Instruction-Guided AI-Driven Predictive Analytics for User-Level CPU Optimization and Overutilization Detection in Cloud Infrastructure Management

R.M.C.P. Rathnayaka

(Reg. No.: MS24016070)

A THESIS

SUBMITTED TO

SRI LANKA INSTITUTE OF INFORMATION TECHNOLOGY

IN PARTIAL FULFILMENT OF THE REQUIREMENTS

FOR THE DEGREE OF

MASTER OF SCIENCE IN INFORMATION TECHNOLOGY

December 2025

I certify that I have read this thesis and that in my opinion it is fully adequate, in scope and in quality, as a thesis for the degree of Master of Science.

Prof. Pradeep Abeygunawardane

Approved for MSc. Research Project:

MSc in IT Programme Co-ordinator, SLIIT

Approved for MSc:

Head of Graduate Studies, FoC, SLIIT

DECLARATION

This is to certify that the work is entirely my own and not of any other person, unless explicitly acknowledged (including citation of published and unpublished sources). The work has not previously been submitted in any form to the Sri Lanka Institute of Information Technology or to any other institution for assessment for any other purpose.

Sign: 
R.M.C.P. Rathnayaka

Date: 30.12.2025

ABSTRACT

Instruction-Guided AI-Driven Predictive Analytics for User-Level CPU Optimization and Overutilization Detection in Cloud Infrastructure Management

R.M.C.P. Rathnayaka
MSc. In Information Technology
Supervisor: Prof. Pradeep Abeygunawardana
December 2025

Cloud computing provides scalable and cost-efficient infrastructure, yet effective CPU resource management continues to pose challenges. Overutilization of CPU resources can result in degraded performance, violations of service-level agreements (SLAs), and increased operational costs. While many existing solutions focus on optimizing system- or virtual machine-level performance, there has been limited exploration of user-level workload variability, where CPU demand often fluctuates significantly across different tenants. To address this gap, the study introduces an AI-driven predictive analytics framework designed to manage CPU overutilization at the user level in cloud environments. The framework combines machine learning with an instruction-guided decision engine that generates actionable optimization strategies for administrators, rather than relying on full automation. This human-in-the-loop approach enhances transparency and trust in operational decisions. The evaluation used two datasets, including synthetic workloads that represented light, medium, and heavy users, as well as real-world traces from the Bitbrains GWA-T-12 dataset. Random Forest and Logistic Regression models were trained using features such as average CPU utilization and provisioned capacity. CPU overutilization was defined as usage exceeding 90 percent of the allocated CPU. The results indicated that the Random Forest model achieved higher predictive accuracy, surpassing 90 percent, and reached an AUC value of 0.99, outperforming Logistic Regression. The decision engine then translated these predictions into optimization instructions, including actions like workload migration and CPU scaling. Overall, the findings demonstrate that combining user-level prediction with instruction-guided optimization improves CPU resource

management, reduces the risk of overutilization, and enhances cloud system performance. This research contributes a practical, lightweight solution that advances predictive cloud resource management while preserving administrative oversight and system interpretability.

Keywords:
Cloud Resource Optimization, Cloud Computing, User-Level CPU Optimization, Overutilization Prediction, Predictive Analytics, Random Forest, Instruction-Guided Optimization, Bitbrains Dataset.

ACKNOWLEDGEMENT

I would like to begin by expressing my heartfelt gratitude to my supervisor, Prof. Pradeep Abeygunawardana, Dean of the Faculty of Computing at SLIIT, for his continuous guidance, encouragement, and invaluable support throughout this research journey. His insightful feedback and constructive advice played a crucial role in shaping my work and ensuring its successful completion.

I am equally thankful to Prof. Dilshan De Silva, MSc in IT Programme Coordinator, and Dr. Prasanna S. Haddela, Dean of the Faculty of Graduate Studies, for their thoughtful guidance and academic leadership during the course of this degree. Their support and direction significantly enriched the quality and progress of this study.

I would also like to sincerely thank the evaluation panel members for their valuable feedback during the progress reviews. Their observations and suggestions helped me refine the scope, methodology, and presentation of this thesis.

A special note of appreciation goes to my family, especially my parents and loved ones, for their unwavering support, patience, and understanding. Their encouragement and sacrifices gave me the strength to persevere through the challenges of balancing academic and personal responsibilities.

Lastly, I am grateful to my colleagues and friends for their continuous motivation and assistance throughout this journey. This research would not have been possible without the collective support, guidance, and encouragement of all these individuals.

TABLE OF CONTENTS

List of Figures

List of Tables