Universitat de Girona
Escola Politècnica Superior

Master's Degree Thesis

Degree: Master in Data Science

Title: The Past and Present of Predictive Models for Anomaly Detection in Smart Cities:
A Systematic Review

Document: Thesis

Student: Andrea Carolina Ramirez Moya

Tutor: Mateu Villaret Auselle
Department: Computer Science, Applied Mathematics and Statistics
Area: Languages and Computer Systems

Co-Tutor: Marc Comas Cufi
Department: Computer Science, Applied Mathematics and Statistics
Area: Statistics and Operations Research

Call: September 2023

Universitat de Girona
Escola Politècnica Superior

MASTER'S DEGREE THESIS

The Past and Present of Predictive Models for Anomaly Detection in Smart Cities: A Systematic Review

Author:
Andrea Carolina RAMIREZ MOYA

September 2023

Master in Data Science

Tutors:
Mateu VILLARET AUSELLE
Marc COMAS CUFI

Abstract

Detecting anomalies in smart cities is a novel area that started being studied in the 21st century. This master's thesis aims to find the most accurate predictive models that can be explainable to scholars and industry stakeholders. With that goal in mind, a PRISMA 2020 for systematic literature reviews methodology is approached to review the papers that have been published in Emerald Insights, IEEE Xplore, Science Direct, and Web of Science with the concepts of Smart Cities, Data Science, and Predictive Models between 2000 and the first half of 2023. The findings show that the algorithms that have been studied the most are for classification, supervised machine learning. This thesis not only took into account the theoretical part, but also attempted addressing those techniques by forecasting the energy consumption in buildings in Barcelona, classifying if those outcomes were an anomaly, and finally clustering to find the consumption patterns. The deliverables are disclosed in a ObservableHQ notebook and a dashboard in Google Data Studio.

Keywords: Systematic Review, Predictive Models, XAI, Anomaly Detection, Smart Cities, Energy Consumption in buildings.

La detecció d'anomalies en ciutats intel·ligents és una àrea nova que va començar a ser estudiada al segle XXI. Aquest treball de màster té com a objectiu trobar els models predictius més precisos que puguin ser explicables als acadèmics i a les parts interessades de la indústria. Amb aquest fi en ment, s'utilitza el PRISMA 2020 com metodologia per la revisió sistemàtica de la literatura dels treballs que han estat publicats en Emerald Insights, IEEE Xplore, Science Direct, i Web of Science amb els conceptes de Ciutats Intel·ligents, Ciència de Dades i Models Predictius entre 2000 i la primera meitat de 2023. Les troballes mostren que els algoritmes que més s'han estudiat són per a la classificació, l'aprenentatge automàtic supervisat. Aquesta tesi no només va tenir en compte la part teòrica, sinó que també va abordar aquestes tècniques mitjançant la predicció del consum d'energia en edificis de Barcelona, classificant si aquests resultats eren una anomalia, i finalment agrupant-se per trobar els patrons de consum. Els lliuraments es revelen en un quadern a ObservableHQ i un tauler de comandaments a Google Data Studio.

Paraules clau: Revisió Sistemàtica, Models Predictius, Detecció dÁnomalies, Ciutats Intel·ligents, Consum d'Energia en edificis, XAI.

Acknowledgement

I want to express my gratitude to my supervisors Mateu Villaret Auselle and Marc Comas Cufi for their support throughout this challenging period. Their commitment to my academic development has been encouraging in the ups and downs of this journey. I value your faith in me and allowing me to gain knowledge from your guidance as I learn and grow. I genuinely thank them for their influence on my work and life.

I also want to acknowledge my professors, who should receive my sincere appreciation. Each lesson has improved my academic knowledge while shaping how I see the world. Their lectures have guided me beyond academia, so I cherish their contribution and assistance as I pursue proficiency and achievement.

I extend a heartfelt acknowledgment to my family and friends, who have been there as the base of my never-ending pursuit of knowledge. Their constant support and willingness to openly share their insights have been priceless treasures. Every conversation, brainstorming session, and word of advice has enriched my outlook and become a never-ending source of inspiration. Their presence on my journey reminds me of the significance of learning and growing together. I am grateful for their time and effort in building my awareness of the world.

Last but not least, I would like to thank my colleagues and supervisors at Nexus Geographics. Without daring to excel in smart cities in Spain, I would not have gained the inspiration to develop this thesis. Daily learning what is going on in the industry inspired me to research this topic and helped me identify gaps in this field.

Contents

List of Figures

List of Tables

CHAPTER 1
Introduction

Cities are facing a digital transformation. Part of it comes from using technology to its advantage to ensure its citizens' safety and quality of life. Governments are seeking tools to predict and alert anomalies to make informed and optimal decisions in cases of emergency to take action and address those problems. However, one limitation is that those forecasting measures must be explainable to ensure transparency and clarity in communicating with stakeholders.

This comprehensive literature review on twenty-first-century studies (2000-2023) will seek papers searched on Emerald Insights, IEEE Xplore, Science Direct, and Web of Science, in the category of Smart Cities, Data Science in the public sector, and Predictive Models. It will follow the checklist provided by the "PRISMA 2020 explanation and elaboration: updated guidance and exemplars for reporting systematic reviews" [Page 2021].

No matter the size, cities need understandable models to predict their data and generate alerts on the grounds that it allows them to make informed and optimal decisions. Explainable models give them a clear and detailed account of how predictions are developing, allowing them to rely on alerts and take appropriate action to address problems detected in the public sector. Additionally, these models are easier to interpret and communicate to other stakeholders, such as citizens and department members, which can foster transparency and trust in local government.

While continuing to be fragmented and interdisciplinary, knowledge production is speeding up dramatically in the Machine Learning world. As a result, it is challenging to assess the amount of information in this field and remain on the leading edge of knowledge when selecting forecasting models. This research aims to identify a series of predictive models that are optimal in detecting and, consequently, issuing alerts in the event of finding any anomaly in the data. In this way, it will contribute significantly to improving the efficiency and effectiveness of public sector data management/governance, translating into better service and well-being for their citizens.

This research will create a solid foundation for current and future research and provide a broader and more rigorous perspective regarding possible solutions and approaches to addressing this problem. To ensure the review is accurate, a systematic review is a methodical approach for this kind of area that has been "conceptualized differently and studied by various groups of researchers within diverse disciplines” [Wong 2013]. It will help identify, analyze, and report patterns.

All the decisions for the search strategy will be written down to enable transparency for the criteria considered when selecting and classifying the papers, the limitations encountered, and the content analysis. For this research to be thorough and robust, it will follow the four-phase guidelines “to assess the quality of a literature review” [Snyder 2019] that include design, conduction, analysis, and writing the review.

Following those outcomes, with the CRISP-DM methodology, there will be an experimentation section with the electricity consumption in Barcelona to model what the authors have stated to forecast them and alert if the values are above or below the historic registers. The data will be accessed from their open data website [Hall 023b]. There’s historical data from January 2019 until March 2023. Suppose there is no access to the data types the papers are evaluating. In that case, there will be a replication section to create a dataset that complies with the aspects needed for modeling and forecasting. Once the model is trained, there will be a communication section to visualize the results found.

In Chapter 2, the preliminaries will explain the concepts to envision this study. In Chapter 3, the planning and methodology will map the road to review the publications related to anomaly detection in smart cities. In Chapter 4, the state of the art will portray the past and present of this field within the literature found. In Chapter 5, the experimentation will perform the anomaly detection models that have been relevant. Chapter 6 will have the results from the evaluation of those approaches, and Chapter 7 will conclude this project with some discussion and future work.

CHAPTER 2
Preliminaries

This chapter presents the foundational knowledge required to comprehend this project.

Anomaly detection. An aspect of intrusion detection involves spotting variations from standard activity that point to intentional or accidental incidents, weaknesses, imperfections, and other problems [Omar 2013].

Big Data. Defined by its 5-v attributes: Volume, Velocity, Veracity, Variety, and Value. It refers to the immense, diverse, and complex environment for data structures (unstructured/semi-structured) that are challenging to index, sort, store, search, analyze, and visualize for use in future processes [Naeem 2022].

Databases of bibliographical references. Platforms to search bibliographic information, such as words, that describe the publication, for instance: the title, authors, abstract, and keywords [University 023].

Data management/governance. The practice of authority and control over the handling of data. It seeks to maximize the value of data while reducing the cost and risk associated with its use [Abraham 2019].

Data visualization. A tool to explain discoveries to others. It helps comprehend, identify patterns, and analyze insights. Its purpose is to provide the information needed in a way that will help them perceive and think clearly with the least biases possible [Andrienko 2020].

Forecasting measures. Projections based on data with historical time stamps. It entails creating models to draw conclusions and guide strategic decision-making in the future. It is not always an accurate prediction, and forecast probabilities might vary considerably—especially when dealing with the variables that frequently vary in time series data and with external factors [Tableau 023].

K-Means. Unsupervised machine learning algorithm that divides points in some dimensions into an established number of clusters [Hartigan 1979].

K-Nearest Neighbor. Supervised machine learning algorithm that is the most straightforward categorization methods that has been employed since the 90s [Laaksonen 1996].

Machine learning. It is a method through which computers run algorithms to carry out tasks while learning from data in a supervised or unsupervised way [Wazid 2022].

Open Data websites. A platform where data is publicly accessible to anyone. It can be utilized for personal or professional reasons. It gathers transport, weather, finance, health, education, policies, and more data [Talukder 2019].

Predictive models. A data-driven model that, given a specific input, makes a probabilistic forecast about the presence of a particular current result in the short or long term [de Hond 2022].

Smart cities. It outlines a local entity, such as an entire city, region, or small area, that employs information technology in an integrated way with real-time analysis to promote sustainable economic growth [Kulkarni 2016].

Systematic review. An assessment that compiles and summarizes information from research that answers a specific topic using specified, methodical techniques [Page 2021].

Table 2.1: List of used acronyms and abbreviations.

Acronyms and abbreviations	**Definition**
BSTS	Bayesian Structural Time Series
CatBoost	Categorical Boosting
COSMO	Consensus self-organized models approach
CRISP-DM	Cross Industry Standard Process for Data Mining
DBSCAN	Density-Based Spatial Clustering of Applications with Noise
DTW	Dynamic Time Warping
GBoost	Gradient Boosted trees
IoT	Internet of Things
K-D tree	K-Dimensional tree
KNN	K-Nearest Neighbor
ML	Machine Learning
NB	Naive Bayes

Acronyms and abbreviations	Definition
LightGBM	Light Gradient Boosting
OBADA	Occupancy Based Anomaly Detection Algorithm
PARX	Poisson Autoregressions with Exogenous Covariates
PRISMA	Preferred Reporting Items for Systematic Reviews and Meta-Analyses
RF	Random Forest
RS	Recommendation System
RUSBoost	Random Under Sampling Boosted trees
SAX	Symbolic Aggregate Approximation
SHAP	Shapley additive explanations
SVM	Support Vector Machine
XAI	Explainable Artificial Intelligence
XGBoost	Extreme Gradient Boosting

CHAPTER 3

Planning and Design for the Methodology

The methodology used in this master's thesis is divided into two distinct approaches that operate independently—one for the systematic literature review, and another for the experimentation with those results.

Planning plays an essential role in this project to provide a clear direction, lower limitations, and ensure timely completion. Defining specific goals and objectives and laying out an approach for achieving them lays a solid basis for this project's success. Having estimated deadlines (table 3.1) ensures a comprehensive layout for all the sections that must be fulfilled for a thorough project.

3.1 Methodical approach for the literature review

A literature review's main objective is to inspect and critically assess prior studies and scholarly writings on a particular subject, in this case, predictive models for anomaly detection in smart cities. This section wants to find the answers to:

- Which area of a city has been studied the most, and which areas are in need of development?
- Which predictive models are being used on anomaly detection in smart cities? Are those models using supervised or unsupervised machine learning techniques?
- Which databases of bibliographical references has the most resources?

As PRISMA [Page 2021] states on its checklist, the first thing to do is to describe the review's criteria for inclusion and exclusion, along with how the studies are categorized for analysis. For that reason, Table 3.2 was created with primary and secondary concepts relevant to this research for filtering purposes, "where in one column we write down each of the terms associated with each concept in all of its variants" [Campos-Asensio 2018]. It will allow for delimiting and classifying the studies more precisely, as well as saving time and effort by not

Description	Estimated time	Actual Effort
Project definition	3 weeks	6 weeks
Research question	-	17/04/23 - 23/05/23
Supervisor's feedback	-	01/05/23 - 08/05/23
Committee acceptance	19/05/23	24/05/23
Literature review	**7 weeks**	**11 weeks**
Planning and design methodology	25/05/23 - 01/06/23	25/05/23 - 08/06/23
Search and extraction on bibliographical databases	01/06/23	05/06/23
Selection of papers to review	01/06/23 - 08/06/23	01/06/23 - 18/06/23
Content analysis	08/06/23 - 21/06/23	18/06/23 - 08/07/23
Results selection for the experimentation	21/06/23 - 01/07/23	08/07/23 - 10/08/23
Supervisor's feedback	08/07/23	20/07/23
Experimentation	**5 weeks**	**7 weeks**
Methodology design	15/07/23 - 21/07/23	15/07/23 - 21/07/23
Field knowledge	21/07/23 - 28/07/23	21/07/23 - 04/08/23
Dataset knowledge	28/07/23 - 04/08/2023	04/08/23 - 20/08/23
Dataset preparation for modeling	04/08/2023 - 18/08/23	11/08/23 - 25/08/23
Dataset training	04/08/23 - 18/08/23	18/08/23 - 31/08/23
Model evaluation	04/08/23 - 18/08/23	18/08/23 - 31/08/23
Data visualization	04/08/23 - 18/08/23	25/08/23 - 31/08/23
Wrapping-up	**1 week**	**1 week**
Conclusions	21/08/23 - 28/08/23	01/09/23 - 05/09/23
Future research	21/08/23 - 28/08/23	01/09/23 - 05/09/23
Abstract	21/08/23 - 28/08/23	01/09/23 - 05/09/23
Supervisor's feedback	01/09/23	06/09/23
Draft corrections and submission	05/09/23	05/09/23
Dissertation	**3 weeks**	**2 weeks**
Designing the slides	01/09/23 - 08/09/23	05/09/23 - 15/09/23
Presentation	18/09/23	18/09/23

Table 3.1: Planning to schedule the performance for this project.

performing a comprehensive review of material that does not fulfill the criteria for this thesis.

The concepts for smart cities were taken from the six-axes that are presented in the paper "*Smart Cities in Europe*", [Allam 2019], and [Singh 2022], where they also state that a city is smart when "investments in human and social capital and traditional and modern (ICT- Information and Communication Technologies) communication infrastructure fuel sustainable economic growth and a high quality of life, with a wise management of natural resources, through participatory governance" [Caragliu 2011].

In order to enhance decision-making and deliver better services to people, city data gathered from many sources, including sensors, internet-connected devices, and other external sources, is extracted for relevant insights and hidden connections [Sarker 2022]. For that reason, and also adding the concepts learned throughout the master is essential for the part of the system environment that houses the large amount of data that a city captures every single day.

Finally, for the third key concept of this project, which is predictive models for anomaly detection, the review will need to center on classifying or clustering data using explainable approaches that enhance transparency when interpreting and communicating them to other stakeholders.

Smart Cities	**Data Science**	**Anomaly detection**
Economy	Sensors	Supervised Machine Learning
Mobility	Internet of Things	Unsupervised Machine Learning
Environment	Data Models	Time Series
People	Data Architecture	
Living	Data Engineering	
Governance		

Table 3.2: Concepts to review.

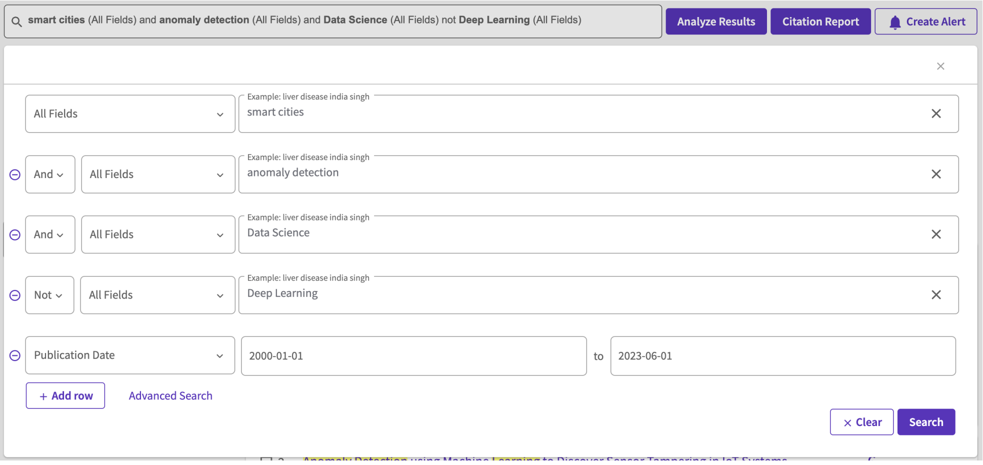
The papers will be retrieved depending on their published date, which fits the limit between January 1st, 2000, until June 1st, 2023 (this century), and the concept to review from the primary sources that are at the bibliographical references databases: *Emerald Insights*, *IEEE Xplore*, *Science Direct* and *Web of Science*. One reason to select these databases is that they have the largest open-access collection, making scientific knowledge and research available to everyone, free from limitations like cost or membership, making reachable information for researchers and the general public.

These databases were selected based on their expertise in a broad range of academic and scientific fields. It was important to extract papers from *IEEE Xplore* for this project since it focuses on engineering and computer science. However, finding only two papers qualified for extraction was shocking. In virtue of that limitation, the conference papers were also selected for review.

Search	Search strategy
#1	Smart Cities AND Data Science AND Predictive Models
#2	Data Science AND Anomaly detection
#3	Smart Cities AND Anomaly detection
#4	Smart Cities AND Data Science AND Anomaly detection
#5	#4 AND (Economy OR Mobility OR Environment OR People OR Living OR Governance OR Sensors OR Internet of Things OR Smart Data Models OR Data Acquisition OR Data Architecture OR Data Engineering OR Data visualization OR Supervised Machine Learning OR Unsupervised Machine Learning)
#6	#5 AND NOT Blockchain [Title/Keywords] AND NOT Deep Learning [Title/Keywords] AND NOT Neural Networks AND NOT Federated Learning AND NOT Reinforcement Learning AND NOT Cloud/Fog Computing [Title/Keywords]

Table 3.3: Search strategy.

With the Boolean operator AND, papers can be identified that include the primary concepts from the search strategy [Campos-Asensio 2018] [Borrego 2014] (table 3.3 and an example of retrieval on figure 3.1). The strategy #4: Smart cities AND Data Science AND Anomaly detection was used because by filtering the other way around, 1,000 papers would be needed to be reviewed. On that iteration of searching on each bibliographical database, 377 papers were downloaded (table 3.4). However, several studies are not relevant to this review since they are related to Blockchain, Cloud, Edge, and Fog Computing, Deep Learning, Reinforcement Learning, Adversarial Learning, and Federated Learning, which are out of the scope of explainable predictive modeling. Therefore, this is the moment to benefit from selecting secondary concepts and filtering those papers that contain at least one relevant keyword or term in their title. That led to 277 papers to review (strategy #5 and strategy #6).


smart cities (All Fields) and anomaly detection (All Fields) and Data Science (All Fields) not Deep Learning (All Fields)

All Fields — Example: liver disease india singh — smart cities

And — All Fields — Example: liver disease india singh — anomaly detection

And — All Fields — Example: liver disease india singh — Data Science

Not — All Fields — Example: liver disease india singh — Deep Learning

Publication Date — 2000-01-01 to 2023-06-01

Figure 3.1: Advanced settings to extract the papers on the *Web of Science* using part of the strategy #5.

"At the end of our search, we need to assess the results obtained, both in exhaustiveness (provision of relevant documents which a strategy has been capable of finding), and precision (number of relevant records retrieved compared with the total number of retrieved records) and relevance (which will be useful) to respond to our question" [Campos-Asensio 2018]. Some papers could be listed in different databases, so it is critical to identify them to measure the studies' precision and relevance. For that reason, those papers were screened by keywords and abstract to ensure the studies were within this project's scope. That left us with 45 papers to review comprehensively.

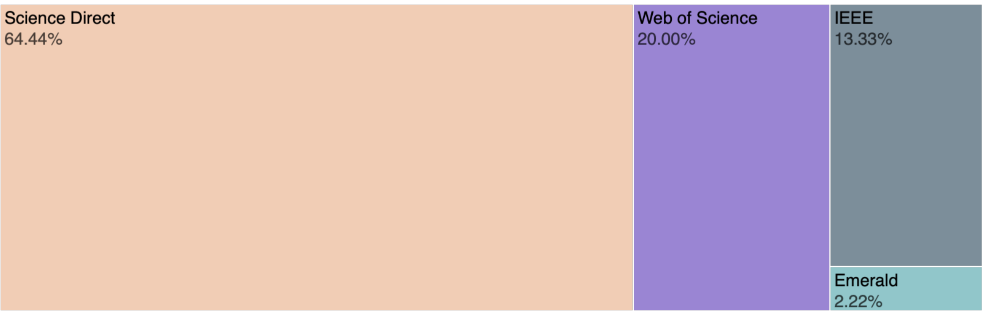


Figure 3.2: Share of reference databases reviewed.

As seen in Table 3.4 and Figure 3.2, the reference database that had the most papers to review was Science Direct (64%), followed by Web of Science (20%), IEEE (13%), and last Emerald (2%).

To perform this systematic review, the four stages of content analysis proposed by [Gaur 2018] were followed. Hence the explanation on collecting the papers (table 3.3), and describing the concepts to review (table 3.2). Now,

Database	Emerald Insights	IEEE Xplore	Science Direct	Web of Science	Total
Records retrieved on strategy #1	91	52	982	47	**1.172**
Records retrieved on strategy #2	53	621	830	510	**2.014**
Records retrieved on strategy #3	13	22	812	134	**981**
Records retrieved on strategy #4	10	58	282	24	**377**
Records retrieved on strategy #5	10	50	200	21	**281**
Records retrieved on strategy #6	6	50	200	21	**277**
Screened by keywords	6	50	67	21	**144**
Screened by abstract	1	6	29	9	**45**

Table 3.4: Papers retrieved.

for the analysis, and interpretation of coded content a *Google Form* (https://shorturl.at/akluK) was created. This tool works as a replicable method based on explicit criteria for downsizing large amounts of text into reduced content categories of the different publications. Using this tool it is more manageable to count the frequency of the keywords and concepts with a manifest focus in the scope of this study. It also allows a qualitative approach by "categorization, summarization, and interpretation of textual data without using statistical interpretation" [Stemler 2000] with a latent focus.

CHAPTER 4

State of the Art - The past and present

The detection of "a thing, situation, etc. that is different from what is normal or expected" [Dictionary 023] has been performed since the beginning of time. The first time it was mentioned along with ML was in 1942 by the Scientific Research Society of North America, Society of the Sigma Xi. They studied the learning behaviors of their students and created an anomaly detection system that is flexible enough to accept "normal" changes, specified by the researchers [of America 1942].

Decades later, there are records from the aeronautical industry [Society 1958] and [Aeronautics 1966], where they developed computer software to approach the detection of anomalies on their several aircraft, and their air force and nuclear projects.

By 1983, there was the first publication on ML and anomaly detection. It covers inductive learning systems, learning by analogy, experimentation, experience, observation, and instruction [Anderson 1983]. From that time forward, there is a new knowledge revolution in science.

It was not until 1995 that the *Mathematical and Computer Modelling* journal published the first research on intelligent transportation systems. [Amin 1995] and [García-Ortiz 1995] explore the complexity of traffic on roads and highways and breakthroughs in computing techniques and computer systems while performing collaborative research between academia, industry, and government.

Ever since the 80s, there has been literature regarding smart cities, where city councils adopt and launch geographic information systems on the internet with data ready to be used. However, by the end of the 20th century, several authors believed that using ML for anomaly detection was a highly neglected subject and that their research "will serve to temper the development and deployment of these Intelligent transportation programs" [García-Ortiz 1995].

It was not until 2014 in Bahrain that the first open access paper published by *Science Direct* that studies anomalies in cyber security on the *Journal Procedia Computer Science* (figure 4.1 and table 4.1). They performed a multi-criterion fuzzy classification method with greedy attribute selection for anomaly-based

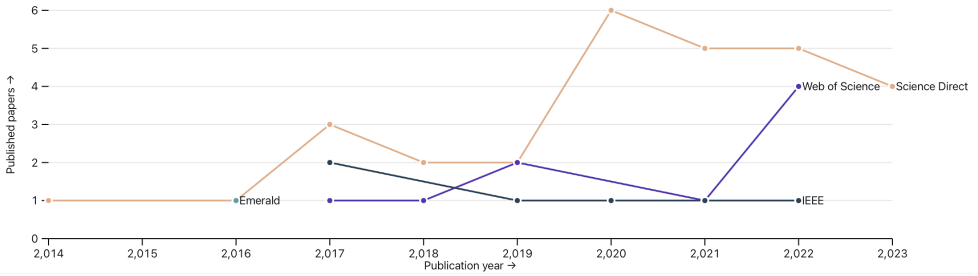


Figure 4.1: Published papers per reference databases reviewed.

intrusion detection. This approach enables it to choose an ideal subset of features most relevant for identifying intrusive events, reducing the dimension of the data set, and improving computing efficiency. "The simplicity of the constructed model allows it to be replicated at various network components in emerging open system infrastructures" [El-Alfy 2014].

In 2016, the Defense Advanced Research Projects Agency introduced the topic of XAI "to create a suite of machine learning techniques that: produce more explainable models, while maintaining a high level of learning performance (prediction accuracy); and enable human users to understand, appropriately trust, and effectively manage the emerging generation of artificially intelligent partners." [Agency 023].

There is an inevitable conflict between explainability and a model learning performance. Despite that, their project discovered evidence that comprehension may boost accuracy when testing it. Their aim was admirable; however, as a system, it faced the issue that "different user types require different types of explanations. This is no different from what we face interacting with other humans" [Gunning 2021].

Even though it started as a tool, it established the ground to perform explainable ML models. In the second decade of the 21st century, it was used for deep learning approaches in healthcare ([Lamy 2019], [Sabol 2020], [Kavya 2021]).

In the case of anomaly detection in smart cities, the first and only time that this term was utilized was in 2022 by researchers in South Korea who experimented on electrical load forecasting of buildings where they "employed the tree SHAP method to improve the explainability of the DT-based ensemble models by computing the contribution of each input variable to the prediction" [Moon 2022]. They also found that the LightGB model trained with external and internal parameters outperformed the other forecasting models in terms of prediction performance.

That said, it is time to dig deeper into the scope at hand: anomaly detection in smart cities. After following the methodology proposed, 45 papers were re-

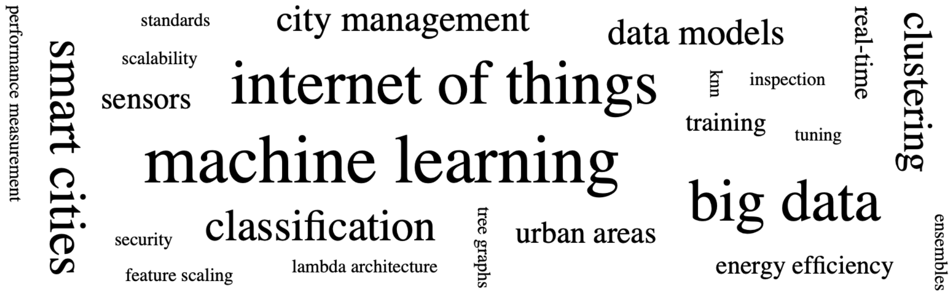


Figure 4.2: Keywords related to the papers reviewed.

viewed. They were published on the four reference databases in the past decade, from 2014 until mid-2023 (figure 4.1).

The peak of publishing so far has been in 2022, where *Science Direct* and *Web of Science* hold the most records available of open access papers. The keywords that repeated the most were *Machine Learning*, *Internet of Things*, *Classification*, and *Clustering* (figure 4.2).

This review found that the area that has been researched the most is *Transportation*, followed by *CyberSecurity* and *Energy*. This scrutiny includes studies in 8 areas of study that also include: *Education*, *Environment*, *Health*, *People*, and *Structures*.

In table 4.1, there is a classification per bibliographical reference database and the paper's area of study. Following that the chapter will be divided into the sections used as key concepts: Section 4.1 Smart Cities, Section 4.2 Data Science, and Section 4.3 Predictive Models for anomaly detection.

Table 4.1: Papers reviewed.

Area of study	**Bibliographical Reference**	**Papers**
CyberSecurity	*Science Direct*	[El-Alfy 2014], [Al-Jarrah 2018], [Mohamudally 2018], [Saranya 2020], [Bangui 2021], [Algani 2022], [Saheed 2022], [Bukhari 2023], [Yadav 2023]
	Web of Science	[Protic 2022]
Energy	*Emerald*	[Gerrish 2017]
	IEEE	[Carbone 2017]

Area of study	Bibliographical Reference	Papers
Energy	*Science Direct*	[Cerquitelli 2017], [Fonseca 2017], [Liu 2018], [Du 2019], [Ali 2020], [Himeur 2020], [Leiria 2021], [Moon 2022], [Alsalemi 2023]
Environment	*Science Direct*	[Vijai 2016]
	Web of Science	[Hangan 2022]
Health	*Web of Science*	[Abu-Alhaija 2022]
People	*IEEE*	[Zhu 2020]
	Science Direct	[Embarak 2021]
Structures	*IEEE*	[Zinno 2022]
Transportation	*IEEE*	[Zhao 2017], [Wang 2017], [Bawaneh 2019], [Xu 2019], [Nugraha 2021]
	Science Direct	[Masino 2017], [Killeen 2019], [Belhadi 2020], [Mondal 2020], [Gomari 2021], [Kyriakou 2021], [Wang 2021], [Bachechi 2022], [Lbazri 2020], [Karanfilovska 2022], [Vidović 2022], [Wu 2023]
	Web of Science	[Zantalis 2019]

4.1 Smart cities

Several scientific methodologies, ML techniques, procedures, and systems are commonly used in data science to research and analyze real-world events utilizing historical data, that way "extracting useful knowledge or actionable insights from city data and building a corresponding data-driven model is the key to making a city system automated and intelligent" [Sarker 2022].

For this section of the review, the latent categories from the concepts will be presented, and how they have developed in this century.

4.1.1 Economy

Most of the publications that have referred to the economy of a smart city are published by the Journal *Procedia Computer Science* that is on *Science Direct* bibliographical reference database (table 4.2).

Table 4.2: Papers related to the economy in smart cities.

Area of study	**Journal**	**Paper**
Energy	*Emerald Publishing Limited*	[Gerrish 2017]
Environment	*Procedia Computer Science*	[Mohamudally 2018]
People	*Procedia Computer Science*	[Embarak 2021]
Transportation	*Engineering Applications of Artificial Intelligence*	[Belhadi 2020]

The first published paper is related to building performance by researchers from Loughborough University, UK, and the professional services firm BuroHappold Engineering. They found "patterns in thermal response across monitored rooms in a single building, to clearly show where rooms are under-performing in terms of their ability to retain heat during unconditioned hours" [Gerrish 2017]. That affects the energy payment bill, especially in an energy crisis scenario when there is demand and costs out of budget.

The following year, at the 15th International Conference on Mobile Systems and Pervasive Computing, researchers from Mauritius [Mohamudally 2018] outlined the basis for building an anomaly detection engine for IoT Smart Applications. They mention examples of anomalies that could be detected that could benefit the administration, such as water leakages to prevent water waste, broken bulbs to save time and fuel for maintenance, or electricity peak and pipe leakage for energy monitoring.

In 2020, researchers from Norway [Belhadi 2020] showcased a case study of urban traffic where they compared Odense, Denmark, and Beijing, China. They don't involve direct economic advantages in a smart city, but they do mention a crucial part that is often overlooked in a budget, computational performance and time could be very consuming particularly while dealing with many categories and huge time series.

Finally, in 2021, at the 8th International Symposium on Emerging Internetworks, Communication and Mobility in Belgium, a professor from the Higher Colleges of Technology, Abu Dhabi, UAE, proposed a new paradigm in sustainable education to identify at-risk students and help them as "academic institutions incur significant costs in order to improve academic success and prevent academic dismissal" [Embarak 2021].

4.1.2 Environment

This concept was studied by more than 50% (25 papers out of 45 papers) scholars in areas related to energy, environment, people, and transportation. Most of the publications are published by *Energy Procedia* that is on *Science Direct* bibliographical reference database (table 4.3).

Table 4.3: Papers related to the environment in smart cities.

Area of study	Journal	Paper
Energy	*Emerald Publishing Limited*	[Gerrish 2017]
	Information Systems	[Liu 2018]
	Energy Procedia	[Cerquitelli 2017] [Fonseca 2017] [Du 2019]
	Applied Energy	[Ali 2020]
	Information Fusion	[Himeur 2020]
	Smart Energy	[Leiria 2021]
	Sustainable Energy Technologies and Assessments	[Moon 2022]
	Environmental Challenges	[Alsalemi 2023]
Environment	*Procedia Computer Science*	[Vijai 2016] [Mohamudally 2018]
	IEEE	[Carbone 2017]
	Water	[Hangan 2022]
Transportation	*Procedia Computer Science*	[Killeen 2019]
	Engineering Applications of Artificial Intelligence	[Belhadi 2020]
	Transportation Research Procedia	[Kyriakou 2021]
	IEEE	[Wang 2021]
Transportation	*Big Data Research*	[Bachechi 2022]

The Amrita School of Engineering was the first one to mention smart city initiatives in India [Vijai 2016]. They specialize in water since the government and organizations are highly interested in managing water sustainably from distribution systems connected to IoT devices. Their study addresses how ML techniques may be employed in elements of smart city management, such as smart water management, which involves anticipating water demand, assessing water quality, and spotting anomalies.

Establishing smart policies to enhance the sustainability and well-being of a city depends critically on analyzing data to track human-related activities. Anomalies in time series can be connected to shorter timescales like days or weeks. The researchers from the University of Padova, Italy, propose the creation of a calendar as "an additional source of information to discriminate between really unwanted anomalies and expected anomalies (e.g., weekends), or even to signal a possible anomaly whenever a “normal” behavior is not expected" [Carbone 2017].

As mentioned in section 4.1, the scholars from Université des Mascareignes, Mauritius, also outline the importance of clearly stating the conditions for detecting anomalous elements as they might be arbitrary or specific. They state that before starting implementing ML techniques, one should define if the anomalies are static or dynamic, if it’s a point or an outlier ("an outlier is not necessarily de facto an anomaly" [Mohamudally 2018]), if an anomaly is unusual in one setting but not necessarily uncommon in another, and if collective anomalies occur when there are ongoing isolated variations in time.

Researchers from Romania [Hangan 2022] published an overview of papers on *Google Scholar*. They found that there is an immense amount of publications that include the keywords "*IoT*" *and* "*water*" and a lack of papers on "*anomaly detection*" *and* "*water*". They state that there is a need for methods that extract useful information from raw data series for use in decision support systems as more data becomes accessible from smart water monitoring devices. The approach must have preprocessing procedures, feature extraction, anomaly identification (to indicate odd occurrences), pipe failure prediction, water demand modeling, and forecasting data.

4.1.3 Governance

This topic was latent throughout the areas of study; there are hints in the discussion limiting the review’s development. The area of *CyberSecurity* (table 4.5) is the one that has the closest approach when referring to the data architecture. Nonetheless, it is never implicit in the text.

4.1.4 Mobility

Most of the publications that have referred to the data governance of a smart city are published by *IEEE* and have been studied in the area of *Transportation* (table 4.4).

Table 4.4: Papers related to mobility in smart cities.

Area of study	**Type**	**Journal**	**Paper**
Cybersecurity	*People*	*Digital Communications and Networks*	[Al-Jarrah 2018]
	Vehicles	*Procedia Computer Science*	[Bangui 2021]
People	*People*	*IEEE*	[Zhu 2020]
Transportation	*People*	*IEEE*	[Zhao 2017]
		Procedia Computer Science	[Vidović 2022]
	Vehicles	*Journal of Traffic and Transportation Engineering*	[Masino 2017]
		IEEE	[Wang 2017]
			[Bawaneh 2019]
			[Nugraha 2021]
			[Wang 2021]
		Future Internet	[Zantalis 2019]
		Procedia Computer Science	[Killeen 2019]
			[Mondal 2020]
		Engineering Applications of Artificial Intelligence	[Belhadi 2020]
		Transportation Research Procedia	[Gomari 2021]
			[Kyriakou 2021]
		Big Data Research	[Bachechi 2022]
Transportation	*Vehicles*	*International Journal of Transportation Science and Technology*	[Wu 2023]

As mentioned in the introduction of this Chapter, the *Transportation* field was the first one to explore anomaly detection. This section explores the development of patterns in mobility on the scope of people and vehicles.

The crowds and people's moving patterns were first studied in 2017 at the Shenzhen Institutes of Advanced Technology. They address travel behaviors at an individual level and develop a method for gathering them using unprocessed smart card transaction data to comprehend the passengers' travel patterns.

Their findings are valuable for transportation researchers and city managers to improve metro and public transportation services. They propose in the future to "consider more factors such as passenger types (regular, student, staff), route choice (there may be several routes connecting two stops) to perform further analysis on individual passengers travel patterns, and build a complete system to distinguish a special type of anomaly passengers from normal passengers" [Zhao 2017].

Later, at Xi'an Jiaotong University, China, researchers identified patterns from cellular networks in Milan to "capture the similarities in activity series dynamics among different geographical areas and segment the city into distinct groups" [Zhu 2020]. First, they recognized the trends of how people move, and then they predicted the traffic in the following week after a month's worth of data.

Contrasted with people's patterns, vehicles have been studied the most; from public transportation ([Wang 2017], [Killeen 2019], [Zantalis 2019], [Bangui 2021], [Wang 2021], [Wu 2023]), to the road maintenance ([Nugraha 2021], [Kyriakou 2021]), and the traffic management ([Bawaneh 2019], [Mondal 2020], [Belhadi 2020], [Gomari 2021], [Bachechi 2022]).

4.2 Data Science

The smart city's plan incorporates information and communication technology (ICT) to gather and find information from their data. This boosts city operations' efficiency, enhances the quality of services, and improves the lives of its inhabitants, all of which contribute to effective results. In order to improve decision-making and deliver better services to people, city-data acquired from many sources, including sensors, IoT devices, and other external sources, is being mined for relevant insights and hidden connections [Sarker 2022].

This section will observe the discussions made on data architecture and data engineering.

4.2.1 Data architecture

Most of the publications that have referred to the data architecture of a smart city have been studied in the area of *CyberSecurity* and *Energy* (table 4.5). The Journal that has specialized the most is *Procedia Computer Science* stored in *Science Direct*.

Table 4.5: Papers related to data architecture.

Area of study	Journal	Paper
CyberSecurity	*Procedia Computer Science*	[Mohamudally 2018] [Bangui 2021] [Bukhari 2023]
	Measurement: Sensors	[Algani 2022]
	Alexandria Engineering Journal	[Saheed 2022] [Yadav 2023]
Energy	*Energy Procedia*	[Cerquitelli 2017]
	Information Systems	[Liu 2018]
	Applied Energy	[Ali 2020]
	Information Fusion	[Himeur 2020]
	Smart Energy	[Leiria 2021]
	Environmental Challenges	[Alsalemi 2023]
Environment	*Procedia Computer Science*	[Vijai 2016]
	Water	[Hangan 2022]
Structures	*IEEE*	[Zinno 2022]
Transportation	*IEEE*	[Wang 2017]
	Procedia Computer Science	[Killeen 2019]
	Transportation Research Procedia	[Kyriakou 2021]
~~Transportation~~	*Big Data Research*	[Bachechi 2022]

Area of study	Journal	Paper
	International Journal of Transportation Science and Technology	[Wu 2023]

In 2016, there was the first framework for an IoT System (figure 4.3). "The main requisite for an IoT system is the solution should be usable for everyone and not just an expert. The data received in the cloud system are stored or processed for discovering patterns and to infer knowledge" [Vijai 2016]. The applications can be data visualization, in a clear way for the user to grasp; and alert systems to provide the right kind of warning to the supervisors. Researchers who replicated this architecture are [Mohamudally 2018], [Himeur 2020], [Kyriakou 2021], [Bachechi 2022], [Zinno 2022], [Alsalemi 2023].

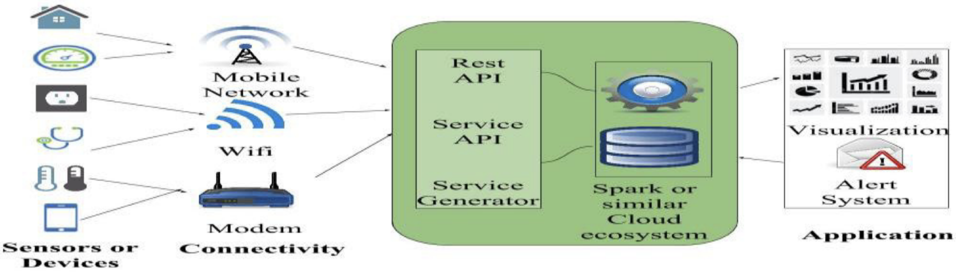


Figure 4.3: First published framework for an IoT System in this scope by [Vijai 2016].

A professor at the Politecnico di Torino in 2017 proposed SPEC (Scalable Predictor of Energy Consumption) distributed and parallel approaches "accompanied with cloud-based services (e.g. Platform-as-a-Service tools) due to the increasing volume of collected data as well as the horizontal scaling in hardware" [Cerquitelli 2017].

Their datasets were "stored in a cluster at our University running Cloudera Distribution of Apache Hadoop (CDH5.3.1). All experiments have been performed on our cluster, which has 8 worker nodes and runs Spark 1.2.0, HDFS 2.5.0, and Yarn 2.5.0. The current implementation of SPEC is a project developed in Scala exploiting the Apache Spark framework" [Cerquitelli 2017].

In figure 4.4 there is the architecture that she proposes that includes the processing of a window of 5 minutes that provides a glance of the most recent energy usage data that was recorded. It delivers a breakdown of the building's recent past energy consumption and, as a result, forecasts the building's upcoming energy needs in the near future. Researchers who replicated this architecture are [Wang 2017], [Liu 2018], [Bangui 2021], [Leiria 2021], [Algani 2022],

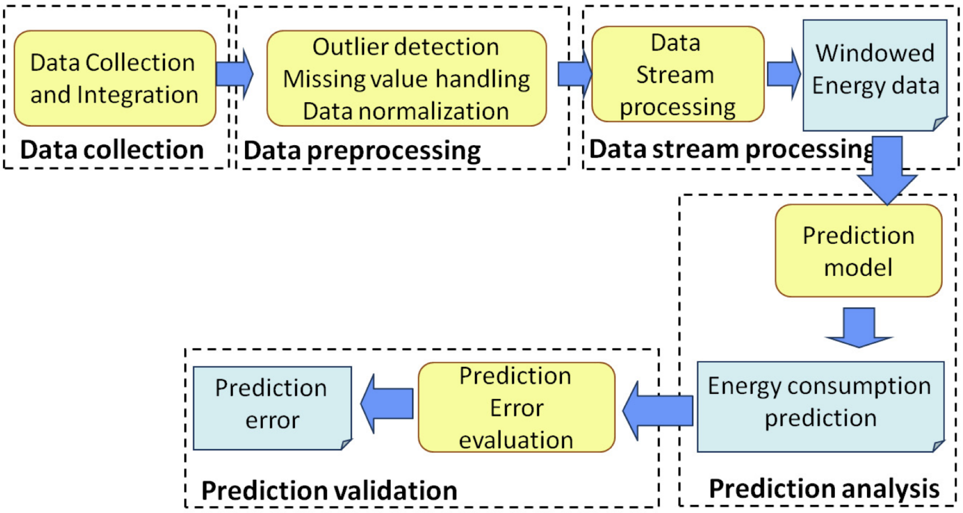


Figure 4.4: SPEC architecture energy-related applications by [Cerquitelli 2017].

[Ali 2020], [Hangan 2022], [Saheed 2022], [Bukhari 2023], [Yadav 2023], [Wu 2023].

Researchers from the Technical University of Denmark in 2018 acknowledge and name for the very first time a *Lambda Architecture* (figure 4.5).

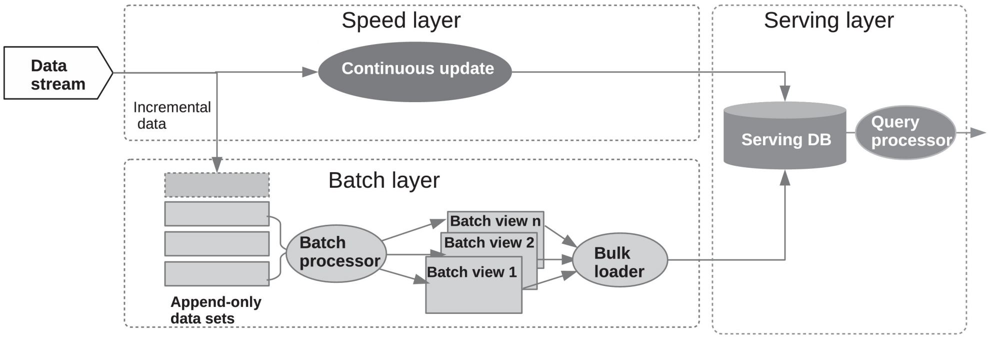


Figure 4.5: Lambda Architecture in [Liu 2018].

"In a lambda system, the data pipeline is broken down into the three layers with clear demarcation of responsibilities. For each layer, there are different technologies that can be used for the implementation. The speed layer performs low-latency computations for incremental data. The streaming technology, such as Spark Streaming, Storm or S4, can be applied to this layer. The batch layer does batch computations for entire data sets, which requires good scalability. The big data processing systems, such as Spark, Hadoop, Pig, and Hive, are the good candidates for this layer. The serving layer needs to respond user queries quickly, which requires a high-performance system. The technologies, including traditional relational data

management system (RDBMS), memory-based data stores (Redis or Memcache), are NoSQL database systems (Cassandra, MongoDB, or HBase), are the good options" [Liu 2018].

In 2019, scholars from the University of Ottawa, Canada, and Harbin University of Science and Technology, China, proposed an improved lambda architecture (figure 4.6). This time it describes what happens on the 3 layers depending on the latency. The *Perception Layer* acquires and gathers the data to send it and store it in the *Middleware Layer*, and then examines it in the *Application Layer*.

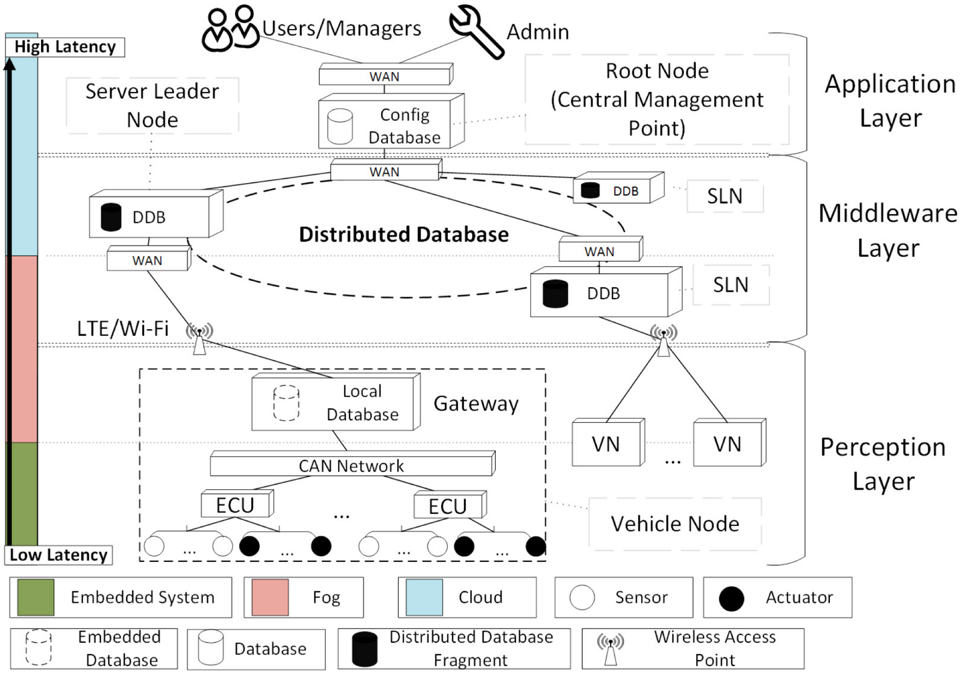


Figure 4.6: Predictive maintenance fleet management system architecture-overview diagram by [Killeen 2019].

4.2.2 Data engineering

Most of the publications that have referred to the data engineering of a smart city are published by *IEEE* and have been studied in the area of *Transportation* (table 4.6).

Table 4.6: Papers related to data engineering.

Area of study	Journal	Paper
CyberSecurity	*Procedia Computer Science*	[Bukhari 2023]
	Measurement: Sensors	[Algani 2022]
	Electronics	[Protic 2022]
	Alexandria Engineering Journal	[Saheed 2022] [Yadav 2023]
Energy	*Energy Procedia*	[Cerquitelli 2017] [Du 2019]
	Applied Energy	[Ali 2020]
	Sustainable Energy Technologies and Assessments	[Moon 2022]
	Environmental Challenges	[Alsalemi 2023]
Environment	*IEEE*	[Carbone 2017]
	Water	[Hangan 2022]
People	*IEEE*	[Zhu 2020]
Structures	*IEEE*	[Zinno 2022]
Transportation	*IEEE*	[Wang 2017] [Zhao 2017] [Bawaneh 2019] [Nugraha 2021] [Wang 2021]
	Big Data Research	[Bachechi 2022]
	International Journal of Transportation Science and Technology	[Wu 2023]

To perform some data engineering, first, there needs to be some data cleaning, as it is necessary given that the raw data frequently has unrelated data and missing data (which could lead to anomalies) [Bawaneh 2019], [Du 2019], [Nugraha 2021], [Bachechi 2022], [Hangan 2022]. That would "assure data accuracy by searching for duplicated or unrealistic information" [Yadav 2023]. According to [Cerquitelli 2017] and [Algani 2022], extreme values, also known as outliers, are not included in the training dataset, despite the fact that they are necessary for the anomaly identification method.

Subsequently, since it "helps to eliminate defects in the dataset" [Saheed 2022]. One way to do it would be by using a feature scaling approach known as normalization to scale all of the attribute values to the same scale. Standardized moment, z-score normalization, and min-max normalization are a few examples of these techniques [Carbone 2017].

A key for data reduction is selecting solely the essential features that are capable of meeting the needs of the model in order to decrease overfitting and increase accuracy [Wang 2017], [Ali 2020], [Protic 2022], [Alsalemi 2023]. A threshold limit is chosen for the datasets using information gain as the foundation in order to choose the features. Information gathering that goes beyond a certain point is always a must for accurately identifying cyberattacks [Bukhari 2023].

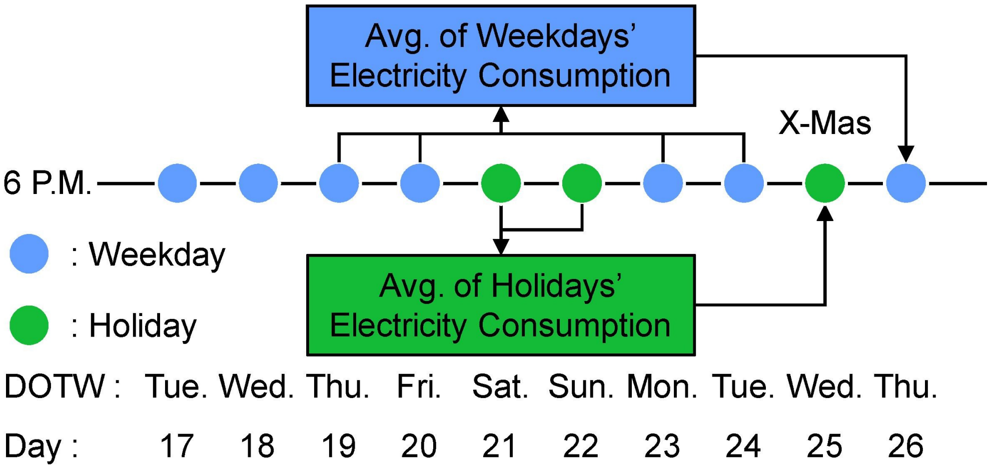

Avg. of Weekdays' Electricity Consumption
X-Mas
6 P.M.
: Weekday
: Holiday
Avg. of Holidays' Electricity Consumption
DOTW : Tue. Wed. Thu. Fri. Sat. Sun. Mon. Tue. Wed. Thu.
Day : 17 18 19 20 21 22 23 24 25 26

Figure 4.7: Example of the weekly average electricity consumption calculation (December 2019). DOTW: days of the week; X-Mas: Christmas. By [Moon 2022].

Data integration/fusion is the process of combining data from different sources. It could be used to also transform the current variables or add new ones [Zhao 2017], [Ali 2020], [Wang 2021], [Zinno 2022], [Wu 2023]. For example: a calendar with holidays that happen on specific days that could alter the trending pattern (figure 4.7).

Finally, data discretization "replaces numerical attributes with nominal values" [Ali 2020]. Equally important, to avoid data sparsity, you can sum values up to a single value that describes the total activity generated by a category [Zhu 2020].

4.3 Predictive models

Anomalies come in many forms, including cyber assaults, data quality, and data values. Consequently, to predict it will change from one anomaly to another. In addition, it could also have a different approach depending on the data. It could perform Supervised or Unsupervised ML.

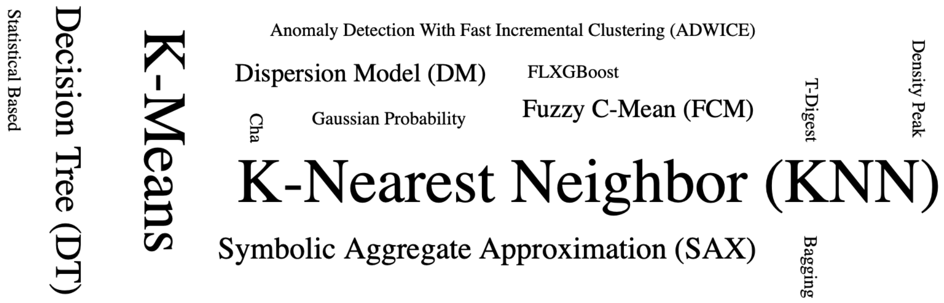


Figure 4.8: Models studied on the reviewed papers.

In figure 4.8, there are the models mentioned in the reviewed, and figure 4.9 shows the share of the frequency that kind of ML approach was utilized.

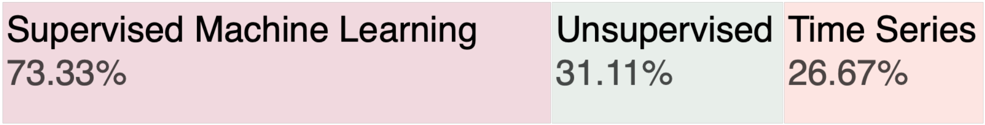


Figure 4.9: Share of type of machine learning models used on the reviewed papers.

4.3.1 Supervised Machine Learning

Most of the publications that have referred to Supervised ML to detect anomalies are classification models like KNN, SVM, and RF. They have been studied the most in the area of *Transportation* (table 4.8).

The only regressions that were performed were *Linear Regression*, one in road traffic [Mondal 2020] and one in energy consumption [Leiria 2021].

Table 4.7: Papers related to supervised ML models.

Area of study	Class	Approach	Paper
CyberSecurity	Classification	RF	[Saranya 2020]
		SVM	[El-Alfy 2014] [Xu 2019] [Bukhari 2023]
		KNN	[Protic 2022]
		XGBoost	[Karanfilovska 2022] [Saheed 2022] [Yadav 2023]
Energy	Classification	GBoost	[Ali 2020]
		KNN	[Himeur 2020]
		LightGBM	[Moon 2022]
		RF	[Cerquitelli 2017]
		RS	[Alsalemi 2023]
	Regression	Linear Regression	[Leiria 2021]
Environment	Classification	KNN	[Hangan 2022]
Structures	Classification	RF	[Zinno 2022]
People	Classification	RS	[Embarak 2021]
Transportation	Classification	Bayesian Model	[Vidović 2022]
		COSMO	[Killeen 2019]
		K-D tree	[Masino 2017]
		Logical Regression	[Wang 2017]
		RUSBoost	[Kyriakou 2021]
		SVM	[Wang 2017]

Area of study	**Class**	**Approach**	**Paper**
Transportation	Classification	SVM	[Zantalis 2019] [Lbazri 2020] [Nugraha 2021]
		XGBoost	[Wang 2021]
	Regression	Linear Regression	[Mondal 2020]

Table 4.8: Findings on papers related to supervised ML models.

Citation	Area of study	Approach	Findings
[El-Alfy 2014]	CyberSecurity	SVM	It explores a new countermeasure approach for anomaly-based intrusion detection using a multicriterion fuzzy classification method combined with a greedy attribute selection.
[Cerquitelli 2017]	Energy	RF	The preliminary version of the SPEC (Scalable Predictor of Energy Consumption) engine to address the fine grain prediction of energy consumption over a sliding time window.
[Masino 2017]	Transportation	K-D tree	The results show that the method and Euclidean distance performs best and is robust in transferring the information of the road surface from one vehicle to another.
[Wang 2017]	Transportation	Logical Regression (LR) and SVM	LR outperforms SVM and decision trees in prediction accuracy and F-score measurement, while SVM is capable of identifying the largest number of unlicensed taxis. The model's learning phase can be improved even further with other heterogeneous datasets like demographics, sites of interest, etc.

Citation	Area of study	Approach	Findings
[Killeen 2019]	Transportation	Consensus self-organized models approach (COSMO)	It proposes a novel IoT architecture for predictive maintenance and proposes a semi-supervised machine learning algorithm that attempts to improve the sensor selection performed in a predictive maintenance system.
[Xu 2019]	CyberSecurity	One Class SVM	The predictive performance of the tuned Local Outlier Factor is comparable to the predictive performance with the best results on the Http and Smtp data, and it outperforms all the other methods on Credit and Mnist data.
[Zantalis 2019]	Transportation	SVM	Given the current applications and infrastructure regarding IoT and ML, a comparatively smaller ML coverage for the smart lighting systems and parking applications are detected.
[Ali 2020]	Energy	Gradient Boosted trees	It extracts the knowledge from available resources to identify the existing building energy performance and formulate retrofit solutions.
[Himeur 2020]	Energy	KNN	They discussed the usefulness of several data fusion strategies that have been implemented or could be deployed to decrease energy wastage and promote sustainability.
[Lbazri 2020]	Transportation	SVM and RF	While training the SVM had a better accuracy score than the RF. However, test results are better than the SVM.

Citation	Area of study	Approach	Findings
[Mondal 2020]	Transportation	Linear Regression	It proposes a technique based on a statistical model which identifies the temporal outliers. It can be used to detect unusual traffic incidents or sensor failures.
[Embarak 2021]	People	RS	All proposed solutions fell within the scope of predictions that result in active or proactive actions to support universities and learners. On the other hand, they fail to comprehend the various forms of education systems and whether it appropriate for the twenty-first century and future generations.
[Saranya 2020]	CyberSecurity	RF	It conveys that the detection rate, false positive rate, and accuracy not only depend on the algorithm but also depend on the application area.
[Kyriakou 2021]	Transportation	RUS Boosted trees	It has already been field-tested for the detection and classification of cracks, rutting, ravelling, patches and potholes, exhibiting accuracy levels higher than 90%.
[Leiria 2021]	Energy	Linear Regression	They analyze measured variables (heat consumption, outdoor temperature, wind speed, and global radiation) to acquire new knowledge on the building characteristics.

Citation	Area of study	Approach	Findings
[Nugraha 2021]	Transportation	One Class SVM	To accurately predict the condition of critical components, it can be started with data collection, followed by detecting normal and abnormal behavior, and continued by training algorithms to make predictions.
[Wang 2021]	Transportation	Focal Loss - XGBoost	For the case of sample imbalance, the model has a high accuracy rate in resisting tampered data domain messages.
[Hangan 2022]	Environment	KNN	The model calculates an anomaly score for each day, based on ten features of daily demand and its historical context. The authors use calendar contexts within the anomaly detection algorithm. A calendar context for a day d is a subset of days from the database D that would be expected to have similar water use as d. This allows them to give possible causes for the detected anomalies. The score and its explanations are posted to users to help them track down the physical causes of anomalies.
[Karanfilovska 2022]	CyberSecurity	XGBoost	Using the approach with StandardScalerWrapper achieved in Azure AML tool, it can be said that machine learning is very applicable in solving anomaly detection problems in IoT networks. This was also shown with the experiments performed with the AE2EML tool, which have resulted with F-Score greater than 95%

Citation	Area of study	Approach	Findings
[Moon 2022]	Energy	LightGBM	They confirmed that the Temperature-Humidity Index (THI) and the Wind Chill Index (WCT) exhibited more influence on forecasting model construction than temperature, humidity, and wind speed in weather information. Because the Shapley additive explanations value was large for THI in summer and WCT in winter, they estimated that they contribute to more accurate day-ahead hourly electricity consumption forecasting for summer and winter, respectively.
[Protic 2022]	CyberSecurity	Weighted KNN	The results show a clear benefit of the Tangent-Hyperbolic normalization used for scaling regarding processing time. Regardless of how accurate the classifiers are, their decisions can sometimes differ.
[Saheed 2022]	CyberSecurity	XGBoost	They performed a dimensionality reduction with Principal Component Analysis (PCA). They infer that using machine learning techniques for successful anomaly detection in the IoT environment is both realistic and practicable.
[Vidović 2022]	Transportation	Bayesian Model	The model when paired with additional data sets (e.g. public transport timetables, location of public transport stations, information on public transport lines, etc.), it can be used for modal split detection.

Citation	Area of study	Approach	Findings
[Zinno 2022]	Structures	RF	Their results showed that less sensors were needed to measure acceleration responses in order to figure out where damage is and how bad it is. Compared to neural network training methods, the proposed method for identifying damage could get good results quickly and with much less computational work and time.
[Alsalemi 2023]	Energy	RS	It presents a modular system for improving domestic household energy savings. It is designed to create customizable sub-components that adapt to the nature of the data and the end-user's preference, such as modules that recommend based on usage patterns, power consumption, and occupancy.
[Bukhari 2023]	CyberSecurity	SVM	The paper goes on to examine the role of ensemble techniques like bagging and boosting to provide an additional security layer to the detection architecture.
[Yadav 2023]	CyberSecurity	XGBoost	They propose a novel combined feature selection method known as the Fast Correlation-based Feature Selection (FCBFS) with the approach, which can successfully minimize the number of features while maintaining a high classification precision and recognition rate.

4.3.2 Unsupervised Machine Learning

Most of the publications that have referred to Unsupervised ML to detect anomalies are clustering models with K-Means. They have been studied the most in the area of *CyberSecurity*, and *Transportation* (table 4.10).

There is one paper that tries DBSCAN for on-street parking behavior in Munich [Gomari 2021], and one that uses Dirichlet Mixture for the behavior of wireless networking [Algani 2022].

Table 4.9: Papers related to unsupervised ML models.

Area of study	Approach	Paper
CyberSecurity	Dirichlet Mixture	[Algani 2022]
	K-Means	[Al-Jarrah 2018] [Bangui 2021] [Karanfilovska 2022]
Energy	K-Means	[Du 2019]
Environment	K-Means	[Mohamudally 2018] [Hangan 2022]
Health	K-Means	[Abu-Alhaija 2022]
People	K-Means	[Zhu 2020]
Structures	Fuzzy C-Means	[Zinno 2022]
Transportation	DBSCAN	[Gomari 2021]
	K-Means	[Zhao 2017] [Belhadi 2020] [Gomari 2021]

Table 4.10: Findings on papers related to unsupervised ML models.

Citation	Area of study	Approach	Findings
[Zhao 2017]	Transportation	K-Means	They looked into the passenger travel distribution patterns and find out the abnormal passengers based on the empirical knowledge. Then, they classified the passengers in terms of the similarity of their travel patterns.
[Al-Jarrah 2018]	CyberSecurity	K-Means	They introduced the use of a weighted Euclidean distance measure based on the observation that different attributes might have a strong impact on the resultant partitions of data. It assigns a weight for each attribute based on its significance in distinguishing between class types. These weighted attributes can lead to a higher probability of obtaining atomic clusters with a lower value of K.
[Mohamudally 2018]	Environment	K-Means	The found that ML in the unsupervised mode is indeed very efficient in situations where datasets are unpredictable. Moreover, cases where data points show non-linear time series require multivariate analysis that makes the process more computing-intensive.

Citation	Area of study	Approach	Findings
[Du 2019]	Energy	K-Means	Users in the same category do not necessarily have the same energy consumption patterns, which potentially leads to unfair prices and many other practical issues. Their results can serve as potential inputs for future energy price models, demand-side management, and load-reshaping strategies.
[Belhadi 2020]	Transportation	K-Means	For each location, they have observed different flow values represented by a time series. Applying space–time series clustering on these data allows the grouping of locations that have similar traffic behaviors.
[Zhu 2020]	People	K-Means	This approach can cluster areal units with similar traffic patterns and segment a city into distinct groups. Then, in grouped areas, they detect anomalous behaviors of the cellular network and verify the accuracy of the results using ground truth information collected from online sources.
[Bangui 2021]	CyberSecurity	Weighted K-Means	They apply the coreset method to deal with computational complexity in clustering and the large volume of vehicular datasets by extracting critical contents without examining all data content, and then enable IDSs to ensure timely network security. This approximate technique is not only just giving a quick viewability of original data, but also helps with scaling Big Data Analytics techniques during data processing.

Citation	Area of study	Approach	Findings
[Gomari 2021]	Transportation	Two-stage: DBSCAN – K-Means	The method can immediately provide first insights on the spatio-temporal parking behavior that exists within a city while employing a random automated data collection by a fleet of vehicles representing normal human mobility behavior, with a bias towards the group of vehicle users.
[Algani 2022]	CyberSecurity	Dirichlet Mixture	It proposes a hybrid anomaly detection method that combines several characteristic selecting strategies with an appropriate mixture approach to recognize each assault form with great precision.
[Abu-Alhaija 2022]	Health	K-Means	Since the probability of false alarms poses a serious impact on the accuracy of cardiac arrhythmia detection, it is the most important factor to keep false alarms to the lowest level.
[Hangan 2022]	Environment	K-Means	They label the clusters according to the predominant peak demand time and discover that patterns that contained multiple peaks during the day were more prone to internal leaks.
[Karanfilovska 2022]	CyberSecurity	K-Means	The results obtained with the PyCaret tool have shown that the silhouette score and distribution of the clusters were improved after applying PCA, while the homogeneity, Rand Index and completeness were better for the clustering without PCA.

Citation	Area of study	Approach	Findings
[Zinno 2022]	Structures	Fuzzy C-Means	They investigated how vibrations may be utilized to detect deterioration in a truss bridge model and suggested a novel approach based on fuzzy clustering and reduced frequency response function (FRF) data using principal component projection.

4.3.3 Time Series

Most of the publications that have referred to Time Series to detect anomalies with SAX model, which has been studied in the areas of *Energy*, and *Environment* (table 4.12).

Table 4.11: Papers related to forecasting time series.

Area of study	**Approach**	**Paper**
Energy	SAX	[Fonseca 2017]
	Average of Heat Loss	[Gerrish 2017]
	PARX	[Liu 2018]
Environment	Least Square SVM	[Vijai 2016]
	SAX	[Carbone 2017]
	DTW	[Bachechi 2022]
	BSTS	[Wu 2023]
Transportation	OBADA	[Bawaneh 2019]

Table 4.12: Findings on papers related to time series predictions.

Citation	Area of study	Approach	Findings
[Vijai 2016]	Environment	Least Square SVM	There is a keen interest from the organizations and government to make proper usage of water. The same can be achieved by proper monitoring and management of water distribution systems.
[Carbone 2017]	Environment	Symbolic Aggregate approXimation (SAX)	They conjecture that different normalization horizons allow to include in the shape of the time series patterns an additional, variable, component from a longer period trend. To support the analysis phase, a calendar can be used as an additional source of information to discriminate between really unwanted anomalies and expected anomalies (e.g. weekends), or even to signal a possible anomaly whenever a "normal" behavior is not expected.
[Fonseca 2017]	Energy	Symbolic Aggregate approXimation (SAX)	The number of clusters and accuracy of SAX highly depends on the highly sensitive input variables related to size. The approach is subjected to three fitness objectives, i.e., maximize data accuracy and compression and minimize complexity.

Citation	Area of study	Approach	Findings
[Gerrish 2017]	Energy	Average of Heat Loss	The response of a single space to changing internal and external temperatures can be used to determine whether it responds differently to other monitored buildings.
[Liu 2018]	Energy	PARX and Gaussian probability models	They propose a system that uses a prediction-based detection method, combined with a novel lambda architecture for iterative model updates and real-time anomaly detection.
[Bawaneh 2019]	Transportation	Occupancy Based Anomaly Detection Algorithm (OBADA)	They searched for subsequence of major changes in values in the occupancy's time series which reflects an inordinate behavior.
[Bachechi 2022]	Environment	Dynamic Time Warping and dispersion model	They demonstrate the potential of a dashboard in identifying trends, seasonal events, abnormal behaviors, and understanding how urban vehicle fleet affects air quality.
[Wu 2023]	Environment	Bayesian Structural Time Series	Since the primary purpose of the paper is to design a practical and ready-to-use data acquisition and processing framework, some other methods, such as Hamiltonian Monte Carlo for outlier detection and Generative Adversarial Network for data imputation, have not been involved. These models should be evaluated and compared in future studies.

Chapter 5

Methodological Contribution

The approach for the experimentation will follow the CRISP-DM methodology. It will get the necessary knowledge in the industry, and then the data will be processed to train the ML algorithms and detect anomalies in the dataset.

In table 5.1, the are three categories that explain the field by the consumption, efficiency, and the load of energy.

Table 5.1: Papers reviewed related to energy.

Model	**Topic**	**Approach**	**Paper**
Time Series	Energy consumption	SAX	[Fonseca 2017]
		Average of Heat Loss	[Gerrish 2017]
		PARX	[Liu 2018]
Supervised ML	Energy efficiency	GBoost	[Ali 2020]
		KNN	[Himeur 2020]
		RS	[Alsalemi 2023]
	Electrical load	LightGBM	[Moon 2022]
	Energy consumption	RF	[Cerquitelli 2017]
		Linear Regression	[Leiria 2021]
Unsupervised ML	Energy consumption	K-Means	[Du 2019]

Energy consumption is the most studied. Some authors have talked regarding classification methods, like [Cerquitelli 2017] and [Leiria 2021]; and clustering, like [Fonseca 2017] and [Du 2019].

After assessing the different approaches, the models that are appropriate to work are:

- for forecasting purposes the SAX with SARIMA, or the Linear Regression,

- for classification effects the RF. Since on other fields the KNN, SVM and XGBoost had high accuracy scores they will be also be taken into account,

- for clustering outcomes the K-Means.

5.1 Data acquisition

The data will be accessed from their open data website (table 5.2). They have the daily electric consumption by postal code, economic sector, and time interval in Barcelona, according to the data provided by the Datadis platform [Hall 023b].

Another thing to take into account is that the records don't come as a time series but as aggregation windows in 5 categories: from 00:00:00 to 05:59:59, from 06:00:00 to 11:59:59, from 12:00:00 to 17:59:59, from 18:00:00 to 23:59:59, and not specified. It is available from January 1, 2019, until March 31st, 2023 at the time of study.

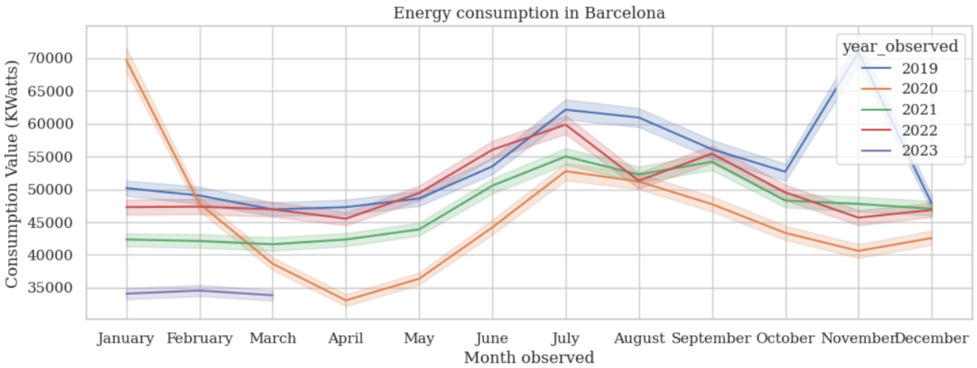


Figure 5.1: Time series of energy consumption in service buildings in Barcelona overlapped by year.

Due to computational performance issues, only the nighttime windows were used as it is when they should be idle, and an anomaly could be more feasibly detected. In figure 5.1 there is a distribution of the energy consumption in that time frame.

They have the accumulated data of 3 economic fields (industry, residential, and services, figure 5.2). However, the scope of this project is to facilitate and assist the job at a governmental level, then the services buildings were selected to perform the modeling.

To sum up the acquisition part, there were initially 1,051,455 records. As a result, performance was weak while using a mere 2GB of memory and limiting the choice to only one type of building. When focusing on public sector facilities, it is selected only to leave data on service buildings in the Gothic borough,

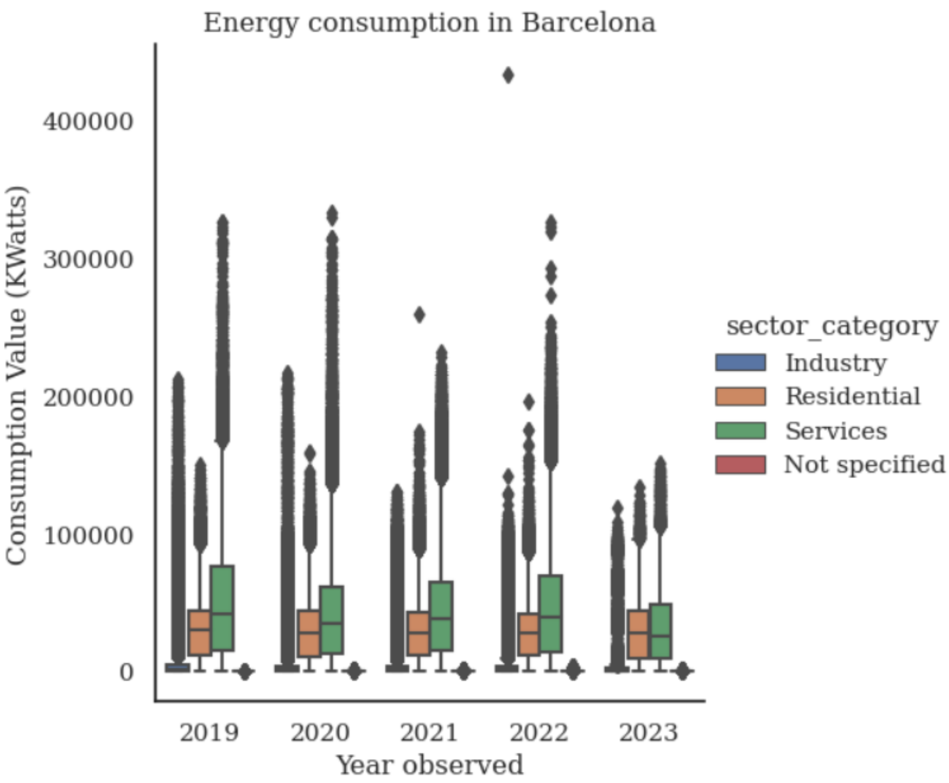


Figure 5.2: Distribution of energy consumption in Barcelona divided by economic field.

leaving 327,180 records to work. This trial is restricted to the overnight period (from 18h to 5h) to lower the size of the dataset, as anomalies can be found when the service buildings should be consuming the least energy. We are now left with 130,872 records to perform ML techniques.

Table 5.2: Information of the dataset: *Electricity consumption by postal code, economic sector, and time interval in the city of Barcelona* by [Hall 023b]

Field	**Description**
Title	Electricity consumption by postal code, economic sector, and time interval in the city of Barcelona
More information	https://datadis.es
Agenda 2030. SDG Principal	SDG 7: Affordable and clean energy
Agenda 2030. SDG Collateral 1	N/A

Field	Description
Agenda 2030. SDG Collateral 2	N/A
Source	Datadis. La plataforma de dades de consum elèctric
Geolocation	No
Long format available	Yes
Historical information	Yes
CKAN API available	Yes
Token required	No
Management	Gerència Municipal
Department	Oficina Municipal de Dades - Departament d'Estadística i Difusió de Dades
Publication Date	20/12/2022
Update frequency	Monthly

5.2 Data processing

The dataset for *electricity consumption in Barcelona* has a quite similar structure to the Swedish city's heat meter records of buildings used by [Du 2019]. In this publication, they propose a data processing procedure (figure 5.3) where they first pre-process the data, then detect the anomalies, extract some features, and then cluster the users' behavior, all based on heat consumption.

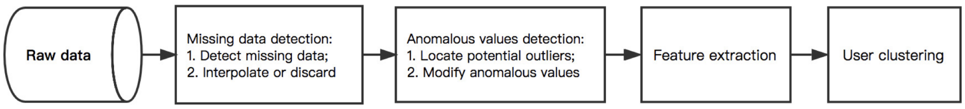

Raw data
Missing data detection:
1. Detect missing data;
2. Interpolate or discard
Anomalous values detection:
1. Locate potential outliers;
2. Modify anomalous values
Feature extraction
User clustering

Figure 5.3: Data processing procedure proposed by [Du 2019].

5.3 Data pre-processing

To manage immense amounts of data, some of Apache Spark's built-in features from its library *pyspark*, as mentioned by [Liu 2018], perform Data Analytics in Big Data. The structure to follow will be the one taken from Chapter 4, Section 4.2.2 Data Engineering.

5.3.1 Data cleaning

The first step was to transform the data that was extracted with an incorrect data type. In this case, *date_observed*, *year_observed* and *postal_code*. Then it was time to translate the records as they were published in Catalan.

5.3.2 Feature scaling

In this part, the attributes were classified between categorical and numerical data. The former was normalized utilizing the MinMax method.

5.3.3 Data integration

From the *date_observed* new attributes were created to find the month and the day of the week the energy consumption tool place.

As mentioned in the review by [Moon 2022], calendars are an amazing tool to guide the difference between abnormal consumption and anomaly detection. It is for that reason that the Data on city festivals in Barcelona(table 5.3) was extracted and pre-processed following the same steps.

Only data was available from 2019 until 2022, which is perfect as that is the range necessary to train the data. Most holidays are periodic, on the same day, like Catalonia's National Day, on September 11th, or The Three Kings Parade on January 5th. A limitation is that the list does not take into account events such as concerts or football matches that move a massive amount of crowd, and that can have an effect on energy consumption and could cause a false anomaly.

Table 5.3: Information of the dataset: *Data on city festivals in the city of Barcelona* by [Hall 023a]

Field	**Description**
Title	Data on city festivals in the city of Barcelona
More information	http://barcelonadadescultura.bcn.cat/festes/dades?lang=en
Agenda 2030. SDG Principal	SDG 10: Reduced inequalities
Agenda 2030. SDG Collateral 1	SDG 11: Sustainable cities and communities
Agenda 2030. SDG Collateral 2	N/A
Source	Secretaria Tècnica. Institut de Cultura. Ajuntament de Barcelona

Field	Description
Geolocation	No
Long format available	Yes
Historical information	Yes
CKAN API available	Yes
Token required	No
Management	Gerència d'empresa, cultura i innovació
Department	Pla de Sistemes
Publication Date	25/06/2014
Update frequency	Annual

5.3.4 Feature Selection

Once the two datasets, energy consumption and activities in the city, have been fused, the event location attribute and the postal code are compared to label if an event has occurred in that area. In this way, only this new feature is selected and the other two are discarded.

5.3.5 Data discretization

To classify if there has been an anomaly in the recorded data. The average of that day in previous years is calculated with a margin of 5% above and below to determine the nature of consumption.

After the selection of categorical features, these were treated to have a one-hot-encoding approach. In figure 5.4 there are the relationships between the selected features after performing the discretization.

5.4 Model building for Anomaly detection

The purpose is to forecast the energy consumption and alert if the values are above or below the historic registers. To do so, the data set was split into 2 groups. One to train the data from 2019 to 2022, and one to test and validate it as there is data in 2023.

The first part was to check if the time series was stationary. To assess it, the ADF (Augmented Dickey–Fuller) test was employed to determine if transforma-

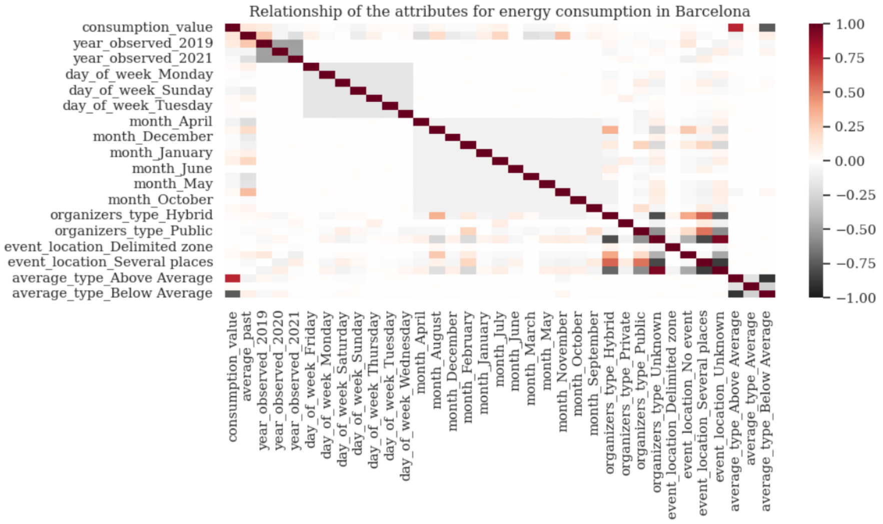


Figure 5.4: Features distribution and intensity in the electricity consumption dataset.

tions should be made to the time series to stabilize it before further modeling or forecasting.

The hypotheses for this test are:

- H0: The time series has a unit root and is therefore non-stationary.

- H1: The time series does NOT have a unit root and is therefore stationary.

Once the time series is forecasted, then the classification and clustering will be performed to detect anomalies and find the pattern of energy consumption respectively.

CHAPTER 6

Results

6.1 Forecasting the energy consumption

The p-value (0.001) is smaller than 0.05. Consequently, there is sufficient statistical evidence to reject the null hypothesis and determine that we are facing a stationary time series.

ADF statistic: -3.9789548077150867

p-value: 0.0015248090038454789

Lags used: 22

Critical values: '1%': -3.4349056408696814, '5%': -2.863552005375758, '10%': -2.5678411776130114

When performing the SAX + SARIMA approach proposed by [Carbone 2017] [Fonseca 2017]. A *UserWarning* was flagged, as there were not enough data points for identifying the seasonal ARMA's initial parameters, and except for variances they were set to zeros.

In view of this, a regression approach was considered that was used in data from smart energy meters to gain knowledge about households connected to the district heating network in Denmark [Leiria 2021].

The forecast was modeled to predict the energy consumption in the Gothic borough from March 2022 until March 2023. In figure 6.1 there are the results overlap with the recorded data in [Hall 023b].

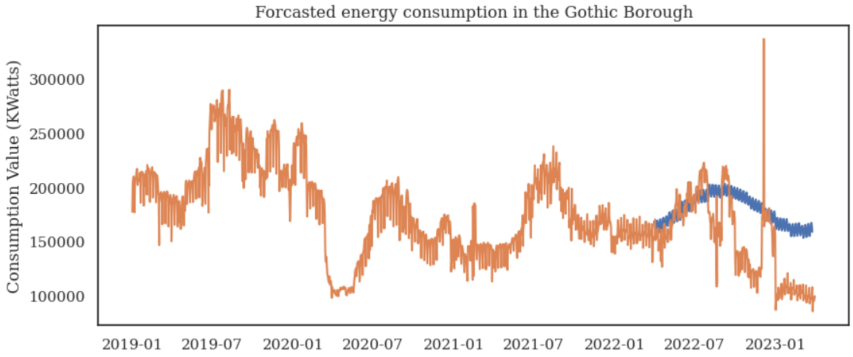


Figure 6.1: Time series of the original data set and the predicted energy consumption.

6.2 Anomaly detection

To classify if there has been an anomaly in the forecasted data. Using the previous consumption calculated in the data discretization with the above and below margin of 5%. With that feature that determines if the recorded data is within the average or not, the training and validation data sets were divided. The first one had the data from January 1st, 2019, until December 31st, 2022, and the second one had the 2023 data.

The approaches followed from the energy field of study were:

- KNN by [Himeur 2020]
- RF by [Cerquitelli 2017]

However, approaches like the XGBoost(used by [Wang 2021], [Karanfilovska 2022], [Saheed 2022]), and SVM (used by [Wang 2017], [Nugraha 2021]) were also considered.

6.2.1 Model evaluation

As a result of so many record reductions due to computational resources, when testing the training set, the accuracy showed an overfitted outcome from KFold Cross-Validation:

- KNN: 99.45%
- SVM → linear: 100.0%
- SVM → Radial Basis Function: 100.0%

In spite of this, the models on the forecasted data were conclusive. In table 6.1 the models have a similar accuracy. Nonetheless, the KNN model was more accurate and had fewer mislabeled points than the others. Still, there is a lot of room for improvement as the predictions are labeled as *Average* non-average records (figure 6.2) which could be harmful.

Table 6.1: Model comparison for classification energy consumption

ML Model	Accuracy	Mislabeled points
KNN	88.09%	10 out of 84
RF	85.71%	12 out of 84
SVM	85.71%	12 out of 84
XGBoost	85.71%	12 out of 84

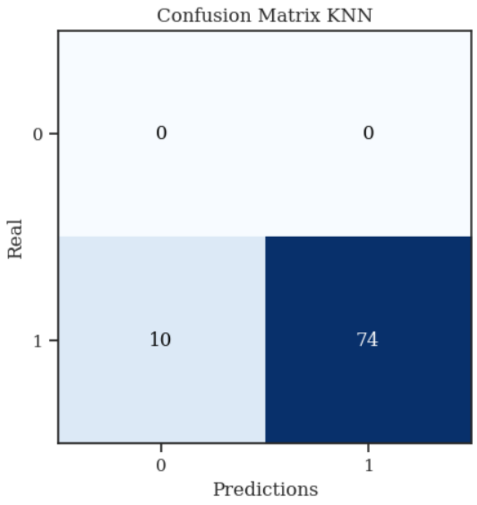


Figure 6.2: Confusion matrix according to the most accurate model trained.

To have an idea of what the other models were taking into account for the classification, in figure 6.3 is the importance of the trained features.

In this case, the most important ones are the average type (below, above and average) and the consumption value despite having more attributes like the month or the activities that were happening that day.

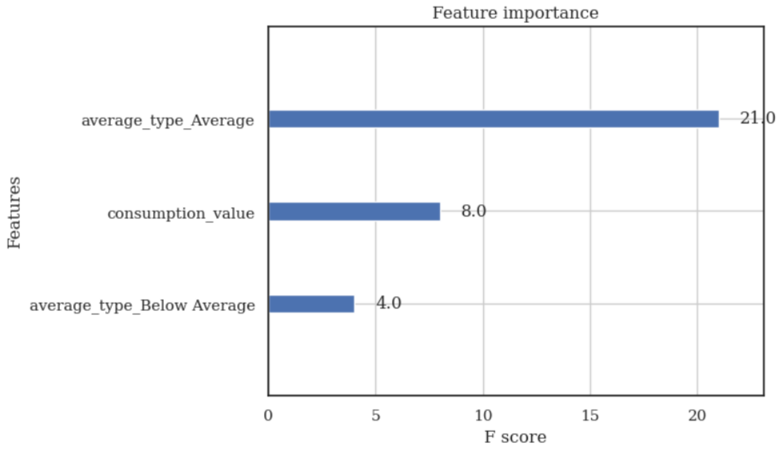
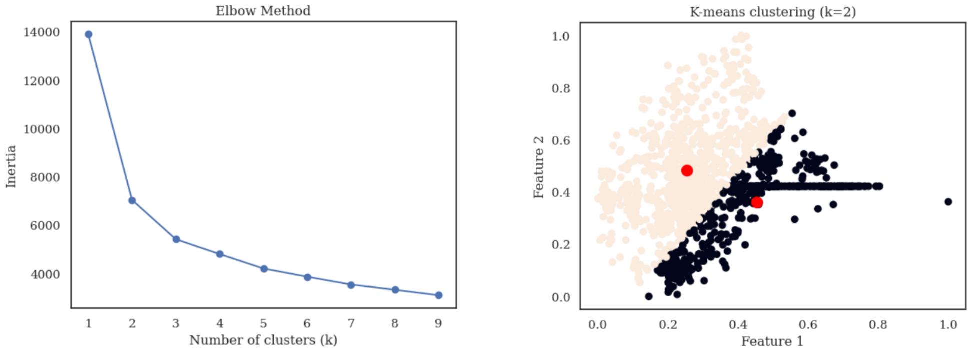


Figure 6.3: Feature importance in XGBoost model.



(a) Elbow Method (b) K-Means with 2 clusters

Figure 6.4: K-Means results

6.3 Pattern consumption

Unsupervised ML is when the algorithm just searches for patterns in the data, therefore, there is no outcome that can be predicted. In the K-Means case, the algorithm estimates the centroid of each set after randomly assigning each observation to a group. Then, in each iteration, it reassigns the data points to the nearest cluster's centroid.

Even though, in figure 6.4a there is a turning like an elbow between k=2 and k=3. For this dataset, the optimal number of clusters is 2 with a prediction strength of 1.0 (figure 6.4b).

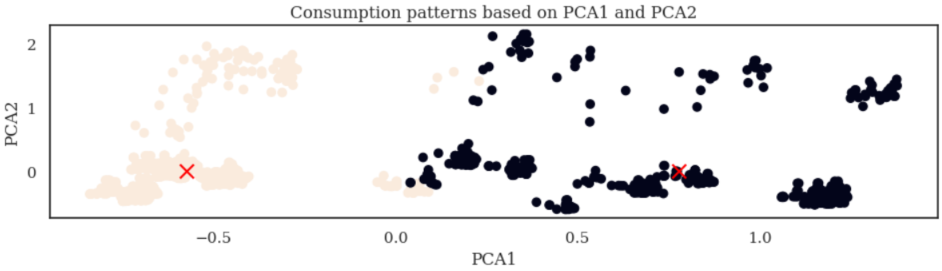


Figure 6.5: Clustering on the PCA components.

The dataset was minimize from 37 features to 2 features using principal component analysis (PCA). In figure 6.5, there is a representation of the clusters differentiated by a color parameter from the model's labels; and in figure 6.6, there is the components' relationship with the training features for energy consumption.

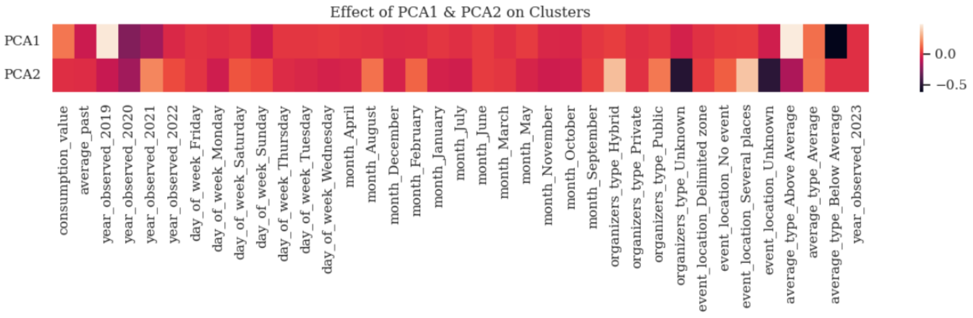
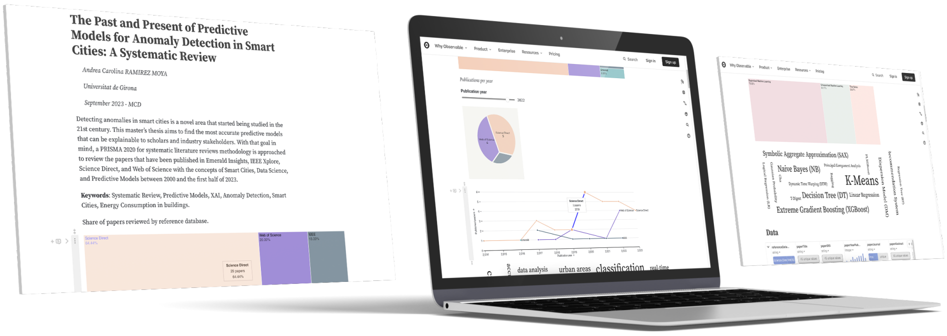


Figure 6.6: PCA on features.

6.4 Communication

The final outcome of this project is to spread the word and make it easier for people in the academia and the industry to gain more knowledge of detecting anomalies in smart cities with predictive models. A way to visualize the results found is by presenting a deliverable.

To use the resources learned throughout the master's degree, an *Obervable HQ* was created with systematic review that can be found in: https://observablehq.com/@andreaudg/anomaly-detection-lit-rev (figure 6.7).


The Past and Present of Predictive Models for Anomaly Detection in Smart Cities: A Systematic Review

Andrea Carolina RAMIREZ MOYA

Universitat de Girona

September 2023 - MCD

Detecting anomalies in smart cities is a novel area that started being studied in the 21st century. This master's thesis aims to find the most accurate predictive models that can be explainable to scholars and industry stakeholders. With that goal in mind, a PRISMA 2020 for systematic literature reviews methodology is approached to review the papers that have been published in Emerald Insights, IEEE Xplore, Science Direct, and Web of Science with the concepts of Smart Cities, Data Science, and Predictive Models between 2000 and the first half of 2023.

Keywords: Systematic Review, Predictive Models, XAI, Anomaly Detection, Smart Cities, Energy Consumption in buildings.

Share of papers reviewed by reference database.

Figure 6.7: Notebook in ObservableHQ.

Then to expose the experimentation results a dashboard was created to grasp the energy consumption in service buildings in the Gothic borough at nighttime. It can be found in https://lookerstudio.google.com/s/j_93RGA8KF0 (figure 6.8).

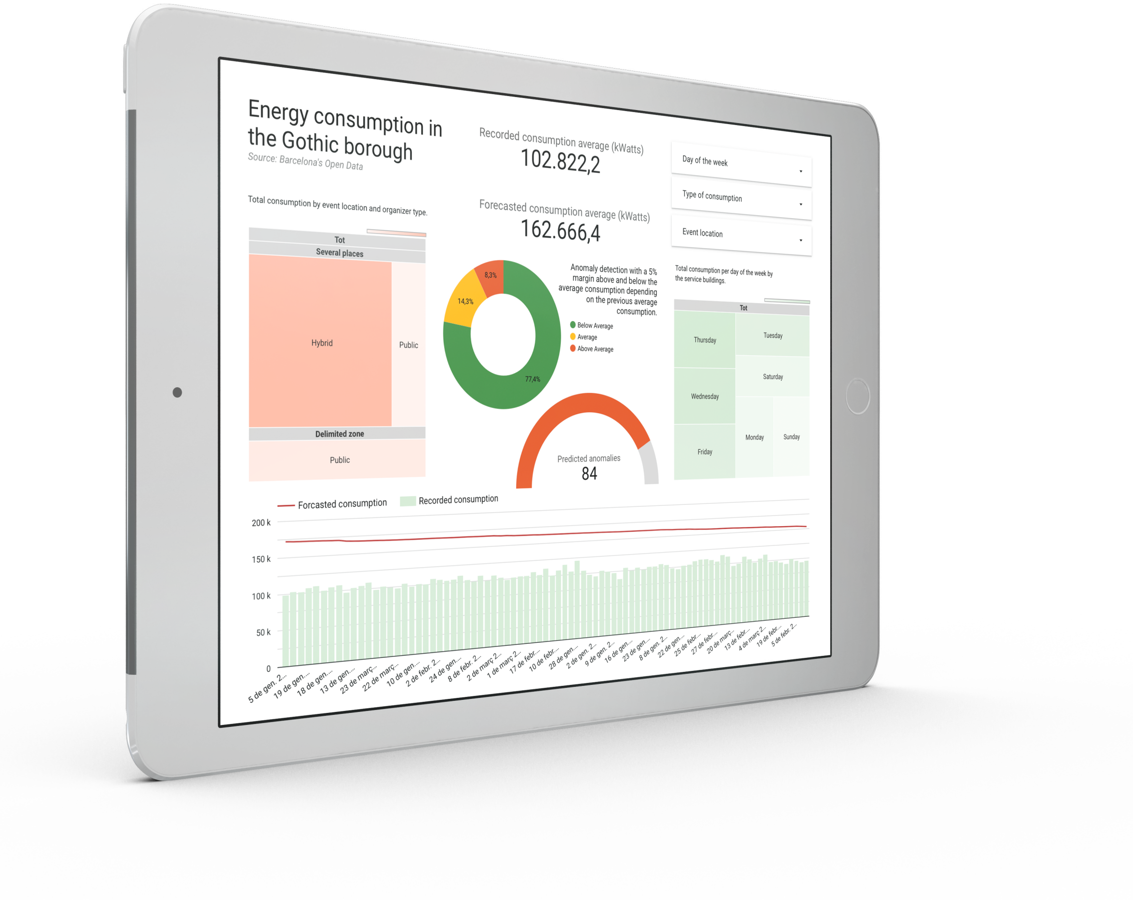


Figure 6.8: Dashboard for energy consumption in service buildings in the Gothic borough at night-time.

Chapter 7
Conclusions and Future work

To answer the initial research questions,

- *Which area of a city has been studied the most, and which areas are in need of development?*

The city area studied the most is **transportation** which follows the trend within the scope of anomaly detection literature. It was closely followed by cybersecurity and energy. The areas that need development are the environment, people's mobility, and structures. There were no studies related to the living and security of the people.

- *Which predictive models are being used on anomaly detection in smart cities? Are those models using supervised or unsupervised ML techniques?*

The predictive models used the most to detect anomalies in smart cities are **classification**, which are **supervised ML techniques** such as RF, XGBoost, SVM, and KNN. However, some studies performed unsupervised clustering using K-Means and DBSCAN, depending on the anomaly type.

- *Which databases of bibliographical references has the most resources?*

The bibliographical reference database with the broader scope of published open access papers is **Science Direct**. *IEEE* has a smaller section of conference papers instead of journals.

Despite referencing key ideas, several disqualified papers on the bibliographical reference databases did so because they correspond to a more theoretical than practical area. They emphasized more on potential solutions and upcoming difficulties. They discuss the possibilities in the area, notably the health sector and cybersecurity, but leave enormous room for future research. They have been used as case studies thus far, but there have yet to be any outcomes in anomaly detection.

Throughout the development of the systematic literature review, I found limitations in the discussion of some topics, as they were mentioned in a latent way and just hinted at the subject. As a result, this opens an door for future research on data governance and the people's security in smart cities. In the future, the systematic review could be carried out using NLP (Natural Language

Processing) tools to be able to address more publications. In this way, it would be more feasible to use search strategies that include more than 2.000 papers.

Papers are stuck in the analysis phase of the data collection and what happens after the models have predicted the anomalies is absent. They also do not present a data mining/acquisition approach; their smart data models are juxtaposed to data architecture, blending it into one topic.

The experimentation utilized 130.872 records, as there were initially more than one million records. For future work, more robust performance with more than 2GB of memory capacity is needed for the models to be trained in all the types of buildings in the city, in all boroughs, and at all times of the day, not only in the overnight period (from 18h to 5h), selected to lower the dataset size.

The operations part of ML and how these types of models can be brought to production can also be addressed.

Bibliography

[Abraham 2019] R Abraham, J Schneider and J Vom Brocke. *Data governance: A conceptual framework, structured review, and research agenda*. International journal of information management, vol. 49, pages 424–438, 2019. Available at https://doi.org/10.1016/j.ijinfomgt.2019.07.008. (Cited on page 3.)

[Abu-Alhaija 2022] M Abu-Alhaija and N Turab. *Automated Learning of ECG Streaming Data Through Machine Learning Internet of Things.* Intelligent Automation & Soft Computing, vol. 32, no. 1, 2022. Available at https://doi.org/10.32604/iasc.2022.021426. (Cited on pages 16, 37 and 40.)

[Aeronautics 1966] United States. National Aeronautics and Space Administration NASA. Guide to the Subject Indexes for Scientific and Technical Aerospace Reports. 1966. Available at https://www.google.es/books/edition/Guide_to_the_Subject_Indexes_for_Scienti/wkoOwReWX6oC?gbpv=1. (Cited on page 13.)

[Agency 023] Defense Advanced Research Projects Agency. *Explainable Artificial Intelligence (XAI)*, (Accessed: July 2023). Available at http://www.darpa.mil/program/explainable-artificial-intelligence. (Cited on page 14.)

[Al-Jarrah 2018] O Al-Jarrah, Y Al-Hammdi, P Yoo, S Muhaidat and M Al-Qutayri. *Semi-supervised multi-layered clustering model for intrusion detection*. Digital Communications and Networks, vol. 4, no. 4, pages 277–286, 2018. Available at https://doi.org/10.1016/j.dcan.2017.09.009. (Cited on pages 15, 20, 37 and 38.)

[Algani 2022] Y Algani, A Vinodhini, R Isabels, C Kaur, M Treve, K Bala, S Balaji and U Devi. *Analyze the anomalous behavior of wireless networking using the big data analytics*. Measurement: Sensors, vol. 23, 2022. Available at https://doi.org/10.1016/j.measen.2022.100407. (Cited on pages 15, 22, 23, 26, 27, 37 and 40.)

[Ali 2020] U Ali, M Shamsi, M Bohacek, C Hoare, K Purcell, E Mangina and J O'Donnell. *A data-driven approach to optimize urban scale energy retrofit decisions for residential buildings*. Applied Energy, vol. 267, 2020. Available at https://doi.org/10.1016/j.apenergy.2020.114861. (Cited on pages 16, 18, 22, 24, 26, 27, 28, 29, 32 and 45.)

[Allam 2019] Z Allam and Z Dhunny. *On big data, artificial intelligence and smart cities*. Cities, vol. 89, pages 80–91, 2019. Available at https://doi.org/10.1016/j.cities.2019.01.032. (Cited on page 9.)

[Alsalemi 2023] A Alsalemi, A Amira, H Malekmohamadi and K Diao. *A Modular Recommender System for Domestic Energy Efficiency*. Environmental Challenges, 2023. Available at https://doi.org/10.1016/j.envc.2023.100741. (Cited on pages 16, 18, 22, 23, 26, 27, 29, 36 and 45.)

[Amin 1995] S.M. Amin, A. García-Ortiz and J.R. Wootton. *Network, control, communication and computing technologies for intelligent transportation systems overview of the special issue*. Mathematical and Computer Modelling, vol. 22, no. 4, pages 1–10, 1995. Available at https://doi.org/10.1016/0895-7177(95)00126-M. (Cited on page 13.)

[Anderson 1983] J.R. Anderson, R.S. Michalski, J.G. Carbonell and T.M. Mitchell. Machine learning: An artificial intelligence approach (volume i). Machine Learning. Elsevier Science, 1983. Available at https://books.google.es/books?id=TWzuUd5gsnkC. (Cited on page 13.)

[Andrienko 2020] G Andrienko, N Andrienko, S Drucker, JD Fekete, D Fisher, S Idreos, T Kraska, G Li, KL Ma, J Mackinlay*et al. Big data visualization and analytics: Future research challenges and emerging applications*. In BigVis 2020-3rd International Workshop on Big Data Visual Exploration and Analytics, 2020. Available at https://inria.hal.science/hal-02568845. (Cited on page 3.)

[Bachechi 2022] C Bachechi, L Po and F Rollo. *Big data analytics and visualization in traffic monitoring*. Big Data Research, vol. 27, 2022. Available at https://doi.org/10.1016/j.bdr.2021.100292. (Cited on pages 16, 18, 20, 21, 22, 23, 26, 27, 42 and 44.)

[Bangui 2021] H Bangui, M Ge and B Buhnova. *A hybrid data-driven model for intrusion detection in VANET*. Procedia Computer Science, vol. 184, pages 516–523, 2021. Available at https://doi.org/10.1016/j.procs.2021.03.065. (Cited on pages 15, 20, 21, 22, 23, 37 and 39.)

[Bawaneh 2019] M Bawaneh and V Simon. *Anomaly Detection in Smart City Traffic Based on Time Series Analysis*. In 2019 International Conference on Software, Telecommunications and Computer Networks (SoftCOM), pages 1–6, 2019. Available at https://doi.org/10.23919/SOFTCOM.2019.8903822. (Cited on pages 16, 20, 21, 26, 27, 42 and 44.)

[Belhadi 2020] A Belhadi, Y Djenouri, K Nørvåg, H Ramampiaro, F Masseglia and J Lin. *Space–time series clustering: Algorithms, taxonomy, and case study on urban smart cities*. Engineering Applications of Artificial Intelligence, vol. 95, 2020. Available at https://doi.org/10.1016/j.engappai.2020.103857. (Cited on pages 16, 17, 18, 20, 21, 37 and 39.)

[Borrego 2014] M Borrego, M Foster and J Froyd. *Systematic Literature Reviews in Engineering Education and Other Developing Interdisciplinary Fields*. Journal of Engineering Education, vol. 103, 1, 2014. Available at https://doi.org/10.1002/jee.20038. (Cited on page 10.)

[Bukhari 2023] O Bukhari, P Agarwal, D Koundal and S Zafar. *Anomaly detection using ensemble techniques for boosting the security of intrusion detection system*. Procedia Computer Science, vol. 218, pages 1003–1013, 2023. Available at https://doi.org/10.1016/j.procs.2023.01.080. (Cited on pages 15, 22, 24, 26, 27, 29 and 36.)

[Campos-Asensio 2018] C Campos-Asensio. *How to develop a bibliographic search strategy*. Biblioteca del Hospital Universitario de Getafe, Getafe, Madrid, Spain, vol. 29, 4, 2018. Available at https://doi.org/10.1016/j.enfie.2018.09.001. (Cited on pages 7, 10 and 11.)

[Caragliu 2011] A Caragliu, C Del Bo and P Nijkamp. *Smart Cities in Europe*. Journal of Urban Technology, vol. 18, 2, 2011. Available at http://dx.doi.org/10.1080/10630732.2011.601117. (Cited on page 9.)

[Carbone 2017] F Carbone, A Cenedese and C Pizzi. *Consensus-based anomaly detection for efficient heating management*. In 2017 IEEE SmartWorld, Ubiquitous Intelligence Computing, Advanced Trusted Computed, Scalable Computing Communications, Cloud Big Data Computing, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI), pages 1–7, 2017. Available at https://doi.org/10.1109/UIC-ATC.2017.8397585. (Cited on pages 15, 18, 19, 26, 27, 42, 43 and 53.)

[Cerquitelli 2017] T Cerquitelli. *Predicting large scale fine grain energy consumption*. Energy Procedia, vol. 111, pages 1079–1088, 2017. Available at https://doi.org/10.1016/j.egypro.2017.03.271. (Cited on pages vii, 16, 18, 22, 23, 24, 26, 27, 29, 31, 45 and 54.)

[de Hond 2022] A de Hond, A Leeuwenberg, L Hooft, I Kant, S Nijman, H van Os, J Aardoom, T Debray, E Schuit, M van Smeden*et al.* *Guidelines and quality criteria for artificial intelligence-based prediction models in healthcare: a scoping review*. NPJ digital medicine, vol. 5, no. 1, page 2, 2022.

Available at https://doi.org/10.1038/s41746-021-00549-7. (Cited on page 4.)

[Dictionary 023] Oxford Advanced Learner's Dictionary. *Anomaly*, (Accessed: July 2023). Available at https://www.oxfordlearnersdictionaries.com/definition/english/anomaly. (Cited on page 13.)

[Du 2019] Y Du, C Wang, H Li, J Song and B Li. *Clustering heat users based on consumption data*. Energy Procedia, vol. 158, pages 3196–3201, 2019. Available at https://doi.org/10.1016/j.egypro.2019.01.1010. (Cited on pages vii, 16, 18, 26, 27, 37, 39, 45 and 48.)

[El-Alfy 2014] M El-Alfy and F Al-Obeidat. *A multicriterion fuzzy classification method with greedy attribute selection for anomaly-based intrusion detection*. Procedia Computer Science, vol. 34, pages 55–62, 2014. Available at https://doi.org/10.1016/j.procs.2014.07.037. (Cited on pages 14, 15, 29 and 31.)

[Embarak 2021] O Embarak. *A new paradigm through machine learning: A learning maximization approach for sustainable education*. Procedia Computer Science, vol. 191, pages 445–450, 2021. Available at https://doi.org/10.1016/j.procs.2021.07.055. (Cited on pages 16, 17, 29 and 33.)

[Fonseca 2017] J A Fonseca, C Miller and A Schlueter. *Unsupervised load shape clustering for urban building performance assessment*. Energy Procedia, vol. 122, pages 229–234, 2017. Available at https://doi.org/10.1016/j.egypro.2017.07.350. (Cited on pages 16, 18, 42, 43, 45 and 53.)

[García-Ortiz 1995] A. García-Ortiz, S.M. Amin and J.R. Wootton. *Intelligent transportation systems—Enabling technologies*. Mathematical and Computer Modelling, vol. 22, no. 4, pages 11–81, 1995. Available at https://doi.org/10.1016/0895-7177(95)00127-N. (Cited on page 13.)

[Gaur 2018] A Gaur and M Kumar. *A systematic approach to conducting review studies: An assessment of content analysis in 25 years of IB research*. Journal of World Business, vol. 53, no. 2, pages 280–289, 2018. Available at https://doi.org/10.1016/j.jwb.2017.11.003. (Cited on page 11.)

[Gerrish 2017] T Gerrish, K Ruikar, M Cook, M Johnson and M Phillip. *Analysis of basic building performance data for identification of performance issues*. Facilities, vol. 35, no. 13/14, pages 801–817, 2017. Available at https://doi.org/10.1108/F-01-2016-0003. (Cited on pages 15, 17, 18, 42, 44 and 45.)

[Gomari 2021] S Gomari, C Knoth and C Antoniou. *Cluster analysis of parking behaviour: A case study in Munich*. Transportation Research Procedia, vol. 52, pages 485–492, 2021. Available at https://doi.org/10.1016/j.trpro.2021.01.057. (Cited on pages 16, 20, 21, 37 and 40.)

[Gunning 2021] D Gunning, E Vorm, Y Wang and M Turek. *DARPA's explainable AI (XAI) program: A retrospective*. Authorea Preprints, 2021. Available at https://doi.org/10.1002/ail2.61. (Cited on page 14.)

[Hall 023a] Barcelona's City Hall. *Open Data BCN - Data on city festivals in the city of Barcelona*, (Accessed: August 2023). Available at https://opendata-ajuntament.barcelona.cat/data/en/dataset/dades-festes-ciutat. (Cited on pages ix and 49.)

[Hall 023b] Barcelona's City Hall. *Open Data BCN - Electricity consumption by postal code, economic sector and time interval in the city of Barcelona*, (Accessed: May 2023). Available at https://opendata-ajuntament.barcelona.cat/data/en/dataset/consum-electricitat-bcn. (Cited on pages ix, 2, 46, 47 and 53.)

[Hangan 2022] A Hangan, C Chiru, D Arsene, Z Czako, D Lisman, M Mocanu, B Pahontu, A Predescu and G Sebestyen. *Advanced techniques for monitoring and management of urban water infrastructures—an overview*. Water, vol. 14, no. 14, 2022. Available at https://doi.org/10.3390/w14142174. (Cited on pages 16, 18, 19, 22, 24, 26, 27, 29, 34, 37 and 40.)

[Hartigan 1979] J Hartigan and M Wong. *Algorithm AS 136: A k-means clustering algorithm*. Journal of the royal statistical society. series c (applied statistics), vol. 28, no. 1, pages 100–108, 1979. Available at https://doi.org/10.2307/2346830. (Cited on page 3.)

[Himeur 2020] Y Himeur, A Alsalemi, A Al-Kababji, F Bensaali and A Amira. *Data fusion strategies for energy efficiency in buildings: Overview, challenges and novel orientations*. Information Fusion, vol. 64, pages 99–120, 2020. Available at https://doi.org/10.1016/j.inffus.2020.07.003. (Cited on pages 16, 18, 22, 23, 29, 32, 45 and 54.)

[Karanfilovska 2022] M Karanfilovska, T Kochovska, Z Todorov, A Cholakoska, G Jakimovski and D Efnusheva. *Analysis and modelling of a ML-based NIDS for IoT networks*. Procedia Computer Science, vol. 204, pages 187–195, 2022. Available at https://doi.org/10.1016/j.procs.2022.08.023. (Cited on pages 16, 29, 34, 37, 40 and 54.)

[Kavya 2021] R Kavya, Cr Jabez, P Subhrakanta and Y Bakthasingh. *Machine Learning and XAI approaches for Allergy Diagnosis*. Biomedical Signal Processing and Control, vol. 69, 2021. Available at https://doi.org/10.1016/j.bspc.2021.102681. (Cited on page 14.)

[Killeen 2019] P Killeen, B Ding, I Kiringa and T Yeap. *IoT-based predictive maintenance for fleet management*. Procedia Computer Science, vol. 151, pages 607–613, 2019. Available at https://doi.org/10.1016/j.procs.2019.04.184. (Cited on pages vii, 16, 18, 20, 21, 22, 25, 29 and 32.)

[Kulkarni 2016] P Kulkarni and T Farnham. *Smart City Wireless Connectivity Considerations and Cost Analysis: Lessons Learnt From Smart Water Case Studies*. IEEE Access, vol. 4, pages 660–672, 2016. Available at https://doi.org/10.1109/ACCESS.2016.2525041. (Cited on page 4.)

[Kyriakou 2021] C Kyriakou, S Christodoulou and L Dimitriou. *Do vehicles sense, detect and locate speed bumps?* Transportation Research Procedia, vol. 52, pages 203–210, 2021. Available at https://doi.org/10.1016/j.trpro.2021.01.023. (Cited on pages 16, 18, 20, 21, 22, 23, 29 and 33.)

[Laaksonen 1996] J Laaksonen and E Oja. *Classification with learning k-nearest neighbors*. In Proceedings of international conference on neural networks (ICNN'96), volume 3, pages 1480–1483. IEEE, 1996. Available at 10.1109/ICNN.1996.549118. (Cited on page 3.)

[Lamy 2019] JB Lamy, B Sekar, G Guezennec, J Bouaud and B Séroussi. *Explainable artificial intelligence for breast cancer: A visual case-based reasoning approach*. Artificial Intelligence in Medicine, vol. 94, pages 42–53, 2019. Available at https://doi.org/10.1016/j.artmed.2019.01.001. (Cited on page 14.)

[Lbazri 2020] S Lbazri, H Jihal, M Azouaziet *al. Predict France trains delays using visualization and machine learning techniques*. Procedia Computer Science, vol. 175, pages 700–705, 2020. Available at https://doi.org/10.1016/j.procs.2020.07.103. (Cited on pages 16, 30 and 32.)

[Leiria 2021] D Leiria, H Johra, A Marszal-Pomianowska, M Pomianowski and P Heiselberg. *Using data from smart energy meters to gain knowledge about households connected to the district heating network: A Danish case*. Smart Energy, vol. 3, 2021. Available at https://doi.org/10.1016/j.segy.2021.100035. (Cited on pages 16, 18, 22, 23, 28, 29, 33, 45 and 53.)

[Liu 2018] X Liu and P Nielsen. *Scalable prediction-based online anomaly detection for smart meter data*. Information Systems, vol. 77, pages 34–47,

2018. Available at https://doi.org/10.1016/j.is.2018.05.007. (Cited on pages vii, 16, 18, 22, 23, 24, 25, 42, 44, 45 and 48.)

[Masino 2017] J Masino, J Thumm, M Frey and F Gauterin. *Learning from the crowd: Road infrastructure monitoring system*. Journal of Traffic and Transportation Engineering (English Edition), vol. 4, no. 5, pages 451–463, 2017. Available at https://doi.org/10.1016/j.jtte.2017.06.003. (Cited on pages 16, 20, 29 and 31.)

[Mohamudally 2018] N Mohamudally and M Peermamode-Mohaboob. *Building an anomaly detection engine (ADE) for Iot smart applications*. Procedia computer science, vol. 134, pages 10–17, 2018. Available at https://doi.org/10.1016/j.procs.2018.07.138. (Cited on pages 15, 17, 18, 19, 22, 23, 37 and 38.)

[Mondal 2020] M Mondal and Z Rehena. *Road traffic outlier detection technique based on linear regression*. Procedia Computer Science, vol. 171, pages 2547–2555, 2020. Available at https://doi.org/10.1016/j.procs.2020.04.276. (Cited on pages 16, 20, 21, 28, 30 and 33.)

[Moon 2022] J Moon, S Rho and S Baik. *Toward explainable electrical load forecasting of buildings: A comparative study of tree-based ensemble methods with Shapley values*. Sustainable Energy Technologies and Assessments, vol. 54, 2022. Available at https://doi.org/10.1016/j.seta.2022.102888. (Cited on pages vii, 14, 16, 18, 26, 27, 29, 35, 45 and 49.)

[Naeem 2022] M Naeem, T Jamal, J Diaz-Martinez, S Butt, N Montesano, M Tariq, E De-la Hoz-Franco and E De-La-Hoz-Valdiris. *Trends and future perspective challenges in big data*. In Advances in Intelligent Data Analysis and Applications: Proceeding of the Sixth Euro-China Conference on Intelligent Data Analysis and Applications, 15–18 October 2019, Arad, Romania, pages 309–325. Springer, 2022. Available at https://doi.org/10.1007/978-981-16-5036-9_30. (Cited on page 3.)

[Nugraha 2021] A Nugraha, S Supangkat, I Nugraha, H Trimadi, A Purwadinata, Sumarni and S Sundari. *Detection of Railroad Anomalies using Machine Learning Approach*. In 2021 International Conference on ICT for Smart Society (ICISS), pages 1–6, 2021. Available at https://doi.org/10.1109/ICISS53185.2021.9533226. (Cited on pages 16, 20, 21, 26, 27, 30 and 34.)

[of America 1942] Scientific Research Society of America. American Scientist. 1942. Available at https://books.google.es/books?id=Y_lUAAAAMAAJ. (Cited on page 13.)

[Omar 2013] S Omar, A Ngadi and H Jebur. *Machine learning techniques for anomaly detection: an overview*. International Journal of Computer Applications, vol. 79, no. 2, 2013. Available at https://doi.org/10.5120/13715-1478. (Cited on page 3.)

[Page 2021] M Page, D Moher, P Bossuyt, I Boutron, T Hoffmann, C Mulrow, L Shamseer, J Tetzlaff, E Akl, S Brennan, R Chou, J Glanville, J Grimshaw, A Hróbjartsson, M Lalu, T Li, E Loder, E Mayo-Wilson, S McDonald, L McGuinness, L Stewart, J Thomas, A Tricco, V Welch, P Whiting and J McKenzie1. *PRISMA 2020 explanation and elaboration: updated guidance and exemplars for reporting systematic reviews*. The BMJ, vol. 372, 2021. Available at http://dx.doi.org/10.1136/bmj.n160. (Cited on pages 1, 4 and 7.)

[Protic 2022] D Protic, L Gaur, M Stankovic and A Rahman. *Cybersecurity in smart cities: Detection of opposing decisions on anomalies in the computer network behavior*. Electronics, vol. 11, no. 22, 2022. Available at https://doi.org/10.3390/electronics11223718. (Cited on pages 15, 26, 27, 29 and 35.)

[Sabol 2020] P Sabol, P Sinčák, P Hartono, P Kočan, Z Benetinová, A Blichárová, L Verbóová, E Štammová, A Sabolová-Fabianová and A Jašková. *Explainable classifier for improving the accountability in decision-making for colorectal cancer diagnosis from histopathological images*. vol. 109, 2020. Available at https://doi.org/10.1016/j.jbi.2020.103523. (Cited on page 14.)

[Saheed 2022] Y Saheed, A Abiodun, S Misra, M Holone and R Colomo-Palacios. *A machine learning-based intrusion detection for detecting internet of things network attacks*. Alexandria Engineering Journal, vol. 61, no. 12, pages 9395–9409, 2022. Available at https://doi.org/10.1016/j.aej.2022.02.063. (Cited on pages 15, 22, 24, 26, 27, 29, 35 and 54.)

[Saranya 2020] T Saranya, S Sridevi, C Deisy, T Chung and A Khan. *Performance analysis of machine learning algorithms in intrusion detection system: A review*. Procedia Computer Science, vol. 171, pages 1251–1260, 2020. Available at https://doi.org/10.1016/j.procs.2020.04.133. (Cited on pages 15, 29 and 33.)

[Sarker 2022] I Sarker. *Smart City Data Science: Towards data-driven smart cities with open research issues*. Internet of Things, vol. 19, 2022. Available at https://doi.org/10.1016/j.iot.2022.100528. (Cited on pages 9, 16 and 21.)

[Singh 2022] T Singh, A Solanki, S Sharma, A Nayyar and A Paul. *A Decade Review on Smart Cities: Paradigms, Challenges and Opportunities*. IEEE Access, 2022. Available at https://doi.org/10.1109/ACCESS.2022.3184710. (Cited on page 9.)

[Snyder 2019] H Snyder. *Literature review as a research methodology: An overview and guidelines*. Journal of Business Research, vol. 104, 2019. Available at https://doi.org/10.1016/j.jbusres.2019.07.039. (Cited on page 2.)

[Society 1958] American Nuclear Society. Transactions of the American Nuclear Society. 1958. Available at https://books.google.es/books?id=e49VAAAAMAAJ. (Cited on page 13.)

[Stemler 2000] S Stemler. *An overview of content analysis*. Practical assessment, research, and evaluation, vol. 7, no. 1, page 17, 2000. Available at https://doi.org/10.7275/z6fm-2e34. (Cited on page 12.)

[Tableau 023] Tableau. *Time Series Forecasting: Definition, Applications, and Examples*, (Accessed: July 2023). Available at https://www.tableau.com/learn/articles/time-series-forecasting. (Cited on page 3.)

[Talukder 2019] S Talukder, L Shen, F Talukder and Y Bao. *Determinants of user acceptance and use of open government data (OGD): An empirical investigation in Bangladesh*. Technology in Society, vol. 56, pages 147–156, 2019. Available at https://doi.org/10.1016/j.techsoc.2018.09.013. (Cited on page 4.)

[University 023] Lund University. *Bibliographic databases*, (Accessed: July 2023). Available at https://libguides.lub.lu.se/c.php?g=677619&p=4829257. (Cited on page 3.)

[Vidović 2022] K Vidović, P Čolić, S Vojvodić and A Blavicki. *Methodology for public transport mode detection using telecom big data sets: case study in Croatia*. Transportation Research Procedia, vol. 64, pages 76–83, 2022. Available at https://doi.org/10.1016/j.trpro.2022.09.010. (Cited on pages 16, 20, 29 and 35.)

[Vijai 2016] P Vijai and B Sivakumar. *Design of IoT systems and analytics in the context of smart city initiatives in India*. Procedia Computer Science, vol. 92, pages 583–588, 2016. Available at https://doi.org/10.1016/j.procs.2016.07.386. (Cited on pages vii, 16, 18, 19, 22, 23, 42 and 43.)

[Wang 2017] Y Wang, X Fan, X Liu, C Zheng, L Chen, C Wang and J Li. *Unlicensed Taxis Detection Service Based on Large-Scale Vehicles Mobility Data*. In 2017 IEEE International Conference on Web Services (ICWS), pages 857–861, 2017. Available at https://doi.org/10.1109/ICWS.2017.106. (Cited on pages 16, 20, 21, 22, 23, 26, 27, 29, 31 and 54.)

[Wang 2021] J Wang and X Mo. *A CAN Bus Anomaly Detection Based on FLXGBoost Algorithm*. In 2021 IEEE 23rd Int Conf on High Performance Computing Communications; 7th Int Conf on Data Science Systems; 19th Int Conf on Smart City; 7th Int Conf on Dependability in Sensor, Cloud Big Data Systems Application (HPCC/DSS/SmartCity/DependSys), pages 1558–1564, 2021. Available at https://doi.org/10.1109/HPCC-DSS-SmartCity-DependSys53884.2021.00231. (Cited on pages 16, 18, 20, 21, 26, 27, 30, 34 and 54.)

[Wazid 2022] M Wazid, A Das, V Chamola and Y Park. *Uniting cyber security and machine learning: Advantages, challenges and future research*. ICT Express, vol. 8, no. 3, pages 313–321, 2022. Available at https://doi.org/10.1016/j.icte.2022.04.007. (Cited on page 3.)

[Wong 2013] G Wong, T Greenhalgh, G Westhorp, J Buckingham and R Pawson. *RAMESES publication standards: Meta-narrative reviews*. BMC Medicine, vol. 11, 20, 2013. Available at https://doi.org/10.1186/1741-7015-11-20. (Cited on page 2.)

[Wu 2023] J Wu, B Du, Z Gong, Q Wu, J Shen, L Zhou and C Cai. *A GTFS data acquisition and processing framework and its application to train delay prediction*. International Journal of Transportation Science and Technology, vol. 12, no. 1, pages 201–216, 2023. Available at https://doi.org/10.1016/j.ijtst.2022.01.005. (Cited on pages 16, 20, 21, 23, 24, 26, 27, 42 and 44.)

[Xu 2019] Z Xu, D Kakde and A Chaudhuri. *Automatic Hyperparameter Tuning Method for Local Outlier Factor, with Applications to Anomaly Detection*. In 2019 IEEE International Conference on Big Data (Big Data), pages 4201–4207, 2019. Available at https://doi.org/10.1109/BigData47090.2019.9006151. (Cited on pages 16, 29 and 32.)

[Yadav 2023] R Yadav, I Sreedevi and D Gupta. *Augmentation in performance and security of WSNs for IoT applications using feature selection and classification techniques*. Alexandria Engineering Journal, vol. 65, pages 461–473, 2023. Available at https://doi.org/10.1016/j.aej.2022.10.033. (Cited on pages 15, 22, 24, 26, 27, 29 and 36.)

[Zantalis 2019] F Zantalis, G Koulouras, S Karabetsos and D Kandris. *A review of machine learning and IoT in smart transportation*. Future Internet, vol. 11, no. 4, 2019. Available at https://doi.org/10.3390/fi11040094. (Cited on pages 16, 20, 21, 30 and 32.)

[Zhao 2017] J Zhao, Q Qu, F Zhang, C Xu and S Liu. *Spatio-Temporal Analysis of Passenger Travel Patterns in Massive Smart Card Data*. IEEE Transactions on Intelligent Transportation Systems, vol. 18, no. 11, pages 3135–3146, 2017. Available at https://doi.org/10.1109/TITS.2017.2679179. (Cited on pages 16, 20, 21, 26, 27, 37 and 38.)

[Zhu 2020] Q Zhu and L Sun. *Big Data Driven Anomaly Detection for Cellular Networks*. IEEE Access, vol. 8, pages 31398–31408, 2020. Available at https://doi.org/10.1109/ACCESS.2020.2973214. (Cited on pages 16, 20, 21, 26, 28, 37 and 39.)

[Zinno 2022] R Zinno, S Haghshenas, G Guido and A VItale. *Artificial Intelligence and Structural Health Monitoring of Bridges: A Review of the State-of-the-Art*. IEEE Access, vol. 10, pages 88058–88078, 2022. Available at https://doi.org/10.1109/ACCESS.2022.3199443. (Cited on pages 16, 22, 23, 26, 27, 29, 36, 37 and 41.)