

**MACHINE LEARNING-BASED PREDICTION OF METFORMIN DISSOLUTION
RATES USING MICRONIR SPECTRAL DATA AND COMPUTATIONAL
ANALYSIS**

By

Antoinetta Kayla Ajani
12202005

BACHELOR'S DEGREE
in

Information Technology: Artificial Intelligence & Data Science
FACULTY OF ENGINEERING & INFORMATION TECHNOLOGY



SWISS GERMAN UNIVERSITY
The Prominence Tower
Jalan Jalur Sutera Barat No. 15, Alam Sutera
Tangerang, Banten 15143 - Indonesia

Revision after the Thesis Defense on 10 June 2026

STATEMENT BY THE AUTHOR

I hereby declare that this submission is my own work and to the best of my knowledge, it contains no material previously published or written by another person, nor material which to a substantial extent has been accepted for the award of any other degree or diploma at any educational institution, except where due acknowledgement is made in the thesis.


Antoinetta Kayla Ajani

Student

19 June 2026

Date

Approved by:

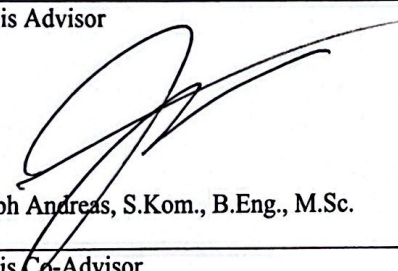


Dr. Ir. Eka Budiarto, S.T., M.Sc.

Thesis Advisor

19 June 2026

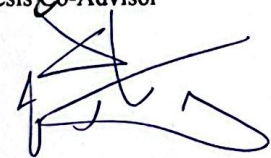
Date


Joseph Andreas, S.Kom., B.Eng., M.Sc.

Thesis Co-Advisor

23 June 2026

Date


Dr. Maulahikmah Galinium, S.Kom, M.Sc

Dean

23.6.2026.

Date

Antoinetta Kayla Ajani

ABSTRACT

MACHINE LEARNING-BASED PREDICTION OF METFORMIN DISSOLUTION RATES USING MICRONIR SPECTRAL DATA AND COMPUTATIONAL ANALYSIS

By

Antoinetta Kayla Ajani

Dr. Ir. Eka Budiarto, S.T., M.Sc., Advisor

Joseph Andreas, S.Kom., B.Eng., M.Sc., Co-Advisor

SWISS GERMAN UNIVERSITY

The dissolution testing of metformin tablets at a nation-wide pharmaceutical industry relies on conventional laboratory analysis, which is slow and destructive. This research develops a machine learning model that predicts tablet dissolution rates from MicroNIR spectral data and integrates it into graphical user interfaces for non-technical users. The model was developed following the CRISP-DM methodology, using 900 spectral scans across 125 wavelengths (908.1–1676.2 nm), comparing tree-based ensembles and an Artificial Neural Network with several preprocessing variants, evaluated using a batch-stratified split with per-fold feature selection, and tuned using Bayesian optimization and Grid Search. The model was developed in both MATLAB and Python, and integrated into a MATLAB desktop application and a Python web application. The MATLAB implementation achieved a test R^2 of 0.93 and MAPE of 0.31% at the 45-minute regulatory time point using Boosted Trees with PCA, while Python achieved R^2 of 0.98 and MAPE of 0.15% using ANN with SNV. Both models showed higher error at the early time points (5–20 minutes) and MAPE below 3% at the 45-minute time point on new validation data. This research concludes that MicroNIR spectral data combined with machine learning provides a viable alternative for fast and non-destructive dissolution prediction.

Keywords: Machine Learning, Dissolution Rate Prediction, MicroNIR Spectral Data, Metformin, Pharmaceutical Quality Control, Regression Models, MATLAB, Python.



© Copyright 2026
by Antoinetta Kayla Ajani
All rights reserved

DEDICATION

This thesis is dedicated to my family, whose unwavering support, patience, and encouragement have carried me throughout my academic journey. I also dedicate this work to the people around me, including my lecturers and friends, who continued to believe in me during my most difficult moments and reminded me to keep going. Their presence and faith have been an essential part of completing this study.

I also dedicated this work to the pharmaceutical industry and everyone striving to make drug quality testing faster and more efficient. It is my hope that this research, by offering a more practical alternative method to conventional dissolution testing, may contribute to advancing pharmaceutical quality control for the better.

ACKNOWLEDGEMENTS

First and foremost, I would like to thank God for His blessings and guidance, which gave me the strength and perseverance to complete this thesis.

My heartfelt thanks go to my family, especially for my parents for their endless support, prayers, understanding, and encouragement throughout this journey. Their support gave me the motivation and grit to continue even during the most challenging parts.

I would like to express my deepest gratitude to my thesis advisor, Dr. Ir. Eka Budiarto, S.T., M.Sc., and my co-advisor, Joseph Andreas, S.Kom., B.Eng., M.Sc., for their continuous guidance, patience, and valuable feedback throughout the completion of this research. Their willingness in providing direction and encouragement has truly been invaluable.

I am also sincerely grateful to partner pharmaceutical company for allowing this research to be conducted, and to all the mentors there for their support. I would also like to thank all my friends; their presence made the process all the more meaningful.

Finally, I acknowledge that this thesis would not have been completed without the help, guidance, and support of many people. Although I may not be able to mention everyone individually, I am truly grateful for every contribution, advice, and encouragement that helped me reach this point.

TABLE OF CONTENTS

STATEMENT BY THE AUTHOR.....	Error! Bookmark not defined.
ABSTRACT.....	3
DEDICATION	5
ACKNOWLEDGEMENTS.....	6
TABLE OF CONTENTS.....	7
LIST OF FIGURES.....	9
LIST OF TABLES	11
CHAPTER 1 – INTRODUCTION.....	13
1.1 Background	13
1.2 Problem Statement	14
1.3 Research Objectives.....	14
1.4 Scope and Limitations.....	15
1.5 Significance of the Study	15
1.6 Research Questions	16
1.7 Hypothesis	16
CHAPTER 2 – LITERATURE REVIEW.....	17
2.1 Pharmaceutical Theoretical Knowledge	17
2.1.1 Metformin: properties and pharmaceutical importance	17
2.1.2 Dissolution rate: concept, testing, and thresholds.....	17
2.1.3 Near-Infrared Spectroscopy and MicroNIR.....	18
2.2 Machine Learning Foundation	18
2.2.1 Random Forest Regression	19
2.2.2 Boosting-Based Regression	20
2.2.3 Artificial Neural Network	20
2.2.4 Feature Selection and Feature Importance Analysis	21
2.3 Development and Tools.....	21
2.3.1 Feature Selection and Feature Importance Analysis	21
2.3.2 Python for machine learning and automation	22
2.3.3 Figma as UI/UX Design Tools.....	22
2.4 Methodology Knowledge	22
2.4.1 CRISP-DM as Machine Learning Methodology	22
2.4.2 Waterfall SDLC as Software Development Methodology.....	23
2.5 Previous Studies	24
2.5.1 Machine Learning-Enabled NIR Spectroscopy for Dissolution Rate Prediction.....	24
2.5.2 NIR-Based Prediction of Metformin Tablet Dissolution.....	25
2.5.3 Spectral Data Processing, Portable NIR, and Computational Tools.....	25

2.6	Research Gap and Novelty	28
CHAPTER 3 – RESEARCH METHODS		29
3.1	Implementation of CRISP-DM as Machine Learning Methodology	29
3.1.1	Business Understanding.....	29
3.1.2	Data Understanding	30
3.1.3	Data Preparation	31
3.1.4	Data Preparation	31
3.1.5	Evaluation.....	32
3.1.6	Deployment.....	32
3.2	Implementation of Waterfall SDLC as Software Development Methodology.....	32
3.2.1	Requirement Analysis	33
3.2.2	System Design.....	34
3.2.3	Software Development.....	42
3.2.4	Software Testing.....	43
3.3	Research Timeline	44
CHAPTER 4 – RESULT AND DISCUSSION		45
4.1	Machine Learning Results – MATLAB	45
4.1.1	Data Understanding	45
4.1.2	Data Preparation and Preprocessing.....	46
4.1.3	Modelling and Evaluation.....	49
4.1.4	Evaluation.....	76
4.1.5	Deployment.....	83
4.1.6	Summary	83
4.1.7	Key Insights.....	84
4.2	Software Development Results – MATLAB	85
4.2.1	Implemented Interface	85
4.2.2	Software Testing Result.....	89
4.3	Machine Learning Results – Python	91
4.3.1	Model Training and Evaluation.....	92
4.3.2	Evaluation.....	100
4.3.3	Deployment.....	107
4.3.4	Summary	108
4.3.5	Key Insights.....	109
4.4	Software Development Result – Python.....	110
4.4.1	Implemented Features.....	110
4.5	Key Insights and Comparison Across Implementations.....	113
4.5.1	Cross-Platform Insights	113
4.5.2	Platform Comparison.....	114
CHAPTER 5 – CONCLUSION AND RECOMMENDATIONS		117
5.1	Conclusion.....	117
5.2	Future Recommendations	118
GLOSSARY		120
REFERENCES		122
CURRICULUM VITAE		124

LIST OF FIGURES

Figure 3.1: Existing Business Process for Dissolution Testing	29
Figure 3.2: Proposed Business Process Using MicroNIR and Machine Learning	30
Figure 3.3: Use Case Diagram – MATLAB	35
Figure 3.4: System Architecture Diagram	36
Figure 3.5: Workflow Diagram - MATLAB	37
Figure 3.6: Software Architecture Design - MATLAB	37
Figure 3.7: Figma Design – MATLAB.....	38
Figure 3.8: Use Case Diagram – Python Machine Learning GUI	39
Figure 3.9: System Architecture Diagram – Python Machine Learning GUI	40
Figure 3.10: Workflow Diagram – Python Machine Learning GUI.....	41
Figure 3.11: Software Architecture Design – Python ML API and Full-Stack GUI.....	41
Figure 3.12: Figma Design – GUI Design for Python Machine Learning.....	42
Figure 4.1: Raw MicroNIR Spectral Data of 10 Random Rows	46
Figure 4.2: Correlation Between Wavelengths and Dissolution Real Time Points	47
Figure 4.3: Correlation Amongst Top Predictors	48
Figure 4.4: Correlation Amongst Dissolution Time Points	49
Figure 4.5: Combined Dissolution Profile - Boosted Trees SNV in Baseline Non-PCA model on Scenario 1 with Data Splitting	54
Figure 4.6: Scatter Plot - Boosted Trees SNV in Baseline Non-PCA model on Scenario 1 with Data Splitting	55
Figure 4.7: Combined Dissolution Profile - Boosted Trees in Baseline Non-PCA Model on Scenario 1 without Data Splitting.....	57
Figure 4.8: Scatter Plot - Boosted SNV in Baseline Non-PCA Model on Scenario 1 without Data Splitting	58
Figure 4.9: Combined Dissolution Profile - Boosted Trees in Baseline Non-PCA Model on Scenario 2 with Data Splitting	62
Figure 4.10: Scatter Plot - Boosted SNV in Baseline Non-PCA Model on Scenario 2 with Data Splitting	63
Figure 4.11: Combined Dissolution Profile – ANN in Baseline Non-PCA Model on Scenario 3 with Data Splitting	67
Figure 4.12: Scatter Plot – ANN in Baseline Non-PCA Model on Scenario 3 with Data Splitting.....	68
Figure 4.13: Combined Dissolution Profile – ANN in Baseline Non-PCA Model on Scenario 3 without Data Splitting	70
Figure 4.14: Scatter Plot – ANN in Baseline Non-PCA Model on Scenario 3 without Data Splitting.....	70
Figure 4.15: Combined Dissolution Profile – Boosted Trees Moving Average (Window 7) with PCA after Grid Search Tuning on Scenario 3.....	75
Figure 4.16: PCA Loading Weights of PC1 and PC2 for the Final MATLAB Model.....	77
Figure 4.17: SHAP Summary of the Final Model (per Time Point).....	78

Figure 4.18: Permutation Importance of the Final Model (per Time Point).....	79
Figure 4.19: Predicted Dissolution Profiles of the Final Model on the New Data (5 Batches)	81
Figure 4.20: PCA Score Scatter Plot of the Training and New Validation Data.....	83
Figure 4.21: Implemented GUI – Main Menu Page	86
Figure 4.22: Implemented GUI – Upload Confirmation Page.....	86
Figure 4.23: Implemented GUI – Run Prediction Model Page	87
Figure 4.24: Implemented GUI – PDF Report Example	88
Figure 4.25: Implemented GUI – List of Reports Page	88
Figure 4.26: Implemented GUI – Activity Log Page	89
Figure 4.27: Combined Dissolution Profile – ANN SNV with Data Splitting Python	95
Figure 4.28: Scatter Plot – ANN SNV with Data Splitting Python.....	96
Figure 4.29: Combined Dissolution Profile – ANN (MLP) with SNV after Bayesian Tuning	100
Figure 4.30: SHAP Summary of the Final Model (per Time Point) - Python.....	103
Figure 4.31: Permutation Summary of the Final Model (per Time Point) - Python	103
Figure 4.32: Individual Scan Comparison of the Training and New Validation Spectra on SNV–Python	107
Figure 4.33: Login Page of the Python Web Application.....	110
Figure 4.34: Role-Based Access Control Python Web Application.....	111
Figure 4.35: Prediction Page Displaying the Dissolution Profile and Status	111
Figure 4.36: Generated PDF Report from the Python Web Application.....	112
Figure 4.37: File Management Page	112
Figure 4.38: User Management Page.....	113
Figure 4.39: Log Activity Page.....	113
Figure 4.40: MATLAB and Python Comparison of Actual vs Predicted MAPE Value of the Final Model on the New Data.....	115

LIST OF TABLES

Table 2.1: Comparison of Related Studies	26
Table 2.2: Indicators for Method Selection Based on Previous Studies.....	28
Table 3.1: Research Timeline	44
Table 4.1. Comparison of the Three Experimental Scenarios	50
Table 4.2. Training Result Baseline Non-PCA Models on Scenario 1 with Data Splitting – Test R ²	52
Table 4.3. Training Result Baseline Non-PCA Models on Scenario 1 with Data Splitting – Test MAPE.....	52
Table 4.4. Training Result Baseline Non-PCA Models on Scenario 1 with Data Splitting – Validation R ²	53
Table 4.5. Training Result Baseline Non-PCA Models on Scenario 1 with Data Splitting – Validation MAPE.....	53
Table 4.6. Training Result Baseline Non-PCA Models on Scenario 1 without Data Splitting – Validation R ²	56
Table 4.7. Training Result Baseline Non-PCA Models on Scenario 1 without Data Splitting – Validation MAPE.....	56
Table 4.8: Performance Comparison of the Split and No-Split Variations in Scenario 1 (at the 45-minute time point)	59
Table 4.9: Training Result Baseline Non-PCA Models on Scenario 2 with Data Splitting – Test R ²	60
Table 4.10: Training Result Baseline Non-PCA Models on Scenario 2 with Data Splitting – Test MAPE.....	61
Table 4.11: Training Result Baseline Non-PCA Models on Scenario 3 with Data Splitting – Test R ²	64
Table 4.12: Training Result Baseline Non-PCA Models on Scenario 3 with Data Splitting – Test MAPE.....	65
Table 4.13: Training Result Baseline Non-PCA Models on Scenario 3 with Data Splitting – Validation R ²	65
Table 4.14: Training Result Baseline Non-PCA Models on Scenario 3 with Data Splitting – Validation MAPE.....	66
Table 4.15: Training Result Baseline Non-PCA Models on Scenario 3 without Data Splitting – Validation R ²	68
Table 4.16: Training Result Baseline Non-PCA Models on Scenario 3 without Data Splitting – Validation MAPE.....	69
Table 4.17: Performance Comparison of the Split and No-Split Variations in Scenario 3 (at the 45-minute time point).....	71
Table 4.18: Training Result with PCA on Scenario 3 with Data Splitting – Test R ²	72
Table 4.19: Training Result with PCA on Scenario 3 with Data Splitting – Test MAPE	73

Table 4.20: Hyperparameter Tuning Results (Bayesian Optimization and Grid Search) on Scenario 3.....	74
Table 4.21. Top 5 Most Important Wavelengths by SHAP and Permutation Importance (MATLAB Moving Average + PCA model, test set).....	79
Table 4.22: Actual vs Predicted MAPE Value of the Final Model on the New Data	81
Table 4.23: User Acceptance Testing Results	89
Table 4.24: Training Result Baseline Non-PCA Models Python - Test R^2	93
Table 4.25: Training Result Baseline Non-PCA Models Python - Test MAPE.....	93
Table 4.26: Training Result Baseline Non-PCA Models Python - Validation R^2	94
Table 4.27: Training Result Baseline Non-PCA Models Python - Validation MAPE.....	94
Table 4.28: Training Result with PCA Python – Test R^2	97
Table 4.29: Training Result with PCA Python – Test MAPE	97
Table 4.30: Hyperparameter Tuning Results (Bayesian Optimization and Grid Search) Python	98
Table 4.31. Selected Wavelengths per Time Point (Spearman $ r > 0.30$, Python non-PCA ANN-SNV model)	102
Table 4.32. Top 5 Most Important Wavelengths by SHAP and Permutation Importance (Python ANN-SNV model, test set).....	104
Table 4.33: Actual vs Predicted MAPE Value of the Final Model on the New Data – Python	107
Table 4.34. Strengths and Weaknesses of the MATLAB and Python Implementations.....	114

