



**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is entirely my own and not of any other person, unless explicitly acknowledged (including citation of published and unpublished sources).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

DECLARATION	ii
ABSTRACT.....	iii
ACKNOWLEDGEMENT	v
TABLE OF CONTENTS.....	vi
List of Figures	x
List of Tables	xi
Chapter 1 Introduction	1
1.1 Introduction	1
1.2 Background	1
1.3 Motivation of the Study.....	2
1.4 Problem Statement	3
1.5 Aim and Objectives.....	5
1.5.1 Aim	5
1.5.2 Objectives	5
1.6 Research Questions	7
1.7 Scope of the Study.....	9
1.7.1 Scope	9
1.7.2 Limitations.....	10
1.8 Significance of the Study	11
1.8.1 Academic Contributions	12
1.8.2 Industrial Relevance	12
1.8.3 Environmental and Sustainability Impact.....	13
1.8.4 Regulatory and Ethical Compliance	13
1.9 Thesis Structure.....	13
1.10 Summary	15
Chapter 2 Literature Review	16
2.1 Introduction	16
2.2 Cloud Resource Management and CPU Challenges	16
2.3 AI-Driven Cloud Resource Optimization	18
2.4 Predictive Analytics for Overutilization Forecasting.....	19
2.5 Dynamic Scaling and Workload Forecasting.....	21
2.6 Autonomous and Self-Healing Cloud Management	22
2.7 User-Level Workloads and the Noisy-Neighbor Problem.....	24
2.8 Instruction-Guided Optimization Approaches	26
2.9 Synthesis of Literature and Identified Gap	28

2.10 Novelty of the Study	29
2.11 Summary	31
Chapter 3 Methodology	33
3.1 Introduction	33
3.2 Research Design.....	33
3.3 Dataset Selection and Acquisition.....	34
3.3.1 Synthetic User-Level Dataset	35
3.3.2 Real-World Bitbrains Dataset.....	36
3.4 Data Preprocessing	36
3.4.1 Final Features	38
3.5 Features and Parameters.....	39
3.5.1 Feature Processing	40
3.6 Machine Learning Models	40
3.6.1 Random Forest Classifier	41
3.6.2 Logistic Regression	41
3.6.3 Gradient Boosting Classifier	42
3.7 Model Training and Evaluation.....	43
3.7.1 Training Setup	43
3.7.2 Evaluation Metrics.....	44
3.7.3 Threshold Adjustment	44
3.8 Instruction-Guided Optimization (Decision Engine)	44
3.8.1 Architecture and Workflow	45
3.8.2 Benefits of Instruction-Guided Optimization	46
3.8.3 Example Scenario	46
3.9 Tools and Environment	47
3.9.1 Development Stack.....	47
3.10 Ethical Considerations.....	49
3.11 Limitations	50
3.12 Chapter Summary.....	51
Chapter 4 Implementation.....	53
4.1 Introduction	53
4.2 System Architecture	53
4.3 Dataset Preparation	54
4.3.1 Synthetic Dataset Implementation.....	54
4.3.2 Bitbrains Dataset Implementation	55
4.4 Preprocessing and Feature Engineering	56
4.5 Model Development.....	58

4.5.1 Random Forest Implementation	58
4.5.2 Logistic Regression Implementation	59
4.5.3 Gradient Boosting Implementation	60
4.5.4 Model Pipeline.....	60
4.5.5 Training Process	60
4.5.6 Validation Strategy and Results	64
4.6 Decision Engine Implementation	64
4.6.1 Purpose and Role	65
4.6.2 Architecture and Flow	65
4.6.3 Optimization Rule Design	68
4.6.4 Output Examples	69
4.6.5 Scalability and Modularity	69
4.7 Simulation in CloudSim	70
4.7.1 Overview of CloudSim Toolkit	70
4.7.2 Simulation Setup and Configuration	71
4.7.3 Integration of Instruction-Guided Actions	71
4.7.4 Sample Implementation Snippet.....	72
4.7.5 Evaluation Metrics.....	73
4.7.6 Expected Simulated Outcomes.....	73
4.7.7 Key Insights from Simulation.....	74
4.7.8 Limitations of Simulation Environment.....	74
4.8 Code Snippets and Screenshots.....	74
4.8.1 Random Forest Model Training	75
4.8.2 ROC Curve Visualization.....	76
4.8.3 Decision Engine Snippet	77
4.9 Summary	78
Chapter 5 Results and Discussion.....	80
5.1 Introduction	80
5.2 Experimental Setup	80
5.2.1 Environment & Tools	80
5.2.2 Dataset Summary.....	81
5.3 Predictive Model Results	81
5.3.1 Random Forest.....	81
5.3.2 Logistic Regression	82
5.3.3 Gradient Boosting.....	82
5.4 Comparative Analysis	83
5.5 Validation Performance Comparison.....	84

5.6 Instruction-Guided Optimization Outcomes	84
5.6.1 Simulation in CloudSim revealed.....	84
5.7 Discussion of Findings	85
5.7.1 Alignment with Literature	85
5.7.2 Research Questions Answered	87
5.7.3 Comparison with Existing Literature	89
5.7.4 Practical Implications	90
5.7.5 Critical Reflections	90
5.7.6 Real-World Impact and Societal Contribution	91
5.8 Limitations of Results	92
5.8.1 Limited Dataset Scale.....	92
5.8.2 CPU-Only Resource Focus.....	93
5.8.3 Simulated Optimization, Not Real-Time Deployment	93
5.8.4 Static Rule-Based Optimization Logic	93
5.8.5 Model Scope	94
5.8.6 No Economic or Energy Cost Modeling	94
5.8.7 No Integration with Existing Cloud Management APIs.....	94
5.9 Summary	95
Chapter 6 Conclusion and Future Work	98
6.1 Introduction	98
6.2 Summary of the Study.....	98
6.3 Key Findings	99
6.4 Contributions of the Study	100
6.4.1 Limitations.....	101
6.5 Future Work	102
6.6 Conclusion.....	103
6.7 Final Reflection	104
6.8 Summary	105
References.....	106
Appendix A - Synthetic Data Generation	108
Appendix B - Bitbrains Setup	113
Appendix C - Decision Engine	117
Appendix D - Colab GUI.....	119
Appendix E - CloudSim Simulation	130
Appendix F - Screenshots	136

List of Figures

Figure 1: Research Work Flow	34
Figure 2: System Architecture Diagram	54
Figure 3: CloudSim CPU provisioning.....	72
Figure 4: Workload Migration	72
Figure 5: Random Forest Model Training	75
Figure 6: ROC curve generation	76
Figure 7: ROC curve - Random Forest.....	76
Figure 8: Decision Engine	77
Figure 9: Synthetic Data Generation.....	112
Figure 10: Bitbrains setup.....	116
Figure 11: Decision Engine	118
Figure 12: Streamlit interface setup.....	129
Figure 13: CloudSim Setup.....	135
Figure 14: Colab GUI	136
Figure 15: Confusion Matrix	136

List of Tables

Table 1: Key Columns	38
Table 2: RandomForest Results in Validation	64
Table 3: Optimization Rules	68
Table 4: Evaluation Metrics.....	73
Table 5: Expected Simulated Outcomes	73
Table 6: Dataset Summary.....	81
Table 7: Random Forest Metrics.....	82
Table 8: Logistic Regression Metrics	82
Table 9: Gradient Boosting Metrics.....	83
Table 10: Comparative Analysis.....	83
Table 11: Validation Performance Comparison	84
Table 12: Comparison with Existing Literature.....	89
Table 13: Summary of the Key Limitations	95