerquote von höchstens 10 % erreicht werden, um der maximalen menschlichen Fehlerquote [241] zu entsprechen. Über den gesamten Informationsextraktionsprozess von der Datenvorbereitung, Merkmalsextraktion über die Klassifizierung und der Gruppierung anhand von Anwendungen in Energiesystemen wird in allen acht Anwendungen ein realer Wert unterhalb von 15 % bzw. sogar unter 10 % Fehlerquote erreicht. Das bedeutet, dass die Informationsextraktion zwar fehlerhaft ist, aber noch innerhalb des menschlichen Fehlerbereichs. Da auf Fehler bei der Informationsextraktion auch Fehler bei der Wissensgenerierung folgen, muss der initiale IE-Prozess möglichst wenige Fehler enthalten.

Es konnte grundsätzlich gezeigt werden, dass mit mehr Daten und Merkmalen bessere Ergebnisse nach Informationsqualitätskriterien erzielt werden können. Es ist aber auch sichtbar, dass praxisbezogen ausgewählt werden muss, wie viel Zeitaufwand in die Aufbereitung von Daten investiert werden soll. Je nach Methode sind hierbei deutlich weniger Daten (ca. 3 %) erforderlich als die 10 %, die in der Literatur als Minimum erachtet werden [468].

Die verwendeten Verfahren erfordern ein unterschiedliches Niveau der manuellen Arbeit. Während Zeitreihen skalierbar auf jedes neue Gebäude und jeden Betreiber angewendet werden können, müssen zeichenkettenbasierte Verfahren an das verwendete Schema des jeweiligen Betreiber angepasst werden. Das zeichenkettenbasierte Verfahren konnte im für die manuelle Arbeit einfacheren Verfahren ca. 91 % und das kompliziertere Verfahren bis zu 97 % F1-Wert erreichen gegenüber 70 % des Verfahrens, das den Stand der Technik abbildet. In der Literatur wurde hier für einen ähnlich herausfordernden Datensatz ein Ergebnis von bis zu 95 % erzielt. Somit konnte der Literaturwert überboten werden.

Das zeitreihenbasierte Verfahren konnte auf einem allgemeinen Mortar-Datensatz als Referenz bis zu 90 % F1-Wert und auf dem speziellen Datensatz aus dem OOM4ABDO-Projekt bis zu 81 % erreichen. Beide Ergebnisse sind im Bereich der Algorithmen aus der Literatur, die hier bei ähnlich komplexer Aufgabe maximal 89 % F1-Wert erreicht haben. Der OOM4ABDO-Datensatz enthält jedoch im Allgemeinen auch im Vergleich zur Referenz mehr schwer zu erkennende binäre Signale wie „ein/aus“. Eingeschränkt auf die für die Anwendung erforderlichen Metadaten erhöht sich der F1-Wert bei der Referenz auf 94 % bzw. 90 % für den OOM4ABDO-Datensatz. Die für die Anwendungen erforderlichen Metadaten konnten somit im Median bis zu 4 % besser erkannt werden als allgemein die Metadaten.

Das bedeutet, dass eine allgemeine Betrachtung der IQ-Kriterien nicht als sinnvoll erachtet wird, sondern die Betrachtung immer auch im Rahmen der Operationalisierbarkeit mithilfe der in dieser Arbeit hergeleiteten operationalisierten Bewertung erfolgen sollte. So wird der Fokus auf die auch

tatsächlich benötigten Metadaten gelegt. Einzelne Klassengruppen wie z.B. Temperaturen können auch bei insgesamt nicht zufrieden stellenden Ergebnissen richtig erkannt werden. Der hier vorgeschlagene Ansatz für die Analyse von Metadaten von Anwendungen bewertet jede Metadatenklasse gleich. So könnten Parameter für die Wichtigkeit von einzelnen Metadatenklassen die Ergebnisse noch genauer an die Anforderungen der Anwendungen anpassen.

Es ist auch aufgedeckt worden, dass einzelne Informationsqualitätskriterien nicht sinnvoll sind weiter zu betrachten, so wie allgemein die Genauigkeit, insbesondere bei Multilabel Verfahren.

5.2 Einordnung Metaextraktionsalgorithmen

Es wurde initial überprüft, welche Metadaten für Datenströme in Energiesystemen in Gebäuden und Quartieren vorhanden sind. So wurde ein Datensatz von ca. 150 Gebäuden mit allen einzelnen enthaltenen Metadaten übersetzt und analysiert. Es wurde im Rahmen der Dissertation das BUDO Schema entwickelt und angewendet für die skalierbare Anwendung von Metadaten als Anlagenkennzeichnungsschlüssel (AKS) in Energiesystemen von Gebäuden und Quartieren. Auch wurden Kennzahlen für die Datensätze entwickelt, die deren Komplexität und Informationsgehalt bewerten können.

Es wurde für die Zeitreihenerkennung eine Unterart der Rocket-Algorithmen (MultiRocket) auf Basis von CNN in die Machine Learning-Bibliothek „sktime“ eingebracht und ein auf Basis von NumPy Arrays basierender schneller klassischer Merkmalsextraktionsalgorithmus (TSFast) entwickelt. TSGlobal hat die existierenden klassischen Merkmalsextraktionsverfahren für Zeitreihen verbunden und mit eigenen Merkmalen erweitert, insbesondere für multivariate Zeitreihen. Es konnte zwar gezeigt werden, dass TSGlobal bessere Ergebnisse liefert als der Referenz-Algorithmus TSFresh, er konnte aber die Rocket-Familie nicht überbieten. Auch konnte keine Dominanz von wöchentlich, monatlich oder jährlich extrahierten Merkmalen ermittelt werden. Somit kann keine Aussage darüber getroffen werden, welches temporale Betriebsverhalten für die Modellbildung entscheidend ist. Es kann davon ausgegangen werden, dass alle Bereiche ähnlich wichtig sind, je nachdem welche Klassen genutzt werden.

TSGlobal bietet aber für die Realität eine Alternative zu komplexen KNN Architekturen, die hohe Rechenleistungen erfordern. So kann in der schnellsten Variante der gesamte Datensatz innerhalb von wenigen Sekunden klassifiziert werden, nur mit wenigen Prozenten unterhalb der Ergebnisse der besten Algorithmen. Insbesondere, wenn eine manuelle, menschliche Überprüfung der Algorithmen z.B. im Rahmen von Human-in-the-loop stattfindet, können besser interpretierbare Merkmale eine Hilfe für den Menschen sein. Im Laufe der Arbeit zeigten beispielsweise die Rocket-Algorithmen Probleme bei der Interpretierbarkeit, die TSGlobal nicht hätte.

Als Klassifizierungsalgorithmus konnten sich CatBoost für beide Verfahren (Zeichenkette und Zeitreihe) und der RidgeClassifier mit zusätzlichem Isolationsbaum für die Zeitreihe durchsetzen. Dies

könnte ein Ansatzpunkt für die zukünftige Weiterentwicklung der Nutzung der Rocket-Algorithmen liefern, die aktuell nur standardmäßig mit RidgeClassifier ohne Isolationsbaum genutzt werden.

Für die Zeichenkette konnte sich zeigen, dass die blockweise Betrachtung von Blöcken auf dem gleichen Datensatz bessere Ergebnisse erzielt als die gesamtheitliche Betrachtung aller Blöcke eines Datenstroms. Die Unterschiede waren jedoch nicht sehr hoch und bei schemaübergreifendem Lernen war die gesamtheitliche Betrachtung sinnvoller, die vor allem von den weiteren entwickelten Merkmalen aus GAS-typischen Eigenschaften profitiert.

Im Rahmen der Dissertation wurden Algorithmen zur Normalisierung auf Basis von Komposita, Synonymen und Korrektur falsch geschriebener Wörter entwickelt, die in der Theorie bessere Ergebnisse der Normalisierung erzielten (z.B. eine höhere Trefferquote bei der Trennung von Komposita). Aber in der realen Anwendung auf den genutzten Daten erzielten aber alle Normalisierungsverfahren keine Verbesserung der Ergebnisse. Auch eine Einschränkung der Datensätze auf häufiger vorkommende Blöcke oder Kennzeichen führte nicht zu verbesserten Ergebnissen. Des weiteren konnte keine große Korrelation von der Komplexität der Datensätze und den Ergebnissen der Algorithmen festgestellt werden. Einzig die genutzte Klassenanzahl hatte einen Einfluss auf das Ergebnis. Dies ist ein überraschendes Ergebnis, da die Literatur tendenziell bessere Ergebnisse bei weniger komplexen Machine Learning Aufgaben erwartet.

Das zeichenkettenbasierte Verfahren ist durch die nötige Anpassung an jedes Anlagenkennzeichnungssystem (AKS) deutlich personalaufwändiger als das zeitreihenbasierte Verfahren, kann aber auch deutlich mehr Metadaten extrahieren. So können in Zeichenketten auch hierarchische Informationen enthalten sein, um welchen Anlagentyp es sich handelt oder auch mehr Signaltypeneigenschaften sichtbar sein. Ein zeitreihenbasiertes Verfahren könnte diese Eigenschaften nur durch stufenweises Herantasten an den schlussendlichen Typ erreichen (z.B. zuerst die Eigenschaft „Zuluft-Temperatur“ und dann darauf aufbauend die Eigenschaft „RLT-Anlage“) und nur in einem weiteren Verfahren die spezifische Anlage. Jede Stufe würde aber eine eigene Fehlerquote in das Verfahren einbringen, die sich bei mehreren gekoppelten Verfahren exponentiell auswirkt, weshalb diese gekoppelten Metadatenextraktionsverfahren in der Realität nur mit großer Vorsicht anwendbar sind. Diese Koppelung, um mit der Zeitreihenklassifizierung z.B. den exakten Anlagentypen zu bestimmen, ist zwar schon im Rahmen der wissenschaftlichen Tätigkeit des Autors für die Unterscheidung zwischen zwei Anlagentypen entstanden, aber bleibt für eine größere Anzahl an Anlagentypen weiterhin eine große Herausforderung.

Bei einer Weiterentwicklung der zeitreihenbasierten Verfahren und größeren Datensätzen hat dieses Verfahren in Zukunft für die Praxis ein größeres Potential genutzt zu werden, da hier keinerlei zusätzlicher manueller Aufwand zum Trainieren erforderlich ist. Für große Gebäudebetreiber, die überall konsequent ein einheitliches AKS verwenden, können beide zeichenkettenbasierte Verfahren verwendet werden. Hierbei kann das eigentlich personalaufwändigere Verfahren („blockweise“), wobei alle Blöcke einzeln benannt werden müssen, durch den deutlich geringeren benötigten Trainingsanteil und die gleichzeitig besseren Ergebnissen als sinnvoller erachtet werden. Die

Nutzung von passenden GAS-basierten Datensätzen zum Training von Large Language Modellen könnte hier die Ergebnisse von beiden Verfahren positiv beeinflussen, wobei nicht zu erwarten ist, dass die Zuweisung aller Blöcke (also personalintensivere Verfahren) besser abschneidet.

5.3 Einordnung Anwendungen

Es wurden insgesamt acht Anwendungen hauptsächlich aus thermischen Energiesystemen in Gebäuden und Quartieren betrachtet. Diese umfassten die „Fehler- und Anomalieerkennung bei raumlufttechnischen Anlagen“, „bei Wärmeübertragern“, und „bei Außentemperatursensoren“, sowie die „PID Regelung der Raumtemperatur“, die „Modellprädiktive Regelung“, das „Raummonitoring“, die „Energieflussanalyse und Topologieerkennung“ und das „Quartiersmonitoring“. Für jede Anwendung wurde ein eigenes Verfahren zum Zusammenführen der Datenströme zu einer Einheit (z.B. RLT-Anlage, Wärmeübertrager oder Raum) entwickelt.

Für zwei Verfahren (Topologieerkennung und Quartiersmonitoring) wurden zeitreihenbasierte Verfahren verwendet. Soweit dem Autor bekannt, wurde für die Topologieerkennung bzw. Relationsinterferenz das erste Mal ein mediumsübergreifendes Verfahren von Wasser-Wasser, Wasser-Luft, Luft-Wasser und Luft-Luft-Verbindungen verwendet. Es konnten F1 Werte für die Verbindungserkennung je nach Mediumskombination von 95,6 % bis 99,1 % erreicht werden. Der universelle Ansatz ohne explizite Betrachtung der Mediumskombination konnte 97,7 % F1 Wert erreichen. Das bedeutet, dass nur 2,3 % für eine perfekte Klassifizierung fehlen, was insbesondere den menschlichen Aufwand für die Verbindungserkennung deutlich senken kann. In der Literatur wurde für einen Ansatz, der nur Luft als Medium verwendet hatte, nur Ergebnisse von bis zu 94 % erreicht, während in dieser Arbeit die 99,1 % erzielt wurde.

Für die restlichen sechs Anwendungen wurden Verfahren verwendet, die auf einer Ähnlichkeit von Zeichenketten basieren und sehr gute Ergebnisse mit Sensitivitäten von in den meisten Fällen bis zu 100 % zeigten. In jeder Anwendung konnte der benötigte menschliche Aufwand stark reduziert werden, da in GAS-Systemen überwiegend hierarchisch benannt wird. Die Clustering-Verfahren müssten für eine weitere Anwendung stärker mit praxisnahen Auswahl-Verfahren gekoppelt werden, um den benötigten menschlichen Aufwand weiter zu reduzieren. Aber eine feste Grenze wie „es existiert immer nur eine Temperaturmessung im Raum“ war hierbei nicht sinnvoll, da auch Räume existieren, die mehrere Temperaturmessungen aufweisen (z.B. ein Veranstaltungssaal).

Jede der Anwendungen zielt auf einen Teil der größten Energienutzungen in thermischen Energiesystemen ab. Diese Arbeit konnte zeigen, dass die Informationsextraktion in diesem Ansatz nur noch wenige menschliche Eingaben benötigt, um die Anwendungen direkt zu unterstützen.

Abbildung 5.1 zeigt die Einordnung der Einsparpotentiale menschlicher Arbeit und das Einsatzpotential der Verfahren für reale Prozesse in Unternehmen. Ein Prozess von der Bereitstellung einer Datenstromliste über deren Aufbereitung und der automatischen Anwendung wird dargestellt.

Die Aufbereitung der Daten wird durch die hier vorgestellten Verfahren sehr gut unterstützt, auch die automatische Anwendung wird gut gefördert, führt aber zu einer erhöhten Fehlereinbringung in die Anwendungen. Insgesamt ist die Anwendbarkeit abhängig von dem Quotienten aus möglicher Energieeinsparung zum benötigten manuellen Aufwand zur Aufbereitung der Daten. Wenn nur wenige Datenströme für eine potentiell hohe Energieeinsparung manuell zusammengebracht werden müssen, so ist ein automatisiertes Verfahren nicht sinnvoll einzusetzen. Dieses würde im Vergleich zum menschlichen Einsatz zu viele Fehler einbauen und somit nicht das volle Energieeinsparpotential ausschöpfen. Bei der Identifikation von vielen Datenströmen für eine niedrige Energieeinsparung muss ein Einsatz von automatisierten Verfahren im Bezug auf seine Wirtschaftlichkeit betrachtet werden. So ist die Einbringung von Entropie bzw. Störgrößen in die Regelung durch die eventuell fehlerhafte Zuweisung von Metadaten ein Risiko für die automatisierte Nutzung eines Metaextraktionsverfahrens für energieeinsparende Anwendungen. Schlussendlich existiert sogar ein Risiko für die Steigerung des Energieverbrauchs.

Die Fehler- und Anomalieerkennung wird insgesamt gut unterstützt, da hier nur Analysefunktionen betrachtet werden, aber auch insgesamt viele Datenströme für ein je Anlage automatisiert nur niedriges bis mittleres Einsparpotential zusammengeführt werden müssen. Außentemperaturen werden hervorragend erkannt, aber es sind in einem einzelnen Gebäude nur wenige Außentempersensoren als zentrale Regelungseingänge erforderlich. Somit ist zwar eine potentiell hohe Energieeinsparung erreichbar, aber der benötigte menschliche Aufwand ist nur sehr gering. Dies ist bei raumlufttechnischen Anlagen oder Wärmeübertragern anders, die jeweils deutlich mehr Datenströme umfassen, aber pro Anlage nur ein tendenziell geringes Einsparpotential besitzen.

Bei der Regelung ist der automatisierte Einsatz der Algorithmen fraglich. Es müssen zwar viele Datenströme einzeln zusammengebracht werden, aber eine fehlerhafte Kombination von Mess- und Sollwerten hat ein hohes Potential die technischen Anlagen zu gefährden. Insbesondere bei der modellprädiktiven Regelung werden bei Betrachtung der Gebäudeebene nur wenige Datenströme benötigt und es existiert ein hoher Einfluss auf den Energieeinsatz. Somit kann hier ein manuelles Verfahren sehr viel sinnvoller eingesetzt werden. Die Bewertung des Einsatzpotentials ist hierbei abhängig von der Ebene des Einsatzes der modellprädiktiven Regelung. Bei der direkten Verteilung der Energie besteht ein höheres Einsatzpotential als bei der Bereitstellung der Energie. Bei Anwendung z.B. auf Raumebene existiert wiederum ein höheres Einsatzpotential für die in dieser Arbeit vorgestellten Algorithmen, da das Energieeinsparpotential je Raum niedrig ist und die benötigten Datenströme viele sind. Beim Energiesystem selbst sind abermals nur Analysefunktionen erforderlich und deren Zusammenbringung kann aufgrund von entweder nicht existierenden, fehlerhaften bzw. schwierig aufzubereitenden Dokumentationen sehr herausfordernd für den menschlichen Einsatz sein. Hier kann insbesondere bei der Topologieerkennung das hier vorgestellte mediumsunabhängige Verfahren eine sinnvolle Aufbereitungsart darstellen.

Das Abgleichen der theoretisch zu verwendeten Metadaten und der praktisch im Datensatz aus dem Projekt OOM4ABDO enthaltenen Metadaten zeigte auf, dass insbesondere bei der Fehler- und

Anomalieerkennung von RLT-Anlagen von 129 theoretisch zu analysierenden Fehlern nur 26 real mithilfe der in der Gebäudeautomation aufgenommenen Daten analysiert werden können. Somit wurde ein großer Forschungsbedarf für eine realitätsnähere Untersuchung von Fehlererkennungsverfahren identifiziert.

In der aktuellen anwendungsbezogenen Berechnung mit den in den Ergebnissen genutzten Parametern werden alle Anwendungen gleich bewertet. Nicht jede Anwendung ist für die Zielsetzung eines Anwendungsportfolios gleich wichtig. Deshalb sollte in Zukunft eine Gewichtung nach Einfluss auf die Ziele des Anwenders erfolgen (z.B. welche Energie- bzw CO₂-Einsparung möglich ist).

Bei der Bewertung der Metadatenextraktionsverfahren konnten vor allem Anwendungen nicht überzeugen, die auch Signale im Bereich 0 bis 1 oder 0 bis 100 benötigen. Diese Signale können Ventil- oder Klappenstellungen, aber auch Betriebsmeldungen umfassen. Diese können von allen Metadatenextraktionsverfahren schlecht erkannt werden, wie insbesondere bei der Fehlererkennung in RLT-Anlagen zu sehen ist. Dagegen konnten die Metadaten von Anwendungen, die hauptsächlich auf Temperaturen basieren, sehr gut erkannt werden. Das bedeutet, dass für diese Anwendungen der Weg deutlich kleiner ist, um in der Praxis eingesetzt werden zu können.

Für die Praxis sollte ein Metadatenextraktionsverfahren in Zukunft immer für mehrere Anwendungen bewertet werden und darauf geachtet werden, ob bestimmte, schlecht erkennbare Signaltypen tatsächlich für die Anwendung relevant sind oder auch ohne diese gute Ergebnisse erzielt werden können. Methodisch bietet die mediumsunabhängige Topologieerkennung die Möglichkeit ein gesamtes Energiesystem in allen Ebenen zu erfassen, die größten Verbraucher zu identifizieren und beispielsweise für die Fehlererkennung die originäre Fehlerquelle zu identifizieren.

5.4 Einordnung Gesamtverfahren

Soweit dem Autor bekannt, wurde in dieser Arbeit das ersten Mal ein klassischer Informationsextraktionsprozess auf die Metadaten von Gebäudeautomationssystem angewendet. Dieser Informationsextraktionsverfahren wurde anhand der sieben Ebenen des DFIG-Modells für die Datenfusion betrachtet. Der Fokus der Arbeit lag hierbei auf den Ebenen 0 (Vorverarbeitung) bis Ebene 2 (Kontextbewertung). Eine Gefährdungsbewertung (Ebene 3) wurde in Kapitel 4.4.10 vorgenommen, um zu überprüfen, welche Anwendungen überhaupt automatisch durchgeführt werden können. Bei der Ressourcenüberprüfung (Ebene 4) kann überprüft werden, ob den Verfahren mehr Ressourcen zur Verfügung gestellt werden könnten, um bessere Ergebnisse zu erzielen. Diese Ressourcen könnten Rechnerleistung, aber auch menschliche Arbeitsleistung umfassen. Jeder der Algorithmen würde wahrscheinlich von mehr Trainingsdaten profitieren, dessen Grenznutzen ist aber begrenzt, wie in der Arbeit gezeigt. Ab einer gewissen Menge an Trainingsdaten könnten wiederum fortschrittliche KNN Algorithmen verwendet werden. Insbesondere der sprachenbasierte Ansatz würde von unterschiedlichen „Sprachen“, also verschiedenen Anlagenkennzeichnungsschlüsseln

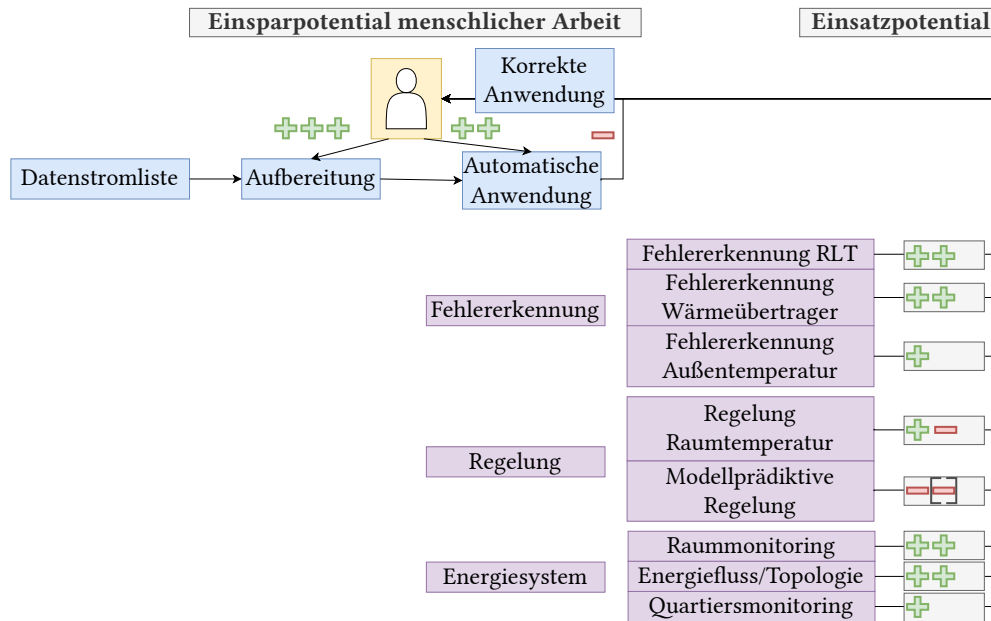


Abbildung 5.1: Einordnung Gesamtverfahren für das Einsparpotential bei menschlichem Einsatz und Einsatzpotential der Anwendungen im Gesamtverfahren. Regelungsanwendungen sind aufgrund ihrer komplexen Zusammenhänge nur eingeschränkt anwendbar.

(AKS), profitieren. Die Analyse der Datensätze weist darauf hin, dass jeder Datensatz unterschiedliche Metadaten aufweist, weshalb ein variantenreicher Datensatz mit unterschiedlichen energiesystemtechnischen Anforderungen helfen würde einen allgemeinen, gebäudeübergreifenden Ansatz zur Optimierung der Energiesysteme zu erreichen.

Bei dem zeichenkettenbasierten Verfahren muss untersucht werden, ob aktuellere Large Language Models wie GPT von OpenAI [341], Claude von Anthropic [21] oder Llama von Meta [445] bessere Ergebnisse liefern. Die LLM könnten von den auch in dieser Arbeit genutzten GAS-fokussierten Datensätzen profitieren. Diese Algorithmen würden jedoch voraussichtlich ein vielfaches des Energieverbrauchs der in dieser Dissertation betrachteten Ansätze benötigen.

In allen Bereichen des IE-Prozesses konnten der Literatur vergleichbare oder bessere Ergebnisse erzielt werden. Zeitreihen sind eine universelle Quelle für die Identifizierung der Metadaten. In den hier vorliegenden Ergebnissen beim F1-Wert war die Zeitreihenerkennung jeweils im Durchschnitt schlechter ($\sim 90\%$) als die Metadatenerkennung anhand der Zeichenketten auf dem gleichen Datensatz (95-99%). Wenn aber nur die Kennzeichen betrachtet wurden, die auch für die Anwendungen relevant sind, ergaben sich auch bei der Zeitreihenerkennung Werte von bis zu 94%. Interessanterweise konnte trotz unterschiedlichster Datensätze selbst beim gebäudeübergreifenden Versuch der Zeichenketten konsequent ein Ergebnis oberhalb von 80% erreicht werden.

Folglich können beide Verfahren die erforderlichen Metadaten-Informationen für die meisten An-

wendungen bereitstellen. Obwohl ein öffentlich zugänglicher Datensatz (Mortar-Datensatz) verwendet wurde, zeigt die Analyse, dass in der Literatur nicht immer alle Metadaten zur Klassifizierung genutzt wurden, sodass ein quantitativer Vergleich der Zeitreihenanalyse nur qualitativ erfolgen kann. Demnach können mithilfe der Zeitreihenanalyse der Dissertation ähnliche Ergebnisse erzielt werden, wenn man davon ausgeht, dass mehr Klassen auch zu einem schlechteren Ergebnis führen. Wie der Nutzer in den Informationsextraktionsprozess besser eingebracht werden kann (Ebene 5), sollte im Weiteren untersucht werden. Da der Informationsextraktionsprozess die benötigten Nutzereingaben minimieren soll, muss dieser mit dem menschlichen Fehler verglichen werden.

Es ist fraglich, ob es eine „allgemeine“ menschliche Fehlergrenze von 10 % existiert oder ob die Fehlerquote von der Aufgabe abhängig ist. So können menschliche Fehlerraten je nach Aufgabe auch höhere Werte aufweisen [131]. Auch wenn in dieser Arbeit ein Wert aus der Literatur ausgewählt wurde, die der Aufgabe der Zuweisung von Metadaten im Energiesystem am ähnlichsten ist, kann nicht ausgeschlossen werden, dass Menschen bei dieser Aufgabe auch entweder bessere oder schlechtere Werte erzielen. Um diese These zu überprüfen, müssten Versuche mit Probanden vorgenommen werden, die genau diese Aufgabe durchführen analog wie diese auch vom Autor dieser Dissertation in van Stiphoudt u. a. [450] als Mitautor für die Fehlerimplementierung in Modelica Modellen demonstriert wurde.

Für die Zielüberprüfung (Ebene 6) ist eine gute Implementierung des IE-Prozesses in die Organisationsstruktur des durchführenden Unternehmens erforderlich. Die Implementierung von Prozessen zur Informationsextraktion in die Organisationsstruktur von Unternehmen kann eine große Herausforderung sein, die insbesondere in einer konservativen Unternehmensumgebung, die in der HLK-Industrie oftmals vorzufinden ist, schwierig zu implementieren sein kann. Algorithmisch liefern die hier vorgestellten Algorithmen gute Leistungen, aber für die Implementierung in der Praxis sind noch viele Schnittstellen erforderlich (z.B. eine einfache Bedienbarkeit mit minimalem Arbeitsaufwand).

Die Bedienbarkeit der Ausgaben der Algorithmen ist hierbei ein entscheidender Faktor. Der Nutzer hat für die Aufarbeitung der Daten nur wenig Zeit und die Kosten steigen mit der benötigten Expertise der Nutzer. Wenn für jeden Datenstrom oder -bestandteil eine einzelne Eingabe erforderlich ist, so kann ein Algorithmus sein Potential nicht entfalten. Für eine skalierbare Anwendung müssten mehrere Datenströme sehr schnell gemeinsam bewertet werden können. Hierbei liefern die in dieser Arbeit betrachteten Algorithmen und Bewertungsmethoden eine sinnvolle Ergänzung zum bestehenden Stand der Technik.

Der insgesamt Prozess muss von professionellen Experten des Machine Learnings und der Gebäudeautomation oder technischen Gebäudeausrüstung (TGA) begleitet werden. So können zwar einzelne Aufgaben wie die Übersetzung von Blöcken von jedem Professionellem in dem Bereich der TGA durchgeführt werden, aber die Steuerung des ganzen Prozesses muss wieder von Experten des Machine Learning in Absprache mit den jeweiligen Fachpersonen übernommen werden.

So muss überprüft werden, ob die jeweiligen Anwendungen auch tatsächlich im Alltag nutzbare Informationen bzw. Wissen generieren.

6 Fazit und Ausblick

In dieser Arbeit wurde anhand eines Informationsextraktionsprozesses (IE-Prozess) für die Gebäudeautomation gezeigt, dass die erforderlichen Metadaten für acht verschiedene Anwendungen von Energiesystemen anhand von Zeitreihen und Zeichenketten aufbereitet werden können. Die Anwendungen sind in den Bereichen „Fehler- und Anomalieerkennung“, „Regelung“ und „Energiesystemanalyse“ einzuordnen. Für die meisten Anwendungen wurde vormalig entweder keine oder eine sehr eingeschränkte Metadatenanalyse vorgenommen. Erstmals wurde ein mediumsunabhängiges Verfahren zur Verknüpfung von zwei Temperatursensoren (Topologieerkennung bzw. Relationsinterferenz) vorgenommen. Insgesamt wurde ein Datensatz aus ca. 150 Gebäuden aus unterschiedlichen Nutzungsbereichen verwendet, das die Nutzbarkeit bei multilingualen und heterogenen Gebäudeautomationssystemen gleichermaßen skalierbar möglich macht, im Vergleich zu den meisten bisherigen Ansätzen.

Es wurden unterschiedliche Informationsqualitätskriterien für die Untersuchung verwendet, wobei der erwartete Kalibrierungsfehler (engl. Expected Calibration Error: ECE) das erste Mal im Kontext von Klassifizierungen in Energiesystemen verwendet wird und sich als sinnvolle Ergänzung für die Bewertung der Zuverlässigkeit der Vorhersagen eines Klassifizierers erwiesen hat. Der ECE bietet die Möglichkeit einzelne Teilergebnisse zu verwerten, obwohl das Gesamtergebnis eines Algorithmus nicht für eine Praxisanwendung geeignet ist. Nicht jedes Informationsqualitätskriterium ist tatsächlich auch für die Anwendung in der Realität nützlich, so dass zur Reduzierung des IE-Prozesses diese Kriterien im Rahmen von Energiedaten nicht weiter betrachtet werden sollten.

Somit konnten alle erforderlichen Aufgaben aus der Fragestellung (Kapitel 2.4) erfolgreich abgeschlossen werden.

Für den Informationsextraktionsprozess wurden die Zeitreihen und Zeichenketten von Datenströmen aus Gebäudeautomationssystemen verwendet und analysiert inwiefern die Anwendungen vom IE-Prozess profitieren können. Für das Zeitreihenverfahren wurde ein schneller klassischer Merkmalsextrahierer (TSFast) entwickelt und ein Merkmalsextrahierer auf Basis von Rocket-Algorithmen (MultiRocket) in eine große Machine Learning Bibliothek für Zeitreihenanalyse (sktime) integriert. Diese Algorithmen besitzen Vorteile aufgrund ihrer zeitlichen und energetischen Effizienz für die Fehlererkennung in Energiesystemen. Die extrahierten Merkmale aus TSPFast können auch leichter von Menschen verstanden werden. In der Zukunft können solche Algorithmen insbesondere die Fehlererkennung in realen Systemen leichter verfügbar machen.

Für das Zeichenkettenverfahren wurden zwei unterschiedlich personalintensive Verfahren entwickelt und aus der natürlichen Sprache mehrere Merkmale eingebracht (wie z.B. Phonetik). Es wurden Verfahren zur Normalisierung eines Datensatzes auf Basis der Trennung von Komposita, Korrektur fehlerhaft geschriebener Wörter und Austausch semantisch ähnlicher Wörter entwickelt. Keines dieser Verfahren bot einen Vorteil gegenüber keiner Normalisierung.

Die zentralen beiden entwickelten Metadatenextraktionsverfahren auf Basis von Zeitreihen und Zeichenketten unterscheiden sich nicht stark in den Ergebnissen, aber stärker in dem benötigten Personalaufwand zur Durchführung. Einzelne Anwendungen konnten im gekoppelten Verfahren (Metadatenextraktion und Clustering-Verfahren) praxistaugliche Ergebnisse erzielen. Insgesamt konnten für sechs der acht Anwendungen im gekoppelten Verfahren Ergebnisse im Bereich des menschlichen Fehlerbereichs erreicht werden. Die anderen Anwendungen verpassen dieses Ziel nur sehr knapp.

Für die Bewertung des Gesamtverfahrens hin zu den Anwendungen ist die Gesamtfehlerrate entscheidend. Bei einer Fehlerrate von 10 % des IE-Prozesses und einer 10 % Fehlerquote einer Fehlererkennung sind bei Kopplung der Ergebnisse nur 81 % der rückgemeldeten Fehler auch reale Fehler. In der Praxis würde solch ein Verfahren wahrscheinlich nicht vollständig akzeptiert werden, wenn fast 20 % der gemeldeten Fehler nicht existieren und Fehlalarme sind. Hier würde es dann auf die Präsentation der Fehler ankommen, so dass schnell eine Entscheidung gefällt werden kann, ob es sich wirklich um einen Fehler handelt. Wie diese Präsentation praxisnah geschehen kann, ist weiterhin eine Forschungslücke.

Nichtsdestotrotz würde ein automatisiertes Vorgehen zur Metadatenextraktion Arbeit abnehmen, da nur ein minimaler einfacher menschlicher Eingriff erforderlich ist, wie das Überprüfen von Zuweisungen. Zukünftig müssten Versuche mit Probanden durchgeführt werden, um die menschliche Fehlerrate und somit die tatsächlichen Grenzen des IE-Prozesses zu definieren. Erst dann könnte eine Bewertung von maschinellen Algorithmen in einem fairen Vergleich mit dem Menschen stattfinden. Auch darf eine Maschine im Zweifel eine etwas höhere Fehlerrate aufweisen, wenn Menschen nur noch die Zuweisungen selbst überprüfen müssen.

Da die Entwicklung der Kapazität des maschinellen Lernens und der verfügbaren Daten aktuell einen exponentiellen Verlauf nimmt, ist zu erwarten, dass zum einen bessere Klassifizierungsalgorithmen, aber auch größere Datensätze genutzt werden können. Neuartige Klassifizierungsalgorithmen wie Large Language Models (LLMs) könnten für zeichenkettenbasierte Verfahren eine revolutionäre Veränderung hervorrufen und die aktuell vorhandenen Lücken in der Performanz und dem benötigten Personalaufwand schließen. Aber auch die aktuell stattfindende Normalisierung von Metadaten in Energiesystemen von Gebäuden und Quartieren kann die Bereitstellung von mehr nutzbaren Daten verbessern.

Es ist einfacher, Metadaten direkt bei initialer Erstellung eines Gebäudes oder Quartiers hinzuzufügen. Die Entwicklung des BUDO Schemas als Anlagenkennzeichnungsschlüssel (AKS) liefert hierzu einen Beitrag, wie die Metadaten initial mit aktuell etablierten Prozessen hinterlegt wer-

den können. Aber auch die Anwendung des BUDO Schemas liefert nicht alle Daten, die für alle Anwendungen erforderlich sind, wie z.B. die Verbindungen zwischen Erzeuger und Verbraucher. Hier könnten pure ontologiebasierte Ansätze wie das Brick Schema eine Lösung sein. Die Analyse der in dieser Arbeit genutzten Metadaten ergab, dass noch eine große Lücke bei den aktuellen Metadatenschemata vorhanden ist, die noch geschlossen werden sollte.

Zusätzlich sollte bei theoretischen und realen Anwendungen immer ein Abgleich mit den Metadaten von größeren Datensätzen aus der Praxis erfolgen. So sollte es von Seiten der Forschung vermieden werden, Anwendungen an der Praxis vorbei zu entwickeln, wie es in dieser Arbeit bei RLT-Anlagen aufgedeckt wurde. Hier wurden nur 20 % der Fehlerbilder durch in realen Anlagen vorhandene Datenströme vollständig unterstützt. Für die Nutzung ist es erforderlich, dass mehr Datensätze aus Energiesystemen in Gebäuden und Quartieren international öffentlich verfügbar gemacht werden, so dass mehrere unterschiedliche regionale Besonderheiten abgedeckt werden können. Trotz hoher Datenschutzanforderungen können auch in der EU mehr Datensätze veröffentlicht werden. Die reale Veröffentlichung scheitert oftmals an nicht eingeräumter Zeit am Ende eines Projektes. Vom Fördergeber von Forschungsprojekten wäre es deshalb sinnvoll, die Förderung von Monitoringprojekten mit einer verpflichtenden Veröffentlichung der Daten zu verbinden.

Die Konzentration auf für Anwendungen erforderliche Metadaten sollte auch in Zukunft standardmäßig bei der Auswertung von Algorithmen der Metadatenextraktion erfolgen. Hierfür liefert die anwendungsorientierte Bewertung von Informationsqualitäten, die in dieser Arbeit entwickelt wurde, einen ersten Ansatz. Dieser Ansatz sollte mit weiteren Anwendungen und den erforderlichen Metadaten ergänzt werden.

Die hier dargestellten Anwendungen haben ein hohes Potential, den Energieeinsatz in Gebäuden und Quartieren zu senken. Aktuell stellt die Integration dieser Anwendungen in der Praxis insbesondere auf Grund der fehlenden und nicht standardisierten Metadaten eine große Hürde dar. Deshalb sollte in Zukunft untersucht werden, ob die entwickelten Ansätze tatsächlich auch die Anwendungen direkt unterstützen bzw. welche menschliche Ergänzungen erforderlich sind und wie der Personaleinsatz mithilfe geschickter Eingabe von Daten gesenkt werden kann.

Literaturverzeichnis

- [1] Measurement of Diversity. In: *Nature* 163 (1949), Nr. 4148, S. 688. <http://dx.doi.org/10.1038/163688a0>. – DOI 10.1038/163688a0. – ISSN 0028–0836
- [2] ABDEL HALEEM, Shadi M. ; PAVLAK, Gregory S. ; BAHNFLETH, William P.: Impact of control loop performance on energy use, air quality, and thermal comfort in building systems with advanced sequences of operation. In: *Automation in Construction* 130 (2021), S. 103837. <http://dx.doi.org/10.1016/j.autcon.2021.103837>. – DOI 10.1016/j.autcon.2021.103837. – ISSN 09265805
- [3] ABDELALIM, Aly ; O'BRIEN, William ; SHI, Zixiao: Development of Sankey diagrams to visualize real HVAC performance. In: *Energy and Buildings* 149 (2017), S. 282–297. <http://dx.doi.org/10.1016/j.enbuild.2017.05.040>. – DOI 10.1016/j.enbuild.2017.05.040. – ISSN 03787788
- [4] ABDELRAHMAN, Mahmoud M. ; ZHAN, Sicheng ; MILLER, Clayton ; CHONG, Adrian: Data science for building energy efficiency: A comprehensive text-mining driven review of scientific literature. In: *Energy and Buildings* (2021), S. 110885. <http://dx.doi.org/10.1016/j.enbuild.2021.110885>. – DOI 10.1016/j.enbuild.2021.110885. – ISSN 03787788
- [5] ABDOLHOSSEINI QOMI, Mohammad J. ; NOSHADRAVAN, Arash ; SOBSTYL, Jake M. ; TOOLE, Jameson ; FERREIRA, Joseph ; PELLENG, Roland J.-M. ; ULM, Franz-Josef ; GONZALEZ, Marta C.: Data analytics for simplifying thermal efficiency planning in cities. In: *Journal of the Royal Society Interface* 13 (2016), Nr. 117. <http://dx.doi.org/10.1098/rsif.2015.0971>. – DOI 10.1098/rsif.2015.0971. – ISSN 1742–5689
- [6] ABIOYE, Sofiat O. ; OYEDELE, Lukumon O. ; AKANBI, Lukman ; AJAYI, Anuoluwapo ; DAVILA DELGADO, Juan M. ; BILAL, Muhammad ; AKINADE, Olugbenga O. ; AHMED, Ashraf: Artificial intelligence in the construction industry: A review of present status, opportunities and future challenges. In: *Journal of Building Engineering* 44 (2021), S. 103299. <http://dx.doi.org/10.1016/j.jobe.2021.103299>. – DOI 10.1016/j.jobe.2021.103299. – ISSN 23527102
- [7] ABUIMARA, Tareq ; HOBSON, Brodie W. ; GUNAY, Burak ; O'BRIEN, William ; KANE, Michael: Current state and future challenges in building management: Practitioner interviews and a literature review. In: *Journal of Building Engineering* 41 (2021), S. 102803. <http://dx.doi.org/10.1016/j.jobe.2021.102803>. – DOI 10.1016/j.jobe.2021.102803. – ISSN 23527102
- [8] ADAMIC, Lada A. ; HUBERMAN, Bernardo A.: Zipf's law and the Internet. In: *Glottometrics* 3 (2002), Nr. 1, S. 143–150

- [9] AGUIRRE ARREDONDO, Luis ; CARRIZO MORENO, Dante: Information quality and quantity-based model to represent the appropriateness of software requirements elicitation techniques. In: *Revista Facultad de Ingeniería Universidad de Antioquia* (2017), Nr. 84, S. 72–83. <http://dx.doi.org/10.17533/udea.redin.n84a09>. – DOI 10.17533/udea.redin.n84a09. – ISSN 01206230
- [10] AHMAD, Ala'a Y. ; ASSAD, Nedal F. ; ALSHURIDEH, Muhammad T.: Running Head: Impact of Smart Buildings on Accounting and Management Control. Version: 2023. http://dx.doi.org/10.1007/978-3-031-12382-5_98. In: ALSHURIDEH, Muhammad (Hrsg.) ; AL KURDI, Barween H. (Hrsg.) ; MASA'DEH, Ra'ed (Hrsg.) ; ALZOUBI, Haitham M. (Hrsg.) ; SALLOUM, Said (Hrsg.): *The Effect of Information Technology on Business and Marketing Intelligence Systems* Bd. 1056. Cham : Springer International Publishing, 2023. – DOI 10.1007/978-3-031-12382-5_98. – ISBN 978-3-031-12381-8, S. 1795–1806
- [11] AHMADZADEH, Azim ; SINHA, Kankana ; AYDIN, Berkay ; ANGRYK, Rafal A.: MVTs-Data Toolkit: A Python package for preprocessing multivariate time series data. In: *SoftwareX* 12 (2020), S. 100518. <http://dx.doi.org/10.1016/j.softx.2020.100518>. – DOI 10.1016/j.softx.2020.100518. – ISSN 23527110
- [12] AHO, Alfred V. ; CORASICK, Margaret J.: Efficient string matching. In: *Communications of the ACM* 18 (1975), Nr. 6, S. 333–340. <http://dx.doi.org/10.1145/360825.360855>. – DOI 10.1145/360825.360855. – ISSN 00010782
- [13] AJIT, Arohan ; ACHARYA, Koustav ; SAMANTA, Abhishek: A Review of Convolutional Neural Networks. In: *2020 International Conference on Emerging Trends in Information Technology and Engineering (ic-ETITE)*, Ieee, 2020. – ISBN 978-1-7281-4142-8, S. 1–5
- [14] ALANNE, Kari: A novel performance indicator for the assessment of the learning ability of smart buildings. In: *Sustainable Cities and Society* 72 (2021), S. 103054. <http://dx.doi.org/10.1016/j.scs.2021.103054>. – DOI 10.1016/j.scs.2021.103054. – ISSN 22106707
- [15] ALANNE, Kari ; SIERLA, Seppo: An overview of machine learning applications for smart buildings. In: *Sustainable Cities and Society* 76 (2022), S. 103445. <http://dx.doi.org/10.1016/j.scs.2021.103445>. – DOI 10.1016/j.scs.2021.103445. – ISSN 22106707
- [16] ALGHANMI, Ashraf ; YUNUSA-KALTUNGO, Akilu ; EDWARDS, Rodger E.: Investigating the influence of maintenance strategies on building energy performance: A systematic literature review. In: *Energy Reports* 8 (2022), S. 14673–14698. <http://dx.doi.org/10.1016/j.egyr.2022.10.441>. – DOI 10.1016/j.egyr.2022.10.441. – ISSN 23524847
- [17] ALMASHOR, Mahathir ; RANA, Mashud ; MCCULLOCH, John ; RAHMAN, Ashfaque ; SETHUVENKATRAMAN, Subbu: What's The Point: AutoEncoding Building Point Names. In: *Proceedings of the 10th ACM International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation*. New York, NY, USA : ACM, 2023. – ISBN 9798400702303, S. 256–260

-
- [18] ALTMAN, Douglas G.: *Practical statistics for medical research*. Chapman and Hall/CRC, 1990. <http://dx.doi.org/10.1201/9780429258589>. <http://dx.doi.org/10.1201/9780429258589>. – ISBN 9780429258589
 - [19] AMINIKHANGHAHI, Samaneh ; COOK, Diane J.: A Survey of Methods for Time Series Change Point Detection. In: *Knowledge and information systems* 51 (2017), Nr. 2, S. 339–367. <http://dx.doi.org/10.1007/s10115-016-0987-z>. – DOI 10.1007/s10115-016-0987-z. – ISSN 0219-1377
 - [20] ANDRIAMAMONJY, Ando ; SAELENS, Dirk ; KLEIN, Ralf: An auto-deployed model-based fault detection and diagnosis approach for Air Handling Units using BIM and Modelica. In: *Automation in Construction* 96 (2018), S. 508–526. – ISSN 09265805
 - [21] ANTHROPIC: *Introducing Claude*. <https://www.anthropic.com/news/introducing-claude>. Version: 14.03.2023
 - [22] ARENDT, Krzysztof ; JOHANSEN, Aslak ; JØRGENSEN, Bo N. ; KJÆRGAARD, Mikkel B. ; MATTERA, Claudio G. ; SANGOGBOYE, Fisayo C. ; SCHWEE, Jens H. ; VEJE, Christian T.: Room-level occupant counts, airflow and CO 2 data from an office building. In: GAO, Jie (Hrsg.) ; ZHANG, Pei (Hrsg.) ; PAN, Shijia (Hrsg.) ; NI, Chien-Chun (Hrsg.): *Proceedings of the First Workshop on Data Acquisition To Analysis*. New York, NY, USA : ACM, 2018. – ISBN 9781450360494, S. 13–14
 - [23] ASKHAM, Nicola ; COOK, Denise ; DOYLE, Martin ; FEREDAY, Helen ; GIBSON, Mike ; LANDBECK, Ulrich ; LEE, Rob ; MAYNARD, Chris ; PALMER, Gary ; SCHWARZENBACH, Julian: The six primary dimensions for data quality assessment. In: *DAMA UK working group* (2013). <https://www.sbctc.edu/resources/documents/colleges-staff/commissions-councils/dgc/data-quality-deminsions.pdf>
 - [24] ÅSTRAND, Nico: *Utilization of Building Management System Design Documents as a Data Source for Automated Point Mapping*. Aalto, Aalto University, Master thesis, 2022. <https://aaltodoc.aalto.fi/handle/123456789/116347>
 - [25] AUER, Thomas ; LAUSS, Lukas ; HEISLER, Karl M. ; MADERSPACHER, Johannes ; REISS, Dirk ; MEHNERT, Jan ; RUMPE, Bernhard ; STÜBER, Sebastian ; HANNEN, Matthias ; PLESSER, Stefan ; PINKERNELL, Claas ; KRÖKER, Alex ; GENTEMANN, Roland: *Big Data in der Gebäudeautomation*
 - [26] BACHMANN, Max: *RapidFuzz*. <https://github.com/rapidfuzz/RapidFuzz>. Version: 02.09.2024
 - [27] BAE, Yeonjin ; BHATTACHARYA, Saptarshi ; CUI, Borui ; LEE, Seungjae ; LI, Yanfei ; ZHANG, Liang ; IM, Piljae ; ADETOLA, Veronica ; VRABIE, Draguna ; LEACH, Matt ; KURUGANTI, Teja: Sensor impacts on building and HVAC controls: A critical review for building energy performance. In: *Advances in Applied Energy* (2021), S. 100068. <http://dx.doi.org/10.1016/j.adapen.2021.100068>. – DOI 10.1016/j.adapen.2021.100068. – ISSN 26667924

- [28] BAGNALL, Anthony ; LARGE, James ; MIDDLEHURST, Matthew: *A tale of two toolkits, report the second: bake off redux. Chapter 1. dictionary based classifiers*. <http://arxiv.org/pdf/1911.12008v1>. Version: 27.11.2019
- [29] BAI, Linge ; BREEN, David: Calculating Center of Mass in an Unbounded 2D Environment. In: *Journal of Graphics Tools* 13 (2008), Nr. 4, S. 53–60. <http://dx.doi.org/10.1080/2151237X.2008.10129266>. – DOI 10.1080/2151237X.2008.10129266. – ISSN 1086–7651
- [30] BAI, Xue ; NUNEZ, Manuel ; KALAGNANAM, Jayant R.: Managing Data Quality Risk in Accounting Information Systems. In: *Information Systems Research* 23 (2012), Nr. 2, S. 453–473. <http://dx.doi.org/10.1287/isre.1110.0371>. – DOI 10.1287/isre.1110.0371. – ISSN 1047–7047
- [31] BALAJI, Bharathan: Towards Large-scale Monitoring of Commercial Buildings using Wireless Sensor Networks, 2012
- [32] BALAJI, Bharathan ; AGARWAL, Yuvraj ; BERGES, Mario ; CULLER, David ; GUPTA, Rajesh ; KJÆRGAARD, Mikkel B. ; SRIVASTAVA, Mani ; WHITEHOUSE, Kamin ; BHATTACHARYA, Arka ; FIERRO, Gabriel ; GAO, Jingkun ; GLUCK, Joshua ; HONG, Dezhi ; JOHANSEN, Aslak ; KOH, Jason ; PLOENNIGS, Joern: Brick: Towards a Unified Metadata Schema For Buildings. In: *BuildSys '16*, 2016, S. 41–50
- [33] BALAJI, Bharathan ; BHATTACHARYA, Arka ; FIERRO, Gabriel ; GAO, Jingkun ; GLUCK, Joshua ; HONG, Dezhi ; JOHANSEN, Aslak ; KOH, Jason ; PLOENNIGS, Joern ; AGARWAL, Yuvraj ; BERGÉS, Mario ; CULLER, David ; GUPTA, Rajesh K. ; KJÆRGAARD, Mikkel B. ; SRIVASTAVA, Mani ; WHITEHOUSE, Kamin: Brick : Metadata schema for portable smart building applications. In: *Applied Energy* 226 (2018), S. 1273–1292. <http://dx.doi.org/10.1016/j.apenergy.2018.02.091>. – DOI 10.1016/j.apenergy.2018.02.091. – ISSN 03062619
- [34] BALAJI, Bharathan ; VERMA, Chetan ; NARAYANASWAMY, Balakrishnan ; AGARWAL, Yuvraj: Zodiac: Organizing Large Deployment of Sensors to Create Reusable Applications for Buildings. In: *Proceedings of the 2nd ACM International Conference on Embedded Systems for Energy-Efficient Built Environments*. 2821674 : ACM, 2015, S. 13–22
- [35] BARANDAS, Marília ; FOLGADO, Duarte ; FERNANDES, Letícia ; SANTOS, Sara ; ABREU, Mariana ; BOTA, Patrícia ; LIU, Hui ; SCHULTZ, Tanja ; GAMBOA, Hugo: TSFEL: Time Series Feature Extraction Library. In: *SoftwareX* 11 (2020), S. 100456. <http://dx.doi.org/10.1016/j.softx.2020.100456>. – DOI 10.1016/j.softx.2020.100456. – ISSN 23527110
- [36] BARJA-MARTINEZ, Sara ; ARAGÜÉS-PEÑALBA, Mònica ; MUNNÉ-COLLADO, Íngrid ; LLORET-GALLEGÓ, Pau ; BULLICH-MASSAGUÉ, Eduard ; VILLAFILA-ROBLES, Roberto: Artificial intelligence techniques for enabling Big Data services in distribution networks: A review. In: *Renewable and Sustainable Energy Reviews* 150 (2021), S. 111459. <http://dx.doi.org/10.1016/j.rser.2021.111459>. – DOI 10.1016/j.rser.2021.111459. – ISSN 13640321

-
- [37] BATINI, Carlo ; SCANNAPIECO, Monica: *Data and Information Quality*. Cham : Springer International Publishing, 2016. <http://dx.doi.org/10.1007/978-3-319-24106-7>. <http://dx.doi.org/10.1007/978-3-319-24106-7>. – ISBN 978-3-319-24104-3
- [38] BECKER, Thealexa: *Discussing Data: Evaluating Data Quality*. <https://www.kansascityfed.org/Technical%20Briefings/documents/4986/tb-Discussing%20Data:%20Evaluating%20Data%20Quality.pdf>. Version: 2019
- [39] BENFER, Rebekka ; MÜLLER, Jochen: Semantic digital twin creation of building systems through time series based metadata inference – A review. In: *Energy and Buildings* 321 (2024), S. 114637. <http://dx.doi.org/10.1016/j.enbuild.2024.114637>. – DOI 10.1016/j.enbuild.2024.114637. – ISSN 03787788
- [40] BERGMANN, Harry ; MOSIMAN, Cory ; SAHA, Avijit ; HAILE, Selam ; LIVINGWOOD, William ; BUSHBY, Steve ; FIERRO, Gabe ; BENDER, Joel ; POPLAWSKI, Michael ; GRANDERSON, Jessica ; PRITONI, Marco: *Semantic Interoperability to Enable Smart, Grid-Interactive Efficient Buildings*
- [41] BERGROTH, L. ; HAKONEN, H. ; RAITA, T.: A survey of longest common subsequence algorithms. In: *Proceedings / Seventh International Symposium on String Processing and Information Retrieval, SPIRE 2000*. Los Alamitos, Calif. : IEEE Computer Society, 2000. – ISBN 0-7695-0746-8, S. 39–48
- [42] BERKOBEN, Keith ; KAED, Charbel ; SODORFF, Trevor: *A Digital Buildings Ontology for Google's Real Estate*. <https://ceur-ws.org/Vol-2721/paper510.pdf>. Version: 2020
- [43] BEUTH VERLAG: *Nautos*. <https://nautos.de>. Version: 2023
- [44] BEUTH VERLAG: *Perinorm*. <https://www.perinorm.com/>. Version: 2023
- [45] BHATTACHARYA, Arka ; CULLER, David E. ; ORTIZ, Jorge ; HONG, Dezhi ; WHITEHOUSE, Kamin: *Enabling Portable Building Applications through Automated Metadata Transformation*. <http://www2.eecs.berkeley.edu/Pubs/TechRpts/2014/EECS-2014-159.html>. Version: 2014
- [46] BHATTACHARYA, Arka ; PLOENNIGS, Joern ; CULLER, David: Short Paper: Analyzing Metadata Schemas for Buildings: The Good, the Bad, and the Ugly. In: *Proceedings of the 2nd ACM International Conference on Embedded Systems for Energy-Efficient Built Environments*. 2821674 : ACM, 2015, S. 33–34
- [47] BHATTACHARYA, Arka A. ; HONG, Dezhi ; CULLER, David ; ORTIZ, Jorge ; WHITEHOUSE, Kamin ; WU, Eugene: Automated Metadata Construction to Support Portable Building Applications. In: *Proceedings of the 2nd ACM International Conference on Embedded Systems for Energy-Efficient Built Environments*. 2821674 : ACM, 2015, S. 3–12
- [48] BLASCH, Erik: One decade of the Data Fusion Information Group (DFIG) model. In: BROOME,

- Barbara D. (Hrsg.) ; HANRATTY, Timothy P. (Hrsg.) ; HALL, David L. (Hrsg.) ; LLINAS, James (Hrsg.): *Next-Generation Analyst III*, SPIE, 2015 (SPIE Proceedings), S. 94990L
- [49] BLASCH, Erik ; STEINBERG, Alan ; DAS, Subrata ; LLINAS, James ; CHONG, Chee ; KESSLER, Otto ; WALTZ, Ed ; WHITE, Frank: Revisiting the JDL model for information exploitation. In: *Proceedings of the 16th International Conference on Information Fusion*, 2013, S. 129–136
- [50] BLASCH, Erik ; STEINBERG, Alan ; DAS, Subrata ; LLINAS, James ; CHONG, Chee ; KESSLER, Otto ; WALTZ, Ed ; WHITE, Frank: Revisiting the JDL model for information exploitation. In: *Proceedings of the 16th International Conference on Information Fusion*, 2013, S. 129–136
- [51] BLASCH, Erik ; VALIN, Pierre ; BOSSE, Eloi ; NILSSON, Maria ; LAERE, Joeri van ; SHAHBAZIAN, Elisa: Implication of culture: User roles in information Fusion for enhanced situational understanding. In: *2009 12th International Conference on Information Fusion*, 2009, S. 1272–1279
- [52] BLASCH, Erik P. ; BRETON, Richard ; VALIN, Pierre: Information fusion measures of effectiveness (MOE) for decision support. In: KADAR, Ivan (Hrsg.): *Signal Processing, Sensor Fusion, and Target Recognition XX*, SPIE, 2011 (SPIE Proceedings), S. 805011
- [53] BLUM, D. ; SPEARS, M. ; PAGE, J. ; GRANDERSON, J.: *When Data Analytics Meet Site Operation: Benefits and Challenges*
- [54] BOGERT, Judith: In Defense of the Fog Index. In: *The Bulletin of the Association for Business Communication* 48 (1985), Nr. 2, S. 9–12. <http://dx.doi.org/10.1177/108056998504800203>. – DOI 10.1177/108056998504800203. – ISSN 8756–1972
- [55] BOLTZMANN, Ludwig: Weitere Studien über das Wärmegleichgewicht unter Gasmolekülen. *Sitzungsberichte der Akademie der Wissenschaften* 66 (1872), 275–370. Translation: Further studies on the thermal equilibrium of gas molecules. In: *Kinetic Theory* 2 (1872), Nr. 88, S. 275–370
- [56] BOTH, Maximilian ; KÄMPER, Björn ; CARTUS, Alina ; BEERMANN, Jo ; FESSLER, Thomas ; MÜLLER, Jochen ; DIEDRICH, Christian: Automated monitoring applications for existing buildings through natural language processing based semantic mapping of operational data and creation of digital twins. In: *Energy and Buildings* 300 (2023), S. 113635. <http://dx.doi.org/10.1016/j.enbuild.2023.113635>. – DOI 10.1016/j.enbuild.2023.113635. – ISSN 03787788
- [57] BOVET, Gerome ; HENNEBERT, Jean: Distributed Semantic Discovery for Web-of-Things Enabled Smart Buildings. In: BADRA, Mohamad (Hrsg.): *2014 6th International Conference on New Technologies, Mobility and Security (NTMS)*. Piscataway, NJ : Ieee, 2014. – ISBN 978–1–4799–3223–8, S. 1–5
- [58] BOYD, S.: Multitone signals with low crest factor. In: *IEEE Transactions on Circuits and Systems* 33 (1986), Nr. 10, S. 1018–1022. <http://dx.doi.org/10.1109/TCS.1986.1085837>. – DOI 10.1109/TCS.1986.1085837. – ISSN 0098–4094

- [59] BOYD, Sarah ; HORVATH, Arpad ; DORNFELD, David: Life-Cycle Assessment of NAND Flash Memory. In: *IEEE Transactions on Semiconductor Manufacturing* 24 (2011), Nr. 1, S. 117–124. <http://dx.doi.org/10.1109/TSM.2010.2087395>. – DOI 10.1109/TSM.2010.2087395. – ISSN 0894–6507
- [60] BOYD, Stephen P. ; VANDENBERGHE, Lieven: *Convex optimization*. Version 29. Cambridge and New York and Melbourne and New Delhi and Singapore : Cambridge University Press, April 2023. <http://dx.doi.org/10.1017/CB09780511804441>. <http://dx.doi.org/10.1017/CB09780511804441>. – ISBN 9780521833783
- [61] BREIMAN, Leo: Random Forests. In: *Machine Learning* 45 (2001), Nr. 1, S. 5–32. <http://dx.doi.org/10.1023/A:1010933404324>. – DOI 10.1023/A:1010933404324. – ISSN 1573–0565
- [62] BREIMAN, Leo: Statistical Modeling: The Two Cultures (with comments and a rejoinder by the author). In: *Statistical Science* 16 (2001), Nr. 3. <http://dx.doi.org/10.1214/ss/1009213726>. – DOI 10.1214/ss/1009213726. – ISSN 0883–4237
- [63] <https://brickschema.org/consortium/>
- [64] BROEDER, Peter ; EXTRA, G. ; VAN HOUT, R.: Richness and variety in the developing lexicon. In: C. PERDUE (Hrsg.): *Adult language acquisition: Cross-linguistic perspectives. Vol. I: Field methods*. United Kingdom : Cambridge University Press, 1993. – ISBN 0521417082, S. 145–163
- [65] BURAK GUNAY, H. ; SHEN, Weiming ; NEWSHAM, Guy: Data analytics to improve building performance: A critical review. In: *Automation in Construction* 97 (2019), S. 96–109. <http://dx.doi.org/10.1016/j.autcon.2018.10.020>. – DOI 10.1016/j.autcon.2018.10.020. – ISSN 09265805
- [66] BUTLER, James F. ; VEELANTURF, Robert: Point Naming Standards. Version:2010. http://www.bacnet.org/Bibliography/BACnet-Today-10/Butler_2010.pdf. In: AMERICAN SOCIETY OF HEATING, REFRIGERATING AND AIR-CONDITIONING ENGINEERS, INC. (Hrsg.): *ASHRAE Journal, November 2010*. 2010
- [67] CANDANEDO, José A. ; VALLIANOS, Charalampos ; DELCROIX, Benoit ; DATE, Jennifer ; SABERI DERAKHTENJANI, Ali ; MOROVAT, Navid ; JOHN, Camille ; ATHIENITIS, Andreas K.: Control-oriented archetypes: a pathway for the systematic application of advanced controls in buildings. In: *Journal of Building Performance Simulation* 15 (2022), Nr. 4, S. 433–444. <http://dx.doi.org/10.1080/19401493.2022.2063947>. – DOI 10.1080/19401493.2022.2063947. – ISSN 1940–1493
- [68] CAO, Yu ; KAMARUZZAMAN, Syahrul N. ; AZIZ, Nur M.: Building Information Modeling (BIM) Capabilities in the Operation and Maintenance Phase of Green Buildings: A Systematic Review. In: *Buildings* 12 (2022), Nr. 6, S. 830. <http://dx.doi.org/10.3390/buildings12060830>. – DOI 10.3390/buildings12060830

- [69] CARLSSON, Oscar ; HEGEDUS, Csaba ; DELSING, Jerker ; VARGA, Pal: Organizing IoT Systems-of-Systems from standardized engineering data. In: *Proceedings of the IECON2016 - 42nd Annual Conference of the IEEE Industrial Electronics Society : Florence (Italy), October 24-27, 2016*. Piscataway, NJ : Ieee, 2016. – ISBN 978–1–5090–3474–1, S. 5277–5282
- [70] CARSTENS, Herman ; XIA, Xiaohua ; YADAVALLI, Sarma: Measurement uncertainty in energy monitoring: Present state of the art. In: *Renewable and Sustainable Energy Reviews* 82 (2018), S. 2791–2805. <http://dx.doi.org/10.1016/j.rser.2017.10.006>. – DOI 10.1016/j.rser.2017.10.006. – ISSN 13640321
- [71] CASH, Michael ; WANG, Shan ; PEARSON, Bryan ; ZHOU, Qun ; FU, Xinwen: On Automating BACnet Device Discovery and Property Identification. In: *ICC 2021 - IEEE International Conference on Communications*, Ieee, 2021. – ISBN 978–1–7281–7122–7, S. 1–6
- [72] CASILLAS, Armando ; LIN, Guanqing ; GRANDERSON, Jessica ; HUANG, Sen ; CHEN, Yan: Development of a Simulated Air Handling Unit Fault Dataset for FDD Tools: Lessons Learned and Considerations for FDD Developers. (2022). <http://dx.doi.org/10.20357/B7K89D>. – DOI 10.20357/B7K89D
- [73] CASTANEDO, Federico: A review of data fusion techniques. In: *TheScientificWorld-Journal* 2013 (2013), S. 704504. <http://dx.doi.org/10.1155/2013/704504>. – DOI 10.1155/2013/704504
- [74] CATBOOST: *CatBoost*. <https://github.com/catboost/catboost>. Version: 2024
- [75] CAVRARO, Guido ; ARGHANDEH, Reza: Power Distribution Network Topology Detection With Time-Series Signature Verification Method. In: *IEEE Transactions on Power Systems* 33 (2018), Nr. 4, S. 3500–3509. <http://dx.doi.org/10.1109/TPWRS.2017.2779129>. – DOI 10.1109/TPWRS.2017.2779129. – ISSN 0885–8950
- [76] CEDERMAN, Lars-Erik: Modeling the Self-Organized Criticality of War Size. (2001)
- [77] CHAREF, Rabia ; EMMITT, Stephen ; ALAKA, Hafiz ; FOUCHAL, Farid: Building Information Modelling adoption in the European Union: An overview. In: *Journal of Building Engineering* 25 (2019), 100777. <http://dx.doi.org/10.1016/j.jobbe.2019.100777>. – DOI 10.1016/j.jobbe.2019.100777. – ISSN 23527102
- [78] CHEN, Long ; GUNAY, H. B. ; SHI, Zixiao ; SHEN, Weiming ; LI, Xiaoping: A Metadata Inference Method for Building Automation Systems With Limited Semantic Information. In: *IEEE Transactions on Automation Science and Engineering* (2020), S. 1–13. <http://dx.doi.org/10.1109/TASE.2020.2990566>. – DOI 10.1109/TASE.2020.2990566. – ISSN 1545–5955
- [79] CHEN, Tianqi ; GUESTRIN, Carlos: XGBoost: A Scalable Tree Boosting System. In: KRISHNAPURAM, Balaji (Hrsg.) ; SHAH, Mohak (Hrsg.) ; SMOLA, Alex (Hrsg.) ; AGGARWAL, Charu (Hrsg.) ; SHEN, Dou (Hrsg.) ; RASTOGI, Rajeev (Hrsg.): *KDD2016*. New York, NY : Association for Computing Machinery Inc. (ACM), 2016. – ISBN 9781450342322, S. 785–794

-
- [80] CHEN, Yimin ; CROWE, Eliot ; LIN, Guanqing ; GRANDERSON, Jessica: What's in a Name? Developing a Standardized Taxonomy for HVAC System Faults. (2020). <http://dx.doi.org/10.20357/B7W02S>. – DOI 10.20357/B7W02S
 - [81] CHEN, Zhe ; XU, Peng ; FENG, Fan ; QIAO, Yifan ; LUO, Wei: Data mining algorithm and framework for identifying HVAC control strategies in large commercial buildings. In: *Building Simulation* 14 (2021), Nr. 1, S. 63–74. <http://dx.doi.org/10.1007/s12273-019-0599-0>. – DOI 10.1007/s12273-019-0599-0. – ISSN 1996–3599
 - [82] CHENG, Na ; YU, Zhongqing ; WANG, Kaixi: String similarity computing based on position and cosine. In: WENZHENG, Li (Hrsg.) ; U, Seng-Pan (Hrsg.) ; HONGDAN, Zhu (Hrsg.) ; SHAOWEN, Ni (Hrsg.): *Proceedings of 2017 IEEE 7th International Conference on Electronics Information and Emergency Communication*. Piscataway, NJ : Ieee, 2017. – ISBN 978–1–5090–3025–5, S. 256–261
 - [83] CHICCO, Davide ; JURMAN, Giuseppe: The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. In: *BMC genomics* 21 (2020), Nr. 1, S. 6. <http://dx.doi.org/10.1186/s12864-019-6413-7>. – DOI 10.1186/s12864-019-6413-7
 - [84] CHITTURI, Bhadrachalam ; SUDBOROUGH, I. H.: Bounding prefix transposition distance for strings and permutations. In: *Theoretical Computer Science* 421 (2012), S. 15–24. <http://dx.doi.org/10.1016/j.tcs.2011.11.018>. – DOI 10.1016/j.tcs.2011.11.018. – ISSN 03043975
 - [85] CHO, Sung-Hwan ; HONG, Young-Ju ; KIM, Won-Tae ; ZAHEER-UDDIN, Mohammed: Multi-fault detection and diagnosis of HVAC systems: an experimental study. In: *International journal of energy research* 29 (2005), Nr. 6, S. 471–483
 - [86] CHO, Sung-Hwan ; YANG, Hoon-Cheol ; ZAHEER-UDDIN, Mohammed ; AHN, Byung-Cheon: Transient pattern analysis for fault detection and diagnosis of HVAC systems. In: *Energy Conversion and Management* 46 (2005), Nr. 18-19, S. 3103–3116. – ISSN 01968904
 - [87] CHOLEWA, Tomasz ; SIUTA-OLCHA, Alicja ; SMOLARZ, Andrzej ; MURYJAS, Piotr ; WOLSZCZAK, Piotr ; ANASIEWICZ, Rafał ; BALARAS, Constantinos A.: A simple building energy model in form of an equivalent outdoor temperature. In: *Energy and Buildings* 236 (2021), S. 110766. <http://dx.doi.org/10.1016/j.enbuild.2021.110766>. – DOI 10.1016/j.enbuild.2021.110766. – ISSN 03787788
 - [88] CHRIST, Maximilian ; BRAUN, Nils ; NEUFFER, Julius ; KEMPA-LIEHR, Andreas W.: Time Series Feature Extraction on basis of Scalable Hypothesis tests (tsfresh – A Python package). In: *Neurocomputing* 307 (2018), S. 72–77. <http://dx.doi.org/10.1016/j.neucom.2018.03.067>. – DOI 10.1016/j.neucom.2018.03.067. – ISSN 09252312
 - [89] CLÉMENÇON, Raymond: The Two Sides of the Paris Climate Agreement. In: *The Journal of Environment & Development* 25 (2016), Nr. 1, S. 3–24. <http://dx.doi.org/10.1177/1070496516631362>. – DOI 10.1177/1070496516631362. – ISSN 1070–4965

- [90] CLEMENTS, K. A. ; DAVIS, P. W.: Detection and identification of topology errors in electric power systems. In: *IEEE Transactions on Power Systems* 3 (1988), Nr. 4, S. 1748–1753. <http://dx.doi.org/10.1109/59.192991>. – DOI 10.1109/59.192991. – ISSN 0885–8950
- [91] COHEN, Jacob: A Coefficient of Agreement for Nominal Scales. In: *Educational and Psychological Measurement* 20 (1960), Nr. 1, S. 37–46. <http://dx.doi.org/10.1177/001316446002000104>. – DOI 10.1177/001316446002000104. – ISSN 0013–1644
- [92] COHEN, William W. ; RAVIKUMAR, Pradeep ; FIENBERG, Stephen E.: A Comparison of String Distance Metrics for Name-Matching Tasks. In: *Proceedings of the 2003 International Conference on Information Integration on the Web*, AAAI Press, 2003 (IIWEB'03), S. 73–78
- [93] COLEMAN, Shirley Y. ; KENETT, Ron S.: The information quality framework for evaluating data science programs. In: *Encyclopedia with Semantic Computing and Robotic Intelligence* 02 (2018), Nr. 02, S. 1730001. <http://dx.doi.org/10.1142/S2529737617300017>. – DOI 10.1142/S2529737617300017. – ISSN 2529–7376
- [94] COSTA, Andrea ; KEANE, Marcus M. ; TORRENS, J. I. ; CORRY, Edward: Building operation and energy performance: Monitoring, analysis and optimisation toolkit. In: *Applied Energy* 101 (2013), S. 310–316. <http://dx.doi.org/10.1016/j.apenergy.2011.10.037>. – DOI 10.1016/j.apenergy.2011.10.037. – ISSN 03062619
- [95] COSTA, I. S. ; LEO, J. A.: Identification of topology errors in power system state estimation. In: *IEEE Transactions on Power Systems* 8 (1993), Nr. 4, S. 1531–1538. <http://dx.doi.org/10.1109/59.260956>. – DOI 10.1109/59.260956. – ISSN 0885–8950
- [96] CROWE, Eliot ; CHEN, Yimin ; GRANDERSON, Jessica ; REEVE, Hayden ; TROUP, Lucas ; YUILL, David ; CHEN, Yuxuan: What We Learned From Analyzing 18 Million Rows of Commercial Buildings HVAC Fault Data. (2022)
- [97] CUESTA-FRAU, David: Slope entropy: A new time series complexity estimator based on both symbolic patterns and amplitude information. In: *Entropy* 21 (2019), Nr. 12, S. 1167
- [98] CUNHA, Carlos ; BESTAVROS, Azer ; CROVELLA, Mark: Characteristics of WWW Client-Based Traces. (1995)
- [99] CUTLER, Dylan ; FRANK, Stephen ; SLOVENSKY, Michelle ; SHEPPY, Michael ; PETERSEN, Anya: Creating an Energy Intelligent Campus: Data Integration Challenges and Solutions at a Large Research Campus. (2016). <https://www.osti.gov/biblio/1324228>
- [100] DA SILVA, Lucian R. ; FLESCH, Rodolfo César C. ; NORMEY-RICO, Julio E.: Controlling industrial dead-time systems: When to use a PID or an advanced controller. In: *ISA transactions* 99 (2020), S. 339–350. <http://dx.doi.org/10.1016/j.isatra.2019.09.008>. – DOI 10.1016/j.isatra.2019.09.008
- [101] DAINTY, Andrew ; LEIRINGER, Roine ; FERNIE, Scott ; HARTY, Chris: BIM and the small construction firm: a critical perspective. In: *Building Research & Information* 45 (2017),

- Nr. 6, S. 696–709. <http://dx.doi.org/10.1080/09613218.2017.1293940>. – DOI 10.1080/09613218.2017.1293940. – ISSN 0961–3218
- [102] DAMERAU, Fred J.: A technique for computer detection and correction of spelling errors. In: *Communications of the ACM* 7 (1964), Nr. 3, S. 171–176. <http://dx.doi.org/10.1145/363958.363994>. – DOI 10.1145/363958.363994. – ISSN 00010782
- [103] DANIELE, Laura ; DEN HARTOG, Frank ; ROES, Jasper: Created in Close Interaction with the Industry: The Smart Appliances REference (SAREF) Ontology. Version: 2015. http://dx.doi.org/10.1007/978-3-319-21545-7_9. In: CUEL, Roberta (Hrsg.) ; YOUNG, Robert (Hrsg.): *Formal Ontologies Meet Industry* Bd. 225. Cham : Springer, 2015. – DOI 10.1007/978-3-319-21545-7_9. – ISBN 978-3-319-21544-0, S. 100–112
- [104] DAU, Hoang A. ; BAGNALL, Anthony ; KAMGAR, Kaveh ; YEH, Chin-Chia M. ; ZHU, Yan ; GHARGHABI, Shaghayegh ; RATANAMAHATANA, Chotirat A. ; KEOGH, Eamonn: The UCR Time Series Archive. (2018). <http://arxiv.org/pdf/1810.07758v2>
- [105] DEMPSTER, Angus ; PETITJEAN, François ; WEBB, Geoffrey I.: ROCKET: Exceptionally fast and accurate time series classification using random convolutional kernels. (2019). <https://arxiv.org/pdf/1910.13051>
- [106] DEMPSTER, Angus ; SCHMIDT, Daniel F. ; WEBB, Geoffrey I.: MINIROCKET: A Very Fast (Almost) Deterministic Transform for Time Series Classification. (2020). <http://arxiv.org/pdf/2012.08791v1>
- [107] DEVLIN, Jacob ; CHANG, Ming-Wei ; LEE, Kenton ; TOUTANOVA, Kristina: *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. <http://arxiv.org/pdf/1810.04805>. Version: 2018
- [108] DHAR, A. ; REDDY, T. A. ; CLARIDGE, D. E.: A Fourier Series Model to Predict Hourly Heating and Cooling Energy Use in Commercial Buildings With Outdoor Temperature as the Only Weather Variable. In: *Journal of Solar Energy Engineering* 121 (1999), Nr. 1, S. 47–53. <http://dx.doi.org/10.1115/1.2888142>. – DOI 10.1115/1.2888142. – ISSN 0199–6231
- [109] DIBOWSKI, Henrik ; MASSA GRAY, Francesco: Applying Knowledge Graphs as Integrated Semantic Information Model for the Computerized Engineering of Building Automation Systems. Version: 2020. http://dx.doi.org/10.1007/978-3-030-49461-2_36. In: HARTH, Andreas (Hrsg.) ; KIRrane, Sabrina (Hrsg.) ; NGONGA NGOMO, Axel-Cyrille (Hrsg.) ; PAULHEIM, Heiko (Hrsg.) ; RULA, Anisa (Hrsg.) ; GENTILE, Anna L. (Hrsg.) ; HAASE, Peter (Hrsg.) ; COCHEZ, Michael (Hrsg.): *The Semantic Web* Bd. 12123. Cham : Springer International Publishing, 2020. – DOI 10.1007/978-3-030-49461-2_36. – ISBN 978-3-030-49460-5, S. 616–631
- [110] DIN DEUTSCHES INSTITUT FÜR NORMUNG E. V.: *DIN EN ISO 16484-2:2004 - Building automation and control systems (BACS) - Part 2: Hardware (ISO 16484-2:2004)*. 2004

- [111] DIN DEUTSCHES INSTITUT FÜR NORMUNG E. V.: *DIN EN 16798-3 - Energetische Bewertung von Gebäuden - Lüftung von Gebäuden - Teil 3: Lüftung von Nichtwohngebäuden - Leistungsanforderungen an Lüftungs- und Klimaanlageanlagen und Raumkühlsysteme (Module M5-1, M5-4); Deutsche Fassung EN 16798-3:2017*. 2017
- [112] DIN DEUTSCHES INSTITUT FÜR NORMUNG E. V.: *DIN EN ISO 16739:2013 - Industry Foundation Classes (IFC) für den Datenaustausch in der Bauindustrie und im Anlagenmanagement*. 2017
- [113] DOROGUSH, Anna V. ; ERSHOV, Vasily ; GULIN, Andrey: *CatBoost: gradient boosting with categorical features support*. <https://arxiv.org/pdf/1810.11363>. Version: 2018
- [114] DRGOÑA, Ján ; ARROYO, Javier ; CUPEIRO FIGUEROA, Iago ; BLUM, David ; ARENDT, Krzysztof ; KIM, Donghun ; OLLÉ, Enric P. ; ORAVEC, Juraj ; WETTER, Michael ; VRABIE, Draguna L. ; HELSEN, Lieve: All you need to know about model predictive control for buildings. In: *Annual Reviews in Control* 50 (2020), S. 190–232. <http://dx.doi.org/10.1016/j.arcontrol.2020.09.001>. – DOI 10.1016/j.arcontrol.2020.09.001. – ISSN 13675788
- [115] DU, Zhimin ; FAN, Bo ; CHI, Jinlei ; JIN, Xinqiao: Sensor fault detection and its efficiency analysis in air handling unit using the combined neural networks. In: *Energy and Buildings* 72 (2014), S. 157–166. – ISSN 03787788
- [116] DU, Zhimin ; FAN, Bo ; JIN, Xinqiao ; CHI, Jinlei: Fault detection and diagnosis for buildings and HVAC systems using combined neural networks and subtractive clustering analysis. In: *Building and Environment* 73 (2014), S. 1–11. – ISSN 0360–1323
- [117] DU, Zhimin ; JIN, Xinqiao ; YANG, Yunyu: Fault diagnosis for temperature, flow rate and pressure sensors in VAV systems using wavelet neural network. In: *Applied Energy* 86 (2009), Nr. 9, S. 1624–1631. – ISSN 03062619
- [118] DUAN, Ruo-Nan ; ZHU, Jia-Yi ; LU, Bao-Liang: Differential entropy feature for EEG-based emotion classification. In: *2013 6th International IEEE/EMBS Conference on Neural Engineering (NER)*, Ieee, 2013. – ISBN 978–1–4673–1969–0, S. 81–84
- [119] DUBOIS, Didier ; PRADE, Henri: Possibility theory in information fusion. In: *Fusion 2000*. Sunnyvale, Calif. : ISIF, 2000. – ISBN 2–7257–0000–0, S. PS6–P19 vol.1
- [120] DWDS ; BERLIN-BRANDENBURGISCHE AKADEMIE DER WISSENSCHAFTEN (Hrsg.): *DWDS – Digitales Wörterbuch der deutschen Sprache. Das Wortauskunftssystem zur deutschen Sprache in Geschichte und Gegenwart*. <https://www.dwds.de>. Version: 2023
- [121] EBRAHIMIFAKHAR, Amir ; YUILL, David ; SMITH, Amanda ; GRANDERSON, Jessica ; CROWE, Eliot ; CHEN, Yimin ; REEVE, Hayden: Analysis of Automated Fault Detection and Diagnostics Records as an Indicator of HVAC Fault Prevalence: Methodology and Preliminary Results. In: *International High Performance Buildings Conference 2021* (2021)

- [122] EINSTEIN, A.: Die Grundlage der allgemeinen Relativitätstheorie. In: *Annalen der Physik* 354 (1916), Nr. 7, S. 769–822. <http://dx.doi.org/10.1002/andp.19163540702>. – DOI 10.1002/andp.19163540702. – ISSN 0003–3804
- [123] ELNOUR, Mariam ; MESKIN, Nader ; AL-NAEMI, Mohammed: Sensor data validation and fault diagnosis using Auto-Associative Neural Network for HVAC systems. In: *Journal of Building Engineering* 27 (2020), S. 100935. – ISSN 23527102
- [124] EPPLER, Martin J.: 11. Information quality and information overload: The promises and perils of the information age. Version:2015. <http://dx.doi.org/10.1515/9783110271355-013>. In: CANTONI, Lorenzo (Hrsg.): *Communication and technology*. Berlin : de Gruyter Mouton, 2015 (Handbooks of communication science / ed. by Peter J. Schulz). – DOI 10.1515/9783110271355-013. – ISBN 9783110266535, S. 215–232
- [125] ERIKOSKARWALLIN ; HAMMAR, Karl ; PERKARLBERG ; LEIFSUNDBOM ; WOTIFI: RealEstateCore/rec: V3.0 – Usability and maintainability refactoring. (2019). <http://dx.doi.org/10.5281/ZENODO.2628367>. – DOI 10.5281/ZENODO.2628367
- [126] ESTELLER, Rosana ; ECHAUZ, Javier ; TCHENG, T. ; LITT, Brian ; PLESS, Benjamin: Line length: an efficient feature for seizure onset detection. In: *2001 Conference Proceedings of the 23rd Annual International Conference of the IEEE Engineering in Medicine and Biology Society* Bd. 2, 2001, S. 1707–1710
- [127] EUROPEAN STATISTICAL SYSTEM COMMITTEE: Quality Assurance Framework of the European Statistical System. (2019). <https://ec.europa.eu/eurostat/documents/64157/4392716/ESS-QAF-V1-2final.pdf/bbf5970c-1adf-46c8-afc3-58ce177a0646>
- [128] EVENSCHOR, Vincent: *Classification of time series data of building automation systems based on machine learning*. Aachen, RWTH Aachen University, Bachelorarbeit, 2019
- [129] FAN, Bo ; DU, Zhimin ; JIN, Xinqiao ; YANG, Xuebin ; GUO, Yibo: A hybrid FDD strategy for local system of AHU based on artificial neural network and wavelet analysis. In: *Building and Environment* 45 (2010), Nr. 12, S. 2698–2708. – ISSN 0360–1323
- [130] FAROUQ, Shiraz ; BYTTNER, Stefan ; BOUGUELIA, Mohamed-Rafik ; NORD, Natasa ; GADD, Henrik: Large-scale monitoring of operationally diverse district heating substations: A reference-group based approach. In: *Engineering Applications of Artificial Intelligence* 90 (2020), S. 103492. <http://dx.doi.org/10.1016/j.engappai.2020.103492>. – DOI 10.1016/j.engappai.2020.103492. – ISSN 09521976
- [131] FARRELL, Christopher-John: Identifying mislabelled samples: Machine learning models exceed human performance. In: *Annals of clinical biochemistry* 58 (2021), Nr. 6, S. 650–652. <http://dx.doi.org/10.1177/00045632211032991>. – DOI 10.1177/00045632211032991
- [132] FATH, Brian D. ; CABEZAS, Heriberto ; PAWLOWSKI, Christopher W.: Regime changes in ecological systems: an information theory approach. In: *Journal of theoretical biology* 222 (2003), Nr. 4, S. 517–530

- [133] FAWAZ, Hassan I. ; FORESTIER, Germain ; WEBER, Jonathan ; IDOUMGHAR, Lhassane ; MULLER, Pierre-Alain: Deep learning for time series classification: a review. In: *CoRR* abs/1809.04356 (2018)
- [134] FIERRO, Gabe: *Self-Adapting Software for Cyberphysical Systems*, EECS Department, University of California, Berkeley, Diss., 2021. <http://www2.eecs.berkeley.edu/Pubs/TechRpts/2021/EECS-2021-159.html>
- [135] FIERRO, Gabe: *Mortar Data*. <https://mortardata.org/>. Version:2022
- [136] FIERRO, Gabe ; KOH, Jason ; NAGARE, Shreyas ; ZANG, Xiaolin ; AGARWAL, Yuvraj ; GUPTA, Rajesh K. ; CULLER, David E.: Formalizing Tag-Based Metadata With the Brick Ontology. In: *Frontiers in Built Environment* 6 (2020). <http://dx.doi.org/10.3389/fbuil.2020.558034>. – DOI 10.3389/fbuil.2020.558034
- [137] FIERRO, Gabe ; PAUWELS, Pieter: *Survey of metadata schemas for data-driven smart buildings (Annex 81): Energy in Buildings and Communities Technology Collaboration Programme*. <https://annex81.iea-ebc.org/Data/publications/IEA%20Annex%2081%20Survey%20of%20Metadata%20Schemas.pdf>. Version:2022
- [138] FIERRO, Gabe ; PRITONI, Marco ; ABDELBAKY, Moustafa ; RAFTERY, Paul ; PEFFER, Therese ; THOMSON, Greg ; CULLER, David E.: Mortar: An Open Testbed for Portable Building Analytics. In: *The 5th ACM International Conference on Systems for Built Environments (BuildSys '18)*, 2018
- [139] FIERRO, Gabe ; SAHA, Avijit ; SHAPINSKY, Tobias ; STEEN, Matthew ; ESLINGER, Hannah: Application-driven creation of building metadata models with semantic sufficiency. In: ORTIZ, Jorge (Hrsg.): *Proceedings of the 9th ACM International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation*. New York, NY, USA : ACM, 2022. – ISBN 9781450398909, S. 228–237
- [140] FISCH, Norbert ; PLESSER, Stefan ; WUSSLER, Maik ; SAUSS, David: *Forschungsinitiative Zukunft Bau. Bd. F 3043: Spec&Check Gebäudeautomation - Entwicklung und Erprobung einer Methodik zur Beschreibung, Abnahme und Überwachung von Funktionen der Gebäudeautomation*. Stuttgart : Fraunhofer IRB Verlag, 2017 <http://www.irbnet.de/daten/rswb/17099012531.pdf>. – ISBN 978-3-7388-0052-4
- [141] FRANCISCO, Abigail ; TRUONG, Hanh ; KHOSROWPOUR, Ardalan ; TAYLOR, John E. ; MOHAMMADI, Neda: Occupant perceptions of building information model-based energy visualizations in eco-feedback systems. In: *Applied Energy* 221 (2018), S. 220–228. <http://dx.doi.org/10.1016/j.apenergy.2018.03.132>. – DOI 10.1016/j.apenergy.2018.03.132. – ISSN 03062619
- [142] FREIESLEBEN, Timo ; GROTE, Thomas: Beyond generalization: a theory of robustness in machine learning. In: *Synthese* 202 (2023), Nr. 4. <http://dx.doi.org/10.1007/s11229-023-04334-9>. – DOI 10.1007/s11229-023-04334-9

- [143] FRIENDLY, E.: Jaro–Winkler Distance Improvement For Approximate String Search Using Indexing Data For Multiuser Application. In: *Journal of Physics: Conference Series* 1361 (2019), Nr. 1, S. 012080. <http://dx.doi.org/10.1088/1742-6596/1361/1/012080>. – DOI 10.1088/1742-6596/1361/1/012080. – ISSN 1742-6588
- [144] FRONTCAK, Monika ; WARGOCKI, Pawel: Literature survey on how different factors influence human comfort in indoor environments. In: *Building and Environment* 46 (2011), Nr. 4, S. 922–937. <http://dx.doi.org/10.1016/j.buildenv.2010.10.021>. – DOI 10.1016/j.buildenv.2010.10.021. – ISSN 0360-1323
- [145] FULCHER, Ben D. ; JONES, Nick S.: hctsa: A computational framework for automated time-series phenotyping using massive feature extraction. In: *Cell systems* 5 (2017), Nr. 5, S. 527–531
- [146] FÜRST, Jonathan ; CHEN, Kaifei ; KATZ, Randy H. ; BONNET, Philippe: Crowd-sourced BMS point matching and metadata maintenance with Babel. In: *2016 IEEE International Conference on Pervasive Computing and Communication workshops (PerCom workshops)*. Piscataway, NJ : Ieee, 2016. – ISBN 978-1-5090-1941-0, S. 1–6
- [147] FÜTTERER, Johannes ; KOCHANOSKI, Maksymilian ; MÜLLER, Dirk: Application of selected supervised learning methods for time series classification in Building Automation and Control Systems. In: *Energy Procedia* 122 (2017). <http://dx.doi.org/10.1016/j.egypro.2017.07.428>. – DOI 10.1016/j.egypro.2017.07.428. – ISSN 18766102
- [148] FÜTTERER, Johannes ; STINNER, Florian ; MÜLLER, Dirk: Simulation-based assessment of an easy-to-apply, automated control tuning method for typical PID control loops in building energy systems. Version:2016. https://publications.rwth-aachen.de/record/687197/files/P386_Simulation-based%20assessment%20of%20an%20easy-to-apply,%20automated%20control%20tuning%20method%20for%20typical%20PID%20control%20loops%20in%20building%20energy%20systems.pdf? In: *Proceedings of ECOS 2016*. Portoroz, Slovenia, 2016
- [149] FÜTTERER, Johannes P.: *Tuning of PID Controllers within Building Energy Systems*. RWTH Aachen University, 2017. <http://dx.doi.org/10.18154/RWTH-2017-10490>. <http://dx.doi.org/10.18154/RWTH-2017-10490>
- [150] GALINDEZ, Laura ; BADAMI, Komail ; VLASSELAER, Jonas ; MEERT, Wannes ; VERHELST, Marian: Dynamic Sensor-Frontend Tuning for Resource Efficient Embedded Classification. In: *IEEE Journal on Emerging and Selected Topics in Circuits and Systems* 8 (2018), Nr. 4, S. 858–872. <http://dx.doi.org/10.1109/JETCAS.2018.2850451>. – DOI 10.1109/JETCAS.2018.2850451. – ISSN 2156-3357
- [151] GAO, Hao ; KOCH, Christian ; WU, Yupeng: Building information modelling based building energy modelling: A review. In: *Applied Energy* 238 (2019), S. 320–343. <http://dx.doi.org/10.1016/j.apenergy.2019.04.044>

- org/10.1016/j.apenergy.2019.01.032. – DOI 10.1016/j.apenergy.2019.01.032. – ISSN 03062619
- [152] GAO, Jingkun ; BERGÉS, Mario: A large-scale evaluation of automated metadata inference approaches on sensors from air handling units. In: *Advanced Engineering Informatics* 37 (2018), S. 14–30. <http://dx.doi.org/10.1016/j.aei.2018.04.010>. – DOI 10.1016/j.aei.2018.04.010. – ISSN 14740346
- [153] GAO, Jingkun ; PLOENNIGS, Joern ; BERGES, Mario: A Data-driven Meta-data Inference Framework for Building Automation Systems. In: *Proceedings of the 2nd ACM International Conference on Embedded Systems for Energy-Efficient Built Environments*. 2821674 : ACM, 2015, S. 23–32
- [154] GASTINGER, Julia ; NICOLAS, Sébastien ; STEPIĆ, Dušica ; SCHMIDT, Mischa ; SCHÜLKE, Anett: *A study on Ensemble Learning for Time Series Forecasting and the need for Meta-Learning*. <http://arxiv.org/pdf/2104.11475v1>. Version: 2021
- [155] GAUTAMA, Temujin ; MANDIC, Danilo P. ; VAN HULLE, Marc M.: A differential entropy based method for determining the optimal embedding parameters of a signal. In: *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP'03)* Bd. 6, 2003, S. VI–29
- [156] GE, Mouzhi ; HELFERT, Markus: Impact of Information Quality on Supply Chain Decisions. In: *Journal of Computer Information Systems* 53 (2013), Nr. 4, S. 59–67. <http://dx.doi.org/10.1080/08874417.2013.11645651>. – DOI 10.1080/08874417.2013.11645651. – ISSN 0887–4417
- [157] GEBAUER, Marcus ; WINDHEUSER, Ulrich: Strukturierte Datenanalyse, Profiling und Geschäftsregeln. Version: 2018. http://dx.doi.org/10.1007/978-3-658-21994-9_5. In: HILDEBRAND, Knut (Hrsg.) ; GEBAUER, Marcus (Hrsg.) ; HINRICHS, Holger (Hrsg.) ; MIELKE, Michael (Hrsg.): *Daten- und Informationsqualität*. Wiesbaden : Springer Fachmedien Wiesbaden, 2018. – DOI 10.1007/978-3-658-21994-9_5. – ISBN 978-3-658-21993-2, S. 87–100
- [158] GERRISH, T. ; RUIKAR, K. ; COOK, M. ; JOHNSON, M. ; PHILLIP, M. ; LOWRY, C.: BIM application to building energy performance visualisation and management: Challenges and potential. In: *Energy and Buildings* 144 (2017), S. 218–228. <http://dx.doi.org/10.1016/j.enbuild.2017.03.032>. – DOI 10.1016/j.enbuild.2017.03.032. – ISSN 03787788
- [159] GHAFARIANHOSEINI, Ali ; TOOKEY, John ; GHAFARIANHOSEINI, Amirhosein ; NAISMITH, Nicola ; AZHAR, Salman ; EFIMOVA, Olia ; RAAHEMIFAR, Kaamran: Building Information Modelling (BIM) uptake: Clear benefits, understanding its implementation, risks and challenges. In: *Renewable and Sustainable Energy Reviews* 75 (2017), S. 1046–1053. <http://dx.doi.org/10.1016/j.rser.2016.11.083>. – DOI 10.1016/j.rser.2016.11.083. – ISSN 13640321

- [160] GIRALDO-SOTO, Catalina ; ERKOREKA, Aitor ; MORA, Laurent ; URIARTE, Irati ; DEL PORTILLO, Luis A.: *Monitoring System Analysis Used for Evaluating the Building's Envelope Energy Performance Through Estimation of Its Heat Lost Coefficient*. 2018. <http://dx.doi.org/10.20944/preprints201805.0474.v1>. <http://dx.doi.org/10.20944/preprints201805.0474.v1>
- [161] GORGIS, David: *Identification of optimization potentials regarding energy usage by automatic detection and evaluation of control loops in complex energy systems*, RWTH Aachen University, Masterarbeit, 2022
- [162] GROEZINGER, Jan ; HERMELINK, Andreas ; MANTEUFFEL, Bernhard von ; OFFERMANN, Markus ; SCHIMSCHAR, Sven ; WAIDE, Paul: *Optimising the energy use of technical building systems: unleashing the power of the EPBD's Article 8*. (2017)
- [163] GUERRA-SANTIN, Olivia ; TWEED, Christopher A.: In-use monitoring of buildings: An overview of data collection methods. In: *Energy and Buildings* 93 (2015), S. 189–207. <http://dx.doi.org/10.1016/j.enbuild.2015.02.042>. – DOI 10.1016/j.enbuild.2015.02.042. – ISSN 03787788
- [164] GUIGNARD, Fabian ; LAIB, Mohamed ; AMATO, Federico ; KANEVSKI, Mikhail: Advanced analysis of temporal data using Fisher-Shannon information: theoretical development and application in geosciences. In: *Frontiers in Earth Science* 8 (2020), S. 255
- [165] GUNNING, Robert: *The technique of clear writing*. Rev. ed. New York : Mcgraw-Hill, 1968. – ISBN 0070252068
- [166] GURSEL, Ezgi ; REDDY, Bhavya ; KHOJANDI, Anahita ; MADADI, Mahboubeh ; COBLE, Jamie B. ; AGARWAL, Vivek ; YADAV, Vaibhav ; BORING, Ronald L.: Using artificial intelligence to detect human errors in nuclear power plants: A case in operation and maintenance. In: *Nuclear Engineering and Technology* 55 (2023), Nr. 2, S. 603–622. <http://dx.doi.org/10.1016/j.net.2022.10.032>. – DOI 10.1016/j.net.2022.10.032. – ISSN 17385733
- [167] HAASE, Martin ; HEITMANN, Kai: *Die Erweiterte Kölner Phonetik*. (2000), S. 526
- [168] HAMIZ AHMED, Muhammad: *Meta Data Inference on Building Sensors Data*. Freiburg, Albert-Ludwig-University Freiburg, Masterarbeit, 2019. https://ad-publications.cs.uni-freiburg.de/theses/Master_Muhammad_Hamiz_Ahmed_2019.pdf
- [169] HAMMING, R. W.: Error Detecting and Error Correcting Codes. In: *Bell System Technical Journal* 29 (1950), Nr. 2, S. 147–160. <http://dx.doi.org/10.1002/j.1538-7305.1950.tb00463.x>. – DOI 10.1002/j.1538-7305.1950.tb00463.x. – ISSN 00058580
- [170] HANCOCK, John T. ; KHOSHGOFTAAR, Taghi M.: CatBoost for big data: an interdisciplinary review. In: *Journal of big data* 7 (2020), Nr. 1, S. 94. <http://dx.doi.org/10.1186/s40537-020-00369-8>. – DOI 10.1186/s40537-020-00369-8. – ISSN 2196-1115

- [171] HANCOX-LI, Leif: Robustness in machine learning explanations. In: HILDEBRANDT, Mireille (Hrsg.): *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. New York, NY, United States : Association for Computing Machinery, 2020 (ACM Digital Library). – ISBN 9781450369367, S. 640–647
- [172] HANNAN, Mahammad A. ; FAISAL, Mohammad ; KER, Pin J. ; MUN, Looe H. ; PARVIN, Khadija ; MAHLIA, Teuku Meurah I. ; BLAABJERG, Frede: A Review of Internet of Energy Based Building Energy Management Systems: Issues and Recommendations. In: *IEEE Access* 6 (2018), S. 38997–39014. <http://dx.doi.org/10.1109/ACCESS.2018.2852811>. – DOI 10.1109/ACCESS.2018.2852811. – ISSN 2169–3536
- [173] HARDSTONE, Richard ; POIL, Simon-Shlomo ; SCHIAVONE, Giuseppina ; JANSEN, Rick ; NIKULIN, Vadim V. ; MANSVELDER, Huibert D. ; LINKENKAER-HANSEN, Klaus: Detrended fluctuation analysis: a scale-free view on neuronal oscillations. In: *Frontiers in physiology* 3 (2012), S. 450. <http://dx.doi.org/10.3389/fphys.2012.00450>. – DOI 10.3389/fphys.2012.00450
- [174] HARISH, V.S.K.V. ; KUMAR, Arun: A review on modeling and simulation of building energy systems. In: *Renewable and Sustainable Energy Reviews* 56 (2016), S. 1272–1292. <http://dx.doi.org/10.1016/j.rser.2015.12.040>. – DOI 10.1016/j.rser.2015.12.040. – ISSN 13640321
- [175] HARRIS, Charles R. ; MILLMAN, K. J. ; VAN DER WALT, Stéfan J. ; GOMMERS, Ralf ; VIRTANEN, Pauli ; COURNAPEAU, David ; WIESER, Eric ; TAYLOR, Julian ; BERG, Sebastian ; SMITH, Nathaniel J. ; KERN, Robert ; PICUS, Matti ; HOYER, Stephan ; VAN KERKWIJK, Marten H. ; BRETT, Matthew ; HALDANE, Allan ; DEL RÍO, Jaime F. ; WIEBE, Mark ; PETERSON, Pearu ; GÉRARD-MARCHANT, Pierre ; SHEPPARD, Kevin ; REDDY, Tyler ; WECKESSER, Warren ; ABBASI, Hameer ; GOHLKE, Christoph ; OLIPHANT, Travis E.: Array programming with NumPy. In: *Nature* 585 (2020), Nr. 7825, S. 357–362. <http://dx.doi.org/10.1038/s41586-020-2649-2>. – DOI 10.1038/s41586-020-2649-2. – ISSN 0028–0836
- [176] HBEICH, Elio ; ROXIN, Ana: Linking BIM and GIS Standard Ontologies with Linked Data. (2020)
- [177] HEIDARI BAFROUI, Hojat ; OHADI, Abdolreza: Application of wavelet energy and Shannon entropy for feature extraction in gearbox fault detection under varying speed conditions. In: *Neurocomputing* 133 (2014), S. 437–445. <http://dx.doi.org/10.1016/j.neucom.2013.12.018>. – DOI 10.1016/j.neucom.2013.12.018. – ISSN 09252312
- [178] HEIMAR ANDERSEN, Kamilla ; POMMERENCKE MELGAARD, Simon ; JOHRA, Hicham ; MARSZAL-POMIANOWSKA, Anna ; LUND JENSEN, Rasmus ; KVOLS HEISELBERG, Per: Barriers and drivers for implementation of automatic fault detection and diagnosis in buildings and HVAC systems: An outlook from industry experts. In: *Energy and Buildings* 303 (2024), S. 113801. <http://dx.doi.org/10.1016/j.enbuild.2023.113801>. – DOI 10.1016/j.enbuild.2023.113801. – ISSN 03787788

- [179] HEINRICH, Bernd ; KLIER, Mathias: Datenqualitätsmetriken für ein ökonomisch orientiertes Qualitätsmanagement. Version:2018. http://dx.doi.org/10.1007/978-3-658-21994-9_3. In: HILDEBRAND, Knut (Hrsg.) ; GEBAUER, Marcus (Hrsg.) ; HINRICH, Holger (Hrsg.) ; MIELKE, Michael (Hrsg.): *Daten- und Informationsqualität*. Wiesbaden : Springer Fachmedien Wiesbaden, 2018. – DOI 10.1007/978-3-658-21994-9_3. – ISBN 978-3-658-21993-2, S. 47–65
- [180] HENRICH, Verena ; HINRICH, Erhard: Determining Immediate Constituents of Compounds in GermaNet. In: *Proceedings of the International Conference Recent Advances in Natural Language Processing 2011*. Hissar, Bulgaria : Association for Computational Linguistics, 2011, 420–426
- [181] HENRY, Miguel ; JUDGE, George: Permutation entropy and information recovery in nonlinear dynamic economic time series. In: *Econometrics* 7 (2019), Nr. 1, S. 10
- [182] HERRANZ, Javier ; NIN, Jordi ; SOLE, Marc: Optimal Symbol Alignment Distance: A New Distance for Sequences of Symbols. In: *IEEE Transactions on Knowledge and Data Engineering* 23 (2011), Nr. 10, S. 1541–1554. <http://dx.doi.org/10.1109/TKDE.2010.190>. – DOI 10.1109/TKDE.2010.190. – ISSN 1041-4347
- [183] HIMEUR, Yassine ; GHANEM, Khalida ; ALSALEMI, Abdullah ; BENSAAI, Faycal ; AMIRA, Abbes: Artificial Intelligence based Anomaly Detection of Energy Consumption in Buildings: A Review, Current Trends and New Perspectives. In: *Applied Energy* 287 (2021), Nr. 3, 116601. <http://dx.doi.org/10.1016/j.apenergy.2021.116601>. – DOI 10.1016/j.apenergy.2021.116601. – ISSN 03062619
- [184] HIPPE, Dwayne R.: *SQLite*. <https://www.sqlite.org/download.html>. Version:2023
- [185] HJORTH, Bo: EEG analysis based on time domain properties. In: *Electroencephalography and clinical neurophysiology* 29 (1970), Nr. 3, S. 306–310
- [186] HOCHREITER, Sepp ; SCHMIDHUBER, Jürgen: Long Short-Term Memory. In: *Neural Computation* 9 (1997), Nr. 8, S. 1735–1780. <http://dx.doi.org/10.1162/neco.1997.9.8.1735>. – DOI 10.1162/neco.1997.9.8.1735. – ISSN 0899-7667
- [187] HOERL, Arthur E. ; KENNARD, Robert W.: Ridge regression: applications to nonorthogonal problems. In: *Technometrics* 12 (1970), Nr. 1, S. 69–82. – ISSN 0040-1706
- [188] HOERL, Arthur E. ; KENNARD, Robert W.: Ridge Regression: Biased Estimation for Nonorthogonal Problems. In: *Technometrics* 12 (1970), Nr. 1, S. 55–67. <http://dx.doi.org/10.1080/00401706.1970.10488634>. – DOI 10.1080/00401706.1970.10488634. – ISSN 0040-1706
- [189] HOLMEGAARD, Emil ; KJAERGAARD, Mikkel B.: Mining building metadata by data stream comparison. In: SUSTAINABILITY, IEEE Conference on Technologies f. (Hrsg.): *2016 IEEE Conference on Technologies for Sustainability (SusTech)*. Piscataway, NJ : Ieee, 2016. – ISBN 978-1-5090-4158-9, S. 28–33

- [190] HOLMES, D. ; MCCABE, M. C.: Improving precision and recall for Soundex retrieval. In: *International Conference on Information Technology, Coding and Computing*. Los Alamitos, Calif : IEEE Computer Society, 2002. – ISBN 0-7695-1506-1, S. 22–26
- [191] HONG, Dezhi: *Towards Automatic Context Inference for Sensors in Commercial Buildings*, University of Virginia, Computer Science - School of Engineering and Applied Science, PHD (Doctor of Philosophy), 2018, Diss., 2018. <https://search.lib.virginia.edu/catalog/zw12z586k>
- [192] HONG, Dezhi ; ORTIZ, Jorge ; BHATTACHARYA, Arka ; WHITEHOUSE, Kamin: *Sensor-Type Classification in Buildings*. <http://arxiv.org/pdf/1509.00498v1>. Version: 2015
- [193] HONG, Dezhi ; WANG, Hongning ; ORTIZ, Jorge ; WHITEHOUSE, Kamin: The Building Adapter: Towards Quickly Applying Building Analytics at Scale. Version: 2015. <http://dx.doi.org/10.1145/2821650.2821657>. In: *BuildSys '15*. 2015. – DOI 10.1145/2821650.2821657
- [194] HONG, Dezhi ; WANG, Hongning ; WHITEHOUSE, Kamin: Clustering-based Active Learning on Sensor Type Classification in Buildings. In: BAILEY, James (Hrsg.) ; MOFFAT, Alistair (Hrsg.) ; AGGARWAL, Charu C. (Hrsg.) ; RIJKE, Maarten de (Hrsg.) ; KUMAR, Ravi (Hrsg.) ; MURDOCK, Vanessa (Hrsg.) ; SELLIS, Timos (Hrsg.) ; YU, Jeffrey X. (Hrsg.): *Proceedings of the 24th ACM International Conference on Information and Knowledge Management & co-located workshops*. New York, NY : Association for Computing Machinery, 2015. – ISBN 9781450337946, S. 363–372
- [195] HSU, Chang F ; WEI, Sung-Yang ; HUANG, Han-Ping ; HSU, Long ; CHI, Sien ; PENG, Chung-Kang: Entropy of entropy: Measurement of dynamical complexity for biological systems. In: *Entropy* 19 (2017), Nr. 10, S. 550
- [196] HU, Zhen-Zhong ; LENG, Shuo ; LIN, Jia-Rui ; LI, Sun-Wei ; XIAO, Ya-Qi: Knowledge Extraction and Discovery Based on BIM: A Critical Review and Future Directions. In: *Archives of Computational Methods in Engineering* (2021). <http://dx.doi.org/10.1007/s11831-021-09576-9>. – DOI 10.1007/s11831-021-09576-9. – ISSN 1134-3060
- [197] HUANG, Zhiheng ; XU, Wei ; YU, Kai: *Bidirectional LSTM-CRF Models for Sequence Tagging*. <http://arxiv.org/pdf/1508.01991v1>. Version: 2015
- [198] HUCHTKOETTER, Jana ; REINHARDT, Andreas: A study on the impact of data sampling rates on load signature event detection. In: *Energy Informatics* 2 (2019), Nr. S1, S. 471. <http://dx.doi.org/10.1186/s42162-019-0096-9>. – DOI 10.1186/s42162-019-0096-9
- [199] HUH, Markus: *Die Gesamtnutzungsdauer in der Immobilienwertermittlung: Eine Analyse der Einflussfaktoren*, Technische Universität Darmstadt, Dissertation, 2019. https://tuprints.ulb.tu-darmstadt.de/9055/1/Dissertation_Markus_Huhn_2019.pdf
- [200] HURST, H. E.: Long-Term Storage Capacity of Reservoirs. In: *Transactions of the American Society of Civil Engineers* 116 (1951), Nr. 1, S. 770–799. <http://dx.doi.org/10.1061/TACEAT.0006518>. – DOI 10.1061/TACEAT.0006518. – ISSN 0066-0604

- [201] HVIID, Jakob ; JOHANSEN, Aslak ; FIERRO, Gabe ; KJÆRGAARD, Mikkel B.: Service Portability and Information Discovery in Building Operating Systems using Semantic Modeling. In: *11th International Conference on the Internet of Things*. New York, NY, USA : ACM, 2021. – ISBN 9781450385664, S. 110–117
- [202] IDDIANOZIE, Chidubem ; PALMES, Paulito: Towards smart sustainable cities: Addressing semantic heterogeneity in Building Management Systems using discriminative models. In: *Sustainable Cities and Society* 62 (2020), S. 102367. <http://dx.doi.org/10.1016/j.scs.2020.102367>. – DOI 10.1016/j.scs.2020.102367. – ISSN 22106707
- [203] IEC INTERNATIONAL ELECTROTECHNICAL COMMISSION: *ISO/IEC 25012 - Software engineering - Software product Quality Requirements and Evaluation (SQuaRE): Data quality model*. 2008
- [204] IM, Piljae ; BAE, Yeonjin ; CUI, Borui ; LEE, Seungjae ; BHATTACHARYA, Saptarshi ; ADETOLA, Veronica ; VRABIE, Draguna ; ZHANG, Liang ; LEACH, Matt: *Literature Review for Sensor Impact Evaluation and Verification Use Cases - Building Controls and Fault Detection and Diagnosis (FDD)*
- [205] IMÖHL, Sören ; GELOWICZ, Svenja: Längstes deutsches Wort: Das ist das längste Wort der deutschen Sprache. In: *Wirtschaftswache* (05.05.2023). <https://www.wiwo.de/politik/deutschland/laengstes-deutsches-wort-das-ist-das-laengste-wort-der-deutschen-sprache/8290960.html>
- [206] INTERNATIONAL ENERGY AGENCY (IEA): *Data centres and energy – from global headlines to local headaches?* <https://www.iea.org/commentaries/data-centres-and-energy-from-global-headlines-to-local-headaches>. Version: 2019
- [207] ISLAM, Aminul ; INKPEN, Diana: Semantic text similarity using corpus-based word similarity and string similarity. In: *ACM Transactions on Knowledge Discovery from Data* 2 (2008), Nr. 2, S. 1–25. <http://dx.doi.org/10.1145/1376815.1376819>. – DOI 10.1145/1376815.1376819. – ISSN 1556–4681
- [208] ISMAIL FAWAZ, Hassan ; FORESTIER, Germain ; WEBER, Jonathan ; IDOUMGHAR, Lhassane ; MULLER, Pierre-Alain: Deep learning for time series classification: a review. In: *Data Mining and Knowledge Discovery* 33 (2019), Nr. 4, S. 917–963. <http://dx.doi.org/10.1007/s10618-019-00619-1>. – DOI 10.1007/s10618-019-00619-1. – ISSN 1384–5810
- [209] ITTEN, René ; HISCHIER, Roland ; ANDRAE, Anders S. G. ; BIESER, Jan C. T. ; CABERNARD, Livia ; FALKE, Annemarie ; FERREBOEUF, Hugues ; HILTY, Lorenz M. ; KELLER, Regula L. ; LEES-PERASSO, Etienne ; PREIST, Chris ; STUCKI, Matthias: Digital transformation—life cycle assessment of digital services, multifunctional devices and cloud computing. In: *The Interna-*

- tional Journal of Life Cycle Assessment* 25 (2020), Nr. 10, S. 2093–2098. <http://dx.doi.org/10.1007/s11367-020-01801-0>. – DOI 10.1007/s11367-020-01801-0. – ISSN 0948-3349
- [210] JAEGER, Ina de ; LAGO, Jesus ; SAELENS, Dirk: A probabilistic building characterization method for district energy simulations. In: *Energy and Buildings* 230 (2021), S. 110566. <http://dx.doi.org/10.1016/j.enbuild.2020.110566>. – DOI 10.1016/j.enbuild.2020.110566. – ISSN 03787788
- [211] JARARWEH, Yaser ; OTOUM, Safa ; RIDHAWI, Ismaeel A.: Trustworthy and sustainable smart city services at the edge. In: *Sustainable Cities and Society* 62 (2020), S. 102394. <http://dx.doi.org/10.1016/j.scs.2020.102394>. – DOI 10.1016/j.scs.2020.102394. – ISSN 22106707
- [212] JARVIS, Scott: Capturing the Diversity in Lexical Diversity. In: *Language Learning* 63 (2013), S. 87–106. <http://dx.doi.org/10.1111/j.1467-9922.2012.00739.x>. – DOI 10.1111/j.1467-9922.2012.00739.x. – ISSN 00238333
- [213] JATON, Florian: Assessing biases, relaxing moralism: On ground-truthing practices in machine learning design and application. In: *Big Data & Society* 8 (2021), Nr. 1, S. 20539517211013569. <http://dx.doi.org/10.1177/20539517211013569>. – DOI 10.1177/20539517211013569. – ISSN 2053-9517
- [214] JIANG, Xiaoqian ; OSL, Melanie ; KIM, Jihoon ; OHNO-MACHADO, Lucila: Smooth isotonic regression: a new method to calibrate predictive models. In: *AMIA Summits on Translational Science Proceedings* 2011 (2011), S. 16–20
- [215] JIAO, Yang ; LI, Jiacheng ; WU, Jiaman ; HONG, Dezhi ; GUPTA, Rajesh ; SHANG, Jingbo: SeNser: Learning Cross-Building Sensor Metadata Tagger. In: COHN, Trevor (Hrsg.) ; HE, Yulan (Hrsg.) ; LIU, Yang (Hrsg.): *Findings of the Association for Computational Linguistics: EMNLP 2020*. Stroudsburg, PA, USA : Association for Computational Linguistics, 2020, S. 950–960
- [216] JIN, Hongyue ; AFIUNY, Peter ; MCINTYRE, Timothy ; YIH, Yuehwern ; SUTHERLAND, John W.: Comparative life cycle assessment of NdFeB magnets: virgin production versus magnet-to-magnet recycling. In: *Procedia CIRP* 48 (2016), S. 45–50
- [217] JO, Hang-Hyun ; HIRAOKA, Takayuki: Bursty Time Series Analysis for Temporal Networks. Version: 2019. http://dx.doi.org/10.1007/978-3-030-23495-9_9. In: HOLME, Petter (Hrsg.) ; SARAMÄKI, Jari (Hrsg.): *Temporal Network Theory*. Cham : Springer, 2019 (Springer eBook Collection). – DOI 10.1007/978-3-030-23495-9_9. – ISBN 9783030234959, S. 161–179
- [218] JOHRA, Hicham ; SCHAFFER, Markus ; CHAUDHARY, Gaurav ; KAZMI, Hussain S. ; LE DRÉAU, Jérôme ; PETERSEN, Steffen: *Coherent description of 48 metrics to compare, validate and assess accuracy of building energy models and indoor environment simulations*
- [219] JOST, Lou: Entropy and diversity. In: *Oikos* 113 (2006), Nr. 2, S. 363–375. <http://dx.doi.org/10.1111/j.2006.0030-1299.14714.x>. – DOI 10.1111/j.2006.0030-1299.14714.x. – ISSN 00301299

- [220] JUNG, Wooyoung ; LEE, Ghang: The status of BIM adoption on six continents. In: *International journal of civil, environmental, structural, construction and architectural engineering* 9 (2015), Nr. 5, 444–448. <https://waset.org/publications/10001095/the-status-of-bim-adoption-on-six-continents>
- [221] KALAUZI, Aleksandar ; BOJIĆ, Tijana ; RAKIĆ, Ljubisav: Extracting complexity waveforms from one-dimensional signals. In: *Nonlinear biomedical physics* 3 (2009), S. 1–11
- [222] KARIM, Fazle ; MAJUMDAR, Somshubra ; DARABI, Houshang: Insights into LSTM Fully Convolutional Networks for Time Series Classification. In: *IEEE Access* 7 (2019), 67718–67725. <http://dx.doi.org/10.1109/ACCESS.2019.2916828>. – DOI 10.1109/ACCESS.2019.2916828. – ISSN 2169–3536
- [223] KARIM, Fazle ; MAJUMDAR, Somshubra ; DARABI, Houshang ; CHEN, Shun: LSTM Fully Convolutional Networks for Time Series Classification. In: *IEEE Access* 6 (2018), S. 1662–1669. <http://dx.doi.org/10.1109/ACCESS.2017.2779939>. – DOI 10.1109/ACCESS.2017.2779939. – ISSN 2169–3536
- [224] KARIM, Fazle ; MAJUMDAR, Somshubra ; DARABI, Houshang ; HARFORD, Samuel: *Multivariate LSTM-FCNs for Time Series Classification*. <http://arxiv.org/pdf/1801.04503v1>. Version: 2018
- [225] KATHIRGAMANATHAN, Anjukan ; ROSA, Mattia D. ; MANGINA, Eleni ; FINN, Donal P: Data-driven Predictive Control for Unlocking Building Energy Flexibility: A Review. In: *Renewable and Sustainable Energy Reviews* 135 (2021), 110120. <http://dx.doi.org/10.1016/j.rser.2020.110120>. – DOI 10.1016/j.rser.2020.110120. – ISSN 13640321
- [226] KATZ, Michael J.: Fractals and the analysis of waveforms. In: *Computers in biology and medicine* 18 (1988), Nr. 3, S. 145–156
- [227] KATZ, Randy H. ; CULLER, David E. ; SANDERS, Seth ; ALSPAUGH, Sara ; CHEN, Yanpei ; DAWSON-HAGGERTY, Stephen ; DUTTA, Prabal ; HE, Mike ; JIANG, Xiaofan ; KEYS, Laura ; KRIOUKOV, Andrew ; LUTZ, Ken ; ORTIZ, Jorge ; MOHAN, Prashanth ; REUTZEL, Evan ; TANEJA, Jay ; HSU, Jeff ; SHANKAR, Sushant: An information-centric energy infrastructure: The Berkeley view. In: *Sustainable Computing: Informatics and Systems* 1 (2011), Nr. 1, S. 7–22. <http://dx.doi.org/10.1016/j.suscom.2010.10.001>. – DOI 10.1016/j.suscom.2010.10.001. – ISSN 22105379
- [228] KAUR, Davinder ; USLU, Suleyman ; RITTICHER, Kaley J. ; DURRESI, Arjan: Trustworthy Artificial Intelligence: A Review. In: *ACM Computing Surveys* 55 (2023), Nr. 2, S. 1–38. <http://dx.doi.org/10.1145/3491209>. – DOI 10.1145/3491209. – ISSN 03600300
- [229] KAZMI, Hussain ; MUNNÉ-COLLADO, Íngrid ; MEHMOOD, Fahad ; SYED, Tahir A. ; DRIESEN, Johan: Towards data-driven energy communities: A review of open-source datasets, models and tools. In: *Renewable and Sustainable Energy Reviews* 148 (2021), S. 111290.

- <http://dx.doi.org/10.1016/j.rser.2021.111290>. – DOI 10.1016/j.rser.2021.111290. – ISSN 13640321
- [230] KEBEDE, Rahel ; MOSCATI, Annika ; TAN, He ; JOHANSSON, Peter: Integration of manufacturers' product data in BIM platforms using semantic web technologies. In: *Automation in Construction* 144 (2022), S. 104630. <http://dx.doi.org/10.1016/j.autcon.2022.104630>. – DOI 10.1016/j.autcon.2022.104630. – ISSN 09265805
- [231] KEIL, Jan M.: Efficient Bounded Jaro-Winkler Similarity Based Search. (2019). <http://dx.doi.org/10.18420/BTW2019-13>. – DOI 10.18420/BTW2019-13
- [232] KENETT, Ron ; SHMUELI, Galit: Dimensions of information quality and InfoQ assessment. Version: 2017. <http://dx.doi.org/10.1002/9781118890622.ch3>. In: KENETT, Ron (Hrsg.) ; SHMUELI, Galit (Hrsg.): *Information Quality*. Chichester, West Sussex : Wiley, 2017. – DOI 10.1002/9781118890622.ch3. – ISBN 9781118874448, S. 31–52
- [233] KENETT, Ron S. ; SHMUELI, Galit: On information quality. In: *Journal of the Royal Statistical Society Series A* 177 (2014), Nr. 1, 3–38. <https://ideas.repec.org/a/bla/jorssa/v177y2014i1p3-38.html>
- [234] KETTUNEN, Kimmo: Can Type-Token Ratio be Used to Show Morphological Complexity of Languages? In: *Journal of Quantitative Linguistics* 21 (2014), Nr. 3, S. 223–245. <http://dx.doi.org/10.1080/09296174.2014.911506>. – DOI 10.1080/09296174.2014.911506
- [235] KEZUNOVIC, M.: Monitoring of Power System Topology in Real-Time. In: SPRAGUE, Ralph H. (Hrsg.): *Proceedings of the 39th Annual Hawaii International Conference on System Sciences, 2006, HICSS '06*. Los Alamitos, Calif. : IEEE Computer Society, 2006. – ISBN 0-7695-2507-5, S. 244b–244b
- [236] KHALILNEJAD, Arash ; FRENCH, Roger H. ; ABRAMSON, Alexis R.: Data-driven evaluation of HVAC operation and savings in commercial buildings. In: *Applied Energy* 278 (2020), S. 115505. <http://dx.doi.org/10.1016/j.apenergy.2020.115505>. – DOI 10.1016/j.apenergy.2020.115505. – ISSN 03062619
- [237] KHALILNEJAD, Arash ; KARIMI, Ahmad M. ; KAMATH, Shreyas ; HADDADIAN, Rojia ; FRENCH, Roger H. ; ABRAMSON, Alexis R.: Automated pipeline framework for processing of large-scale building energy time series data. In: *PloS one* 15 (2020), Nr. 12, S. e0240461. <http://dx.doi.org/10.1371/journal.pone.0240461>. – DOI 10.1371/journal.pone.0240461
- [238] KHORASGANI, Hamed ; BISWAS, Gautam: A methodology for monitoring smart buildings with incomplete models. In: *Applied Soft Computing* 71 (2018), S. 396–406
- [239] KIM, Ryunhee ; HONG, Yejin ; CHOI, Youngwoong ; YOON, Sungmin: System-level fouling detection of district heating substations using virtual-sensor-assisted building automation system. In: *Energy* 227 (2021), S. 120515. <http://dx.doi.org/10.1016/j.energy.2021.120515>. – DOI 10.1016/j.energy.2021.120515. – ISSN 03605442

- [240] KIM, Woohyun ; KATIPAMULA, Srinivas: A review of fault detection and diagnostics methods for building systems. In: *Science and Technology for the Built Environment* 24 (2018), Nr. 1, S. 3–21. <http://dx.doi.org/10.1080/23744731.2017.1318008>. – DOI 10.1080/23744731.2017.1318008. – ISSN 2374–4731
- [241] KLEIN, Barbara D. ; GOODHUE, Dale L. ; DAVIS, Gordon B.: Can Humans Detect Errors in Data? Impact of Base Rates, Incentives, and Goals. In: *MIS Quarterly* 21 (1997), Nr. 2, S. 169. <http://dx.doi.org/10.2307/249418>. – DOI 10.2307/249418. – ISSN 02767783
- [242] KNUTH, Donald E.: *Art of Computer Programming, Volume 4, Fascicle 4, The: Generating All Trees–History of Combinatorial Generation*. Addison-Wesley Professional, 2013
- [243] KNUTH, Donald E. ; MORRIS, Jr. James H. ; PRATT, Vaughan R.: Fast Pattern Matching in Strings. In: *SIAM Journal on Computing* 6 (1977), Nr. 2, S. 323–350. <http://dx.doi.org/10.1137/0206024>. – DOI 10.1137/0206024. – ISSN 0097–5397
- [244] KOH, Jason ; BALAJI, Bharathan ; AKHLAGHI, Vahideh ; AGARWAL, Yuvraj ; GUPTA, Rajesh: *Quiver: Using Control Perturbations to Increase the Observability of Sensor Data in Smart Buildings*. <http://arxiv.org/pdf/1601.07260v1>. Version: 2016
- [245] KOH, Jason ; BALAJI, Bharathan ; SENGUPTA, Dhiman ; MCAULEY, Julian ; GUPTA, Rajesh ; AGARWAL, Yuvraj: Scrabble: Transferrable Semi-Automated Semantic Metadata Normalization using Intermediate Representation. In: RAMACHANDRAN, Gowri S. (Hrsg.) ; BATRA, Nipon (Hrsg.): *BuildSys'18*. New York, New York : The Association for Computing Machinery, 2018. – ISBN 9781450359511, S. 11–20
- [246] KOJIMA, Masumi ; YAMASHITA, Junko: Reliability of lexical richness measures based on word lists in short second language productions. In: *System* 42 (2014), S. 23–33. <http://dx.doi.org/10.1016/j.system.2013.10.019>. – DOI 10.1016/j.system.2013.10.019. – ISSN 0346251X
- [247] KOKAR, Mieczyslaw M. ; MATHEUS, Christopher J. ; BACLAWSKI, Kenneth: Ontology-based situation awareness. In: *Information Fusion* 10 (2009), Nr. 1, S. 83–98. <http://dx.doi.org/10.1016/j.inffus.2007.01.004>. – DOI 10.1016/j.inffus.2007.01.004. – ISSN 15662535
- [248] KOLODENKOVA, Anna E. ; KHALIKOVA, Elena A. ; VERESHCHAGINA, Svetlana S.: Development of a Diagnostic Data Fusion Model of the Electrical Equipment at Industrial Enterprises. Version: 2020. http://dx.doi.org/10.1007/978-3-030-50097-9_50. In: KOVALEV, Sergey (Hrsg.) ; TARASSOV, Valery (Hrsg.) ; SNASEL, Vaclav (Hrsg.) ; SUKHANOV, Andrey (Hrsg.): *Proceedings of the Fourth International Scientific Conference “Intelligent Information Technologies for Industry” (IITI'19)* Bd. 1156. Cham : Springer International Publishing and Imprint Springer, 2020. – DOI 10.1007/978-3-030-50097-9_50. – ISBN 978-3-030-50096-2, S. 499–506
- [249] KOSHELEVA, Olga ; KREINOVICH, Vladik ; AUTCHARIYAPANIKUL, Kittawit: Commonsense ex-

- planations of sparsity, Zipf law, and Nash's bargaining solution. In: *Prediction and Causality in Econometrics and Related Topics* (2022), S. 67–74
- [250] KOSONEN, R. ; TAN, E.: The effect of perceived indoor air quality on productivity loss. In: *Energy and Buildings* 36 (2004), Nr. 10, S. 981–986. <http://dx.doi.org/10.1016/j.enbuild.2004.06.005>. – DOI 10.1016/j.enbuild.2004.06.005. – ISSN 03787788
- [251] KRÄUCHI, Philipp ; JURT, Dominic ; DAHINDEN, Christoph ; BUNDESAMT FÜR ENERGIE BFE, PROGRAMM ENERGIESCHWEIZ, SCHWEIZ (Hrsg.): *Eigenenergieverbrauch der Gebäudeautomation (EEV-GA): Ergebnisbericht*. <https://pubdb.bfe.admin.ch/de/publication/download/8307>. Version: 2016
- [252] KRAWCZYK, Malgorzata J. ; MALARZ, Krzysztof: *Recovering Zipf's law in intercontinental scientific collaboration*. <https://arxiv.org/pdf/2303.07356.pdf>. Version: 2023
- [253] KRIZHEVSKY, Alex ; SUTSKEVER, Ilya ; HINTON, Geoffrey E.: ImageNet Classification with Deep Convolutional Neural Networks. Version: 2012. <http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf>. In: F. PEREIRA (Hrsg.) ; C. J. C. BURGESS (Hrsg.) ; L. BOTTOU (Hrsg.) ; K. Q. WEINBERGER (Hrsg.): *Advances in Neural Information Processing Systems* 25. Curran Associates, Inc, 2012, 1097–1105
- [254] KROTTMAIER, Johannes: *Versuchsplanung: Der Weg zur Qualität des Jahres 2000 ; [vom Preisträger des 2. European Taguchi Award für die beste Taguchi-Versuchsplanung und -analyse des Jahres 1989. 2. überarb. Aufl. Zürich and Köln : Verl. Industrielle Organisation and Verl. TÜV Rheinland, 1991 (Praxiswissen für Ingenieure). – ISBN 3857439459*
- [255] KULLBACK, S. ; LEIBLER, R. A.: On Information and Sufficiency. In: *The Annals of Mathematical Statistics* 22 (1951), Nr. 1, S. 79–86. <http://dx.doi.org/10.1214/aoms/1177729694>. – DOI 10.1214/aoms/1177729694. – ISSN 0003–4851
- [256] KUMAR, Ananya ; LIANG, Percy ; MA, Tengyu: *Verified Uncertainty Calibration*. <http://arxiv.org/pdf/1909.10155v2>. Version: 2019
- [257] KYLE, Kristopher: Measuring Lexical Richness. In: WEBB, Stuart (Hrsg.): *The Routledge handbook of vocabulary studies*. Abingdon and New York : Routledge Taylor & Francis Group, 2019 (Routledge handbooks in linguistics). – ISBN 9780429291586, S. 454–476
- [258] LANDIS, J. R. ; KOCH, Gary G.: The Measurement of Observer Agreement for Categorical Data. In: *Biometrics* 33 (1977), Nr. 1, 159. <http://dx.doi.org/10.2307/2529310>. – DOI 10.2307/2529310. – ISSN 0006341X
- [259] LANGE, Henrik ; JOHANSEN, Aslak ; KJÆRGAARD, Mikkel B.: Evaluation of the opportunities and limitations of using IFC models as source of building metadata. In: RAMACHANDRAN, Gowri S. (Hrsg.) ; BATRA, Nipon (Hrsg.): *BuildSys'18*. New York, New York : The Association for Computing Machinery, 2018. – ISBN 9781450359511, S. 21–24

- [260] LEE, Bruce W. ; JANG, Yoo S. ; LEE, Jason Hyung-Jong: *Pushing on Text Readability Assessment: A Transformer Meets Handcrafted Linguistic Features*. <https://arxiv.org/pdf/2109.12258.pdf>. Version: 2021
- [261] LEE, Bruce W. ; LEE, Jason Hyung-Jong: *LFTK: Handcrafted Features in Computational Linguistics*. <https://arxiv.org/pdf/2305.15878.pdf>. Version: 2023
- [262] LEE, Yang W. ; STRONG, Diane M. ; KAHN, Beverly K. ; WANG, Richard Y.: AIMQ: a methodology for information quality assessment. In: *Information & Management* (2002), Nr. 40, S. 133–146
- [263] LEI, Xuechen ; CHEN, Yan ; BERGÉS, Mario ; AKINCI, Burcu: Formalized control logic fault definition with ontological reasoning for air handling units. In: *Automation in Construction* 129 (2021), S. 103781. <http://dx.doi.org/10.1016/j.autcon.2021.103781>. – DOI 10.1016/j.autcon.2021.103781. – ISSN 09265805
- [264] LEIDEN, Kenneth ; LAUGHERY, K. R. ; KELLER, John ; FRENCH, Jon ; WARWICK, Walter ; WOOD, Scott D.: A Review of Human Performance Models for the Prediction of Human Error. In: *Ann Arbor* 1001 (2001), 48105. <https://hsi.arc.nasa.gov/groups/HCSL/publications/HumanErrorModels.pdf>
- [265] LEITAO, Paulo ; KARNOUSKOS, Stamatis ; RIBEIRO, Luis ; MOUTIS, Panayiotis ; BARBOSA, Jose ; STRASSER, Thomas. I.: Common practices for integrating industrial agents and low level automation functions. In: *IECON 2017 - 43rd Annual Conference of the IEEE Industrial Electronics Society*, Ieee, 2017. – ISBN 978–1–5386–1127–2, S. 6665–6670
- [266] LEONARDI, Francesco ; M REEVE, Hayden ; C WAGNER, Timothy ; XIONG, Ziyu ; PARK, June Y.: Assisted Point Mapping to Enable Cost-effective Deployment of Intelligent Building Applications. In: *International High Performance Buildings Conference 2016*, 2016
- [267] LEONARDO, Brinardi ; HANSUN, Seng: Text Documents Plagiarism Detection using Rabin-Karp and Jaro-Winkler Distance Algorithms. In: *Indonesian Journal of Electrical Engineering and Computer Science* 5 (2017), Nr. 2, S. 462. <http://dx.doi.org/10.11591/ijeecs.v5.i2.pp462-471>. – DOI 10.11591/ijeecs.v5.i2.pp462–471. – ISSN 2502–4752
- [268] LEVENSHTAIN, Vladimir I.: Binary codes capable of correcting deletions, insertions, and reversals. In: *Soviet physics doklady* Bd. 10, 1966, S. 707–710
- [269] LI, Dan ; ZHOU, Yuxun ; HU, Guoqiang ; SPANOS, Costas J.: Optimal sensor configuration and feature selection for AHU fault detection and diagnosis. In: *Ieee Transactions on Industrial Informatics* 13 (2016), Nr. 3, S. 1369–1380. – ISSN 1551–3203
- [270] LI, Dan ; ZHOU, Yuxun ; HU, Guoqiang ; SPANOS, Costas J.: Identifying unseen faults for smart buildings by incorporating expert knowledge with data. In: *IEEE Transactions on Automation Science and Engineering* 16 (2018), Nr. 3, S. 1412–1425. – ISSN 1545–5955
- [271] LI, Dan ; ZHOU, Yuxun ; HU, Guoqiang ; SPANOS, Costas J.: Handling incomplete sensor mea-

- surements in fault detection and diagnosis for building HVAC systems. In: *IEEE Transactions on Automation Science and Engineering* 17 (2019), Nr. 2, S. 833–846. – ISSN 1545–5955
- [272] LI, Guannan ; HU, Haonan ; GAO, Jiajia ; FANG, Xi: Dynamic Calibration Method of Sensor Drift Fault in HVAC System Based on Bayesian Inference. In: *Sensors (Basel, Switzerland)* 22 (2022), Nr. 14. <http://dx.doi.org/10.3390/s22145348>. – DOI 10.3390/s22145348
- [273] LI, Guannan ; HU, Yunpeng ; LIU, Jiangyan ; FANG, Xi ; KANG, Jing: Review on Fault Detection and Diagnosis Feature Engineering in Building Heating, Ventilation, Air Conditioning and Refrigeration Systems. In: *IEEE Access* 9 (2021), S. 2153–2187. <http://dx.doi.org/10.1109/ACCESS.2020.3040980>. – DOI 10.1109/ACCESS.2020.3040980. – ISSN 2169–3536
- [274] LI, Guoyan ; YUAN, Chenxi ; KAMARTHI, Sagar ; MOGHADDAM, Mohsen ; JIN, Xiaoning: Data science skills and domain knowledge requirements in the manufacturing industry: A gap analysis. In: *Journal of Manufacturing Systems* 60 (2021), S. 692–706. <http://dx.doi.org/10.1016/j.jmsy.2021.07.007>. – DOI 10.1016/j.jmsy.2021.07.007. – ISSN 02786125
- [275] LI, Shuheng ; HONG, Dezhi ; WANG, Hongning: Relation Inference among Sensor Time Series in Smart Buildings with Metric Learning. In: *Proceedings of the AAAI Conference on Artificial Intelligence* 34 (2020), Nr. 04, S. 4683–4690. <http://dx.doi.org/10.1609/aaai.v34i04.5900>. – DOI 10.1609/aaai.v34i04.5900. – ISSN 2159–5399
- [276] LI, Shun ; WEN, Jin: Application of pattern matching method for detecting faults in air handling unit system. In: *Automation in Construction* 43 (2014), S. 49–58. – ISSN 09265805
- [277] LI, Shun ; WEN, Jin: A model-based fault detection and diagnostic methodology based on PCA method and wavelet transform. In: *Energy and Buildings* 68 (2014), S. 63–71. – ISSN 03787788
- [278] LI, Wen-Syan ; CLIFTON, Chris: Semantic Integration in Heterogeneous Databases Using Neural Networks. In: *Proceedings of the 20th International Conference on Very Large Data Bases*. San Francisco, CA, USA : Morgan Kaufmann Publishers Inc, 1994 (VLDB '94). – ISBN 1558601538, 1–12
- [279] LI, Yanfei ; O'NEILL, Zheng ; ZHANG, Liang ; CHEN, Jianli ; IM, Piljae ; DEGRAU, Jason: Grey-box modeling and application for building energy simulations - A critical review. In: *Renewable and Sustainable Energy Reviews* 146 (2021), S. 111174. <http://dx.doi.org/10.1016/j.rser.2021.111174>. – DOI 10.1016/j.rser.2021.111174. – ISSN 13640321
- [280] LI, Yuchao ; ZHAO, Qin ; LIU, Yunhe ; HEI, Xinhong ; LI, Zongjian: Semiautomatic Generation of Code Ontology Using ifcOWL in Compliance Checking. In: *Advances in Civil Engineering* 2021 (2021), S. 1–18. <http://dx.doi.org/10.1155/2021/8861625>. – DOI 10.1155/2021/8861625. – ISSN 1687–8086
- [281] LIANG, Jian ; DU, Ruxu: Model-based fault detection and diagnosis of HVAC systems using support vector machine method. In: *International Journal of Refrigeration* 30 (2007), Nr. 6, S. 1104–1114. – ISSN 01407007

-
- [282] LIFE4: *TextDistance*. <https://github.com/life4/textdistance>. Version: 2024
 - [283] LIN, Guanjing ; PRITONI, Marco ; CHEN, Yimin ; GRANDERSON, Jessica ; MOROMISATO, Ricardo ; KOZLEN, Stephen: *Can We Fix It Automatically? Development of Fault Auto-Correction Algorithms for HVAC and Lighting Systems*
 - [284] LIN, Guanjing ; PRITONI, Marco ; CHEN, Yimin ; WEYANDT, Christopher ; VITTI, Raphael ; GRANDERSON, Jessica: Fault "Auto-correction" for HVAC Systems: A Preliminary Study. In: *International High Performance Buildings Conference 2021* (2021)
 - [285] LIN, Jessica ; KEOGH, Eamonn ; WEI, Li ; LONARDI, Stefano: Experiencing SAX: a novel symbolic representation of time series. In: *Data Mining and Knowledge Discovery* 15 (2007), Nr. 2, S. 107–144. <http://dx.doi.org/10.1007/s10618-007-0064-z>. – DOI 10.1007/s10618-007-0064-z. – ISSN 1573–756X
 - [286] LIN, Lu ; LUO, Zheng ; HONG, Dezhi ; WANG, Hongning: Sequential Learning with Active Partial Labeling for Building Metadata. In: *Proceedings of the 6th ACM International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation*. New York, NY, USA : ACM, 2019. – ISBN 9781450370059, S. 189–192
 - [287] LINDBLAD, Gustav: *Evaluating Performance of Pattern Searching Algorithms on Wildcard Patterns*. Umeå, Umeå University, Department of Computing Science, Bachelor thesis, 2023
 - [288] LITTLE, Christopher: *abydos*. <https://github.com/chrislit/abydos>. Version: 2020
 - [289] LIU, Huiheng ; LIANG, Jinrui ; LIU, Yanchen ; WU, Huijun: A Review of Data-Driven Building Energy Prediction. In: *Buildings* 13 (2023), Nr. 2, S. 532. <http://dx.doi.org/10.3390/buildings13020532>. – DOI 10.3390/buildings13020532
 - [290] LIU, Xiaofeng ; JIANG, Aimin ; XU, Ning ; XUE, Jianru: Increment entropy as a measure of complexity for time series. In: *Entropy* 18 (2016), Nr. 1, S. 22
 - [291] LIU, Yinhan ; OTT, Myle ; GOYAL, Naman ; DU JINGFEI ; JOSHI, Mandar ; CHEN, Danqi ; LEVY, Omer ; LEWIS, Mike ; ZETTMAYER, Luke ; STOYANOV, Veselin: *RoBERTa: A Robustly Optimized BERT Pretraining Approach*. <http://arxiv.org/pdf/1907.11692>. Version: 2019
 - [292] LJUNG, Lennart: System Identification: LiTH-ISY-R-2809. (2007)
 - [293] LONG, Zijun ; ZHUANG, Lipeng ; KILICK, George ; MCCREADIE, Richard ; CAMARASA, Gerardo A. ; HENDERSON, Paul: *Understanding and Mitigating Human-Labeling Errors in Supervised Contrastive Learning*. <http://arxiv.org/pdf/2403.06289v1>
 - [294] LÖNING, Markus ; BAGNALL, Anthony ; GANESH, Sajaysurya ; KAZAKOV, Viktor ; LINES, Jason ; KIRÁLY, Franz J.: *sktime: A Unified Interface for Machine Learning with Time Series*. <http://arxiv.org/pdf/1909.07872v1>. Version: 2019
 - [295] LU, Xiaofei: The Relationship of Lexical Richness to the Quality of ESL Learners' Oral Narratives. In: *The Modern Language Journal* 96 (2012), Nr. 2, S. 190–208. <http://dx.doi.org/>

- 10.1111/j.1540-4781.2011.01232.x. – DOI 10.1111/j.1540-4781.2011.01232.x. – ISSN 00267902
- [296] LUBBA, Carl H. ; SETHI, Sarab S. ; KNAUTE, Philip ; SCHULTZ, Simon R. ; FULCHER, Ben D. ; JONES, Nick S.: *catch22: CAnonical Time-series CHaracteristics*. <http://arxiv.org/pdf/1901.10200v2>. Version: 2019
- [297] LUO, Na ; PRITONI, Marco ; HONG, Tianzhen: An overview of data tools for representing and managing building information and performance data. In: *Renewable & Sustainable Energy Reviews* 147 (2021), 111224. <http://dx.doi.org/10.1016/j.rser.2021.111224>. – DOI 10.1016/j.rser.2021.111224. – ISSN 1364-0321
- [298] MA, Jing ; HONG, Dezhi ; WANG, Hongning: Selective Sampling for Sensor Type Classification in Buildings. In: *2020 19th ACM/IEEE International Conference on Information Processing in Sensor Networks (IPSN)*, Ieee, 4/21/2020 - 4/24/2020. – ISBN 978-1-7281-5497-8, S. 241–252
- [299] MADITHIYAGASTHENNA, Dumindu ; JAYARAMAN, Prem P. ; MORSHED, Ahsan ; FORKAN, Abdur Rahim M. ; GEORGAKOPOULOS, Dimitrios ; KANG, Yong-Bin ; PIECHOWSKI, Mirek: A solution for annotating sensor data streams - An industrial use case in building management system. In: *2020 21st IEEE International Conference on Mobile Data Management (MDM)*, Ieee, 62020. – ISBN 978-1-7281-4663-8, S. 194–201
- [300] MAKABATE, Choeu T. ; MUSONDA, Innocent ; OKORO, Chioma S. ; CHILESHE, Nicholas: Scientometric analysis of BIM adoption by SMEs in the architecture, construction and engineering sector. In: *Engineering, Construction and Architectural Management* 29 (2022), Nr. 1, S. 179–203. <http://dx.doi.org/10.1108/ECAM-02-2020-0139>. – DOI 10.1108/ECAM-02-2020-0139. – ISSN 0969-9988
- [301] MAKOWSKI, Dominique ; PHAM, Tam ; LAU, Zen J. ; BRAMMER, Jan C. ; LESPINASSE, François ; PHAM, Hung ; SCHÖLZEL, Christopher ; CHEN, S. H. A.: NeuroKit2: A Python toolbox for neurophysiological signal processing. In: *Behavior research methods* 53 (2021), Nr. 4, S. 1689–1696. <http://dx.doi.org/10.3758/s13428-020-01516-y>. – DOI 10.3758/s13428-020-01516-y
- [302] MALIK, P.: Governing Big Data: Principles and practices. In: *IBM Journal of Research and Development* 57 (2013), Nr. 3/4, S. 1:1–1:13. <http://dx.doi.org/10.1147/JRD.2013.2241359>. – DOI 10.1147/JRD.2013.2241359. – ISSN 0018-8646
- [303] MANDELROT, Benoit: An informational theory of the statistical structure of language. In: *Communication theory* 84 (1953), S. 486–502
- [304] MANIS, George ; AKTARUZZAMAN, Md ; SASSI, Roberto: Bubble Entropy: An Entropy Almost Free of Parameters. In: *IEEE transactions on bio-medical engineering* 64 (2017), Nr. 11, S. 2711–2718. <http://dx.doi.org/10.1109/TBME.2017.2664105>. – DOI 10.1109/TBME.2017.2664105. – ISSN 0018-9294

- [305] MARCILIO, Wilson E. ; ELER, Danilo M.: From explanations to feature selection: assessing SHAP values as feature selection mechanism. In: *2020 33rd SIBGRAPI Conference on Graphics, Patterns and Images*. Piscataway, NJ : Ieee, 2020. – ISBN 978–1–7281–9274–1, S. 340–347
- [306] MARIANO-HERNÁNDEZ, D. ; HERNÁNDEZ-CALLEJO, L. ; ZORITA-LAMADRID, A. ; DUQUE-PÉREZ, O. ; SANTOS GARCÍA, E: A review of strategies for building energy management system: Model predictive control, demand side management, optimization, and fault detect & diagnosis. In: *Journal of Building Engineering* 33 (2021), S. 101692. <http://dx.doi.org/10.1016/j.jobe.2020.101692>. – DOI 10.1016/j.jobe.2020.101692. – ISSN 23527102
- [307] MATARNEH, Sandra T. ; DANSO-AMOAKO, Mark ; AL-BIZRI, Salam ; GATERELL, Mark ; MATARNEH, Rana: Building information modeling for facilities management: A literature review and future research directions. In: *Journal of Building Engineering* 24 (2019), S. 100755. <http://dx.doi.org/10.1016/j.jobe.2019.100755>. – DOI 10.1016/j.jobe.2019.100755. – ISSN 23527102
- [308] MATILLA-GARCIA, Manuel: Selection of embedding dimension and delay time in phase space reconstruction via symbolic dynamics. In: *Entropy* 23 (2021), Nr. 2, S. 221
- [309] MATTERA, Claudio G. ; QUEVEDO, Joseba ; ESCOBET, Teresa ; SHAKER, Hamid R. ; JRADI, Muhyiddine: A method for fault detection and diagnostics in ventilation units using virtual sensors. In: *Sensors* 18 (2018), Nr. 11, S. 3931. – ISSN 1424–8220
- [310] MATTHEWS, B. W.: Comparison of the predicted and observed secondary structure of T4 phage lysozyme. In: *Biochimica et biophysica acta* 405 (1975), Nr. 2, S. 442–451. [http://dx.doi.org/10.1016/0005-2795\(75\)90109-9](http://dx.doi.org/10.1016/0005-2795(75)90109-9). – DOI 10.1016/0005-2795(75)90109-9. – ISSN 0006–3002
- [311] MAVROKAPNIDIS, Dimitris: Enabling scalable deployment of data-driven applications across building portfolios. In: *Proceedings of the 2023 European Conference on Computing in Construction and the 40th International CIB W78 Conference*, European Council for Computing in Construction, 2023 (Computing in Construction)
- [312] MEDIANO, Pedro am ; ROSAS, Fernando E. ; TIMMERMAN, Christopher ; ROSEMAN, Leor ; NUTT, David J. ; FEILDING, Amanda ; KAELEN, Mendel ; KRINGELBACH, Morten L. ; BARRETT, Adam B. ; SETH, Anil K. u. a.: Effects of external stimulation on psychedelic state neurodynamics. In: *ACS Chemical Neuroscience* 15 (2024), Nr. 3, S. 462–471
- [313] MEGGERS, Forrest ; RYSANEK, Adam ; PANTELIC, Jovan ; CHEN, Kian ; GUO, Hongshan ; HOUCOIS, Nicholas ; TEITELBAUM, Eric: *Globe thermometer free convection error potentials*. 2019
- [314] MEHRABI, Mehdi ; YUILL, David: Fouling and its effects on air-cooled condensers in split system air conditioners (RP-1705). In: *Science and Technology for the Built Environment* 25 (2019), Nr. 6, S. 784–793. – ISSN 2374–4731

- [315] MENG, Tong ; JING, Xuyang ; YAN, Zheng ; PEDRYCZ, Witold: A survey on machine learning for data fusion. In: *Information Fusion* 57 (2020), S. 115–129. <http://dx.doi.org/10.1016/j.inffus.2019.12.001>. – DOI 10.1016/j.inffus.2019.12.001. – ISSN 15662535
- [316] MERTENS, Noah ; WILDE, Andreas: Automated Classification of Datapoint Types in Building Automation Systems Using Time Series. In: NOËL, Frédéric (Hrsg.) ; NYFFENEGGER, Felix (Hrsg.) ; RIVEST, Louis (Hrsg.) ; BOURAS, Abdelaziz (Hrsg.): *Product Lifecycle Management. PLM in Transition Times: The Place of Humans and Transformative Technologies*. Cham : Springer Nature Switzerland, 2023. – ISBN 978–3–031–25182–5, S. 495–505
- [317] MICHAEL, Jörg: Doppelgänger gesucht. Ein Programm für die kontextsensitive phonetische Stringumwandlung. In: *c't Magazin für Computer und Technik* 25 (1999), S. 252
- [318] MILLER, Clayton ; NAGY, Zoltán ; SCHLUETER, Arno: A review of unsupervised statistical learning and visual analytics techniques applied to performance analysis of non-residential buildings. In: *Renewable and Sustainable Energy Reviews* (2017). <http://dx.doi.org/10.1016/j.rser.2017.05.124>. – DOI 10.1016/j.rser.2017.05.124. – ISSN 13640321
- [319] MILLER, Holmes: The Multiple Dimensions of Information Quality. In: *Information Systems Management* 13 (1996), Nr. 2, S. 79–82. <http://dx.doi.org/10.1080/10580539608906992>. – DOI 10.1080/10580539608906992. – ISSN 1058–0530
- [320] MILLS, Evan: Building commissioning: a golden opportunity for reducing energy costs and greenhouse gas emissions in the United States. In: *Energy Efficiency* 4 (2011), Nr. 2, S. 145–173. <http://dx.doi.org/10.1007/s12053-011-9116-8>. – DOI 10.1007/s12053-011-9116-8. – ISSN 1570–646X
- [321] MISHRA, Sakshi ; GLAWS, Andrew ; CUTLER, Dylan ; FRANK, Stephen ; AZAM, Muhammad ; MOHAMMADI, Farzam ; VENNE, Jean-Simon: Unified architecture for data-driven metadata tagging of building automation systems. In: *Automation in Construction* 120 (2020), S. 103411. <http://dx.doi.org/10.1016/j.autcon.2020.103411>. – DOI 10.1016/j.autcon.2020.103411. – ISSN 09265805
- [322] MONTAZER, Maryam: *A Concept for an Integrated Building Model for Energy Management*, TU Wien, Diss., 2024. <http://dx.doi.org/10.34726/hss.2024.50354>. – DOI 10.34726/hss.2024.50354
- [323] MORITZ, Philipp ; NISHIHARA, Robert ; WANG, Stephanie ; TUMANOV, Alexey ; LIAW, Richard ; LIANG, Eric ; ELIBOL, Melih ; YANG, Zongheng ; PAUL, William ; JORDAN, Michael I. ; STOICA, Ion: *Ray: A Distributed Framework for Emerging AI Applications*. <http://arxiv.org/pdf/1712.05889>. Version: 2017
- [324] MULLER, Marina F. ; ESMANIOTO, Filipe ; HUBER, Natan ; LOURES, Eduardo R. ; CANGIOLIERI, Osiris: A systematic literature review of interoperability in the green Building Information Modeling lifecycle. In: *Journal of Cleaner Production* 223 (2019), S. 397–412. <http://dx.doi.org/10.1016/j.jclepro.2019.03.101>. – DOI 10.1016/j.jclepro.2019.03.101. – ISSN 0959–6526

- doi.org/10.1016/j.jclepro.2019.03.114. – DOI 10.1016/j.jclepro.2019.03.114. – ISSN 09596526
- [325] MULUMBA, Timothy ; AFSHARI, Afshin ; YAN, Ke ; SHEN, Wen ; NORFORD, Leslie K.: Robust model-based fault diagnosis for air handling units. In: *Energy and Buildings* 86 (2015), S. 698–707. – ISSN 03787788
- [326] MYTTON, David: Hiding greenhouse gas emissions in the cloud. In: *Nature Climate Change* 10 (2020), Nr. 8, S. 701. <http://dx.doi.org/10.1038/s41558-020-0837-6>. – DOI 10.1038/s41558-020-0837-6. – ISSN 1758-6798
- [327] NAJAFI, Massieh ; AUSLANDER, David M. ; BARTLETT, Peter L. ; HAVES, Philip ; SOHN, Michael D.: Application of machine learning in the fault diagnostics of air handling units. In: *Applied Energy* 96 (2012), S. 347–358. – ISSN 03062619
- [328] NAJAFI, Massieh ; AUSLANDER, David M. ; HAVES, Philip ; SOHN, Michael D.: A statistical pattern analysis framework for rooftop unit diagnostics. In: *HVAC&R Research* 18 (2012), Nr. 3, S. 406–416. – ISSN 1078-9669
- [329] NAJEH, Houda ; SINGH, Mahendra P. ; PLOIX, Stephane: Faults and Failures in Smart Buildings: A New Tool for Diagnosis. Version: 2021. http://dx.doi.org/10.1007/978-3-030-76477-7_14. In: PLOIX, Stephane (Hrsg.) ; AMAYRI, Manar (Hrsg.) ; BOUGUILA, Nizar (Hrsg.): *Towards Energy Smart Homes*. Cham : Springer International Publishing and Imprint Springer, 2021. – DOI 10.1007/978-3-030-76477-7_14. – ISBN 978-3-030-76476-0, S. 433–462
- [330] NASH, John: Non-cooperative games. In: *Annals of mathematics* (1951), S. 286–295
- [331] NAUMANN, Felix: *Similarity measures*. https://hpi.de/fileadmin/user_upload/fachgebiete/naumann/folien/SS13/DPDC/DPDC_12_Similarity.pdf. Version: 2013
- [332] NEHMADI, Y. ; GUTERMAN, H.: A comparison of neural network and Bayes recognition approaches to the diagnostic of multiple sclerosis. In: *Program and papers / The Eighteenth Convention of Electrical and Electronics Engineers in Israel, Tel Aviv, March 7-8, 1995*. Piscataway, NJ : IEEE Service Center, 1995. – ISBN 0-7803-2498-6, S. 1.2.3/1–1.2.3/5
- [333] NEUMAYER, Martin ; STECHER, Dominik ; GRIMM, Sebastian ; MAIER, Andreas ; BÜCKER, Dominikus ; SCHMIDT, Jochen: Fault and anomaly detection in district heating substations: A survey on methodology and data sets. In: *Energy* 276 (2023), S. 127569. <http://dx.doi.org/10.1016/j.energy.2023.127569>. – DOI 10.1016/j.energy.2023.127569. – ISSN 03605442
- [334] NEWMAN, M. E.: Power laws, Pareto distributions and Zipf's law. In: *Contemporary Physics* 46 (2005), Nr. 5, 323–351. <http://dx.doi.org/10.1080/00107510500052444>. – DOI 10.1080/00107510500052444. – ISSN 0010-7514

- [335] NIKLAUS, Joel: *CompoundSplit*. <https://github.com/JoelNiklaus/CompoundSplit>. Version: 2020
- [336] NORDHAUS, William: Climate Change: The Ultimate Challenge for Economics. In: *American Economic Review* 109 (2019), Nr. 6, S. 1991–2014. <http://dx.doi.org/10.1257/aer.109.6.1991>. – DOI 10.1257/aer.109.6.1991. – ISSN 0002–8282
- [337] O'GRADY, Timothy ; CHONG, Heap-Yih ; MORRISON, Gregory M.: A systematic review and meta-analysis of building automation systems. In: *Building and Environment* 195 (2021), S. 107770. <http://dx.doi.org/10.1016/j.buildenv.2021.107770>. – DOI 10.1016/j.buildenv.2021.107770. – ISSN 0360–1323
- [338] OLIVEIRA, Paulo ; RODRIGUES, Fátima ; HENRIQUES, Pedro R.: A formal definition of data quality problems. In: *ICIQ*, 2005
- [339] OMIDVARNIA, Amir ; MESBAH, Mostefa ; PEDERSEN, Mangor ; JACKSON, Graeme: Range entropy: A bridge between signal complexity and self-similarity. In: *Entropy* 20 (2018), Nr. 12, S. 962
- [340] O'NEILL, Zheng ; LI, Yanfei ; WILLIAMS, Keith: HVAC control loop performance assessment: A critical review (1587-RP). In: *Science and Technology for the Built Environment* 23 (2017), Nr. 4, S. 619–636. <http://dx.doi.org/10.1080/23744731.2016.1239466>. – DOI 10.1080/23744731.2016.1239466. – ISSN 2374–4731
- [341] OPENAI: *OpenAI Official Website: Introducing ChatGPT - Learn more*. <https://openai.com/blog/chatgpt/>. Version: 30.11.2022
- [342] PADILLA, Miguel ; CHOINIERE, Daniel: A combined passive-active sensor fault detection and isolation approach for air handling units. In: *Energy and Buildings* 99 (2015), S. 214–219. – ISSN 03787788
- [343] PAIHO, Satu ; SAASTAMOINEN, Heidi ; KARJALAINEN, Sami: Promoting energy services in offices: How do office workers see their role? In: *Journal of Building Engineering* 33 (2021), S. 101602. <http://dx.doi.org/10.1016/j.jobee.2020.101602>. – DOI 10.1016/j.jobee.2020.101602. – ISSN 23527102
- [344] PANG, Zhihong ; CHEN, Yan ; ZHANG, Jian ; O'NEILL, Zheng ; CHENG, Hwakong ; DONG, Bing: Nationwide HVAC energy-saving potential quantification for office buildings with occupant-centric controls in various climates. In: *Applied Energy* 279 (2020), S. 115727. <http://dx.doi.org/10.1016/j.apenergy.2020.115727>. – DOI 10.1016/j.apenergy.2020.115727. – ISSN 03062619
- [345] PARHIZKAR, Tarannom ; ARAMOUN, Fereshteh ; SABOOHI, Yadollah: Efficient health monitoring of buildings using failure modes and effects analysis case study: air handling unit system. In: *Journal of Building Engineering* 29 (2020), S. 101113. – ISSN 23527102

-
- [346] PARK, Hoonseok ; JUNG, Jae-Yoon: SAX-ARM: Deviant event pattern discovery from multi-variate time series using symbolic aggregate approximation and association rule mining. In: *Expert Systems with Applications* 141 (2020), S. 112950. <http://dx.doi.org/10.1016/j.eswa.2019.112950>. – DOI 10.1016/j.eswa.2019.112950. – ISSN 09574174
 - [347] PARK, June Y. ; LASTERNAS, Bertrand ; AZIZ, Azizan: Data-Driven Framework to Find the Physical Association between AHU and VAV Terminal Unit-Pilot Study. In: *2018 ASHRAE Winter Conference*
 - [348] PARK, June Y. ; OUF, Mohamed M. ; GUNAY, Burak ; PENG, Yuzhen ; O'BRIEN, William ; KJÆRGAARD, Mikkel B. ; NAGY, Zoltan: A critical review of field implementations of occupant-centric building controls. In: *Building and Environment* 165 (2019), S. 106351. <http://dx.doi.org/10.1016/j.buildenv.2019.106351>. – DOI 10.1016/j.buildenv.2019.106351. – ISSN 0360-1323
 - [349] PARK, Unkyu ; SILVA, Fabio ; HOU, Mengzhao ; HEIDEMANN, John ; GUILIANO, Genevive ; WANG, Xi ; PRASHAR, Nikhil: Single-and multi-sensor techniques to improve accuracy of urban vehicle classification. In: *ISI-TR* (2006), 13. <https://ant.isi.edu/~johnh/PAPERS/Park06a.pdf>
 - [350] PARVIN, K. ; LIPU, M. S. H. ; HANNAN, M. A. ; ABDULLAH, Majid A. ; JERN, Ker P. ; BEGUM, R. A. ; MANSUR, Muhamad ; MUTTAQI, Kashem M. ; MAHLIA, T. M. I. ; DONG, Zhao Y.: Intelligent Controllers and Optimization Algorithms for Building Energy Management Towards Achieving Sustainable Development: Challenges and Prospects. In: *IEEE Access* 9 (2021), S. 41577–41602. <http://dx.doi.org/10.1109/ACCESS.2021.3065087>. – DOI 10.1109/ACCESS.2021.3065087. – ISSN 2169-3536
 - [351] PEARSON, K.: Notes on regression and inheritance in the case of two parents proceedings of the royal society of london, 58, 240-242. In: *K Pearson* (1895)
 - [352] PEDREGOSA, Fabian ; VAROQUAUX, Gaël ; GRAMFORT, Alexandre ; MICHEL, Vincent ; THIRION, Bertrand ; GRISEL, Olivier ; BLONDEL, Mathieu ; MÜLLER, Andreas ; NOTHMAN, Joel ; LOUPPE, Gilles ; PRETTENHOFER, Peter ; WEISS, Ron ; DUBOURG, Vincent ; VANDERPLAS, Jake ; PASSOS, Alexandre ; COURNAPEAU, David ; BRUCHER, Matthieu ; PERROT, Matthieu ; DUCHESNAY, Édouard: Scikit-learn: Machine Learning in Python. (2011). <http://arxiv.org/pdf/1201.0490v4>
 - [353] PENNINGTON, Jeffrey ; SOCHER, Richard ; MANNING, Christopher D.: GloVe: Global Vectors for Word Representation. In: *Empirical Methods in Natural Language Processing (EMNLP)*, 2014, 1532–1543
 - [354] PEREIRA, Rui ; COUTO, Marco ; RIBEIRO, Francisco ; RUA, Rui ; CUNHA, Jácome ; FERNANDES, Joao P. ; SARAIVA, João: Energy efficiency across programming languages: how do energy, time, and memory relate? In: COMBEMALE, Benoit (Hrsg.): *Proceedings of the 10th ACM SIG-*

- PLAN International Conference on Software Language Engineering*. New York, NY : ACM, 2017. – ISBN 9781450355254, S. 256–267
- [355] PETROSIAN, Arthur: Kolmogorov complexity of finite sequences and recognition of different preictal EEG patterns. In: *Proceedings eighth IEEE symposium on computer-based medical systems*, 1995, S. 212–217
- [356] PINCUS, S. M.: Approximate entropy as a measure of system complexity. In: *Proceedings of the National Academy of Sciences of the United States of America* 88 (1991), Nr. 6, S. 2297–2301. <http://dx.doi.org/10.1073/pnas.88.6.2297>. – DOI 10.1073/pnas.88.6.2297. – ISSN 0027–8424
- [357] PLOENNIGS, Joern: Automating Analytics: How to Learn Metadata Such That Our Buildings Can Learn from Us. In: *2016 IEEE International Conference on Sensing, Communication and Networking (SECON Workshops)*. Piscataway, NJ : Ieee, 2016. – ISBN 978–1–5090–2429–2, S. 1–6
- [358] PLOENNIGS, Joern ; GORMAN, Bernard ; BRADY, Niall ; SCHUMANN, Anika: Demo Abstract: BEAD - Building Energy Asset Discovery Tool for Automating Smart Building Analytics. Version: 2014. <http://dx.doi.org/10.1145/2674061.2675030>. In: *BuildSys '14*. 2014. – DOI 10.1145/2674061.2675030
- [359] PÖRTNER, H. O. (Hrsg.) ; ROBERTS, D. C. (Hrsg.) ; TIGNOR, M. (Hrsg.) ; POLOCZANSKA, E. S. (Hrsg.) ; MINTENBECK, K. (Hrsg.) ; ALEGRÍA, A. (Hrsg.) ; CRAIG, M. (Hrsg.) ; LANGSDORF, S. (Hrsg.) ; LÖSCHKE, S. (Hrsg.) ; MÖLLER, V. (Hrsg.) ; OKEM, A. (Hrsg.) ; RAMA, B. (Hrsg.): *Climate Change 2022: Impacts, Adaptation and Vulnerability. Contribution of Working Group II to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge, UK and New York, USA : Cambridge University Press, 2022. – ISBN 9781009325844
- [360] POSTEL, Hans J.: Die Kölner Phonetik. Ein Verfahren zur Identifizierung von Personennamen auf der Grundlage der Gestaltanalyse. In: *IBM-Nachrichten* 19 (1969), S. 925–931
- [361] POURZOLFAGHAR, Zohreh ; HELFERT, Markus: Utilising Latent Data in Smart Buildings: A Process Model to Collect, Analyse and Make Building Data Accessible for Smart Industries. In: *Position Papers of the 2017 Federated Conference on Computer Science and Information Systems*, PTI, 2017 (Annals of Computer Science and Information Systems), S. 195–201
- [362] POVEDA-VILLALÓN, María ; GARCIA-CASTRO, Raúl: Extending the SAREF ontology for building devices and topology. In: *Proceedings of the 6th Linked Data in Architecture and Construction Workshop (LDAC 2018)*, Vol. CEUR-WS Bd. 2159, 2018, 16–23
- [363] POWERS, David ; AILAB: Evaluation: From precision, recall and F-measure to ROC, informedness, markedness & correlation. In: *J. Mach. Learn. Technol* 2 (2011), S. 2229–3981. <http://dx.doi.org/10.9735/2229-3981>. – DOI 10.9735/2229–3981

- [364] PRITONI, Marco ; BHATTACHARYA, Arka A. ; CULLER, David ; MODERA, Mark: Short Paper: A Method for Discovering Functional Relationships Between Air Handling Units and Variable-Air-Volume Boxes From Sensor Data. In: CULLER, David (Hrsg.) ; AGARWAL, Yuvraj (Hrsg.) ; MANGHARAM, Rahul (Hrsg.): *Proceedings of the 2nd ACM International Conference on Embedded Systems for Energy-Efficient Built Environments*. New York, NY : ACM, 2015. – ISBN 9781450339810, S. 133–136
- [365] PRITONI, Marco ; LIN, Guanjin ; CHEN, Yimin ; VITTI, Raphael ; WEYANDT, Christopher ; GRANDERSON, Jessica: From fault-detection to automated fault correction: A field study. In: *Building and Environment* (2022), S. 108900. <http://dx.doi.org/10.1016/j.buildenv.2022.108900>. – DOI 10.1016/j.buildenv.2022.108900. – ISSN 0360–1323
- [366] PRITONI, Marco ; LIN, Guanjin ; CHEN, Yimin ; HOUSE, John ; CROWE, Eliot ; JESSICA, Granderson: *Market Barriers and Drivers for the Next Generation Fault Detection and Diagnostic Tools*
- [367] PRITONI, Marco ; PAINE, Drew ; FIERRO, Gabriel ; MOSIMAN, Cory ; POPLAWSKI, Michael ; SAHA, Avijit ; BENDER, Joel ; GRANDERSON, Jessica: Metadata Schemas and Ontologies for Building Energy Applications: A Critical Review and Use Case Analysis. In: *Energies* 14 (2021), Nr. 7, S. 2024. <http://dx.doi.org/10.3390/en14072024>. – DOI 10.3390/en14072024. – ISSN 1996–1073
- [368] PRÍVARA, Samuel ; CIGLER, Jiří ; VÁŇA, Zdeněk ; OLDEWURTEL, Frauke ; SAGERSCHNIG, Carina ; ŽÁČEKOVÁ, Eva: Building modeling as a crucial part for building predictive control. In: *Energy and Buildings* 56 (2013), S. 8–22. <http://dx.doi.org/10.1016/j.enbuild.2012.10.024>. – DOI 10.1016/j.enbuild.2012.10.024. – ISSN 03787788
- [369] PROJECT HAYSTACK: *Project Haystack*. <https://project-haystack.org/>. Version: 2022
- [370] PROKHORENKOVA, Liudmila ; GUSEV, Gleb ; VOROBEOV, Aleksandr ; DOROGUSH, Anna V. ; GULIN, Andrey: *CatBoost: unbiased boosting with categorical features*. <https://arxiv.org/pdf/1706.09516>. Version: 2017
- [371] QUASTHOFF, Uwe ; RICHTER, Matthias: Projekt Deutscher Wortschatz. Version: 2005. http://babylonia.ch/fileadmin/user_upload/documents/2005-3/Baby3_05x.pdf. In: RÜDIGER-HARPER, Jean (Hrsg.): *Babylonia 3*. 2005
- [372] QUINN, Caroline ; MCARTHUR, J. J.: A case study comparing the completeness and expressiveness of two industry recognized ontologies. In: *Advanced Engineering Informatics* 47 (2021), 101233. <http://dx.doi.org/10.1016/j.aei.2020.101233>. – DOI 10.1016/j.aei.2020.101233. – ISSN 14740346
- [373] QUINN, Caroline ; SHABESTARI, Ali Z. ; MISIC, Tony ; GILANI, Sara ; LITOIU, Marin ; MCARTHUR, J. J.: Building automation system - BIM integration using a linked data structure. In: *Automation in Construction* 118 (2020), S. 103257. <http://dx.doi.org/10.1016/j.autcon.2020.103257>. – DOI 10.1016/j.autcon.2020.103257. – ISSN 09265805

- [374] RAFTERY, Paul ; VERNON, David ; SINGLA, Rupam ; NAKAJIMA, Mia: Measured Space Heating Hot Water Distribution Losses in Large Commercial Buildings. (2023). <https://escholarship.org/uc/item/46h4h28q#author>
- [375] RAHMANI ASL, Mohammad ; ZARRINMEHR, Saied ; BERGIN, Michael ; YAN, Wei: BPOpt: A framework for BIM-based performance optimization. In: *Energy and Buildings* 108 (2015), S. 401–412. <http://dx.doi.org/10.1016/j.enbuild.2015.09.011>. – DOI 10.1016/j.enbuild.2015.09.011. – ISSN 03787788
- [376] RAMANATHAN, Ganesh ; HUSMANN, Maria ; NIEDERMEIER, Christoph ; VICARI, Norbert ; GARCIA, Kimberly ; MAYER, Simon: Assisting automated fault detection and diagnostics in building automation through semantic description of functions and process data. In: JIANG, Xiaofan (Hrsg.) ; GNAWALI, Omprakash (Hrsg.) ; NAGY, Zoltan (Hrsg.): *Proceedings of the 8th ACM International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation*. New York, NY, USA : ACM, 2021. – ISBN 9781450391146, S. 228–229
- [377] RAMSEIER, Livia ; FRISCHKNECHT, Rolf: Umwelfußabdruck von Gebäuden in Deutschland. (2020)
- [378] RANA, Mashud ; RAHMAN, Ashfaqur ; ALMASHOR, Mahathir ; MCCULLOCH, John ; SETHUVENKATRAMAN, Subbu: Automatic Classification of Sensors in Buildings: Learning from Time Series Data. Version: 2024. http://dx.doi.org/10.1007/978-981-99-8388-9_30. In: LIU, Tongliang (Hrsg.) ; WEBB, Geoff (Hrsg.) ; YUE, Lin (Hrsg.) ; WANG, Dadong (Hrsg.): *AI 2023: Advances in Artificial Intelligence* Bd. 14471. Singapore : Springer Nature Singapore, 2024. – DOI 10.1007/978-981-99-8388-9_30. – ISBN 978-981-99-8387-2, S. 367–378
- [379] RANADE, Akshay ; PROVAN, Gregory ; MADY, Alie El-Din ; O’SULLIVAN, Dominic: A computationally efficient method for fault diagnosis of fan-coil unit terminals in building Heating Ventilation and Air Conditioning systems. In: *Journal of Building Engineering* 27 (2020), S. 100955. – ISSN 23527102
- [380] RAO, M. ; CHEN, Y. ; VEMURI, B. C. ; WANG, F.: Cumulative Residual Entropy: A New Measure of Information. In: *IEEE Transactions on Information Theory* 50 (2004), Nr. 6, S. 1220–1228. <http://dx.doi.org/10.1109/TIT.2004.828057>. – DOI 10.1109/TIT.2004.828057. – ISSN 0018-9448
- [381] REEVE, Hayden ; WAGNER, Tim ; XIONG, Ziyou: *Volttron application for assisted discovery of semantic meaning of BAS points*. <http://cbei.psu.edu/wp-content/uploads/2016/07/Volttron-application-for-assisted-discovery.pdf>. Version: 2016
- [382] RÉHAULT, Nicolas ; OHR, Felix ; ZEHNLÉ, Sebastian ; MÜLLER, Thorsten ; RIST, Tim ; JACOB, Dirk ; LICHTENBERG, Gerwald ; PANGALOS, Georg ; KRUPPA, Kai ; SCHMIDT, Fritz ; ZUZEL, Arkadius ; HARMSSEN, Arnold ; SEWE, Erik: *Projektbericht: Modellbasierte Qualitätssicherung des energetischen Gebäudebetriebs (ModQS): Abschlussbericht Förderkenn-*

- zeichen 0327893. http://www.modqs.de/uploads/tx_buildingeq/20140703_ModQS_Abschlussbericht_rev1.pdf. Version: 2013
- [383] REISS & ROMMERICH: *Berufspolitische Befragung 2021, Teil 7: BIM (Building Information Modeling)*. https://bak.de/wp-content/uploads/2021/10/2021_BAK_Berufspolitische-Befragung_Bericht_Themenblock-7_BIM.pdf. Version: 2021
- [384] RÉNYI, Alfréd: On measures of entropy and information. In: *Proceedings of the fourth Berkeley symposium on mathematical statistics and probability, volume 1: contributions to the theory of statistics* Bd. 4, 1961, S. 547–562
- [385] REPODIAC: *german_compound_splitter*. https://github.com/repodiac/german_compound_splitter. Version: 2020
- [386] RESENDES, Sílvia ; CARREIRA, Paulo ; SANTOS, André C.: Conflict detection and resolution in home and building automation systems: a literature review. In: *Journal of Ambient Intelligence and Humanized Computing* 5 (2014), Nr. 5, S. 699–715. <http://dx.doi.org/10.1007/s12652-013-0184-9>. – DOI 10.1007/s12652-013-0184-9. – ISSN 1868-5137
- [387] RETH, H. P. ; SCHEK, H. J.: *Eine Zugriffsmethode für die phonetische Ähnlichkeitssuche*. Wiss. Zentrum, 1977 (Heidelberg Scientific Center Technical reports). <https://books.google.de/books?id=9aQavwEACAAJ>
- [388] RIZZO, Maria L. ; SZÉKELY, Gábor J.: Energy distance. In: *WIREs Computational Statistics* 8 (2016), Nr. 1, S. 27–38. <http://dx.doi.org/10.1002/wics.1375>. – DOI 10.1002/wics.1375. – ISSN 1939-5108
- [389] ROBERTS, D. C. ; TURCOTTE, D. L.: Fractality and Self-Organized Criticality of Wars. In: *Fractals* 06 (1998), Nr. 04, S. 351–357. <http://dx.doi.org/10.1142/S0218348X98000407>. – DOI 10.1142/S0218348X98000407. – ISSN 0218-348X
- [390] ROBERTS, Stephen J. ; PENNY, William ; REZEK, Iead: Temporal and spatial complexity measures for electroencephalogram based brain-computer interfacing. In: *Medical & biological engineering & computing* 37 (1999), S. 93–98
- [391] ROHILA, Ashish ; SHARMA, Ambalika: Phase entropy: A new complexity measure for heart rate variability. In: *Physiological measurement* 40 (2019), Nr. 10, S. 105006
- [392] ROHWEDER, Jan P. ; KASTEN, Gerhard ; MALZAHN, Dirk ; PIRO, Andrea ; SCHMID, Joachim: Informationsqualität – Definitionen, Dimensionen und Begriffe. Version: 2021. http://dx.doi.org/10.1007/978-3-658-30991-6_2. In: HILDEBRAND, Knut (Hrsg.) ; GEBAUER, Marcus (Hrsg.) ; MIELKE, Michael (Hrsg.): *Daten- und Informationsqualität*. Wiesbaden and Heidelberg : Springer Vieweg, 2021. – DOI 10.1007/978-3-658-30991-6_2. – ISBN 978-3-658-30990-9, S. 23–43
- [393] RONTHALER, Marc: *Dialogschnittstellen an Online-Informationssystemen: Notwendigkeit,*

- Leistungsfähigkeit und Entwicklungsmöglichkeiten am Beispiel des OSIRIS-Systems*, Ph. D. thesis, Universität Osnabrück, Osnabrück, Germany, Doktorarbeit, 2000
- [394] ROSENBERG, Andrew ; BINKOWSKI, Ed: Augmenting the kappa statistic to determine inter-annotator reliability for multiply labeled data points. In: *Proceedings of HLT-NAACL 2004: Short Papers*. Boston, Massachusetts, USA : Association for Computational Linguistics, 2004, 77–80
- [395] ROSENSTEIN, Michael T. ; COLLINS, James J. ; LUCA, Carlo J.: A practical method for calculating largest Lyapunov exponents from small data sets. In: *Physica D: Nonlinear Phenomena* 65 (1993), Nr. 1-2, S. 117–134
- [396] ROSTAGHI, Mostafa ; AZAMI, Hamed: Dispersion Entropy: A Measure for Time-Series Analysis. In: *IEEE Signal Processing Letters* 23 (2016), Nr. 5, S. 610–614. <http://dx.doi.org/10.1109/LSP.2016.2542881>. – DOI 10.1109/LSP.2016.2542881. – ISSN 1070–9908
- [397] ROTGER-GRIFUL, Sergi ; WELLING, Ubbe ; JACOBSEN, Rune H.: Implementation of a building energy management system for residential demand response. In: *Microprocessors and Microsystems* 55 (2017), S. 100–110. <http://dx.doi.org/10.1016/j.micpro.2017.10.006>. – DOI 10.1016/j.micpro.2017.10.006. – ISSN 01419331
- [398] ROTH, Kurt ; WESTPHALEN, Detlef ; FENG, Michael ; LLANA, Patricia: Energy Impact of Commercial Building Controls and Performance Diagnostics: Market Characterization, Energy Impact of Building Faults and Energy Savings Potential. (2005)
- [399] SAGIROGLU, Seref ; TERZI, Ramazan ; CANBAY, Yavuz ; COLAK, Ilhami: Big data issues in smart grid systems. In: APPLICATIONS, International Conference on Renewable Energy Research a. (Hrsg.): *The 5th IEEE International Conference on Renewable Energy Research and Applications (ICRERA 2016)*. [Piscataway, NJ] : Ieee, 2016. – ISBN 978–1–5090–3388–1, S. 1007–1012
- [400] SAYED, Aya N. ; HIMEUR, Yassine ; BENSAAALI, Faycal: Deep and transfer learning for building occupancy detection: A review and comparative analysis. In: *Engineering Applications of Artificial Intelligence* 115 (2022), S. 105254. <http://dx.doi.org/10.1016/j.engappai.2022.105254>. – DOI 10.1016/j.engappai.2022.105254. – ISSN 09521976
- [401] SCHÄFER, Patrick ; HÖGQVIST, Mikael: SFA: a symbolic fourier approximation and index for similarity search in high dimensional datasets. In: RUNDENSTEINER, Elke (Hrsg.): *Proceedings of the 15th International Conference on Extending Database Technology*. New York, NY : ACM, 2012 (ACM Other conferences). – ISBN 9781450307901, S. 516–527
- [402] SCHERER, Raimar J. ; KATRANUSCHKOV, Peter: BIMification: How to create and use BIM for retrofitting. In: *Advanced Engineering Informatics* 38 (2018), S. 54–66. <http://dx.doi.org/10.1016/j.aei.2018.05.007>. – DOI 10.1016/j.aei.2018.05.007. – ISSN 14740346
- [403] SCHLEUTER, D. ; DAUFRESNE, M. ; MASSOL, F. ; ARGILLIER, C.: A user's guide to functional diversity indices. In: *Ecological Monographs* 80 (2010), Nr. 3, S. 469–484. <http://dx.doi.org/10.1890/08-2225.1>. – DOI 10.1890/08-2225.1. – ISSN 0012–9615

-
- [404] SCHREIBER, Jan: *Free German Dictionary*. <https://sourceforge.net/projects/germandict/>. Version: 2021
- [405] SCHUMANN, Anika ; PLOENNIGS, Joern ; GORMAN, Bernard: Towards automating the deployment of energy saving approaches in buildings. In: SRIVASTAVA, Mani (Hrsg.): *Proceedings of the 1st ACM Conference on Embedded Systems for Energy-Efficient Buildings - BuildSys '14*. New York, New York, USA : ACM Press, 2014. – ISBN 9781450331449, S. 164–167
- [406] SELENIUM: *Selenium: Webdriver Documentation*. <https://www.selenium.dev/documentation/webdriver/>. Version: 2023
- [407] SEVCIK, Carlos: A procedure to estimate the fractal dimension of waveforms. In: *arXiv pre-print arXiv:1003.5266* (2010)
- [408] SHAHCHERAGHIAN, Amir ; MADANI, Hatef ; ILINCA, Adrian: From White to Black-Box Models: A Review of Simulation Tools for Building Energy Management and Their Application in Consulting Practices. In: *Energies* 17 (2024), Nr. 2, S. 376. <http://dx.doi.org/10.3390/en17020376>. – DOI 10.3390/en17020376. – ISSN 1996–1073
- [409] SHAHNAZARI, Hadi ; MHASKAR, Prashant ; HOUSE, John M. ; SALSBUURY, Timothy I.: Heating, ventilation and air conditioning systems: Fault detection and isolation and safe parking. In: *Computers & Chemical Engineering* 108 (2018), S. 139–151. – ISSN 00981354
- [410] SHAHNAZARI, Hadi ; MHASKAR, Prashant ; HOUSE, John M. ; SALSBUURY, Timothy I.: Distributed fault diagnosis of heating, ventilation, and air conditioning systems. In: *AIChE Journal* 65 (2019), Nr. 2, S. 640–651
- [411] SHAHNAZARI, Hadi ; MHASKAR, Prashant ; HOUSE, John M. ; SALSBUURY, Timothy I.: Modeling and fault diagnosis design for HVAC systems using recurrent neural networks. In: *Computers & Chemical Engineering* 126 (2019), S. 189–203. – ISSN 00981354
- [412] SHANNON, Claude E.: A mathematical theory of communication. In: *The Bell system technical journal* 27 (1948), Nr. 3, 379–423. <http://people.math.harvard.edu/~ctm/home/text/others/shannon/entropy/entropy.pdf>
- [413] SHAW, Conor ; ANDRADE PEREIRA, Flávia de ; McNALLY, Ciaran ; FARGHALY, Karim ; HARTMANN, Timo ; O'DONNELL, James: Information management in the facilities domain: investigating practitioner priorities. In: *Facilities* (2022). <http://dx.doi.org/10.1108/F-02-2022-0033>. – DOI 10.1108/F-02-2022-0033. – ISSN 0263–2772
- [414] SHEARER, Colin: The CRISP-DM model: the new blueprint for data mining. In: *Journal of data warehousing* 5 (2000), Nr. 4, S. 13–22
- [415] SHI, Zixiao ; O'BRIEN, William: Development and implementation of automated fault detection and diagnostics for building systems: A review. In: *Automation in Construction* 104 (2019), S. 215–229. <http://dx.doi.org/10.1016/j.autcon.2019.04.002>. – DOI 10.1016/j.autcon.2019.04.002. – ISSN 09265805

- [416] SHOLAHUDIN, S. ; HAN, Hwataik: Simplified dynamic neural network model to predict heating load of a building using Taguchi method. In: *Energy* 115 (2016), S. 1672–1678. <http://dx.doi.org/10.1016/j.energy.2016.03.057>. – DOI 10.1016/j.energy.2016.03.057. – ISSN 03605442
- [417] SIDI, Fatimah ; SHARIAT PANAHY, Payam H. ; AFFENDEY, Lilly S. ; JABAR, Marzanah A. ; IBRAHIM, Hamidah ; MUSTAPHA, Aida: Data quality: A survey of data quality dimensions. In: *2012 International Conference on Information Retrieval & Knowledge Management*, Ieee, 2012. – ISBN 978–1–4673–1091–8, S. 300–304
- [418] SILVOLA, Risto ; HARKONEN, Janne ; VILPPOLA, Olli ; VEHKAPERÄ, Hanna K. ; HAAPASALO, Harri: Data quality assessment and improvement. In: *International Journal of Business Information Systems* 22 (2016), Nr. 1, S. 62. <http://dx.doi.org/10.1504/IJBIS.2016.075718>. – DOI 10.1504/IJBIS.2016.075718. – ISSN 1746–0972
- [419] SINGER-VINE, Jeremy ; JAIN, Samkit: *pdfplumber*. <https://github.com/jsvine/pdfplumber>. Version: 2022
- [420] SKANDHAKUMAR, Nimalaprakasan ; SALIM, Farzad ; REID, Jason ; DROGEMULLER, Robin ; DAWSON, Ed: Graph theory based representation of building information models for access control applications. In: *Automation in Construction* 68 (2016), S. 44–51. <http://dx.doi.org/10.1016/j.autcon.2016.04.001>. – DOI 10.1016/j.autcon.2016.04.001. – ISSN 09265805
- [421] SMALL, Melvin ; SINGER, J. D.: *Resort to arms : international and civil wars, 1816-1980*. [2nd ed.]. Beverly Hills, Calif. : Sage Publications Beverly Hills, Calif, 1982. – ISBN 0803917767
- [422] SMITH, E. A. ; SENTER, R. J.: Automated readability index. In: *AMRL-TR. Aerospace Medical Research Laboratories (U.S.)* (1967), S. 1–14
- [423] SONG, Li ; WANG, Gang: Using a hybrid method to construct a computational efficient cooling coil model for an automated single-duct variable air volume system fault detection and diagnosis. In: *Energy and Buildings* 92 (2015), S. 363–373. – ISSN 03787788
- [424] SOUNDARARAJAN, Kamal ; HO, Hiang K. ; SU, Bin: Sankey diagram framework for energy and exergy flows. In: *Applied Energy* 136 (2014), S. 1035–1042. <http://dx.doi.org/10.1016/j.apenergy.2014.08.070>. – DOI 10.1016/j.apenergy.2014.08.070. – ISSN 03062619
- [425] SPEKTRUM: *Lexikon der Psychologie - Diversifikationsquotient*. <https://www.spektrum.de/lexikon/psychologie/diversifikationsquotient/3560>. Version: 2000
- [426] SPRING, Ryan ; JOHNSON, Matthew: The possibility of improving automated calculation of measures of lexical richness for EFL writing: A comparison of the LCA, NLTK and SpaCy tools. In: *System* 106 (2022), S. 102770. <http://dx.doi.org/10.1016/j.system.2022.102770>. – DOI 10.1016/j.system.2022.102770. – ISSN 0346251X

-
- [427] STINNER, Florian: *Entwicklung einer simulativen Testumgebung zur Einstellung von typischen gebäudetechnischen Regelkreisen und Anwendung am E.ON-ERC-Hauptgebäude*. Aachen, RWTH Aachen, Masterarbeit, 2015
- [428] STINNER, Florian ; BARANSKI, Marc ; YANG, Yingying ; MÜLLER, Dirk: Anwendung eines Virtuellen Labors zur Integration von virtuellen Sensoren in Bestandsgebäude. In: *DKV-Tagung 2018*. 2018
- [429] STINNER, Florian ; GORGIS, David ; KÜMPEL, Alexander ; MÜLLER, Dirk: Automatic detection of control loops in existing buildings. In: *Journal of Physics: Conference Series* 2600 (2023), Nr. 8, S. 082013. <http://dx.doi.org/10.1088/1742-6596/2600/8/082013>. – DOI 10.1088/1742-6596/2600/8/082013. – ISSN 1742-6588
- [430] STINNER, Florian ; KORNAS, Alina ; BARANSKI, Marc ; MÜLLER, Dirk: Structuring building monitoring and automation system data. Version: 2018. https://www.rehva.eu/fileadmin/user_upload/10-15_RJ1804_WEB.pdf. In: REHVA (Hrsg.): *The REHVA European HVAC Journal - August 2018*. 2018 (REHVA Journal), 10–15
- [431] STINNER, Florian ; KÜMPEL, Alexander ; MÜLLER, Dirk: Comparison of supervised algorithms for automated data analysis in existing buildings. In: *Journal of Physics: Conference Series* 2600 (2023), Nr. 8, S. 082026. <http://dx.doi.org/10.1088/1742-6596/2600/8/082026>. – DOI 10.1088/1742-6596/2600/8/082026. – ISSN 1742-6588
- [432] STINNER, Florian ; LLOPIS-MENGUAL, Belén ; STOREK, Thomas ; KÜMPEL, Alexander ; MÜLLER, Dirk: Comparative study of supervised algorithms for topology detection of sensor networks in building energy systems. In: *Automation in Construction* 138 (2022), S. 104248. <http://dx.doi.org/10.1016/j.autcon.2022.104248>. – DOI 10.1016/j.autcon.2022.104248. – ISSN 09265805
- [433] STINNER, Florian ; NEISSER-DEITERS, Paul ; BARANSKI, Marc ; MÜLLER, Dirk: Aikido: Structuring data point identifiers of technical building equipment by machine learning. In: *Journal of Physics: Conference Series* 1343 (2019), 012039. <http://dx.doi.org/10.1088/1742-6596/1343/1/012039>. – DOI 10.1088/1742-6596/1343/1/012039. – ISSN 1742-6588
- [434] STINNER, Florian ; RASSPE-LANGE, Lukas ; BARANSKI, Marc ; MÜLLER, Dirk: Takeshi: Application of unsupervised machine learning techniques for topology detection in building energy systems. In: *Journal of Physics: Conference Series* 1343 (2019), 012041. <http://dx.doi.org/10.1088/1742-6596/1343/1/012041>. – DOI 10.1088/1742-6596/1343/1/012041. – ISSN 1742-6588
- [435] STINNER, Florian ; WIECEK, Martin ; KÜMPEL, Alexander ; BARANSKI, Marc ; MÜLLER, Dirk: Automatic digital twin data model generation of building energy systems from piping and instrumentation diagrams. In: *Proceedings of ECOS 2021*, 2021
- [436] STOFFEL, Phillip: *Learning Strategies for Data-Driven Model Predictive Control of Building Energy Systems*, RWTH Aachen University, Dissertation, 2024

- [437] SYNNERGREN, Jane ; GAMALIELSSON, Jonas ; OLSSON, Bjorn: Mapping of the JDL data fusion model to bioinformatics. In: *2007 IEEE International Conference on Systems, Man and Cybernetics*. Piscataway, NJ : IEEE Service Center, 2007. – ISBN 978–1–4244–0990–7, S. 1506–1511
- [438] TAGUCHI, Gen'ichi ; KONISHI, Shōzō: *Orthogonal Arrays and linear graphs: Tools for quality engineering*. Dearborn, Mich. : American Supplier Institute, 1987. – ISBN 094124301X
- [439] TAN, Chang W. ; DEMPSTER, Angus ; BERGMEIR, Christoph ; WEBB, Geoffrey I.: *MultiRocket: Multiple pooling operators and transformations for fast and effective time series classification*. <https://arxiv.org/pdf/2102.00457>. Version: 2021
- [440] TANG, Shu ; SHELDEN, Dennis R. ; EASTMAN, Charles M. ; PISHDAD-BOZORGI, Pardis ; GAO, Xinghua: A review of building information modeling (BIM) and the internet of things (IoT) devices integration: Present status and future trends. In: *Automation in Construction* 101 (2019), S. 127–139. <http://dx.doi.org/10.1016/j.autcon.2019.01.020>. – DOI 10.1016/j.autcon.2019.01.020. – ISSN 09265805
- [441] TATA, Sandeep ; PATEL, Jignesh M.: Estimating the selectivity of tf-idf based cosine similarity predicates. In: *ACM SIGMOD Record* 36 (2007), Nr. 2, S. 7–12. <http://dx.doi.org/10.1145/1328854.1328855>. – DOI 10.1145/1328854.1328855. – ISSN 0163–5808
- [442] TIMM, Stephanie N. ; DEAL, Brian M.: Effective or ephemeral? The role of energy information dashboards in changing occupant energy behaviors. In: *Energy Research & Social Science* 19 (2016), S. 11–20. <http://dx.doi.org/10.1016/j.erss.2016.04.020>. – DOI 10.1016/j.erss.2016.04.020. – ISSN 22146296
- [443] TORABI, Narges ; BURAK GUNAY, H. ; O'BRIEN, William: A review of common human errors in design, installation, and operation of multiple-zone VAV AHU systems. In: *Journal of Physics: Conference Series* 2042 (2021), Nr. 1, S. 012130. <http://dx.doi.org/10.1088/1742-6596/2042/1/012130>. – DOI 10.1088/1742-6596/2042/1/012130. – ISSN 1742–6588
- [444] TÖRNGREN, Martin ; BENSALAM, Saddek ; CENGARLE, María V. ; CHEN, De-Jiu ; MCDERMID, John ; PASSERONE, Roberto ; SANGIOVANNI-VINCENTELLI, Alberto ; RUNKLER, Thomas: *Cy-PhERS - Cyber-Physical European Roadmap & Strategy*. <http://www.cyphers.eu/sites/default/files/D5.1.pdf>. Version: 2014
- [445] TOUVRON, Hugo ; LAVRIL, Thibaut ; IZACARD, Gautier ; MARTINET, Xavier ; LACHAUX, Marie-Anne ; LACROIX, Timothée ; ROZIÈRE, Baptiste ; GOYAL, Naman ; HAMBRO, Eric ; AZHAR, Faisal ; RODRIGUEZ, Aurelien ; JOULIN, Armand ; GRAVE, Edouard ; LAMPLE, Guillaume: *LLaMA: Open and Efficient Foundation Language Models*. <http://arxiv.org/pdf/2302.13971>. Version: 2023
- [446] TSALLIS, Constantino: *Introduction to nonextensive statistical mechanics: approaching a complex world*. Bd. 1. Springer, 2009

-
- [447] TUGGENER, Don: *Incremental Coreference Resolution for German*. Zurich, University of Zurich, Diss., 2016. https://www.cl.uzh.ch/dam/jcr:b2212d28-6248-47dc-a4e3-04206ff4c6db/tuggener_diss.pdf
- [448] UHLENBECK, G. E. ; ORNSTEIN, L. S.: On the Theory of the Brownian Motion. In: *Physical Review* 36 (1930), Nr. 5, S. 823–841. <http://dx.doi.org/10.1103/PhysRev.36.823>. – DOI 10.1103/PhysRev.36.823. – ISSN 0031–899X
- [449] VAN DER DONCKT, Jonas ; VAN DER DONCKT, Jeroen ; DEPROST, Emiel ; VAN HOECKE, Sofie: tsflex: Flexible time series processing & feature extraction. In: *SoftwareX* 17 (2022), S. 100971. <http://dx.doi.org/10.1016/j.softx.2021.100971>. – DOI 10.1016/j.softx.2021.100971. – ISSN 23527110
- [450] VAN STIPHOUDT, Christine ; STINNER, Florian ; BODE, Gerrit ; KÜMPEL, Alexander ; MÜLLER, Dirk: Fault detection and diagnosis in building energy systems: A tool chain for the automated generation of training data. In: *Journal of Physics: Conference Series* 2042 (2021), Nr. 1, S. 012083. <http://dx.doi.org/10.1088/1742-6596/2042/1/012083>. – DOI 10.1088/1742-6596/2042/1/012083. – ISSN 1742–6588
- [451] VASERSTEIN, Leonid N.: Markov processes over denumerable products of spaces, describing large systems of automata. In: *Problemy Peredachi Informatsii* 5 (1969), Nr. 3, S. 64–72
- [452] VBIS: *VBIS - Virtual Building Information System*. <https://vbis.com.au/>. Version: 2024
- [453] VERBERT, Kim ; BABUŠKA, R. ; SCHUTTER, Bart de: Combining knowledge and historical data for system-level fault diagnosis of HVAC systems. In: *Engineering Applications of Artificial Intelligence* 59 (2017), S. 260–273. – ISSN 09521976
- [454] VEREGIN, Howard: Data quality parameters. In: *Geographical information systems* 1 (1999), S. 177–189
- [455] VIEIRA, Susana M. ; KAYMAK, Uzay ; SOUSA, Joao M. C.: Cohen's kappa coefficient as a performance measure for feature selection. In: *International Conference on Fuzzy Systems (FUZZ 2010)*. Piscataway, NJ : Ieee, 2010. – ISBN 978–1–4244–6919–2, S. 1–8
- [456] VOLK, Rebekka ; STENGEL, Julian ; SCHULTMANN, Frank: Building Information Modeling (BIM) for existing buildings — Literature review and future needs. In: *Automation in Construction* 38 (2014), S. 109–127. <http://dx.doi.org/10.1016/j.autcon.2013.10.023>. – DOI 10.1016/j.autcon.2013.10.023. – ISSN 09265805
- [457] VOLLMER, Michael ; ENGLHARDT, Adrian ; TRITTENBACH, Holger ; BIELSKI, Pawel ; KARRARI, Shahab ; BÖHM, Klemens: Energy Time-Series Features for Emerging Applications on the Basis of Human-Readable Machine Descriptions. In: *Proceedings of the Tenth ACM International Conference on Future Energy Systems*. New York, NY, USA : ACM, 2019. – ISBN 9781450366717, S. 474–481

- [458] VOSECKY, Jan ; HONG, Dan ; SHEN, Vincent Y.: User identification across multiple social networks. In: *2009 First International Conference on Networked Digital Technologies*, Ieee, 2009. – ISBN 978-1-4244-4614-8, S. 360–365
- [459] WAIDE, Paul ; URE, Jim ; KARAGIANNI, Nadia ; SMITH, Graham ; BORDASS, Bill ; WAIDE STRATEGIC EFFICIENCY (Hrsg.): *The scope for energy and CO2 savings in the EU through the use of building automation technology: Final Report*. http://neu.eubac.org/fileadmin/eu.bac/BACS_studies_and_reports/2014.06.13_Waide_ECI_-_Energy_and_CO2_savings_BAT.pdf. Version: 2014
- [460] WANG, Haitao ; CHEN, Youming: A robust fault detection and diagnosis strategy for multiple faults of VAV air handling units. In: *Energy and Buildings* 127 (2016), S. 442–451. – ISSN 03787788
- [461] WANG, Haitao ; CHEN, Youming ; CHAN, Cary W. H. ; QIN, Jianying: A robust fault detection and diagnosis strategy for pressure-independent VAV terminals of real office buildings. In: *Energy and Buildings* 43 (2011), Nr. 7, S. 1774–1783. – ISSN 03787788
- [462] WANG, Haitao ; CHEN, Youming ; CHAN, Cary W. H. ; QIN, Jianying: An online fault diagnosis tool of VAV terminals for building management and control systems. In: *Automation in Construction* 22 (2012), S. 203–211. – ISSN 09265805
- [463] WANG, Haitao ; CHEN, Youming ; CHAN, Cary W. H. ; QIN, Jianying ; WANG, Jinhua: Online model-based fault detection and diagnosis strategy for VAV air handling units. In: *Energy and Buildings* 55 (2012), S. 252–263. – ISSN 03787788
- [464] WANG, Shengwei ; XIAO, Fu: Sensor fault detection and diagnosis of air-handling units using a condition-based adaptive statistical method. In: *HVAC&R Research* 12 (2006), Nr. 1, S. 127–150. – ISSN 1078-9669
- [465] WANG, Weimin ; BRAMBLEY, Michael R. ; KIM, Woohyun ; SOMASUNDARAM, Sriram ; STEVENS, Andrew J.: Automated point mapping for building control systems: Recent advances and future research needs. In: *Automation in Construction* 85 (2018), S. 107–123. <http://dx.doi.org/10.1016/j.autcon.2017.09.013>. – DOI 10.1016/j.autcon.2017.09.013. – ISSN 09265805
- [466] WANG, Zhe: How frequent should we measure the indoor thermal environment. In: *Building and Environment* 222 (2022), S. 109464. <http://dx.doi.org/10.1016/j.buildenv.2022.109464>. – DOI 10.1016/j.buildenv.2022.109464. – ISSN 0360-1323
- [467] WATERWORTH, David ; SETHUVENKATRAMAN, Subbu ; SHENG, Quan Z.: Advancing Smart Building Readiness: Automated Metadata Extraction Using Neural Language Processing Methods. In: *Advances in Applied Energy* (2021), S. 100041. <http://dx.doi.org/10.1016/j.adapen.2021.100041>. – DOI 10.1016/j.adapen.2021.100041. – ISSN 26667924

- [468] WATERWORTH, David ; SETHUVENKATRAMAN, Subbu ; SHENG, Quan Z.: Deploying data driven applications in smart buildings: Overcoming the initial onboarding barrier using machine learning. In: *Energy and Buildings* (2023), S. 112699. <http://dx.doi.org/10.1016/j.enbuild.2022.112699>. – DOI 10.1016/j.enbuild.2022.112699. – ISSN 03787788
- [469] WECHSELAUER, Josef: *Towards an Automatic Configuration Process for Building Automation Systems*. Wien, TU Wien, Diplomarbeit, 2021. <https://repositum.tuwien.at/bitstream/20.500.12708/17582/1/Wechselauer%20Josef%20-%202021%20-%20Towards%20an%20Automatic%20Configuration%20Process%20for...pdf>
- [470] WEINER, Peter: Linear pattern matching algorithms. In: *14th Annual Symposium on Switching and Automata Theory (swat 1973)*, Ieee, 10/15/1973 - 10/17/1973, S. 1–11
- [471] WELSCH, Merten ; LÜTZKENDORF, Thomas ; BUNDESINSTITUT FÜR BAU-, STADT- UND RAUMFORSCHUNG: *Bundesinstitut für Bau-, Stadt- und Raumforschung*. Bd. Nr. 2020, 17; 2020, Nr. 17: *Umweltfußabdruck von Gebäuden in Deutschland: Kurzstudie zu sektorübergreifenden Wirkungen des Handlungsfelds "Errichtung und Nutzung von Hochbauten" auf Klima und Umwelt*. Stand: Dezember 2020. Bonn : Bundesinstitut für Bau-, Stadt- und Raumforschung (BBSR) im Bundesamt für Bauwesen und Raumordnung (BBR), 2020 https://www.bbsr.bund.de/BBSR/DE/veroeffentlichungen/bbsr-online/2020/bbsr-online-17-2020-dl.pdf?__blob=publicationFile&v=2
- [472] WHITE, Franklin E.: *Data fusion lexicon*. <https://apps.dtic.mil/dtic/tr/fulltext/u2/a529661.pdf>. Version: 1991
- [473] WIENS, Jenna ; GUTTAG, John V. ; HORVITZ, Eric: Patient risk stratification for hospital-associated C. diff as a time-series classification task. In: *Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1*. Red Hook, NY, USA : Curran Associates Inc, 2012 (NIPS'12), S. 467–475
- [474] WILLS, Nadine ; BARTELS, Niels: Bedarfsorientierte FM-Leistungserbringung im Rahmen der lebenszyklusübergreifenden BIM-Methode unter Berücksichtigung offener Datenaustauschformate, 2021
- [475] WU, Jiaman ; HONG, Dezhi ; GUPTA, Rajesh ; SHANG, Jingbo: *Sensei: Self-Supervised Sensor Name Segmentation*. <http://arxiv.org/pdf/2101.00130v1>. Version: 2021
- [476] XIAO, Fu ; FAN, Cheng: Data mining in building automation system for improving building operational performance. In: *Energy and Buildings* 75 (2014), S. 109–118. <http://dx.doi.org/10.1016/j.enbuild.2014.02.005>. – DOI 10.1016/j.enbuild.2014.02.005. – ISSN 03787788
- [477] XIAO, Fu ; WANG, Shengwei ; XU, Xinhua ; GE, Gaoming: An isolation enhanced PCA method with expert-based multivariate decoupling for sensor FDD in air-conditioning systems. In: *Applied Thermal Engineering* 29 (2009), Nr. 4, S. 712–722. – ISSN 13594311

- [478] XIAO, Fu ; ZHAO, Yang ; WEN, Jin ; WANG, Shengwei: Bayesian network based FDD strategy for variable air volume terminals. In: *Automation in Construction* 41 (2014), S. 106–118. – ISSN 09265805
- [479] XU, Hongjiang: What Are the Most Important Factors for Accounting Information Quality and Their Impact on AIS Data Quality Outcomes? In: *Journal of Data and Information Quality* 5 (2015), Nr. 4, S. 1–22. <http://dx.doi.org/10.1145/2700833>. – DOI 10.1145/2700833. – ISSN 1936–1955
- [480] XUE, Puning ; ZHOU, Zhigang ; FANG, Xiumu ; CHEN, Xin ; LIU, Lin ; LIU, Yaowen ; LIU, Jing: Fault detection and operation optimization in district heating substations based on data mining techniques. In: *Applied Energy* 205 (2017), S. 926–940. – ISSN 03062619
- [481] YAN, Chang ; LI, Peng ; LIU, Changchun ; WANG, Xinpei ; YIN, Chunyan ; YAO, Lianke: Novel gridded descriptors of poincaré plot for analyzing heartbeat interval time-series. In: *Computers in biology and medicine* 109 (2019), S. 280–289
- [482] YAN, Ke ; JI, Zhiwei ; LU, Huijuan ; HUANG, Jing ; SHEN, Wen ; XUE, Yu: Fast and accurate classification of time series data using extended ELM: Application in fault diagnosis of air handling units. In: *IEEE Transactions on Systems, Man, and Cybernetics: Systems* 49 (2017), Nr. 7, S. 1349–1356. – ISSN 2168–2216
- [483] YAN, Ke ; ZHONG, Chaowen ; JI, Zhiwei ; HUANG, Jing: Semi-supervised learning for early detection and diagnosis of various air handling unit faults. In: *Energy and Buildings* 181 (2018), S. 75–83. – ISSN 03787788
- [484] YAN, Rui ; MA, Zhenjun ; ZHAO, Yang ; KOKOGIANNAKIS, Georgios: A decision tree based data-driven diagnostic strategy for air handling units. In: *Energy and Buildings* 133 (2016), S. 37–45. – ISSN 03787788
- [485] YAN, Ruqiang ; GAO, Robert X.: Approximate Entropy as a diagnostic tool for machine health monitoring. In: *Mechanical Systems and Signal Processing* 21 (2007), Nr. 2, S. 824–839. <http://dx.doi.org/10.1016/j.ymssp.2006.02.009>. – DOI 10.1016/j.ymssp.2006.02.009. – ISSN 08883270
- [486] YAN, Ying ; LUH, Peter B. ; PATTIPATI, Krishna R.: Fault diagnosis of HVAC air-handling systems considering fault propagation impacts among components. In: *IEEE Transactions on Automation Science and Engineering* 14 (2017), Nr. 2, S. 705–717. – ISSN 1545–5955
- [487] YAN, Ying ; LUH, Peter B. ; PATTIPATI, Krishna R.: Fault diagnosis of components and sensors in HVAC air handling systems with new types of faults. In: *IEEE Access* 6 (2018), S. 21682–21696. – ISSN 2169–3536
- [488] YANG, Jiawei ; CHOUDHARY, Gulraiz I. ; RAHARDJA, Susanto ; FRÄNTI, Pasi: Classification of Interbeat Interval Time-Series Using Attention Entropy. In: *IEEE Transactions on Affective Computing* 14 (2023), Nr. 1, S. 321–330. <http://dx.doi.org/10.1109/TAFFC.2020.3031004>. – DOI 10.1109/TAFFC.2020.3031004

-
- [489] YANG, Xue-Bin ; JIN, Xin-Qiao ; DU, Zhi-Min ; FAN, Bo ; ZHU, Yong-Hua: Optimum operating performance based online fault-tolerant control strategy for sensor faults in air conditioning systems. In: *Automation in Construction* 37 (2014), S. 145–154. – ISSN 09265805
- [490] YANG, Xue-Bin ; JIN, Xin-Qiao ; DU, Zhi-Min ; ZHU, Yong-Hua: A novel model-based fault detection method for temperature sensor using fractal correlation dimension. In: *Building and Environment* 46 (2011), Nr. 4, S. 970–979. – ISSN 0360–1323
- [491] YANG, Xue-Bin ; JIN, Xin-Qiao ; DU, Zhi-Min ; ZHU, Yong-Hua ; GUO, Yi-Bo: A hybrid model-based fault detection strategy for air handling unit sensors. In: *Energy and Buildings* 57 (2013), S. 132–143. – ISSN 03787788
- [492] YIN, Wenpeng ; KANN, Katharina ; YU, Mo ; SCHÜTZE, Hinrich: *Comparative Study of CNN and RNN for Natural Language Processing*. <http://arxiv.org/pdf/1702.01923v1>. Version: 2017
- [493] YOON, Sungmin: Virtual sensing in intelligent buildings and digitalization. In: *Automation in Construction* 143 (2022), S. 104578. <http://dx.doi.org/10.1016/j.autcon.2022.104578>. – DOI 10.1016/j.autcon.2022.104578. – ISSN 09265805
- [494] YU, Yuebin ; WORADECHJUMROEN, Denchai ; YU, Daihong: A review of fault detection and diagnosis methodologies on air-handling units. In: *Energy and Buildings* 82 (2014), S. 550–562. <http://dx.doi.org/10.1016/j.enbuild.2014.06.042>. – DOI 10.1016/j.enbuild.2014.06.042. – ISSN 03787788
- [495] ZADEH, Puyan A. ; WANG, Guan ; CAVKA, Hasan B. ; STAUB-FRENCH, Sheryl ; POTTINGER, Rachel: Information Quality Assessment for Facility Management. In: *Advanced Engineering Informatics* 33 (2017), S. 181–205. <http://dx.doi.org/10.1016/j.aei.2017.06.003>. – DOI 10.1016/j.aei.2017.06.003. – ISSN 14740346
- [496] ZHAN, Sicheng ; CHONG, Adrian: Data requirements and performance evaluation of model predictive control in buildings: A modeling perspective. In: *Renewable and Sustainable Energy Reviews* 142 (2021), S. 110835. <http://dx.doi.org/10.1016/j.rser.2021.110835>. – DOI 10.1016/j.rser.2021.110835. – ISSN 13640321
- [497] ZHAN, Sicheng ; CHONG, Adrian ; LASTERNAS, Bertrand: Automated recognition and mapping of building management system (BMS) data points for building energy modeling (BEM). In: *Building Simulation* 68 (2020), S. 671. <http://dx.doi.org/10.1007/s12273-020-0612-7>. – DOI 10.1007/s12273-020-0612-7. – ISSN 1996–8744
- [498] ZHANG, Liang ; ALAHMAD, Mahmoud ; WEN, Jin: Comparison of time-frequency-analysis techniques applied in building energy data noise cancellation for building load forecasting: A real-building case study. In: *Energy and Buildings* 231 (2021), S. 110592. <http://dx.doi.org/10.1016/j.enbuild.2020.110592>. – DOI 10.1016/j.enbuild.2020.110592. – ISSN 03787788

- [499] ZHANG, Liang ; LEACH, Matt ; BAE, Yeonjin ; CUI, Borui ; BHATTACHARYA, Saptarshi ; LEE, Seungjae ; IM, Piljae ; ADETOLA, Veronica ; VRABIE, Draguna ; KURUGANTI, Teja: Sensor impact evaluation and verification for fault detection and diagnostics in building energy systems: A review. In: *Advances in Applied Energy* 3 (2021), S. 100055. <http://dx.doi.org/10.1016/j.adapen.2021.100055>. – DOI 10.1016/j.adapen.2021.100055. – ISSN 26667924
- [500] ZHANG, Yan ; BAI, Xuemei ; MILLS, Franklin P. ; PEZZEY, John C.: Rethinking the role of occupant behavior in building energy performance: A review. In: *Energy and Buildings* 172 (2018), S. 279–294. <http://dx.doi.org/10.1016/j.enbuild.2018.05.017>. – DOI 10.1016/j.enbuild.2018.05.017. – ISSN 03787788
- [501] ZHANG, Yuxi: *CNN-based Symbol Recognition and Detection in Piping Drawings*. West Lafayette, USA, Purdue University, Master thesis, 2019. <http://dx.doi.org/10.25394/PGS.8301080.v1>. – DOI 10.25394/PGS.8301080.v1
- [502] ZHAO, Futao ; FAN, James ; MIJANOVIC, Stevo: PI auto-tuning and performance assessment in HVAC systems. In: *2013 American Control Conference (ACC 2013)*. Piscataway, NJ : Ieee, 2013. – ISBN 978-1-4799-0178-4, S. 1783–1788
- [503] ZHAO, Yang ; LI, Tingting ; FAN, Cheng ; LU, Jingchao ; ZHANG, Xuejun ; ZHANG, Chaobo ; CHEN, Shuqin: A proactive fault detection and diagnosis method for variable-air-volume terminals in building air conditioning systems. In: *Energy and Buildings* 183 (2019), S. 527–537. – ISSN 03787788
- [504] ZHAO, Yang ; LI, Tingting ; ZHANG, Xuejun ; ZHANG, Chaobo: Artificial intelligence-based fault detection and diagnosis methods for building energy systems: Advantages, challenges and the future. In: *Renewable and Sustainable Energy Reviews* 109 (2019), S. 85–101. <http://dx.doi.org/10.1016/j.rser.2019.04.021>. – DOI 10.1016/j.rser.2019.04.021. – ISSN 13640321
- [505] ZHAO, Yang ; WEN, Jin ; WANG, Shengwei: Diagnostic Bayesian networks for diagnosing air handling units faults–Part II: Faults in coils and sensors. In: *Applied Thermal Engineering* 90 (2015), S. 145–157. – ISSN 13594311
- [506] ZHAO, Yang ; WEN, Jin ; XIAO, Fu ; YANG, Xuebin ; WANG, Shengwei: Diagnostic Bayesian networks for diagnosing air handling units faults–part I: Faults in dampers, fans, filters and sensors. In: *Applied Thermal Engineering* 111 (2017), S. 1272–1286. – ISSN 13594311
- [507] ZHAO, Zhenge ; MOTTA, Danilo ; BERGER, Matthew ; LEVINE, Joshua A. ; KUZUCU, Ismail B. ; FLEISCHMAN, Robert B. ; PAIVA, Afonso ; SCHEIDEGGER, Carlos: STFT-LDA: An algorithm to facilitate the visual analysis of building seismic responses. In: *Information Visualization* 20 (2021), Nr. 4, S. 263–282. <http://dx.doi.org/10.1177/14738716211038618>. – DOI 10.1177/14738716211038618. – ISSN 1473–8716
- [508] ZHONG, Chaowen ; YAN, Ke ; DAI, Yuting ; JIN, Ning ; LOU, Bing: Energy efficiency solutions

- for buildings: Automated fault diagnosis of air handling units using generative adversarial networks. In: *Energies* 12 (2019), Nr. 3, S. 527. – ISSN 1996–1073
- [509] ZHOU, Qiuju ; LEYDESDORFF, Loet: The normalization of occurrence and Co-occurrence matrices in bibliometrics using Cosine similarities and Ochiai coefficients. In: *Journal of the Association for Information Science and Technology* 67 (2016), Nr. 11, S. 2805–2814. <http://dx.doi.org/10.1002/asi.23603>. – DOI 10.1002/asi.23603. – ISSN 23301635
- [510] ZHU, Lei ; SHAN, Ming ; XU, Zhao: Critical review of building handover-related research in construction and facility management journals. In: *Engineering, Construction and Architectural Management* 28 (2021), Nr. 1, S. 154–173. <http://dx.doi.org/10.1108/ECAM-10-2018-0442>. – DOI 10.1108/ECAM-10-2018-0442. – ISSN 0969–9988
- [511] ZHU, Yueying ; ZHANG, Benwei ; WANG, Qiuping A. ; LI, Wei ; CAI, Xu: The principle of least effort and Zipf distribution. In: *Journal of Physics: Conference Series* 1113 (2018), S. 012007. <http://dx.doi.org/10.1088/1742-6596/1113/1/012007>. – DOI 10.1088/1742-6596/1113/1/012007. – ISSN 1742–6588
- [512] ZINS, Chaim: Conceptual approaches for defining data, information, and knowledge. In: *Journal of the American Society for Information Science and Technology* 58 (2007), Nr. 4, S. 479–493. <http://dx.doi.org/10.1002/asi.20508>. – DOI 10.1002/asi.20508. – ISSN 1532–2882
- [513] ZIPF, George K.: The Psycho-Biology of Language. In: *Language* 12 (1936), Nr. 3, S. 196. <http://dx.doi.org/10.2307/408930>. – DOI 10.2307/408930. – ISSN 00978507
- [514] ZIPF, George K.: National Unity and Disunity. The Nation as a Bio-Social Organism. By George Kingsley Zipf. Bloomington Indiana: Principia Press 1941 pp. xix 408. Index. 3.50. In: *American Journal of International Law* 35 (1941), Nr. 3, S. 593–594. <http://dx.doi.org/10.2307/2192485>. – DOI 10.2307/2192485

Anhang

A Informationsqualität

A.1 Learning Ability Index (LAI)

Alanne u. Sierla [15] schlägt den Learning Ability Index (LAI) für die Bewertung von autonomen Prozessen in Gebäudeenergiemanagement-Systemen vor. Diese stellt die Fähigkeit eines Smart Building zu Lernen in den Zusammenhang mit seiner Basis-Fähigkeit. Sie stellen die Gebäudeintelligenz BI (engl. building intelligence) formal als Summe von verschiedenen KPIs (engl. key performance indicator) in der Zeit wie in Formel A.1 dargestellt dar [14].

$$BI(t) = f(KPI_1(t) \dots KPI_n(t)) \quad (A.1)$$

Allgemeiner kann die Zeit auch als Ausdruck für den benötigten Aufwand (Index: eff für engl. effort) gesehen werden, so verändert sich die Grundformel der BI zur Formel A.2.

$$BI(eff) = f(KPI_1(eff) \dots KPI_n(eff)) \quad (A.2)$$

Die Lerngeschwindigkeit LR kann allgemein wie in Formel A.3 als Veränderung der Gebäudeintelligenz über die Zeit definiert werden.

$$LR = \frac{dBI}{dt} \quad (A.3)$$

Relevanter für die Praxis ist die durchschnittliche Lernrate (ALR: engl. average learning rate), die durch die Differenz einer Basisintelligenz BI_{pre} zu seiner Intelligenz nach einem Lernvorgang BI_{post} im Vergleich zur benötigten Zeit oder Aufwand definiert ist.

Der normalisierte Lernerfolg (NLG: engl. normalized learning gain) basiert auf einem Basiswert, der auf einem vordefinierten Aufwand (e_{ref} basiert (siehe Formel A.4). Insgesamt wird somit ein Verhältnis des realen Ergebnisses zu einem optimalen Basiswert ermittelt.

$$NLG_{e_{ref}} = \frac{BI_{post} - BI_{pre}}{BI_{target} - BI_{pre}} \quad (A.4)$$

Auf die KPIs bezogen, kann der NLG für die i -te KPI wie im Folgenden in Formel A.5 beschrieben berechnet werden. So wird eine KPI berechnet, die mit dem wenigsten Aufwand berechnet werden kann ($KPI_{i,0}$). Diese wird als Basis genommen und eine KPI berechnet, die bei unendlichem

Aufwand bzw. vollem Wissen erreicht werden kann ($KPI_{i,\infty}$). Diese werden dann mit dem errechnetem KPI des untersuchten Ansatzes KPI_i ins Verhältnis gesetzt werden. Schlussendlich zeigt der NLG den Wert an, die ein Ansatz im Vergleich zu seinem idealen Wert erreicht werden kann.

$$NLG_i = \frac{KPI_i - KPI_{i,0}}{KPI_{i,\infty} - KPI_{i,0}} \quad (A.5)$$

Wenn mehrere KPIs genutzt werden, so kann der Gesamt-NLG mithilfe folgender Formel und verschiedenen spezifischen Gewichten der KPIs w_i wie folgt berechnet werden:

$$NLG_{e_{ref}} = \sum_{i=1}^n w_i \cdot NLG_i \quad (A.6)$$

Die Gewichte w_i müssen insgesamt eine Summe von 1 bilden ($\sum w_i = 1$).

Zusätzlich zu den KPIs müssen zusätzlich die Qualität ($Q \in [0, 1]$) und Diversität ($D \in [0, 1]$) der verwendeten Datensätze berücksichtigt werden. Dieser wird in einem Erfahrungskoeffizienten $X \in [0, 1]$ abgebildet ($X = f(Q, D) = \sum_{j=1}^m D_j \cdot Q_j$).

Der Lernkompetenzindex (LAI: engl. learning ability index) ergibt sich somit zu Formel A.7.

$$LAI_{e_{ref}} = X \cdot NLG_{e_{ref}} \quad (A.7)$$

Für jeden Schritt in dieser Arbeit wird ein optimaler Basiswert ermittelt, an dem die anderen Ansätze gemessen werden müssen.

A.2 Herleitung Effektivität

Der „Informationsgewinn“ IG kann in Formel A.8 ausgedrückt werden [15], wobei I_{base} ein Informationsstand ist bei einem Basisalgorithmus und I_{post} nach Durchführung des IE-Algorithmus.

$$IG = I_{post} - I_{base} \quad (A.8)$$

Die Fähigkeit kann auch als Lernfähigkeit (siehe Kapitel A.1) aufgefasst werden:

$$LAI_{e_{ref}} = X \cdot NLG_{e_{ref}} \quad (A.9)$$

Hierbei ist X ein Korrekturfaktor für die initiale Datenqualität und Komplexität des evaluierten Datensatzes und $NLG_{e_{ref}}$ der normalisierte Lernerfolg als gewichtete normalisierte Summe der Informationsqualität gegenüber der Basisqualität I_{base} , wie in Formel A.10 ausgedrückt.

$$IG_{norm} = \frac{I_{post} - I_{base}}{I_{base}} \quad (A.10)$$

Die Robustheit wird formell sehr unterschiedlich definiert [142]. Ziel ist immer einen Unterschied zur Generalisierbarkeit herzustellen [142]. Der Rashomon Effekt sagt aus, dass mehrere Klassifizierer ähnliche Ergebnisse auch aus anderen Perspektiven (anderen Merkmalswichtigkeiten) erzielen können [62, 171]. Somit kann die Robustheit auch als vergleichbare Ergebnisse der Algorithmen bei unterschiedlichen Datensätzen mit der gleichen Aufgabenstellung und Zielsetzung verstanden werden.

Als Basis wird deshalb die Standardabweichung in einem IQ-Kriterium σ_I im Vergleich zum mittleren IQ-Abschneiden des Algorithmus I_{av} als Robustheit Rob wie in Formel A.11 definiert.

$$\text{Rob} = \frac{1 - 3 \cdot \sigma_I}{I_{av}} \quad (\text{A.11})$$

Wenn alle in diesem Abschnitt definierten Formeln in Formel 3.12 für die Effektivität eingesetzt werden, ergibt sich folgende Gesamtformel A.12 für die Effektivität:

$$\text{Effektivität} = X \cdot \frac{(I_{\text{post}} - I_{\text{base}})^2}{I_{\text{base}}} \cdot \frac{1 - 3 \cdot \sigma_I}{I_{av}} \quad (\text{A.12})$$

B Metadatenextraktionsverfahren in Energiesystemen

C Sprachverarbeitung

C.1 Wortfilter

Alle numerischen Werte werden a priori entfernt. Auch alle Wörter mit einem numerischen Buchstaben werden entfernt. Unter diesen Wörtern befinden sich sehr viele Wörter, die nur fehlerhaft als Wörter erkannt wurden. Für den Wortfilter werden alle Wörter entfernt, die zwei oder mehr Buchstaben in Großschreibung enthalten. Der erste Buchstabe jedes Wortes muss in Großschreibung auftauchen. Außerdem werden nur Wörter aufgenommen, die nur die deutschen Buchstaben a-z, A-Z, ß, ä, Ä, ö, Ö, ü, Ü enthalten.

Alle Wörter werden anhand von allen Sonderzeichen getrennt. So werden vom Wortfilter anhand von Bindestrich zusammengesetzte Wörter wie „Nordrhein-Westfalen“ in die Wörter „Nordrhein“

Quelle	Verwendete Algorithmen	Anzahl der Merkmale	Supervised(SV) / Unsupervised(USV)	Zeichenketten(ZK) / Zeitreihen(ZR) / Topologieerkennung(TE)	Gebäudeübergreifend
[34]	RFC, hierarchical clustering	30	SV	ZK, ZR	Ja
[45]	eigene	NaN	SV	ZK	Ja
[78]	RFC	23	SV	ZR, TE	Nein
[146]	eigene	NaN	SV	ZR	Nein
[147]	13 Klassifizierer	8	SV	ZR	Nein
[152]	AdaBoost, Decision Tree, kNN, LDA, Logistic, Naive BaJa, RFC	6 Gruppen (max. 19 Merkmale)	SV	ZR	Ja
[168]	RFC, CNN-Bi-LSTM	20	SV	ZR	Ja
[189]	DTW, EMD, Slope Compare	2	SV	ZR	Nein
[193]	SVM, Logistic Regression (LR), RFC, GMM	11	SV	ZR	Ja
[194]	NaN	NaN	SV	ZK	Ja
[191]	Markov Event Model, EM	NaN	USV	TE	Ja
[215]	Char-LSTM-CRF, CNN, MLP	NaN	SV	ZK	Ja
[244]	6 Klassifizierer	2	SV	TE	Nein
[245]	CRF, MLP	NaN	SV	ZK	Ja
[266]	Jaro-Winkler	NaN	SV	ZK	Ja
[275]	STFT + CNN (STN)	NaN	SV	TE	Ja
[286]	Cosinus-Ähnlichkeit	NaN	SV	ZK	Ja
[298]	k-mers, RF, SVM, LR	11	SV	ZK, ZR	Ja
[299]	13 verschiedene Methoden	NaN	USV	ZK	Ja
[321]	SVM, RFC	NaN	SV	ZK, ZR	Ja
[475]	Char-CNN	NaN	USV	ZK	Ja
[497]	DBSCAN, Fuzzy string matching	NaN	SV	ZK	Ja
[17]	Auto Encoder	NaN	SV	ZK	Nein
[378]	XGBoost, als Vergleich: RFC, NN, SVM	44	SV	ZR	Nein
[467]	RoBERTa	NaN	USV (ME), SV (Klass)	ZK	Ja
[56]	gBERT	NaN	SV	ZK	Ja
[316]	TSfresh + 7 Klassifizierer	NaN	SV	ZR	Ja

Tabelle B.1: Übersicht über alle untersuchten Algorithmen. ZR als Anzahl der Merkmale ist die gesamte Zeitreihe. NaN kennzeichnet Werte, die aus den Ansätzen nicht ersichtlich sind. „ME“ ist Merkmalsextraktion, „Klass“ zeigt die Klassifizierung an.

	Anzahl Wörter	Überschneidung mit OOM4ABDO Datensatz (in %)	Wortdichte (Treffer/ 100.000 Wörter)
Wikipedia	15,01 Mio	77,46	19,32
Wortschatzportal (Leipzig)	6,27 Mio	72,81	
Free German Dictionary	2,15 Mio	48,40	
Webcrawling	0,73 Mio	78,90	406,74
Normen	0,11 Mio	59,94	2.011,82
Datenstromnamen	0,01 Mio	33,68	18.341,82
Total	15,82 Mio	82,75	19,58

Tabelle C.1: Statistiken der verwendeten Korpora.

und „Westfalen“ aufgeteilt. Bei der Überprüfung, ob ein Wort in dem jeweiligen Wortkorpus enthalten ist, werden die Wörter als Kleinbuchstaben verwendet.

C.2 Korpora Statistiken

In Tabelle C.1 sind die einzelnen Statistiken der Korpora aufgeführt. Die Korpora, die insbesondere Inhalte über Gebäudeautomation enthalten, haben eine deutlich höhere Wortdichte mit dem OOM4ABDO Datensatz. Bis zu 18,3 % aller im Korpus enthaltenen Wörter befinden sich auch im OOM4ABDO Datensatz.

C.3 Synonyme

Im folgenden sind die Synonyme in der Gebäudeautomation aufgelistet:

- Anlage, System, Einrichtung, Gerät, Teil, Einheit, Technik, Station, Werk, Maschine
- Steuerung, Regelung, Betrieb, Funktion, Kontrolle, Anwendung, Folge, Vorgang, Programm, Plan, Logik, Ausführung
- Zustand, Modus, Stufe, Signal
- Sensor, Fühler, Zähler, Messer, Messung, Meter, Detektor
- Klasse, Typ, Art
- Strom, Strömung, Fluss

- Bereich, Fläche, Zone, seitig
- Kreislauf, Verteiler, Teilsystem, Zirkulation, Umlauf, Netz
- Zahl, Nummer, Wert, Objekt
- Parameter, Beiwert, Profil, Vorgabe, (Feld)-Element, Eigenschaften, Kennzeichnung, Register
- Zeit, Dauer
- Schalter, Taster
- Masse, Menge
- Anzeige, Ausgabe, Status, Meldung, Tableau
- Änderung, Übergang
- Evaluierung, Bewertung, Ergebnis, Kalkulation, Berechnung, Analyse
- Speicher, Tank, Behälter
- Mittel, Lösung
- Art, Weise

Unklare Synonyme: Werte, Versorgung, Anschluss, Bereich

Die folgenden Zeichenketten werden am Ende eines Wortes zusätzlich als Suffixe verwendet: bare, bar, keit, en, ung, t, lich, heit, ts, ern, falls, isch, dem, den, ens, ig, et, end. Die folgenden Zeichenketten werden am Anfang eines Wortes zusätzlich als Prefixe verwendet: Be, ge, Ve, An. Die folgende Zeichenkette wird in der Mitte eines Wortes zusätzlich als Middlefix verwendet: zu.

C.4 Komposita

Ein großes Problem bei der Anwendung automatischer Algorithmen für die Erkennung von Metadaten in Gebäuden ist die Sprachabhängigkeit. So haben beispielsweise Algorithmen, die mit englischen semantischen Daten gefüttert werden, Schwierigkeiten, deutsche Semantik zu erkennen und umgekehrt.

In der deutschen Sprache werden häufig zusammengesetzte Wörter verwendet (Komposita). Im Englischen wird zum Beispiel die Vorlauftemperatur als zwei Wörter geschrieben: „flow temperature“. Im Deutschen hingegen wird ein Wort verwendet: „Vorlauftemperatur“. Die Komposita können theoretisch beliebig viele Wortbestandteile enthalten, deren Anzahl in der Realität jedoch limitiert ist. So kommt das laut Duden längste Wort auf maximal fünf Wortbestandteile: Aufmerksamkeitsdefizit-Hyperaktivitätsstörung (Aufmerksamkeit|Defizit|Hyper|Aktivität|Störung) [205].

Im Folgenden wird ein Ansatz zur Trennung von Komposita auf Basis verschiedener bestehender Ansätze und eigenen Datensätzen erstellt.

C.4.1 Wortdatenbank

Wenn einem ein Datensatz aus der Gebäudeautomation vorliegt, so sind viele unterschiedliche Wörter enthalten. Leider werden nicht alle Wörter korrekt geschrieben (Typos). Außerdem werden Abkürzungen verwendet. Diese können aus unterschiedlich vielen Buchstaben bestehen. So können nur einzelne Buchstaben wie „T“ für Temperatur, aber auch ein Teilaspekt des Wortes verwendet werden wie „Temp“ für Temperatur. Um herauszufinden, was ein gültiges Wort ist, können Wortdatenbanken herangeführt werden.

Die Normen-Datenbank Perinorm [44] (heute Nautos [43]) bietet eine Vielzahl an Standards für die Gebäudeautomation. Mithilfe eines selbstentwickelten Webcrawlers auf Basis von Selenium Webdriver [406] wurden 229 Normen mithilfe des Suchbegriffs „Gebäudeautomation“ identifiziert. Aus den Dokumenten wurde mithilfe von pdfplumber [419] der Text der Normen extrahiert. Der extrahierte Text wird dann in seine Wörter überführt. Zusätzlich zu eigenen Wortdatenbanken, können Datenbanken für allgemeine deutsche Wörter herangezogen werden.

Die Free German Dictionary [404] bietet über 2 Millionen verschiedene deutsche Wörter und ist laut eigenen Angaben der größte deutsche Wortkorpus. Alternativen sind hierbei GermaNet der Abteilung „Allgemeine Sprachwissenschaft und Computerlinguistik“ der Universität Tübingen mit ungefähr 121.700 deutschen Komposita [180]. Das Digitale Wörterbuch der deutschen Sprache [120] bietet einen Wortkorpus von 584.000 Einträgen. Der Wortkorpus ist jedoch nicht in einem Stück herunterladbar. Das Wortschatzportal der Universität Leipzig [371] begrenzt seine öffentlich verfügbaren Wortkorpora auf maximal 1 Millionen Sätze pro Korpus. Jedoch bietet das Portal verschiedene Korpora, wovon insgesamt 49 verschiedene Korpora verwendet werden. Diese enthalten jeweils bis zu 1 Millionen Sätze (22,4 GB purer Text). Die verwendeten Korpora basieren hauptsächlich auf Texten aus Nachrichten und Wikipedia. Aus diesen Sätzen werden alle enthaltenen Wörter extrahiert.

Zusätzlich wird ein Datensatz von 121 Standards, Vorschlägen und Normen zur Beschreibung von Datenstromnamen für den Wortkorpus herangeführt. Diese stammen aus der Recherche zur Entwicklung des BUDO Schemas und bilden einen kompakten Wortkorpus für typische Begriffe der Gebäudeautomation und -technik. Außerdem können alle bei der Recherche zur Entwicklung des BUDO Schemas heruntergeladenen Dokumente herangeführt werden. Dies umfasst insgesamt 4.924 Dokumente. Auch aus diesen Dokumenten werden alle Wörter extrahiert.

Aus allen Wortkorpora wird ein Gesamtkorpora gebildet, der durch eine Vorverarbeitung auf signifikante Wörter reduziert wird. Die verwendeten Filter sind im Anhang in Kapitel C.1 dargestellt.

Der Ansatz von Tuggener [447] ist von der Anzahl der nutzbaren Wörter anhand des verwendeten Arbeitsspeicher begrenzt. Deshalb müssen entweder die Wörter begrenzt werden, die in den Algorithmus zur Erzeugung der n-Gramme gespeist werden oder eine technische Möglichkeit entwickelt werden, wie eine Arbeitsspeicherbegrenzung umgangen werden kann. In Kapitel C.4.2 wird eine Methode entwickelt, wie diese Herausforderung bewältigt wird.

Zusätzlich wird ein Ansatz von repodiac [385] überprüft, der einen Ansatz auf Basis von Aho-Corasick-Zustandsautomaten [12] entwickelt hat. In dieser Dissertation wird dieser Ansatz verwendet und um weitere Verfahren erweitert. Schlussendlich werden alle drei Ansätze (Tuggener, repodiac und Stinner) miteinander verglichen und aus diesen ausgewählt.

Die genauen Ansätze werden in den folgenden Kapiteln C.4.2 (Tuggener), C.4.3 (Repodiac) und C.4.4 (Stinner) vorgestellt.

C.4.2 Tuggener

Der ursprüngliche Ansatz von Tuggener [447] basiert auf der Bewertung von Zeichenketten, ob diese am Anfang, in der Mitte oder am Ende von Wörtern auftauchen. Alle möglichen Zeichenketten eines Wortes werden an den einzelnen Stellen bewertet. Wenn die Wahrscheinlichkeit hoch ist, dass an dieser Stelle eine Trennung vorhanden ist, dann wird eine Trennung ausgegeben.

Wichtig für einen funktionierenden Ansatz ist, dass er nicht mit zu vielen Komposita als Eingangswerte trainiert wird. Sonst werden einzelne Wörter als in der Mitte zugehörig definiert, obwohl diese eigenständige Wörter sind. Deshalb wird ein kleiner Datensatz zum Training des Splitters verwendet. Zusätzlich wird ein großer Datensatz von technischen Wörtern verwendet. Somit werden die technischen Wörter im gesamten Datensatz überbewertet, was für die hier vorliegende Nutzung sinnvoll ist.

Technisch wurde der ursprüngliche Ansatz von Niklaus [335] erweitert, aber der grundlegende Algorithmus blieb gleich. Die Basis von Niklaus [335] wird weiter verwendet. In dieser Dissertation wurde der grundlegende Algorithmus technisch um eine lokale Datenbank auf Basis von SQLite [184] erweitert. So kann das Problem des benötigten zu großen Speicherplatzes umgangen werden, da nur Festplattenspeicher verwendet werden muss.

Zusätzlich wurde alternativ eine umfangreiche Nachbearbeitung eingeführt, die die Reihenfolge der Wahrscheinlichkeiten des ursprünglichen Ansatzes von Tuggener verwendet, aber auch alle anderen Varianten überprüft, egal ob ermittelt wurde, ob hier eine tatsächliche Trennung vorliegt oder nicht. Die genaue Nacharbeitung wird in Kapitel C.4.4 vorgestellt.

Ein Problem vom Ansatz von Tuggener ist, dass er immer nur an einer Stelle im aktuell verwendeten Datenblock trennen kann. Dies ist insbesondere problematisch, wenn ein Komposita aus mehr als zwei Bestandteilen besteht, die auch nicht als zwei Bestandteile im Lexikon bestehen.

C.4.3 Repodiac

Repodiac [385] entwickelte einen Ansatz auf Basis von Aho-Corasack-Zustandsautomaten [12] mit dem Free German Dictionary [404] als Lexikon. Aho-Corasack-Zustandsautomaten [12] sind bei Suchpattern von weniger als 20 Zeichen und Suchtexten kleiner als 30 Zeichen schneller als der

Knuth-Morris-Pratt-Algorithmus [243] als beliebte String-Suchalternative [287]. Der Ansatz von Repodiac wurde um eine Nachbearbeitung ergänzt, das Artefakte, die Prefixe sein könnten wie „Vor-“, „Durch-“ oder „Ent-“ entsprechend ihren Nachfolgenden Datenblöcken zuweist. Aber auch allgemeine Suffixe wie „-ung“, „-tion“, „-schaft“ oder Plural Suffixe wie „-en“ oder „-s“ werden als Artefakte gesondert betrachtet. Auch die Änderung von Vokalen zu Umlauten (z.B. Vogel → Vögel) bei der Pluralform finden Beachtung bei Artefakten. Der Algorithmus überprüft bei jedem Wort, welche Teilwörter in diesem enthalten sind. Dann geht er vom Ende des Wortes aus und überprüft vom kleinsten Treffer von diesem Zeichen ausgehend, ob der Anfang des Wortes sich auch in vorhandenen Wörtern befindet (siehe Grafik C.1). Ersatzweise wird noch überprüft, ob sich eine potentielle Pluralform mit einem Zeichenabstand von eins (z.B. „s“) im Lexikon befindet. Falls Artefakte im Algorithmus gefunden werden, so können diese im Nachbearbeitungsprozess den passenden Wörtern angefügt werden.

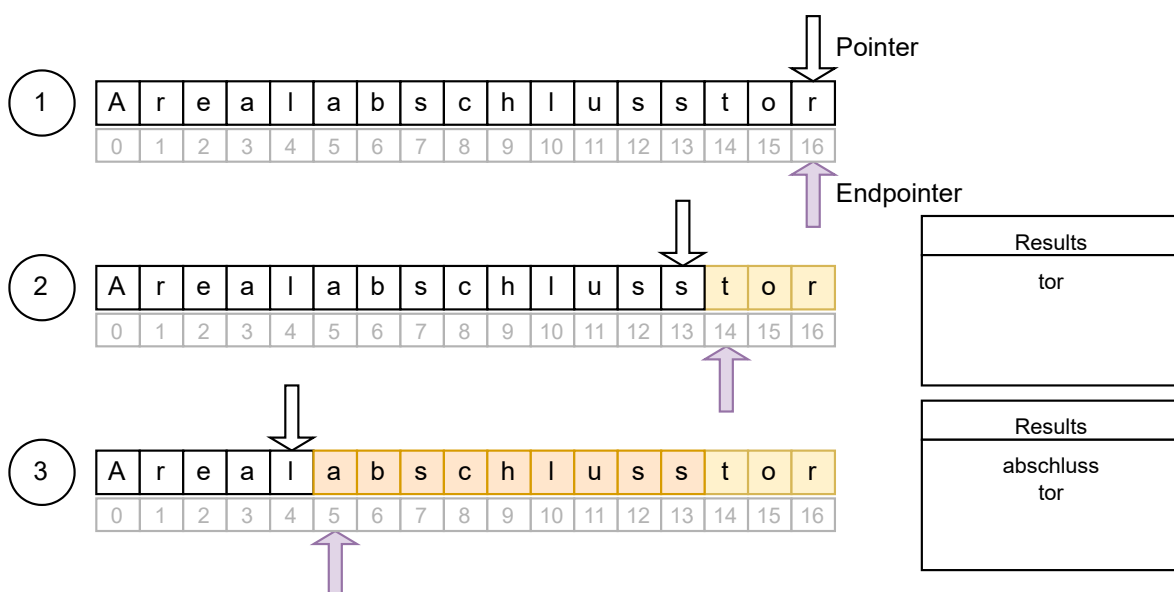


Abbildung C.1: Ablauf der Detektion von Repodiac. Vom letzten Zeichen ausgehend wird jeweils überprüft, ob sich am Ende eines gefundenen Wortes ein neues Wort befindet.

C.4.4 Stinner

Als Basis für den hier in der Dissertation verwendeten Algorithmus wird der vorher vorgestellte Ansatz von Repodiac verwendet. Zum einen kann dieser Ansatz sehr viel schneller ausgeführt werden als der Algorithmus von Tuggener (Faktor 10-20x schneller), zum anderen bietet dieser für Optimierungen größeres Potential, da er Wort für Wort durchgeht. Auch die Begrenzung von Tuggener auf zwei Bestandteile von Komposita kann bei der realen Trennung problematisch sein, weshalb der Ansatz von Repodiac als Basis sinnvoll ist.

Bestandteil des eigenen Algorithmus ist ein Postprocessing, in dem jeder gefundene Datenblock

nochmal überprüft wird, ob er auch gültig ist, d.h. auch eine Variante eines Wortes im Wörterbuch sein kann. Hierzu wird erst der Datenblock in Rohform untersucht, darauf folgend wird überprüft, ob etwaige Umlaute ausgeschrieben im Wort vorhanden sind (ae → ä) oder Umlautverschiebungen (z.B. Vogel → Vögel) zusammen mit Pluralformen auftreten könnten. Die Besonderheit im Gegensatz zum Ansatz von Repodiac ist, dass hier das gesamte Wort direkt betrachtet wird und nicht nur etwaige Artefakte. So wird in diesem Ansatz auch ein Plural direkt erkannt, wenn er sich noch nicht im Wörterbuch befindet. Auch mittlere Fixe wie „zu“ bei „einzusehen“ werden ersetzt und überprüft, ob die Grundform im Wörterbuch vorhanden ist. Die Länge der Datenblöcke wird auf minimal vier Zeichen begrenzt. Dem Algorithmus können aber auch einzelne Wörter beigebracht werden, die unter dieser Grenze liegen (z.B. „Öl“). Auch können im Teilwörter beigebracht werden, die nicht getrennt werden können bzw. im Kontext keine sinnvollen Wörter sind. Insbesondere Verben können fehlerhaft erkannt werden. Repodiac hat deshalb empfohlen Verben ganz auszuschließen, aber dies würde insbesondere bei Kurzformen der Verben („Ausström“ für „ausströmen“) ein Problem darstellen. Zusätzlich arbeitet der Algorithmus rekursiv und nutzt nach drei Iterationen in Ausnahmefällen den Algorithmus von Tuggener, falls die Datenblöcke nicht getrennt werden können. Die maximale Rekursionsdurchläufe werden pro Wort auf fünf festgelegt.

In Kapitel C.5 werden die Ergebnisse der Trennung der Komposita auf Basis der vorher vorgestellten Datensätze beschrieben. Insgesamt konnte in Vorversuchen beim Algorithmus „Stinner“ eine Genauigkeit von bis zu 98,2 % im Gegensatz zum aktuell besten Ansatz mit maximal 82,9 % Genauigkeit auf dem Datensatz auf Basis von Normen aus der Gebäudeautomation erreicht werden.

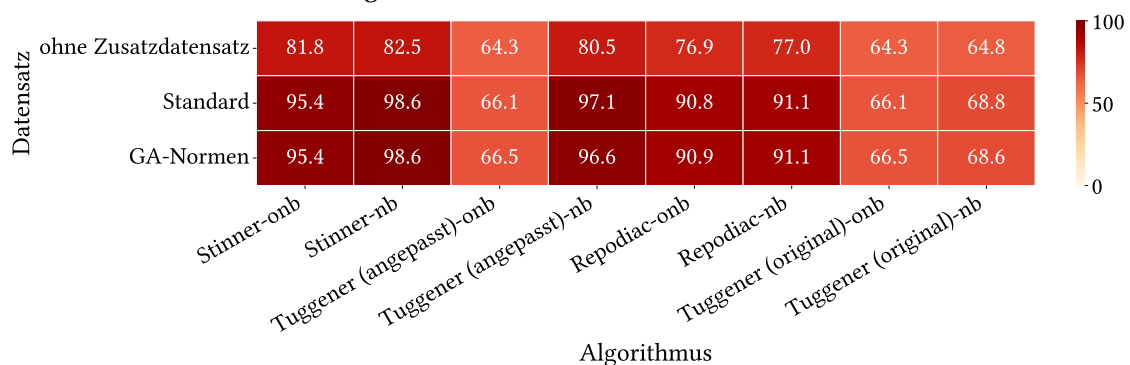
C.5 Ergebnisse Komposita Trennung

In Abbildung C.2 sind die Ergebnisse der Komposita Trennung mit verschiedenen Datensätzen und verschiedenen Algorithmen zu sehen. Zuerst werden die Ergebnisse auf dem Datensatz „Standard“ vorgestellt, darauf aufbauend auf dem Datensatz „Gebäudeautomation-Normen“. Die Nachbearbeitung („nb“) ist immer die in dieser Dissertation entwickelte Nacharbeitung der Ergebnisse des Algorithmus. Das bedeutet, dass nur jeweils die Algorithmen ohne Nachbearbeitung („onb“) dem jeweiligen Original-Algorithmus entsprechen.

Wie in Grafik C.2 zu sehen ist, führte die Nachbearbeitung der Ergebnisse beim Trainingsdatensatz „Standard“ zu Verbesserungen von bis zu 31 % Genauigkeit (Tuggener (angepasst) - ohne Nachbearbeitung im Verhältnis zu mit Nachbearbeitung). Die Anpassung des Algorithmus von Tuggener mithilfe des in dieser Arbeit entwickelten Auswahlverfahrens von Trennungen im Verhältnis zum Originalverfahren war um bis zu 28,3 % Genauigkeit besser (Tuggener (angepasst)-mit Nachbearbeitung im Verhältnis zu Tuggener (original)-mit Nachbearbeitung). Der originale Tuggener-Ansatz war mit Ergebnissen von 68,8 % Genauigkeit der Algorithmus mit den schlechtesten Ergebnissen. Der von repodiac entwickelte Algorithmus folgte mit Ergebnissen bis zu 91,1 % Genauigkeit. Der angepasste Tuggener Algorithmus zeigte Ergebnisse bis zu 97,1 % Genauigkeit. Der Algo-

rithmus von Stinner (Kapitel 3.3.5.2.3) zeigte in allen Fällen die besten Ergebnisse mit teils 98,6 % Genauigkeit. Der Unterschied zum angepassten Tuggener Algorithmus beträgt im Maximum nur 1,5 % Genauigkeit, jedoch wird die Fehlerrate von 2,9 % (angepasster Tuggener) auf 1,4 % (Stinner) halbiert. Wenn nur die Ergebnisse ohne die in dieser Dissertation entwickelten Nachbearbeitung betrachtet werden, kann das Ergebnis von maximal 90,9 % beim Stand der Technik (SdT) auf 98,6 % beim Algorithmus dieser Dissertation gesteigert werden.

Abbildung C.2: Genauigkeit der Ergebnisse des Dokumentensplitters für den Datensatz „Standard“. „nb“ steht für eine durchgeführte Nachbearbeitung, „onb“ steht für keine Nachbearbeitung.



Der Testdatensatz „Gebäudeautomation-Normen“ bestätigte qualitativ weitestgehend die Ergebnisse des Trainingsdatensatzes „Standard“ (siehe Abbildung C.3). Die Ergebnisse wenn auch der eigene Datensatz als Trainingsdatensatz miteinbezogen wird, sind auch quantitativ mit bis zu 98,2 % Genauigkeit in einem ähnlichen Bereich wie der Trainingsdatensatz (98,6 %). Dies ist aber ein unrealistisches Nutzerszenario, denn eine umfangreiche Nachverarbeitung ist bei dem Trennungsvorgang nicht sinnvoll, wenn dieses für jedes Gebäude einzeln durchgeführt werden müsste. Die Ergebnisse mit der gleichen Nachverarbeitung wie beim Testdatensatz erreicht eine Genauigkeit von 93,2 %. Interessant ist hierbei, dass der originale Tuggener-Ansatz und der Ansatz von repodiac ca. 10 % schlechter abschneiden als auf dem Trainingsdatensatz. Der Unterschied zwischen angepasstem Tuggener-Ansatz und dem Stinner-Ansatz ist weitgehend gleich wie beim Trainingsdatensatz (2 bis 2,5 %).

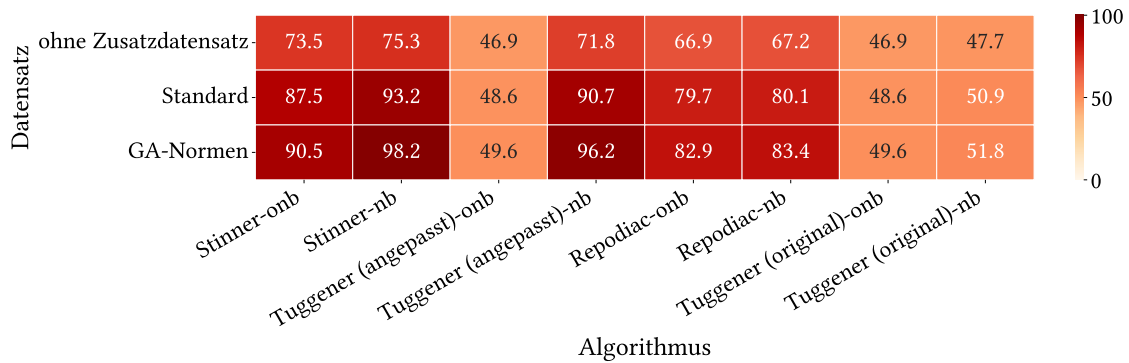
Wenn beachtet wird, dass die Nachbearbeitung in dieser Dissertation entwickelt wurde, konnte bei diesem Experiment der Stand der Technik von 82,9 % auf 98,2 % gesteigert werden.

Da alle Ergebnisse darauf hindeuten, dass der Ansatz von Stinner der aktuell beste ist. Deshalb wird dieser in dieser Arbeit entwickelte Ansatz auch für den Metadaten-Algorithmus verwendet.

C.5.1 Korrektur fehlerhaft geschriebener Wörter

Das Ziel ist einen allgemeinen Algorithmus zu trainieren, der gültige Komposita mit einer hohen Sicherheit trennen kann und nicht gültige Komposita mit einer hohen Sicherheit nicht trennt. Um

Abbildung C.3: Genauigkeit der Ergebnisse des Dokumentensplitters für den Datensatz „Gebäudeautomation-Normen“: „nb“ steht für eine durchgeführte Nachbearbeitung, „onb“ steht für keine Nachbearbeitung.



dieses Problem zumindest für lange Wörter bzw. Komposita zu lösen, wird der Algorithmus von Stinner (siehe Kapitel 3.3.5.2.3) verwendet. Jedes Wort, was dieser Algorithmus nicht trennen kann und länger als 10 Zeichen ist, wird als Kandidat für ein fehlerhaft geschriebenes Wort erachtet.

Falls ein potentiell fehlerhaftes Wort erkannt wurde, so wird zuerst geschaut, ob Fall 1 eintritt, der der am häufigsten anzutreffende Fehler ist. Also werden alle Wörter gesucht, die den Kandidaten beinhalten. In unserem Beispiel würden alle Wörter gesucht, die „Bürogeb“ enthalten. Beispielhaft könnten dies die Wörter „Bürogebäude“, „Bürogebäuden“ und „Bürogebäudekomplex“ sein. Falls es mögliche Wörter gibt, so wird das mit den wenigsten Zeichen ausgewählt (in unserem Beispiel: „Bürogebäude“).

Falls kein potentielles Wort gefunden wird, so wird das Wort mithilfe eines iterativen Trenners getrennt. Der Trenner geht von vorne bis hinten und von hinten nach vorne ein Wort durch und überprüft welcher Ansatz am meisten Trennungen durchführt. Auf Basis dieser Trennungen wird jeder Datenblock überprüft, ob er in dieser Form existiert. Falls diese Form nicht existiert, werden zuerst alle Wörter mithilfe der Levenshtein Distanz (siehe Kapitel 2.3.4) überprüft, welche die ersten drei Buchstaben teilen. Die Levenshtein Distanz wird verwendet, da diese für die Fälle 3-9 eine passende Distanz darstellt, da hier jeweils nur einzelne Zeichen ausgetauscht werden müssen. Die Levenshtein Distanz wird so verwendet, dass das Ersetzen von Zeichen mit dem Faktor 2 bewertet wird, das entfernen und einfügen wird jeweils mit dem Faktor 1 bewertet. Diese Ungleichbehandlung soll dem Sorge tragen, dass die Behebung der Fehler 3-8 keine direkte Ersetzung von Zeichen benötigt, sondern immer ein Löschen und Einfügen von Zeichen braucht.

Wenn auch bei der Anwendung der Levenshtein Distanz kein Treffer erzielt wird, so werden dann alle Wörter verwendet, die die ersten zwei bzw. den ersten Buchstaben mit dem Kandidat teilen. Falls auch hier kein Wort gefunden wird, so wird das erste Zeichen als potentiell fehlerhaftes Zeichen erachtet und ein Wort gesucht, dass alle anderen Zeichen bis auf den ersten mit einem Wort teilt.

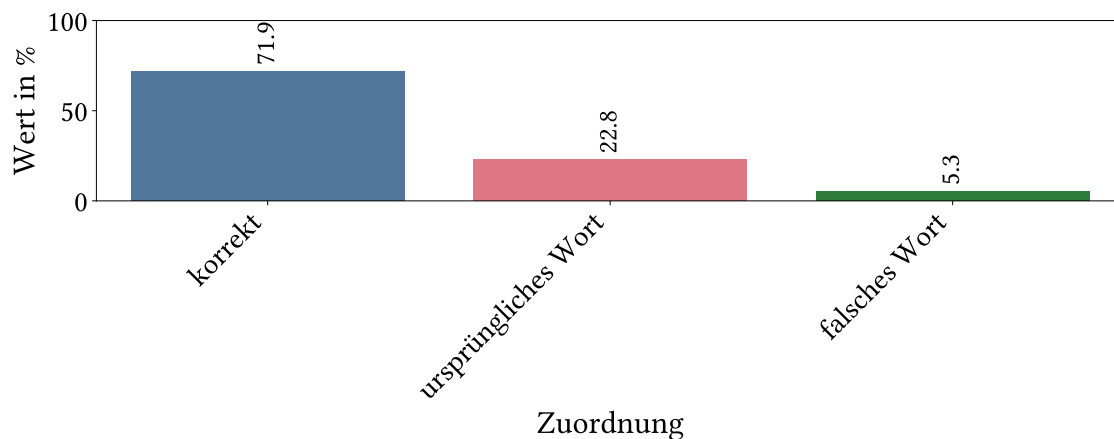
Gesetz dem Fall, dass auch hier kein Wort gefunden wird, so wird das Wort nicht getrennt. Es kann auch ein Fachwort oder Eigenname sein, das oder der sich noch nicht im Wörterbuch befinden. Auch kann nicht ausgeschlossen werden, dass es sich um fremdsprachige Wörter handelt.

In den Ergebnissen werden zuerst die Ergebnisse anhand des Test-Datensatzes und darauf aufbauend anhand des OOM4ABDO Datensatzes gezeigt.

C.5.2 Ergebnisse Korrektur fehlerhaft geschriebener Wörter

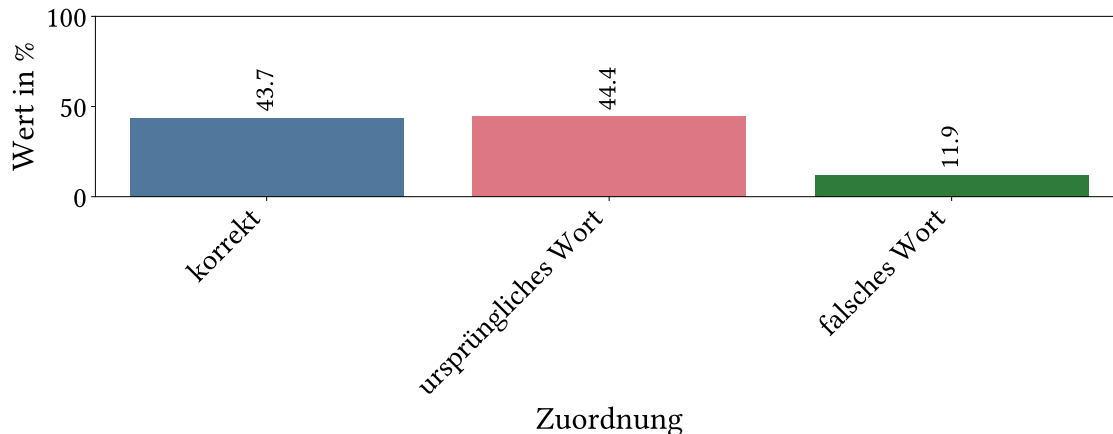
Es wurden im Trainingsdatensatz insgesamt 281 fehlerhaft geschriebene Wörter untersucht. Im Trainingsdatensatz wurden fehlerhafte Wörter zu 71,9 % durch richtige Wörter ersetzt, während in 5,3 % der Fälle ein fehlerhaftes neues Wort ausgewählt wurde. In den restlichen 22,8 % der Fälle wurde das ursprüngliche Wort beibehalten. Somit ergibt sich ein Verhältnis von richtigen Wörtern zu falschen Wörtern von 13,57.

Abbildung C.4: Ergebnisse der Tippfehlerkorrektur anhand von gesplitteten Wörtern auf Basis des Training-Datensatzes. „Korrekt“ bezeichnet, wenn das Wort korrekt geändert wurde. „Ursprüngliches Wort“ heißt, dass kein neues Wort gefunden wurde und das fehlerhafte ursprüngliche Wort weiter verwendet wird. „Falsches Wort“ bezeichnet die Änderung eines Wortes auf ein neues immer noch fehlerhaftes Wort.



Der OOM4ABDO Datensatz als Testdatensatz umfasst insgesamt 142 fehlerhaft geschriebene Wörter. Hierbei wurden nur 43,7 % der fehlerhaften Wörter korrekt zugeordnet und 11,9 % der Wörter fehlerhaft zugeordnet. In 44,4 % der Fälle wurde kein besseres Wort gefunden.

Abbildung C.5: Ergebnisse der Tippfehlerkorrektur anhand von gesplitteten Wörtern anhand des OOM4ABDO-Datensatzes. „Korrekt“ bezeichnet, wenn das Wort korrekt geändert wurde. „Ursprüngliches Wort“ heißt, dass kein neues Wort gefunden wurde und das fehlerhafte ursprüngliche Wort weiter verwendet wird. „Falsches Wort“ bezeichnet die Änderung eines Wortes auf ein neues immer noch fehlerhaftes Wort.



D Merkmalsextraktion von Zeitreihen

Im Folgenden werden Methoden zur Merkmalsextraktion bei Zeitreihen vorgestellt, die in der Literatur gute Ergebnisse bei Zeitreihenanalyse bzw. explizit bei Energiedaten geliefert haben.

Iddianozie u. Palmes [202] verwenden bildbasierte Verfahren wie Gramian Angular Fields, aber auch Cosinus Ähnlichkeit und Clustering um Zeitreihen voneinander zu unterscheiden. Es werden 10 bzw. 12 semantische Typen voneinander unterschieden bei 2200 und 1500 Datenströmen. Bildbasierte Verfahren sind sehr aufwändig und liefern nicht unbedingt bessere Ergebnisse als signalbasierte Verfahren. Es wird aber ein Verfahren auf Basis der Jaccard Distanz als Alternative der Cosinus Ähnlichkeit für die Merkmalsextraktion integriert.

Wavelet Transformationen sind sehr beliebt bei der Merkmalsextraktion, da sie die Frequenzebene und die Zeitebene betrachten [289], weshalb diese hier genutzt werden.

Approximate Entropy, Wavelet Energy und Shannon Entropy liefern gute Features für die Fehleranalyse bei elektrischen Maschinen [485, 177, 236] und zeigen die Systemkomplexität der Zeitreihen an [356]. Auch weitere Entropien, wie die Phase Entropie [391] sind in der Bibliothek NeuroKit2 [301] enthalten, die für die weiteren Entropien verwendet wurde. Außerdem wurde sie genutzt, wenn die Implementierung schneller war als die von TSFresh.

Taguchi entwickelte ein Signal-Störgrößenverhältnis (signal to noise) [438, 254], das in der Vorhersage vom Wärmebedarf von Gebäuden gute Ergebnisse zeigte [416]. Deshalb wurde dieses als Merkmal mit integriert.

Short-Time Fourier Transformation (STFT) wird insbesondere angewandt, damit die Zeitfrequenz

mit der Frequenzebene verbunden wird, was insbesondere bei ähnlichen Nutzungsmustern einen Vorteil bietet [507]. Li u. a. [275] zeigten auch mit STFT bei der Topologieerkennung von Gebäudedaten (z.B. Verbindungen von Raumtemperaturen und Luftvolumenströmen zu Räumen) gute Ergebnisse. Die STFT wurde auch als Merkmal extrahiert.

Vollmer u. a. [457] stellen 67 Merkmale vor, die sie für die Vorhersage von Fehlern in der elektrischen Energiedomäne genutzt haben, die alle in TSFast integriert wurden.

Merkmalsextrahierer wie Symbolic Fourier Approximation (SFA) [401] oder Symbolic Aggregate approximation (SAX) [285], die eine Zeitreihe auf Basis von Wörterbüchern in eine Textfolge überführen, zeigen eine hohe Genauigkeit, haben aber auch eine hohe Ausführungsdauer [28]. SAX wurde in einer eigens entwickelten Version in TSFast übernommen, die 10.000-fach schneller ist als die bisherige Implementierung. Für die Senkung der Ausführungszeit wurde der Algorithmus konsequent in Numpy umgeschrieben und auf numerische Werte statt Zeichen basiert.

Auch die verwendete Programmiersprache hat einen entscheidenden Einfluss auf die Energieeffizienz einer Anwendung [354]. Je näher die Ausführung am Maschinencode ist, desto ressourcenschonender ist sie. Deshalb wurde bei der Programmierung darauf geachtet, dass der größte Teil des Codes möglichst maschinennah in C ausgeführt wird.

D.1 TSFast

TSFast basiert auf Ray [323] zur verteilten Ausführung der Merkmalsextrahierer. Jedes NumPy-Array kann sequentiell genutzt werden, so dass die Ausführung nicht auf dem Arbeitsspeicher beschränkt ist bzw. an diesen angepasst werden kann.

In Tabelle D.1 werden die Merkmale des im Rahmen dieser Dissertation entstandenen Werkzeug zur Merkmalsextraktion „TSFast“ abgebildet. Es sind die Merkmale, deren Anzahl und die entsprechenden wissenschaftlichen Quellen abgebildet.

In Abbildung D.1 ist ein Verhältnis der Ausführungszeiten vom Originalalgorithmus gegenüber TSFast bei unterschiedlichen Zeitreihenarten (multivariat und univariat) aufgeführt.

Es gab insgesamt in 84 % der Fälle eine Verbesserung der Ausführungszeit, das aber auch bedeutet, dass es in 16 % der Fälle keine Verbesserung gab. In diesen Experimenten war meistens eine Zeitreihenart nahe an der gleichen Ausführungszeit wie der Originalalgorithmus und bei der anderen Zeitreihenart jedoch deutlich unterhalb der Ausführungszeit des Originalalgorithmus. In 30 % der Fälle konnte die Zeit auf 10 % der Zeit des Originalalgorithmus gesenkt werden, in 6 % sogar auf nur 1 % der Ausführungszeit. Im Mittel konnte die Ausführungszeit auf 2,5 % des Originals gesenkt werden, das

Abbildung D.1: Verhältnis der Ausführungszeiten des Werkzeuges „TSFast“. Die x-Achse (Anzahl der Observationen) ist logarithmisch aufgetragen.

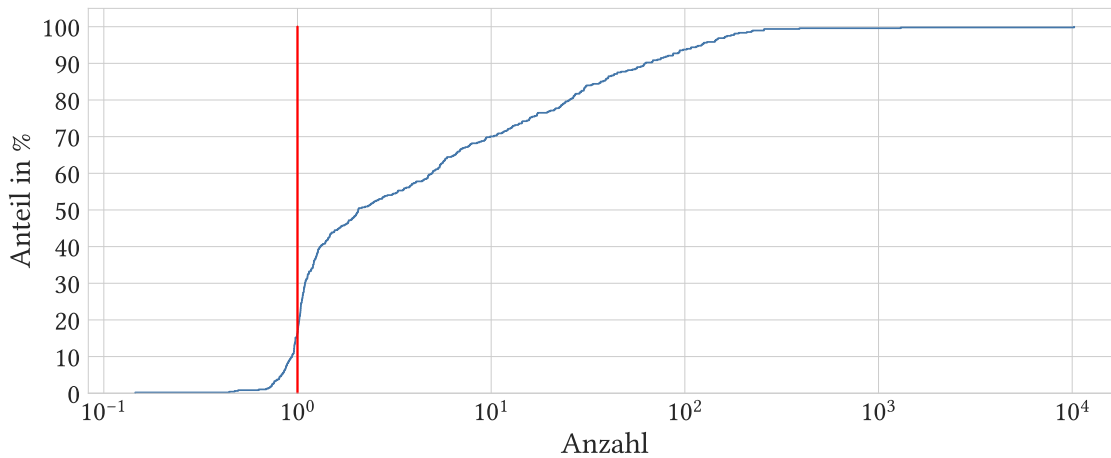


Tabelle D.1: Merkmale, die bei TSTask enthalten sind und speziell für TSTask implementiert wurden. Die Merkmale von TSFresh und TSFEL wurden in dieser Tabelle nicht gesondert aufgeführt.

Merkmals	Herkunft	n Merkmale	Wiss. Quellen
abs_diff_einsum	tsfast	1	[122]
agg_corrcoef	tsfast	5	[351]
agg_cov	tsfast	5	[210]
agg_digit	tsfast	6	
agg_divide	tsfast	5	
agg_second_derivation	tsfast	1	
agg_stft	tsfast	1	[275, 498]
agg_subtract	tsfast	5	
attention_entropy	tsfast	1	[488, 301]
average_mono_decrease_slope	tsfast	1	[11]
average_mono_increase_slope	tsfast	1	[11]
average_stationarity_slope	tsfast	1	[11]
bubble_entropy	tsfast	1	[304]
burstiness	tsfast	1	[217]
center_of_gravity	tsfast	1	[29]
changepoint	tsfast	1	[19]
corrcoef	tsfast	1	[351]
cov	tsfast	1	[210]
crest_factor	tsfast	1	[58, 457]

Fortsetzung auf nächster Seite

Tabelle D.1: Merkmale, die bei TSFast enthalten sind und speziell für TSFast implementiert wurden. Die Merkmale von TSFresh und TSFEL wurden in dieser Tabelle nicht gesondert aufgeführt.

Merkmals	Herkunft	n Merkmale	Wiss. Quellen
crm	tsfast	1	[218]
cum_hist	tsfast	1	
cumulativeresidual_entropy	tsfast	1	[380, 301]
cvmse	tsfast	1	[218]
decompose	tsfast	1	
dfa	tsfast	1	[173, 145]
difference_maxs_half	tsfast	1	[11]
difference_means_half	tsfast	1	[11]
difference_medians_half	tsfast	1	[11]
difference_mins_half	tsfast	1	[11]
difference_stds_half	tsfast	1	[11]
difference_vars_half	tsfast	1	[11]
differential_entropy	tsfast	1	[118, 301]
dispersion_entropy	tsfast	1	[396, 301]
energy_distance	tsfast	1	[388]
events	tsfast	1	
fractal_hurst	tsfast	1	[200, 301]
fractal_katz	tsfast	1	[226, 301]
fractal_linelength	tsfast	1	[126, 301]
fractal_nld	tsfast	1	[221, 301]
fractal_petrosian	tsfast	1	[355, 301]
fractal_sevcik	tsfast	1	[407, 301]
grid_entropy	tsfast	1	[481, 301]
hjorth	tsfast	1	[185, 301]
increment_entropy	tsfast	1	[290, 301]
index_of_agreement	tsfast	1	[218]
kendall_rank_correlation_coefficient	tsfast	1	[218]
kl_entropy	tsfast	1	[155, 301]
knee	tsfast	1	[301]
kurtosis_difference	tsfast	1	
kurtosis_gradient	tsfast	1	[11]
last	tsfast	1	[11]
last_k	tsfast	1	[11]
linear_weighted_average	tsfast	1	[11]

Fortsetzung auf nächster Seite

Tabelle D.1: Merkmale, die bei TSFast enthalten sind und speziell für TSFast implementiert wurden. Die Merkmale von TSFresh und TSFEL wurden in dieser Tabelle nicht gesondert aufgeführt.

Merkmal	Herkunft	n Merkmale	Wiss. Quellen
longest_group	tsfast	1	[301]
longest_monotonic_decrease	tsfast	1	[11]
longest_monotonic_increase	tsfast	1	[11]
longest_negative_run	tsfast	1	[11]
longest_positive_run	tsfast	1	[11]
longest_stationarity	tsfast	1	
longest_strike_above_mean_position	tsfast	1	
longest_strike_below_mean_position	tsfast	1	
lyapunov	tsfast	1	[395, 301]
matrix_repeat	tsfast	1	
max_difference	tsfast	1	
max_gradient	tsfast	1	[11]
maximum_absolut_error	tsfast	1	[218]
maximum_absolute_percentage_error	tsfast	1	
maximum_entropy	tsfast	1	[412]
mean_absolut_error	tsfast	1	[218]
mean_absolute_normalized_mae	tsfast	1	[218]
mean_absolute_percentage_error	tsfast	1	[218]
mean_bias_error	tsfast	1	[218]
mean_difference	tsfast	1	
mean_gradient	tsfast	1	[11]
mean_last_k	tsfast	1	[11]
mean_local_maxima_value	tsfast	1	[11]
mean_local_minima_value	tsfast	1	[11]
mean_square_error	tsfast	1	[218]
median_absolute_percentage_error	tsfast	1	
median_difference	tsfast	1	
median_gradient	tsfast	1	[11]
min_difference	tsfast	1	
min_gradient	tsfast	1	[11]
mode	tsfast	1	
modelling_efficiency	tsfast	1	[218]
moment	tsfast	3	
normalised_mean_absolute_error	tsfast	1	[218]

Fortsetzung auf nächster Seite

Tabelle D.1: Merkmale, die bei TSFast enthalten sind und speziell für TSFast implementiert wurden. Die Merkmale von TSFresh und TSFEL wurden in dieser Tabelle nicht gesondert aufgeführt.

Merkmale	Herkunft	n Merkmale	Wiss. Quellen
normalised_mean_bias_error	tsfast	1	[218]
normalised_mean_square_error	tsfast	1	[218]
nrmse	tsfast	1	[218]
number_crossing_mean	tsfast	1	[457]
number_extrema	tsfast	1	[11]
number_local_maxima	tsfast	1	[11]
number_local_minima	tsfast	1	[11]
number_maxima	tsfast	1	[457]
number_mean_local_maxima_upsurges	tsfast	1	[11]
number_mean_local_minima_downslides	tsfast	1	[11]
number_minima	tsfast	1	[457]
ofentropy_entropy	tsfast	1	[195, 301]
outlier	tsfast	1	[301]
peaks	tsfast	1	[301]
pearson_correlation_coefficient	tsfast	1	[351, 218]
percentage_in_confidence_interval	tsfast	3	
permutation_entropy_nk	tsfast	1	[181, 301]
phase_entropy	tsfast	1	[391, 301]
plateau	tsfast	1	[301]
power_entropy	tsfast	1	[164, 301]
quadratic_weighted_average	tsfast	1	[473, 457, 11]
r_squared	tsfast	1	[218]
range_entropy	tsfast	1	[339, 301]
range_evolve	tsfast	1	[145]
range_normalised_mean_bias_error	tsfast	1	[218]
range_normalized_mean_absolute_error	tsfast	1	[218]
range_normalized_root_mean_square_error	tsfast	1	[218]
rate_entropy	tsfast	1	[312, 301]
ratio_off	tsfast	1	[457]
ratio_on	tsfast	1	[457]
ratio_on_off	tsfast	1	[457]
renyi_entropy	tsfast	1	[384, 301]
rmseiqr	tsfast	1	[218]
rmsle	tsfast	1	[218]

Fortsetzung auf nächster Seite

Tabelle D.1: Merkmale, die bei TSFast enthalten sind und speziell für TSFast implementiert wurden. Die Merkmale von TSFresh und TSFEL wurden in dieser Tabelle nicht gesondert aufgeführt.

Merkmal	Herkunft	n Merkmale	Wiss. Quellen
root_mean_square_error	tsfast	1	[218]
root_mean_square_error_percentage	tsfast	1	[218]
sax	tsfast	1	[285, 346]
sd_ratio	tsfast	1	[218]
signal_to_noise_taguchi	tsfast	1	[438, 254, 416]
skew_difference	tsfast	1	
skew_gradient	tsfast	1	[11]
slope_entropy	tsfast	1	[97, 301]
slope_of_longest_mono_decrease	tsfast	1	[11]
slope_of_longest_mono_increase	tsfast	1	[11]
spearman_correlation_coefficient	tsfast	1	[218]
standard_deviation_difference	tsfast	1	
standard_deviation_gradient	tsfast	1	[11]
stationarity_percentage	tsfast	1	
stationary_average	tsfast	2	
string_distance	tsfast	1	[285, 268]
sum_difference	tsfast	1	
sum_gradient	tsfast	1	[11]
sum_last_k	tsfast	1	[11]
sum_of_squared_errors	tsfast	1	[218]
svd_entropy	tsfast	1	[390, 301]
symbolicdynamic_entropy	tsfast	1	[308, 301]
theils_coefficients_of_inequality	tsfast	1	[218]
tmax	tsfast	4	
tmean	tsfast	6	
tmin	tsfast	6	
torque	tsfast	1	
trends	tsfast	1	
tsallis_entropy	tsfast	1	[446, 301]
tsem	tsfast	6	
tstd	tsfast	6	
tvar	tsfast	6	
wasserstein	tsfast	1	[451]
zero_point_five_percentage_error	tsfast	1	[218]
Fortsetzung auf nächster Seite			

Tabelle D.1: Merkmale, die bei TSFast enthalten sind und speziell für TSFast implementiert wurden. Die Merkmale von TSFresh und TSFEL wurden in dieser Tabelle nicht gesondert aufgeführt.

Merkmals	Herkunft	n Merkmale	Wiss. Quellen
zmae	tsfast	1	[218]
zrmse	tsfast	1	[218]

E Statistiken

E.1 Statistische Merkmale Datensatz

Tabelle E.1: Statistische Merkmale der Datensätze. Für kursiv markierte Merkmale sind Übersetzungen der Blöcke notwendig (höherer Personalaufwand).

Statistisches Merkmal	ERC	Aachen_1	Aachen_2	Leipzig	München
Gunning Fog	10.93	7.59	7.47	7.52	7.82
ARI	10.26	6.99	6.23	4.26	4.69
Shannon_text	8.46	8.06	8.54	8.14	8.44
<i>Shannon_trans</i>	5.92	6.07	6.05	5.87	6.20
Simpson_text	0.99	0.98	0.99	0.99	0.99
<i>Simpson_trans</i>	0.96	0.96	0.97	0.96	0.97
Linsear_Write	14.34	10.19	10.21	8.77	10.50
Smog	12.46	7.30	8.02	7.55	8.65
Spache	11.80	10.92	10.63	11.25	10.78
TTR	0.04	0.01	0.02	0.02	0.02
CorrTTR	12.50	6.08	10.93	7.86	6.03
<i>lex. Dichte</i>	0.62	0.50	0.35	0.42	0.57
<i>lex. Satzdichte</i>	6.50	3.50	4.00	5.00	10.00
Zipf_max_diff	0.94	0.92	0.95	0.93	0.94
Zipf_txt_var	0.01	0.01	0.01	0.01	0.01
<i>Zipf_trans_var</i>	0.01	0.01	0.01	0.01	0.01
$ n_{w,i} $	0.03	0.01	0.01	0.01	0.00
KW_i	1.66	1.80	1.26	0.69	0.98
KWD_i	0.07	0.12	0.05	0.09	0.15
KD_i	27.30	21.63	22.57	13.23	15.74
KDD_i	0.04	0.02	0.01	0.04	0.04
WD_i	16.45	12.00	17.86	19.14	16.03
WDD_i	0.63	0.17	0.31	0.42	0.24
ZW_i	3.99	4.12	3.78	3.39	3.63
ZK_i	2.40	2.28	2.99	4.90	3.69
ZD_i	65.64	49.39	67.42	64.85	58.13

Tabelle E.2: Statistische Merkmale der Datensätze. Für kursiv markierte Merkmale sind Übersetzungen der Blöcke notwendig (höherer Personalaufwand).

Datensatz	ERC	Aachen_1	Aachen_2	Leipzig	München
F1 Wert	0.98	0.99	0.43	0.61	0.53
flesch_kincaid	9.31	4.96	4.67	3.43	5.47
flesch	71.61	91.72	92.68	102.20	87.20
gunning_fog	10.93	7.59	7.47	7.52	7.82
coleman_liaw	6.78	6.35	5.76	3.09	3.74
dale_chall	18.81	19.38	18.51	19.92	19.09
ari	10.26	6.99	6.23	4.26	4.69
linsear_write	14.34	10.19	10.21	8.77	10.50
smog	12.46	7.30	8.02	7.55	8.65
spache	11.80	10.92	10.63	11.25	10.78
lex_dichte	0.62	0.50	0.35	0.42	0.57
lex_satzdichte	6.50	3.50	4.00	5.00	10.00
TTR	0.04	0.01	0.02	0.02	0.02
CorrTTR	12.50	6.08	10.93	7.86	6.03
Guiraud	17.68	8.59	15.45	11.11	8.52
Herdan	0.73	0.67	0.70	0.69	0.67
Uber	46.14	38.54	45.54	40.64	38.31
var_freq_types	62330.20	377930.36	499724.49	173769.13	208920.40
coeff_of_var	9.63	8.83	12.13	9.04	6.88
coeff_of_var_mean	0.02	0.01	0.01	0.02	0.01
single_vocabulary	0.03	0.01	0.01	0.01	0.00
shannon_text	8.46	8.06	8.54	8.14	8.44
shannon_translation	5.92	6.07	6.05	5.87	6.20
simpson_text	0.99	0.98	0.99	0.99	0.99
simpson_translation	0.96	0.96	0.97	0.96	0.97
Zipf_txt_var	0.01	0.01	0.01	0.01	0.01
Zipf_trans_var	0.01	0.01	0.01	0.01	0.01
Zipf_txt_var_abs	0.01	0.01	0.01	0.01	0.01
Zipf_trans_var_abs	0.01	0.01	0.01	0.01	0.01
Zipf_txt_sum	5.13	5.08	5.12	5.09	5.11
Zipf_trans_sum	4.90	4.90	4.90	4.90	4.91
Zipf_txt_std	0.08	0.08	0.08	0.08	0.08
Zipf_trans_std	0.08	0.07	0.08	0.07	0.08
Zipf_txt_std_abs	0.08	0.08	0.08	0.08	0.08
Fortsetzung auf nächster Seite					

Tabelle E.2: Statistische Merkmale der Datensätze. Für kursiv markierte Merkmale sind Übersetzungen der Blöcke notwendig (höherer Personalaufwand).

Datensatz	ERC	Aachen_1	Aachen_2	Leipzig	München
Zipf_trans_std_abs	0.08	0.07	0.08	0.07	0.08
Zipf_max_diff	0.94	0.92	0.95	0.93	0.94
Zipf_euclidean	1.18	1.16	1.19	1.17	1.19
Zipf_chebyshev	0.94	0.92	0.95	0.93	0.94
Zipf_manhattan	5.13	5.08	5.12	5.09	5.11
Zipf_minkowski	1.18	1.16	1.19	1.17	1.19
Zipf_jaccard	1.00	1.00	1.00	1.00	1.00
Zipf_hamming	1.00	1.00	1.00	1.00	1.00
Zipf_dice	0.96	0.95	0.96	0.96	0.96
Zipf_cosine	0.06	0.02	0.10	0.05	0.09
Zipf_jensenshannon	0.12	0.11	0.14	0.11	0.17
Zipf_nmae	7.75	7.80	7.87	7.51	7.64
Zipf_n_max_error	1.42	1.41	1.46	1.37	1.41
Zipf_nmse	0.01	0.01	0.01	0.01	0.01
Zipf_nrmse	0.13	0.13	0.13	0.12	0.13
tags_to_blocks_ratio	1.66	1.80	1.26	0.69	0.98
tags_to_blocks_diversity_ratio	0.07	0.12	0.05	0.09	0.15
tags_to_datastream_ratio	27.30	21.63	22.57	13.23	15.74
tags_to_datastream_diversity_ratio	0.04	0.02	0.01	0.04	0.04
blocks_to_datastream_ratio	16.45	12.00	17.86	19.14	16.03
blocks_to_datastream_diversity_ratio	0.63	0.17	0.31	0.42	0.24
chars_to_blocks_ratio	3.99	4.12	3.78	3.39	3.63
chars_to_tag_ratio	2.40	2.28	2.99	4.90	3.69
chars_to_datastream_ratio	65.64	49.39	67.42	64.85	58.13
own_median	0.77	0.55	0.01	-0.80	-0.53
own_mean	0.87	0.64	0.04	-0.95	-0.59
own_pearson	0.19	0.23	-0.22	-0.08	-0.13

E.2 Bewertung von statistischen Merkmalen als prädiktive Vorhersage von Ergebnissen

Zur Bewertung der statistischen Merkmale wird geprüft, ob sie eine zuverlässige Vorhersage des F1-Werts für die Metadatenanalyse ermöglichen. Die fünf Datensätze werden jeweils in vier Trainings- und einen Testdatensatz aufgeteilt. Zusätzlich wird jeweils ein Datensatz entfernt und die ver-

bleibenden vier in drei Trainings- und einen Testdatensatz unterteilt, was insgesamt zehn Datensatzkombinationen ergibt. Diese Vorgehensweise dient dazu, Ausreißer unter den Datensätzen zu identifizieren.

Die Trainings- und Testdaten werden als Eingabe für die in Kapitel 2.3.3 beschriebenen Regressionsalgorithmen genutzt. Jeder Algorithmus wird zehnmal ausgeführt, um Variabilitäten auszugleichen. Die Differenz zwischen vorhergesagtem und tatsächlichem F1-Wert bestimmt die Vorhersagequalität. Liegt die Differenz in allen fünf Datensätzen bei maximal 10 %, besteht ein statistischer Zusammenhang.

E.3 Ergebnis von statistischen Merkmalen als prädiktive Vorhersage von Ergebnissen

In diesem Kapitel wird überprüft, ob die Metadaten-Kennzeichnung bzw. deren statistische Merkmale einen Einfluss auf die Ergebnisse des Metadaten Algorithmus hat. Merkmale ohne Kennzeichen können auch ohne menschlichen Aufwand im Vorhinein berechnet und überprüft werden. Ziel ist es zu überprüfen, ob im Vorhinein erkannt werden kann, ob die Nützlichkeit der Investition teurer menschlicher Arbeit in die Übersetzung eines Datensatzes a priori bewertet werden kann.

Deshalb wird in einem ersten Schritt überprüft, ob die Nutzung der Kennzeichen, die einen hohen Personalaufwand haben, eine Verbesserung bietet. In den Abbildungen E.1 und E.2 sind die absoluten Differenzen von Vorhersage und realen Ergebnissen mit bzw. ohne Verwendung der Kennzeichen abgebildet. Es ist erkennbar, dass in zwei von sechs Algorithmen der Algorithmus ohne Kennzeichen, in einem der Algorithmus mit Kennzeichen und zwei mal die Ergebnisse annähernd konstant ausfielen. Somit kann gesagt werden, dass zumindest die Nutzung von kennzeichenbezogenen Merkmalen aufgrund des erhöhten Personalaufwands fraglich ist. Insgesamt ist die Abweichung von Vorhersage und realen Ergebnissen mit im Median 25,5 % bis 41,6 % F1-Wert zu hoch.

Um zu Überprüfen, ob einzelne Datensätze als Ausreißer einen großen Einfluss auf das Ergebnis haben, wird jeweils ein Datensatz gelöscht. Das heißt, dass nur noch drei Datensätze zum Training und ein Datensatz zum Testen verwendet werden. Durch dieses Vorgehen kann der Median der Differenz zum realen F1 Wert durch die Löschung des Datensatzes „Stadt Aachen 1“ auf minimal 13,4 bis maximal 18,4 % reduziert werden, was die größte Senkung ist.

Somit kann zwar gezeigt werden, dass es insgesamt Ausreißer bei den Ergebnissen der Anwendung von Regression bei der Metadatenkennzeichnung und deren Ergebnissen existieren. Aber die Ergebnisse weichen trotz Löschung um minimal 13,4 % ab. Das zeigt, dass die statistischen Merkmale auf Basis der Metadaten keine gute a priori Vorhersage liefern wie gut ein Algorithmus zur Metadatenenerkennung auf einem Datensatz abschneidet. Diese Ergebnisse bestätigen weitestgehend die Ergebnisse aus Kapitel 4.1.1.4 bei der Eingrenzung der Nutzung der Merkmale. Die Aussage ist aber

aufgrund der begrenzten Anzahl an verfügbaren Datensätzen nur beschränkt auf viele Datensätze anwendbar.

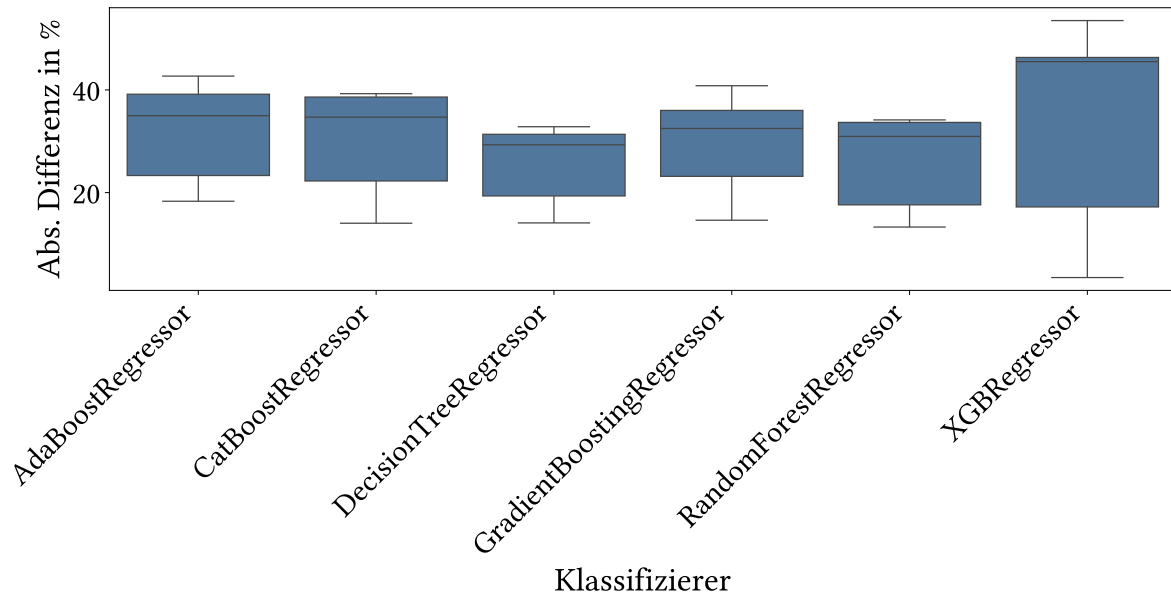


Abbildung E.1: Ergebnisse der Modelle bei Training der statistischen Merkmale der Metadaten der Datensätze mit Nutzung von kennzeichenenthaltenden Merkmalen.

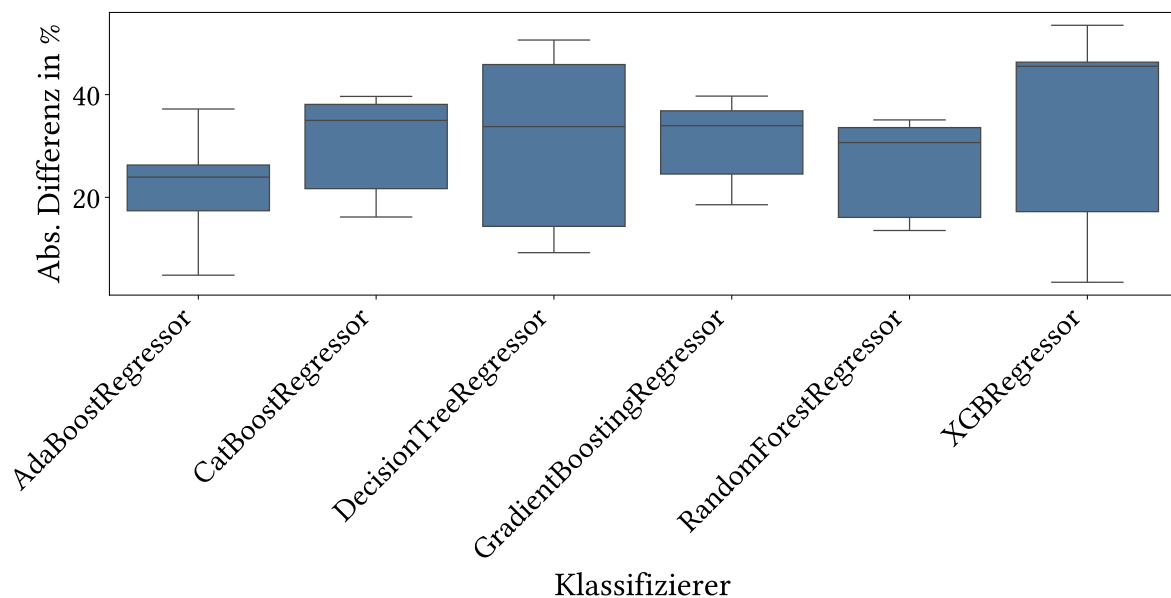


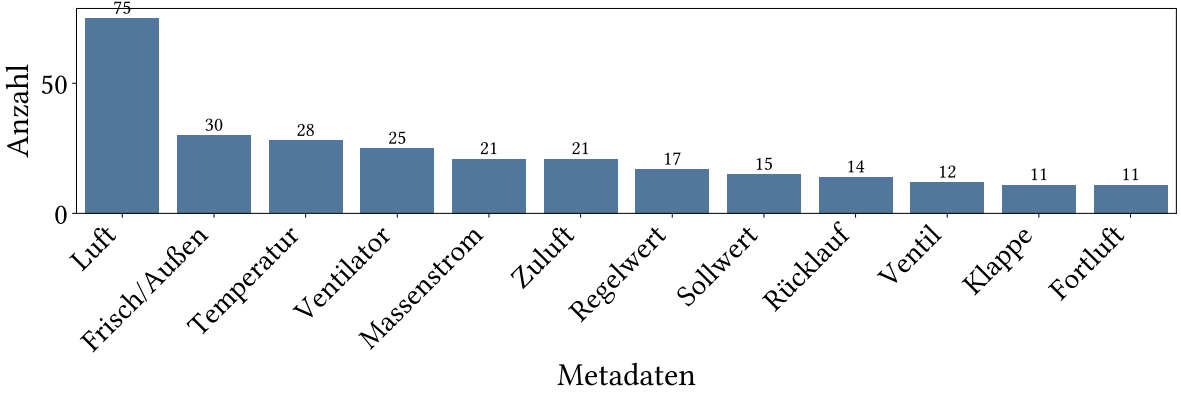
Abbildung E.2: Ergebnisse der Modelle bei Training der statistischen Merkmale der Metadaten der Datensätze ohne die Nutzung von kennzeichenenthaltenden Merkmalen.

E.4 Statistiken Anwendungen

E.4.1 RLT Anlagen

Im Datensatz von Fehlern in RLT Anlagen von Li u. a. [273] tauchen mit über 10 Observationen die in Abbildung E.3 abgebildeten Metadaten auf. Für die RLT Anlage typisch wird die Eigenschaft „Luft“ mit 75-mal am häufigsten eingesetzt. Die Außenbedingungen folgen mit 30-mal. Auch Temperaturen spielen direkt danach eine große Rolle. Die hohe Anzahl an Massenströmen (21-mal) ist in der Praxis kritisch zu bewerten. In der Praxis wird der Massenstrom nicht in RLT-Anlagen gemessen. Dort wird höchstens der Volumenstrom gemessen bzw. errechnet.

Abbildung E.3: Statistiken über die enthaltenen Metadaten der Quellen für die Fehlererkennung aus Li u. a. [273]



In Tabelle E.3 sind die benötigten Metadaten und die entsprechenden Quellen für die Fehlererkennung bei RLT Anlagen aus Li u. a. [273] abgebildet.

Tabelle E.3: Benötigte Metadaten für die einzelnen Fehlererkennungsansätze aus Li u. a. [273].		
Index	Quellen	Benötigte Metadaten
1	[86, 85]	(Leistung, Erhitzer),
2	[86, 85]	(Frisch/Außen, Variable Luftvolumeneinheit, Klappe),
3	[86, 85]	(Volumenstrom, Luft, Zuluft),
		(Frisch/Außen, Variable Luftvolumeneinheit, Klappe),
4	[86, 85]	(Leistung, Erhitzer),
		(Temperatur, Luft, Zuluft),
5	[86, 85]	(Temperatur, Luft, Ventilator, Zuluft),
6	[86, 85]	(Leistung, Erhitzer),
Fortsetzung auf nächster Seite		

Tabelle E.3: Benötigte Metadaten für die einzelnen Fehlererkennungsansätze aus Li u. a. [273].

Index	Quellen	Benötigte Metadaten
7	[86, 85]	(Temperatur, Luft, Zuluft), (Leistung, Erhitzer), (Volumenstrom, Luft, Zuluft), (Temperatur, Luft, Zuluft),
8	[86, 85]	(Leistung, Erhitzer), (Temperatur, Luft, Zuluft), (Temperatur, Luft, Ventilator, Zuluft),
9	[314]	(Differenz, Druck, Luft, Register), (Volumenstrom, Luft, Register),
10	[117]	(Massenstrom, Luft, Frisch/Außen), (Massenstrom, Luft, Zuluft), (Elektrische Leistung, Luft, Zuluft), (Massenstrom, Luft, Rücklauf),
11	[117]	(Temperatur, Luft, Zuluft), (Massenstrom, Luft, Zuluft), (Temperatur, Luft, Rücklauf), (Temperatur, Kühlwasser, Zuluft), (Temperatur, Kühlwasser, Rücklauf), (Massenstrom, Kühlwasser),
12	[117]	(Massenstrom, Luft, Frisch/Außen), (Temperatur, Luft, Zuluft), (Massenstrom, Luft, Zuluft), (Elektrische Leistung, Luft, Zuluft), (Temperatur, Luft, Rücklauf), (Massenstrom, Luft, Rücklauf),
13	[487, 486]	(Massenstrom, Luft, Austritt), (Massenstrom, Luft, Zuluft), (Drehzahl, Luft, Ventilator, Zuluft), (Massenstrom, Luft, Rücklauf), (Drehzahl, Luft, Ventilator, Rücklauf),
14	[487, 486]	(Massenstrom, Luft, Austritt), (Massenstrom, Luft, Zuluft), (Drehzahl, Luft, Ventilator, Zuluft), (Massenstrom, Luft, Rücklauf), (Drehzahl, Luft, Ventilator, Rücklauf),
		Fortsetzung auf nächster Seite

Tabelle E.3: Benötigte Metadaten für die einzelnen Fehlererkennungsansätze aus Li u. a. [273].

Index	Quellen	Benötigte Metadaten
15	[487, 486]	(Massenstrom, Luft, Zuluft), (Differenz, Temperatur, Luft, Zuluft, Istwert, Sollwert), (Massenstrom, Kühlwasser), (Differenz, Temperatur, Kühlwasser, Verdampfer), (Differenz, Enthalpie, Luft, Variable Luftvolumeneinheit, Mischluft, Zuluft), (Differenz, Temperatur, Luft, Zone, Istwert, Sollwert), (, Kühlregister),
16	[487, 486]	(Massenstrom, Luft, Austritt), (Elektrische Leistung, Luft, Zuluft), (Drehzahl, Luft, Ventilator, Zuluft), (Massenstrom, Luft, Rücklauf), (Drehzahl, Luft, Ventilator, Rücklauf),
17	[487, 486]	(Massenstrom, Luft, Zuluft), (Differenz, Temperatur, Luft, Zuluft, Istwert, Sollwert), (Drehzahl, Luft, Ventilator, Zuluft), (Massenstrom, Luft, Rücklauf), (Drehzahl, Luft, Ventilator, Rücklauf),
18	[487, 486]	(Massenstrom, Luft, Zuluft), (Differenz, Temperatur, Luft, Zuluft, Istwert, Sollwert), (Drehzahl, Luft, Ventilator, Zuluft), (Massenstrom, Luft, Rücklauf), (Drehzahl, Luft, Ventilator, Rücklauf),
19	[281]	(Temperatur, Kühlwasser, Rücklauf), (Regelwert, Kühlwasser, Ventil), (Temperatur, Luft, Mischluft),
20	[281]	(Temperatur, Kühlwasser, Rücklauf), (Regelwert, Kühlwasser, Ventil),
21	[281]	(Temperatur, Luft, Zuluft), (Temperatur, Kühlwasser, Rücklauf), (Regelwert, Kühlwasser, Ventil),
22	[309]	(Massenstrom, Luft, Zuluft), (Temperatur, Luft, Rücklauf), (Temperatur, Luft, Wärmeübertrager, nach Wärmeübertrager), (Temperatur, Luft, Wärmeübertrager, vor Wärmeübertrager),
23	[309]	(Temperatur, Luft, Wärmeübertrager, nach Wärmeübertrager),
		Fortsetzung auf nächster Seite

Tabelle E.3: Benötigte Metadaten für die einzelnen Fehlererkennungsansätze aus Li u. a. [273].

Index	Quellen	Benötigte Metadaten
		(Temperatur, Luft, Wärmeübertrager, vor Wärmeübertrager), (Massenstrom, Heizwasser, Zuluft), (Temperatur, Heizwasser, Zuluft), (Temperatur, Heizwasser, Rücklauf), (Massenstrom, Luft, Zuluft), (Massenstrom, Luft, Zuluft), (Drehzahl, Luft, Ventilator, Zuluft), (Massenstrom, Luft, Zuluft), (Drehzahl, Luft, Ventilator, Zuluft), (Drehzahl, Luft, Ventilator, Zuluft), (, Luft, Ventilator, Rücklauf), (, Luft, Ventilator, Rücklauf), (Temperatur, Luft, Frisch/Außen), (Temperatur, Luft, Zuluft), (Temperatur, Luft, Zuluft, Sollwert), (Temperatur, Kühlwasser, Zuluft), (Temperatur, Kühlwasser, Zuluft, Sollwert), (Temperatur, Kühlwasser, Rücklauf), (Temperatur, Luft, Mischluft), (Temperatur, Luft, Zone), (Temperatur, Luft, Zone, Sollwert), (Massenstrom, Heizwasser, Kessel), (Regelwert, RLT-Anlage, Ventil), (Regelwert, Heizkörper, Ventil), (Massenstrom, Heizwasser, Kessel), (Regelwert, RLT-Anlage, Ventil), (Regelwert, Heizkörper, Ventil), (Temperatur, Luft, Zuluft), (Temperatur, Kühlwasser, Zuluft), (Temperatur, Luft, Mischluft), (Regelwert, RLT-Anlage, Ventil), (Massenstrom, Luft, Zuluft), (Elektrische Leistung, Luft, Zuluft), (Differenz, Elektrische Leistung, Filter), (Frisch/Außen, Luft, Frisch/Außen, Klappe),
24	[309]	
25	[309]	
26	[309]	
27	[309]	
28	[453]	
29	[453]	
30	[453]	
31	[506, 505]	
32	[506, 505]	
33	[506, 505]	
34	[506, 505]	
Fortsetzung auf nächster Seite		

Tabelle E.3: Benötigte Metadaten für die einzelnen Fehlererkennungsansätze aus Li u. a. [273].

Index	Quellen	Benötigte Metadaten
35	[506, 505]	(Frisch/Außen, Luft, Rücklauf, Klappe),
36	[506, 505]	(Frisch/Außen, Luft, Fortluft, Klappe),
37	[506, 505]	(Drehzahl, Luft, Ventilator, Zuluft), (Leistung, Luft, Zuluft, Ventilator), (Drehzahl, Luft, Ventilator, Rücklauf), (Leistung, Luft, Ventilator, Rücklauf),
38	[506, 505]	(Drehzahl, Luft, Ventilator, Zuluft), (Leistung, Luft, Zuluft, Ventilator), (Drehzahl, Luft, Ventilator, Rücklauf), (Leistung, Luft, Ventilator, Rücklauf),
39	[506, 505]	(Massenstrom, Luft, Zuluft),
40	[506, 505]	(Differenz, Elektrische Leistung, Filter),
41	[506, 505]	(Frisch/Außen, Heizregister, Klappe),
42	[506, 505]	(Frisch/Außen, Kühlregister, Ventil),
43	[506, 505]	(Energie, Heizregister, RLT-Anlage),
44	[506, 505]	(Differenz, Elektrische Leistung, Heizwasser, Pumpe),
45	[506, 505]	(Temperatur, Heizwasser, Zuluft),
46	[506, 505]	(Energie, Heizregister, RLT-Anlage),
47	[506, 505]	(Temperatur, Luft, Zuluft),
48	[506, 505]	(Temperatur, Luft, Rücklauf),
49	[506, 505]	(Temperatur, Luft, Mischluft),
50	[506, 505]	(Temperatur, Luft, Frisch/Außen),
51	[506, 505]	(Frisch/Außen, Heizregister, Ventil),
52	[460]	(Temperatur, Luft, Zuluft), (Temperatur, Luft, Zuluft, Sollwert), (Leistung, Kühlwasser, Pumpe),
53	[460]	(Temperatur, Luft, Zuluft), (Frisch/Außen, Kühlregister, Ventil), (Energie, Kühlregister, RLT-Anlage), (Temperatur, Kühlwasser, Zuluft),
54	[460]	(Temperatur, Luft, Zuluft), (Frisch/Außen, Kühlregister, Ventil), (Temperatur, Luft, Mischluft),
55	[460]	(Temperatur, Luft, Zuluft), (Temperatur, Luft, Zuluft, Sollwert),
Fortsetzung auf nächster Seite		

Tabelle E.3: Benötigte Metadaten für die einzelnen Fehlererkennungsansätze aus Li u. a. [273].

Index	Quellen	Benötigte Metadaten
		(Frisch/Außen, Kühlregister, Ventil), (Regelwert, Kühlregister, Ventil), (Differenz, Elektrische Leistung, Filter), (Temperatur, Luft, Frisch/Außen), (Temperatur, Luft, Zuluft), (Temperatur, Luft, Rücklauf), (Temperatur, Luft, Mischluft), (Stellantrieb, Temperatur, Luft, Mischluft), (Temperatur, Luft, Frisch/Außen), (Stellantrieb, Temperatur, Luft, Frisch/Außen), (Leistung, Luft, Ventilator, Rücklauf), (Temperatur, Luft, Rücklauf), (Stellantrieb, Temperatur, Luft, Rücklauf), (Elektrische Leistung, Luft, Zuluft), (Temperatur, Luft, Zuluft), (Frisch/Außen, Kühlregister, Ventil), (Energie, Kühlregister, RLT-Anlage), (Temperatur, Kühlwasser, Zuluft), (Stellantrieb, Temperatur, Luft, Zuluft), (Elektrische Leistung, Luft, Zuluft), (Leistung, Luft, Zuluft, Ventilator), (Elektrische Leistung, Luft, Zuluft), (Leistung, Luft, Zuluft, Ventilator), (Temperatur, Luft, Zuluft), (Temperatur, Luft, Zuluft, Sollwert), (Frisch/Außen, Kühlregister, Ventil), (Regelwert, Kühlregister, Ventil), (Temperatur, Luft, Zuluft), (Temperatur, Kühlwasser, Zuluft), (Temperatur, Luft, Zuluft), (Temperatur, Kühlwasser, Zuluft), (Temperatur, Luft, Zuluft), (Temperatur, Luft, Zuluft, Sollwert), (Frisch/Außen, Kühlregister, Ventil), (Regelwert, Kühlregister, Ventil),
56	[460]	
57	[460]	
58	[460]	
59	[460]	
60	[460]	
61	[460]	
62	[460]	
63	[460]	
64	[460]	
65	[460]	
66	[460]	
67	[460]	
68	[460]	
69	[460]	
70	[460]	
71	[460]	
72	[460]	
		Fortsetzung auf nächster Seite

Tabelle E.3: Benötigte Metadaten für die einzelnen Fehlererkennungsansätze aus Li u. a. [273].

Index	Quellen	Benötigte Metadaten
73	[460]	(Temperatur, Luft, Zuluft), (Frisch/Außen, Kühlregister, Ventil), (Energie, Kühlregister, RLT-Anlage), (Temperatur, Kühlwasser, Zuluft),
74	[478]	(Temperatur, Luft, Zone), (Temperatur, Luft, Zone, Sollwert),
75	[478]	(Massenstrom, Luft),
76	[478]	(Massenstrom, Luft), (Massenstrom, Luft, Sollwert),
77	[478]	(Massenstrom, Luft), (Massenstrom, Luft, Sollwert),
78	[478]	(Elektrische Leistung, Luft, Zuluft), (Elektrische Leistung, Luft, Zuluft, Sollwert),
79	[478]	(Temperatur, Luft, Zuluft), (Temperatur, Luft, Zuluft, Sollwert),
80	[478]	(Energie, Kühlregister),
81	[478]	(Energie, Kühlregister),
82	[478]	(Temperatur, Luft, Zone), (Temperatur, Luft, Zone, Sollwert),
83	[478]	(Temperatur, Luft, Zone), (Temperatur, Luft, Zone, Sollwert),
84	[20]	(Regelwert, Luft, Frisch/Außen, Klappe), (Regelwert, Luft, Frisch/Außen, Klappe, Bypass), (Temperatur, Luft, Zuluft), (Temperatur, Luft, Fortluft), (Regelwert, Luft, Fortluft, Klappe),
85	[20]	(Regelwert, Luft, Frisch/Außen, Klappe), (Regelwert, Luft, Fortluft, Klappe), (Regelwert, Kühlregister, Ventil),
86	[20]	(Regelwert, Luft, Frisch/Außen, Klappe), (Regelwert, Luft, Frisch/Außen, Klappe, Bypass), (Temperatur, Luft, Zuluft), (Temperatur, Luft, Fortluft), (Regelwert, Luft, Fortluft, Klappe),
87	[20]	(Regelwert, Luft, Frisch/Außen, Klappe),
		Fortsetzung auf nächster Seite

Tabelle E.3: Benötigte Metadaten für die einzelnen Fehlererkennungsansätze aus Li u. a. [273].

Index	Quellen	Benötigte Metadaten
88	[20]	(Regelwert, Luft, Frisch/Außen, Klappe, Bypass), (Temperatur, Luft, Zuluft), (Temperatur, Luft, Fortluft), (Regelwert, Luft, Fortluft, Klappe), (Regelwert, Luft, Frisch/Außen, Klappe), (Regelwert, Luft, Frisch/Außen, Klappe, Bypass), (Temperatur, Luft, Zuluft), (Temperatur, Luft, Fortluft), (Regelwert, Luft, Fortluft, Klappe),
89	[20]	(Regelwert, Luft, Frisch/Außen, Klappe), (Regelwert, Luft, Frisch/Außen, Klappe, Bypass), (Temperatur, Luft, Zuluft), (Temperatur, Luft, Fortluft), (Regelwert, Luft, Fortluft, Klappe),
90	[20]	(Regelwert, Luft, Frisch/Außen, Klappe), (Temperatur, Luft, Zuluft), (Temperatur, Luft, Fortluft), (Regelwert, Luft, Fortluft, Klappe), (Regelwert, Kühlregister, Ventil),
91	[463, 462, 461]	(Frisch/Außen, Klappe), (Temperatur, Luft, Zuluft), (Temperatur, Luft, Zuluft, Sollwert), (Regelwert, Luft, Ventilator, Zuluft), (Temperatur, Luft, Zone), (Temperatur, Luft, Zone, Sollwert), (Massenstrom, Luft, Zone), (Massenstrom, Luft, Zone, Sollwert), (Massenstrom, Luft, Zone, Sollwert, Maximum), (Massenstrom, Luft, Zone, Sollwert, Minimum), (Massenstrom, Luft, Variable Luftvolumeneinheit, Sollwert), (Massenstrom, Luft, Variable Luftvolumeneinheit, Sollwert, Maximum), (Massenstrom, Luft, Sollwert, Minimum),
92	[327, 328]	(Regelwert, Klappe), (Regelwert, Ventil), (Massenstrom, Luft, Frisch/Außen),
		Fortsetzung auf nächster Seite

Tabelle E.3: Benötigte Metadaten für die einzelnen Fehlererkennungsansätze aus Li u. a. [273].

Index	Quellen	Benötigte Metadaten
		(Temperatur, Luft, Frisch/Außen), (Temperatur, Luft, Zuluft), (Massenstrom, Luft, Zuluft), (Temperatur, Luft, Rücklauf), (Massenstrom, Luft, Rücklauf), (Temperatur, Luft, Fortluft), (Temperatur, Kühlwasser, Zuluft), (Temperatur, Luft, Mischluft), (Temperatur, Heizwasser, Zuluft),
93	[477, 464]	(Massenstrom, Luft, Frisch/Außen), (Temperatur, Luft, Frisch/Außen), (Feuchtigkeit, Luft, Frisch/Außen), (Temperatur, Luft, Zuluft), (Massenstrom, Luft, Zuluft), (Feuchtigkeit, Luft, Zuluft), (Temperatur, Luft, Rücklauf), (Massenstrom, Luft, Rücklauf), (Feuchtigkeit, Luft, Rücklauf), (Regelwert, Kühlwasser, Ventil),
94	[409, 410, 411]	(Frisch/Außen, Klappe), (Frisch/Außen, Ventil), (Temperatur, Luft, Zuluft), (Elektrische Leistung, Luft, Zuluft),
95	[409, 410, 411]	(Massenstrom, Luft, Zuluft), (Temperatur, Luft, Rücklauf),
96	[423]	(Feuchtigkeit, Luft, Zuluft), (Feuchtigkeit, Luft, Rücklauf),
97	[342]	(Regelwert, Klappe), (Temperatur, Luft, Frisch/Außen), (Relative Feuchte, Luft, Frisch/Außen), (Temperatur, Luft, Zuluft), (Massenstrom, Luft, Zuluft), (Elektrische Leistung, Luft, Zuluft), (Regelwert, Luft, Ventilator, Zuluft), (Relative Feuchte, Luft, Zuluft),
		Fortsetzung auf nächster Seite

Tabelle E.3: Benötigte Metadaten für die einzelnen Fehlererkennungsansätze aus Li u. a. [273].

Index	Quellen	Benötigte Metadaten
99	[379]	(Temperatur, Luft, Rücklauf), (Massenstrom, Luft, Rücklauf), (Regelwert, Luft, Ventilator, Rücklauf), (Relative Feuchte, Luft, Rücklauf), (Regelwert, Heizregister, Ventil), (Regelwert, Kühlregister, Ventil), (Temperatur, Luft, Mischluft), (Massenstrom, Luft, Mischluft), (Temperatur, Luft, Zuluft), (Temperatur, Luft, Fortluft), (Temperatur, Kühlwasser, Rücklauf), (Temperatur, Luft, Zone), (Temperatur, Luft, Zuluft), (Temperatur, Luft, Fortluft), (Temperatur, Kühlwasser, Rücklauf), (Temperatur, Luft, Zone), (Massenstrom, Luft, Frisch/Außen), (Temperatur, Luft, Frisch/Außen), (Relative Feuchte, Luft, Frisch/Außen), (Temperatur, Luft, Zuluft), (Massenstrom, Luft, Zuluft), (Temperatur, Luft, Rücklauf), (Massenstrom, Luft, Rücklauf), (Relative Feuchte, Luft, Rücklauf), (Temperatur, Register), (Temperatur, Kühlwasser, Zuluft), (Temperatur, Kühlwasser, Rücklauf), (Massenstrom, Kühlwasser), (Relative Feuchte, Luft, Mischluft), (Massenstrom, Heizwasser, Zuluft), (Temperatur, Heizwasser, Zuluft), (Temperatur, Heizwasser, Rücklauf), (Massenstrom, Luft, Frisch/Außen), (Massenstrom, Ventilator, Luft, Frisch/Außen), (Drehzahl, Ventilator, Frisch/Außen),
100	[379]	
101	[345]	
103	[238]	
Fortsetzung auf nächster Seite		

Tabelle E.3: Benötigte Metadaten für die einzelnen Fehlererkennungsansätze aus Li u. a. [273].

Index	Quellen	Benötigte Metadaten
		(Elektrische Leistung, Luft, Ventilator, Frisch/Außen), (Regelwert, Ventilator, Fortluft), (Drehzahl, Ventilator, Fortluft), (Elektrische Leistung, Luft, Ventilator, Fortluft), (Massenstrom, Ventilator, Fortluft), (Regelwert, Ventilator, Filter, Fortluft), (Massenstrom, Ventilator, Luft, Frisch/Außen), (Elektrische Leistung, Luft, Ventilator, Frisch/Außen), (Elektrische Leistung, Luft, Ventilator, Fortluft), (Massenstrom, Ventilator, Fortluft), (Frisch/Außen, Klappe), (Temperatur, Luft, Frisch/Außen), (Temperatur, Luft, Zuluft), (Massenstrom, Luft, Zuluft), (Elektrische Leistung, Luft, Zuluft), (Drehzahl, Luft, Ventilator, Zuluft), (Leistung, Luft, Zuluft, Ventilator), (Drehzahl, Luft, Ventilator, Rücklauf), (Leistung, Luft, Ventilator, Rücklauf), (Frisch/Außen, Heizregister, Ventil), (Frisch/Außen, Kühlregister, Ventil), (Temperatur, Luft, Mischluft), (Temperatur, Luft, Frisch/Außen), (Relative Feuchte, Luft, Frisch/Außen), (Frisch/Außen, Klappe, Zuluft), (Relative Feuchte, Luft, Fortluft), (Temperatur, Kühlwasser, Zuluft), (Temperatur, Kühlwasser, Rücklauf), (Frisch/Außen, Kühlwasser, Ventil), (Temperatur, Heizwasser, Zuluft), (Temperatur, Heizwasser, Rücklauf), (Massenstrom, Heizwasser, Zuluft), (Energie, Heizregister), (Temperatur, Luft, Raum/Innen), (Temperatur, Luft, Zuluft),
104	[238]	
105	[276, 277]	
106	[490, 491, 489]	
107	[480]	
109	[480]	
110	[480]	
112	[123]	
		Fortsetzung auf nächster Seite

Tabelle E.3: Benötigte Metadaten für die einzelnen Fehlererkennungsansätze aus Li u. a. [273].

Index	Quellen	Benötigte Metadaten
113	[503]	(Regelwert, Kühlwasser, Ventil), (Temperatur, Kühlwasser, Speicher), (Temperatur, Luft, Zone), (Regelwert, Variable Luftvolumeneinheit, Ventil), (Massenstrom, Luft), (Massenstrom, Luft, Sollwert), (Frisch/Außen, Klappe), (Temperatur, Luft, Zone), (Temperatur, Luft, Zone, Sollwert), (Massenstrom, Luft, Raum/Innen), (Massenstrom, Luft, Rücklauf), (Leistung, Luft, Ventilator, Rücklauf), (Temperatur, Heizwasser, Heizregister, Austritt), (Frisch/Außen, Kühlregister, Ventil), (Frisch/Außen, Variable Luftvolumeneinheit, Klappe), (Massenstrom, Luft, Frisch/Außen), (Frisch/Außen, Luft, Frisch/Außen, Klappe), (Temperatur, Heizwasser, Heizregister, Austritt), (Temperatur, Heizwasser, Heizregister, Mischluft), (Energie, Heizregister, RLT-Anlage), (Leistung, Luft, Zuluft, Ventilator), (Temperatur, Heizwasser, Heizregister, Austritt), (Temperatur, Register), (Temperatur, Kühlwasser, Zuluft), (Temperatur, Luft, Zone), (Temperatur, Luft, Frisch/Außen), (Temperatur, Luft, Zuluft), (Feuchtigkeit, Luft, Zuluft), (Temperatur, Luft, Rücklauf), (Feuchtigkeit, Luft, Rücklauf), (Energie, Kühlregister), (Temperatur, Kühlwasser, Zuluft), (Temperatur, Luft, Mischluft), (Temperatur, Luft, Frisch/Außen), (Temperatur, Luft, Zuluft),
115	[269, 270, 271]	
116	[269, 270, 271]	
117	[269, 270, 271]	
118	[483, 482]	
119	[483, 482]	
		Fortsetzung auf nächster Seite

Tabelle E.3: Benötigte Metadaten für die einzelnen Fehlererkennungsansätze aus Li u. a. [273].

Index	Quellen	Benötigte Metadaten
120	[483, 482]	(Feuchtigkeit, Luft, Zuluft), (Temperatur, Luft, Rücklauf), (Feuchtigkeit, Luft, Rücklauf), (Energie, Heizregister), (Temperatur, Luft, Mischluft), (Temperatur, Heizwasser, Zuluft), (Temperatur, Luft, Frisch/Außen), (Temperatur, Luft, Zuluft), (Massenstrom, Luft, Zuluft), (Energie, Luft, Ventilator, Zuluft), (Feuchtigkeit, Luft, Zuluft), (Temperatur, Luft, Rücklauf), (Massenstrom, Luft, Rücklauf), (Feuchtigkeit, Luft, Rücklauf), (Energie, Kühlregister), (Temperatur, Kühlwasser, Zuluft), (Temperatur, Luft, Mischluft), (Temperatur, Luft, Frisch/Außen), (Temperatur, Luft, Zuluft), (Massenstrom, Luft, Zuluft), (Energie, Luft, Ventilator, Zuluft), (Feuchtigkeit, Luft, Zuluft), (Temperatur, Luft, Rücklauf), (Massenstrom, Luft, Rücklauf), (Feuchtigkeit, Luft, Rücklauf), (Energie, Kühlregister), (Temperatur, Kühlwasser, Zuluft), (Temperatur, Luft, Mischluft), (Massenstrom, Luft, Frisch/Außen), (Massenstrom, Ventilator, Luft, Frisch/Außen), (Elektrische Leistung, Luft, Ventilator, Frisch/Außen), (Frisch/Außen, Luft, Ventilator, Frisch/Außen), (Regelwert, Luft, Ventilator, Frisch/Außen), (Frisch/Außen, Luft, Ventilator, Filter, Frisch/Außen), (Regelwert, Ventilator, Fortluft),
121	[508]	
123	[238]	
		Fortsetzung auf nächster Seite

Tabelle E.3: Benötigte Metadaten für die einzelnen Fehlererkennungsansätze aus Li u. a. [273].

Index	Quellen	Benötigte Metadaten
124	[484]	(Frisch/Außen, Luft, Ventilator, Fortluft), (Frisch/Außen, Luft, Ventilator, Filter, Fortluft), (Drehzahl, Luft, Ventilator), (Temperatur, Luft, Frisch/Außen), (Temperatur, Luft, Zuluft), (Massenstrom, Luft, Zuluft), (Regelwert, Luft, Ventilator, Rücklauf), (Elektrische Leistung, Luft, Ventilator, Rücklauf), (Regelwert, Kühlregister, Ventil), (Temperatur, Luft, Mischluft), (Temperatur, Luft, Zuluft), (Temperatur, Kühlwasser, Zuluft), (Massenstrom, Kühlwasser), (Regelwert, Kühlwasser, Ventil), (Temperatur, Luft, Zuluft), (Temperatur, Kühlwasser, Zuluft), (Massenstrom, Kühlwasser), (Regelwert, Kühlwasser, Ventil), (Massenstrom, Luft, Frisch/Außen), (Massenstrom, Luft, Frisch/Außen, Sollwert), (Massenstrom, Luft, Zuluft), (Regelwert, Luft, Ventilator, Zuluft), (Massenstrom, Luft, Rücklauf), (Temperatur, Luft, Zuluft), (Feuchtigkeit, Luft, Zuluft), (Temperatur, Luft, Rücklauf), (Energie, Heizregister), (Energie, Kühlregister), (Temperatur, Luft, Zuluft), (Feuchtigkeit, Luft, Zuluft), (Massenstrom, Luft, Mischluft),
125	[115, 116]	
126	[129]	
127	[129]	
128	[482]	
129	[325]	

E.5 Anwendungen

Im Folgenden werden die Parameter für die betrachteten Anwendungen beschrieben.

E.5.1 Raumlufthtechnische Anlage

Benötigte Eigenschaften für die Anwendung „Raumlufthtechnische Anlage“:

- (, Kühlregister),
- (, Luft, Ventilator, Rücklauf),
- (Differenz, Druck, Luft, Register),
- (Differenz, Elektrische Leistung, Filter),
- (Differenz, Elektrische Leistung, Heizwasser, Pumpe),
- (Differenz, Enthalpie, Luft, Variable Luftvolumeneinheit, Mischluft, Zuluft),
- (Differenz, Temperatur, Kühlwasser, Verdampfer),
- (Differenz, Temperatur, Luft, Zone, Istwert, Sollwert),
- (Differenz, Temperatur, Luft, Zuluft, Istwert, Sollwert),
- (Drehzahl, Luft, Ventilator),
- (Drehzahl, Luft, Ventilator, Rücklauf),
- (Drehzahl, Luft, Ventilator, Zuluft),
- (Drehzahl, Ventilator, Fortluft),
- (Drehzahl, Ventilator, Frisch/Außen),
- (Elektrische Leistung, Luft, Ventilator, Fortluft),
- (Elektrische Leistung, Luft, Ventilator, Frisch/Außen),
- (Elektrische Leistung, Luft, Ventilator, Rücklauf),
- (Elektrische Leistung, Luft, Zuluft),
- (Elektrische Leistung, Luft, Zuluft, Sollwert),
- (Energie, Heizregister),
- (Energie, Heizregister, RLT-Anlage),
- (Energie, Kühlregister),
- (Energie, Kühlregister, RLT-Anlage),
- (Energie, Luft, Ventilator, Zuluft),
- (Feuchtigkeit, Luft, Frisch/Außen),
- (Feuchtigkeit, Luft, Rücklauf),
- (Feuchtigkeit, Luft, Zuluft),

- (Frisch/Außen, Heizregister, Klappe),
- (Frisch/Außen, Heizregister, Ventil),
- (Frisch/Außen, Klappe),
- (Frisch/Außen, Klappe, Zuluft),
- (Frisch/Außen, Kühlregister, Ventil),
- (Frisch/Außen, Kühlwasser, Ventil),
- (Frisch/Außen, Luft, Fortluft, Klappe),
- (Frisch/Außen, Luft, Frisch/Außen, Klappe),
- (Frisch/Außen, Luft, Rücklauf, Klappe),
- (Frisch/Außen, Luft, Ventilator, Filter, Fortluft),
- (Frisch/Außen, Luft, Ventilator, Filter, Frisch/Außen),
- (Frisch/Außen, Luft, Ventilator, Fortluft),
- (Frisch/Außen, Luft, Ventilator, Frisch/Außen),
- (Frisch/Außen, Variable Luftvolumeneinheit, Klappe),
- (Frisch/Außen, Ventil),
- (Leistung, Erhitzer),
- (Leistung, Kühlwasser, Pumpe),
- (Leistung, Luft, Ventilator, Rücklauf),
- (Leistung, Luft, Zuluft, Ventilator),
- (Massenstrom, Heizwasser, Kessel),
- (Massenstrom, Heizwasser, Zuluft),
- (Massenstrom, Kühlwasser),
- (Massenstrom, Luft),
- (Massenstrom, Luft, Austritt),
- (Massenstrom, Luft, Frisch/Außen),
- (Massenstrom, Luft, Frisch/Außen, Sollwert),
- (Massenstrom, Luft, Mischluft),
- (Massenstrom, Luft, Raum/Innen),
- (Massenstrom, Luft, Rücklauf),

- (Massenstrom, Luft, Sollwert),
- (Massenstrom, Luft, Sollwert, Minimum),
- (Massenstrom, Luft, Variable Luftvolumeneinheit, Sollwert),
- (Massenstrom, Luft, Variable Luftvolumeneinheit, Sollwert, Maximum),
- (Massenstrom, Luft, Zone),
- (Massenstrom, Luft, Zone, Sollwert),
- (Massenstrom, Luft, Zone, Sollwert, Maximum),
- (Massenstrom, Luft, Zone, Sollwert, Minimum),
- (Massenstrom, Luft, Zuluft),
- (Massenstrom, Ventilator, Fortluft),
- (Massenstrom, Ventilator, Luft, Frisch/Außen),
- (Regelwert, Heizkörper, Ventil),
- (Regelwert, Heizregister, Ventil),
- (Regelwert, Klappe),
- (Regelwert, Kühlregister, Ventil),
- (Regelwert, Kühlwasser, Ventil),
- (Regelwert, Luft, Fortluft, Klappe),
- (Regelwert, Luft, Frisch/Außen, Klappe),
- (Regelwert, Luft, Frisch/Außen, Klappe, Bypass),
- (Regelwert, Luft, Ventilator, Frisch/Außen),
- (Regelwert, Luft, Ventilator, Rücklauf),
- (Regelwert, Luft, Ventilator, Zuluft),
- (Regelwert, RLT-Anlage, Ventil),
- (Regelwert, Variable Luftvolumeneinheit, Ventil),
- (Regelwert, Ventil),
- (Regelwert, Ventilator, Filter, Fortluft),
- (Regelwert, Ventilator, Fortluft),
- (Relative Feuchte, Luft, Fortluft),
- (Relative Feuchte, Luft, Frisch/Außen),

- (Relative Feuchte, Luft, Mischluft),
- (Relative Feuchte, Luft, Rücklauf),
- (Relative Feuchte, Luft, Zuluft),
- (Stellantrieb, Temperatur, Luft, Frisch/Außen),
- (Stellantrieb, Temperatur, Luft, Mischluft),
- (Stellantrieb, Temperatur, Luft, Rücklauf),
- (Stellantrieb, Temperatur, Luft, Zuluft),
- (Temperatur, Heizwasser, Heizregister, Austritt),
- (Temperatur, Heizwasser, Heizregister, Mischluft),
- (Temperatur, Heizwasser, Rücklauf),
- (Temperatur, Heizwasser, Zuluft),
- (Temperatur, Kühlwasser, Rücklauf),
- (Temperatur, Kühlwasser, Speicher),
- (Temperatur, Kühlwasser, Zuluft),
- (Temperatur, Kühlwasser, Zuluft, Sollwert),
- (Temperatur, Luft, Fortluft),
- (Temperatur, Luft, Frisch/Außen),
- (Temperatur, Luft, Mischluft),
- (Temperatur, Luft, Raum/Innen),
- (Temperatur, Luft, Rücklauf),
- (Temperatur, Luft, Ventilator, Zuluft),
- (Temperatur, Luft, Wärmeübertrager, nach Wärmeübertrager),
- (Temperatur, Luft, Wärmeübertrager, vor Wärmeübertrager),
- (Temperatur, Luft, Zone),
- (Temperatur, Luft, Zone, Sollwert),
- (Temperatur, Luft, Zuluft),
- (Temperatur, Luft, Zuluft, Sollwert),
- (Temperatur, Register),
- (Volumenstrom, Luft, Register),
- (Volumenstrom, Luft, Zuluft),

E.5.2 Wärmeübertrager

Benötigte Eigenschaften für die Anwendung „Wärmeübertrager“:

- Wärmeübertrager
- Sensor
- Messwert
- Geregelter Wert
- Temperatur
- Temperaturdifferenz
- Sollwert
- Sollwert Heizen
- Sollwert Kühlen
- Volumenstrom
- vor Wärmeübertrager
- nach Wärmeübertrager
- Wasser
- Luft

Distanz	c_i
Jaro-Winkler-Distanz	0.0
Cosinus-Distanz	0.0
Damerau-Levenshtein-Distanz	0.0
Optimal String Alignment (OSA)	0.0
Longest Common Subsequence (LCS)	0.0
Longest Common Substring (LCSS)	0.0
Prefix	0.0
Beginn	1.0

Tabelle E.4: Anteile c_i der einzelnen Distanzen für die Ähnlichkeitsberechnung von SSC_{total} bei Wärmeübertragern, Raum- und Quartiersmonitoring.

E.5.3 Außentemperatursensor

Benötigte Eigenschaften für die Anwendung „Außentemperatursensor“:

- Sensor
- Messwert

- Temperatur
- Luft
- Frisch/Außen

E.5.4 PID Regelung der Raumtemperatur

Benötigte Eigenschaften für die Anwendung „PID Regelung der Raumtemperatur“:

- PID Regelung
- Sensor
- Messwert
- Geregelter Wert
- Frisch/Außen
- Temperatur
- Raum/Innen
- Sollwert
- Sollwert Heizen
- Sollwert Kühlen
- Sollwertverstellung
- Sollwertverstellung, Comfort
- Sollwertverstellung, Economy
- Sollwertverstellung, Pre-comfort
- Relative Feuchte
- Absolute Feuchte
- Feuchtigkeit
- Volumenstrom
- Anwesenheit
- Fenster

Distanz	c_i
Jaro-Winkler-Distanz	0.4
Cosinus-Distanz	0.4
Damerau-Levenshtein-Distanz	0.2
Optimal String Alignment (OSA)	0.0
Longest Common Subsequence (LCS)	0.0
Longest Common Substring (LCSS)	0.0
Prefix	0.0

Tabelle E.5: Anteile c_i der einzelnen Distanzen für die Ähnlichkeitsberechnung von SSC_{total} bei PID Regelung von Raumtemperaturen.

E.5.5 Modellprädiktive Regelung

Benötigte Eigenschaften für die Anwendung „modellprädiktive Regelung“:

- Sensor
- Messwert
- Geregelter Wert
- Energie
- Elektrische Energie
- Wärmemenge
- Kältemenge
- Heizleistung
- Kühlleistung
- Heiz-/Kühlleistung
- Frisch/Außen
- Temperatur
- Raum/Innen
- Sollwert
- Sollwert Heizen
- Sollwert Kühlen
- Sollwertverstellung
- Sollwertverstellung, Comfort
- Sollwertverstellung, Economy

- Sollwertverstellung, Pre-comfort
- Relative Feuchte
- Absolute Feuchte
- Feuchtigkeit
- Kohlenstoffdioxid (CO₂)
- Totale flüchtige organische Verbindungen

E.5.6 Raummonitoring

Benötigte Eigenschaften für die Anwendung „Raummonitoring“:

- PID Regelung
- Sensor
- Messwert
- Geregelter Wert
- Frisch/Außen
- Temperatur
- Raum/Innen
- Sollwert
- Sollwert Heizen
- Sollwert Kühlen
- Sollwertverstellung
- Sollwertverstellung, Comfort
- Sollwertverstellung, Economy
- Sollwertverstellung, Pre-comfort
- Relative Feuchte
- Absolute Feuchte
- Feuchtigkeit
- Kohlenstoffdioxid (CO₂)
- Totale flüchtige organische Verbindungen
- Volumenstrom
- Anwesenheit
- Fenster

E.5.7 Energiefluss/Topologie

Benötigte Eigenschaften für die Anwendung „Energiefluss/Topologie“:

- Sensor
- Messwert
- Temperatur

E.5.8 Quartiersmonitoring

Benötigte Eigenschaften für die Anwendung „Quartiersmonitoring“:

- Sensor
- Messwert
- Temperatur
- Wirkarbeit
- Energie
- Elektrische Energie
- Elektrische Leistung
- Heizleistung
- Wärmemenge
- Leistung
- Volumenstrom

E.6 Metadaten-Kennzeichnung

Die manuelle Beschriftung von Gebäudedaten ist, wie bereits erwähnt, sehr zeitaufwendig. Um festzustellen, ob der Zeitaufwand bzw. die Ergebnisse bereits a priori nur auf Basis der vorhandenen Daten abgeschätzt werden kann, werden verschiedene Kennzahlen aus dem Bereich des NLP und der Informationstheorie verwendet. Hierbei wird der Umstand untersucht, der im Spracherwerb von Menschen als Kind oder als Zweitsprache die Fähigkeit eines Sprechers innerhalb dieser Sprache einschätzt. In dieser Arbeit wird die Bewertung der Komplexität einer Sprache im Zusammenhang mit der Fähigkeit von Algorithmen untersucht, ob diese in der Lage sind die Sprache zu übersetzen. Es wird untersucht, ob es eine statistische Korrelation zwischen der Komplexität der Sprache und der Ergebnisse der Algorithmen vorhanden ist. Dies entspricht der Spieltheorie nach

John Nash [330] nach der Spieler oder Sprecher sparsam mit der Anzahl der gesprochenen Wörtern umgehen [303, 511] und die Anzahl der verwendeten Wörter mit der Fähigkeit eine Sprache zu beherrschen korreliert[249].

Da natürliche Sprache anders aufgebaut ist als die künstliche Sprache der Gebäudeautomation, müssen vor der Vorstellung der einzelnen Kennzahlen die Begriffe aus der natürlichen Sprache in die Sprache der Gebäudeautomation überführt werden. Wenn ein Kennzeichen eines Datenstroms „0431-A-Ahu01-TOa“ beträgt, dann sind die Blöcke „0431“, „A“, „Ahu“, „01“, „T“ und „Oa“ einzelne Wörter. Jedes Zeichen, das kein Trennungszeichen (z.B. „-“) ist, wird als Zeichen gezählt. Jeder Datenstromname wird als Satz interpretiert. Komplexe Wörter sind Wörter die höchstens n mal in 1000 Datenstromnamen vorkommen.

E.6.1 Gunning Fog Index

Der Gunning Fog Index (GFI) [165, 54] wird zur Bewertung von Texten und ihrer Komplexität verwendet. Ursprünglich wurden dieser und andere Indizes für die Lesbarkeit der englischen Sprache entwickelt. Der GFI (siehe Formel E.1) setzt die Anzahl der Wörter und die Anzahl der Sätze in Beziehung. Darüber hinaus wird die Anzahl der komplexen Wörter (mindestens drei Silben) mit der Anzahl der Wörter in Beziehung gesetzt.

$$GFI = 0.4 \cdot \left[\left(\frac{\text{Wörter}}{\text{Sätze}} \right) + 100 \cdot \frac{\text{Komplexe Wörter}}{\text{Wörter}} \right] \quad (\text{E.1})$$

E.6.2 Automated readability index

Der automatische Lesbarkeitsindex (Automated readability index: ARI)[422] setzt die Anzahl der Zeichen ins Verhältnis zur Anzahl der Wörter und die Anzahl der komplexen Wörter zur Anzahl der Sätze[37] (siehe Formel E.2).

$$ARI = 4.71 \cdot \left[\left(\frac{\text{Zeichen}}{\text{Wörter}} \right) + 0.5 \cdot \frac{\text{Komplexe Wörter}}{\text{Sätze}} - 21.43 \right] \quad (\text{E.2})$$

In unserem Beispiel sind die Datenstrom-Labels jeweils ein Satz, die Chunks sind jeweils ein Wort, und alle Zeichen innerhalb des Labels werden als Zeichen gezählt.

E.6.3 Semantische Komplexität

Mithilfe von der Messung der lexikalischen Reichhaltigkeit kann das Wortschatzwissen von einem Sprecher bewertet und die Komplexität der verwendeten Sprache beurteilt werden[246]. Lu [295] stellten ein System vor zur Berechnung der lexikalischen Reichhaltigkeit. Laut Spring u. Johnson

[426] können auch NLP Werkzeuge wie SpaCy oder NLTK die lexikalische Reichhaltigkeit berechnen. In dieser Arbeit wird die Idee von Spring u. Johnson [426] übernommen, dass meist gebräuchliche Wörter nicht von der Berechnung ausgeschlossen werden sollten. Das System von Lu [295] wird angepasst in dieser Arbeit verwendet. Kyle [257] liefert hierzu einen Überblick über gebräuchliche Formeln für die lexikalische Reichhaltigkeit.

Im Folgenden werden verschiedene Bewertungskennzahlen für die Komplexität von Texten vorgestellt.

E.6.4 Semantische Bewertung

Lee u. a. [260] entwickelten ausgewählte Merkmale, die selbst SdT Algorithmen bei der Kategorisierung von Texten helfen. Um die Komplexität von Texten zu bewerten, entwickelten sie drei Kennzahlen (Semantische Vielfalt, Semantische Deutlichkeit, Semantisches Rauschen) [260, 261]:

- Semantische Vielfalt: $\sum_{i=1}^n p_i \cdot i$
- Semantische Deutlichkeit: $\frac{1}{n} \cdot \sum_{i=1}^n \max(p) - p_i$
- Semantisches Rauschen: $n \cdot \frac{\sum_{i=1}^n (p_i - \bar{p})^4}{(\sum_{i=1}^n (p_i - \bar{p})^2)^2}$

E.6.4.1 Lexikalische Dichte

Nach Kyle [257] wird die lexikalische Dichte anhand der Formel E.3 berechnet. Hierbei werden die Anzahl der Autosemantika (Worte mit Bedeutung) ins Verhältnis zu der Gesamtzahl der Wörter.

$$D_{lex} = \frac{\text{Anzahl Autosemantika}}{\text{Gesamtanzahl Wörter}} \quad (\text{E.3})$$

E.6.4.2 Type-Token-Ratio

Beim Type-Token-Ratio (TTR) oder Diversifikationsquotient [425] (siehe Formel E.4) werden die Anzahl der einzigartigen Wörter in einem Sprachbeispiel in das Verhältnis zur Gesamtzahl der Wörter gesetzt [64, 234]. Der TTR wird in vielen Fällen eingesetzt, um die Reichhaltigkeit einer Sprache zu bestimmen [257]. In der Gebäudeautomation besitzt jeder Betreiber oder Gebäudeautomatisierer eine eigene Sprache, um Datenströme zu benennen. Somit kann die Reichhaltigkeit der jeweiligen „Gebäudeautomationssprache“ anhand des TTR bestimmt werden.

$$TTR = \frac{\text{Anzahl einzigartiger Wörter}}{\text{Gesamtanzahl Wörter}} \quad (\text{E.4})$$

E.6.5 Informationsentropie

In der Informationstheorie wird der Informationsgehalt einer Zeichenkette durch die Informationsentropie ausgedrückt. Für die Informationsentropie existieren zusätzlich zum klassischen Ansatz von Shannon mehrere Ansätze als Diversitätsindizes, die vor allem ihre Anwendung in der Biodiversität finden[403], jedoch auch in der Datenanalyse abseits der Biologie verwendet werden. Deshalb werden diese zusätzlich vorgestellt.

E.6.5.1 Shannon-Index

Der gebräuchlichste ist der Ansatz der Informationsentropie von Claude E. Shannon von 1948[412] und gilt nach wie vor als Standard für den Informationsgehalt von Zeichenketten.

Er ordnet jedem Zeichen eine Wahrscheinlichkeit für sein Auftreten zu. In einem gleichmäßig verteilten Buchstabensystem sind die Zeichen unabhängig von den anderen. In der Realität ist dies nicht der Fall. In der deutschen und englischen Sprache kommt beispielsweise der Buchstabe E häufiger vor als die übrigen Buchstaben.

Shannon definiert den Informationsgehalt einer Zeichenkette danach, wie unwahrscheinlich es ist, dass ein Zeichen in der Zeichenkette vorkommt. Wenn ein Zeichen selten vorkommt, dann ist ein hoher Informationsgehalt erreicht.

Er berechnet die Informationsentropie H (benannt nach dem Boltzmann H-Theorem[55]) mit der folgenden Gleichung:

$$H(X) = - \sum_{i=0}^n p(x_i) \cdot \log_b(p(x_i)) \quad (\text{E.5})$$

Dabei entspricht $p(x_i)$ der Wahrscheinlichkeit, dass ein bestimmtes Zeichen oder eine bestimmte Zeichenkette auftritt. n ist die Anzahl der Zeichen oder Zeichenketten. b entspricht der Basis des Logarithmus. Shannon definierte b als 2 und ging dabei von einem bitbasierten System aus. Die Definition von b ist entscheidend für die Skalierung der Entropieberechnung.

E.6.5.2 Gini-Simpson-Index

Als alternativer Informationsentropie-Index wird auch der Gini-Simpson-Index verwendet[1, 132]. Er setzt die Individuen einer Klasse in Zusammenhang mit der Gesamtheit an verfügbaren Individuen aller Klassen[219]. Er ist in Formel E.6 bzw. vereinfacht in E.7 definiert.

$$D = \sum_{i=1}^n \frac{n_i(n_i - 1)}{n(n - 1)} \quad (\text{E.6})$$

$$D = \sum_{i=1}^n p_i^2 \quad (\text{E.7})$$

E.6.5.3 Kullback-Leibler-Divergenz

Die Kullback-Leibler-Divergenz (KL)[255] oder relative Entropie gibt an, wie unterschiedlich zwei Wahrscheinlichkeitsverteilungen sind[60].

Bei zwei Wahrscheinlichkeitsverteilungen U und V mit den Wahrscheinlichkeiten u und v , wird die Kullback-Leibler-Divergenz KL mit folgender Formel berechnet:

$$KL = \sum_{i=1}^n p(x) \cdot \left(\log\left(\frac{p(x)}{q(x)}\right)\right) \quad (\text{E.8})$$

Die Kullback-Leibler-Divergenz kann verwendet werden, um insbesondere die Verteilung der Übersetzungen in den verschiedenen Datensätzen zu überprüfen.

E.6.5.4 Zipfsches Gesetz

Nach dem Zipfschen Gesetz [513] verteilen sich die in einer Sprache verwendeten Wörter in ihrer Anzahl nach ihrem Rang [212]. Dieses Gesetz trifft nicht nur auf Sprachen zu, sondern auch auf Zitationen von wissenschaftlichen Papern [252], der Intensität von Kriegen[421, 389, 76, 334], der Zugriffe auf Webseiten[98, 8] oder der Population von den größten Städten eines Landes [514, 334]. Das Zipfsche Gesetz sagt aus, dass das n . häufigste Wort $1/n$ mal so häufig verwendet wird wie das erste Wort. Also bei $n = 3$ wird das dritthäufigste Wort $1/3$ mal so häufig verwendet wie das erste Wort. Somit ergibt sich ein Zusammenhang für die Wahrscheinlichkeit $p(n)$ des Auftretens eines Wortes im Rang n nach der Formel $p(n) \sim \frac{1}{n}$. Bei höheren n ist das Gesetz nur noch eingeschränkt anwendbar[334, 212]. Deshalb wird n maximal bis zu $n = 200$ untersucht. In dieser Arbeit wird das Zipfsche Gesetz bei allen Datensätzen überprüft, um zu ermitteln, inwiefern eine Verzerrung der Datensätze vorliegt und inwiefern diese Verzerrung das Ergebnis des Algorithmus beeinflusst.

E.7 Zusätzliche Ergebnisse

Im Folgenden werden zusätzliche Ergebnisse aufgeführt, die die Ergebnisse der Dissertation in einem größeren Maßstab darstellen sollen:

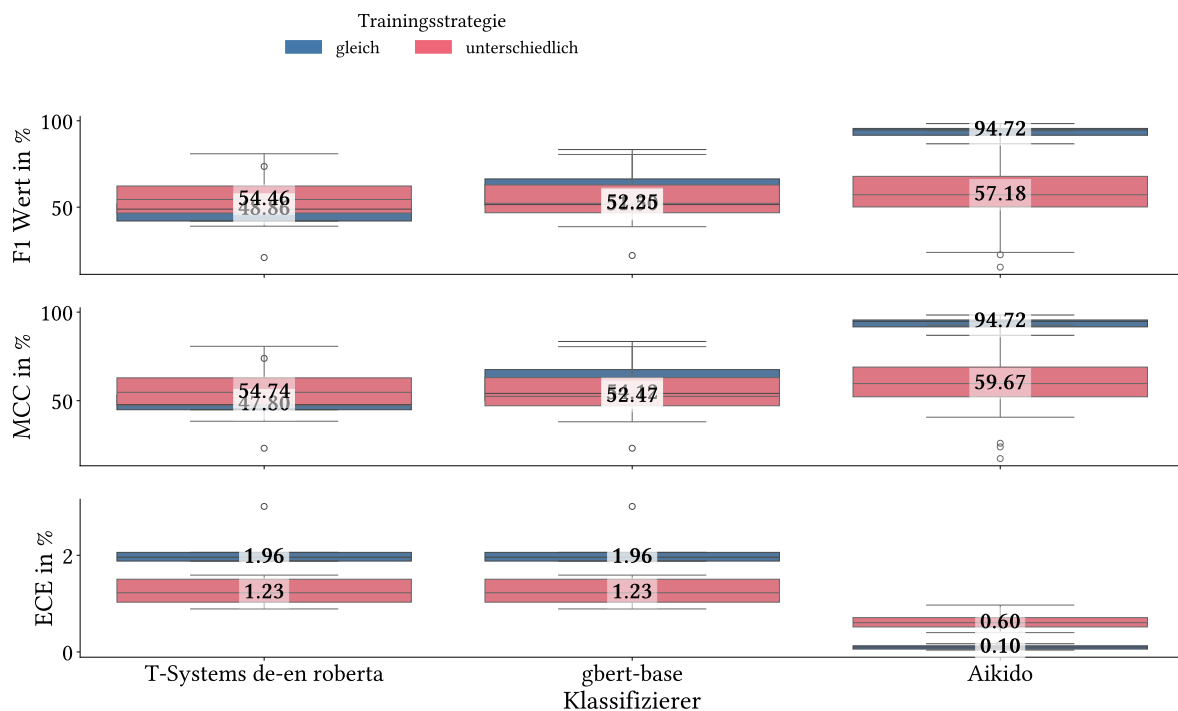
E.7.1 Vergleich Aikido mit RoBERTa

Im Folgenden wird ein Vergleich von Aikido 2.0, dem in dieser Arbeit entwickelten Algorithmus, zur Klassifizierung von Zeichenketten mit RoBERTa und gbert durchgeführt, die in der Literatur als Algorithmen für GAS-Daten verwendet wurden. Hierzu wurde Huggingface als Schnittstelle für NLP-Algorithmen verwendet. Als RoBERTa-Algorithmus wurde die Variante von T-Systems verwendet

(„T-Systems-onsite/cross-en-de-roberta-sentence-transformer“), die auf der deutschen und englischen Sprache trainiert wurde, was am nächsten an den im OOM4ABDO Datensatz vorliegenden Datensatz liegt. Für gbert wurde der allgemeine Fall „gbert-base“ verwendet. In Abbildung E.4 sind die Ergebnisse im Vergleich zu dem in dieser Arbeit entwickelten Aikido Algorithmus dargestellt. Es ist sichtbar, dass in jedem Fall der Aikido-Algorithmus die Ergebnisse des RoBERTa- und gbert-Algorithmus übersteigt. Die Abhängigkeit von der Anzahl der Trainingsdaten ist in den besseren Ergebnissen von unterschiedlichen Trainings- und Testdaten ersichtlich. Dies ist darauf zurückzuführen, dass nur 10 % des gleichen Datensatzes bei gleicher Trainingsstrategie verwendet werden, aber 100 % des anderen Datensatzes bei unterschiedlicher Trainingsstrategie.

Es ist fraglich, welche weiteren Schritte durchgeführt wurden, so dass die Algorithmen diese Ergebnisse auf GAS-Daten erzielt haben.

Abbildung E.4: Vergleich des allgemeinen F1 Wertes, MCC und ECE bei BERT-Varianten und dem in dieser Arbeit entwickelten Aikido Algorithmus bei Training und Testen auf gleichen und unterschiedlichen Datensätzen.



F Verwendete Umgebung

Für die Berechnung wurde eine virtuelle Maschine mit 50 virtuellen CPU-Kernen und 128 GB RAM verwendet.

F.1 Verwendete Umgebung Zeitreihenklassifizierung

Die folgenden Pakete wurden für die Zeitreihenklassifizierung verwendet:

Tabelle F.1: Genutzte Pakete für die Zeitreihenklassifizierung.

Paket	Version
bottleneck	1.3.7
catboost	1.2
cython	0.29.34
EMD-signal	1.5.1
EntropyHub	0.2
fastdtw	0.3.4
keras	2.12.0
lightgbm	3.3.5
mlflow	2.3.2
neurokit2	0.2.4
numba	0.57.0
numpy	1.23.5
opencv-python	4.7.0
opt-einsum	3.3.0
prophet	1.1.3
rapidfuzz	3.1.1
ray	2.4.0
ruptures	1.1.8
scikit-learn	1.2.2
scipy	1.10.1
stumpy	1.11.1
tensorflow	2.12.0
torch	2.2.2
torchmetrics	1.3.2
tsfel	0.1.5

Fortsetzung auf nächster Seite
