SCALABLE CLOUD-DRIVEN PIPELINE FOR MACHINE LEARNING ALGORITHMS DEPLOYMENT

Ranasinghe Weerakkodi Pubudu Mahesh Meththananda

219368R

Degree of Master of Science in Computer Science

Department of Computer Science and Engineering
Faculty of Engineering

University of Moratuwa
Sri Lanka

March 2025

SCALABLE CLOUD-DRIVEN PIPELINE FOR MACHINE LEARING ALGORITHMS DEPLOYMENT

Ranasinghe Weerakkodi Pubudu Mahesh Meththananda

219368R

Thesis/Dissertation submitted in partial fulfillment of the requirements for Master of Science in Computer Science specializing in Cloud Computing

Department of Computer Science and Engineering
Faculty of Engineering

University of Moratuwa
Sri Lanka

March 2025

DECLARATION

I declare that this is my own work, and this thesis/dissertation does not incorporate without acknowledgement any material previously submitted for a degree or diploma in any other University or Institute of higher learning and to the best of my knowledge and belief it does not contain any material previously published or written by another person except where the acknowledgement is made in the text.

Also, I hereby grant to the University of Moratuwa the non-exclusive right to reproduce and distribute my dissertation, in whole or in part, in print, electronic or other medium. I retain the right to use this content in whole or part in future works (such as articles or books).

Signature: Date: 10/07/2025

Name: Pubudu Mahesh Meththananda

The above candidate has carried out research for the Master of Science thesis under my supervision.

Signature of the supervisor: …… Date: …21/07/2025……

Name: Dr. Sunimal Rathnayake

ABSTRACT

As machine learning (ML) continues to permeate industry and research, the need for robust, scalable, and maintainable deployment pipelines has become critical. Despite advances in model development, operationalizing ML models remains a non-trivial challenge due to the complexity of managing infrastructure, ensuring reproducibility, and integrating with heterogeneous data sources and consumers. This thesis proposes a generalized, cloud-oriented deployment framework for ML algorithms, addressing these challenges through modular and extensible architecture.

The framework is built upon containerized microservices, enabling environment-agnostic deployment and seamless transition from local development to production. A core component of the system is its pluggable inference interface, which abstracts model-specific logic, allowing arbitrary ML algorithms to be integrated with minimal engineering overhead. The pipeline supports dynamic parameterization of models, automatic resource provisioning, and concurrent execution, promoting scalability and adaptability to varying workloads.

A lightweight, web-based interface facilitates user interaction with the system, supporting dataset upload, model parameter configuration, execution of inference tasks, and retrieval of outputs in multiple formats. Workflow orchestration, monitoring, and logging are integrated into the pipeline to ensure operational transparency, fault tolerance, and streamlined debugging. All system components are loosely coupled, promoting maintainability and compatibility with multiple cloud service providers and on-premises environments.

The generality of the framework is validated through the integration of the Growing Self-Organizing Map (GSOM) algorithm and deployment on Amazon Web Service (AWS), demonstrating its ability to support computationally intensive models while preserving system responsiveness and reliability. However, architecture is agnostic to specific ML methods or cloud platforms, making it applicable across a wide range of deployment scenarios. This work contributes to a reusable and scalable Machine Learning Operations (MLOps) aligned infrastructure that bridges the gap between ML development and production deployment, supporting continuous delivery, reproducibility, and operational efficiency in modern ML workflows.

Keywords: Machine Learning Deployment, Cloud Computing, MLOps, CI/CD Pipeline, RESTful APIs, Containerization, Scalable Architecture, Model Operationalization

ACKNOWLEDGMENT

I would like to express my gratitude to **Prof. Damminda Alahakoon (Professor- La Trobe University, Australia)** and Mr. Prasad Maduranga (PhD. Candidate- La Trobe University, Australia), for initiating this project and providing guidance and motivation throughout the development of this cloud-driven ML pipeline. Their vision to enhance the accessibility and scalability of the GSOM algorithm as a cloud service has shaped the objectives and implementation of this research.

I extend my appreciation to my supervisor, **Dr. Sunimal Rathnayake (Senior Lecturer- Department of Computer Science and Engineering, University of Moratuwa)**, for his continuous support, insightful feedback, and technical guidance. His expertise in cloud computing and ML deployment has been instrumental in refining the methodologies and ensuring the success of this project.

Furthermore, I acknowledge the University of Moratuwa for providing the necessary resources and academic support to undertake this research. Special thanks to my colleagues and others who contributed through discussions, testing, and valuable suggestions, which greatly improved the quality of this work.

Finally, I would like to express my gratitude to my family for their encouragement and unwavering support throughout this journey. Their belief in my abilities has been a constant source of motivation in successfully completing this project.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

Abbreviation	Description
AWS	Amazon Web Service
API	Application Programming Interface
BMU	Best Matching Unit
CD	Continuous Development
CI	Continuous Integration
DevOps	Developer Operations
ECR	Elastic Container Registry
GCP	Google Cloud Platform
GSOM	Growing Self-Organizing Map
HTTP	Hyper Text Transfer Protocol
IaaS	Infrastructure as a Service
IT	Information Technology
KMS	Key Management Service
ML	Machine Learning
MLOps	Machine Learning Operations
OCI	Open Container Initiative
OCI	Open Container Initiative
RBAC	Role-Based Access Control
REST API	Representational State Transfer Application Programming Interface
S3	Simple Storage Service
SOM	Self-Organizing Map
UI	User Interface