

CONTENTS

	Page
ABSTRACT IN THAI	I
ABSTRACT IN ENGLISH	II
CONTENTS	V
LIST OF TABLES	XI
LIST OF FIGURES	XIV
CHAPTER	
I INTRODUCTION	1
1.1 Research Objectives	5
1.2 Scope and Limitations	5
1.3 Research Procedure	6
1.4 Expected Result	6
II LITERATURE REVIEW	7
2.1 History and Development of Linear Modeling	7
2.1.1 Simple Linear Modeling	7
2.1.2 Multiple Linear Modeling	8
2.1.3 Response Transformation	8
2.1.4 Classical Linear Modeling	9
2.2 Generalized Linear Model (GLM)	10
2.2.1 Random Component	10
2.2.2 Systematic Component	11
2.2.3 The Link Function	11
2.2.4 Parameter Estimation	12
2.3 Double Generalized Linear Model	14
2.3.1 Parameter Estimation for Double Generalized Linear Models	15

CONTENTS (Continued)

	Page
2.4 Generalized Additive Model	18
2.4.1 Penalized Estimation	19
2.5 Methods for Parameter Estimation	22
2.5.1 Smoothing Splines	22
2.5.2 Explainable Boosting Machine	22
2.5.3 XGBoost	23
2.5.4 Random Forest	24
2.5.5 Support Vector Regression	25
2.6 Model Evaluation	27
2.6.1 Point Prediction	27
Mean Absolute Error	27
Root Mean Squared Error	28
Coefficient of Determination	28
2.6.2 Likelihood Based	29
Akaike Information Criterion	29
Negative Log-Likelihood	29
2.6.3 Probabilistic Forecasts	30
Continuous Ranked Probability Score	30
Prediction Interval Coverage Probability at 95%	30
2.7 Related Researches	31
III RESEARCH METHODOLOGY	34
3.1 Phase I: Theoretical and Algorithmic Development	34
3.1.1 Conceptual Foundation: From Classical GAMs to DGAMs	34
Limitations of Classical GAMs	34
The DGAM Framework	35
3.1.2 Mathematical Formulations and Algorithmic Specifications	36
3.1.3 Theoretical Properties and Convergence	37

CONTENTS (Continued)

		Page
3.1.4	Modeling Framework Comparison	37
	GAM with Constant Dispersion	38
	DGAM with Flexible Dispersion	38
3.1.5	Summary of Phase I	39
3.2	Phase II: Computational Implementation and Empirical Validation . .	39
3.2.1	Step 1: Data Acquisition and Preparation	40
	Dataset Selection and Conceptual Framework	40
	Data Quality Assessment and Preprocessing	40
3.2.2	Step 2: Data Partitioning Strategy	41
3.2.3	Step 3: Feature Preprocessing	42
3.2.4	Step 4: Model Training Procedures	42
	GAM Training Protocol	42
	DGAM Training Protocol	43
3.2.5	Step 5: Hyperparameter Configuration	44
	Rationale for Fixed Hyperparameters	45
3.2.6	Step 6: Model Evaluation Framework	46
	Point Prediction Metrics	46
	Likelihood-Based Metrics	46
	Probabilistic Calibration Metrics	47
3.2.7	Step 7: Statistical Analysis and Result Compilation	48
3.2.8	Implementation and Computational Infrastructure	48
3.3	Research Framework Summary	49
IV	RESULTS	53
4.1	Overview and Structure	53
4.1.1	Structure of This Chapter	53
4.1.2	Purpose and Contribution	54
4.1.3	Relationship to Empirical Validation	55

CONTENTS (Continued)

	Page
4.1.4 Key Contributions	55
4.1.5 Reading Guide	56
4.2 Double Generalized Additive Models: Mathematical Framework . . .	56
4.2.1 Random Component	57
4.2.2 Mean Structure (Location Model)	57
4.2.3 Dispersion Structure (Scale Model)	58
4.2.4 Link Functions	58
4.2.5 Model Specification Summary	60
4.3 Smoothing Splines	60
4.3.1 Basis Representation and Matrix Formulation	61
4.3.2 Parameter Estimation for DGAM with Smoothing Splines . .	61
4.3.3 Algorithm: Classical Spline DGAM via Coupled PIRLS	62
4.4 Explainable Boosting Machine	64
4.4.1 EBM Model Structure	64
4.4.2 Key Features of EBM	65
4.4.3 Parameter Estimation for DGAM with EBM	65
4.4.4 Algorithm: EBM-DGAM via Alternating Boosting	66
4.5 XGBoost	67
4.5.1 XGBoost Model Structure	68
4.5.2 Objective Function	68
4.5.3 Second-Order Taylor Approximation	69
4.5.4 Parameter Estimation for DGAM with XGBoost	69
4.5.5 Algorithm: XGBoost-DGAM via Alternating Second-Order Boosting	70
4.6 Random Forest	72
4.6.1 Random Forest Model Structure	72
4.6.2 Random Forest Algorithm	72

CONTENTS (Continued)

	Page
4.6.3	Parameter Estimation for DGAM with Random Forest 74
4.6.4	Algorithm: Random Forest DGAM via Alternating Residual Forests 74
4.7	Support Vector Regression 76
4.7.1	SVR Formulation 76
4.7.2	Kernel Methods and Dual Formulation 77
4.7.3	SVR Model Structure 78
4.7.4	Parameter Estimation for DGAM with SVR 78
4.7.5	Algorithm: SVR-DGAM via Alternating Optimization 79
4.8	Results (Phase II): Model Implementation and Empirical Validation . . 82
4.8.1	Target Variables 83
4.8.2	Modeling Techniques 84
4.8.3	Data Splitting Ratios 84
4.8.4	Experimental Grid and Paired-Comparison Protocol 84
4.8.5	Overall Comparative Performance: DGAM versus GAM 85
4.8.6	Robustness Across Splitting Ratios 86
4.8.7	Heterogeneity Across Datasets 87
4.8.8	Most Frequently Selected Dispersion Technique Under DGAM 87
4.8.9	Magnitude of Improvements and Representative Examples 88
V	DISCUSSION AND CONCLUSION 91
5.1	Discussion 91
5.1.1	Smoothing Splines for DGAM 91
5.1.2	Explainable Boosting Machine (EBM) for DGAM 92
5.1.3	XGBoost for DGAM 92
5.1.4	Random Forest (RF) for DGAM 93
5.1.5	Support Vector Regression (SVR) for DGAM 93

CONTENTS (Continued)

	Page
5.1.6 Recommendations	94
5.2 Conclusion	95
REFERENCES	98
APPENDICES	
APPENDIX A EXAMPLE PYTHON CODE FOR DGAM IMPLEMENTATIONS . .	103
A.1 Fixed Hyperparameter Configuration (Used by All Snippets)	104
A.2 Classical Spline DGAM via coupled PIRLS (Algo- rithm 1)	105
A.3 EBM-DGAM via Alternating Boosting for Mean and Dispersion	107
A.4 Random Forest DGAM via Alternating Residual Forests	109
APPENDIX B MODEL PERFORMANCE EVALUATION	111
CURRICULUM VITAE	128

LIST OF TABLES

Table		Page
2.1	Common link functions in generalized linear models.	12
3.1	Data Splitting Configurations	41
3.2	Complete Model Configuration Matrix.	44
3.3	Fixed Hyperparameter Settings for All Experiments.	45
3.4	Comprehensive Evaluation Metrics.	48
4.1	Target Variables by Dataset and Life Cycle Stage.	83
4.2	Overall win-rate of DGAM relative to GAM across all paired comparisons ($n = 240$). For AIC, NLL, CRPS, MAE, and RMSE, lower is better. For R^2 and PICP95, higher is better.	86
4.3	DGAM win-rate (%) relative to GAM by train-validation-test split (each split has $n = 80$ paired comparisons).	86
4.4	DGAM win-rate (%) relative to GAM by dataset (each dataset has $n = 60$ paired comparisons).	87
4.5	Frequency with which each dispersion technique is selected as the best DGAM dispersion learner (selection by minimum NLL).	88
4.6	Effect-size summary for DGAM—GAM across all paired comparisons ($n = 240$). Negative values indicate improvements for metrics where lower is better.	89
4.7	Representative high-impact examples (largest improvements in NLL; most negative ΔNLL).	90
4.8	Representative failure-mode examples (worst ΔNLL ; DGAM degrades NLL relative to GAM).	90
B.1	Predictive performance of GAM, DGAM, and machine learning models for abnormal short-term variability (STV) across data splitting strategies using the Fetal Health Monitoring Dataset.	112

LIST OF TABLES (Continued)

Table		Page
B.2	Predictive performance of GAM, DGAM, and machine learning models for percentage of time with abnormal long-term variability across data splitting strategies using the Fetal Health Monitoring Dataset. . .	113
B.3	Predictive performance of GAM, DGAM, and machine learning models for mean short-term variability of fetal heart rate across data splitting strategies using the Fetal Health Monitoring Dataset.	114
B.4	Predictive performance of GAM, DGAM, and machine learning models for mean long-term variability of fetal heart rate across data splitting strategies using the Fetal Health Monitoring Dataset.	115
B.5	Predictive performance of GAM, DGAM, and machine learning models for Calories burned across data splitting strategies using lifestyle data collected for health recommendation systems and mobile applications.	116
B.6	Predictive performance of GAM, DGAM, and machine learning models for body fat percentage across data splitting strategies using lifestyle data collected for health recommendation systems and mobile applications.	117
B.7	Predictive performance of GAM, DGAM, and machine learning models for average heart rate maintained during the session across data splitting strategies using lifestyle data collected for health recommendation systems and mobile applications.	118
B.8	Predictive performance of GAM, DGAM, and machine learning models for Maximum heart rate recorded during a workout session across data splitting strategies using lifestyle data collected for health recommendation systems and mobile applications.	119

LIST OF TABLES (Continued)

Table	Page
B.9	Predictive performance of GAM, DGAM, and machine learning models for serum creatinine levels across data splitting strategies using the National Health Checkup dataset of South Korea (2002–2020). 120
B.10	Predictive performance of GAM, DGAM, and machine learning models for serum gamma-glutamyl transferase (GGT) levels across data splitting strategies using the National Health Checkup dataset of South Korea (2002–2020). 121
B.11	Predictive performance of GAM, DGAM, and machine learning models for serum aspartate aminotransferase (AST/SGOT) levels across data splitting strategies using the National Health Checkup dataset of South Korea (2002–2020). 122
B.12	Predictive performance of GAM, DGAM, and machine learning models for serum alanine aminotransferase (SGPT) levels across data splitting strategies using the National Health Checkup dataset of South Korea (2002–2020). 123
B.13	Predictive performance of GAM, DGAM, and machine learning models for life expectancy across data splitting strategies using the World Health Organization life expectancy dataset. 124
B.14	Predictive performance of GAM, DGAM, and machine learning models for adult mortality across data splitting strategies using the World Health Organization life expectancy dataset. 125
B.15	Predictive performance of GAM, DGAM, and machine learning models for under-five deaths across data splitting strategies using the World Health Organization life expectancy dataset. 126
B.16	Predictive performance of GAM, DGAM, and machine learning models for infant mortality outcomes across data splitting strategies using the National Health Checkup dataset of South Korea (2002–2020). 127

LIST OF FIGURES

Figure		Page
3.1	Overall Phase I Research Framework: Theoretical and Algorithmic Development.	50
3.2	Overall Phase II Research Framework: Model Implementation.	51