

School of Electrical Engineering, Computing and Mathematical Sciences

**Machine Learning as a Service (MLaaS) Selection for IoT
Environments**

Keyaben Mukeshbhai Patel

0000-0003-3987-9828

This thesis is presented for the Degree of

Master of Philosophy

Of

Curtin University

October 2024

Declaration

To the best of my knowledge and belief, this thesis contains no material previously published by any other person except where due acknowledgment has been made.

This thesis contains no material which has been accepted for the award of any other degree or diploma in any university.

Keya Patel

School of Electrical, Computing,
and Mathematical Sciences

Curtin University

Signature:

07 October 2024

Acknowledgements

I would like to express my sincere gratitude to my supervisor, Dr. Sajib Mistry, for his invaluable guidance, feedback, and support throughout my research. His profound knowledge and experience were instrumental in the shaping of this thesis. I am especially grateful for his willingness to offer unconditional meetings whenever I needed guidance. His impact on refining my critical thinking, writing, and presentation skills has been exceptional. Dr. Sajib Mistry is among the most inspiring mentors I have ever encountered.

I am also extremely thankful to my co-supervisor, A/Prof. Aneesh Krishna, for his insightful suggestions and motivation in helping me enter the research field. His mentorship provided me with the foundational knowledge and confidence needed to navigate the complexities of academic research.

I am deeply grateful to Co-Author, Deepak for his significant contributions in helping me develop the preliminary experiments. His expertise, commitment, and collaborative efforts were instrumental in shaping the direction of this research. I deeply appreciate his support and contributions throughout this process.

I must acknowledge my husband and greatest supporter, Bhavin, whose unwavering encouragement, love, and understanding empowered me to pursue my dream of completing my MPhil. I want to thank my most diligent daughter, Khushi and caring son, Happy; your smiles, laughter, and endless energy have been a constant reminder of the joy and purpose in life. I am very thankful for your patience and for always brightening my days. This thesis is much yours as it is mine. Furthermore, I am very grateful to my parents and parents-in-law, brother-in-law, and sister-in-law for their kind support.

This research is supported by an Australian Government Research Training Program (RTP) Scholarship. The partial project is supported by a grant from the Defence Science Centre under the Research Higher Degree Student Grant program.

Authorship Acknowledgment

The main results presented in this thesis are based on works that were published in conference proceedings or submitted during the author's MPhil study. They are listed as follows:

- **Patel, K., Mistry, S., Kanneganti, S. K. D., & Krishna, A. (2023).** Machine Learning as a Service (MLaaS) Selection with Incomplete QoS Information. <https://aisel.aisnet.org/acis2023/39/>
(ERA ranking: A, Acceptance rate: 24%)
- **Patel, K., Mistry, S., Kanneganti, S. K. D., & Krishna, A. (2024).** Context-Aware Selection of Machine Learning as a Service (MLaaS) in an IoT Environment. Web Information Systems Engineering (WISE). (Submitted).

The copyright information to reuse the published work has been provided in Appendix A.

Abstract

Machine Learning as a Service (MLaaS) holds significant importance in technology and business due to its pivotal role in democratising and simplifying the deployment of machine learning models. MLaaS refers to providing machine learning tools, infrastructure, and algorithms as cloud-based services, allowing users to build, train, deploy and manage machine learning models without needing to handle the underlying technical complexities. Selecting the right Machine Learning as a Service (MLaaS) provider for organisations requires a strategic approach considering the business's unique needs, user needs, and goals. However, the MLaaS selection process is complex due to the need for complete information on the quality of services and MLaaS latent features such as model accuracy, explainability and intrinsic biases. Also, integrating MLaaS into Internet of Things (IoT) environments, the process of selecting the appropriate MLaaS in an IoT setting is complex due to the various contextual dimensions, such as user preferences, locations, IoT device capabilities, and application requirements. This research aims to develop an MLaaS service selection framework where MLaaS providers reveal limited QoS information about their services and user changes in contextual information. First, we propose a novel MLaaS Selection Framework (MSF) using incomplete QoS information available through service advertisement. We develop the knowledge-based bias detection and Explainable (B-XAI) framework to discover MLaaS latent features. The proposed framework builds a complete QoS profile of the providers using MLaaS advertisements, other user experiences, and short-term trial experiences. We apply the nearest neighbour algorithm to select the optimal MLaaS providers based on user preference models. Second, we propose a novel framework for context-aware selection of MLaaS in IoT settings, aimed at optimising the interaction between IoT users' activities and machine learning services to develop a dynamic selection process. By employing context-aware algorithms, our approach seeks to enhance the efficiency, accuracy, and responsiveness of IoT systems. We propose a context change analysis algorithm based on support vector machines (SVM). We develop a

contextual bandit algorithm and skyline queries to achieve optimal mapping between abstract MLaaS services and concrete MLaaS services for quality of service (QoS) attributes. Experiments with real-world datasets demonstrated the effectiveness of the proposed approaches in enhancing the efficiency, accuracy and responsiveness of IoT systems while facilitating informed MLaaS service selection.

Contents

Declaration	iii
Acknowledgments	v
Authorship Acknowledgment	vi
Abstract	vii
Contents	ix
List of Figures	xii
List of Tables	xii
Chapter 1 Introduction	1
1.1 Key Research Challenges.....	4
1.2 Research Contributions.....	5
1.2.1 Machine Learning as a Service (MLaaS) Selection.....	6
1.2.2 Machine Learning as a Service (MLaaS) Selection with Incomplete QoS Information.....	6
1.2.2 Context-Aware Selection of Machine Learning as a Service (MLaaS) in IoT Environments.....	6
1.3 Outline of the Thesis Chapters.....	6
Chapter 2 Related Works	8
2.1 Types of Machine Learning as a Service (MLaaS).....	10
2.1.1 Inference-based MLaaS.....	10
2.1.2 Platform-based MLaaS.....	10
2.1.3 Data-based MLaaS.....	11
2.2 A Review of Service Composition across Different Domains.....	11
2.3 Machine Learning as a Service (MLaaS) Selection Approaches.....	12
2.4 Service Selection in the context of Web, Cloud and Edge computing.....	13
2.5 Selection of Context-Aware Service Approaches in the context of Web, Cloud and IoT.....	15
2.6 Understanding the Implications of MLaaS.....	18

2.6.1 Short-Term Implications.....	18
2.6.2 Long-Term Implications.....	19
2.6.3 Economic Implications.....	20
2.7 Conclusion.....	20
Chapter 3 Machine Learning as a Service (MLaaS) Selection with Incomplete QoS Information	21
3.1 Introduction.....	21
3.2 Motivation Scenario.....	24
3.3 Building MLaaS Service Selection Framework.....	25
3.3.1 MLaaS Advertisements.....	27
3.3.2 Performing own trials.....	27
3.3.3 Measure the QoS information from the other user experience.....	28
3.3.4 Data Selection.....	29
3.3.5 Discovering Latent Feature – Bias.....	30
3.3.6 Discover latent feature- Explainability.....	32
3.3.7 MLaaS Service Selection.....	34
3.4 Experiments and Results.....	34
3.4.1 Experiment Setup.....	34
3.4.2 Evaluation.....	35
3.5 Conclusion.....	36
Chapter 4 Context-Aware Selection of Machine Learning as a Service (MLaaS) in IoT Environments	38
4.1 Introduction.....	38
4.2 Context-Aware MLaaS Selection Framework.....	41
4.2.1 Context Change Analysis.....	43
4.2.2 Mapping User Context to Abstract MLaaS Services.....	44
4.2.3 Mapping Abstract MLaaS Service to Concrete MLaaS Services based on QoS.....	45
4.3 Experiment and Results.....	48
4.3.1 Data Set Description.....	48
4.3.2 Baseline Approaches.....	49
4.3.2 Performance Analysis.....	50
4.4 Conclusion.....	54

Chapter 5 Conclusion and Future Works	55
5.1 Discussion.....	56
5.2 Limitations.....	56
5.3 Future Work.....	57
Appendices	
A. Copyright Information	59
B. Statement of Attribution	62
C. Research Output Authorship Attribution	63
Bibliography	66

List of Figures

1.1 Key Contributions.....	5
3.1 MLaaS Service Selection Framework.....	27
3.2 A B-XAI framework (a knowledge-based graph for Bias detection methods and Explainable service methods).....	32
3.3 Case 1- Limited QoS information, Case 2- collect data from different aspects (full knowledge of QoS data), and Case 3- Random data (b) Preference ranking for all 3 cases according to user preference (C) 3 cases with evaluation of Results.....	36
4.1 MLaaS Service Selection based on Single user’s context.....	39
4.2 Context-Aware MLaaS Selection Framework (CAMSF).....	42
4.3 The effectiveness and scalability of CAMSF (a) SVM context detection accuracy, (b) Computation Time Comparison, (c) Accuracy Comparison, and (d) F1 Score Comparison Across Models.....	53

List of Tables

2.1 A summary of service selection studies - strengths and weaknesses.....	17
4.1 Statistics of Dataset for Context-Aware MLaaS in IoT.....	48
4.2 Test cases of mapping MLaaS abstract and concrete service with QoS.....	49
4.3 Performance measurement result.....	51

CHAPTER 1

INTRODUCTION

Machine Learning as a Service (MLaaS) refers to a cloud-based service that delivers machine learning tools and infrastructure on a service basis [[Sahi, 2022](#)]. MLaaS has emerged as a transformative advancement in the cloud computing landscape, providing organisations with on-demand access to powerful machine learning tools without needing in-depth expertise and substantial investments in infrastructure [[Ribeiro et al., 2015](#); [Sahi, 2022](#)]. MLaaS offers pre-built models, intuitive APIs, development tools, and robust computational resources, enabling businesses to accelerate innovation while adapting to evolving users and organisation demands [[Sun et al., 2014](#)]. These platforms typically include features such as data visualisation, model training, deployment and monitoring, making it easier to integrate AI-driven solutions into various applications for industries such as healthcare, finance, education and retail [[Ribeiro et al., 2015](#)]. These services not only simplify the deployment of sophisticated solutions but also lower barriers, making cutting-edge machine-learning (ML) technology accessible to consumers and fostering innovation across various domains [[Ribeiro et al., 2015](#)]. Popular MLaaS providers include Google Cloud AI, Microsoft Azure, Amazon and IBM, offering various services to consumers and organisations.

Various industries are leveraging MLaaS to optimise their operations and deliver better outcomes. For instance, Health care providers are utilising MLaaS to analyse medical data to improve diagnostics and predict the likelihood of diseases, which, as a result, enables preventive care and improves patient health outcomes through early disease detection [[Lupo, 2016](#)]. Similarly, the financial sector relies heavily on MLaaS to identify fraud detection. These services analyse real-time transaction patterns by continuously learning from new data to detect suspicious activities and potential fraud [[Ribeiro et al., 2015](#)]. This approach protects financial institutes and customers from significant financial losses. By adopting the MLaaS, these organisations can drive innovation and autonomous growth in their respective fields. MLaaS offers several key benefits. First, MLaaS is cost-effective. MLaaS allows users to select services based on their specific demands, and users pay only for what they would like to use, which allows them to adjust services up or down as preferences change [[Zhang et](#)

[al., 2020](#)]. Second, MLaaS platforms provide services with the flexibility to adapt to changing user needs. It offers pre-built models, development tools, and robust computational resources, enabling businesses to accelerate innovation while adapting to evolving users and organisation demands [[Ribeiro et al., 2015](#)]. Third, MLaaS easily integrates with existing systems, allowing seamless adoption and enabling organisations to enhance operations without disruptions [[Philipp et al., 2020](#)]. MLaaS deliver low-cost, flexible and easily integrable solutions that allow businesses to scale services according to consumer demands.

The selection of MLaaS is a significant issue in several domains. This selection process involves identifying and selecting the most appropriate MLaaS provider from a range of functional services and non-functional services (which can be called Quality of Services (QoS)). *Function services* include data storage and virtual machines, while *Non-functional services* include response time, throughput, reliability, availability cost, latency, and usability. QoS attributes help a user to select high-performing services from a range of functionality-equivalent services [[Huang et al., 2018](#)]. For instance, while two MLaaS providers may offer comparable predictive analysis services in the healthcare domain, one might provide low biasness of MLaaS service with high explainability. This difference can make one provider more suitable in health where a clinical decision-making system is required. This means that the effective MLaaS service selection of QoS ensures that the chosen MLaaS provider not only meets the functional requirement but also performs better in terms of scalability and efficiency. This leads to greater user satisfaction.

However, most of the MLaaS service providers do not disclose much information about the QoS attributes of their services. The primary reasons for these behaviours are market competition, business confidentiality and potential conflict of interest [[Amazon Web Services, n.d.](#)]. Hence, selecting an MLaaS service that aligns well with an organisation's QoS requirements becomes challenging for two main reasons:

1) Incomplete or Insufficient Advertisements of MLaaS

MLaaS service providers typically advertise minimal QoS information in their advertisements and contain a limited number of QoS attributes. For instance, explainability, and biasness are unavailable in most advertisements [[Amazon Web Services, n.d.](#)]. The QoS information of the advertised attribute may not be helpful to a user for selection. For example, a user or organisation may want to know how the service's model explainability and bias detection quality, yet the advertised QoS information is hidden for an MLaaS service. Additionally, the

advertised QoS information may not provide a clear understanding of service performance as it often lacks detailed latent features of biasness and explainability. For instance, AWS Sagemaker does not reveal complete information about the provided service's model explainability and bias detection quality. MLaaS providers advertise either an average or maximum QoS performance of their services. For example, Google Cloud AI's Predictive Service allows users to obtain predictions and target values [[Google Cloud, n.d.](#)]. With ongoing global concerns about bias in technologies such as online recruiting apps and criminal justice algorithms, a study assessing MLaaS services for biasness is crucial [[Akter et al., 2022](#)]. Additionally, explainability is a criterion for users to understand how a model makes decisions. Therefore, relying only on MLaaS advertisement is insufficient to select an MLaaS service.

2) **Dynamic contextual changes to meet QoS requirements**

Dynamic contextual changes in IoT environments require continuous adaption to meet QoS requirements. Traditional MLaaS selection focuses on functional and non-functional properties without incorporating dynamic context awareness for the selection of MLaaS [[Matos, 2020](#); [Rhayem et al., 2021](#)]. For instance, pre-defined rule-based systems fail to provide long-term adaptability and are unsuitable for meeting users' evolving needs where complexity is in changing IoT contexts. These approaches are not applicable in the smart environment, for example, where a user's changing contexts, such as fall incidents, emotional distress, or security threats, demand timely transitions between relevant services. The capability for ongoing improvement and adjustment to shifting needs and contexts in the traditional service selection approaches may lack the degree of flexibility and responsiveness [[Matos, 2020](#)]. MLaaS service providers do not address the evolving QoS requirements for changing contexts. Therefore, the adaptability of MLaaS is important in selecting an MLaaS service for the long term in an IoT environment.

Existing research in the context of web and cloud service selection mainly focuses on short-term selection methods [[Kumar, Kumari & Kumar, 2021](#); [Wu et al., 2022](#)]; however, these practicalities are inapplicable to MLaaS selection due to the incompleteness of advertisement. Similarly, existing context-aware service selection approaches for Web, Cloud, and Edge computing typically consider short-term contextual information such as user preferences, service environment settings, and advertisements [[Rhayem et al., 2021](#); [Xu et al., 2016](#)]. These approaches increase the complexity of updating rules and limit their adaptability and

scalability in the diversity of contexts. **The objective of our research is to propose a novel framework to select the optimal MLaaS service according to the user’s preference for QoS attributes and the user’s context changes where MLaaS service providers reveal limited QoS information and functionally similar MLaaS services, respectively.**

1.1 Key Research Challenges

The most effective approach to address incomplete QoS information is to leverage the free trials that service providers offer [[Fattah, Bouguettaya, & Mistry, 2020](#)]. Most MLaaS service providers offer free trials and promote users to use the MLaaS service for a limited time. For instance, Microsoft Azure offers a 30-day free trial of MLaaS services to its customers. MLaaS users may get a trial experience before selecting a service. To the best of our knowledge, existent works do not explore the effective use of free trials for MLaaS service selection. Our goal is to utilise a free short-term trial to reveal hidden QoS information of MLaaS service selection. However, selecting an MLaaS service based on a free trial is challenging. The short-term free trial may not provide a QoS profile for MLaaS service selection. We identify the following key challenges in the MLaaS selection using the free trial:

- **Multiple MLaaS Providers:** Several MLaaS providers may meet a user’s QoS requirements, yet their performance can vary greatly due to differences in business strategies and infrastructures. Conducting a free trial with each MLaaS provider is impractical. Thus, a more efficient selection strategy is needed to identify the most promising candidates for trial evaluation.
- **Variability in QoS Performance:** Predicting the QoS performance based on a free trial is challenging without further insight or additional information on the MLaaS service’s performance. It means the QoS information obtained from a 30-day trial may not accurately represent the service performance for an extended period, which makes it difficult to assess the provider’s reliability. For example, in the first year, MLaaS service providers offer better biasness of the service; however, in the second year, they are focusing more on explainable MLaaS service due to market competition.

To the best of our knowledge, existing context-aware service selection approaches for Web, Cloud, and Edge computing typically consider short-term contextual information such as user preferences, service environment settings, and advertisements [[Rhayem et al., 2021](#)]. We aim

to select the most suitable context-aware MLaaS in dynamic IoT environments by considering changing contextual features, including long-term QoS, context duration, adaptability, and service evolution. However, selecting an MLaaS service based on a functionally similar service remains challenging. These functionally equivalent services often fail to account for non-functional aspects that are critical for the selection of MLaaS services in changing user contexts. The following key challenges have been in the context awareness selection of MLaaS.

- **Adaptability to Dynamic Contexts:** Several MLaaS providers offer functional services depending on the application requirements. The services offer similar functionality but vary significantly in non-functional nature. This selection process involves identifying and choosing the most suitable MLaaS provider from a range of similar functional services to meet specific QoS requirements where the organisation requires how each MLaaS service adapts to changes over time, especially in use cases such as IoT and the health care domain where user contexts are dynamic.

1.2 Research Contributions

A key objective of this work is to help a user and an organisation to make an informed MLaaS selection. We initially propose a novel we propose a novel MLaaS Selection Framework (MSF). In this approach, we assume that we have incomplete QoS information available through MLaaS service advertisement. We then introduce a context-aware MLaaS selection where we employ context-aware algorithms. Our approach seeks to enhance the efficiency, accuracy, and responsiveness of IoT systems aimed at optimising the interaction between IoT user’s activities and machine learning capabilities.

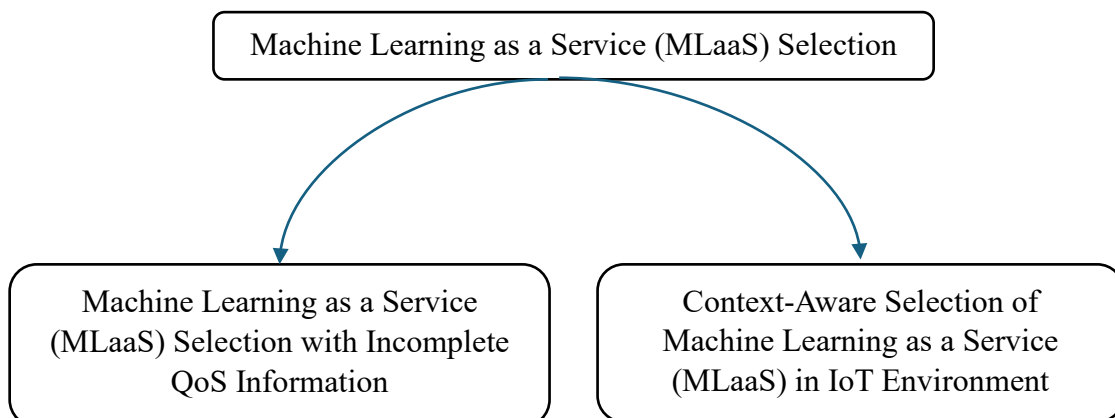


FIGURE 1.1: Key contributions

1.2.1 Machine Learning as a Service (MLaaS) Selection

1.2.1 Machine Learning as a Service (MLaaS) Selection with Incomplete QoS Information

We propose a novel MLaaS Selection Framework (MSF) that operates with the incomplete Quality of Service (QoS) information typically provided through MLaaS service advertisements. Discovering hidden features such as explainability and intrinsic biases, we design a knowledge-based bias detection and Explainable (B-XAI) framework. Our framework generates a comprehensive QoS profile by integrating data from service advertisements, user reviews, and short-term trial experiences. We then apply a nearest neighbour algorithm to select the most appropriate MLaaS providers, tailored to the user preference model.

1.2.2 Context-Aware Selection of Machine Learning as a Service (MLaaS) in IoT Environment

We propose a novel, cutting-edge framework for context-aware MLaaS selection in IoT environments, aimed at exploring dynamic MLaaS service mapping for users in varying contexts in smart health monitoring. We developed intelligent MLaaS services to manage real-time data from advanced sensors, focusing on context change analysis and personalised MLaaS service mapping. We present a context change analysis using a support vector machine (SVM) and design a contextual bandit algorithm integrated with skyline queries to achieve optimal alignment between abstract and concrete MLaaS services, considering important QoS attributes: accuracy, biasness and explainability. Our work addressed the challenge of selecting appropriate MLaaS services for a user within a smart hospital environment, ensuring the user receives services suited to their unique contexts.

1.3 Outline of the Thesis Chapters

We present the contribution of the research in two sections. The first section showcases our work on Machine Learning as a Service (MLaaS) Selection with incomplete QoS information, and the second part represents our work on Context-Aware Machine Learning as a Service (MLaaS) in IoT Environments. The thesis is structured as follows:

- Chapter 2 examines related work in the context of web service, cloud service, and IoT service selection methods and key distinctions between the current studies and methods proposed in this study.

- Chapter 3, we present a novel MLaaS Selection Framework (MSF) using incomplete QoS information available through service advertisement. First, we develop the knowledge-based bias detection and Explainable (B-XAI) framework to discover MLaaS latent features. Then, the proposed framework builds a complete QoS profile of the providers using MLaaS advertisements, other user experiences, and short-term trial experiences. Finally, we apply the nearest neighbour algorithm to select the optimal MLaaS providers based on the users' preference models.
- In Chapter 4, we extend our work to explore context-aware MLaaS selection. In this work, we propose a novel framework for context-aware selection of MLaaS in IoT settings, aimed at optimising the interaction between IoT users' activities and machine learning services. In our framework, we consider various contextual dimensions, such as user preferences, locations, IoT device capabilities, and application requirements, to develop a dynamic selection process. First, we propose a context change analysis algorithm based on support vector machines (SVM). Then, we develop a contextual bandits algorithm along with skyline queries to achieve optimal mapping between abstract MLaaS services and concrete MLaaS services for quality of service (QoS) attributes. This combined approach ensures that the selected MLaaS services align closely with the dynamic user needs and constraints of the IoT environment.
- In Chapter 5, we provide a conclusion to this thesis and briefly address the limitations and potential areas for further research.

CHAPTER 2

RELATED WORKS

Cloud computing has recently become a foundation technology, revolutionising modern IT infrastructure by providing several services. This transformation enables businesses and organisations to streamline operations, enhance collaboration, and drive digital transformation across different sectors. Cloud computing services, including Infrastructure as a Service (IaaS), Platform as a Service (PaaS), Software as a Service (SaaS), and Machine Learning as a Service (MLaaS) which, are designed to handle large-scale computational demands by utilising distributed resources.

MLaaS is a critical component of the cloud computing ecosystem. It offers powerful machine learning tools and algorithms to industries in the IoT space without requiring significant investment in complex hardware or specialised expertise. Major companies like Google, Amazon, and Microsoft provide subscription-based MLaaS, making it increasingly popular in E-commerce, Healthcare, and manufacturing [[Lupo, 2016](#)]. Also, the rapid growth of the Internet of Things (IoT) has further accelerated this trend, as businesses in healthcare, finance, retail, manufacturing, and smart cities increasingly rely on MLaaS to optimise operations, enhance customer experiences, and foster innovation through advanced data analytics and predictive modelling.

MLaaS delivers transformative *advantages* for IoT applications across multiple industries by offering scalable, cost-effective, and efficient solutions. For instance, in the healthcare sector, Philips leverages Amazon Web Services (AWS) to monitor patient health through wearable devices. AWS's robust infrastructure, including Amazon EC2 for computing power and Amazon S3 for storage, facilitates real-time patient condition analysis and prediction, ensuring timely medical interventions. Similarly, in the manufacturing industry, General Electric (GE) utilises Google Cloud's AutoML to forecast equipment failures and optimise maintenance schedules, effectively reducing downtime and lowering operations costs. These examples illustrate how MLaaS empowers various sectors to enhance operations and achieve greater efficiency through advanced machine learning capabilities.

The emergence of MLaaS services enables new ML models to tackle challenging engineering, medical, and social problems. ML algorithms have garnered increased attention

from both academia and industry for their role in empowering intelligence within edge devices. Studies have demonstrated the remarkable success of M-based IoT applications across various fields, including autonomous vehicles, clinical diagnosis, and face recognition [Jeong, Son, & Lee, 2019; Kute et al., 2022; Thara, Anusha, & Bharath, 2024]. Research indicates that MLaaS can tackle domain-specific challenges, such as wood identification, by leveraging deep learning techniques. For example, it demonstrated this capability with Xylorix, which shows how ML services can offer specialised solutions in the fields of forestry and material science that require precise identification and classification for informed decision-making [Tay, 2019]. A study on urban modelling workflows leverages ML to propose Urban Modelling as a service (UMaaS), aiming to enhance collaboration and data quality standards [Milton & Roumpani, 2019]. Additionally, a decision support system to assist doctors in assessing patients' health risks, advocating for the use of Risk Prediction as a Service. This study underscores concerns regarding data privacy and security [Mariani et al., 2019].

Despite its advantages, MLaaS must overcome two major challenges: Data Dynamicity and Service Degradation. Data dynamicity refers to the continuous changes in data over time, which can affect the model's performance. Service degradation involves a decline in model accuracy and effectiveness. This typically occurs as the model becomes outdated due to changes in the underlying data patterns, a phenomenon known as "data drift". As the data evolves – whether due to user behaviours, market conditions or environmental factors, the model may no longer be as effective at making accurate predictions. Therefore, addressing these challenges is crucial for maintaining the reliability and relevance of MLaaS solutions.

The existing literature shows limited work addressing these challenges in the MLaaS domain. A study explores the complexity of model degradation caused by polynomial approximation activations and pooling layers in prior LHECNN implementation on MLaaS platforms. They proposed a new method, Shift-accumulation-based LHE-enabled deep neural network (SHE), which uses binary-operations-friendly encryption schemes and logarithmic quantisation to address these issues [Lou & Jiang, 2019]. This method is a cutting-edge solution for secure and efficient ML in emerging MLaaS frameworks. A notable contribution addressed the problem of model degradation caused by membership inference attacks (MIA) on MLaaS. They introduced a novel approach called MIASec, which ensures the indistinguishability of training data, offering a defence against MIA in MLaaS [Hu et al., 2023]. However, these approaches primarily focus on security and privacy, potentially overlooking the critical

challenges of maintaining MLaaS efficiency and resilience in a dynamic IoT environment, for example, significant data changes in MLaaS can lead to model degradation, and rebuilding the model, including the processes of data collection and retaining, can be highly time-consuming.

2.1 Types of Machine Learning as a Service (MLaaS)

MLaaS can be primarily categorised into three main types: Inference-based MLaaS, Platform-based MLaaS, and Data-based MLaaS services. These offer distinct capabilities tailored to various needs. These services are seamlessly delivered and consumed in real-time over the Internet, enabling organisations to harness the power of ML. Examining these categories from a broader perspective allows for a deeper understanding of how MLaaS empowers businesses to integrate cutting-edge AI solutions into their operations, driving innovation and efficiency across diverse industries.

2.1.1 Inference-based MLaaS

Inference-based MLaaS revolutionise how organisations leverage AI by providing pre-trained models ready for deployment, offering efficient, scalable, and cost-effective solutions for making real-time predictions. This is also called *Trained Models as Service*. This type of service is particularly beneficial for applications requiring immediate, accurate responses, such as personalised recommendations, fraud detection, and predictive maintenance. By abstracting the complexities of model deployment and infrastructure management, inference-based MLaaS enables businesses to integrate ML into their operations. However, the “black-box” nature of this pre-trained model- where the service provider controls the training process- raises concerns about data privacy and trust. Users must weigh the convenience and performance benefits against the potential risks of entrusting their data to third-party providers. Despite the challenges, the accessibility of inference-based MLaaS makes it an invaluable and effective tool for organisations.

2.1.2 Platform-based MLaaS

Platform-based MLaaS represents a transformative approach to leveraging AI, offering a comprehensive ecosystem where users can easily manage the entire ML lifecycle. This is also called as *ML models as Services*. These platforms provide end-to-end solutions, from data pre-processing and model training to development and monitoring, all within an integrated environment that simplifies the complexity of ML. Offering scalability, pre-built algorithms,

and seamless integration with other cloud services, platform-based MLaaS helps organisations build, tune and deploy models efficiently, regardless of ML expertise. Additionally, robust collaboration tools and stringent security measures ensure that these platforms can support complex projects while maintaining the integrity and privacy of data. Services such as Google Cloud AI, Amazon SageMaker, and Microsoft Azure exemplify the power of platform-based MLaaS. This service allows businesses to use pre-trained models and APIs to solve specific problems, making ML more accessible and practical for diverse industries.

2.1.3 Data-based MLaaS

Data-based MLaaS represents a powerful fusion of advanced data management and ML, offering organisations the tools to turn vast amounts of data into actionable insights. This is also called *Collaborative Trained models as service*. These platforms are tailored to handle the complexities of big data, providing seamless integrations with diverse data sources, robust data pre-processing, and storage solutions within a unified environment. Using advanced analytics and automated ML(AutoML), data-based MLaaS empowers users to explore, visualise, and model data. These platforms are not only scalable enough to efficiently process large datasets, making them ideal in various domains, but also emphasise data security and compliance to safeguard sensitive information. Platforms such as IBM Watson and Azure Synapse Analytics enable organisations to make data-driven decisions, supporting from predictive analytics to real-time processing for immediate insights. For example, in collaborative settings, such as health care, these platforms allow for the sharing and training of models across multiple data owners to improve diagnostic tools while maintaining data privacy. This approach to MLaaS not only enhances the capabilities of organisations to leverage their data but also fosters collaborations and innovation in data-intensive fields.

2.2 A Review of Service Composition across Different Domains

Service Composition is a key area of research that is widely applied across various applications such as web-based services, cloud computing and IoT to improve scalability, flexibility, and efficiency. A solution for dynamic web service composition using domain ontology, where user requirements are decomposed into abstract services and matched with existing composite services through semantic matching to create an executable web service composition [[Wang, Tang, & Zhang, 2009](#)]. A selection and composition of web services by analysing their functional and non-functional attributes, proposing an improved method for

evaluating Web service QoS attributes using a variable weight vector for dynamic adjustment of indicator weights. Additionally, it introduces a particle swarm optimisation algorithm with linearly decreasing inertia weight and learning factors to enhance the speed and global search ability for web service composition. It emphasised quick decision-making but faced overfitting issues with optimised QoS results. [Liao, Wang, & Wu, 2023]. In the context of cloud computing, a hybrid artificial neural network-based particle swarm optimisation (ANN-PSO) algorithm designed to improve QoS factors, validate the algorithm's effectiveness and enhance the reachability rate of candidate services by a formal verification method using a labelled transition system is proposed to check linear temporal logics (LTL) formulas. Despite its service composition reliability, it lacked adaption to real-time IoT changes [Hosseinzadeh et al., 2020]. An Optimal Service Selection and Ranking framework for Cloud Computing Services (CCS-OSSR) is developed to assist cloud customers in comparing services based on QoS criteria. It proposes a hybrid multi-criteria decision-making approach, using the Best-Worst method to rank and prioritise QoS criteria and the Technique for Order of Preference by Similarity to the Ideal Solution (TOPSIS) to determine the final ranking of cloud services. It demonstrated robustness through sensitivity analysis through limited adaptability [Kumar, Kumari, & Kumar, 2021]. Another proposed framework for IoT service selection identifies key components, including communication, computing and things and defines QoS metrics for each, utilising a multi-criteria group decision-making method to rank services across different interaction models: push, pull, and hybrid. This framework addresses issues like rank reversal and fuzziness in decision-making, and its effectiveness is validated through a healthcare application case study and comparison with existing methods; however, biased results affect ranking [Baranwal, Singh, & Vidyarthi, 2020]. A rapid energy-focused and QoS-aware service composition approach (FSCA-EQ) for IoT services using hierarchical optimisation. It proposes a Compromise Ratio Method (CRM) that pre-selects services based on QoS requirements, followed by the method of relative dominance to choose the composite service for energy efficiency and extended IoT device lifetime, considering energy consumption profiles, QoS attributes, and user preferences. This approach has failed to adjust to changing user demands [Chai, Du, & Song, 2021].

2.3 Machine Learning as a Service (MLaaS) Selection Approaches

MLaaS service selection is a topical research issue in cloud computing [Sahi, 2022]. There is limited research on MLaaS service selection approaches. A novel MLaaS Selection Framework (MSF) that leverages incomplete QoS information available from service

advertisements. The framework incorporates a knowledge-based detection and Explainable AI(B-XAI) approach to uncover latent MLaaS features, building a comprehensive QoS profile using service advertisements, user experience, and short-term trials. Finally, the nearest neighbour algorithm is employed to select the optimal MLaaS providers based on user preference [Patel et al., 2023]. A comprehensive article has explored how MLaaS is applied in marketing, focusing on its customised platform featuring modules such as churn prediction, personalised product recommendations, and send frequency prediction [Pereira et al., 2024]. It discusses the benefits of AI-driven campaigns in improving Open Rate and Click Rate, enhancing customer engagement and retention, and enabling data-driven decision-making for businesses in a competitive market driven by consumer insights, through its generalisation beyond marketing is limited. Another study introduces a multi-criteria method for comparing and ranking various MLaaS providers across cloud service platforms. By integrating the Analytics Hierarchy Process (AHP) with the Technique for Order Preference by Similarity to the Ideal Solution (TOPSIS), it evaluates different MLaaS options. However, subjectivity involved in determining criteria weight leads to potential bias in service selection [Bhol, Mohanty, & Pattnaik, 2024].

In this chapter, we will discuss the existing approaches for selecting services in the context of Web, Cloud and Edge computing and emphasise their work with our research on the Selection of MLaaS with Incomplete QoS information. Moreover, in our work on Context-Aware Selection of Machine Learning as a Service (MLaaS) in IoT Environment, we will discuss how context awareness can be applied in the different domain applications of recommendation systems in the context of Web, IoT and cloud computing.

2.4 Service Selection in the context of Web, Cloud and Edge computing

Several studies have been published that present the service selection process. The study developed QoS ontology that expressed QoS information with constraints and adopted the AHP approach to select the set web service using predefined QoS metrics [Tran et al., 2009]. The research works consider non-functional attributes such as throughput, reliability, response time, availability, and price of service requesters to choose web service. It focuses on static service composition but lacks adaptability for dynamic environments [Huang et al., 2009]. A study quantitatively analysed service selection based on agents' preferences and proposed the QoS-based service selection algorithm for multiple agents with partial predictions. Despite its semantic interoperability, dependence on domain ontology limits its applicability [Wang et al., 2009]. Moreover, published works analysed and compared various

MLaaS providers to select the most efficacious one. They have compared the Natural Language Processing APIs of all different vendors and measured the right service provider in terms of cost, time, ease of use, and accuracy according to user requirements [Xie et al., 2022]. However, the mentioned studies did not consider the analysis of explainability and bias detection approaches for the QoS selection of MLaaS.

QoS-aware service recommendation is becoming significant for selecting services for an MLaaS that offers different services and associated applications with differentiated QoS requirements [Zeng et al., 2004]. In the context of web service selection, proposed a method that computed an optimal set of web services for each process based on a weighting combination of QoS measures and applied local and global approaches to select web services by maximising user satisfaction [Zeng et al., 2004]. A simulation-based method for QoS-aware dynamic service selection mobile edge computing systems is proposed, where stochastic system models and mathematical analyses are used to formulate the problem as a dynamic optimisation challenge and applied goal softening and developing service selection algorithms through ordinal optimisation approaches [Huang, Lan, & Xu, 2018]. However, additional research is needed for inclusion in the MLaaS selection, particularly in addressing the bias and enhancing the explainability of the MLaaS service. Most research comparing well-known providers' services regarding quality, price, and feature availability was published. One study represented the Armol framework for acquiring quality data measurement in MLaaS service selection. According to the study, they utilised a deep combinatorial reinforcement learning method to maximise accuracy and gave evidence of accurate results by inference of 67% less cost [Xie et al., 2022].

A Service Diversity Adjustment algorithm selects alternative services from outside the initial recommendation list by replacing those currently recommended and enhances the chances of meeting the user's QoS preferences more effectively; however, passing more data reduces the chances of recommendation efficiency [Kang et al., 2024]. A novel proposed approach that combines adaptive neuro-fuzzy inference systems(ANFIS) with metaheuristic optimisation methods to enhance the model's ability to solve complex problems such as it uses COOT bird optimisation to select parameters for ANFIS, which creates ANFIS-COOT model. This model is then applied to predict the QoS characteristics of web services; but limits adaptability with contexts [Jithendra et al., 2024]. A study addresses minimising response time in the selection of mobile edge computing by formulating an optimisation problem and proposing a heuristic

algorithm, GAMEEC, which combines Genetic and Simulated Annealing algorithms [Wu et al., 2019].

2.5 Selection of Context-Aware Service Approaches in the context of Web, Cloud and IoT

Various studies have explored context-aware service selection based on contextual information. A study has explored various methods for enhancing context-aware recommender systems (CARS) by considering contextual information (i.e., user, day place), including prefiltering, post-filtering, and contextual modelling, highlighting the potential for improved recommendation accuracy and user satisfaction [Adomavicius & Tuzhilin, 2010]. The authors introduced two innovative prediction models for web service recommendation and selection, incorporating user and service context information. They utilised geographical data to establish similarity among user neighbourhoods and incorporated company and country affiliations on the service side. These models predict QoS values by analysing historical records from users and services and data from their neighbours, aiming to enhance accuracy and reliability in service recommendations; however, these models are limited in adaptability in time-varying conditions [Xu et al., 2016]. A selection approach, Partial Historical Records-based service evaluation (Partial-HR), is implemented in context-aware cloud computing that assigns weights to each historical QoS record based on service invocation context. By prioritising the most relevant records, Partial-HR enhances accuracy and efficiency in the quality evaluation process, optimising resource utilisation and improving decision-making capabilities. This selective approach aims to maximise the utility of historical data while minimising computational overhead, offering a robust framework for evaluating service performance in dynamic environments but not utilised for changing user needs [Qi et al. 2015].

An IoT Medicare system is designed as a semantic-based context-aware system with Medical Connected Objects (MCOs), leveraging the HealthIoT ontology to describe heterogeneous MCO semantics. This system enables efficient knowledge management across contexts through SWRL rules, facilitating MCO functionality verification and health data analysis in a case study of gestational diabetes management [Rhayem et al., 2021]. The system produces a decentralised authentication architecture that enhances local authentication while considering context information from network elements, supported by Markov and random walk mobility models, demonstrating through simulations its ability to achieve a balanced trade-off between

network operating cost and reliability but faced scalability issues [Han et al., 2019]. A dynamic skyline operator to improve efficiency in multi-criteria decision-making for context-aware sensor selection in IoT architectures. The system features distributed gateways that respond to user requests locally, with results aggregated by a central service to obtain the final answer. It improves the response time and scalability but relies on sufficient contextual data [Kertiou et al., 2018].

A context-aware decision support system introduces Context Processing Rules designed to significantly enhance personalisation and decision-making support [Matos, 2020]. The selection of MLaaS is challenging to achieve solely through the use of Semantic Web Rule Language and Context Processing Rules within a context-aware decision support system. While these rules enhance personalisation and decision-making by employing flexible inference mechanisms and different comparison operators, they are primarily designed for achieving predictable outcomes in dynamic contexts. MLaaS selection, however, involves evaluating diverse machine learning models and considering factors such as scalability, performance metrics, data security, and compliance with industry standards. These factors extend beyond the capabilities of context-aware rules to manage comprehensively. Additionally, MLaaS selection requires adaptability to evolving technologies and datasets, which are not fully addressed by existing rule-based approaches. Thus, while valuable in specific scenarios, these rules alone may not suffice for the complex and multifaceted process of MLaaS selection.

The service selection studies enhance adaptability to user preferences, improved QoS prediction and effective management of diverse applications; however, they also exhibit limitations in scalability, challenges in maintaining relevance in dynamic environments, and issues related to the reliability and trustworthiness of data, which hinder their overall applicability in real-world scenarios. A summary of strategic aspects of service selection studies with strengths and notable weaknesses is defined in Table 2.1.

Table 2.1 A summary of service selection studies - strengths and weaknesses

Strategic aspects of service selection	Study of	Strengths	Weaknesses
MLaaS Selection	Pereira et al. (2024)	Enhancing user engagement through predictive analytics	-Relies on quality and quantity of data -limited generalisation beyond marketing
	Bhol, Mohanty, & Pattnaik (2024)	Recognises the MLaaS market's dynamic nature and rapidly evolving environment	Determining criteria weights introduces subjectivity and bias
Service Composition	Liao, Wang, & Wu (2023)	Quick decision-making and responsiveness	Optimised QoS results cause overfitting
	Hosseinzadeh et al. (2020)	Service composition correctness and reliability	Does not account for dynamic and real-time changes in the IoT environment
	Kumar, Kumari, & Kumar (2021)	Demonstrated robustness through sensitivity analysis	Limited adaptability
	Baranwal, Singh, & Vidyarthi (2020)	Demonstrated effectiveness through sensitivity analysis	Biased results cause inconsistent rankings
	Chai, Du, & Song (2021)	Energy-efficient service selection that balances QoS and energy consumption	Does not adapt to dynamically changing user demands
Service Selection across different domains	Tran et al. (2009)	Enabling fine-grained service customisation	Reliant on predefined QoS metrics
	Huang et al. (2009)	Focuses on static service composition	Lacking the adaptability needed for real-time or frequently changing service environments
	Wang et al. (2009)	Facilitates semantic understanding and interoperability between services	Reliance on domain ontology and OWL-S may limit applicability to specific domain
	Xie et al. (2022)	Federated object detection service	Scalability in large-scale MLaaS deployments remains unevaluated
	Huang, Lan, & Xu (2018)	Tackles challenges of state explosion and high variability in simulation-based optimisation	Reliance on real data limits findings' generalizability
	Kang et al. (2024)	Recommendations align	More user data reduces

		with users' diverse preferences	recommendation efficiency
	Jithendra et al. (2024)	Applicable for both time series and non-linear complex problems	Limit adaptability to diverse contexts
Context-Aware Service Selection	Xu et al. (2016)	Using context information to predict QoS	Limit adaptability to environments with time-varying conditions
	Qi et al. (2015)	Addresses the critical issue of trustworthiness in QoS information	Limit adaptability to changing conditions or user needs
	Rhayem et al. (2021)	Manage diverse medical objects for personalised health analysis	Require frequent updates to remain relevant in a rapidly changing IoT context
	Han et al. (2019)	Model balance network cost and reliability	Limit scalability
	Kertiou et al. (2018)	Improve response times and scalability	Limit applicability if contextual data is insufficient or unreliable.
	Matos (2020)	Flexible inference mechanisms for dynamic decision-making	maintaining and updating rules become cumbersome

2.6 Understanding the Implications of MLaaS

As MLaaS represents a transformative shift, it provides a comprehensive view of how it impacts businesses both immediately and over the long term, as well as the economic considerations involved, which are as follows:

2.6.1 Short-Term Implications

- **Rapid Adoption:** Organisations and users can quickly integrate machine learning into products without the need for in-house expertise or infrastructure, enabling faster innovations. For example, a healthcare provider utilises MLaaS to implement predictive analytics for patient risk assessment. By adopting services such as Microsoft Azure or IBM, an organisation can deploy models that predict patient outcomes and identify high-risk individuals in a matter of weeks [[Philipp et al., 2020](#)]. This rapid deployment allows healthcare providers to personalise patient care plans and allocate resources more efficiently.
- **Cost-Efficiency:** Users only need to pay for specific tools and services, which allows them immediate savings and scalability, which is particularly helpful for smaller

organisations [[Patel et al., 2023](#)]. This approach helps businesses manage their budget effectively while scaling their operations.

- **Focus on Use Cases:** MLaaS offers benefits and allows organisations to concentrate on applying AI to business problems (e.g., predictive analytics) rather than managing infrastructure, which, as a result, organisations can allocate resources to enhance business problems.
- **Easy Deployment:** MLaaS offering pre-built APIs, models, and tools speed up deployment, making it easier to adopt AI technology. For example, a social media, utilising pre-built sentiment analysis APIs offered by MLaaS service providers to quickly incorporate sentiment tracking into their platform. This facilitates them to deliver new features to their clients with minimal development time [[Pereira et al., 2024](#)].

2.6.2 Long-Term Implications

- **Data Dependency and Vendor Lock-In:** Relying on MLaaS service providers can reduce flexibility and increase long-term costs, making it more difficult to switch platforms or adapt to alternative solutions in the future. For example, a financial service becomes heavily dependent on an MLaaS fraud detection service; as integrating deeply with one of the MLaaS ecosystems, switching to a different provider would require significant reengineering and incur high costs.
- **Advanced Customisation:** As MLaaS platforms evolve, they offer more powerful tools. However, using them in an organisation requires deeper expertise, which, as a result, creates a barrier for less technical teams.
- **Security and Privacy Concerns:** Managing a large volume of sensitive data in the cloud presents significant challenges, particularly in maintaining data security and ensuring compliance with evolving privacy regulations [[Hu et al., 2023](#)]. Failure to adhere to these regulations not only compromises the security of the data but also results in severe financial penalties and operational disruptions.
- **The democratisation of AI:** In the long run, MLaaS will make AI accessible to a wide range of businesses, accelerating innovation across industries [[Fortuna et al., 2023](#)]. For example, a start-up business can leverage MLaaS to implement precision techniques. This access to advanced AI tools allows them to compete with larger, established companies, driving innovations across the sector.

2.6.3 Economic Implications

- **Cost Management:** MLaaS provides services at relatively low initial costs, which is highly attractive for businesses seeking to adopt ML capabilities without upfront capital investment. However, as operations scale up, particularly for large organisations that demand high computing power and storage, operational expenses can grow significantly. Continuous usage of advanced models and large datasets may lead to increased costs over time, requiring careful financial planning to maintain cost-effectiveness [[Grigoriadis et al., 2023](#)].
- **Increased Innovation:** MLaaS reduces the barriers to entry for advanced ML technologies, making them accessible to a broader range of industries. This accessibility promotes economic growth and innovation by enabling businesses of all sizes to harness AI capabilities. As more companies adopt MLaaS, they can drive transformation in their respective industries, leading to enhanced products, services, and operational efficiencies [[Lupo, 2016](#)].
- **Resource Efficiency:** By outsourcing machine learning infrastructure, businesses can optimise their internal resources. This allows them to allocate more focus on the core competencies rather than managing complex IT infrastructure. As a result, this approach enhances economic efficiency by reducing the need for expensive in-house hardware. In short, it lowers operational costs and improves productivity [[Pereira et al., 2024](#)].

2.7 Conclusion

This chapter provided an overview of the fundamental service frameworks of MLaaS and other domains. We discussed the formal definition of MLaaS and explored its different types, providing insight into the diversity of services offered within this paradigm. We discussed the service composition approaches in the context of Web service, Edge and Cloud Computing and IoT domains. We also discussed the current studies of MLaaS Selection approaches that focus on the different strategies used to select MLaaS services based on specific criteria. We then discussed service selection in the context of Web, Cloud and Edge computing and the selection of context-aware service approaches in the context of Web, Cloud and IoT. Finally, we have highlighted short-term, long-term and economic aspects of MLaaS that offer insights into how these factors influence decision-making in service adoption and sustainability.

CHAPTER 3

MACHINE LEARNING AS A SERVICE (MLAAS) SELECTION WITH INCOMPLETE QoS INFORMATION

3.1 Introduction

As we gradually move towards the future, artificial intelligence with machine learning has become a game-changer in computing [Sahi 2022]. Machine learning (ML) is a subspecialty of Artificial Intelligence (AI) that is considered a most important innovation, especially called the industrial revolution [Sahi, 2022]. According to PwC's study, the AI revolution will be the new drive, and by 2030, it will contribute \$16 trillion to the world economy [Sahi, 2022]. Many researchers published advanced reports on these fields in the last 5-6 years, but data scientists and engineers are slowly engaging with that advancement [Pugliese et al., 2021]. According to their studies, new machine learning-based innovated technologies would cater to materialising industries. Today, cloud services offer several services delivered to companies and valuable customers as these services are easy to maintain, affordable and applicable without catering hardware or software programs [Sun et al., 2014]. This means that organisations can select the cloud service that best fits their business and satisfies user requirements. Large companies can afford to build machine learning services to process and analyse vast amounts of data; however, small companies need help with fundamental factors, time, cost, and technical expertise. To cope with situations, cloud service providers offered readily available Machine Learning as a Service delivery model (a range of machine learning tools) where one can use MLaaS services without writing a single line of code [Ribeiro et al., 2015]. Machine learning engineers, data engineers and other professionals pay more attention to MLaaS as it helps ML teams in various ways, including data pre-processing, model training and tuning, and predictive analysis [Pop et al., 2016]. The popular MLaaS are Google Cloud AI, Microsoft Azure, Amazon Sage Maker, IBM, and Watson machine learning. There are three types of MLaaS. The definition of each is below:

1. **ML models as Services** (users provide training data, and they tune the models applicable to a single user). MLaaS service provider solves the company's problem using customers' applications. MLaaS services consist of trained models that do not require uploading any training data. The service provider can fulfil customers' needs as they use models via API calls to get predictions. Examples include Google Cloud

Vision API [[Google Cloud, n.d.](#)] that detects objects and ModelScope for Alibaba Cloud [[Alibaba Cloud, n.d.](#)].

2. **Trained models as Service** (complete black box to the users, trained by others). In this type of MLaaS, clients have no control over the training process; the service provider keeps the Service in a black box. Here, the customer would be concerned about the data and insights they provide to train new models placed in the service provider. It creates the issue of trust. Examples include AWS Rekognition [[Amazon Web Services, Inc., n.d.](#)] and Google AutoML [[Google Cloud, n.d.](#)].
3. **Collaborative trained models as Service** (users collaborate and share the models). In this type of MLaaS, the machine learning model is shared and collaboratively trained. Here, training the ML model requires a large amount of data. For example, healthcare service training models help clinicians make a diagnosis based on other collected data. They combine data from multiple hospitals, and data owners cannot share it openly because of privacy concerns.

Many big organisations, such as health care, education sector, and research facilities, use MLaaS services in the long term. As a result, selecting an exemplary MLaaS service is a significant decision for long-term customers. Customers determine long-term service requirements based on budget, history, and internal revenue [[Ye et al., 2014](#)]. MLaaS service consists of functional and non-functional; functional attributes are data storing and virtual machines, and non-functional attributes are the Quality of Service (QoS), such as response time and throughput. These QoS attributes help customers select the best MLaaS-performing services from similar services.

Selecting the best MLaaS is only possible by getting complete information on a provider's long-term QoS information [[Ye et al. 2014](#)]. A study mentions that service selection is vital in the healthcare sector due to its unique nature, as certain services may contain risks, and selecting the effective service that prioritises the patient's safety and reduces potential harm is an essential criterion [[Lupo, 2016](#)]. Also, the best MLaaS service selection offers the students support service flexibility. Published work analysed personalised and intelligent systems such as a chatbot catering to students' support services in a single interface. It connects students on different mobile and desktop applications and addresses students' queries [[Srimathi & Krishnamoorthy, 2019](#)]. MLaaS providers typically contain limited, incomplete information because of the QoS stringent management policies in a dynamic environment, related market

competitors and conflicts of interest. *We identify an incomplete advertisement from service providers as a challenge in the MLaaS service selection.* MLaaS advertisements typically contain limited QoS attributes such as speed, cost-effectiveness, explainability, bias, and availability, which are unavailable in most MLaaS advertisements [[Amazon Web Services, Inc., n.d.](#)]. For instance, AWS Sagemaker should offer complete information about the provided service's model explainability and bias detection quality. It keeps the latent QoS features, such as explainability and bias, private. In short, the advertised QoS information may only be representative for a short time. However, providers often advertise their services' average or maximum measurement. For instance, the predictive service of Google Cloud AI allows one to request a prediction from the model and get the target values [[Google Cloud, n.d.](#)]. Today, there is a debate around the globe about the present bias in online recruiting apps, Facebook ads, facial recognition tech and even in the criminal justice algorithm, which, as a result, shows unfairness to the communities [[Akter et al., 2022](#)]. As a result, to respond to this debate, measuring the MLaaS service quality in terms of bias is an essential criterion.

Similarly, an explainable service is also a significant requirement from the user to know why the model has come to a particular decision. The bias and explainability of MLaaS services are not embedded in any current work for service selection. One of the research papers presents a web service selection according to user QoS requirements and preferences and then defines ranking through semantic matching where complete information is provided [[Makhlughian, 2012](#)]. As a result, incomplete information such as bias and explainability need to be explored for the service selection. To solve this, we propose a novel generated framework to select the best MLaaS selection.

Our contributions to this paper are as follows:

- Using the weight function, we perform a data selection process to retrieve the data from different MLaaS aspects: MLaaS advertisements, trials, and past users' trial experiences.
- We are creating the B-XAI framework for bias and explainability to measure the level of bias and how explainable the service is from the past available methods.
- This study proposes a k-nearest neighbour algorithm for the MLaaS service selection.

3.2 Motivation Scenario

Let us assume that a private bank's financial and credit department wants to buy a loan decision-making system. The credit department wants the decision-making system to be ML models as service system types of MLaaS service where the MLaaS service provider can fulfil the department's needs. It means the department can use the provider's model via API calls to get predictions for Bias and explainability of the services. As with the increasing demand for MLaaS service in the market and several MLaaS providers offering various services according to user requirements, a department would like to buy an automated decision-making system MLaaS service at a low cost. As we mentioned before, academic centres and businesses do not have enough resources in terms of technical expertise, time, and fundamental factors; therefore, for example, universities and some small companies always prefer to buy readily available MLaaS services according to their needs rather than invest high-cost money in establishing. As for the qualitative preferences from user requirements, different users have different preferences; for example, a bank or university prefers more biases and explainability of the services and is ready to invest the high cost of money; however, a small organisation, for example, spend low-cost money to buy virtual assistance system for customer support. In our case, the banks' credit department's qualitative requirement for the MLaaS service requires the QoS attributes, such as Explainability, Bias, Price and Availability, and Response time of MLaaS service. Many MLaaS service providers are available in the market and offer similar services; however, providers do not advertise all pieces of information related to the QoS. Therefore, selecting the best MLaaS service provider for the bank's credit department is challenging. The banks' credit department always wants to select services that match their qualitative requirements the most. According to the bank's credit department, the target is to discover the latent qualitative features, explainability, and biases of services from MLaaS providers.

The department's qualitative preferences on explainability, bias, price, and availability can be interpreted in two semantic levels: high and low. The department has different preference ranks for each quality attribute throughout the duration. The bank's credit department prefers high-quality QoS attributes to improve the model explainability of the service and bias detection. Hence, the bank's credit department prefers "high" explainability to "low" explainability of the MLaaS service. Similarly, the department prefers "low" bias to "high" bias in the MLaaS service. However, the department's preference for the Price and Availability of the QoS attributes are conditional on the combination of Explainability and

the Bias of the MLaaS service. According to user preference, the credit department prefers MLaaS services that provide high explainability and low Bias. Also, the department would invest a high cost of money if the MLaaS service provider delivered a highly explainable service with a low level of Bias. In addition, the department's requirement for service availability depends on the price attribute as the bank's credit department requires a high availability of the MLaaS service if a high amount of money is invested in the decision-making system. The bank's credit department knowledge shows more inter-dependencies within criteria or attributes.

3.3 Building MLaaS Service Selection Framework

We are formulating the MLaaS service selection using the following formal definitions.

- **User:** A new MLaaS user wants to select an MLaaS service based on the required QoS attribute information. In our case, we will consider a bank's credit department as a new user.
- **Non-functional requirements:** Non-functional requirements are defined in terms of the quality of a particular service. For example, a service's quality attributes are availability, response time, speed, and ease of use.
- **QoS Requirements:** A QoS requirement of a user is a set of QoS attributes and their average or minimum values.
- **Provider:** A provider is an MLaaS provider who advertises the QoS attributes for MLaaS services.
- **QoS Advertisement:** A QoS advertisement is a set of QoS attributes for MLaaS service and the values of the QoS attributes.
- **Trial Periods:** A user can use some services with restricted conditions for a short time for free.

Several MLaaS providers may advertise the services that meet the bank's credit department's needs. Let us assume three MLaaS providers, Amazon, Microsoft, and Google, fulfil the credit department's non-functional requirements. Let us assume that P1 and P2 are two MLaaS service providers. Both providers advertise the service's response time, availability, and throughput information in the MLaaS advertisements. However, information about bias and Explainability is hidden and unavailable in their advertisements. It means no provider advertises or has incomplete information about the Explainability and Bias of the service. The bank's credit department will select the P1 or P2 provider based on the response time,

availability, accuracy, and throughput information. However, this selection would not be a good decision as it needs to consider the bank's credit department requirements regarding the types of Explainability and the bias of the service. We assume each provider offers a two-month free trial period, allowing the bank's credit department to use it for some duration for each service. For example, the bank's credit department may run 8 hours every day in a two-month free trial and measure the QoS service attributes of each provider to make the best selection. However, the MLaaS provider's service may fluctuate in the free trial duration. Specifically, measuring the provider's service behaviour in the given trial period is difficult. The bank's credit department requires effective trial strategies to measure the quality of service to understand the MLaaS service. When MLaaS service providers offer limited QoS attributes and some of the pieces of information are hidden, as well as different MLaaS providers available in the market, our goal is how we can fulfil the bank's credit department's QoS requirements, especially for types of bias of services and Explainability as this information is hidden in the MLaaS advertisements. Figure 1 shows the proposed MLaaS service selection framework. For example, the framework requires a user's qualitative QoS preference as the input. The other input to the frameworks is the MLaaS advertisements from the MLaaS providers who can fulfil the non-functional requirements of the user. A detailed description of the provided framework follows below.

To make informed MLaaS service selection, the banks' credit department has three sources to get QoS information:

- Most of the MLaaS providers advertise decision-making services with some functional and non-functional attribute information. Based on available MLaaS advertisements, the bank's credit department can get QoS attribute information.
- Let us assume that the bank's credit department can access the experiences of past trial users, so based on other users' reviews, the bank's credit department can measure QoS attributes according to their preferences.
- Let us assume that the bank's credit department can perform its trial and get QoS attributes that match its qualitative preferences.

With the help of the above-described sources, it is easy for a bank's credit department to select the best MLaaS provider that offers the best service and meets its credit department's needs. Below, we will describe the three sources mentioned in detail.

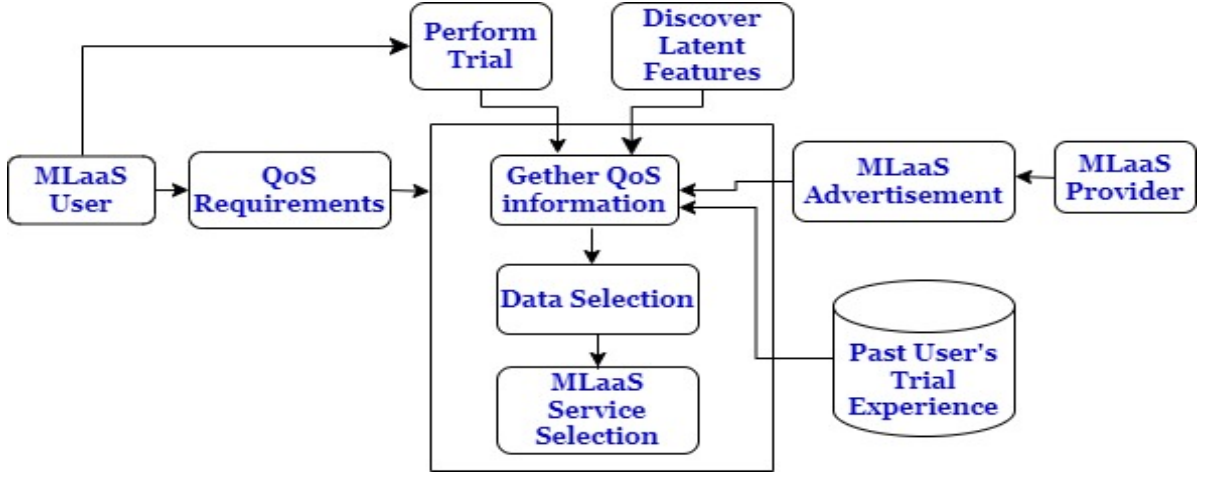


Figure 3.1: MLaaS Service Selection Framework

3.3.1 MLaaS Advertisements.

With the given MLaaS advertisement, the department wants to measure the quality of service to make a service selection. Let us assume the bank's credit department wants to measure QoS attributes for an MLaaS service. We assume that the user has full knowledge of Quality-of-Service attributes. There are N numbers of MLaaS providers who can fulfil the non-functional requirements of the bank's credit department. The MLaaS provider set is $P = \{P_1, P_2, \dots, P_N\}$. The QoS attributes of the provider P_i are denoted as $A = \{A_1, A_2, \dots, A_k\}$, where k is the number of QoS parameters in A . We assume that $k < l$, for example, the number of QoS attributes in the advertisements is always less than the number of the QoS attributes in the user preference. The advertisements provide insufficient information (i.e., $k < l$). We denote the QoS attribute's measurement of the provider P_i as $Q_i = \{q_{i1}, q_{i2}, \dots, q_{il}\}$, where l is the number of QoS parameters in Q_i . The bank's credit department wants to select the service based on its required QoS attributes closely matching its MLaaS provider's QoS attributes. Given the bank's credit department's QoS requirements Q_d and the provider's QoS attribute is Q_i , we use a measuring function distance (Q_d, Q_i) to find the most matched QoS attribute of an MLaaS service using the below-given formula:

$$M_a = \operatorname{argmin}(i = 1..n) \operatorname{distance}(Q_d, Q_i)$$

3.3.2 Performing own trials.

Another source of information for the bank's credit department is trialling with the MLaaS provider to know about QoS parameters. Let us assume the department can perform its trials to make qualitative MLaaS service selections. We consider that the bank's credit department

QoS requirements (Q_d) contain only one QoS parameter; for example, $|Q_d|=1$, we can apply the filtering method based on the single criterion decision-making. The filtering process can be modelled as a single or multiple criteria decision-making based on the user's requirements [Ye et al. 2014]. In our case, we are considering a single-criteria decision-making process. For example, if the department only cares about the availability of the MLaaS service, the service is selected based on the availability without considering other MLaaS QoS attributes. In this case, we compare the bank's credit department's QoS requirement with each MLaaS provider using time series similarity matching approaches. One of the most effective, fast and easy-to-implement similarity-matching techniques is Mean Absolute Error (MAE) distance [Fattah, Bouguettaya & Mistry, 2020]. Using the below equation, we can calculate the similarity between the department's QoS requirement Q_d and the MLaaS provider's advertisement (P_a) for a single QoS attribute. It will provide the bank's credit department's required QoS attributes about an MLaaS service.

$$\text{MAE}(Q_d, P_a) = \frac{1}{n} E_{t=1..n} | Q_d^t - P_a^t |$$

In the above equation, n is the number of timestamps, Q_d^t is the value of the QoS attribute Q_d , at time t . Some service attributes can be selected using the Top-K technique [Zheng et al., 2012]. The Top-K method selects the best MLaaS service that contains the minimum distance from the department's QoS requirement. If the numbers of selected MLaaS are too large or small, K can be adjusted to increase or decrease the number of MLaaS services for the trial period. In short, using the MAE approach, the bank's credit department can measure the MLaaS QoS attribute about one parameter in each free trial period.

3.3.3 Measure the QoS information from the other user experience.

The experiences of other trial users might not be directly measured to predict the MLaaS service provider's attributes for a new user. This is because each user may have performed a trial with different MLaaS services according to their preference and contains different experiences. Collaborative filtering-based approaches are widely known for predicting QoS information from similar user experiences; for example, the study of Zheng et al. (2012) measures user similarities between two users based on the consumer's QoS experience. Let us assume that other users have performed similar trials to the bank's credit department at such time, for example, in the month of "December". We denote the number of timestamps at

a particular time is n. For each timestamp t, the bank's credit department can predict the QoS attribute's values:

$$E_{t=1..n}(Q_d) = \frac{\sum_{i=1..k} Q_{dk}^t}{k}, t \in \Delta T$$

where k is the similar user of the department, Q_{dk}^t is the value observed by the other user c_k at time t. $E_{t=1..n}(Q_d)$ measures the QoS in the trial period T based on average observed measurement by other similar users. The above Equation measures the QoS attributes of MLaaS service using the average of past trial experiences. Here, we will consider all users equally without considering the degree of similarity of each user. For example, a particular user might have the highest or lowest similarity if we consider four past users. It will provide a poor accuracy measurement of QoS attributes.

3.3.4 Data Selection

The above-performed service selection using three sources: MLaaS advertisements, performing trials, and getting the MLaaS QoS service information from other users; we now assume that the bank's credit department has multiple information for each QoS attribute for an MLaaS service except bias and explainability of the MLaaS service, as these are the hidden features in MLaaS advertisement. This means that some of the QoS information meets the bank's credit department's preference or requirements, or some do not. One of the best ways we can do this is statistical analysis. Several methods exist to refine data based on multiple criteria, including conditional preference, skyline, and utility function [Fattah, Bouguettaya & Mistry, 2020]. Among these utility functions is computing the score of each service provider according to the user's preference on each QoS attribute. With the help of the utility function, the bank's credit department may put the highest weight on the most trusted or important service information and set the lowest weight on the least preferred attribute information. For example, suppose the users have multiple information for each QoS attribute. In that case, the user can assign either the highest weight on trustable information for required attributes or the lowest weight that is not important QoS. This means that the utility function computes a score for each MLaaS service based on the given weights of the QoS attributes and their respective values of attributes. The utility function computes the score as $Pref *$ equation:

$$Pref * = \sum_{q_d \in Q_d, p_a \in P_a} W_q * MAE(Q_d, P_a) + W_q * E_{t=1..n}(Q_d) + W_q * M_a$$

where W_q is the weight of the QoS attribute q given by the department, Q_d is the bank's credit department's QoS preference, P_a is the MLaaS advertisements as utility function can make the problem in a single criteria decision-making problem.

3.3.5 Discovering Latent Feature - Bias

Let us assume that the bank's credit department needs to get the QoS attribute information about the Bias of the MLaaS service. The reason for not disclosing information about latent features is business competency. Our study facilitates well-informed MLaaS service selection by uncovering latent features like biases and explainability. It measures the quality of the MLaaS service so the bank's credit department can select the service. We have performed the static analysis to find the bias detection methods from past research. A bias is a prejudice against a group of people. It is considered unfair to the group due to prejudiced assumptions in training data or when developing an algorithmic process. The causes of Bias are historical human Bias, selection bias, active Bias, latent Bias, biased labels, and imbalanced representation. The Bias occurs in the whole AI life cycle, including the pre-training phase, model training and validation process, and even model deployment and monitoring. Let us consider the classification problem where many published works are used in domains such as NLP, image classification and finance. From the above debate about Bias, the decision-making system is considered biased for the favoured and disfavoured groups. Let us take an example: the ML-based application that assesses customers' financial information and decides whether the customer is eligible or not; therefore, in that case, the model trained on the data set may favour certain groups of customers, which as a result, leads to reduce the chance of other customers' eligibility means not getting the loan. We are creating a knowledge-based graph for a loan approval decision-making system to learn about past bias detection methods for the loan approval process. Figure 3.2 shows the B-XAI framework (knowledge graph). The B-XAI framework focuses on enabling the selection of MLaaS services by identifying latent features such as bias and explainability in decision-making systems. It aims to enhance applications such as loan approval processes by detecting bias at various stages of the AI lifecycle. A study used the human-agent interaction method to investigate the agent's explainable behaviour by introducing a bias in the human decision-making process. Specifically, the study has performed a qualitative analysis where users can detect the Bias in the agent's decision and make noticeable differences between explanation and non-XAI recommendations as black box recommendations by utilising LIME and SHAP tools for the loan application [Malhi et al., 2020]. The published work mentioned that

explanation-based recommendation is beneficial to reduce Bias in the human decision-making system. The work has applied the random forest classification method to identifying good borrowers in the world's biggest social lending platform for the financial domain. Also, they compared different algorithms for the cross-validation process. Studies have reduced the selection bias and considered a random forest a scalable and robust approach to choosing the best borrower for social lending [[Malekipirbazari & Akasakali, 2015](#)]. Several comparison algorithms, including random forest, nearest neighbour, support vector machine and logistic regression, give good accuracy scores such as 78%, 70%, 63.3% and 54.5%, respectively. However, a novel method called GRC to learn about application representation and detect loan fraud based on these representations [[Xu et al., 2021](#)]. To detect loan fraud, they have validated the method by using varying numbers of labelled samples from 1000 to 12,000, where applying the SVM algorithm model performed at a low level at 57.5 % accuracy (for 1000 samples) and gives 60.51% accuracy when processing with the 12,000 samples [[Xu et al., 2021](#)]. Studies have applied bias detection with several different impacts and statically parity differences mitigation processes to achieve fairness in the decision-making system by using several classifiers, including XGBoost, LightGBM, and RF, for bank customer data [[Wardani et al., 2023](#)]. During pre-processing, in-processing and post-processing stages, the reweighing and equalised odds methods with the mentioned classifier failed to achieve the level of Bias and made the model more biased. At the same time, only adversarial debiasing shows good performance in mitigating Bias [[Wardani et al., 2023](#)]. Another methodology that controls the Bias is using a probabilistic network that exploits structural equation modelling. Studies have shown better responses on a loan approval data set and highlighted the effect of tuning parameters on the Bias [[Barbierato et al., 2022](#)]. Based on the above bias detection methods comparisons, we have derived the value of the biased model (i.e., high or low) according to the respective model's performance. Low Bias can be categorised as 0% to 4%, and a highly biased model is measured in 5% to 10% value to get the values of models' performance about Bias. The bias detection methods, which are high and low-biased models, highlighted performance can be shown in the knowledge graph (see Figure 3.2).

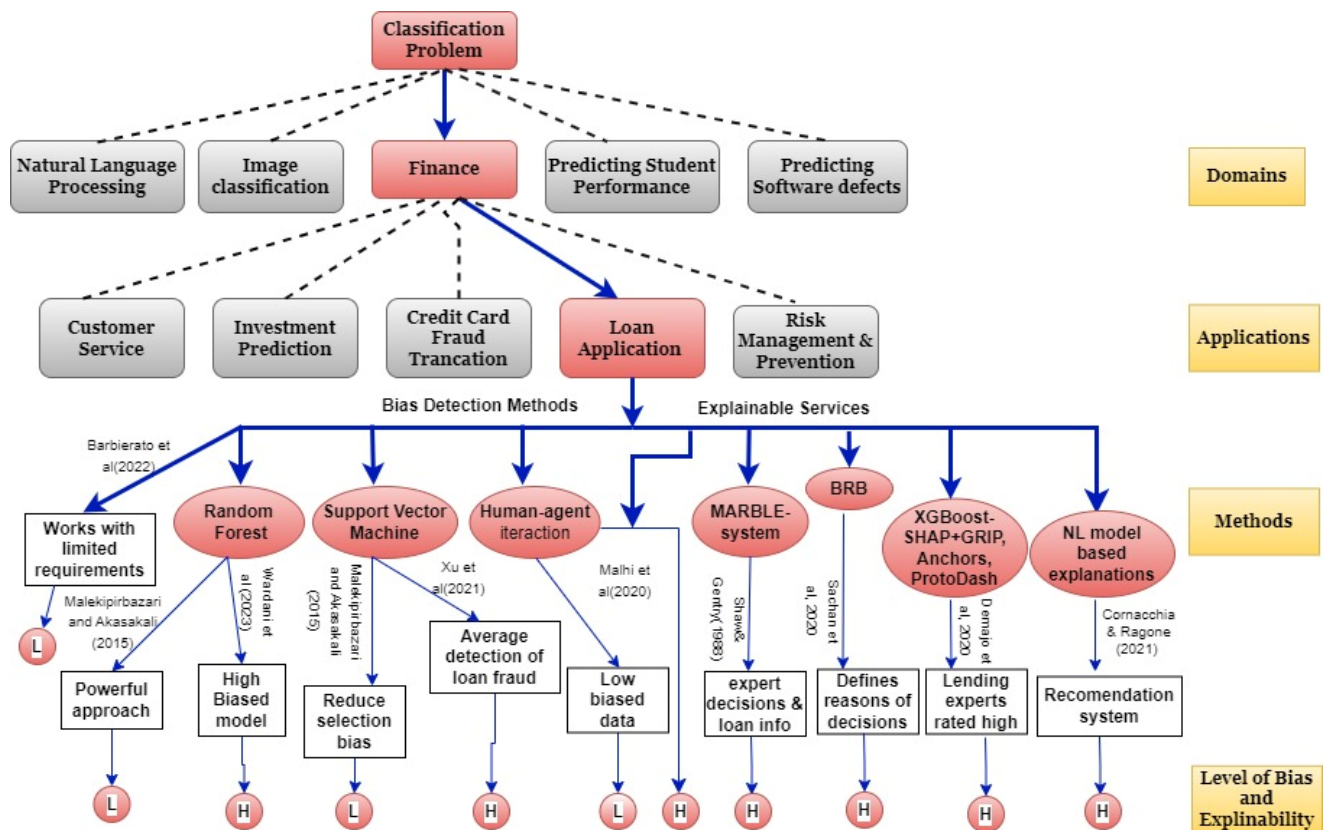


Figure 3.2: A B-XAI framework (a knowledge-based graph for Bias detection methods and Explainable service methods)

3.3.6 Discover latent feature- Explainability.

Here, we assume that the MLaaS provider needs to provide information about the explainability of the MLaaS service, meaning the information is hidden for business competency. Therefore, we need to discover the latent feature explainability and measure the quality of the same service so the bank's credit department can select the service based on preference. In the banking department, credit scoring models are the decision models that help lenders decide whether to accept loan applications based on their decision-making system or the model's decisions. There are several AI and ML-based techniques proposed to solve this research problem. Some black box nature methods do not provide the reasons or explanations behind the decisions. Consequently, the inability of humans to interpret such predictions leads professionals to place trust in model outcomes without conducting proper evaluation or assessment. In our proposed B-XAI framework (see Fig 3.2), through the integration of explainable AI techniques developing knowledge graphs, the framework empowers informed decision-making in MLaaS selection, particularly for critical financial use cases.

Model explainability has become an emerging field in computing and IT area. XAI focus on opening the black box and improves the reasons behind the logic behind the decisions or predictions. In the loan approval decision-making system, lenders should understand the models' decisions to ensure the system makes decisions correctly. A decision-support system using a Belief rule-based method offers a better trade-off between prediction accuracy and explainability [Sachan et al., 2020]. The study mentions that in the rule-based system, the activated rules and related attributes are the significant factors in understanding the reasons behind the decisions. Their system defines the reasons for rejecting loan applications: textual explanations are sent from a factual rule base to a heuristic rule base [Sachan et al., 2020]. This study compared the performance of the BRB system with other machine learning models, DT, RF, XGBoost, and SVM, where the BRB provides high accuracy, 0.9550, where others reside between 0.87 and 0.94. The research ensures that it can be comprehended by individuals without technical expertise and readily implemented within enterprises. Similarly, MARBLE, an expert system, aids loan officers, credit assessors, and loan reviewers in enhancing the loan evaluation process [Shaw & Gentry, 1988]. MARBLE incorporates expert decisions and loan applicant information; both are considered to decide on loan repayment [Shaw & Gentry, 1988]. It means that the judgement of lending experts is a significant process for the decision-making system. Another published work published proposed a credit scoring model by incorporating XGBoost, which balances type I and Type II errors. Also, the XGBoost-based enhanced 360-degree explanation framework provides human-understandable explanations using post-hoc explanation methods [Demajo et al., 2020]. It provides different explanations, such as global explanations using the SHAP+GIRP method both local feature-based and local- instance-based explanations using Anchors and ProtoDash methods [Demajo et al., 2020]. The evaluation process shows that all three explanations are correct, complete, adequate, scalable, and trustworthy so that users can implement them in the bank's decision-making system. Furthermore, a study developed a natural language-based explanation model for the loan recommendation system [Cornacchia, Narducci, & Ragone-2, n.d.]. This model takes user requests, and the developed loan recommendation platform compares available offers of loans and provides explainability of each offer using counterfactual explanation. The system also accepts the user's help to modify the request to be eligible for the loan. In short, the help of a set of actions of counterfactual explanation makes the loan application accepted. Based on the above explainable AI methods, we will

consider the value of the explainable service (i.e., High or Low) according to the model's results and expert judgement, shown in the knowledge graph (see Figure 3.2). To assess the performance of models regarding the explainability of MLaaS services, we assign a value of 0 to represent low explainability and a value of 1 to denote high explainability services. The methods that offer explainable services are highlighted in the knowledge graph (see Figure 3.2).

3.3.7 MLaaS Service Selection

The proposed framework ranks MLaaS service providers using the K-nearest neighbour (KNN) algorithm, which aligns with a user preference model based on QoS attributes. This model captures the user's QoS attribute requirements, such as response time, availability, latency, bias, and explainability, along with their relative importance through weighted preferences. This method computes the distance between service providers' QoS attribute information and the user's QoS attribute requirements, and the rank of each provider is computed based on their nearest distance from the user's requirements [Peng et al., 2020]. MLaaS providers are ranked based on their proximity to the user's requirements, with those closer to the user's preference vector receiving higher ranks. This approach ensures the selection of MLaaS providers that best align with the user's prioritised QoS expectations. The preference ranking algorithm is the follows:

Algorithm: Preference Ranking Algorithm

Input: M; number of MLaaS providers, N; a number of user requests

Output: Rank; Ranking the MLaaS provider according to user preference

Rank = \emptyset

For i= 1 to n do

For j = 1 to m do

Compute the Euclidean distance between i and j

End for

Sort array based on nearer distances

Select the best k nearest neighbours

Assign the rank to the nearest neighbours in ascending order

3.4 Experiments and Results

3.4.1 Experiment Setup

Finding real-world datasets that meet our experiment requirements for an extended period is challenging. We require MLaaS providers' Quality of Service datasets for a long-term period.

To our knowledge, long-term QoS datasets of MLaaS providers are not publicly available. Therefore, we leverage existing Quality of Service of Web services datasets (QWS) to synthesise datasets for our experiments. MLaaS is running over web services and with 1000 web services, we will use the Quality of Service of Web services dataset (QWS) to create experimental datasets [Giommi, 2023]. In the real world, the big MLaaS provider, Amazon, advertises information about deployment and availability. Also, finding MLaaS datasets for bias and explainability is very challenging. Therefore, we have created a B-XAI framework to discover those features, and we will take the data from the knowledge base graph. These values are then augmented with the QWS dataset for MLaaS providers. Finally, to build the complete profile QoS of MLaaS, the user ranks the provider based on the help QoS profiles of each provider. We will use randomly generated user preferences for service selection.

We have implemented the discussed aspects to evaluate the feasibility of the proposed framework. We are considering three cases with three different datasets to measure the proposed approach. For the experiment, we are taking $n = 20$ samples (see example Figure 3.3(a)) where the QoS attributes are Response Time, Availability, Latency, Bias and Explainability. In Figure 3.3(a), case 1, we have generated a dataset with insufficient or limited information. Let us assume the QoS information comes only from the advertisements. In Figure 3.3(a), case 2, we have generated data from the different aspects, such as MLaaS advertisements, trials, and other user experiences and combined those sources of QoS information. Also, bias and explainability of service values are derived from the knowledge-based graph (see sections 3.3.5, 3.3.6). Here, we consider the QoS web service dataset in case 2, response time, availability, and latency. We have combined all those data to measure the feasibility of the proposed approach. In Figure 3.3(a), case 3, we have generated a dataset using random data; for example, we assume that past research used the random data to detect the bias without generating any knowledge system. After creating all the mentioned datasets, we used the k-nearest neighbour algorithm to rank the $k=5$ (i.e.) algorithm according to the user preference. Based on the user query, [Response time=500, Availability=93, Latency=20, Bias=3, XAI=1], we have generated a ranking for MLaaS providers for all three cases (see example Figure 3.3(b)).

3.4.2 Evaluation

After ranking the nearest MLaaS providers based on user requirements, we performed the evaluation process with the help of two evaluation matrices.

- **Spearman's Rank Correlation:** The Spearman Rank Correlation coefficient measures the monotonicity of the relationship between two variables. The Spearman Correlation between two variables will be high when the observation contains a similar rank between two variables and gives low when there is dissimilarity.
- **Kendall's Tau:** This coefficient measures the association between two variables. Gives a value near one means the ranking is similar, and a value near -1 means the ranking is dissimilar.

To evaluate the process, we have compared the three cases' rankings. The ground truth is a user preference, [Response time=500, Availability=93, Latency=20, Bias=3, XAI=1]. Later, considering Rank 1 of each case, we measured the proposed approach in terms of the monotonicity of the relation between two variables and the association between two variables by Spearman's Rank Correlation coefficient and Kendall's Tau coefficient, respectively. The result (See Figure 3.3 (c)) shows that case 2, which contains complete knowledge of data (combination of data), performed well with Spearman's Rank Correlation and Kendall's Tau, giving value 1 ((See Figure 3 (c)). This means that observation contains a similar rank between two variables. However, case 1 gives 0.71 and 0.63 for Spearman's Rank Correlation coefficient and Kendall's Tau coefficient, respectively, and case 3 has results of 0.90 for Spearman's Rank Correlation coefficient and 0.80 for Kendall's Tau coefficient.

Case 1					Case 2					Case 3				
Res Time	Availability	Latency	Bias	Explainability	Res Time	Availability	Latency	Bias	Explainability	Res Time	Availability	Latency	Bias	Explainability
146.5	0	0	0	1	146.5	72	3.5	2	1	146.5	72	3.5	5	0
382.6	87	0	0	0	382.6	87	4.8	3	1	0	0	0	7	0
0	0	0	0	1	121.46	61	9.15	6	1	121.46	61	9.15	5	0
102	91	0	0	0	102	91	1	1	1	102	91	1	5	0
253.33	0	0	0	0	253.33	88	50.33	7	1	0	88	0	7	0

(a)

User Preference
Response Time= 500.0, Availability= 93.0, Latency=20.0, Bias=3.0, XAI=1.0
Case 1: Preference Ranking
Rank 1: Response Time= 505.0, Availability= 0.0, Latency=0.0, Bias=0.0, XAI=0.0
Case 2: Preference Ranking
Rank 1: Response Time= 505.0, Availability= 89.0, Latency=32.67, Bias=8.0, XAI=1.0
Case 3: Preference Ranking
Rank 1: Response Time= 581.0, Availability= 83.0, Latency=116.0, Bias=3.0, XAI=0.0

(b)

3 Different Cases	Results
Case 1 Rank1: [505.0, 0, 0, 0, 0]	Spearman's Rank Correlation: 0.71 Kendall's Tau: 0.63
Case 2 Rank1: [505.0, 89.0, 32.67, 8.0, 1.0]	Spearman's Rank Correlation: 1.00 Kendall's Tau: 1.00
Case 3 Rank1: [581.0, 83.0, 116.0, 3.0, 0]	Spearman's Rank Correlation: 0.90 Kendall's Tau: 0.80

(c)

Figure 3.3 (a) Case 1- Limited QoS information, Case 2- collect data from different aspects (full knowledge of QoS data), and Case 3- Random data (b) Preference ranking for all 3 cases according to user preference (C) 3 cases with evaluation of Results

3.5 Conclusion

We proposed an MLaaS service selection framework to select the optimal MLaaS service for a user. The proposed framework helps the user to make informed decisions in the selection as the user depends not only on advertisements from the MLaaS service providers. The proposed framework augments data through MLaaS advertisements, free trials, and past user experiences. We also discover the latent features bias and explainability by creating a B-XAI framework to get the data. Finally, we have discovered the experiments with full knowledge of the information's dataset compared to incomplete information. We found that Spearman's Rank Correlation coefficient and Kendall's Tau coefficient matrices give better results, value 1, with complete information than incomplete information.

One of the key limitations of this work is the use of synthetic datasets in the experiments. There are some challenges with using synthetic data, as it gives us biased or deceptive results due to a lack of variability and correlation. However, where the real data does not exist, synthetic data is the only solution. We leverage existing Quality of Service of Web services

datasets (QWS) to synthesise datasets for our experiments. The synthetic dataset is then augmented with the values of bias and explainability (data taken from the knowledge base graph) to capture the characteristics of real MLaaS providers. Therefore, the result would be similar if the experiments were conducted using real-world datasets. Also, our proposed work needs to consider temporal user aspects, which will be addressed in future work. The proposed framework also does not consider the arrival of incoming user requests and the probabilistic qualitative user preferences in a natural language. A possible extension of this research is to extend the MLaaS service selection using a conditional preference network (CP-Net) where users can express their preferences more qualitatively to make informed selections.

CHAPTER 4

CONTEXT-AWARE SELECTION OF MACHINE LEARNING AS A SERVICE (MLAAS) IN IOT ENVIRONMENTS

4.1 Introduction

Machine Learning (ML) plays a crucial role in enhancing the capabilities of the Internet of Things (IoT) by providing advanced analytics, predictive capabilities, and intelligent decision-making [[Pereira et al., 2024](#)]. One effective way to utilise ML is through Machine Learning as a Service (MLaaS), a cloud-based platform that offers machine learning tools and services without the need for users to invest in their infrastructure [[Patel et al., 2023](#)]. Leading companies such as Amazon, IBM, Microsoft Azure, and Google Cloud provide MLaaS services that can be integrated with IoT solutions [[Pereira et al., 2024](#)]. For example, smart thermostat IoT sensors gather temperature, humidity, and occupancy data. This data is sent to AWS IoT Core (an IoT service). AWS Lambda (an MLaaS service) processes the data, which is then analysed by ML models in AWS Sage Maker (another MLaaS service). These models predict optimal ventilation settings, which are applied in real-time to reduce energy consumption and lower costs [[Amazon Web Services, 2023](#)].

The *selection of Machine Learning as a Service (MLaaS)* is a significant issue in the IoT domain. This selection process involves identifying and choosing the most suitable MLaaS provider from a range of similar *functional* services to meet specific *Quality of Service (QoS)* requirements. In any IoT environment, selecting the right MLaaS is a complex and crucial task that requires several considerations to ensure the chosen service aligns with the application's specific needs. For instance, consider a smart home system, where IoT devices such as wearable technology, smart sensors, and environmental sensors collect vast amounts of user data. The goal is to leverage MLaaS for predictive analytics, such as detecting anomalies in smart home activities, predicting potential risks, and optimising home operations. Functionally compatible MLaaS providers like AWS Sage Maker and Google Cloud Auto ML each offer unique features. One provider might excel in robust scalability, while another might offer more customisation options. Hence, each MLaaS provider must be selected based on factors such as scalability, integration capabilities, security features, cost, and ease of use.

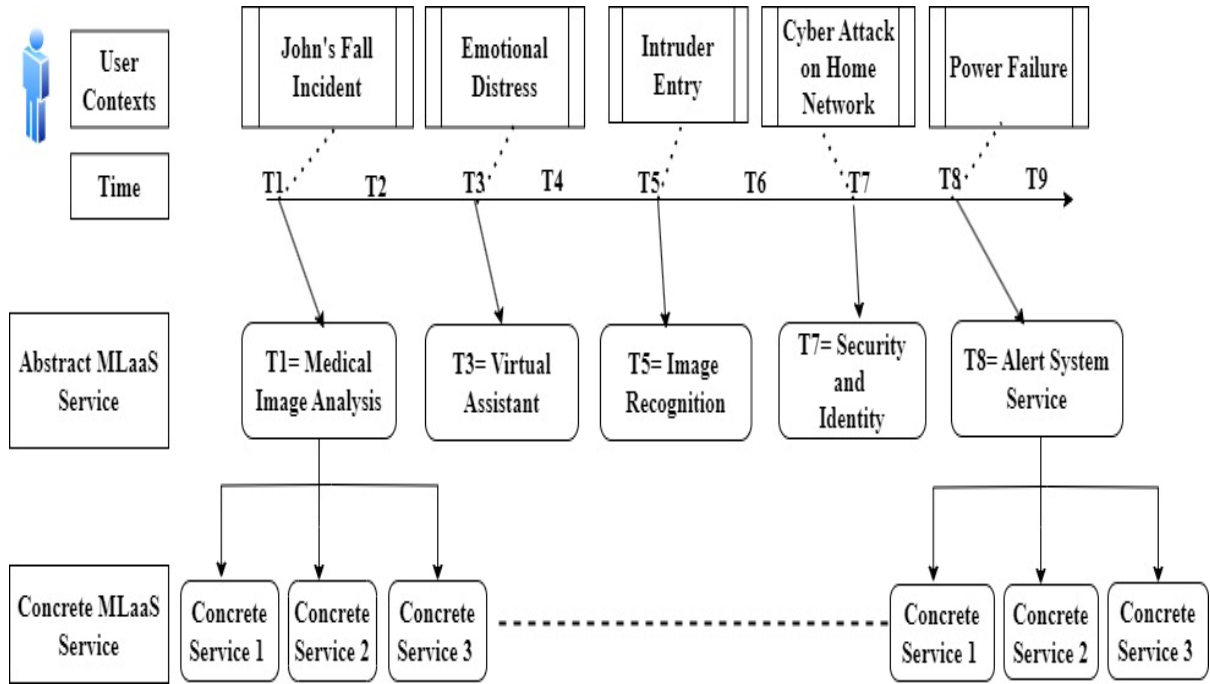


Fig. 4.1: MLaaS Service Selection based on Single user's context

In this paper, we focus on the *context-aware* MLaaS selection in IoT environments. A “context” is any information that can be utilised to characterise the situation of an entity, where the entity could be a person, place, physical, or computational object [Adomavicius & Tuzhilin, 2010]. Existing research mainly addresses MLaaS selection based on functional and non-functional (QoS) properties [Pereira et al., 2024, Patel et al., 2023]. However, our approach goes further by incorporating contextual information to enhance the selection process for both *abstract* and *concrete* MLaaS services. An abstract MLaaS service refers to a high-level representation of a service that encapsulates the general functionalities and capabilities required to meet specific user needs. This helps identify the type of MLaaS needed to effectively meet contextual needs. While abstract services represent high-level functionalities and performance characteristics needed by an application, concrete services are the actual, operational services that fulfill these requirements.

We consider context awareness as the *long-term applicability* of the MLaaS, ensuring that the chosen MLaaS provider can *adapt* to evolving needs and sustain performance over time. For example, in a smart home (see Fig. 4.1), let us assume a user (i.e., John), a tech-savvy, is

adapting with varying contexts over time $t = \{1, 2, 3, \dots, 9\}$ such as sudden fall, emotional distress, unauthorised entry, cyber-attacks on a home network, and power failure. John may fall in the initial context (time $t=1$), where it triggers the smart home system to initiate the medical image analysis abstract service. However, multiple MLaaS providers offer concrete services for medical image analysis, each with different QoS features such as varying levels of bias, explainability, and accuracy. In context awareness of MLaaS selection, the system must continuously adapt to the user's changing contexts. For instance, if a user shows signs of emotional distress (feels frustrated and upset) after a fall, the smart system transitions from fall detection to offering virtual companion service, ensuring timely and accurate responses. It means a system that efficiently and accurately switches between health assessment, virtual assistant services, security, and cybersecurity to ensure user safety and well-being. In this paper, *we only consider the MLaaS selection from a single user's contextual information*. The multi-user context-based MLaaS selection is out of the scope of this paper.

Existing context-aware service selection approaches for Web, Cloud, and Edge computing typically consider *short-term contextual information* such as user preferences, service environment settings, and advertisements [Rhayem et al., 2021, Xu et al., 2016]. Predefined rule mining approaches are used to match contexts' applicability with services [Matos, 2020]. However, these approaches face significant limitations in IoT environments. Maintaining and updating rules becomes cumbersome as the complexity and diversity of contexts increase, limiting their adaptability and scalability. Ontology-based approaches have been developed to enhance context awareness for IoT-based smart monitoring systems [Rhayem et al., 2021]. While these methods provide a structured framework for context modelling, they require frequent updates to remain relevant in *rapidly changing* IoT contexts.

We propose a novel framework for selecting the most suitable context-aware MLaaS in dynamic IoT environments. The proposed approach considers rapidly changing contextual features, including long-term quality of service (QoS), context duration, adaptability, and service evolution. Traditional ontology-based and fixed rule-based approaches face challenges in this domain [Rhayem et al., 2021; Matos, 2020], as MLaaS must account for the duration of the context, i.e., how long a particular context remains relevant. The adaptability of MLaaS is also crucial, as it can apply to multiple contexts and *uniquely evolve through feedback and interactions*. These continuous improvements and adaptation to changing requirements and contexts, distinguish our approach from traditional service

selections such as web and cloud services [Qi et al., 2015; Xu et al., 2016]. The main contributions of this paper are as follows:

- Development of an IoT context analysis framework for pattern identification using Support Vector Machines with semi-supervised learning.
- Enabling mapping user context to *abstract MLaaS services* using a novel contextual bandits approach.
- Selecting *{concrete MLaaS services}* through skyline queries for optimal context-aware service selection.

4.2 Context-Aware MLaaS Selection Framework

We formalise context-aware MLaaS selection with the following definitions:

User Context: The user contexts UC are represented by combinations of features and values, denoted as $UC = \{UC^t | \forall f_i^t \in UC^t, f_i^t \in F_i\}$. Here, UC captures all possible combinations of features f_i over time, where F_i represents the set of values for the feature i . For example, John’s fall incident, emotional distress, etc., are the user contexts and contain features and values.

Abstract Service: An Abstract Service (AS) outlines key functionalities to meet specific user needs. For example, Medical Image Analysis is an AS that helps healthcare professionals diagnose and monitor conditions during fall incidents.

Concrete Service: A Concrete Service (CS) implement the requirements of AS by being selected based on QoS criteria such as biasness and explainability.

The proposed *Context-aware MLaaS Service Selection Framework (CAMSF)* is designed to enhance the efficiency and effectiveness of selecting Machine Learning as a Service (MLaaS) solutions by incorporating three distinct layers (see Figure 4.2). The first layer, *changing user context*, involves dynamically assessing and understanding the evolving needs and conditions of the user, such as location, time, and specific task requirements. The second layer, *selecting abstract MLaaS services based on user contexts*, utilises the information gathered from the first layer to identify and match suitable abstract MLaaS services that align with the user’s context. This step ensures that the services considered are relevant and capable of meeting the user’s general needs. Finally, the third layer, *selecting concrete MLaaS services to ensure*

optimal *QoS* (Quality of Service), focuses on the practical implementation by choosing specific MLaaS providers that not only fit the abstract requirements but also offer the best

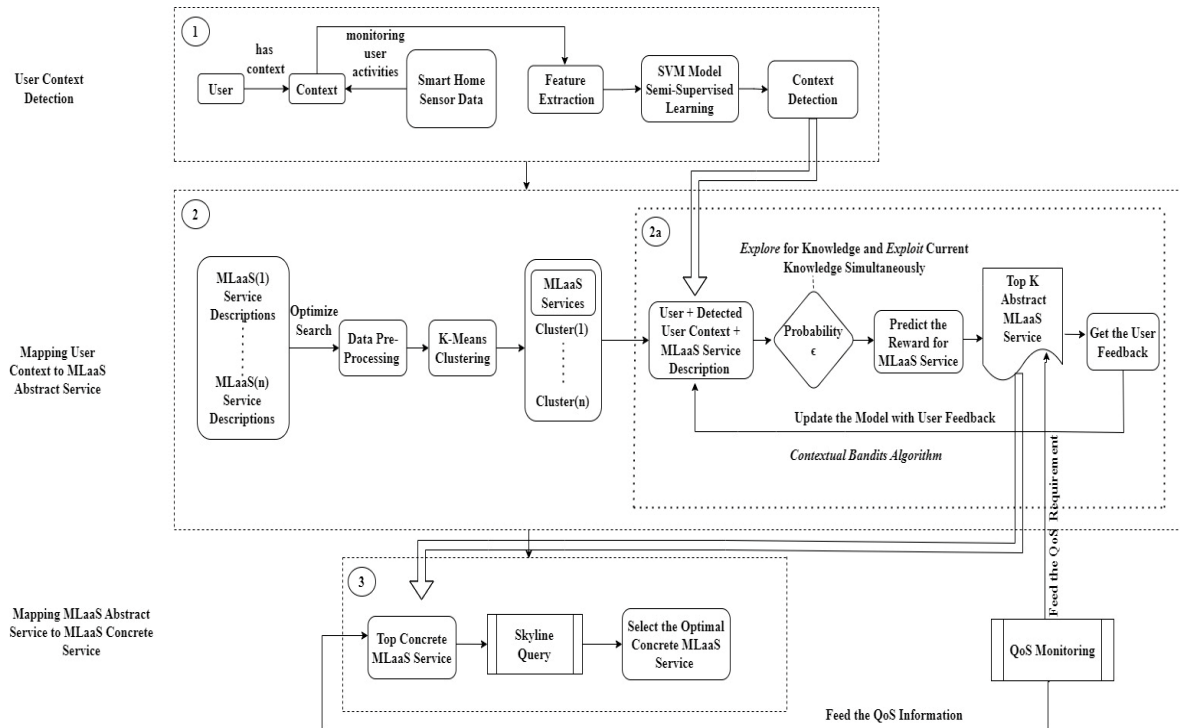


Figure 4.2: Context-Aware MLaaS Selection Framework (CAMSF)

performance, reliability, and other QoS metrics. Together, these layers form a cohesive framework that adapts to user contexts and optimises service selection for superior outcomes. *QoS monitoring* continuously evaluates MLaaS performance based on key metrics and serves two roles: (1) feeding QoS requirements to inform abstract MLaaS selection based on user needs and (2) providing real-time QoS data for selecting the best-performing concrete services. Our framework utilises Support Vector Machines (SVMs) to detect user contexts. It deploys contextual bandit algorithms to dynamically select abstract services, ensuring adaptive service selection and optimisation in dynamic IoT environments. This approach concludes with a skyline query method that filters and identifies concrete MLaaS services that meet user quality of service (QoS) requirements across various attributes.

4.2.1 Context Change Analysis

This section explores how changing user contexts can enhance MLaaS selection using a *Support Vector Machine* to classify and predict contextual patterns. First, we design a feature extraction process. Feature extraction reduces data dimensionality by selecting relevant features, aiding context analysis and revealing meaningful patterns [Rhayem et al., 2021]. We standardise sensor data to a mean of 0 and a standard deviation of 1 by calculating $Z = X - \mu / \sigma$; where X is the data value, μ is the mean, and σ is the standard deviation. The Z score indicates how many standard deviations a particular observation is from the mean, providing a standardised interpretation of the data.

Let us assume that X is the feature matrix of size $n * m$, where n is the number of data samples, and m is the number of features. We need to identify a single user context by labelling a small subset of the data instances with the appropriate user context. Therefore, we are implementing a Support Vector Machine (SVM) to determine the user's current context detection (Figure 4.2(1)). SVMs are a supervised learning model for classification and regression, aiming to find a hyperplane in feature space that maximises the margin between classes, known as the Maximal Margin classifier [Mohd et al., 2023].

We train the SVM model with labelled and unlabelled user context data to maximise the margin between context classes. Regularisation ensures a smooth decision boundary, formulated similarly to supervised learning with an added smoothness term. The mathematical formulation for the optimisation problem in SVM with semi-supervised learning is similar to the supervised learning case with an additional term representing the smoothness constraint using the below-given formula:

$$\min_{w,b} \frac{1}{2} \|w\|^2 + C \sum_{i=1}^n \delta_i + \epsilon \sum_{i=1}^n \delta_{\text{unlabeled}}$$

Subject to constraints:

$$y_i = (w * x_i + b) \geq 1 - \delta_i, \quad \forall i \in \text{labeled data}$$

$$\delta_i \geq 0, \quad \forall i \in \text{labelled data}$$

$$|w * x_i + b| \leq M \delta_{\text{unlabelled}}, \quad \forall i \in \text{unlabelled data}$$

where δ_i is the slack variable for the labelled data, $\delta_{\text{unlabeled}}$ are the slack variables for the unlabelled data, C is the regularisation parameter, and ϵ controls the importance of the

smoothness constraints. M is the constant representing the margin for unlabelled data. After that, the trained SVM model will be used to get the user context for new data instances. Given a new data instance x_{new} , the predicted user context can be obtained using the below-mentioned decision function:

$$f(x_{new}) = \text{sign}(w * x_{new} + b)$$

If $f(x_{new}) > 0$, the predicted user context belongs to one class, and if $f(x_{new}) < 0$, it belongs to another. If $f(x_{new}) > 0$ and if $(x_{new} < 0$, the SVM model represents an identified user context. However, values close to zero indicate uncertainty, termed *unidentified user context, which is out of the focus of this paper.*

4.2.2 Mapping User Context to Abstract MLaaS Services

We use the *Contextual Bandit* approach, a class of *reinforcement learning method* that integrates contextual information for decision-making [Varatharajah & Berry, 2022]. To map user context to abstract MLaaS services, first, textual descriptions of MLaaS advertisements are processed by removing punctuation, stop words, tokenisation, stemming, and vectorisation, to narrow down relevant services. K-means clustering groups services by advertisement description to simplify mapping [Purohit & Kumar, 2016]. Finally, the contextual bandit algorithm selects the appropriate abstract service based on user context (see Fig. 4.2 (2a)). We define the following key terms for the contextual bandit modelling:

- **State (s):** Represents the user context, including contextual features.
- **Action (a):** Represents a decision, such as selecting an MLaaS abstract service.
- **State-action pair (s_i, a_i):** Combines the current user context (s_i) with a specific action (a_i) selecting an MLaaS abstract service aiming to learn the best actions per state for maximising cumulative rewards.

Given S, A and sequences of observed context-reward pairs $\{(s_t, a_t)\}$, the goal is to find a policy $\pi = S \rightarrow A$ that maximises the expected cumulative reward can be expressed as follows:

$$\max_{\pi} E [r_i | x_i]$$

The mapping begins by observing the current user context, encompassing user-specific features and contextual information fed into the decision-making system. We then create a comprehensive feature vector by concatenating user features (e.g., age and location) and their

context information (e.g., fall downtime and type of fall) with the pre-processed MLaaS service features. This vector is input into our model to predict the likelihood of selecting abstract MLaaS service. The algorithm employs an exploration-exploitation strategy, referred to as a ϵ -greedy policy. This strategy performs the MLaaS abstract service selection with a probability ϵ , the algorithm explores service selection randomly, while with probability $1 - \epsilon$, it exploits by choosing the MLaaS service based on current knowledge [Varetharajah & Berry, 2022]. The predicted reward for selecting the MLaaS service M_i in the state represented by X_i using the below calculation:

$$\hat{R}(X_i) = \beta_0 + \sum_{j=1}^m \beta_j x_{ij}$$

Where $\hat{R}(X_i)$ is the predicted reward for a state-action pair X_i , β_0 is the intercept term, β_j are coefficients corresponding to each feature x_{ij} , m is the total number of features in X_i . The algorithm (Algorithm 1) updates its predictive model and policy based on user feedback to improve decision-making and maximise cumulative rewards. This contextual bandit approach allows for adaptive MLaaS abstract service selection based on user context by balancing exploration and exploitation and iterative updating the model with user feedback, improving the offering of abstract services at expected rewards.

Algorithm 1: User Context to Abstract MLaaS Service Mapping

Input: User features $U = \{u_1, u_2, \dots, u_n\}$, User context features $UC = \{uc_1, uc_2, \dots, uc_m\}$ and MLaaS service descriptions $M_i = \{mi_1, mi_2, \dots, mi_p\}$

Output: Select the MLaaS abstract service m_i with $\hat{R}(x_i)$

Initialize $X_i = [U, UC, M_i]$

For each interaction:

While $M_i \neq \emptyset$:

$\pi(X_i) = P(M_i | X_i)$

Select mi_1 using ϵ -greedy policy

Count $\hat{R}(X_i) = \beta_0 + \sum_{j=1}^m \beta_j x_{ij}$

Top-K $\leftarrow \operatorname{argmax}_{i \in n} \hat{R}(X_i)$

Get user feedback r_i

$\pi \leftarrow \operatorname{update}(\pi, (x_i, m_i, r_i))$

End while

End for

4.2.3 Mapping Abstract MLaaS Service to Concrete MLaaS Services based on QoS

This section maps abstract services to concrete services based on non-functional attributes (see Fig. 4.2(3)). Generating non-functional attributes from the context is important to selecting the best *QoS* concrete MLaaS service. Let us assume that the user has a context where it contains additional information, for example, variations due to environmental factors such as high heart rate, which may affect the non-functional attributes. Mapping based on non-functional attributes ensures that the MLaaS service chosen aligns with the user's *QoS* requirements in varying contexts.

QoS Monitoring: To select an optimal *QoS* concrete MLaaS service, we focus on key indicators, including biasness and explainability. We assume that *QoS* Monitoring delivers these services and feeds to CAMSF for selection. In a smart home system, identifying bias is crucial, as a model trained on biased datasets may favour certain user groups, which can lead to some individuals not receiving the necessary services. For example, elderly people often have different service requirements than typical adults. To align concrete MLaaS services with *QoS*, we utilise the established Adaptive Boosting model [Javed et al., 2021], which effectively classifies human activity data, identifies biases, and differentiates between the activities of elderly individuals and typical adults. Furthermore, accurate model predictions are essential. To achieve this, we use an established SHAP-based explanation (Shapley Additive Explanations) model developed by [Das et al., 2023], which generates meaningful and interpretable explanations for the model predictions. This approach enhances user trust and confidence in the system's decisions by providing insights into how specific inputs influence outcomes. The method to assess accuracy is not part of the focus. When multiple providers offer a service, we choose one by calculating the *QoS* score for each, weighting relevant attributes (e.g., biasness, explainability) based on the context of the user. The service with the highest utility score is selected.

$$Utility\ Score = \frac{\sum_{i=1}^n w_i \cdot QoS_i(S)}{\sum_{i=1}^n w_i}$$

where n is the number of *QoS* attributes (e.g., biasness, explainability), w_i is the weight assigned to the i^{th} *QoS* attributes. $QoS_i(S)$ is the *QoS* value of the i^{th} attribute for service S . S is the MLaaS service being evaluated.

We propose a skyline method to filter services based on bias and explainability *QoS* attributes, ensuring optimal selection. Inspired by real-world skylines, this method prioritises services not dominated by others across multiple attributes [Fattah, 2021]. We assume there

are N concrete MLaaS services meeting user non-functional requirements. We aim to find a subset CS' of M services ($CS' \subset CS$) based on QoS attributes and user preferences Q_{UP} . Each service has QoS attributes $QoS = \{QoS_1, QoS_2, \dots, QoS_n\}$. The selected subset $QoS' \subset QoS$ should match Q_{UP} , making this a multi-criteria selection problem. We are applying a temporal skyline to select concrete MLaaS services based on multiple QoS criteria. This method identifies superior services in a dataset by extending to a time series for optimal selection over time T [Fattah, 2021].

Solving the Skyline Query

A concrete MLaaS service CS_i is said to dominate another service CS_j if it provides equal or superior QoS across all attributes and exceed the QoS in at least one attribute. A skyline of MLaaS services includes optimal services that are not dominated by others across all QoS attributes. For example, if a user prioritises explainability and biasness in a Medical Image Analysis service, only services with these attributes are considered and weighted equally. We are considering the concepts below for QoS selection.

- Dominant Concrete MLaaS Service: A concrete MLaaS service CS_i dominates another concrete service CS_j , denoted as $CS_i > CS_j$. If CS_i offers as equal or superior QoS information in Q_D . Specifically for every quality metric q in quality domention Q_D , CS_i must meet or exceed CS_j (*i.e* $CS_i \geq CS_j$) with at least one q' where CS_i distinctly outperform CS_j (*i.e* $CS_i > CS_j$).
- MLaaS Skyline: The skyline is denoted as SK_{CS} , is a subset of concrete MLaaS service CS , that are not dominated by any other concrete service. It means a concrete service cs is included in the skyline if there is no other service cs' in the set CS that outperforms it across all QoS criteria.
- Temporal MLaaS Skyline: We utilise the temporal skyline, where QoS parameters are represented as time series that evolve over time.
- Dominant QoS Time Series: A QoS time series Q_i is said to dominate another QoS time series Q_j over the time period T , represented as $Q_i > Q_j$, if $\forall t \in T, Q_i \geq Q_j$, and $\forall t' \in T, Q_i > Q_j$.

– Temporal *QoS* Skyline: The temporal *QoS* skyline of a set of *QoS* time series Q , denoted as ST_Q , consists of those time series that are not dominated by any other series at any timestamp t , i.e., $ST_Q = q \in Q | \neg \exists q' \in Q : q' > q$.

By applying the *MLaaS* skyline approach it filters concrete *MLaaS* services based on multiple *QoS* attributes, ensuring that the selected service provides the best trade-off for the abstract *MLaaS* service in the given user context. The algorithm starts by initialising a list S that contains all the services to be evaluated. It then uses a nested loop structure where each service C_i in the list is compared against every other service C_j . For each comparison, the algorithm checks if the service C_i “dominates” service C_j . In this context, “dominates” means that C_i is at least as good as C_j in all relevant attributes (such as quality, cost, or efficiency) and better in at least one attribute. If C_i dominates C_j , then C_j is removed from the list S . Ultimately, the algorithm returns the refined list S , which represents the skyline set of services.

4.3 Experiment and Results

We develop a Python environment on a Windows 11 system with 8 *CPU* cores and 500 *GB* of storage for experiments. *Table 4.1* presents the details of the experiment set up where it evaluated *MLaaS* service selection involving ten users with five distinct contexts (i.e., health conditions and environmental factors) on 500 *MLaaS* service descriptions. The focus is on assessing services based on two *QoS* preferences: biases and explainability (XAI), along with accuracy, response time (up to 5,000ns per 1,000 samples), and availability.

Table 4.1: Statistics of Dataset for Context-Aware *MLaaS* in IoT

Statistics	Values	Statistics	Values
Users	10	User Contexts	5
<i>MLaaS</i> Services	5000	IoT Devices	10
User Preferences	2	Response Time Range	5000ns
(<i>QoS</i>) Attributes	5	Accuracy Range	80-99%
XAI	0-1	Bias	0-1%

4.3.1 Data Set Description

Our experimental setup integrated data from ten IoT devices across three diverse datasets: the Smart Human Fall Dataset from Kaggle [[Sakib, 2024](#)], used to monitor physical movements; the CAUCAFall dataset [[Eraso et al., 2022](#)] focused on detecting object interactions such as

picking an object or potential theft incidents; and the Synthetic Network Traffic Dataset [Waghela, 2024] employed to analyse network behaviour for identifying security threats and unusual activities. Due to the unavailability of public MLaaS service description datasets, we curated service descriptions from leading providers such as AWS Sagemaker, Google Cloud, and Azure Machine Learning [AltexSoft, n.d.]. This dataset facilitated our analysis of mapping user contexts to abstract MLaaS services. Additionally, leveraging a comprehensive knowledge dataset [Patel et al., 2023], we optimised the selection of QoS for MLaaS concrete services derived from abstract MLaaS services. To evaluate each approach, we assessed several test cases (see examples in Table 4.2).

Table 4.2: Test cases of mapping MLaaS abstract and concrete service with QoS

No.	Examples of Test Cases
1.	Fall is normal and occurs indoors; prefer explainable MLaaS service.
2.	Intruders enter at night, intention unclear; prefer low-biased and explainable MLaaS.
3.	A cyber-attack entry point is via an unsecured password; prefer a more explainable MLaaS.
4.	Simultaneous anomalies detected in network behaviour across several users; low biases and explainable MLaaS service are needed.

4.3.2 Baseline Approaches

To evaluate the performance of the proposed *CAMSF*, we choose two service selection methods.

– **Pre-defined or Fixed-Rules** [Matos, 2020; Rhayem et al., 2021]: Traditional service selection system uses Event-Condition-Action (ECA) rules in the form of: "On <event>, If <condition>, Do <action> ", where on detecting an event, typically monitored through wearable sensors and smart devices. Later, validating the event against predefined criteria in a condition part and initiating the appropriate responses as an Action. This approach allows automated, consistent decision-making based on monitored events that match the conditions and trigger the corresponding actions.

– **Brute-Force Approach** [Garba, Mohamad, & Saadon, 2022]: The brute-force approach comparing candidate services involves evaluating every possible pair of services to determine how closely they match their attributes. This method ensures a thorough assessment by

comparing each service with every other candidate, providing comprehensive results; this approach can be computationally intensive due to the exhaustive nature of the comparisons, particularly when dealing with many candidate services.

Evaluations Metrics: We evaluate each service selection approach using below key performance metrics: accuracy, precision, recall and F1-score.

$$Precision = TP / (TP + FP)$$

$$Recall = TP / (TP + FN)$$

$$Accuracy = (TP + TN) / (TP + FP + TN + FN)$$

$$F1 - Score = 2 * Precision * \frac{Recall}{Precision + Recall}$$

where TP (True Positive) represent the correctly identified optimal MLaaS service, FP (False Positive) represents the incorrectly identified optimal MLaaS service, TN (True Negative) represents the correctly identified as not optimal MLaaS service, and FN (False Negative) represents the incorrectly identified as not optimal MLaaS service.

4.3.2 Performance Analysis

We evaluate the different selection approaches by two sets of experiments. First, we evaluate the efficiency of CAMSF in user context detection regarding how accurately SVM detects context. Second, we measure the effectiveness of CAMSF by comparing the accuracy and F1-score of each approach for selecting MLaaS based on user context. We also evaluate the scalability of each approach by measuring the computation time from user context detection to the MLaaS concrete service. During the evaluation, we vary the size of the MLaaS services description. Starting with 10-20 MLaaS service descriptions, we incrementally increased the samples across different experimental phases.

The effectiveness of our framework in identifying user context is demonstrated by the correlation between the number of contextual data samples and accuracy scores (Fig 4.3(a)). Initially, with smaller numbers of contextual data, the accuracy score is relatively low, around 0.50. However, increasing the sample size enhances accuracy, ultimately reaching an impressive score of 0.90. This trend underscores the efficiency of our framework in accurately identifying user context, highlighting its robustness and reliability as more contextual data is incorporated. To evaluate CAMSF's efficiency against the fixed-rule

approach in abstract service mapping and select the best concrete service based on given preferences, we measure it in terms of the success rate that evaluates the system’s efficiency in selecting services to maximise expected rewards, reflecting its ability to identify optimal choices and enhance decision-making through observed rewards in interactions using below formula:

$$SuccessRate = \frac{Number\ of\ times\ optimal\ service\ selected}{Total\ number\ of\ interactions}$$

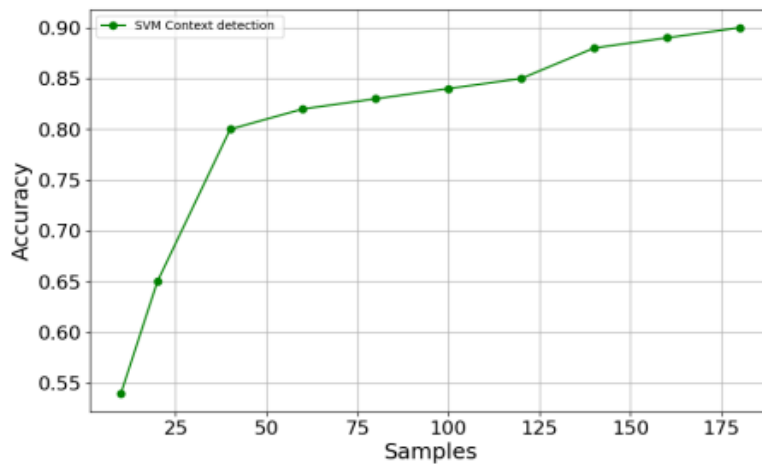
Our experimental results show that the CAMSF performs robustly in selecting optimal services based on user context (Table 4.3). Results indicate that with smaller samples, mapping accuracy is around 0.57 while increasing samples have an accuracy of 0.91(Fig 4.3(c)); the framework correctly identifies the optimal choice 91% of the time, indicating high reliability in aligning services with user context. Precision, at 0.90, signifies that when the framework selects a service, it is accurate 90% of the time, emphasising the correctness of positive predictions. A recall score of 0.93 highlights the framework’s ability to include 93% of all truly required services. The F1 score of 0.89 (Fig 4.3(d)) demonstrates a well-balanced performance between precision and recall. This could be attributed to success being driven by dynamic, learning-based strategies that continuously evolve the service selection process based on real-time feedback and context adaptability. This ensures that the CAMSF framework enhances long-term efficiency and effectiveness in service delivery.

In contrast to the CAMSF, the Fixed-Rule method demonstrates a low score in all metrics (Table 4.3) with an accuracy of 0.50 (Fig 4.3(c)) and an F1-score of 0.62 (Fig 4.3(d)) in several samples. These metrics underscore their dependence on predetermined rules designed for specific domains, ensuring predictability within those boundaries. However, these systems frequently encounter challenges adapting to changing contexts, resulting in inefficiencies when applied beyond their designated domains. In Brute-Force, despite its ability to measure similarity accurately, this approach (Table 4.3) has a lower accuracy of 0.45 (slowly increased

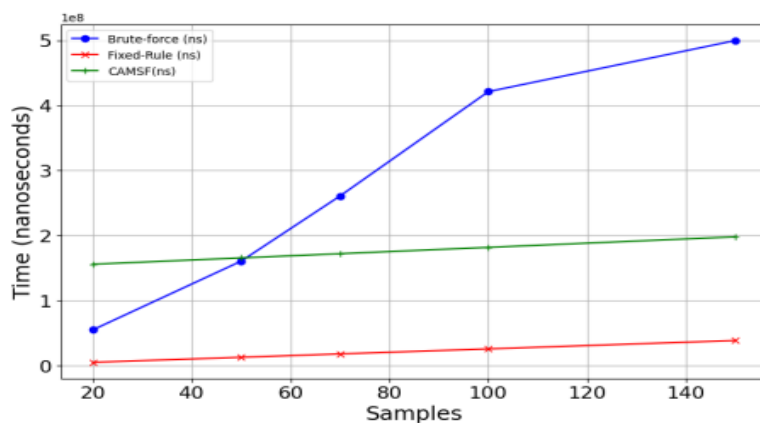
Table 4.3: Performance measurement results

	Accuracy	Precision	Recall	F1 Score
CAMSF	0.91	0.90	0.93	0.89
Fix Rule Approach	0.50	0.57	0.57	0.62
Brute Force Approach	0.45	0.34	0.35	0.34

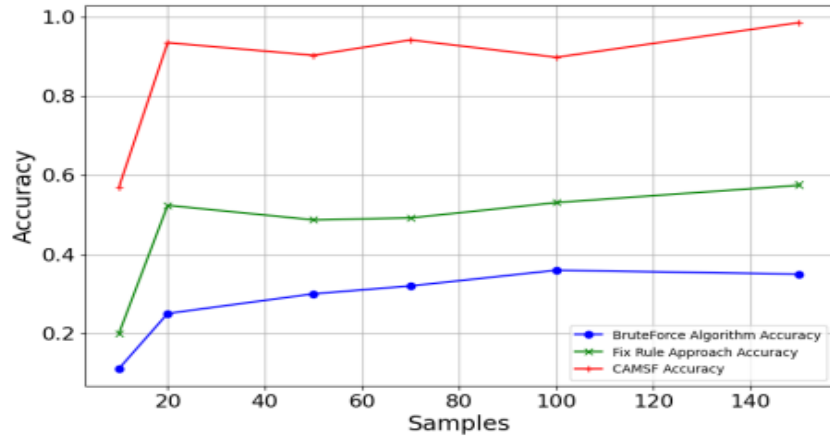
with samples; Fig 4.3(c)), indicating the correctness of its service predictions. However, with a precision of 0.34, it selects only a fraction of the services relevant to the user’s needs. The recall rate of 0.35 shows that it captures some relevant services, with an F1 score of 0.34 (Fig 4.3(d)), which assesses its effectiveness overall. This limitation makes it impractical for large-scale applications requiring quick service selection. Similarly, we started with smaller samples to evaluate the *scalability* by measuring the computation time of different approaches for selecting MLaaS services based on user context and increased slowly. Figure 4.3(b) illustrates the average computation time for all approaches.



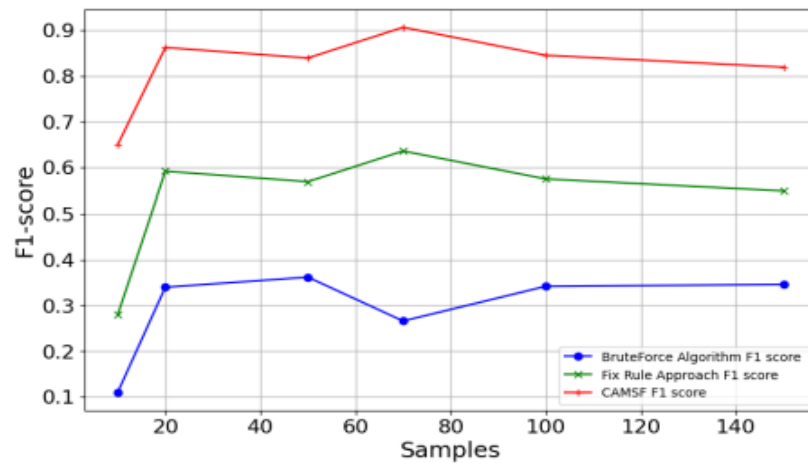
(a)



(b)



(c)



(d)

Fig. 4.3: The effectiveness and scalability of CAMSF (a) SVM context detection accuracy, (b) Computation Time Comparison, (c) Accuracy Comparison, and (d) F1 Score Comparison Across Models

Initially, the Fixed-Rule and Brute-Force approaches demonstrated significantly shorter service selection times with smaller datasets. However, with increased samples, the computing time for the Brute-Force algorithm increased by five nanoseconds. Conversely, the Fixed-Rule approach maintained a consistently low computation time of around 0.4 nanoseconds. This time efficiency is attributable to its reliance on pre-defined criteria for service selection. However, our proposed CAMSF, which dynamically identifies user context and maps it to MLaaS abstract services, showed a computation time of 1.5 to 2 nanoseconds with increased samples. Although slightly higher than the fixed rule, CAMSF optimises

service selection based on evolving contextual factors, ensuring timely and accurate alignment with user preferences while enhancing overall scalability and responsiveness.

4.4 Conclusion

The proposed context-aware MLaaS service selection framework (CAMSF) significantly enhances the optimisation of selecting the best Quality of Service (QoS) MLaaS offerings based on user context. CAMSF uses Support Vector Machines (SVM) with semi-supervised learning to detect and adapt to changing user contexts, ensuring responsiveness to diverse needs. We then utilise contextual bandits to map user contexts to abstract MLaaS services, enabling dynamic optimisation that improves service delivery in machine learning applications. This approach continuously refines service selection to maintain high performance in evolving environments. To validate CAMSF, we incorporate a skyline query method to select optimal concrete MLaaS services, aligning selections with user preferences while ensuring scalability, runtime efficiency, and effectiveness. Our results show a 61.65% performance improvement in context-aware MLaaS selection compared to the Fixed-Rule method and a 74.80% improvement compared to the Brute-Force approach. These findings underscore CAMSF's robust capabilities in various MLaaS service selection scenarios, highlighting its potential as a powerful tool for enhancing machine learning service delivery. Future work will focus on developing models that address sustainability issues, particularly cost efficiency and time optimisation.

CHAPTER 5

CONCLUSION AND FUTURE WORKS

MLaaS has become a widely popular form of cloud computing, which provides several advantages to consumers, such as reducing low maintenance costs for large-scale infrastructure and high scalability. There are two types of subscriptions offered by MLaaS providers: 1) pay-as-you-go and 2) a 12-month free trial. Large organisations tend to utilise MLaaS services for a long period of time, for instance, several months to years. The selection of MLaaS service is, therefore, a crucial decision for MLaaS users. Selection of MLaaS that does provide good performance can lead to significant financial for an organisation. MLaaS providers offer a free trial to MLaaS users. However, the selection of an MLaaS service based on a free trial is particularly challenging. This thesis tackles the major challenges involved in the selection of MLaaS services considering a user's preference for QoS attributes. The primary contribution of this research is divided into two main sections. In the first section, we propose a service selection framework to help a new user in MLaaS selection to address the two key challenges: 1) candidate MLaaS service provider selection for the free trials and 2) QoS performance variability. In the second section, we concentrate on context-awareness MLaaS selection. This part addresses the key challenge in selection: 1) adaptability to dynamic user contexts.

Chapter 3 proposes an MLaaS selection framework (MSF) to select the closest matched MLaaS provider according to a user's preference when QoS information is incomplete. The free trial offered by the MLaaS providers is used to discover both known and hidden aspects of their QoS performance. To select the optimal MLaaS service, enabling users to make informed decisions beyond solely relying on MLaaS advertisements from service providers. We propose integrating data from MLaaS advertisements, free trials and users' past experiences. Additionally, we identify biasness and explainability information from the knowledge graph to uncover hidden QoS information. Experiment results show that with complete QoS information, Spearman's Rank Correlation and Kendall's Tau coefficients can effectively provide more accurate results than scenarios with incomplete QoS information.

In Chapter 4, we extend the concept of MLaaS selection into context awareness. Here, we propose a context-aware MLaaS service selection framework (CAMSF) for selecting the best

QoS tailored to user context. A support vector machine (SVM) with semi-supervised learning to identify and adapt to changing user needs, ensuring responsiveness to varying requirements. We then leverage the contextual bandits approach to map these user contexts to abstract MLaaS services for dynamic optimisation. Finally, the framework incorporates a skyline query method to select the optimal concrete MLaaS service and user preferences while maintaining scalability, efficiency and effectiveness. Experimental results indicate that relying solely on MLaaS advertisements is insufficient for service selection, as it frequently results in incorrect decisions.

5.1 Discussion

This research centres on the selection of MLaaS services according to the user's QoS. It proposes a range of strategies to effectively utilise free MLaaS service trial periods in alignment with these QoS requirements. The concept of MLaaS advertisement, utilising free trials and other users' experiences are proposed ideas to tackle with the QoS performance for the MLaaS selection. Identifying hidden QoS information of MLaaS is also proposed using the concept of a knowledge base graph. Experimental results indicate that relying solely on MLaaS advertisements is insufficient for the selection as it results in incorrect decisions. We consider user contextual information when selecting an MLaaS concrete service based on changing user context. We propose approaches from identifying user context using SVM with semi-supervised learning to mapping user context in abstract MLaaS service and concrete MLaaS service using contextual bandit approach and skyline query method, respectively. Although this research helps a user make an informed MLaaS selection; however, several limitations could lead to new research directions. We provide a brief overview of the key limitations and potential future work below.

5.2 Limitations

To make an informed MLaaS selection using incomplete QoS information work, we assumed that the user-preferred MLaaS service would be available by offering a free trial. While this assumption is realistic since most of the MLaaS providers would offer free trials for the MLaaS services, however; if the service is not available for free trial, the proposed MLaaS selection framework would be unable to perform the selection.

To make an informed, context-aware selection of MLaaS work, we assumed that all the user wearable sensor data and QoS information would be available and accurate for selection. The

effectiveness of the proposed methods, including SVM-based context change analysis and skyline queries, largely depends on the quality and relevance of real-world datasets. In real-world scenarios, data may be incomplete or unavailable due to network issues, sensor malfunctions, or privacy restrictions. Incomplete data can affect the responsiveness of the context-aware selection of MLaaS, leading to inaccurate context recognition and suboptimal MLaaS concrete selection, which could be critical in dynamic environments where timely decisions are essential.

A significant limitation of these works is the reliance on synthetic datasets for experiments. We utilised publicly available Quality of Service of Web services datasets (QWS) to conduct the experiments due to the unavailability of suitable MLaaS quality of service datasets. This synthetic dataset is augmented with the values of bias and explainability (data taken from the knowledge base graph) to capture the characteristics of real MLaaS providers. Consequently, we believe that the key findings would be similar if the experiments were conducted using real datasets.

5.3 Future Work

The limitations of this work could lead to open new research directions. In the proposed incomplete QoS information, the MLaaS selection framework does not account for incoming user requests or qualitative user preferences in a natural language. A potential extension involves integrating a Conditional Preference Network (CP-Net) into the MLaaS selection process. CP-Nets allow users to articulate their preferences in a more qualitative manner, enabling the framework to make more informed selections of MLaaS based on these expressed preferences.

We then introduced a context-aware selection of MLaaS in the smart health monitoring scenario. This work does not adequately address how MLaaS services perform over time and how to solve issues related to service degradation. Also, there is a lack of focus on ethical issues such as transparency and explainability in MLaaS selection. By focusing on long-term sustainability and ethical integrity, the proposed framework will provide organisations with the tools necessary to select MLaaS providers that can maintain high standards of accuracy, transparency, and fairness. This is particularly vital in high-stakes environments such as healthcare, where the reliability of machine learning systems can directly impact patient outcomes.

To be specific, while current selection processes often emphasise short-term factors such as cost, functionality, and initial performance, they tend to overlook potential issues like service degradation and ethical concerns, which can significantly impact the reliability and trustworthiness of these services in the future. Key gaps in the literature include Service Degradation and Ethical concerns. Existing research does not adequately address how MLaaS services perform over time and how to solve issues related to service degradation, which can affect model accuracy and reliability. Also, there is a lack of focus on ethical issues such as transparency and explainability in MLaaS selection. For instance, understanding how ML models make decisions and ensuring that these decisions align with ethical standards is crucial for maintaining trust and accountability.

A possible extension of this research is the development of robust and ethically sound MLaaS solutions that can meet the evolving needs of various industries by guiding organisations in making better long-term decisions.

Appendix A

Copyright Information

In this appendix, the copyright agreements for the published paper, **Machine Learning as a Service (MLaaS) Selection with Incomplete QoS Information**, allow authors to reproduce an extract of all works in an MPhil thesis. The permission to reproduce published material was obtained from AIS eLibrary in which the author has published.

Copyright © 2023 Patel et al. This is an open-access article licensed under a Creative Commons Attribution-Non-Commercial 3.0 Australia License, which permits non-commercial use, distribution, and reproduction in any medium, provided the original author and ACIS are credited.



Notice

This is an older version of this license. Compared to previous versions, the 4.0 versions of all CC licenses are [more user-friendly and more internationally robust](#). If you are [licensing your own work](#), we strongly recommend the use of the 4.0 license instead: [Deed - Attribution-NonCommercial 4.0 International](#)

Canonical URL: <https://creativecommons.org/licenses/by-nc/3.0/au/>

[See the legal code](#)


You are free to:


Share — copy and redistribute the material in any medium or format

Adapt — remix, transform, and build upon the material

The licensor cannot revoke these freedoms as long as you follow the license terms.

Under the following terms:

 **Attribution** — You must give [appropriate credit](#), provide a link to the license, and [indicate if changes were made](#). You may do so in any reasonable manner, but not in any way that suggests the licensor endorses you or your use.

 **NonCommercial** — You may not use the material for [commercial purposes](#).

No additional restrictions — You may not apply legal terms or [technological measures](#) that legally restrict others from doing anything the license permits.

Notices:

You do not have to comply with the license for elements of the material in the public domain or where your use is permitted by an applicable [exception or limitation](#).

No warranties are given. The license may not give you all of the permissions necessary for your intended use. For example, other rights such as [publicity, privacy, or moral rights](#) may limit how you use the material.

Creative Commons is the nonprofit behind the open licenses and other legal tools that allow creators to share their work. Our legal tools are free to use.

Appendix B

Statement of Attribution

Chapter 3 is based on work published in a conference throughout the author's MPhil study. This chapter is a reproduction of a published manuscript, except for formatting consistency with the thesis.

The research presented in Chapter 3 was published in the ACIS 2023 Proceedings, 'Australasian Conferences on Information Systems (ACIS)', Wellington, 12-Dec-2023:

Patel, K., Mistry, S., Kanneganti, S. K. D., & Krishna, A. (2023). Machine Learning as a Service (MLaaS) Selection with Incomplete QoS Information. <https://aisel.aisnet.org/acis2023/39/>

All authors contributed to the revision of the manuscript and approved the final manuscript.

Appendix C

Research Output Authorship Attribution

Research Output Authorship Attribution

Title of Research Output: *Machine Learning as a Service (MLaaS) Selection with Incomplete QoS Information*

Field of Activity	Conception and Design	Acquisition of Data and Method	Data Conditioning and Manipulation	Analysis and Statistical Method	Interpretation and Discussion
Author 1: Keya Patel (Primary Author)	✓	✓	✓	✓	✓
Author 1 Acknowledgment: I acknowledge that these represent my contribution to the above research output, and I have approved the final version.					
Signed:					
Co-Author 2: Dr. Sajib Mistry (Primary Supervisor)	✓			✓	✓
Co-Author 2 Acknowledgment: I acknowledge that these represent my contribution to the above research output, and I have approved the final version.					
Signed:					
Co-Author 3: Deepak Kanneganti			✓		
Co-Author 3 Acknowledgment: I acknowledge that these represent my contribution to the above research output, and I have approved the final version.					
Signed:					
Co-Author 4: A/Prof Aneesh Krishna (Co-Supervisor)	✓				✓
Co-Author 4 Acknowledgment: I acknowledge that these represent my contribution to the above research output, and I have approved the final version.					
Signed:					

Title of Research Output: *Context-Aware Selection of Machine Learning as a Service (MLaaS) in IoT Environments*

Field of Activity	Conception and Design	Acquisition of Data and Method	Data Conditioning and Manipulation	Analysis and Statistical Method	Interpretation and Discussion
Author 1: Keya Patel (Primary Author)	✓	✓	✓	✓	✓
Author 1 Acknowledgment: I acknowledge that these represent my contribution to the above research output, and I have approved the final version.					
Signed:					
Co-Author 2: Dr. Sajib Mistry (Primary Supervisor)	✓			✓	✓
Co-Author 2 Acknowledgment: I acknowledge that these represent my contribution to the above research output, and I have approved the final version.					
Signed:					
Co-Author 3: Deepak Kanneganti			✓		
Co-Author 3 Acknowledgment: I acknowledge that these represent my contribution to the above research output, and I have approved the final version.					
Signed:					
Co-Author 4: A/Prof Aneesh Krishna (Co-Supervisor)	✓				✓
Co-Author 4 Acknowledgment: I acknowledge that these represent my contribution to the above research output, and I have approved the final version.					
Signed:					

Bibliography

Amazon Web Services, Inc. (n.d.). *Machine learning and artificial intelligence - Amazon Web Services*. Amazon Web Services. <https://aws.amazon.com/machine-learning/>

Amazon Web Services. (2023). *Internet of things services for sensors*. <https://docs.aws.amazon.com/whitepapers/latest/aws-overview/internet-of-things-services.html>

Adomavicius, G., & Tuzhilin, A. (2010). Context-aware recommender systems. In *Recommender systems handbook* (pp. 217-253). Boston, MA: Springer US.

Akter, S., Dwivedi, Y. K., Sajib, S., Biswas, K., Bandara, R. J., & Michael, K. (2022). Algorithmic bias in machine learning-based marketing models. *Journal of Business Research*, 144, 201-216.

Alibaba Cloud. (n.d.). Digital transformation for your business in Australia & New Zealand with Alibaba Cloud. *Alibaba Cloud*. <https://au.alibabacloud.com/>

AltexSoft. (n.d.). Comparing machine learning as a service: Amazon, Microsoft Azure, Google Cloud AI, IBM Watson. *AltexSoft*. <https://www.altexsoft.com/blog/comparing-machine-learning-as-a-service-amazon-microsoft-azure-google-cloud-ai-ibm-watson/>

Amazon Web Services, Inc. (n.d.). Image recognition software - ML image & video analysis - Amazon Rekognition - AWS. *Amazon Web Services*. <https://aws.amazon.com/rekognition/>

Barbierato, E., Vedova, M. L. D., Tessera, D., Toti, D., & Vanoli, N. (2022). A methodology for controlling bias and fairness in synthetic data generation. *Applied Sciences*, 12(9), 4619.

Baranwal, G., Singh, M., & Vidyarthi, D. P. (2020). A framework for IoT service selection. *The Journal of Supercomputing*, 76(4), 2777-2814.

Bhol, S. G., Mohanty, S., & Pattnaik, P. K. (2024). Machine Learning as a Service Cloud Selection: An MCDM Approach for Optimal Decision Making. *Procedia Computer Science*, 233, 909-918.

Chai, Z. Y., Du, M. M., & Song, G. Z. (2021). A fast energy-centered and QoS-aware service composition approach for Internet of Things. *Applied Soft Computing*, 100, 106914.

Cornacchia, G., Narducci, F., and Ragone-2, A. (n.d.). A General Model for Fair and Explainable Recommendation in the Loan Domain. (<https://doi.org/ceur-ws.org/Vol-2960>).

Das, D., Nishimura, Y., Vivek, R. P., Takeda, N., Fish, S. T., Ploetz, T., & Chernova, S. (2023). Explainable activity recognition for smart home systems. *ACM Transactions on Interactive Intelligent Systems*, 13(2), 1-39.

Demajo, L. M., Vella, V., & Dingli, A. (2020). Explainable ai for interpretable credit scoring. *arXiv preprint arXiv:2012.03749*.

Eraso, J. C., Muñoz, E., Muñoz, M., & Pinto, J. (2022). *Dataset Caucafall* (Version 4) [Data set]. Mendeley Data. <https://data.mendeley.com/datasets/7w7fccy7ky/4>

Fattah, S. M. M. (2021). *Long-term IaaS cloud service selection* (Ph.D. thesis). University of Melbourne.

Fattah, S. M. M., Bouguettaya, A., & Mistry, S. (2020). Long-term IaaS selection using performance discovery. *IEEE Transactions on Services Computing*, 15(4), 2129-2143.

Fortuna, C., Mušić, D., Cerar, G., Čampa, A., Kapsalis, P., & Mohorčič, M. (2023). On-premises artificial intelligence as a service for small and medium size setups. In *Advances in Engineering and Information Science Toward Smart City and Beyond* (pp. 53-73). Cham: Springer International Publishing.

Garba, S., Mohamad, R., & Saadon, N. A. (2022). Self-adaptive mobile web service discovery approach based on modified negative selection algorithm. *Neural Computing and Applications*, 34(3), 2007-2029.

Google Cloud. (n.d.). Google Cloud AutoML - Train models without ML expertise. *Google Cloud*. <https://cloud.google.com/automl>

Google Cloud. (n.d.). *Vertex AI* | *Google Cloud*. Google Cloud. <https://cloud.google.com/vertex-ai>

Google Cloud. (n.d.). *Vision AI* | *Cloud Vision API* | *Google Cloud*. Google Cloud. <https://cloud.google.com/vision>

Giommi, L. (2023). Machine learning as a service for high energy physics (MLaaS4HEP): a service for ML-based data analyses.

- Grigoriadis, I., Vrochidou, E., Tsiatsiou, I., & Papakostas, G. A. (2023, February). Machine learning as a service (MLaaS)—an enterprise perspective. In *Proceedings of International Conference on Data Science and Applications: ICDSA 2022, Volume 2* (pp. 261-273). Singapore: Springer Nature Singapore.
- Han, B., Wong, S., Mannweiler, C., Crippa, M. R., & Schotten, H. D. (2019). Context-awareness enhances 5G multi-access edge computing reliability. *IEEE Access*, 7, 21290-21299.
- Hosseinzadeh, M., Tho, Q. T., Ali, S., Rahmani, A. M., Souri, A., Norouzi, M., & Huynh, B. (2020). A hybrid service selection and composition model for cloud-edge computing in the Internet of Things. *IEEE Access*, 8, 85939-85949.
- Huang, A. F., Lan, C. W., & Yang, S. J. (2009). An optimal QoS-based Web service selection scheme. *Information Sciences*, 179(19), 3309-3322.
- Huang, J., Lan, Y., & Xu, M. (2018). A Simulation-Based Approach of QoS-Aware Service Selection in Mobile Edge Computing. *Wireless Communications and Mobile Computing*, 2018(1), 5485461.
- Hu, L., Yan, A., Yan, H., Li, J., Huang, T., Zhang, Y., & Yang, C. (2023). Defenses to membership inference attacks: A survey. *ACM Computing Surveys*, 56(4), 1-34.
- Javed, A. R., Fahad, L. G., Farhan, A. A., Abbas, S., Srivastava, G., Parizi, R. M., & Khan, M. S. (2021). Automated cognitive health assessment in smart homes using machine learning. *Sustainable Cities and Society*, 65, 102572.
- Jeong, Y., Son, S., & Lee, B. (2019). The lightweight autonomous vehicle self-diagnosis (LAVS) using machine learning based on sensors and multi-protocol IoT gateway. *Sensors*, 19(11), 2534.
- Jithendra, T., Khan, M. Z., Basha, S. S., Das, R., Divya, A., Chowdhary, C. L., ... & Alahmadi, A. H. (2024). A novel QoS prediction model for web services based on an adaptive neuro-fuzzy inference system using COOT optimisation. *IEEE Access*.
- Kang, G., Liang, B., Ding, L., Liu, J., Cao, B., & Kang, Y. (2024). QoS-aware web service recommendation via exploring the users' personalised diversity preferences. *Engineering Reports*, 6(1), e12695.

- Kertiou, I., Benharzallah, S., Kahloul, L., Beggas, M., Euler, R., Laouid, A., & Bounceur, A. (2018). A dynamic skyline technique for a context-aware selection of the best sensors in an IoT architecture. *Ad Hoc Networks*, *81*, 183-196.
- Kumar, R. R., Kumari, B., & Kumar, C. (2021). CCS-OSSR: a framework based on hybrid MCDM for optimal service selection and ranking of cloud computing services. *Cluster Computing*, *24*(2), 867-883.
- Kute, S. S., Shreyas Madhav, A. V., Kumari, S., & Aswathy, S. U. (2022). Machine learning-based disease diagnosis and prediction for E-healthcare system. *Advanced analytics and deep learning models*, 127-147.
- Liao, L., Wang, S., & Wu, J. (2023, May). Research on web service composition selection based on QoS metrics. In *2023 15th International Conference on Advanced Computational Intelligence (ICACI)* (pp. 1-8). IEEE.
- Lou, Q., & Jiang, L. (2019). She: A fast and accurate deep neural network for encrypted data. *Advances in neural information processing systems*, *32*.
- Lupo, T. (2016). A fuzzy framework to evaluate service quality in the healthcare industry: An empirical case of public hospital service evaluation in Sicily. *Applied Soft Computing*, *40*, 468-478.
- Makhluhian, M., Hashemi, S. M., Rastegari, Y., & Pejman, E. (2012). Web service selection based on ranking of qos using associative classification. *arXiv preprint arXiv:1204.1425*.
- Malekipirbazari, M., & Aksakalli, V. (2015). Risk assessment in social lending via random forests. *Expert Systems with Applications*, *42*(10), 4621-4631.
- Malhi, A., Knapic, S., & Främling, K. (2020). Explainable agents for less bias in human-agent decision making. In *Explainable, Transparent Autonomous Agents and Multi-Agent Systems: Second International Workshop, EXTRAAMAS 2020, Auckland, New Zealand, May 9–13, 2020, Revised Selected Papers 2* (pp. 129-146). Springer International Publishing.
- Mariani, S., Zambonelli, F., Tenyi, A., Cano, I., & Roca, J. (2019, June). Risk prediction as a service: a DSS architecture promoting interoperability and collaboration. In *2019 IEEE 32nd International Symposium on Computer-Based Medical Systems (CBMS)* (pp. 300-305). IEEE.
- Matos, É. D. (2020). Edge-centric context sharing architecture for the internet of things: context interoperability and context-aware security.

- Milton, R., & Roumpani, F. (2019, May). Accelerating urban modelling algorithms with artificial intelligence. In *Proceedings of the 5th International Conference on Geographical Information Systems Theory, Applications and Management* (Vol. 1, pp. 105-116). INSTICC.
- Mohd, M., Javeed, S., Wani, M. A., Khanday, H. A., Wani, A. H., Mir, U. B., & Nasrullah, S. (2023). Poliweet—Election prediction tool using tweets. *Software Impacts*, 17, 100542.
- Patel, K., Mistry, S., Kanneganti, S. K. D., & Krishna, A. (2023). Machine Learning as a Service (MLaaS) Selection with Incomplete QoS Information.
- Pereira, I., Madureira, A., Bettencourt, N., Coelho, D., Rebelo, M. Â., Araújo, C., & de Oliveira, D. A. (2024, April). A Machine Learning as a Service (MLaaS) Approach to Improve Marketing Success. In *Informatics* (Vol. 11, No. 2, p. 19). MDPI.
- Peng, X., Chen, R., Yu, K., Ye, F., & Xue, W. (2020). An improved weighted K-nearest neighbor algorithm for indoor localization. *Electronics*, 9(12), 2117.
- Pop, D., Iuhasz, G., & Petcu, D. (2016). Distributed platforms and cloud services: Enabling machine learning for big data. *Data Science and Big Data Computing: Frameworks and Methodologies*, 139-159.
- Philipp, R., Mladenow, A., Strauss, C., & Völz, A. (2020, November). Machine learning as a service: Challenges in research and applications. In *Proceedings of the 22nd International Conference on Information Integration and Web-based Applications & Services* (pp. 396-406).
- Pugliese, R., Regondi, S., & Marini, R. (2021). Machine learning-based approach: Global trends, research directions, and regulatory standpoints. *Data Science and Management*, 4, 19-29.
- Purohit, L., & Kumar, S. (2016). Exploring K-means clustering and skyline for web service selection. In *2016 International conference on industrial information system (ICIIS)* (pp. 1-5).
- Ribeiro, M., Grolinger, K., & Capretz, M. A. (2015, December). MLaaS: Machine learning as a service. In *2015 IEEE 14th international conference on machine learning and applications (ICMLA)* (pp. 896-902). IEEE.

- Qi, L., Dou, W., Hu, C., Zhou, Y., & Yu, J. (2015). A context-aware service evaluation approach over big data for cloud applications. *IEEE Transactions on Cloud Computing*, 8(2), 338-348.
- Rhayem, A., Mhiri, M. B. A., Drira, K., Tazi, S., & Gargouri, F. (2021). A semantic-enabled and context-aware monitoring system for the internet of medical things. *Expert Systems*, 38(2), e12629.
- Sakib, S. (2024). *Smartphone accelerometer and gyroscope data for human activity recognition* [Data set]. Kaggle. <https://www.kaggle.com/datasets/saadmansakib/smartphone-human-fall-dataset>
- Sachan, S., Yang, J. B., Xu, D. L., Benavides, D. E., & Li, Y. (2020). An explainable AI decision-support-system to automate loan underwriting. *Expert Systems with Applications*, 144, 113100.
- Sahi, S. M. (2022). The artificial intelligence and its global economic growth impact. *World Economics and Finance Bulletin*, 9, 16-24.
- Shaw, M. J., & Gentry, J. A. (1988). Using an expert system with inductive learning to evaluate business loans. *Financial Management*, 45-56.
- Srimathi, H., & Krishnamoorthy, A. (2019). Personalization of student support services using chatbot. *International Journal of Scientific & Technology*, 8(9), 1744-1747.
- Sun, L., Dong, H., Hussain, F. K., Hussain, O. K., & Chang, E. (2014). Cloud service selection: State-of-the-art and future research directions. *Journal of Network and Computer Applications*, 45, 134-150.
- Tay, Y. H. (2019, May). XYLORIX: An AI-as-a-service platform for wood identification. In *Proceedings of the IAWA-IUFRO international symposium for updating wood identification, Beijing, China* (pp. 20-22).
- Tran, V. X., Tsuji, H., & Masuda, R. (2009). A new QoS ontology and its QoS-based ranking algorithm for Web services. *Simulation Modelling Practice and Theory*, 17(8), 1378-1398.
- Varatharajah, Y., & Berry, B. (2022). A contextual-bandit-based approach for informed decision-making in clinical trials. *Life*, 12(8), 1277.

Waghela, V. (2024). *Network traffic data* [Data set]. Kaggle. <https://www.kaggle.com/datasets/vidhikishorwaghela/synthetic-network-traffic/data>

Wang, H., Shao, S., Zhou, X., Wan, C., & Bouguettaya, A. (2009). Web service selection with incomplete or inconsistent user preferences. In *Service-Oriented Computing: 7th International Joint Conference, ICSOC-ServiceWave 2009, Stockholm, Sweden, November 24-27, 2009. Proceedings 2* (pp. 83-98). Springer Berlin Heidelberg.

Wang, Q. M., Tang, Y., & Zhang, Z. B. (2009, May). Research in enterprise applications of dynamic web service composition methods and models. In *2009 Second International Symposium on Electronic Commerce and Security* (Vol. 1, pp. 146-150). IEEE.

Wardani, B. S., Sa'adah, S., & Nurjanah, D. (2023). Measuring and Mitigating Bias in Bank Customers Data with XGBoost, LightGBM, and Random Forest Algorithm. *Jurnal Ilmiah Teknik Elektro Komputer dan Informatika (JITEKI)*, 9(1), 142-155.

Wu, D., Zhang, P., He, Y., & Luo, X. (2022). A double-space and double-norm ensembled latent factor model for highly accurate web service QoS prediction. *IEEE Transactions on Services Computing*, 16(2), 802-814.

Wu, H., Deng, S., Li, W., Yin, J., Li, X., Feng, Z., & Zomaya, A. Y. (2019, July). Mobility-aware service selection in mobile edge computing systems. In *2019 IEEE international conference on web services (ICWS)* (pp. 201-208). IEEE.

Xie, S., Xue, Y., Zhu, Y., & Wang, Z. (2022, May). Cost-effective MLaaS federation: A combinatorial reinforcement learning approach. In *IEEE INFOCOM 2022-IEEE Conference on Computer Communications* (pp. 2078-2087). IEEE.

Xu, B., Shen, H., Sun, B., An, R., Cao, Q., & Cheng, X. (2021, May). Towards consumer loan fraud detection: Graph neural networks with role-constrained conditional random field. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 35, No. 5, pp. 4537-4545).

Xu, Y., Yin, J., Deng, S., Xiong, N. N., & Huang, J. (2016). Context-aware QoS prediction for web service recommendation and selection. *Expert Systems with Applications*, 53, 75-86.

Ye, Z., Mistry, S., Bouguettaya, A., & Dong, H. (2014). Long-term QoS-aware cloud service composition using multivariate time series analysis. *IEEE Transactions on Services Computing*, 9(3), 382-393.

Zhang, C., Yu, M., Wang, W., & Yan, F. (2020). Enabling cost-effective, slow-aware machine learning inference serving on public cloud. *IEEE Transactions on Cloud Computing*, 10(3), 1765-1779.

Zeng, L., Benatallah, B., Ngu, A. H., Dumas, M., Kalagnanam, J., & Chang, H. (2004). QoS-aware middleware for web services composition. *IEEE Transactions on software engineering*, 30(5), 311-327.

Zheng, Z., Wu, X., Zhang, Y., Lyu, M. R., & Wang, J. (2012). QoS ranking prediction for cloud services. *IEEE transactions on parallel and distributed systems*, 24(6), 1213-1222.

Every reasonable effort has been made to acknowledge the owners of copyright material. I would be pleased to hear from any copyright owner who has been omitted or incorrectly acknowledged.