

ISTANBUL TECHNICAL UNIVERSITY ★ GRADUATE SCHOOL

**PENETRATION RATE OPTIMIZATION IN HETEROGENEOUS
FORMATIONS WITH SUPPORT VECTOR MACHINES METHOD**



Ph.D. THESIS

Korhan KOR

Department of Petroleum and Natural Gas Engineering

Petroleum and Natural Gas Engineering Programme

SEPTEMBER 2021

ISTANBUL TECHNICAL UNIVERSITY ★ GRADUATE SCHOOL

**PENETRATION RATE OPTIMIZATION IN HETEROGENEOUS
FORMATIONS WITH SUPPORT VECTOR MACHINES METHOD**



Ph.D. THESIS

**Korhan KOR
(505152501)**

Department of Petroleum and Natural Gas Engineering

Petroleum and Natural Gas Engineering Programme

Thesis Advisor: Assoc. Prof. Dr. Gürşat ALTUN

SEPTEMBER 2021

İSTANBUL TEKNİK ÜNİVERSİTESİ ★ LİSANSÜSTÜ EĞİTİM ENSTİTÜSÜ

**DESTEK VEKTÖR MAKİNESİ YÖNTEMİ İLE HETEROJEN
FORMASYONLARDA İLERLEME HIZI OPTİMİZASYONU**

DOKTORA TEZİ

**Korhan KOR
(505152501)**

Petrol ve Doğal Gaz Mühendisliği Anabilim Dalı

Petrol ve Doğal Gaz Mühendisliği Programı

Tez Danışmanı: Doç. Dr. Gürşat ALTUN

EYLÜL 2021

Korhan KOR, a Ph.D. student of ITU Graduate School student ID 505152501, successfully defended the dissertation entitled “PENETRATION RATE OPTIMIZATION IN HETEROGENEOUS FORMATIONS WITH SUPPORT VECTOR MACHINES METHOD”, which he prepared after fulfilling the requirements specified in the associated legislations, before the jury whose signatures are below.

Thesis Advisor : **Assoc. Prof. Dr. Gürşat ALTUN**
Istanbul Technical University

Jury Members : **Asst. Prof. Dr. Şenol YAMANLAR**
Istanbul Technical University

.....
Prof. Dr. Hanifi ÇOPUR
Istanbul Technical University

.....
Assoc. Prof. Dr. Şeyda ERTEKİN
Middle East Technical University

.....
Prof. Dr. Serhat AKIN
Middle East Technical University

Date of Submission : 30 July 2021
Date of Defense : 14 September 2021



To my beloved family,



FOREWORD

This dissertation represents the reflections of a Ph.D. journey initiated in September 2015 and finished in July 2021. Assoc. Prof. Dr. Gürsat ALTUN has been my supervisor since my master's education. I owe him many thanks for leading me towards high-quality science.

The following appreciation is for my mother and father. One of the most significant contributors to this dissertation is my parents, for sure. I don't know if I would have been the person I am today without you. I am eternally grateful to you for many things. I love you all so much and thank you for being there for me.

The study was mainly conducted in Istanbul Technical University - Department of Petroleum and Natural Gas Engineering, but with close interactions with Turkish Petroleum Corporation (TPAO). TPAO has supported this study by releasing data and granting permission to publish the findings of this thesis. Therefore, I wish to thank Melih Han BİLGİN, CEO; Enver ÇETİNKAYA, Head of Engineering Department; Onur ÇETİN, Well Design and Planning Manager; and Sezer SEVİNÇ, a dear colleague, who has a lot of support for the geomechanical studies.

The work of this dissertation intersects with various engineering disciplines. Therefore, I have relied on the expertise and guidance from respectable academicians. Thanks to Dr. Şeyda ERTEKİN and Dr. Şenol YAMANLAR for their precious assistance.

And finally, my lovely wife Ezgi BİLGİÇ KOR... Thank you for your patience, for your massive support, and for your warm love.

July 2021

Korhan KOR
(Petroleum and Natural Gas Engineer)

TABLE OF CONTENTS

	<u>Page</u>
FOREWORD	ix
TABLE OF CONTENTS	xi
ABBREVIATIONS	xiii
SYMBOLS	xv
LIST OF TABLES	xvii
LIST OF FIGURES	xix
SUMMARY	xxvii
ÖZET	xxix
1. INTRODUCTION	1
2. LITERATURE REVIEW	7
2.1 Drilling and Fluid Parameters Optimization	7
2.2 Hardware Design and Laboratory Experiments	9
2.3 Mathematical Optimization and Statistics	11
2.4 Data-Driven and Machine Learning Methods	14
2.5 ROP Optimization in Heterogeneous Formations.....	24
2.6 General Overview and Interpretations	26
3. STATEMENT OF THE PROBLEM AND SCOPE OF THE THESIS	29
4. BOURGOYNE AND YOUNG METHOD	33
4.1 Default BYM Parameters.....	34
4.1.1 Effect of drillability	34
4.1.2 Effect of normal compaction	34
4.1.3 Effect of under compaction	35
4.1.4 Effect of overbalance	35
4.1.5 Effect of weight on bit per bit diameter.....	36
4.1.6 Effect of rotary speed.....	37
4.1.7 Effect of bit tooth wear	38
4.1.8 Effect of hydraulics	39
4.2 Additional Features from Geomechanical Properties	40
4.2.1 Shale index.....	41
4.2.2 Uniaxial compressive strength.....	42
4.2.3 Brittleness index.....	42
4.2.4 Shear failure gradient.....	43
4.2.5 Torque	44
5. MACHINE LEARNING AND PREDICTIVE MODELS	47
5.1 Predictive Models.....	49
5.1.1 Multiple linear regression.....	49
5.1.2 Support vector machines	51
5.1.2.1 Support vector machines for classification	51
5.1.2.2 Kernel functions.....	54
5.1.2.3 Support vector regression	55
5.1.3 Artificial Neural Networks.....	57

5.1.4 Random Forests	60
5.2 Overfitting and Cross-Validation.....	61
5.3 Outlier Labeling and Standardization.....	62
5.3.1 Tukey's method	63
5.3.2 z-score analysis.....	63
5.3.3 Data normalization.....	64
5.4 Statistical Performance Measures	64
5.4.1 Root mean squared error	65
5.4.2 Correlation coefficient	65
5.4.3 Statistical significance test.....	65
5.5 Software and Toolboxes.....	66
5.5.1 R statistical programming language	66
5.5.2 IBM SPSS statistics	66
6. APPLICATION, RESULTS, AND DISCUSSIONS.....	67
6.1 Application.....	67
6.1.1 Data preparation	69
6.1.2 Data preprocessing.....	72
6.1.3 Feature selection	73
6.1.4 Model training and validation	74
6.1.5 Model testing and post-processing of the results.....	74
6.2 Results	74
6.2.1 Case 1: data taken from the literature (homogeneous).....	74
6.2.1.1 Case 1.1: calculations with data from Bourgoyne Jr. and Young Jr. (1974)	75
6.2.1.2 Case 1.2: calculations with data from Irawan et al. (2012).....	79
6.2.1.3 Case 1.3: calculations with data from Al-Betairi et al. (1988).....	83
6.2.2 Case 2: data taken from the well C-1	87
6.2.2.1 Case 2.1: calculations with data from F-4.....	87
6.2.2.2 Case 2.2: calculations with data from F-5.....	92
6.2.2.3 Case 2.3: calculations with data from F-6.....	96
6.2.3 Case 3: data taken from the well C-2	101
6.2.3.1 Case 3.1: calculations with data from F-4.....	101
6.2.3.2 Case 3.2: calculations with data from F-5.....	106
6.2.3.3 Case 3.3: calculations with data from F-6.....	111
6.3 Discussions and General Evaluations.....	116
7. CONCLUSIONS	119
REFERENCES	123
APPENDICES	139
APPENDIX A: Datasets Taken from the Literature	141
APPENDIX B: Derivation of Support Vector Machines	143
APPENDIX C: R Script for the Calculations	151
APPENDIX D: Graphical Results of Intermediate Steps	157
CURRICULUM VITAE	245

ABBREVIATIONS

AI	: Artificial Intelligence
ANFIS	: Adaptive Neuro-Fuzzy Inference Systems
ANN	: Artificial Neural Networks
ASCII	: American Standard Code for Information Interchange
BA	: Bat Algorithm
BHA	: Bottom Hole Assembly
BI	: Brittleness Index
BYM	: Bourgoyne and Young Method
CCS	: Confined Compressive Strength
CV	: Cross-validation
DOT	: Drill-off Test
ECD	: Equivalent Circulation Density
ELM	: Extreme Learning Machines
ES	: Expert Systems
GA	: Genetic Algorithms
IQR	: Interquartile Range
kNN	: k-Nearest Neighbors
LF	: Loss Function
LS	: Least-squares
ML	: Machine Learning
MLP	: Multi-layer Perceptron
MLR	: Multiple Linear Regression
MLU	: Mud Logging Unit
MNR	: Multiple Non-linear Regression
MSE	: Mechanical Specific Energy
OCSH	: Optimal Canonical Separating Hyperplane
OOB	: Out of Bag
PDC	: Polycrystalline Diamond Compact
QA	: Quality Assurance
QC	: Quality Control
RBF	: Radial Basis Functions
RF	: Random Forests
RMSE	: Root Mean Square Error
ROP	: Rate of Penetration
RPM	: Rotary Speed
SCI	: Science Citation Index
SFG	: Shear Failure Gradient
SI	: Shale Index
SV	: Support Vector
SVM	: Support Vector Machines
SVR	: Support Vector Regression
SWR	: Step-wise Regression
TCI	: Tungsten Carbide Insert
TRQ	: Torque

UCS	: Uniaxial (Unconfined) Compressive Strength
USA	: Upper-layer Solution-aware
VC	: Vapnik-Chervonenkis
WOB	: Weight on Bit



SYMBOLS

C	: Penalty parameter
C₀	: Unconfined compressive strength
D	: Well depth
E	: Young's modulus
F_j	: Jet impact force
G	: Shear modulus
GR	: Gamma-ray log reading
I_{GR}	: Gamma-ray index
K	: Kernel
L_w	: ANN loss function
L_ε	: ε-insensitive loss function
M	: Margin
P	: Probability
Q₁	: 1 st quartile
Q₃	: 3 rd quartile
T	: Tree
T₀	: Tensile strength
V_{sh}	: Shale index
W_b	: Weight on bit
a	: Learning rate
a_{1,...,8}	: Regional regression coefficients of BYM
b	: Bias
c	: Coefficient of internal friction angle
cov	: Covariance
d_b	: Bit diameter
e	: Residual
g_p	: Pore pressure gradient
h	: Fractional bit tooth wear
l	: Number of training data
r	: Correlation coefficient
s	: Standard deviation of sample differences
t	: t-score
v_p	: Compressional wave velocity
v_s	: Shear wave velocity
w	: Weight
x₂	: Effect of normal compaction
x₃	: Effect of under compaction
x₄	: Effect of overbalance
x₅	: Effect of weight on bit per bit diameter
x₆	: Effect of rotary speed
x₇	: Effect of bit tooth wear
x₈	: Effect of hydraulics

$\mathbf{x}_{\text{norm}}$	: Normalized input
z_i	: z-score
ΔC	: Bulk compressibility
Δt_s	: Shear sonic log reading
Θ	: Regression error term
N	: Poisson's ratio
Φ	: Objective function
β_i	: Regression coefficients
γ	: RBF kernel radius controller parameter
ε	: SVR tolerance
ζ	: Slack variables
θ	: Internal friction angle
ρ	: Pairwise correlation
ρ_b	: Bulk density
ρ_c	: Equivalent circulation density
σ	: Standard deviation
σ_1	: Maximum principal stress
σ_2	: Intermediate principal stress
σ_3	: Minimum principal stress
σ^2	: Sample variance
μ	: Mean
μ_0	: Hypothesized mean
ν	: SVR tolerance controller parameter
ξ_i	: Characterizer of the i^{th} random forest
ω	: Rotary speed

LIST OF TABLES

	<u>Page</u>
Table 2.1 : Comparison matrix for the studies regarding ML applications for ROP prediction.....	27
Table 6.1 : Formation thickness and lithological composition for the wells C-1 and C-2.	71
Table 6.2 : Descriptive statistics of the dataset for Case 1.1.	75
Table 6.3 : SVR hyper-parameters for Case 1.1.....	77
Table 6.4 : Descriptive statistics of the dataset for Case 1.2.	79
Table 6.5 : SVR hyper-parameters for Case 1.2.....	81
Table 6.6 : Descriptive statistics of the dataset for Case 1.3.	83
Table 6.7 : SVR hyper-parameters for Case 1.3.....	85
Table 6.8 : Descriptive statistics of the dataset for Case 2.1.	87
Table 6.9 : SVR hyper-parameters for Case 2.1.....	89
Table 6.10 : Descriptive statistics of the dataset for Case 2.2.	92
Table 6.11 : SVR hyper-parameters for Case 2.2.....	94
Table 6.12 : Descriptive statistics of the dataset for Case 2.3.	97
Table 6.13 : Descriptive statistics of the dataset for Case 3.1.	102
Table 6.14 : RMSE values of the feature selection for Case 3.1.	104
Table 6.15 : Descriptive statistics of the dataset for Case 3.2.	107
Table 6.16 : RMSE values of the feature selection for Case 3.2.	109
Table 6.17 : Descriptive statistics of the dataset for Case 3.3.	112
Table 6.18 : RMSE values of the feature selection for Case 3.3.	114
Table 6.19 : Summary of the overall results.	116
Table A.1 : Data from Bourgoyne Jr. and Young Jr. (1974).	141
Table A.2 : Data from Irawan et al. (2012).	142
Table A.3 : Data from Al-Betairi et al. (1988).	142
Table D.1 : SVR hyper-parameters for Case 2.3.1.	157
Table D.2 : SVR hyper-parameters for Case 2.3.2.	161
Table D.3 : SVR hyper-parameters for Case 2.3.3.	164
Table D.4 : SVR hyper-parameters for Case 2.3.4.	167
Table D.5 : SVR hyper-parameters for Case 2.3.5.	170
Table D.6 : SVR hyper-parameters for Case 2.3.6.	173
Table D.7 : SVR hyper-parameters for Case 2.3.7.	176
Table D.8 : SVR hyper-parameters for Case 3.1.1.	179
Table D.9 : SVR hyper-parameters for Case 3.1.2.	182
Table D.10 : SVR hyper-parameters for Case 3.1.3.	185
Table D.11 : SVR hyper-parameters for Case 3.1.4.	188
Table D.12 : SVR hyper-parameters for Case 3.1.5.	191
Table D.13 : SVR hyper-parameters for Case 3.1.6.	194
Table D.14 : SVR hyper-parameters for Case 3.1.7.	197
Table D.15 : SVR hyper-parameters for Case 3.2.1.	200
Table D.16 : SVR hyper-parameters for Case 3.2.2.	203

Table D.17 : SVR hyper-parameters for Case 3.2.3.	206
Table D.18 : SVR hyper-parameters for Case 3.2.4.	209
Table D.19 : SVR hyper-parameters for Case 3.2.5.	212
Table D.20 : SVR hyper-parameters for Case 3.2.6.	215
Table D.21 : SVR hyper-parameters for Case 3.2.7.	218
Table D.22 : SVR hyper-parameters for Case 3.3.1.	221
Table D.23 : SVR hyper-parameters for Case 3.3.2.	224
Table D.24 : SVR hyper-parameters for Case 3.3.3.	227
Table D.25 : SVR hyper-parameters for Case 3.3.4.	230
Table D.26 : SVR hyper-parameters for Case 3.3.5.	233
Table D.27 : SVR hyper-parameters for Case 3.3.6.	236
Table D.28 : SVR hyper-parameters for Case 3.3.7.	239
Table D.29 : SVR hyper-parameters for Case 3.3.8.	242



LIST OF FIGURES

	<u>Page</u>
Figure 1.1 : Exploration costs.....	1
Figure 1.2 : Maximum and minimum values of a function.....	2
Figure 1.3 : Number of publications related to AI found in the OnePetro database. .	4
Figure 4.1 : Parameters of the BYM.	33
Figure 4.2 : Effect of compaction.	34
Figure 4.3 : Effect of overbalance.....	35
Figure 4.4 : ROP and WOB relationship.....	36
Figure 4.5 : ROP and rotary speed relationship.	37
Figure 4.6 : Bit tooth wear model comparison.....	38
Figure 4.7 : Dull grading chart.	39
Figure 4.8 : Effect of hydraulic models.....	40
Figure 4.9 : Gamma-ray responses for different types of lithology.....	41
Figure 4.10 : Borehole failure mechanism.....	43
Figure 4.11 : The forces acting on a drill string component.	45
Figure 5.1 : General concept of ML.....	47
Figure 5.2 : A sample illustration of ML types.	48
Figure 5.3 : An example of geometrical distribution in 2-D space.....	48
Figure 5.4 : An example of a decision tree.	49
Figure 5.5 : Linear classification.	51
Figure 5.6 : The OCSH.....	52
Figure 5.7 : Input and feature space mapping.	54
Figure 5.8 : Different loss functions.	55
Figure 5.9 : ϵ -insensitive loss function and slack variables.	56
Figure 5.10 : A sample of a perceptron.	57
Figure 5.11 : Activation functions.	59
Figure 5.12 : An MLP with one hidden layer.	59
Figure 5.13 : Overfitting due to degree of the polynomial.....	61
Figure 5.14 : Overfitting due to the number of input data.	62
Figure 5.15 : Outlier example.....	62
Figure 5.16 : Box-Whisker plot and outlier detection.	63
Figure 5.17 : Relationship between the z-scores and data centiles.	64
Figure 6.1 : General utilization flowchart.....	68
Figure 6.2 : ROP values of the datasets from the literature.	69
Figure 6.3 : ROP values of the datasets from the wells C-1 and C-2.....	70
Figure 6.4 : Correlation plot of the dataset for Case 1.1.....	75
Figure 6.5 : Boxplots of the dataset for Case 1.1.	76
Figure 6.6 : ANN hidden layer sensitivity for Case 1.1.....	76
Figure 6.7 : 5-fold CV grid search plots for Case 1.1.....	77
Figure 6.8 : ROP predictions for Case 1.1.....	77
Figure 6.9 : Error histograms for Case 1.1.	78
Figure 6.10 : Statistical comparison for Case 1.1.....	78

Figure 6.11 : Correlation plot of the dataset for Case 1.2	79
Figure 6.12 : Boxplots of the dataset for Case 1.2.....	80
Figure 6.13 : ANN hidden layer sensitivity for Case 1.2.....	80
Figure 6.14 : 5-fold CV grid search plots for Case 1.2.....	81
Figure 6.15 : ROP predictions for Case 1.2.....	81
Figure 6.16 : Error histograms for Case 1.2.....	82
Figure 6.17 : Statistical comparison for Case 1.2.....	82
Figure 6.18 : Correlation plot of the dataset for Case 1.3.	83
Figure 6.19 : Boxplots of the dataset for Case 1.3.....	84
Figure 6.20 : ANN hidden layer sensitivity for Case 1.3.....	84
Figure 6.21 : 5-fold CV grid search plots for Case 1.3.....	85
Figure 6.22 : ROP predictions for Case 1.3.....	85
Figure 6.23 : Error histograms for Case 1.3.....	86
Figure 6.24 : Statistical comparison for Case 1.3.....	86
Figure 6.25 : Correlation plot of the dataset for Case 2.1.	87
Figure 6.26 : Boxplots of the dataset for Case 2.1.....	88
Figure 6.27 : ANN hidden layer sensitivity for Case 2.1.....	88
Figure 6.28 : 10-fold CV grid search plots for Case 2.1.....	89
Figure 6.29 : Model training for Case 2.1.....	89
Figure 6.30 : Training errors for Case 2.1.....	90
Figure 6.31 : ROP predictions with testing subset for Case 2.1.	90
Figure 6.32 : Error histograms for Case 2.1.....	91
Figure 6.33 : Statistical comparison for Case 2.1.....	91
Figure 6.34 : Correlation plot of the dataset for Case 2.2.	92
Figure 6.35 : Boxplots of the dataset for Case 2.2.....	93
Figure 6.36 : ANN hidden layer sensitivity for Case 2.2.....	93
Figure 6.37 : 5-fold CV grid search plots for Case 2.2.....	93
Figure 6.38 : Model training for Case 2.2.....	94
Figure 6.39 : Training errors for Case 2.2.....	94
Figure 6.40 : ROP predictions with testing subset for Case 2.2.	95
Figure 6.41 : Error histograms for Case 2.2.....	95
Figure 6.42 : Statistical comparison for Case 2.2.....	96
Figure 6.43 : Correlation plot of the dataset for Case 2.3.	96
Figure 6.44 : Boxplots of the dataset for Case 2.3.....	98
Figure 6.45 : Cumulative statistical comparison for Case 2.3.....	98
Figure 6.46 : Training errors of different models with ϵ -SVR RBF for Case 2.3. ...	99
Figure 6.47 : Statistical performance comparison of Function x_5 for Case 2.3.....	100
Figure 6.48 : Residual box-plots for SFG model with and without Function x_5 for Case 2.3.	100
Figure 6.49 : Correlation plot of the dataset for Case 3.1.	101
Figure 6.50 : Boxplots of the dataset for Case 3.1.....	103
Figure 6.51 : Cumulative statistical comparison for Case 3.1.....	103
Figure 6.52 : Relative feature importance for Case 3.1.	104
Figure 6.53 : Statistical performance comparison of feature selection for Case 3.1.	105
Figure 6.54 : Residual box-plots of feature selection for Case 3.1.	105
Figure 6.55 : Correlation plot of the dataset for Case 3.2.	106
Figure 6.56 : Boxplots of the dataset for Case 3.2.....	108
Figure 6.57 : Cumulative statistical comparison for Case 3.2.....	108
Figure 6.58 : Relative feature importance for Case 3.2.	109

Figure 6.59 : Statistical performance comparison of feature selection for Case 3.2.	110
Figure 6.60 : Residual box-plots of feature selection for Case 3.2.	110
Figure 6.61 : Correlation plot of the dataset for Case 3.3.	111
Figure 6.62 : Boxplots of the dataset for Case 3.3.	113
Figure 6.63 : Cumulative statistical comparison for Case 3.3.	113
Figure 6.64 : Relative feature importance for Case 3.3.	114
Figure 6.65 : Statistical performance comparison of feature selection for Case 3.3.	115
Figure 6.66 : Residual box-plots of feature selection for Case 3.3.	115
Figure D.1 : ANN hidden layer sensitivity for Case 2.3.1.	157
Figure D.2 : 10-fold CV grid search plots for Case 2.3.1.	157
Figure D.3 : Model training for Case 2.3.1.	158
Figure D.4 : Training errors for Case 2.3.1.	158
Figure D.5 : ROP predictions with testing subset for Case 2.3.1.	159
Figure D.6 : Error histograms for Case 2.3.1.	159
Figure D.7 : Statistical comparison for Case 2.3.1.	160
Figure D.8 : ANN hidden layer sensitivity for Case 2.3.2.	160
Figure D.9 : 10-fold CV grid search plots for Case 2.3.2.	160
Figure D.10 : Model training for Case 2.3.2.	161
Figure D.11 : Training errors for Case 2.3.2.	161
Figure D.12 : ROP predictions with testing subset for Case 2.3.2.	162
Figure D.13 : Error histograms for Case 2.3.2.	162
Figure D.14 : Statistical comparison for Case 2.3.2.	163
Figure D.15 : ANN hidden layer sensitivity for Case 2.3.3.	163
Figure D.16 : 10-fold CV grid search plots for Case 2.3.3.	163
Figure D.17 : Model training for Case 2.3.3.	164
Figure D.18 : Training errors for Case 2.3.3.	164
Figure D.19 : ROP predictions with testing subset for Case 2.3.3.	165
Figure D.20 : Error histograms for Case 2.3.3.	165
Figure D.21 : Statistical comparison for Case 2.3.3.	166
Figure D.22 : ANN hidden layer sensitivity for Case 2.3.4.	166
Figure D.23 : 10-fold CV grid search plots for Case 2.3.4.	166
Figure D.24 : Model training for Case 2.3.4.	167
Figure D.25 : Training errors for Case 2.3.4.	167
Figure D.26 : ROP predictions with testing subset for Case 2.3.4.	168
Figure D.27 : Error histograms for Case 2.3.4.	168
Figure D.28 : Statistical comparison for Case 2.3.4.	169
Figure D.29 : ANN hidden layer sensitivity for Case 2.3.5.	169
Figure D.30 : 10-fold CV grid search plots for Case 2.3.5.	169
Figure D.31 : Model training for Case 2.3.5.	170
Figure D.32 : Training errors for Case 2.3.5.	170
Figure D.33 : ROP predictions with testing subset for Case 2.3.5.	171
Figure D.34 : Error histograms for Case 2.3.5.	171
Figure D.35 : Statistical comparison for Case 2.3.5.	172
Figure D.36 : ANN hidden layer sensitivity for Case 2.3.6.	172
Figure D.37 : 10-fold CV grid search plots for Case 2.3.6.	172
Figure D.38 : Model training for Case 2.3.6.	173
Figure D.39 : Training errors for Case 2.3.6.	173
Figure D.40 : ROP predictions with testing subset for Case 2.3.6.	174

Figure D.41 : Error histograms for Case 2.3.6.	174
Figure D.42 : Statistical comparison for Case 2.3.6.	175
Figure D.43 : ANN hidden layer sensitivity for Case 2.3.7.	175
Figure D.44 : 10-fold CV grid search plots for Case 2.3.7.....	175
Figure D.45 : Model training for Case 2.3.7.	176
Figure D.46 : Training errors for Case 2.3.7.....	176
Figure D.47 : ROP predictions with testing subset for Case 2.3.7.....	177
Figure D.48 : Error histograms for Case 2.3.7.	177
Figure D.49 : Statistical comparison for Case 2.3.7.	178
Figure D.50 : ANN hidden layer sensitivity for Case 3.1.1.	178
Figure D.51 : 10-fold CV grid search plots for Case 3.1.1.....	178
Figure D.52 : Model training for Case 3.1.1.	179
Figure D.53 : Training errors for Case 3.1.1.....	179
Figure D.54 : ROP predictions with testing subset for Case 3.1.1.....	180
Figure D.55 : Error histograms for Case 3.1.1.	180
Figure D.56 : Statistical comparison for Case 3.1.1.	181
Figure D.57 : ANN hidden layer sensitivity for Case 3.1.2.	181
Figure D.58 : 10-fold CV grid search plots for Case 3.1.2.....	181
Figure D.59 : Model training for Case 3.1.2.	182
Figure D.60 : Training errors for Case 3.1.2.....	182
Figure D.61 : ROP predictions with testing subset for Case 3.1.2.....	183
Figure D.62 : Error histograms for Case 3.1.2.	183
Figure D.63 : Statistical comparison for Case 3.1.2.	184
Figure D.64 : ANN hidden layer sensitivity for Case 3.1.3.	184
Figure D.65 : 10-fold CV grid search plots for Case 3.1.3.....	184
Figure D.66 : Model training for Case 3.1.3.	185
Figure D.67 : Training errors for Case 3.1.3.....	185
Figure D.68 : ROP predictions with testing subset for Case 3.1.3.....	186
Figure D.69 : Error histograms for Case 3.1.3.	186
Figure D.70 : Statistical comparison for Case 3.1.3.	187
Figure D.71 : ANN hidden layer sensitivity for Case 3.1.4.	187
Figure D.72 : 10-fold CV grid search plots for Case 3.1.4.....	187
Figure D.73 : Model training for Case 3.1.4.	188
Figure D.74 : Training errors for Case 3.1.4.....	188
Figure D.75 : ROP predictions with testing subset for Case 3.1.4.....	189
Figure D.76 : Error histograms for Case 3.1.4.	189
Figure D.77 : Statistical comparison for Case 3.1.4.	190
Figure D.78 : ANN hidden layer sensitivity for Case 3.1.5.	190
Figure D.79 : 10-fold CV grid search plots for Case 3.1.5.....	190
Figure D.80 : Model training for Case 3.1.5.	191
Figure D.81 : Training errors for Case 3.1.5.....	191
Figure D.82 : ROP predictions with testing subset for Case 3.1.5.....	192
Figure D.83 : Error histograms for Case 3.1.5.	192
Figure D.84 : Statistical comparison for Case 3.1.5.	193
Figure D.85 : ANN hidden layer sensitivity for Case 3.1.6.	193
Figure D.86 : 10-fold CV grid search plots for Case 3.1.6.....	193
Figure D.87 : Model training for Case 3.1.6.	194
Figure D.88 : Training errors for Case 3.1.6.....	194
Figure D.89 : ROP predictions with testing subset for Case 3.1.6.....	195
Figure D.90 : Error histograms for Case 3.1.6.	195

Figure D.91 : Statistical comparison for Case 3.1.6.....	196
Figure D.92 : ANN hidden layer sensitivity for Case 3.1.7.....	196
Figure D.93 : 10-fold CV grid search plots for Case 3.1.7.	196
Figure D.94 : Model training for Case 3.1.7.....	197
Figure D.95 : Training errors for Case 3.1.7.....	197
Figure D.96 : ROP predictions with testing subset for Case 3.1.7.	198
Figure D.97 : Error histograms for Case 3.1.7.....	198
Figure D.98 : Statistical comparison for Case 3.1.7.....	199
Figure D.99 : ANN hidden layer sensitivity for Case 3.2.1.	199
Figure D.100 : 10-fold CV grid search plots for Case 3.2.1.	199
Figure D.101 : Model training for Case 3.2.1.....	200
Figure D.102 : Training errors for Case 3.2.1.....	200
Figure D.103 : ROP predictions with testing subset for Case 3.2.1.....	201
Figure D.104 : Error histograms for Case 3.2.1.....	201
Figure D.105 : Statistical comparison for Case 3.2.1.	202
Figure D.106 : ANN hidden layer sensitivity for Case 3.2.2.	202
Figure D.107 : 10-fold CV grid search plots for Case 3.2.2.	202
Figure D.108 : Model training for Case 3.2.2.....	203
Figure D.109 : Training errors for Case 3.2.2.....	203
Figure D.110 : ROP predictions with testing subset for Case 3.2.2.....	204
Figure D.111 : Error histograms for Case 3.2.2.....	204
Figure D.112 : Statistical comparison for Case 3.2.2.	205
Figure D.113 : ANN hidden layer sensitivity for Case 3.2.3.	205
Figure D.114 : 10-fold CV grid search plots for Case 3.2.3.	205
Figure D.115 : Model training for Case 3.2.3.....	206
Figure D.116 : Training errors for Case 3.2.3.....	206
Figure D.117 : ROP predictions with testing subset for Case 3.2.3.....	207
Figure D.118 : Error histograms for Case 3.2.3.....	207
Figure D.119 : Statistical comparison for Case 3.2.3.	208
Figure D.120 : ANN hidden layer sensitivity for Case 3.2.4.	208
Figure D.121 : 10-fold CV grid search plots for Case 3.2.4.	208
Figure D.122 : Model training for Case 3.2.4.....	209
Figure D.123 : Training errors for Case 3.2.4.....	209
Figure D.124 : ROP predictions with testing subset for Case 3.2.4.....	210
Figure D.125 : Error histograms for Case 3.2.4.....	210
Figure D.126 : Statistical comparison for Case 3.2.4.	211
Figure D.127 : ANN hidden layer sensitivity for Case 3.2.5.	211
Figure D.128 : 10-fold CV grid search plots for Case 3.2.5.	211
Figure D.129 : Model training for Case 3.2.5.....	212
Figure D.130 : Training errors for Case 3.2.5.....	212
Figure D.131 : ROP predictions with testing subset for Case 3.2.5.....	213
Figure D.132 : Error histograms for Case 3.2.5.....	213
Figure D.133 : Statistical comparison for Case 3.2.5.	214
Figure D.134 : ANN hidden layer sensitivity for Case 3.2.6.	214
Figure D.135 : 10-fold CV grid search plots for Case 3.2.6.	214
Figure D.136 : Model training for Case 3.2.6.....	215
Figure D.137 : Training errors for Case 3.2.6.....	215
Figure D.138 : ROP predictions with testing subset for Case 3.2.6.....	216
Figure D.139 : Error histograms for Case 3.2.6.....	216
Figure D.140 : Statistical comparison for Case 3.2.6.	217

Figure D.141 : ANN hidden layer sensitivity for Case 3.2.7.....	217
Figure D.142 : 10-fold CV grid search plots for Case 3.2.7.....	217
Figure D.143 : Model training for Case 3.2.7.	218
Figure D.144 : Training errors for Case 3.2.7.....	218
Figure D.145 : ROP predictions with testing subset for Case 3.2.7.....	219
Figure D.146 : Error histograms for Case 3.2.7.	219
Figure D.147 : Statistical comparison for Case 3.2.7.	220
Figure D.148 : ANN hidden layer sensitivity for Case 3.3.1.....	220
Figure D.149 : 10-fold CV grid search plots for Case 3.3.1.....	220
Figure D.150 : Model training for Case 3.3.1.	221
Figure D.151 : Training errors for Case 3.3.1.....	221
Figure D.152 : ROP predictions with testing subset for Case 3.3.1.....	222
Figure D.153 : Error histograms for Case 3.3.1.	222
Figure D.154 : Statistical comparison for Case 3.3.1.	223
Figure D.155 : ANN hidden layer sensitivity for Case 3.3.2.....	223
Figure D.156 : 10-fold CV grid search plots for Case 3.3.2.....	223
Figure D.157 : Model training for Case 3.3.2.	224
Figure D.158 : Training errors for Case 3.3.2.....	224
Figure D.159 : ROP predictions with testing subset for Case 3.3.2.....	225
Figure D.160 : Error histograms for Case 3.3.2.	225
Figure D.161 : Statistical comparison for Case 3.3.2.	226
Figure D.162 : ANN hidden layer sensitivity for Case 3.3.3.....	226
Figure D.163 : 10-fold CV grid search plots for Case 3.3.3.....	226
Figure D.164 : Model training for Case 3.3.3.	227
Figure D.165 : Training errors for Case 3.3.3.....	227
Figure D.166 : ROP predictions with testing subset for Case 3.3.3.....	228
Figure D.167 : Error histograms for Case 3.3.3.	228
Figure D.168 : Statistical comparison for Case 3.3.3.	229
Figure D.169 : ANN hidden layer sensitivity for Case 3.3.4.....	229
Figure D.170 : 10-fold CV grid search plots for Case 3.3.4.....	229
Figure D.171 : Model training for Case 3.3.4.	230
Figure D.172 : Training errors for Case 3.3.4.....	230
Figure D.173 : ROP predictions with testing subset for Case 3.3.4.....	231
Figure D.174 : Error histograms for Case 3.3.4.	231
Figure D.175 : Statistical comparison for Case 3.3.4.	232
Figure D.176 : ANN hidden layer sensitivity for Case 3.3.5.....	232
Figure D.177 : 10-fold CV grid search plots for Case 3.3.5.....	232
Figure D.178 : Model training for Case 3.3.5.	233
Figure D.179 : Training errors for Case 3.3.5.....	233
Figure D.180 : ROP predictions with testing subset for Case 3.3.5.....	234
Figure D.181 : Error histograms for Case 3.3.5.	234
Figure D.182 : Statistical comparison for Case 3.3.5.	235
Figure D.183 : ANN hidden layer sensitivity for Case 3.3.6.....	235
Figure D.184 : 10-fold CV grid search plots for Case 3.3.6.....	235
Figure D.185 : Model training for Case 3.3.6.	236
Figure D.186 : Training errors for Case 3.3.6.....	236
Figure D.187 : ROP predictions with testing subset for Case 3.3.6.....	237
Figure D.188 : Error histograms for Case 3.3.6.	237
Figure D.189 : Statistical comparison for Case 3.3.6.	238
Figure D.190 : ANN hidden layer sensitivity for Case 3.3.7.....	238

Figure D.191 : 10-fold CV grid search plots for Case 3.3.7.	238
Figure D.192 : Model training for Case 3.3.7.	239
Figure D.193 : Training errors for Case 3.3.7.	239
Figure D.194 : ROP predictions with testing subset for Case 3.3.7.	240
Figure D.195 : Error histograms for Case 3.3.7.	240
Figure D.196 : Statistical comparison for Case 3.3.7.	241
Figure D.197 : ANN hidden layer sensitivity for Case 3.3.8.	241
Figure D.198 : 10-fold CV grid search plots for Case 3.3.8.	241
Figure D.199 : Model training for Case 3.3.8.	242
Figure D.200 : Training errors for Case 3.3.8.	242
Figure D.201 : ROP predictions with testing subset for Case 3.3.8.	243
Figure D.202 : Error histograms for Case 3.3.8.	243
Figure D.203 : Statistical comparison for Case 3.3.8.	244





PENETRATION RATE OPTIMIZATION IN HETEROGENEOUS FORMATIONS WITH SUPPORT VECTOR MACHINES METHOD

SUMMARY

The exploration of petroleum and natural gas resources is at topmost importance because of the unending demand for energy resources. Due to the oil companies' cost-efficient policy, the importance of reducing cost and increasing performance has ascended. Accordingly, there have been significant advances in drilling technology. Nowadays, cost and performance can be optimized thoroughly by using developing technology and computer science.

The primary optimization problem in oil and gas exploration has always been related to minimize costs due to market volatility related to the decreasing oil prices. Drilling operations constitute a significant part of the exploration costs. Thus, the main objective of cost optimization is to reduce total well costs. One way to achieve this goal is to optimize or maximize the rate of penetration (ROP). The ROP starts to decrease as a drill bit wears within a run that cause additional cost.

Practically, if the lithology is homogeneous, the optimum ROP can be achieved by adjusting controllable drilling parameters and considering the drilling cost starts to increase after a minimum value. On the other hand, this approach usually fails when drilling through complex lithological formations known as a heterogeneous environment distinctly non-uniform in lithological composition.

There is no widely accepted model for defining the optimum ROP since various variables affect the cost and time relationship. It is clear that maximizing ROP or minimizing drilling costs is a multi-parameter optimization problem that requires optimization techniques. There exist some mathematical and statistical models to optimize ROP and various drilling parameters. The most commonly used ROP model in the drilling industry is Bourgoyne and Young Method (BYM). In BYM, the aim is to calculate eight regression coefficients related to standard drilling parameters by applying multiple linear regression (MLR) analysis to drilling data for predicting ROP.

The BYM involves eight drilling parameters and requires statistically at least thirty inputs obtained from the formations with uniform lithology. The number of input data required to obtain such a large number of parameters may not always be possible. This is especially the case in drilling environments where there are not many shale zones, and where there are few wells in a field, or where complex lithology is dominant such as in the fields of Turkey. Thus, the results obtained by using the BYM in these limited conditions are meaningless, observations obtained from the field practice, and method results give very different values. On the other hand, the support vector machines (SVM) method effectively predicts ROP with higher sensitivity and conservation without decreasing the number of parameters.

The main purpose of this study is to implement a different type of regression model as an alternative to multiple linear regression by modifying BYM by introducing several geomechanical parameters to estimate a relationship between the rock mechanical properties and the heterogeneity, which are shale index, uniaxial compressive strength,

brittleness index, shear failure gradient, and torque. In this study, several predictive methods are used to build a model for ROP prediction: MLR, support vector regression (SVR), and artificial neural networks (ANN). Feature selection is performed via random forests method. The prediction accuracy of each predictive model is compared by using several statistical comparison criteria: root mean square error (RMSE), correlation coefficient, and statistical significance (p-value).

The main focus of the applications in this study is to perform rate of penetration (ROP) predictions in heterogeneous formations. With a field data taken recently, it is possible to show the prediction performances of MLR, SVR, and ANN on the dataset taken from heterogeneous lithology. Since data quality is one of the essential process of data analysis. Hence, data cleansing and data selection are made precisely while constructing the cases and the datasets.

In the results, it is clearly shown that BYM performs well in terms of ROP prediction in homogeneous formations. In heterogeneous formations, SVR with radial basis function (RBF) kernels gives better prediction results in terms of RMSE. On the contrary, for the homogeneous formations, linear models produce lower errors. ANN also performs well in heterogeneous formations in terms of ROP prediction. Adding some extra features related to the rock mechanics have different effects on prediction performances for all models. Torque is found as the key parameter to define the relationship between ROP and heterogeneity. Depth is another important parameter on ROP prediction in heterogeneous formations. Moreover, feature selection based on the random forests algorithm is applied. The cut-off value is found as formation-specific. Hence, it is stated that feature selection should be performed for each data set exclusively.

The studies' findings in this dissertation are expected to provide significant time and cost savings since it is expected that faster and more accurate results will be achieved than those of currently available methods. The results will provide a basis for monitoring the condition of a drill bit on a real-time basis during any drilling operations and determining the best time to change the drill bit while avoiding a possible increase in the overall drilling cost.

DESTEK VEKTÖR MAKİNESİ YÖNTEMİ İLE HETEROJEN FORMASYONLARDA İLERLEME HIZI OPTİMİZASYONU

ÖZET

Petrol ve doğal gaz kaynaklarının araştırılması, enerji kaynaklarına olan bitmeyen talep nedeniyle büyük önem taşımaktadır. Daha önce sondaj yapılmamış bir sahada açılacak olan bir kara arama kuyusunun toplam derinliği 7.000 metreyi bulabilmektedir. Açık deniz kuyularında ise su derinliği 2.000 metreye kadar ulaşabilmektedir. Sondaj maliyetleri, arama sahalarındaki yüksek petrol ve gaz arama maliyetlerinin yaklaşık %70'ini oluşturmaktadır.

Petrol fiyatlarındaki dalgalanma ve şirketlerin düşük maliyet politikası sebebiyle, sondaj operasyonlarında maliyeti düşürmenin ve performansı artırmanın önemi gün geçtikçe artmaktadır. Dolayısı ile sondaj teknolojisinde önemli stratejik ve teknolojik gelişmeler yaşanmaktadır. Günümüzde maliyet ve performans, gelişen teknoloji ve bilgisayar bilimi kullanılarak kapsamlı bir şekilde en uygun hale getirilebilmektedir.

Optimizasyon, mühendisliğin temel bir yöntemi olup, tanım olarak bir fonksiyonun minimum veya maksimum değerini veren koşulları belirlemektir. Optimizasyon, en uygun koşulları bulmayla ilgili çok sayıda kavramı kapsadığı gibi, tüm optimizasyon problemleri için benzersiz bir teknik veya çözüm yoktur. Bir optimizasyon probleminin uygunluğu, kısıtlamalara ve amaç fonksiyonuna bağlıdır. Amaç fonksiyonu tanımlandığında ve matematiksel olarak formüleleştirildiğinde, en uygun çözüm farklı sayısal teknikler kullanılarak hesaplanabilmektedir. Bu teknikler, matematiksel programlama yöntemleri, stokastik süreçler, istatistiksel yöntemler ve modern makine öğrenimi modelleri olarak örneklendirilebilir.

Petrol ve doğal gaz aramacılığındaki birincil optimizasyon problemi, düşen petrol fiyatlarına bağlı olarak piyasa dalgalanması sebebi ile her zaman maliyetleri en aza indirmekle ilgili olmuştur. Maliyet optimizasyonunun temel amacı, toplam kuyu maliyetlerini düşürmektir. Sondaj operasyonlarının ise arama maliyetlerinin önemli bir bölümünü oluşturduğu göz önüne alındığında, yüksek kuyu maliyetlerini düşürmenin etkin bir yolu, sondaj ilerleme hızını en uygun hale getirmek veya mümkün olan en büyük değere yükseltmektir. Sondaj operasyonları esnasında matkap dişleri aşındıkça ilerleme hızı azalmaya başlamakta, bu durum da ekstra maliyete sebep olmaktadır. Bu sebeple sondaj esnasında matkabın durumunun ve ilerleme hızının dikkatli bir şekilde takip edilmesi, maliyet optimizasyonu problemi için yüksek derecede önem arz etmektedir.

Pratikte eğer litoloji homojen ise, kontrol edilebilir sondaj parametreleri ayarlanarak ve sondaj maliyetinin minimum bir değerden sonra artmaya başladığını düşünerek en uygun ilerleme hızı elde edilebilir. Öte yandan bu yaklaşım, stratigrafik anlamda belirgin bir şekilde düzenli olmayan (heterojen bir ortam olarak bilinen) karmaşık litolojik formasyonlarda sondaj yaparken genellikle başarısız olmaktadır.

Çeşitli değişkenler maliyet ve zaman ilişkisini etkilediği için en uygun ilerleme hızını hesaplamak için yaygın olarak kabul edilen bir model yoktur. İlerleme hızını mümkün olan en yüksek değere çıkarmanın veya sondaj maliyetlerini en aza indirmenin

optimizasyon teknikleri gerektiren çok parametrelili bir optimizasyon problemi olduğu açıktır. En yüksek ilerleme hızı ve en düşük maliyeti belirleyebilmek için çeşitli yöntemler ile çalışmalar yapılarak literatüre kazandırılmıştır. İlerleme hızını ve çeşitli sondaj parametrelerini en uygun hale getirmek için bir takım matematiksel ve istatistiksel modeller geliştirilmiştir. Sondaj endüstrisinde en yaygın olarak kullanılan ilerleme hızı tahmin modeli Bourgoynne ve Young Metodu'dur (BYM). BYM'deki temel amaç, ilerleme hızını tahmin etmek için sondaj verilerine çoklu doğrusal regresyon yöntemi uygulayarak standart sondaj parametreleriyle ilgili sekiz adet regresyon katsayısını hesaplamaktır. BYM yönteminde ele alınan sondaj parametreleri formasyon kazılabilirliği, derinlik, eşdeğer sirkülasyon yoğunluğu, gözenek basıncı gradyanı, matkap üzerine verilen ağırlık ve matkap çapı, dizi döndürme hızı, matkap dış aşınımı ve hidroliktir.

Teorik olarak, sekiz değişken (veya parametre) için sekiz korelasyon katsayısını (veya sekiz denklemi) tahmin etmek için sekiz girdi verisi yeterlidir. Ancak bu durum, gerçekleşen sondajın sekiz denklem ile yalnızca kusursuz bir şekilde temsil edildiği durumlar için geçerlidir. Bu durumun pratikte asla gerçekleşmeyeceği ise bilinen bir gerçektir. Bununla birlikte, az girdi verisi kullanmak, regresyonda negatif katsayı değerleri elde edilmesine neden olacaktır. Matematiksel olarak, doğru modeli oluşturmak ve korelasyon katsayılarını en iyi doğrulukla tahmin etmek için belirli sayıda parametre için minimum sayıda girdi verisi mevcut olmalıdır. Buna bağlı olarak, parametre sayısı azaldıkça minimum girdi verisi ihtiyacı da azalmaktadır.

BYM için en kritik durum, sekiz sondaj parametresini temsil eden sondaj verilerinin belirli bir saha içerisinde birden fazla kuyudaki tek tip bir formasyondan (özellikle şeyl) alınması gerektiğidir. BYM, matematiksel olarak doğru regresyon sonuçları elde edebilmek için en az 30 veri setine ihtiyaç duymaktadır. Bu veriler, BYM'nin orijinal çalışmasında da belirtilen Offshore Louisiana bölgesi gibi kalın şeyl formasyonlarının egemen olduğu sahalardan rahatlıkla elde edilebilir. Ancak, arama sahalarında (özellikle Türkiye'de) litolojik olarak aşırı karmaşıklığa sahip bir formasyonun kazılması durumu ile sıklıkla karşılaşılabilir. Bu tür ortamlar, BYM'nin tahmin doğruluğunu önemli ölçüde düşürmekte ve formasyon heterojenliği nedeniyle BYM'yi kullanışsız hale getirmektedir. Bu sebeple, ilerleme hızını daha iyi tahmin edebilmek için çoklu doğrusal regresyon yöntemi dışında başka bir tahmin yöntemi araştırılmalı ve kullanılmalıdır.

Geleneksel matematiksel optimizasyon tekniklerini kağıt üzerinde büyük boyutlu bir veri setine uygulamak mümkün değildir. Ayrıca bu teknikler uzun hesaplama süresi gerektirmektedir. Bununla birlikte, bilgisayar teknolojisinin donanım ve yazılım bakımından geliştirilmesi, zorlu optimizasyon problemlerinin çözülmesini oldukça kolaylaştırmaktadır. Makine öğrenimi kavramı, eldeki verilerden bir model oluşturarak yeni veriler veya amaç fonksiyonu hakkında tahminlerde bulunma amacıyla geliştirilmiştir. Makine öğrenimi, birçok yönden verilerden bir model oluşturmak ve oluşturulan bu model analiz etmek için oldukça kullanışlı bir araçtır.

Makine öğrenmesi yöntemleri, petrol mühendisliğinde 1980'lerin ortalarından beri kullanılmaktadır. Günlük hayatımızın hemen her alanında kullanılan yapay zeka yöntemleri, petrol endüstrisi uygulamalarında geniş bir yelpazeye sahiptir. Petrol mühendisliğinde kullanılan makine öğrenmesi yöntemleri arasında en sık kullanılan yöntemler Destek Vektör Makineleri (DVM) ve Yapay Sinir Ağları (YSA)'dır.

Literatürde karmaşık litolojilerin hakim olduğu durumlar için ilerleme hızı tahmininde çoklu doğrusal regresyon yönteminin doğru ve anlamlı sonuçlar vermediği

belirtilmiştir. Bunun yanında araştırmacılar, YSA yönteminin çoklu doğrusal regresyon yöntemine göre ilerleme hızı tahmini için daha anlamlı ve doğru sonuçlar verdiğini ifade etmişlerdir. Ancak YSA yönteminde kullanılan girdi verisi sayısı, DVM yönteminin gerektirdiği veri sayısının yaklaşık 2-3 katı kadardır. Ayrıca literatürde diğer konular için gerçekleştirilen araştırmalarda DVM'nin daha az girdi verisi kullanarak YSA'dan daha hızlı ve daha doğru sonuçlar verdiğini belirtilmiştir. Bununla birlikte DVM'nin BYM'de tanımlanan parametre fonksiyonlarını kullanarak heterojen formasyonlarda ROP tahmini için kullanılmadığı literatür araştırması sonucunda ortaya çıkmıştır. Bu bağlamda, DVM'nin doğrudan BYM fonksiyonlarına uygulanarak heterojen formasyonlarda ilerleme hızı tahminini içeren bu doktora tezi, literatürde bir ilk olması sebebiyle benzersiz bir yöne sahiptir.

Bu doktora tezinin temel amacı, mekanik kayaç özellikleri ile heterojenlik arasındaki bir ilişkiyi belirlemek için çeşitli jeomekanik parametreleri tanıtarak BYM'yi değiştirmek ve çoklu doğrusal regresyon yöntemine alternatif olarak farklı bir regresyon modeli uygulamaktır. BYM'ye ilave edilen jeomekanik parametreler şeyl indeksi, tek eksenli basınç dayanımı, kırılgenlik indeksi, kesme kırılması gradyanı ve torktur. Bu çalışmada, heterojen formasyonlarda ilerleme hızı tahmini amacıyla bir model oluşturmak için farklı tahmin yöntemleri kullanılmıştır. Bu yöntemler çoklu doğrusal regresyon yöntemi, destek vektör regresyonu yöntemi ve yapay sinir ağları yöntemidir. Verinin kalitesi, veri analizinin temel niteliklerinden biridir ve model doğruluğu için yüksek önem arz etmektedir. Bu nedenle, veri setleri oluşturulurken veri temizleme ve veri seçimi işlemleri hassas bir şekilde yapılmıştır. Öznitelik seçimi rastgele ormanlar yöntemi ile gerçekleştirilmiştir. Her bir tahmin modelinin doğruluğu ortalama kare hatası, korelasyon katsayısı ve istatistiksel anlamlılık (p-değeri) değerleri kullanılarak değerlendirilmiş ve birbirleri ile karşılaştırılmıştır.

Sonuçlarda BYM'nin homojen formasyonlarda ilerleme hızı tahmini açısından iyi performans verdiğini gösterilmiştir. Heterojen formasyonlarda ise Gauss çekirdek fonksiyonlu DVR, diğer tahmin yöntemlerine göre daha düşük hatalar ile daha iyi tahmin sonuçları vermektedir. Diğer taraftan, homojen formasyonlarda ilerleme hızı tahmini için doğrusal modeller daha düşük hata ile tahmin gerçekleştirmektedir. Ayrıca, YSA yönteminin ilerleme hızı tahmini açısından heterojen formasyonlarda iyi performans gösterdiği görülmüştür. BYM yöntemine jeomekanik parametreler eklemenin farklı modellerin ilerleme hızı tahmin performansları üzerinde farklı etkileri olduğu gözlenmiştir. İlerleme hızı ve heterojenlik arasındaki ilişkiyi tanımlayan anahtar parametre tork olarak belirlenmiştir. Heterojen formasyonlarda ilerleme hızı tahmininde diğer bir önemli parametrenin ise derinlik olduğu görülmüştür. Bunun yanı sıra, rastgele ormanlar algoritmasına dayalı öznitelik seçimi uygulamasında eşik değeri, formasyona özgü olarak değişkenlik göstermektedir. Bu nedenle, öznitelik seçiminin her bir veri seti için özel olarak yapılması gerektiği belirtilmiştir.

Elde edilen çıktılar ışığında, bu çalışmada kullanılan yaklaşımın mevcut yöntemlerden daha hızlı ve daha doğru sonuçlara ulaştığı görülmüştür. Dolayısı ile bu doktora tezindeki bulguların sondaj operasyonlarında önemli ölçüde zaman ve maliyet tasarrufu sağlaması beklenmektedir. Elde edilen sonuçlar, sondaj operasyonları esnasında maktabın dış aşınım durumunu gerçek zamanlı olarak izleme ve sondaj maliyetinde olası bir artışı önlemek için en uygun matkap değiştirme zamanını belirleme hususunda da bir temel sağlayacaktır.



1. INTRODUCTION

Petroleum and natural gas exploration is at topmost importance because of the unending demand for energy sources. Total depth of an onshore wildcat well for oil and gas exploration can reach up to 7,000 meters. Likewise, in offshore wells, water depth can reach up to 2,000 meters. Drilling costs constitute up to 70% of the current high costs of oil and gas exploration in unexplored fields, which is illustrated by a data taken from Norway in Figure 1.1.

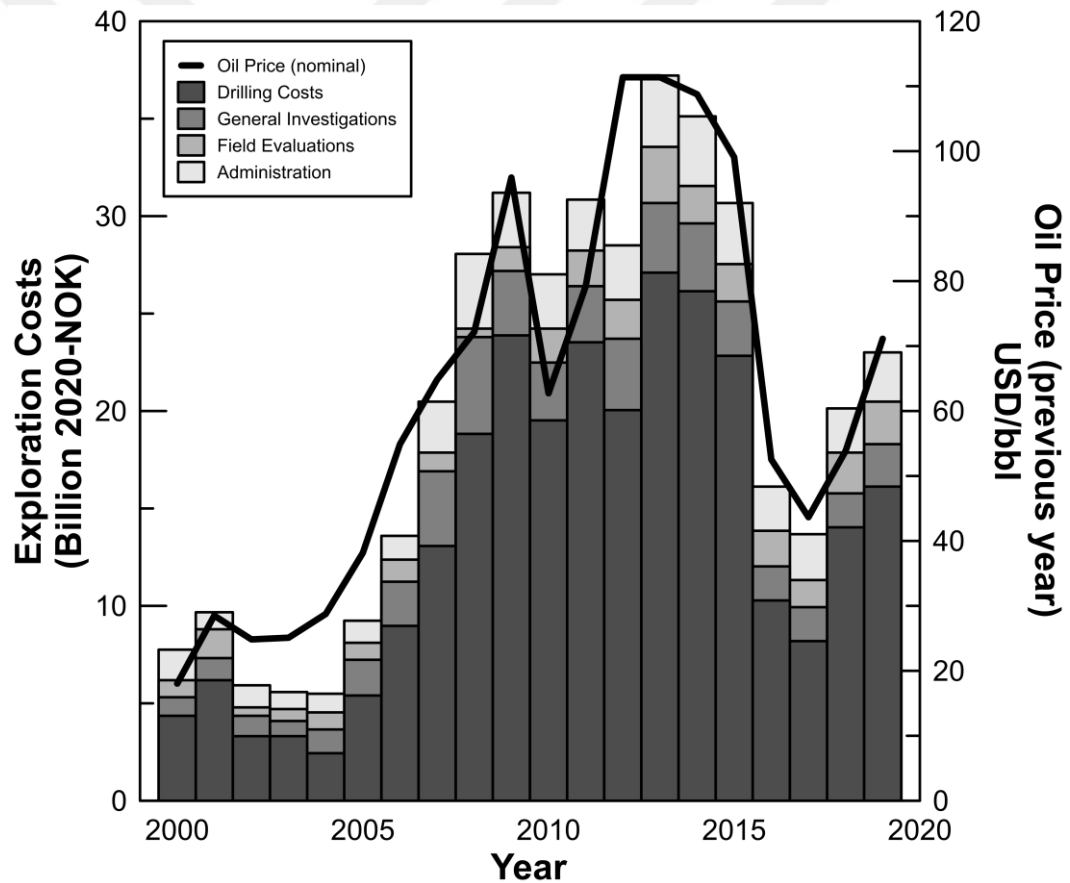


Figure 1.1 : Exploration costs, adapted from (NPD, 2020).

Due to the oil companies' cost efficiency policy, the importance of reducing cost and increasing performance has ascended. Accordingly, there have been significant advancements in drilling technology. Nowadays, cost and performance can be optimized thoroughly by using developing technology and computer science.

Optimization is an essential aspect of engineering. Optimization is to find the conditions which give the minimum or maximum value of a function (Rao, 2009, p. 2). An example of the maximum and minimum values of a function is illustrated in Figure 1.2.

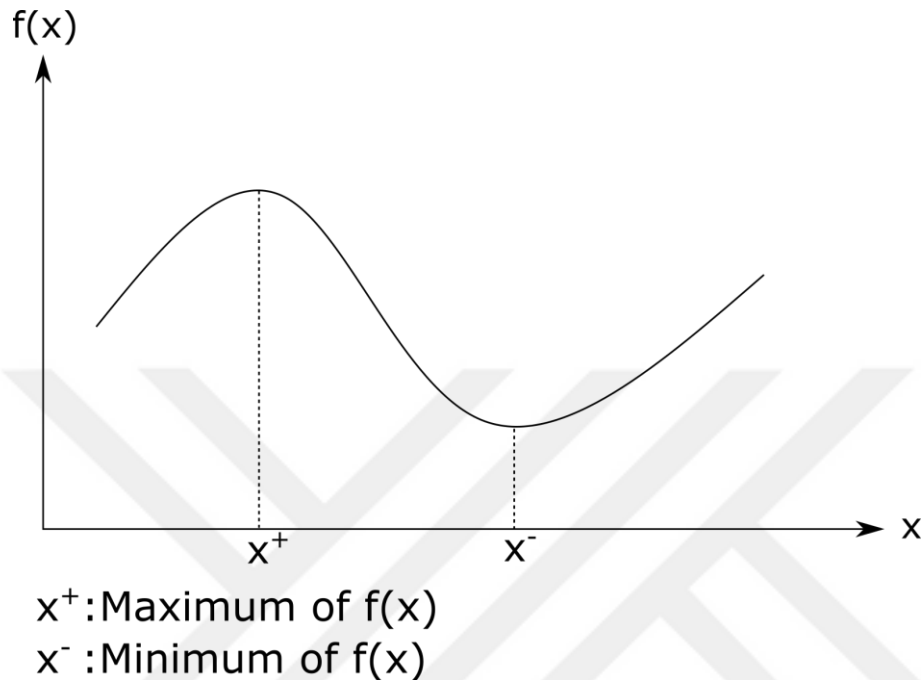


Figure 1.2 : Maximum and minimum values of a function.

Optimization covers numerous concepts about finding specific optimality. There is no unique technique or solution for all of the optimization problems. However, the compliancy of an optimization problem is hinged on the constraints and the objective functions (Yang, 2010, p. 15). When the objective function is defined and mathematically formulated, the optimum solution can be found by using several numerical techniques. These techniques can be mathematical programming approaches (linear programming, quadratic programming), stochastic processes (Markov processes, simulation methods), statistical methods (regression analysis, clustering, discriminant analysis), and modern machine learning models (support vector machines, artificial neural networks, genetic algorithms).

The primary optimization problem in oil and gas exploration has always been related to minimizing costs due to market volatility related to the decreasing oil prices. Drilling operations constitute a significant part of the exploration costs. Thus, the main objective of cost optimization is to reduce total well costs. One way to achieve this goal is to optimize (or maximize) bit penetration rate (ROP) among the approaches.

The ROP starts to decrease as a drill bit wears off within a run, which results in extra cost.

Practically, if the lithology is homogeneous, the optimum ROP can be achieved by adjusting controllable drilling parameters (weight on bit, rotary speed, drilling hydraulics, etc.) and considering the drilling cost starts to increase after a minimum value. On the other hand, this approach usually fails when drilling through distinctly non-uniform complex lithological formations called heterogeneous environments (Bourgoyne Jr. and Young Jr., 1974).

There is no widely accepted model for defining the optimum ROP since various variables affect the cost and time relationship. It is clear that maximizing ROP or minimizing drilling costs is a multi-parameter optimization problem that requires optimization techniques. Several studies were conducted to determine the maximum ROP and minimum cost. There exist some mathematical and statistical models to optimize ROP and various drilling parameters. The most commonly used ROP model in the drilling industry is Bourgoyne and Young Method (BYM) (Bourgoyne Jr. and Young Jr., 1974). In BYM, the aim is to calculate eight regression coefficients related to standard drilling parameters by applying Multiple Linear Regression (MLR) analysis to drilling data for predicting ROP.

A critical and vital issue for BYM considering eight drilling parameters is that the drilling data must be taken from one type of uniform formation (especially shale) from more than one well in a particular field. Mathematically speaking, BYM requires at least 30 data sets to obtain accurate regression results. Besides, these data sets must be taken from different depths of uniform lithology in various wells in a field. This number of inputs can be acquired from the fields with thick shale sections such as Offshore Louisiana Area indicated in the original study. However, drilling through a formation with extreme complexity (non-uniform) in lithology can be confronted frequently worldwide, particularly in the exploration fields such as Turkey (Okay, 2008). These kinds of drilling environments make the prediction accuracy of BYM decrease substantially and make the BYM impractical because of the formation heterogeneity. Hence, another predictive method other than MLR must be searched and used to predict ROP better.

It is impossible to apply conventional mathematical optimization techniques to a large-sized data set by using pen and paper. In addition, these techniques necessitate long calculation time. However, by improving computer technology with hardware and software capabilities, it is now easy to solve challenging optimization problems. Machine Learning (ML) concept is developed to learn and build a model from data to make predictions about new (unseen) data or objective function. ML is a convenient tool to analyze and build a model from data in many aspect.

ML methods have been used in petroleum engineering since the mid-1980s. Artificial intelligence (AI) methods, which are used in almost every aspect of our daily lives today, have a wide range of petroleum industry applications (Kor and Altun, 2020). Among the ML methods used in petroleum engineering since the 1980s, the most frequently used methods are Support Vector Machines (SVM) and Artificial Neural Networks (ANN). Figure 1.3 shows the number of studies and publications on SVM and ANN found in the OnePetro database (main publication database of petroleum engineering related subjects).

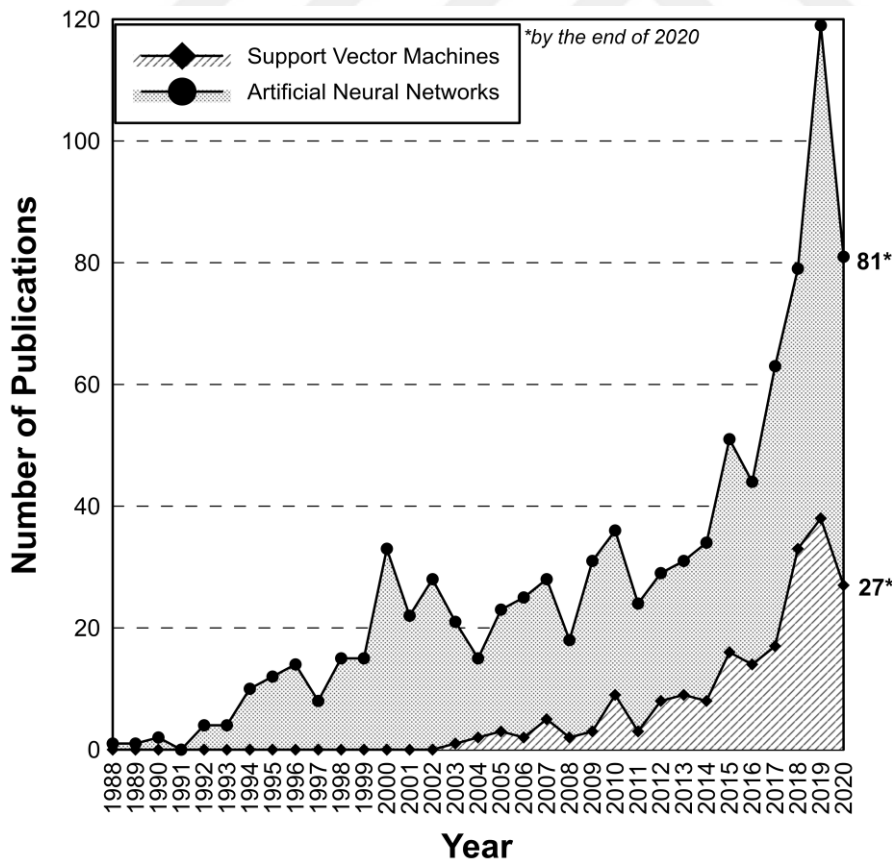


Figure 1.3 : Number of publications related to AI found in the OnePetro database (OnePetro, 2021).

In this study, two of the commonly used ML methods are implemented to predict ROP in heterogeneous formations along with the BYM: Support Vector Regression (SVR) and ANN. To summarize, this dissertation is organized as follows:

- Chapter 2 is the literature review based on ROP prediction models.
- Chapter 3 represents the problem statement and the scope of this dissertation.
- Chapter 4 is the detailed explanation of the BYM and additional geomechanical parameters to define heterogeneity.
- Chapter 5 contains the explanation of methods and predictive models used in the study.
- Chapter 6 is the results and discussions of the calculations performed in this study.
- Chapter 7 includes the conclusions and future recommendations.

Chapters 2 and 3 are introductory for the scope of the study. Chapters 4 and 5 include the theoretical concepts and the mathematical background of BYM, MLR, SVR, and ANN. The data taken from the literature is given in Appendix A. The detailed derivation of the SVR method is given in Appendix B. The code written in R Statistical Programming Language used for the calculations is provided in the Appendix C. Graphical presentation of the calculations in Chapter 6 is given in Appendix D.



2. LITERATURE REVIEW

Several studies were performed to predict and optimize the drilling penetration rate. Essentially, the goal is to minimize the cost and achieve the maximum rate of penetration (ROP). Research, as an ongoing and always challenging subject, in ROP optimization will sustain its priority in the drilling industry because of the volatility and dependency on oil prices.

There are various studies in the literature on ROP optimization by using different approaches. These approaches are grouped and investigated in this study under the following topics: (1) drilling and fluid parameters optimization, (2) hardware design and laboratory experiments, (3) mathematical optimization and statistics, (4) data-driven and machine learning methods, and (5) ROP optimization in heterogeneous formations.

2.1 Drilling and Fluid Parameters Optimization

The very first drilling parameters optimization was studied by Speer (1959). The author developed a primary method to determine the optimum combination of weight on bit (WOB), rotary speed (RPM), and hydraulics giving the minimum cost by establishing empirical correlations. It was showed that the optimum WOB depends on the formation drillability and RPM for a given WOB. The relationships were combined and placed on a chart to determine the optimum parameters for desired conditions.

The other primary drilling parameters optimization was performed by Graham and Muench (1959). In this study, an analytical model was developed by taking RPM and WOB under consideration to minimize the cost. It is indicated that either WOB or RPM are increased, the cost decreases because of ROP increment. Nevertheless, either WOB and RPM is selected excessively, the bit life decreases, which causes trip time and bit cost increase. Hence, empirical correlations were derived to determine expected bit life and ROP parameters using WOB, RPM, and depth data.

One of the most critical ROP optimization models was developed by Galle and Woods (1963). Several charts were generated to find optimum parameters that minimize the

cost for various cases, including the optimum constant WOB and RPM, the optimum WOB for variable RPM, and the optimum RPM for variable WOB. This model considers milled tooth bits only by evaluating both tooth wear and bearing life.

The importance of drilling fluids and hydraulics on ROP optimization was investigated by Lummus (1971). It is indicated in the study that it is possible to optimize drilling operations thoroughly by being able to provide sufficient hydraulic force, RPM, and WOB.

Warren (1987) indicated that the ROP limitation for the roller cone bits was due to the occurrence and removal of the cuttings during drilling. An ROP model was introduced, including the effect of formation chipping and cuttings removal. It is also mentioned that the ROP could decrease in high overbalance because of the cratering and borehole cleaning.

Iqbal (2008) performed an analysis on Galle and Woods method to determine the optimum WOB and RPM for the maximum ROP. The author implemented software to find and adjust the optimum parameters to maximize ROP. It is reported that the cost decreases, as well as the ROP increases.

He et al. (2016) found a global correlation between rock brittleness and petrophysical and geomechanical properties. A system was designed for estimating ROP for polycrystalline diamond compact (PDC) bits using several log data such as gamma-ray, neutron, density, and sonic logs. The results showed a good match between observed and estimated ROP values in homogeneous formations.

Lee et al. (2018) divided the abrasiveness of drilling bits into several steps. They used the bit life expectancy equation to develop an algorithm to calculate the bit changing trip time. A BYM-based methodology classifies the depth based on the formation characteristics and drilling parameters introduced in the study.

Eren and Kok (2018) studied the data of 40 wells drilled in the Middle East region. It is showed that previous drillability comparisons were not reasonable for benchmarking. It is stated that a new methodology using advanced technology and data acquisition systems was required in the drilling industry to benchmark the drilling performance of the wells accurately.

Alkinani et al. (2018) investigated the effect of mud weight on ROP. A method to determine the optimal mud weight window by using rock hardness was introduced.

Field data were acquired from more than 800 wells to examine the relationship between the mud weight and the ROP based on the rock hardness. The results showed that mud weight has an inverse relationship with ROP for weak formations. On the contrary, it is stated that mud weight directly relates to ROP for the medium and hard formations.

Al Rubaii et al. (2018) studied the effect of bottom hole cleaning on the ROP to assure optimum drilling performance. It is mentioned that having more information about the bottom hole cleaning can improve the ROP. Fluid carrying capacity index, cuttings concentration, and mechanical specific energy (MSE) were used to estimate the cleaning condition and the ROP performance. The method was validated in the challenging sections, where the ROP was improved by 55% as MSE is degraded by 54%.

2.2 Hardware Design and Laboratory Experiments

One of the early laboratory experiments with micro bits was performed by Eckel (1967) in a low-permeability limestone. The tests were operated at a constant WOB and RPM along with variable flow rate and nozzle configuration. It is showed that ROP with a constant flow rate and a given nozzle configuration is related to the kinematic viscosity. Also, a Reynolds number function explains the effect of drilling fluid properties and hydraulics on ROP.

Laboratory tests to investigate the effects of hydraulics and hole cleaning on ROP with soft formation insert bits were conducted by Doiron and Deane (1982). Shale samples in certain fixed conditions were used in tests. It is stated that ROP is increased with increasing hydraulic force, and ROP is decreasing with the increasing nozzle size or number. Additionally, it is mentioned that hydraulic horsepower over limits can be cost-effective due to high overbalance when ROP is low.

Pessier and Fear (1992) conducted simulator tests under laboratory conditions to develop an MSE drilling model considering hydrostatic pressure. Field data analysis shows a good match between the simulator and actual results. It is also indicated that MSE can be a good performance indicator for accurate drilling operations.

Caicedo et al. (2005) developed a new method to calculate confined compressive strength (CCS) regarding rock mechanics. A methodology was introduced to select the

input parameters for optimum bit performance based on CCS and MSE, validated with laboratory tests. It is mentioned that the new method is robust and requires almost no calibration to improve drilling performance and minimize drilling costs by determining optimum parameters.

Judzis et al. (2009) described the limiting factors in ROP as depth, circulation systems, and drill bit design. Numerous high-pressure tests were performed in a laboratory with four different same-sized drill bits, three different types of rock samples, and five different types of drilling fluids. It is stated that PDC bits perform better in ROP rather than roller cone bits.

Hareland et al. (2010) performed tests with tungsten carbide inserts (TCI) for the rock failure mechanism. A new ROP model for TCI bits was developed. It is mentioned that ROP and compressive strength can be predicted by the new ROP model with offset data.

Esmaeili et al. (2012) carried out tests within a lab-scaled drilling rig to demonstrate the effect of vibration on the ROP. A sensor placed on different locations on the rig recorded the vibration and various drilling parameters while drilling through a sandstone sample with different WOB and RPM values. Several models were constructed by using artificial neural networks (ANN) to analyze data and optimize ROP.

Anemangely et al. (2018) used punch penetration to determine the rock drillability index. In the study, several samples were taken from the two wells drilled in the same formation. It is stated that the relationships between different features and the measured ROP were fit to the exponential and power-law model. Evaluation of the models between the features and actual ROP showed that the exponential model fit the calculated values. A comparison between the estimated and the measured ROP values indicated that the estimated ROP from the proposed index is accurate and can represent the rock drillability.

Al Dabyah et al. (2018) mentioned the importance of real-time geomechanics using a downhole tool in a case study. It is indicated that usage of the right tool with expert engineers made real-time parameter interaction possible. It is also stated that downhole optimization via real-time geomechanics improved the performance. Additionally, it

is mentioned that including geomechanical studies in the well planning phase and the operations resulted in higher ROP and less well instability problems.

Millan and Ringer (2018) presented a workflow to observe real-time drilling efficiency. It is stated that the method could be used to predict the ROP, optimize the drilling parameters, and estimate the optimum bit pulling time. The method was first initiated in a lab-scale drilling rig. Then, the results were validated in drilling operations. It is mentioned that the method is powerful enough to estimate drilling efficiency in real-time and predicting bit wearing and the ROP.

Mazen et al. (2020) proposed a model to build the relationship between the ROP and drilling parameters to estimate the PDC performance and wearing during operations. The model analyzed the effect of rock strength, bit design, and hydraulics on cutter wearing by calculating MSE, torque, and depth of cut. Then the model was validated by predicting the wearing condition of the PDC cutters in the field.

Han et al. (2020) established a relationship between the MSE measured from micro-drilling tests and scratch tests. A new model to predict the rock drillability of PDC bits is established. Several types of rocks with various ranges of hardness were used in the scratch tests. It is stated that the results from the scratch test model had a good match that those of micro-drilling tests. It is also mentioned that the model can be used to characterize the rock in terms of mineral composition.

2.3 Mathematical Optimization and Statistics

Reed (1972) used a Monte Carlo approach to solve the ROP optimization problem of variable WOB and RPM. The results of the study were compared and matched to those of the method of Galle and Woods. It is stated that variable WOB and RPM optimization give slightly better results when compared to the constant WOB and RPM approach. As a result, computer software was designed, which used the conjugate gradient technique to solve the constant WOB and RPM problem.

One of the most widely used statistical ROP models was developed by Bourgoyne Jr. and Young Jr. (1974). The Bourgoyne and Young Model (BYM) was primarily designed to detect abnormally pressured intervals while drilling, which can also predict ROP. The model is based on multiple regression analysis to determine the effects of several properties such as formation strength, depth, compaction, overbalance, WOB,

RPM, bit tooth wear, and hydraulics. This ROP model is the fundamental of this dissertation and will be investigated in detail in later chapters.

Tansev (1975) presented an experimental approach including regression and optimization with raw data. Equations to predict ROP was found by regressing data granted from several bit runs. WOB, RPM, and hydraulics are selected as control parameters. ROP equations were implemented to the cost equation to minimize drilling cost by adjusting controlling parameters. It is also indicated that the approach is suitable for drilling a homogeneous formation with similar types of bits.

Al-Betairi et al. (1988) presented an application and verification of the BYM, which predicts optimum ROP, WOB, and RPM by using field data gathered from three wells. It is indicated that the regression coefficients are data-sensitive, and it is crucial to perform a multicollinearity sensitivity analysis for each parameter.

Barragan et al. (1997) introduced a Monte Carlo approach to determine optimum drilling conditions for a specific bit run and compared the results with other ROP estimation models. It is indicated that the drilling cost can be minimized for an entire well by using and implementing the approach.

Bahari and Seyed (2007) applied four statistical models to field data to predict ROP. The models were multiple regression, least-squares curve fitting, Gauss-Newton, and trust-region, which are applied directly to the data converted into the BYM. The trust-region approach is the best model to estimate parameter coefficients, which can also be used when the data is limited and exceeds recommended ranges.

Osgouei (2007) modified BYM by developing a bit tooth model for roller cone and PDC drill bits. The new model also included the effects of hole cleaning, fluid loss, and well deviation on ROP prediction. The model was tested in a field. It is mentioned that the modified Bourgoyne and Young model can predict the ROP with a 25% error.

Hamad-Allah and Ismael (2008) applied a mathematical ROP model, modified BYM to data taken from forty wells in different fields in Southern Iraq. It is stated that MLR results indicated a high correlation coefficient among the actual and predicted ROP values. A non-linear optimization method is applied to the model to estimate the optimum drilling parameters. It is mentioned that the optimization resulted in a remarkable reduction in the drilling cost.

Bahari and Seyed (2009) studied several cases from a gas field in Iran. Mathematical methods were applied to cost per foot and MSE approaches, and the results were compared. It is indicated that a proper combination of the theory and the practice contributes to notable savings in time and money for the drilling operations.

Bataee et al. (2010) performed an optimization on drilling parameters using BYM, Bingham, and Warren methods. This study aimed to determine the best ROP model for both roller cone and PDC bits using these methods and validate each method using actual field data. In conclusion, the optimization of several parameters has a remarkable effect on both ROP and the drilling cost.

Rahimzadeh et al. (2011) proposed a new method to determine the parameters constants of the BYM. It is stated that several probable outputs are selected within the recommended range, and the best solution was chosen and compared with new probable outputs. The iteration was continued until the convergence criteria had been met. It is stated in the results that the new method was found suitable for determining the coefficients in the BYM.

Irawan et al. (2012) applied BYM to a field dataset to investigate various drilling parameters' effects during drilling operations. MLR was used to determine the optimum WOB during drilling by predicting the ROP value within a given interval. The results were compared and proved with an industry-standard drilling simulator called Drillsim 500.

Arabjamaloei and Karimi Dehkordi (2012) tested and compared several models for ROP prediction. A new ROP prediction model based on fuzzy logic is introduced. It is stated that ANN can perform better ROP predictions if a large-sized data set is available. Additionally, it is stated that current mathematical equations are not sufficient in terms of ROP prediction, although they are considered as easy to apply.

Amjad et al. (2015) optimized drilling ROP to decrease drilling time and minimize drilling cost. The method used in optimization included BYM as the basis. As a result of this study, it is stated that possible problems could be overcome to improve future wells' drilling efficiency.

Wiktorowski et al. (2017) extended the BYM by adding the dogleg severity as an input parameter and applied it to field data. It is stated that the ROP prediction accuracy was

increased in all cases. It is also mentioned that the model is both suitable for ROP optimization and directional well trajectory design.

Alsubaih et al. (2018) performed a drilling performance optimization study in the Mishrif reservoir section. Data gathered from twenty-five wells and analyzed via MSE and statistical point of view. The optimization process resulted in an excessive increase in the ROP by monitoring the MSE. It is stated that several drilling parameters can be optimized to reduce the cost.

Diaz et al. (2018) investigated the drilling parameters to improve ROP prediction accuracy for the enhanced geothermal systems. Data taken from a 4200-meter pilot well was analyzed. The Fourier filter smoothed the oscillating ROP values. The effect of smoothing was controlled by multiple regression and ANN. The smoothing process improved ROP prediction accuracy by 90% correlation for the low degrees of filtering. It is suggested that the smoothness degree should be optimized for each dataset individually.

Al-AbdulJabbar et al. (2019) introduced a new ROP prediction model for carbonate formations using numerous drilling parameters with about 7000 data rows. It is indicated that ROP is directly affected by WOB, RPM, torque, and flow rate. It is also mentioned that the new ROP model surpassed the current mathematical ROP models, such as BYM, Bingham, and Maurer.

2.4 Data-Driven and Machine Learning Methods

One of the earliest data-driven and computerized ROP optimization studies was conducted by Young Jr. (1969). An on-site data acquisition and rig controlling system was developed in to control WOB and RPM by any means of optimizing ROP in terms of minimum-cost drilling.

A milestone of the real-time drilling optimization was studied by Simmons (1986) by developing a method based on d-exponent with several drilling parameters. The method lets the driller and the drilling engineer tweak and modify drilling parameters such as WOB, RPM, and hydraulics to achieve optimum performance in real-time.

Bilgesu et al. (1997) developed a new method to model the ROP and bit wear. An ANN simulator was designed to gather drilling data and remove the noise to find the

drilling parameters' relationship. The validity of the method was compared with the data taken from a field.

Fear (1999) introduced a curve-fitting method to identify the factors affecting ROP. The input parameters were gathered from mud logging unit (MLU), geological properties, and drill bit specs. Correlations were used to develop instructions to maximize ROP in real-time drilling.

Millheim and Gaebler (1999) presented a new empirical concept called Virtual Experience Simulation by acquiring data from 22 different wells and processing them into different features such as geology, logging, drilling, etc. The data subsets were gained value and knowledge based on previous drilling experience. The system works as a learning expert system and lets the individuals use the drilling experience for future drilling operations.

John et al. (2002) facilitated a remote drilling operations system called InterACT. The InterACT system uses modern information technology, enabling the geologists and drilling engineers located in offices to control and monitor the drilling operations remotely. The system was designed as a wellsite monitoring and controlling system based on the rig data transferred through the internet and accessed via web browsers.

Another remote operations system called Real-Time Operations Center was demonstrated by Ursem et al. (2003). The system was based on learning from experience and making operational decisions itself.

An innovative drilling automation and simulation system was introduced by Rommetveit et al. (2004). The system was designed to acquire all surface and downhole data in real-time to optimize the whole drilling operations, where computer controllable drilling equipment was available.

Bjornsson et al. (2004) performed an optimization using a bit selection expert system (ES) and an ROP prediction algorithm to overcome drilling performance uncertainties. The results were compared with actual drilling data, and it is stated that ROP estimation showed good accuracy within lithological intervals.

Dupriest and Koederitz (2005) designed a real-time drilling optimization system that lets rig site personnel achieve the maximum ROP by monitoring the MSE. It is stated that the pilot results from six selected rigs show a 133% increment in average ROP by monitoring MSE and optimizing ROP.

Monden and Chia (2007) performed several case studies in an operation center by modeling drilling parameters, including shock and vibration, to improve ROP. It is stated in the study that the ROP is increased when optimum WOB and RPM are selected.

Bahari et al. (2008) applied Genetic Algorithm (GA) to BYM to estimate regression coefficients more accurately. It is mentioned that GA improved the accuracy of estimating the coefficients where BYM gives nominal values. The GA approach was tested with actual field data. It is stated that field results validated the excellent approximation of coefficient estimation with GA.

Eren and Ozbayoglu (2010) established a real-time ROP optimization system based on the BYM by gathering data from the rig data acquisition system. The ROP model was optimized for the functions of BYM for each data point. It is mentioned that the ROP can be estimated and optimized in real-time as dependent on various drilling parameters.

Moradi et al. (2010) presented a soft computing method based on k-mean clustering to predict ROP. Prediction accuracy was resulted in better than BYM when the method was applied to field data.

Bataee and Mohseni (2011) introduced an Artificial Intelligence (AI) based optimization program to optimize the drilling parameters accurately with ANN. It is stated that ANN is capable of estimate the accurate values for the parameters to maximize ROP without any need for an equation. It is also indicated that the optimum ROP is chosen by learning from existing ROP values and applying optimum drilling parameters.

Arabjamaloei and Shadizadeh (2011) introduced a new model based on ANN to predict the ROP. The data was gathered from the Ahwaz field in Southern Iran. GA was integrated into the model for estimating the optimum parameters for the maximum ROP. It is stated that the new model was suitable for predicting the ROP in shaly formations.

Amar and Ibrahim (2012) compared the multiple regression with two ANN approaches; extreme learning machines (ELM) and radial basis function networks. The comparison was made based on the field data with no correction, then with WOB correction, and finally with the interpolated WOB and RPM correction. BYM

functions were used as the input parameters. The approaches were compared in terms of training/testing time and accuracy. Simulations showed that ELM resulted in the best accuracy and the shortest processing time.

Gidh et al. (2012) introduced a drilling optimization system with ANN to maximize the drilling footage and maximize the ROP by using offset data. Moreover, prediction residuals were imported again to the algorithm to increase the prediction accuracy for later datasets. It is mentioned the system is highly beneficial for improving bit life and eliminating unplanned or unnecessary trips.

Monazami et al. (2012) used ANN to predict ROP with field data. It is mentioned in the study that ANN is a powerful tool to correlate complex drilling parameters. Furthermore, it is stated that the method can be used either in the planning phase or during operations. A three-layer feed-forward network was created to estimate the correlation coefficients. It is emphasized that ANN can predict the ROP better than BYM.

Jamshidi and Mostafavi (2013) introduced two ANN and GA-based models. The first model was developed to select the appropriate drill bit for optimum ROP. The second model was developed to estimate the optimum drilling parameters acquired from the first model to maximize the ROP. The model could allow the drilling system to maximize the drilling efficiency by eliminating the anticipation of bit wearing sooner than expected.

Yi et al. (2014) optimized the ROP using the Shuffled Frog Leaping Algorithm (SFLA). In this study, real-time gamma-ray data are used to evaluate the rock strength and tooth wear. The other parameters used in this study are WOB, RPM, and flow rate. It is stated that this method was found effective in calculating optimum parameters for different formation characteristics.

Valisevich et al. (2015) developed a real-time optimization system for drill bits to improve the ROP and reduce bit wearing. The focus of the study was to minimize human-related errors. An ANN-based analysis was performed on several drilling parameters of the offset wells. The selected parameters were conducted to the rig to make the driller adjust the parameters to reach optimized ROP. It is stated that ROP was increased minimum of 6% as a result.

Wallace et al. (2015) introduced a machine learning (ML) based system for the ROP estimation. This system's basic statistical model included Random Forests (RF), ANN, and ensemble techniques to perform predictive analytics on real-time and offset data. The model runs on different formations separately. It is mentioned that this method does not need data sources in terms of rock strength, drillability, etc.

Evangelatos and Payne (2016) conducted a new bottom hole assembly (BHA) and ROP model using ANN on drilling data. Several drilling parameters were used as input data for the model to correctly determine the rock and the bit interaction. As a result, this model could let the projection of bit response with different BHAs and different drilling parameters.

Kristjansson et al. (2016) used offset data to enhance drilling performance by using pattern recognition techniques. Several drilling parameters were analyzed from offset data to be created real-time drilling parameter schemes. These schemes are then used in comparison with trial wells. The results indicated a significant increase in the ROP and remarkable savings on drilling costs.

Mantha and Samuel (2016) developed a new model including regression analysis and ANN to estimate ROP using actual drilling data in horizontal wells and one particular formation type. This model used predictive methods such as step-wise regression (SWR), ANN, SVR, classification regression trees, and RF. The structure of the model is constructed to predict the values of future wells with remarkable accuracy.

Moraveji and Naderi (2016) performed and optimized the ROP prediction using the response surface method and Bat Algorithm (BA). Response surface methodology was used to engender the mathematical expression between the ROP and six parameters: depth, WOB, RPM, jet impact force, yield point/plastic viscosity ratio, and 10-min/10-sec gel strength ratio. Then, a BA was used to determine the ranges of these parameters to maximize the ROP. The results show that ROP can be predicted accurately within a 95% confidence interval.

Jiang and Samuel (2016) made an ROP optimization by combining the ANN and ant colony optimization in complex drilling groundwork. Bayes type ANN algorithm was compared to the Warren ROP model. It is stated that ANN was found acceptable to evaluate ROP from parameters such as WOB, RPM, flow rate, depth, and gamma-ray.

Taylor et al. (2016) developed a benchmark process focused on ROP to improve the drilling efficiency. An autonomous process including real-time analysis was built to estimate the best drilling performance. A case study was presented to show the improvement in drilling efficiency. It is stated that the process allowed rig site personnel to take immediate action to decrease drilling inefficiency.

Shi et al. (2016) implemented the extreme learning machine (ELM) and upper-layer-solution-aware methods (USA) for ROP prediction. Formation type and mechanical properties, hydraulics, and bit properties are used as input parameters. ANN was used as a benchmark method to compare the results. All the methods are suitable for ROP prediction, while ELM and USA methods are more advantageous in terms of prediction accuracy, learning speed, and generalization.

Eskandarian et al. (2017) performed a data mining approach to predict ROP to minimize drilling operation costs. The authors used the Cubist method, RF, and monotone multi-layer perceptron (MLP) models to estimate the ROP. They verified the reliance of the model by using a 10-fold cross-validation technique and stated an excellent accuracy.

Hegde et al. (2017) compared physics-based and data-driven modeling approaches in terms of ROP estimation. The models used drilling surface parameters to predict the ROP. Data-driven models were based on ML algorithms. It is stated that data-driven models gave better results rather than other models.

Elkatatny et al. (2017) made an optimization on the ROP using ANN by taking mud properties under consideration as the basis. The authors developed a new empirical ROP model by using more than 3,000 data points of several parameters. They stated the model could estimate the ROP with very high accuracy. Furthermore, the authors indicated the importance of combining both mechanical and drilling fluid parameters for ROP estimation.

Amer et al. (2017) studied ROP prediction and bit selection in Nile Delta Area using ANN. The model inputs in this study were formation lithology, drilling parameters, and bit data. Actual field data were used in statistical analysis to expel outliers and low-quality data. This model's results showed relatively low values in terms of average absolute percent error, standard deviation, maximum error, and root mean squared error.

Bezminabadi et al. (2017) derived mechanical rock properties from well logs and used them to develop a new ROP prediction model with ANN and multiple non-linear regression (MNR). It is stated that the prediction accuracy of the model was increased after implementing the parameters regarding mechanical rock properties. It is also mentioned that ANN performance was slightly better than MNR in ROP prediction.

Hegde and Gray (2018) evaluated three models for drilling optimization based on several criteria. Models for ROP, torque on bit, and MSE were developed with a data-driven approach using the RF algorithm for several drilling parameters. A meta-heuristic optimization algorithm optimized the models to determine the optimum drilling parameters. It is stated that model optimization increased the average ROP by 28% and increased the MSE and torque. Then, optimizing the model based on MSE resulted in ROP increment while observing a decrease in MSE and torque, resulting in longer bit life and cost savings. Furthermore, hypothesis testing was used to prove the simulations' validity by showing statistically significant results.

Hegde et al. (2018a) applied several optimization algorithms to an objective function derived from a data-driven ROP model. WOB, RPM, flow rate, and rock strength are stated as the input parameters. It is indicated that data-driven models are appropriate for real-time ROP optimization with proper algorithms.

Nadhan et al. (2018) studied the digital twin concept of a drilled well and analyzed the effect of the concept on planning and operations by introducing several case studies. It is stated that digital twinning directly influences both rig site and office engineers by making the personnel decide faster and more accurately. It is also mentioned that the concept helps achieve the maximum ROP and minimum non-productive time by allowing real-time optimization.

Elkatatny (2018) introduced the usage of the self-adaptive differential evolution method to optimize ANN parameters. The method was used to predict the ROP by using various drilling and fluid parameters. It is stated that ROP had a strong relationship with RPM, WOB, and torque. The validity of the model was shown by the importance assignment of the drilling and fluid parameters.

Hassan et al. (2018) investigated different approaches to calculate the torque and drag and develop an ROP model. In the study, ROP was estimated with real-time torque data and MSE calculations. The importance of several parameters is investigated

through statistical analysis. An ANN model was built to predict ROP and torque. It is mentioned that the model lets the drilling parameters being optimized in real-time.

Al-AbdulJabbar et al. (2018) built an ROP model with ANN using the minimum number of inputs. More than 4,500 data points were gathered from the field for modeling. The model inputs were stated as flow rate, WOB, RPM, torque, standpipe pressure, and Uniaxial Compressive Strength (UCS). Data from one well was used to train and test the model. Unseen data from two other wells are used to test the model. It is stated that ANN is an effective tool to estimate ROP with a high value of correlation coefficient.

Nasir and Rickabaugh (2018) applied several predictive models such as random forests, SVM, and boosted trees to develop an ROP model. It is mentioned that random forests work faster and more accurately than ANN. It is also stated that the models used in the study require less time and resources when compared to ANN.

Abdulmalek et al. (2018) predicted ROP by using several drilling parameters and drilling fluid properties. Data was gathered from shale sections of more than 400 wells to estimate the ROP with SVM. The SVM method surpassed the theoretical ROP models in terms of high correlation coefficient and low error.

Hegde et al. (2018b) explored deterministic, data-driven, and hybrid approaches to model the ROP. The hybrid model was formulated by using ensembles of single and deterministic models to introduce the physics of drilling to the ML algorithms, such as trees, k-Nearest Neighbors (kNN). It is mentioned that hybrid models perform ROP predictions more accurately than the deterministic model and have higher interpretability than the data-driven model. Moreover, the hybrid model was used to obtain optimum drilling parameters, such as WOB, RPM, and flow rate in specific types of formations.

Abbas et al. (2019a) performed a bit selection optimization to achieve the maximum ROP. An ROP prediction model was developed using ANN with numerous drilling parameters. Optimization was performed via GA to estimate the optimum parameters for choosing the appropriate bit that achieves maximum ROP. It is stated that the method lets the rig site engineer choose the best option among several cases to maximize ROP.

Hegde et al. (2019) introduced a methodology, which combines the ROP optimization with an ML-based vibration modeling that estimates optimum drilling parameters. Analytical models were used to model the ROP while optimizing the WOB and RPM with a gradient ascent algorithm. The model was validated in a field by showing an increment in ROP by about 14% on average.

Abbas et al. (2019b) proposed a computational approach to predict the ROP with ANN. The feature importance of several input parameters was ranked and selected accordingly. The method was also used to examine the effect of WOB, RPM, and flow rate on the ROP. The model results were compared with the field measurements.

Sabah et al. (2019) compared several ML algorithms to assess the performance in ROP prediction. The noise was removed by using a smoothing filter. GA was used to fine-tune the hyperparameters of ML algorithms by using numerous drilling parameters as inputs. It is stated that ROP was predicted accurately by all of the algorithms. It is also mentioned that prediction accuracy was increased when filtered data was used in the models.

Ashrafi et al. (2019) developed hybrid AI models to predict ROP with high accuracy. Data was gathered from a well drilled in Southern Iran, and the noise was removed by applying a smoothing filter. GA was applied to perform the feature selection. Eight hybrid ANN models were built and trained by several algorithms. MLR and MNR were used as the benchmark to compare the results. It is stated that hybrid ANN models are more efficient than conventional neural networks in terms of ROP prediction accuracy.

Gan et al. (2019a) introduced an AI model to predict the ROP. The geological properties were introduced to the model. The noise was detected by frequency spectrum analysis and then reduced by filtering. The model inputs were determined, and a hybrid bat algorithm was applied to tune the SVM hyperparameters. The model was verified in field operations.

Soares and Gray (2019) investigated the real-time ROP prediction potential of several models. Data were acquired simultaneously during drilling, yet it is assumed that models could learn to improve the prediction accuracy. It is stated that ML models could produce low testing errors in ROP prediction. It is also mentioned that cross-validation is a suitable approach to determine the best real-time ROP prediction model.

Gan et al. (2019b) proposed a new hybrid bat algorithm to maximize the ROP by establishing a data-driven model with the combination of wavelet filtering and SVM. Several modifications were applied to the model to improve the global optimization performance. The model was validated through a field operation and showed about a 35% increment in the ROP.

Zhao et al. (2020) developed several ANN-based real-time ROP prediction models. The best ROP model was selected by comparing the results with various statistical comparison criteria. Additionally, an artificial bee colony algorithm was implemented to optimize the drilling parameters. The model was verified in field operations.

Mehrad et al. (2020) introduced a combination of several predictive algorithms such as least-squares SVM, cuckoo optimization algorithm, particle swarm optimization, and GA for developing a model ROP prediction with standard MLU parameters and several geomechanical parameters. Outliers were detected via histograms and Tukey's method. Feature selection was performed with GA. It is indicated that modeling error was increased when the number of input parameters was increased. Hence, only six parameters were used as inputs. It is stated that least-squares SVM outperformed the other models by resulting in high prediction accuracy.

Elkatatny (2020) evaluated the suitability of adaptive neuro-fuzzy inference systems (ANFIS), functional neural networks, RF, and SVM for real-time ROP prediction in S-type directional wells. Data were gathered from three wells. Two of the wells were used for training and testing, and another well was used for validation purposes. As a conclusion, ANFIS, functional neural networks, and RF were stated as the best ROP prediction models in terms of minimum error.

Kor and Altun (2020) showed the applicability of SVR to predict ROP when the number of input data is limited. In the study, two kinds of regression techniques were used (MLR and SVR), and the results were compared to each other. For the calculations, a field data set taken from the literature was used. The dataset was trained and tested based on different data selection cases. The results of several predicting methods for each dataset were compared to each other to demonstrate the effectiveness of the support vector regression method subjected to fewer input data. The results showed the significant effect of data selection on the accuracy of penetration rate prediction. This study is published within the scope and studies of this dissertation.

One of the most powerful methods in machine learning, SVR, was used for ROP optimization for the first time in the literature with this study. In this way, the chance for further investigations and studies on the practicability of SVR on penetration rate optimization is created.

2.5 ROP Optimization in Heterogeneous Formations

Al-Saleh (1999) conducted a study in Kuwait, including optimization of the whole drilling process starting from the drilling design. It is stated that the optimization resulted in less average cost per footage.

Mensa-Wilmot et al. (2004) introduced a new PDC bit technology by presenting the improvements in the analysis of formation drillability in highly heterogeneous formations. It is addressed the effectiveness of the improvement of the drilling performance as a result of this technology.

Rastegar et al. (2008) optimized multiple bit runs established on ROP models and cost per foot equation in Persian Gulf carbonate fields. UCS was used in terms of footage drilled to calculate the ROP for using in bit design and construct the learning curve to optimize bit hydraulics and drilling parameters. It is stated that this optimization can reduce the drilling costs if there are offset wells in the same field.

Thomson et al. (2011) used a hybrid bit to improve Brazil's heterogeneous formations' drilling efficiency. It is indicated that a hybrid bit can achieve higher ROP than standard or PDC bits. Furthermore, it is stated that a 35% decrease in time spent on drilling in heterogeneous formations with this hybrid bit.

Zhu et al. (2012) tested several mineral components to investigate the lithological properties of heterogeneous deep-sea formations of the oil fields in the south-western China Sea. The authors developed a prediction model for rock mechanics parameters formed on UCS, drillability, hardness, plasticity, and core samples' abrasiveness taken from interest formation. Based on three characteristics, a bit structure was designed and used. The results showed a significant increase in average ROP and footage.

Okutucu et al. (2013) performed a field experience of a new generation of a hybrid bit used in the south-eastern region of Turkey. The authors indicated that the bit runner economized several TCI bit runs and achieved two times faster ROP than that of the

most successful TCI bit run in the same field. The same well is drilled faster with this new generated drilling bit within the same highly complex formations.

Pinheiro et al. (2015) built a new drilling optimization methodology that works in pre-salt formations in Brazil, highly heterogeneous. This methodology was based on simulators, drill string vibration, MSE, reamer and shock absorber usage, drilling fluids optimization, and real-time monitoring. It is indicated that the target depth arrived 60 days earlier if the scheduled program was based upon this methodology.

Kutas et al. (2015) made an application of BYM in Brazilian pre-salt formations. It is indicated the importance of the normalization factors within the functions of the method. Also, it is stated that this method can be used in limited functionality in these formations because of the complex characteristics of pre-salt layers.

Miao et al.(2016) established a field run of a hybrid drilling bit used in heterogeneous formations of China oil fields. The authors stated that several TCI runs were conserved and drilled 100% more and reached two times faster ROP within the same well in the same oil field.

Guo et al. (2017) generated an ROP optimization by analyzing formation characteristics and offset well data in poor-drillable southern China Sea formations. By considering the previous wells drilled, an ROP optimization design including PDC development was conducted to overcome the drillability problems in these heterogeneous formations. It is stated in the results that this optimization had a significant effect on reducing the drilling costs.

Ahmed et al. (2019) explored the ROP prediction efficiency of several predictive models such as ANN, ELM, SVM, and least-squares SVM with the data taken through two wells. All the data is normalized in order to eliminate data scatterness due to litgological differentiation. It is stated that all of the models performed ROP prediction within an acceptable range of accuracy, where least-squares SVM outperformed the other models in terms of a higher correlation coefficient.

Yuswandari et al. (2019) used ANN to predict ROP in a geothermal field containing heterogeneous formations. ANN was conducted for training until the convergence criteria, while the relationship between the drilling parameters was established. It is stated that the prediction accuracy of ANN was affected mainly by depth, WOB, and RPM. It is also mentioned that modeling the parameters using the data from one well

and predicting the ROP in another well resulted in high errors. However, it is stated that this process generated a general idea about the ROP estimation between nearby wells.

Gan et al. (2019c) introduced an AI method to predict ROP in complex environments by taking missing data, duplication, and non-linearity under consideration. Missing data were completed by cubic Hermite interpolation. The formation drillability model was built with ELM by determining the highly correlated parameters with the ROP. ANN-based ROP prediction model was optimized by particle swarm algorithm. The model was validated in the field and stated that the study's method outperformed conventional ROP prediction methods.

Elkatatny (2021) developed an empirical equation to predict ROP in real-time with several AI methods, such as ANN, ANFIS, and SVM. PDC design parameters, total flow area, mud weight, gamma-ray log readings, and numerous drilling parameters were used as inputs in the model. Data were acquired from two wells. One well was used for training, and the other well was used for validation. All the methods could perform ROP prediction in heterogeneous formations with high accuracy, while ANN was selected as the best method.

2.6 General Overview and Interpretations

In the light of the literature review within the scope of ROP prediction and optimization, the reviewed studies are mainly based on analytical, empirical, and statistical optimization methods. Regarding computer and information technology improvement, real-time data monitoring and analysis have gained importance and have been used in drilling automation to achieve the maximum ROP. Hardware designing looks like the superior subject on ROP maximization studies for heterogeneous formations. On the other hand, there is not any generalized model to optimize the drilling parameters. Nevertheless, BYM is one of the most commonly used ROP prediction and optimization models, which creates a basis for most recent ROP optimization studies.

There are numerous studies regarding ML applications with different approaches for ROP prediction. A comparison matrix for the studies considering the applications of the ML methods for ROP prediction is given in Table 2.1.

Table 2.1 : Comparison matrix for the studies regarding ML applications for ROP prediction.

Reference	ML Methods	# of Inputs	Compared w/ BYM	Considering Heterogeneity	Considering Real-Time
Bilgesu et al., 1997	ANN	10	No	No	No
Bjornsson et al., 2004	ES	8	No	No	No
Bahari et al., 2008	GA	8	Yes	No	No
Moradi et al., 2010	Fuzzy	8	Yes	No	No
Bataee and Mohseni, 2011	ANN	5	No	No	No
Arabjamaloei and Shadizadeh, 2011	ANN,GA	7	Yes	No	No
Arabjamaloei and Dehkordi, 2012	ANN,ANFIS	11	No	No	No
Esmaili et al., 2012	ANN	18	No	No	No
Amar and Ibrahim, 2012	ANN	7	Yes	No	No
Gidh et al., 2012	ANN	12	No	No	Yes
Monazami et al., 2012	ANN	13	No	No	No
Jamshidi and Mostafavi, 2013	ANN,GA	6	No	No	No
Yi et al., 2014	SFLA	3	No	No	Yes
Valisevich et al., 2015	ANN	5	No	No	Yes
Wallace et al., 2015	RF,ANN	4	Yes	No	Yes
Evangelatos and Payne, 2016	ANN	13	No	No	No
Kristjansson et al., 2016	Pat. Rec.	9	No	No	Yes
Mantha and Samuel, 2016	ANN,SVR,SWR,RF	4	No	No	No
Moraveji and Naderi , 2016	BA	6	No	No	No
Jiang and Samuel, 2016	ANN	5	No	No	No
Shi et al., 2016	ANN	10	No	No	No
Eskandarian et al., 2017	RF,ANN	6	No	No	No
Hegde et al., 2017	RF	4	No	No	No
Elkatatny et al., 2017	ANN	7	Yes	No	No
Amer et al., 2017	ANN	20	No	No	No
Bezzminabadi et al., 2017	ANN	9	No	No	No
Diaz et al., 2018	ANN	8	Yes	No	No
Hegde and Gray, 2018	RF	4	No	No	No
Hegde et al., 2018a	RF	4	No	No	Yes
Hegde et al., 2018b	RF,kNN	3	Yes	No	Yes
Elkatatny, 2018	ANN	8	No	No	No
Hassan et al., 2018	ANN	6	No	No	No
Al-AbdulJabbar et al., 2018	ANN	6	Yes	No	No
Nasir and Rickabaugh, 2018	RF,SVR	3	No	No	No
Abdulmalek et al., 2018	ANN,SVM	12	Yes	No	No
Abbas et al., 2019a	ANN,GA	19	No	No	No
Abbas et al., 2019b	ANN	19	No	No	No
Hegde et al., 2019	Gradient	4	Yes	No	Yes
Sabah et al., 2019	ANN,RF,SVR	11	No	No	No
Ashrafi et al., 2019	ANN,GA	11	No	No	No
Gan et al., 2019a	ANN,RF,SVR	6	No	No	No
Soares and Gray, 2019	ANN,SVM,RF	4	Yes	No	Yes
Gan et al., 2019b	Bat	4	No	No	Yes
Gan et al., 2019c	ANN	7	No	Yes	Yes
Zhao et al., 2020	ANN	7	No	No	Yes
Mehrad et al., 2020	SVM,GA	6	No	No	No
Elkatatny, 2020	ANFIS,ANN,SVM,RF	5	Yes	No	Yes
Kor and Altun, 2020	SVR	8	Yes	No	No
Ahmed et al., 2019	ANN,SVR	8	No	Yes	No
Yuswandari et al., 2019	ANN	7	No	Yes	No
Elkatatny, 2021	ANN,ANFIS,SVM	11	No	Yes	No

As seen in Table 2.1, there are total of 51 studies using ML methods for ROP optimization and prediction. Nevertheless, only 4 of these studies consider formation heterogeneity. Plus, none of them is neither compared with nor based on BYM.

Heterogeneous drilling environments make prediction accuracy of BYM decrease substantially and make the BYM impractical. Most of the studies for heterogeneous formations consider of designing PDC cutters and hybrid drilling bits to obtain maximum achievable ROP within a single bit run. However, these approaches do not

consider the functional relationships of various drilling parameters defined in BYM. In the light of these explanations, implementing a different type of regression model as an alternative to MLR by modifying BYM to predict ROP in heterogeneous formations is the main purpose of this study. This is done by introducing several geomechanical properties and torque data to the BYM to estimate a relationship between the rock mechanical properties and the heterogeneity. Therefore, this dissertation has its unique aspect in the literature.



3. STATEMENT OF THE PROBLEM AND SCOPE OF THE THESIS

It is acknowledged that there is a strong relationship between maximizing the penetration rate (ROP) and minimizing the drilling costs. Hence, it is crucial to predict the ROP accurately in both well planning stage and during operations. There is no widely accepted model for defining the optimum ROP since various variables affect the cost and time relationship.

It is clear that maximizing ROP and minimizing drilling costs is a multi-parameter optimization problem that requires advanced optimization techniques. Several studies were conducted to determine the maximum ROP and minimum cost. There exist some mathematical and statistical models to optimize ROP and various drilling parameters. The most commonly used ROP model in the drilling industry is Bourgoyne and Young Method (BYM) (Bourgoyne Jr. and Young Jr., 1974). In the BYM, the aim is to calculate eight regression coefficients related to standard drilling parameters by applying Multiple Linear Regression (MLR) analysis to drilling data for predicting ROP. The drilling parameters considered in the BYM are drillability, depth, equivalent circulation density, pore pressure gradient, weight on bit, rotary speed, bit tooth wear, and hydraulics.

In theory, eight input data are sufficient to estimate eight correlation coefficients for eight variables (or parameters). This situation is only valid for the cases representing the drilling process with perfection, which is a known fact that this never occurs in actual drilling. Moreover, using too few input data will result in negative coefficient values (Bourgoyne Jr. and Young Jr., 1974, p. 374). Therefore, a minimum number of input data for a given number of parameters must be present to build the model and estimate the correlation coefficients with maximum accuracy. Accordingly, as the number of parameters decreases, the minimum number of input data also decreases.

A critical and vital issue for BYM considering eight drilling parameters is that minimum of thirty data sets must be acquired within short intervals (2 to 5 meters) in one type of homogeneous formation (especially shale) from a single well or from different depths of the homogeneous formations (shale sections) in various wells

(Bourgoyne Jr. and Young Jr., 1974). This can be achievable for fields with thick shale sections such as Offshore Louisiana area in which the original study was performed. However, drilling through formations with extreme complexity (non-uniform) in lithology is confronted frequently worldwide, particularly in the exploration fields such as Turkey (Okay, 2008). These kinds of drilling environments make prediction accuracy of BYM decrease substantially and simply make the BYM impractical because of the formation heterogeneity, such as: (i) absence of requisite shale sections in a single well to collect enough amount of drilling data, (ii) an insufficient number of wells in one particular field containing enough shale sections, (iii) drilling through a heterogeneous formation having extreme complex lithology. Encountering either one of these cases will be eventuated in data restriction. When drilling data is not sufficient to correlate accurately, the reliability of predictions with MLR decreases. To overcome this problem, another predictive method rather than MLR must be used to predict ROP better.

Machine learning methods have been used in petroleum engineering since the mid-1980s. Artificial intelligence methods have a wide range of petroleum industry applications (Kor and Altun, 2020). Among the machine learning methods used in petroleum engineering studies, the most frequently used methods are Support Vector Machines (SVM) and ANN.

According to the literature, the MLR does not give accurate and meaningful results in ROP prediction where complex lithologies dominate. Artificial Neural Network (ANN) in the petroleum literature mainly performed the work done using the machine learning techniques. Researchers have expressed that ANN gives more meaningful and correct results than MLR. However, the number of input data used in the ANN method is about 2-3 times higher than the number of data required by the SVM method (Al-Baiyat and Heinze, 2012). Moreover, the Support Vector Machine (SVM), another machine learning model, has not been used in ROP prediction using the parameter functions defined in BYM until now. It is also reported that SVM gives faster and more accurate results than ANN by using fewer input data in researches for other subjects (Jahanbakhshi et al., 2012). In this context, this study, which includes the SVM's use to predict the ROP in heterogeneous formations by directly applying to BYM functions, will be a first and unique aspect in the literature.

The BYM involves eight parameters and requires sufficient amount of input data points obtained from the formations with uniform lithology with statistically at least thirty different properties. The number of input data required to obtain such a large number of parameters may not always be possible. This is especially the case in drilling environments where there are not many shale zones, and where there are few wells in a field, or where complex lithology is dominant such as in the fields of Turkey. Thus, the results obtained by using the BYM in these limited conditions are meaningless, observations obtained from the field practice, and method results give very different values. On the other hand, the SVM method effectively predicts ROP with higher sensitivity and conservation without decreasing the parameters. It is the main reason that establishes the purpose of this dissertation.

As a summary of the literature survey, it is clarified that the recent studies about the ROP prediction and optimization consider uniform lithology, i.e., homogeneous shale formations. Consequently, none of these studies brings a solution to the ROP prediction problems encountered in non-uniform lithology, i.e., heterogeneous formations. Therefore, this study aims to close this gap by introducing an approach to predict ROP more accurately in the well programming stage for complex or non-uniform lithologies by using offset drilling data.

The main purpose of this study is to implement a different type of regression model as an alternative to multiple linear regression by modifying BYM by introducing several additional geomechanical parameters to estimate a relationship between the rock mechanical properties and the heterogeneity, which are shale index (SI), uniaxial compressive strength (UCS), brittleness index (BI), shear failure gradient (SFG), and torque. In this study, several predictive methods are used to build a model for ROP prediction: multiple linear regression (MLR), support vector regression (SVR), and artificial neural networks (ANN). Feature selection is performed via random forests (RF) method. The prediction accuracy of each predictive model is compared by using several statistical comparison criteria: root mean square error (RMSE), correlation coefficient, and statistical significance (p-value).

The studies' findings in this dissertation are expected to provide significant time savings and cost savings since it is expected that faster and more accurate results will be achieved than those of currently available methods. The results will provide a basis for monitoring the condition of a drill bit on a real-time basis during any drilling

operations and determine the best time to change the drill bit while avoiding a possible increase in the overall drilling cost.



4. BOURGOYNE AND YOUNG METHOD

There are numerous rate of penetration (ROP) prediction models. However, Bourgoyne and Young ROP model (Bourgoyne Jr. and Young Jr., 1974) is the most commonly used ROP prediction and optimization method in the industry. Bourgoyne and Young Method (BYM) takes a non-linear relationship of various drilling parameters under consideration to predict ROP. The parameters in BYM are (1) drillability, (2) depth, (3) pore pressure gradient, (4) equivalent circulation density (ECD), (5) weight on bit (WOB), (6) rotary speed, (7) bit tooth wear, and (8) hydraulics. The parameters consist of the BYM is shown in Figure 4.1.

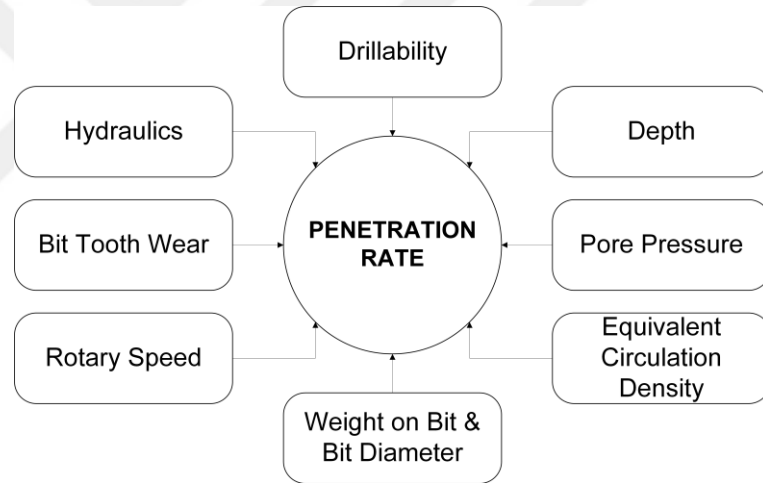


Figure 4.1 : Parameters of the BYM (Kor and Altun, 2020).

The aim is to estimate ROP for desired conditions by multiple regressing the offset data consisting of the eight parameters in Figure 4.1. ROP equation in the form of the functional relationship of the drilling parameters can be expressed as in equation 4.1.

$$R = (f_1)(f_2) \dots (f_n) \quad (4.1)$$

Each f term in equation 4.1 stands for a function of relevant drilling parameters. According to the least-squares principle, equation 4.1 can be written as equation 4.2 in MLR form.

$$R = \exp \left(a_1 + \sum_{j=2}^8 a_j x_j \right) \quad (4.2)$$

In equation 4.2, a_1 through a_8 are regression coefficients that belong to one particular field. x_2 through x_8 are the functions belonging to several drilling parameters.

4.1 Default BYM Parameters

In this section, default BYM parameters that affect ROP are introduced.

4.1.1 Effect of drillability

In the BYM, constant a_1 directly represents the drillability (formation strength). It is mentioned that drillability is inversely proportional to the natural logarithm of the square of the constant described by Maurer (1962) (Bourgoyne Jr. and Young Jr., 1974). The a_1 constant also corresponds to other drilling parameters not modeled in BYM, such as mud type and solid content.

4.1.2 Effect of normal compaction

Murray and Cunningham (1955) introduced the exponential relationship between the ROP and well depth in a normal-compacted formation through micro-drilling bit experiments. The exponential trend was also mentioned by Combs (1968) via actual field data. The ROP and depth relationship is shown in Figure 4.2.

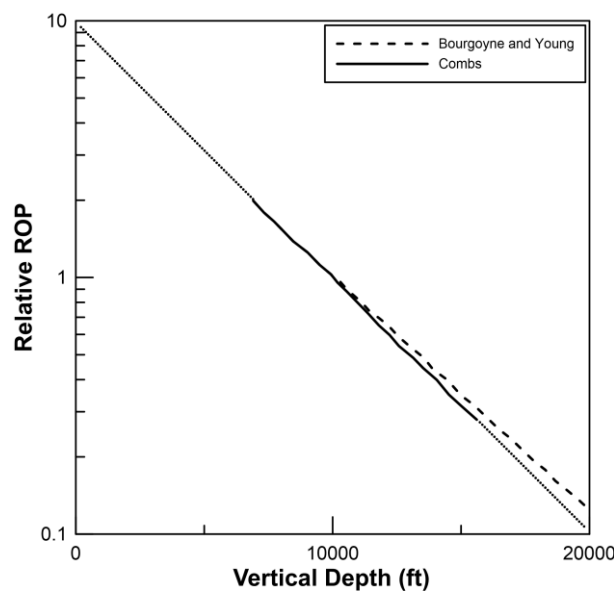


Figure 4.2 : Effect of compaction, adapted from (Bourgoyne Jr. and Young Jr., 1974, p. 372).

The effect of normal compaction is expressed as in equation 4.3.

$$x_2 = 10000 - D \quad (4.3)$$

In equation 4.3, D is well depth (ft), which is normalized for 10,000 ft.

4.1.3 Effect of under compaction

Under compaction is the consequence of the pore pressure. It is accepted that ROP is increased exponentially when the pore pressure gradient is increased. The compaction theory explains the phenomenon. The effect of under compaction is expressed as in equation 4.4.

$$x_3 = D^{0.69} (g_p - 9) \quad (4.4)$$

In equation 4.4, D is well depth (ft), g_p is the pore pressure gradient (pounds per gallon, ppg or lbm/ft), which is normalized for 9 ppg.

4.1.4 Effect of overbalance

Overbalance is the pressure differential across the bottom hole. As indicated in several studies, there exists an exponential relationship between the ROP and the bottom hole pressure (Cunningham and Eenink, 1959; Garnier and van Lingen, 1959; Combs, 1968; Vidrine and Benit, 1968). The ROP and overbalance relationship is shown in Figure 4.3.

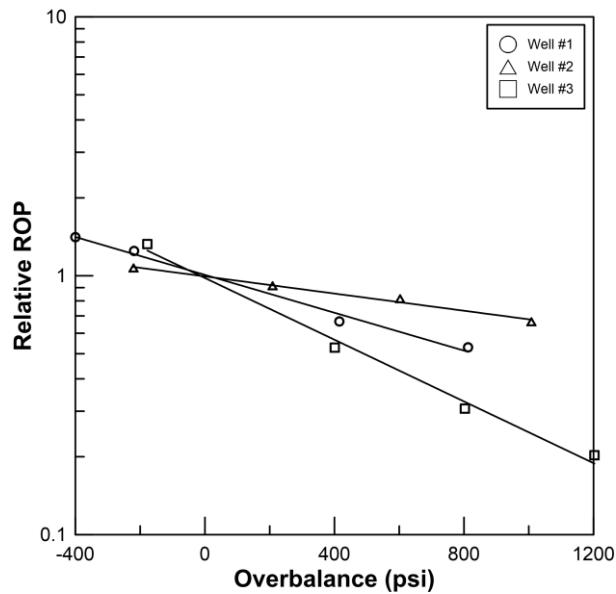


Figure 4.3 : Effect of overbalance, adapted from (Bourgoyne Jr. and Young Jr., 1974, p. 372).

The effect of overbalance is expressed as in equation 4.5.

$$x_4 = D(g_p - \rho_c) \quad (4.5)$$

In equation 4.5, D is well depth (ft), g_p is the pore pressure gradient (ppg), and ρ_c is equivalent circulation density (ECD) (ppg).

4.1.5 Effect of weight on bit per bit diameter

The effect of WOB relative to bit diameter on the ROP was investigated thoroughly in the literature. The general idea is to plot the ROP corresponding to variable WOB while the remaining drilling parameters stay constant. The ROP and WOB relationship is shown in Figure 4.4.

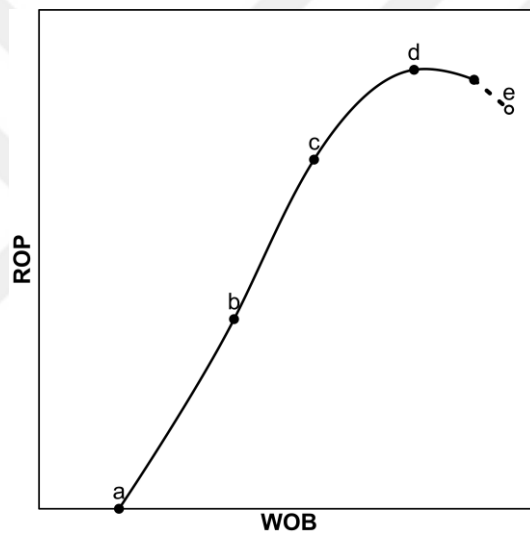


Figure 4.4 : ROP and WOB relationship, adapted from (Bourgoyne Jr. et al., 1991, p. 226).

In Figure 4.4, point (a) indicates the threshold WOB, section (a-b) shows the increment in ROP with increasing WOB, section (b-c) is the linear trend, section (c-d) represents the slight increment in ROP with higher WOB. Finally, section (d-e) shows the bit floundering with excessive WOB values (Bourgoyne Jr. et al., 1991).

The effect of WOB per bit diameter is expressed as in equation 4.6.

$$x_5 = \ln \left(\frac{(W_b/d_b) - (W_b/d_b)_t}{4 - (W_b/d_b)_t} \right) \quad (4.6)$$

In equation 4.6, W_b/d_b is the WOB per bit diameter (lb/in), $(W_b/d_b)_t$ is the threshold WOB per bit diameter (lb/in), which are normalized for 4,000 lbs/in.

The threshold WOB value can be determined with drill-off tests (DOT) performed at drill bit or formation changes to estimate the optimum WOB and rotary speed combination for the maximum ROP with a reasonable tooth wear rate (Lyons, 2010, p. 315). A DOT is performed by applying a high amount of WOB with a constant rotary speed, locking the brake, monitoring the decline in WOB, and recording time. The procedure is repeated by using a couple of different constant rotary speed values in short intervals. The optimum WOB and rotary speed selection are determined by selecting the combination that gives the shortest time, which means the highest ROP.

Moreover, it is stated that the threshold WOB can be neglected for the bits used in wells drilled through less than 5000 ft with relatively soft formations (Bourgoyne Jr. and Young Jr., 1974). Furthermore, since excessive WOB can make the well deviate unintentionally, an alternative controllable parameter, such as RPM should be optimized. The threshold WOB value is taken as zero in this study, as in the original study of BYM.

4.1.6 Effect of rotary speed

Rotary speed is relatively directly proportional to the ROP. The relationship between the ROP and rotary speed can be shown by plotting the ROP corresponding to variable rotary speed while setting other drilling parameters to constant. The ROP and rotary speed relationship is shown in Figure 4.5.

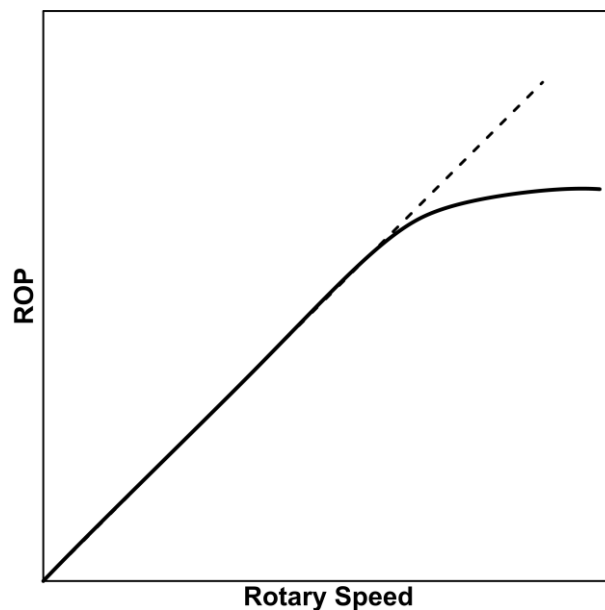


Figure 4.5 : ROP and rotary speed relationship, adapted from (Bourgoyne Jr. et al., 1986, p. 226).

In Figure 4.5, there is a linear response in the ROP with increasing rotary speed up to some values. The ROP starts to decrease after exceeding a limit. This limit is called the bit flounder point.

The effect of rotary speed is expressed as in equation 4.7.

$$x_6 = \ln\left(\frac{\omega}{100}\right) \quad (4.7)$$

In equation 4.7, ω is the rotary speed (rpm), which is normalized for 100 rpm.

4.1.7 Effect of bit tooth wear

Several studies about the effect of bit tooth wear on the ROP are present in the literature (Galle and Woods, 1963; Edwards, 1964). The previous models were much complex to implement in a regression analysis (Bourgoyne Jr. and Young Jr., 1974, p. 372). A comparison of the bit tooth wear models is shown in Figure 4.6.

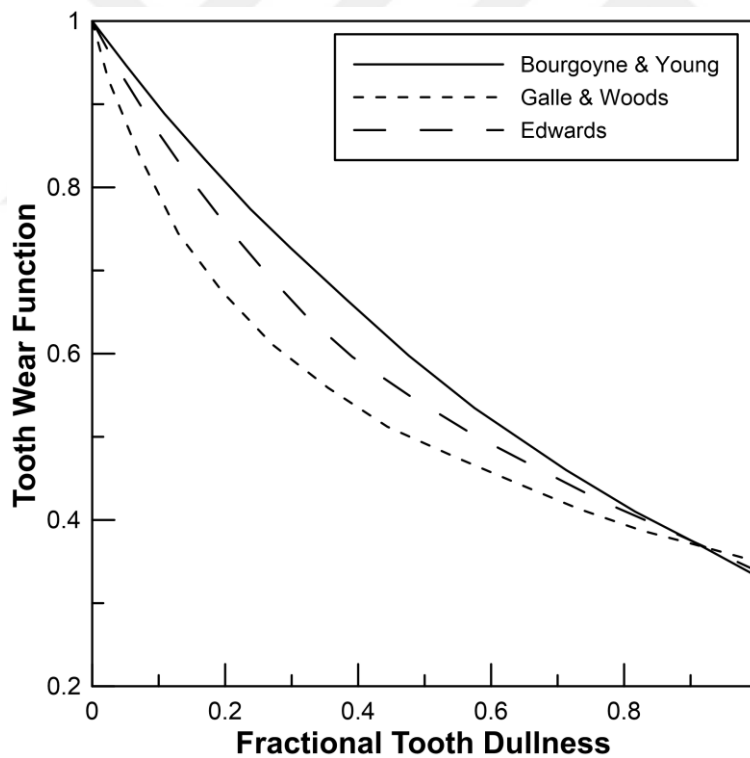


Figure 4.6 : Bit tooth wear model comparison, adapted from (Bourgoyne Jr. and Young Jr., 1974, p. 373).

In Figure 4.6, the most consistent bit tooth wear model is the Bourgoyne & Young model, in which the wearing exponent is bit and formation dependent (Bourgoyne Jr. and Young Jr., 1974, p. 372).

Essentially, fractional tooth wear is determined in the 0-8 scale via measuring the worn tooth height of a roller cone bit or estimating worn portion of the tungsten carbide insert (TCI) bits. A sample scheme of dull tooth grading is given in Figure 4.7.

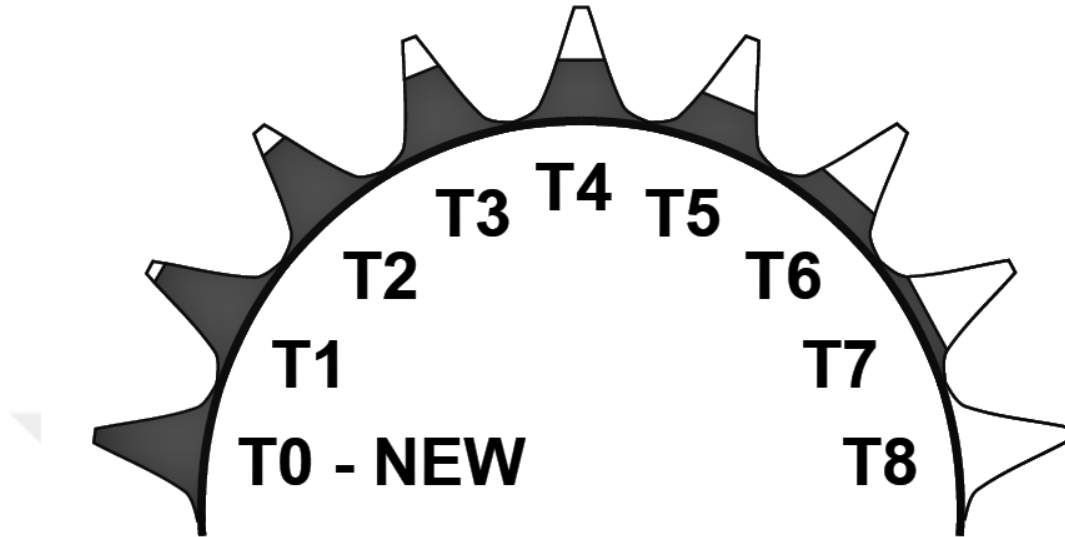


Figure 4.7 : Dull grading chart (Smith, 2008).

The effect of bit tooth wear is expressed as in equation 4.8.

$$x_7 = -h \quad (4.8)$$

In equation 4.8, h is the fractional tooth wear.

Additionally, the bit tooth wear model in the BYM is valid only in the presence of roller cone bits. It is stated the effect of tooth wear on the ROP can be negligible for polycrystalline diamond compact (PDC) or TCI bits (Bourgoyne Jr. and Young Jr., 1974, p. 372). Hence, the x_7 term should be neglected in these kinds of situations.

4.1.8 Effect of hydraulics

The effect of hydraulics on the ROP was first investigated by Eckel (1967) by introducing a Reynolds Number function that is directly proportional to the ROP. When the WOB increases, the behavior shifts upward by preserving its trend. Nevertheless, Eckel's model does not involve the phenomenon in the flounder point. Hence, this model is not recommended to use in field practice (Bourgoyne Jr. et al., 1991, p. 232).

It is known that hydraulic horsepower and jet impact force approaches are the most common hydraulic models used in the industry (Bourgoyne Jr. et al., 1991; Wright et

al., 2003). Additionally, Reynolds Number function, hydraulic horsepower, and jet impact force models give similar trends in ROP and hydraulics correlation, given in Figure 4.8.

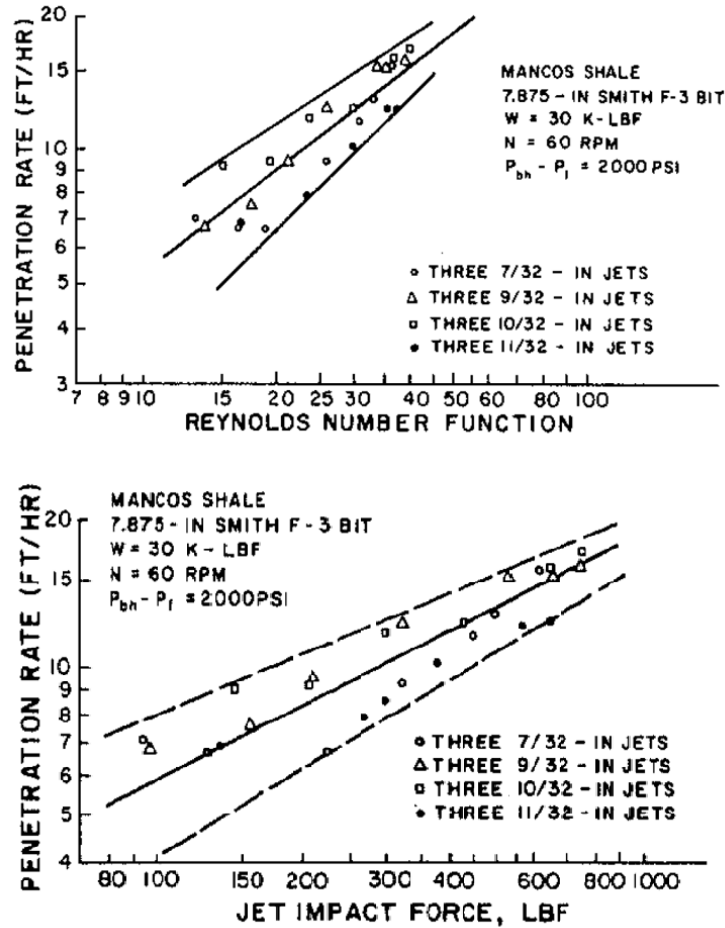


Figure 4.8 : Effect of hydraulic models (Bourgoyne Jr. et al., 1991, p. 231).

The effect of hydraulics is expressed as in equation 4.9.

$$x_8 = \ln \left(\frac{F_j}{1000} \right) \quad (4.9)$$

In equation 4.9, F_j is the jet impact force (lbf), which is normalized for 1,000 lbf.

4.2 Additional Features from Geomechanical Properties

Several additional features should be added to the ROP prediction model to improve the prediction accuracy for the BYM in heterogeneous formations. In this study, four new features related to geomechanics and rock mechanics are introduced and

evaluated: shale index, uniaxial compressive strength, brittleness index, shear failure gradient, and torque.

4.2.1 Shale index

Gamma-ray logs can be used to estimate shale volume since shales are more radioactive than sand and carbonate rocks. Figure 4.9 represents the response of gamma-ray readings to different types of lithology.

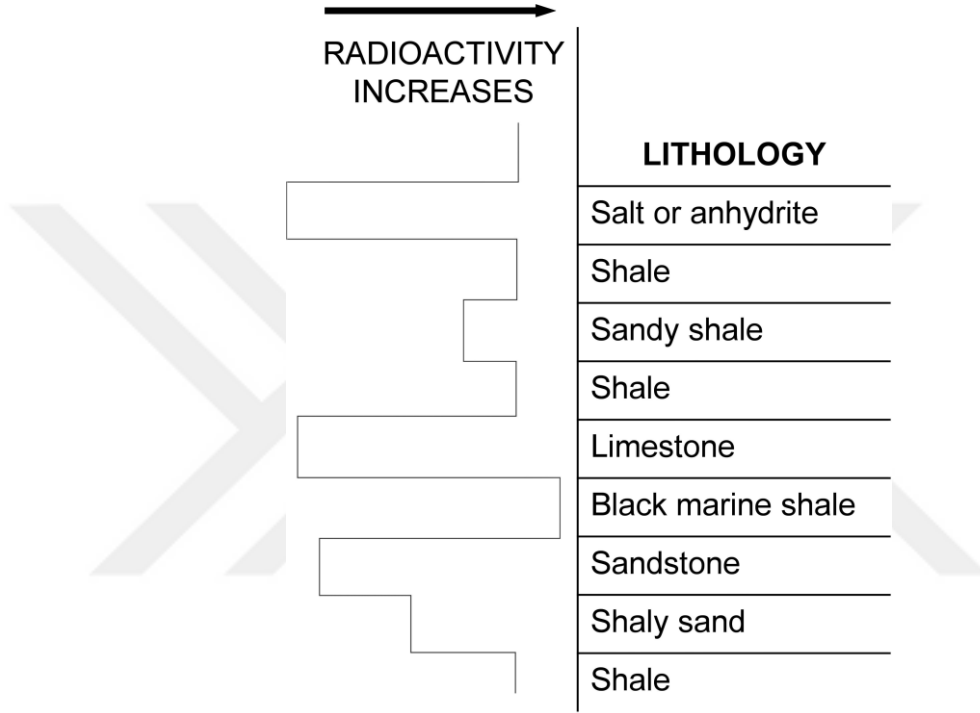


Figure 4.9 : Gamma-ray responses for different types of lithology, adapted from (Bassiouni, 1994).

The first step needed to determine V_{sh} from a gamma-ray log is to calculate the gamma-ray index (I_{GR}). There are several non-linear empirical correlations for estimating I_{GR} from a gamma-ray log. These non-linear correlations are established by taking geographic area or formation age under consideration. In this study, a non-linear shale index modeling (Larionov older rocks model) is used, which can be expressed as equations 4.10 and 4.11.

$$I_{GR} = \frac{GR - GR_{\min}}{GR_{\max} - GR_{\min}} \quad (4.10)$$

$$V_{sh} = 0.33(2^{2I_{GR}} - 1) \quad (4.11)$$

In equation 4.10, I_{GR} is the gamma-ray index, GR is the gamma-ray log reading (API), and GR_{min} and GR_{max} are the minimum and maximum of the log readings of a specific formation (API). In equation 4.11, V_{sh} is the shale index (or shale volume).

4.2.2 Uniaxial compressive strength

Uniaxial compressive strength (UCS) is a measure of the strength of a material. UCS is also named as the unconfined compressive strength since confining pressure is zero.

UCS is measured by an unconfined compression test. During the test, the right-cylindrical sample is only applied compression lengthwise. The confining stress is set to zero, while the axial stress is applied at a constant rate. The maximum stress recorded during the test resulted in the UCS (Agustawijaya, 2007). Based on these loading tests, several empirical correlations have been developed to estimate UCS using different input variables such as elastic modulus, porosity, formation hardness, compressional wave velocity, density, etc. (Briševac et al., 2016).

In this study, UCS is calculated via equation 4.12 (Şahin, 1995).

$$C_0 = 3.464 \times \frac{0.026 \times E \times 10^6}{\Delta C} \times (0.008 \times V_{sh} + 0.0045 \times (1 - V_{sh})) \quad (4.12)$$

In equation 4.12, C_0 is UCS (psi), E is Young's modulus (Mpsi), ΔC is bulk compressibility (μ psi), V_{sh} is shale index.

4.2.3 Brittleness index

The brittleness index (BI) concept is developed from a rock's tendency to fail in brittle type, which means the opposite of ductile failure where inelastic strain overcomes elastic strain that causes catastrophic deformation (Mathia et al., 2016). BI is defined to classify rocks based on the tendency to fail brittle or being ductile. There are several empirical equations to define BI. In this study, BI related to Young's modulus and Poisson's ratio is used, which can be calculated via equation 4.13 (Rickman et al., 2008).

$$BI = 0.5 \times \left[\left(\frac{E - E_{min}}{E_{max} - E_{min}} \right) + \left(\frac{N - N_{max}}{N_{max} - N_{min}} \right) \right] \quad (4.13)$$

In equation 4.13, BI is brittleness index, E is Young's modulus (Mpsi), N is Poisson's ratio.

Young's modulus and Poisson's ratio can be estimated from compressional/shear sonic logs by using equations 4.14, 4.15, and 4.16.

$$N = \frac{0.5(v_p/v_s)^2 - 1}{(v_p/v_s)^2 - 1} \quad (4.14)$$

$$E = 2G(1 + N) \quad (4.15)$$

$$G = 1.34 \times \frac{\rho_b}{\Delta t_s^2} \times 10^{10} \quad (4.16)$$

Here E is Young's modulus (Mpsi), N is Poisson's ratio, v_p and v_s are compressional and shear wave velocities (m/s), G is shear modulus (Mpsi), ρ_b is bulk density (g/cm^3), and Δt_s is shear sonic log reading ($\mu\text{s/ft}$).

4.2.4 Shear failure gradient

In the beginning, rock strength is in balance with the overburden and confining stresses. During drilling, this equilibrium state is deteriorated by the interaction between the formation fluids and drilling mud, resulting in borehole instability. Mechanical instability occurs when the stresses exerting on a rock surpass the compressive or the tensile strength. Additionally, under the action of combined stresses, failure occurs when the energy of distortion reaches the same energy of failure of the rock under tensional force (McLean and Addis, 1990). A sample illustration of the borehole failure mechanism is given in Figure 4.10.

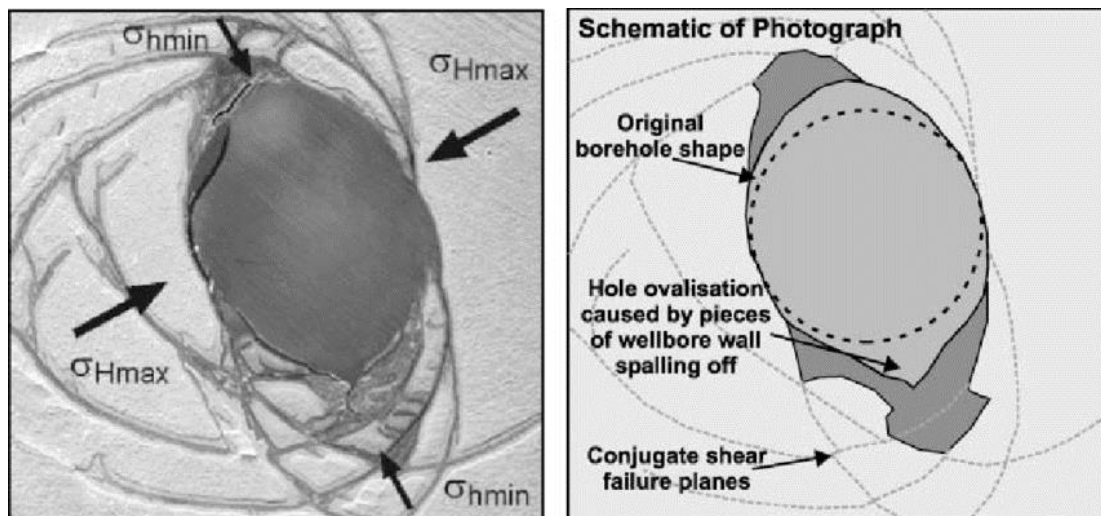


Figure 4.10 : Borehole failure mechanism (Tingay et al., 2005).

Shear stresses cause compressive failure as a result of the usage of low mud weight. Shear strength is the minimum tensile strength that encounters the borehole stresses during drilling. The safe mud weight window to prevent any fluid loss or kick can be defined as the envelope between the pore pressure and the fracture gradients. Moreover, the mechanical safety window to eliminate the shear and tensile failure can be expressed as the envelope between minimum horizontal stress and shear stress (Sevinç, 2020, p. 34).

In compressive failure, the most common criteria to predict instability is the Mohr-Coulomb criterion. However, it is recommended to use the Stassi-D'Alia failure criterion (Stassi-D'Alia, 1967) for the failure calculations in shale formations (Stevenson et al., 2013; Sevinç, 2020). Stassi-D'Alia failure criterion was derived in terms of tensile strength and UCS, as in equation 4.17 (Rahimi and Nygaard, 2015).

$$(\sigma_1 - \sigma_3)^2 + (\sigma_1 - \sigma_2)^2 + (\sigma_2 - \sigma_3)^2 = 2(C_0 - T_0)(\sigma_1 + \sigma_2 + \sigma_3) + 2C_0T_0 \quad (4.17)$$

In equation 4.17, σ_1 , σ_2 , and σ_3 are maximum, intermediate, and minimum principal stresses (psi), respectively, C_0 is UCS (psi), and T_0 is tensile strength (psi). The tensile requirement can be assumed as in equation 4.18.

$$\frac{C_0}{T_0} = \frac{4}{\sqrt{c^2 + 1} - c} \quad (4.18)$$

Here c is the coefficient of internal friction angle, which can be calculated via equation 4.19.

$$c = \tan \theta \quad (4.19)$$

In equation 4.19, θ is internal friction angle (degree).

4.2.5 Torque

Torque can be expressed as the difference between the power applied from the rotary table and available at the bit (Çağlayan, 2014). In other words, torque is the rotational impulse between the formation and the bit regarding the friction and side forces acting on the drill string. Friction occurs due to the drill string motion. Side forces, including normal forces, exist among the wellbore and any component of the drill string. The forces acting on a drill string component is shown in Figure 4.11.

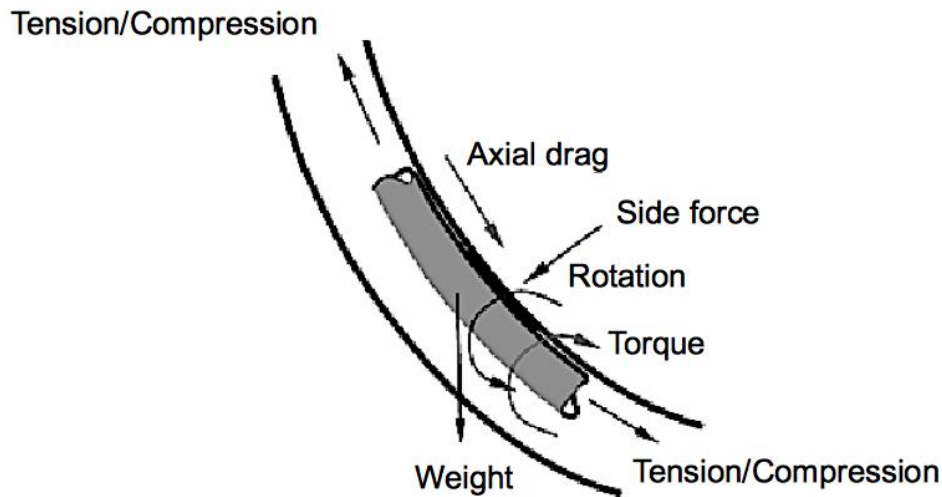


Figure 4.11 : The forces acting on a drill string component, adapted from (Wu et al., 2011).

Poor hole cleaning and wellbore geometry also cause torque during drilling. Hole tortuosity, which is the ratio of the actual path and the length of the travelled distance, establishes the friction level that affect the surface torque and drilling efficiency. Additionally, cuttings removal issues due to high tortuosity hinders the ability of the drill string to rotate, which results in decrease in ROP and increase in surface torque (Maehs et al., 2010).

As indicated in the literature review, the studies using BYM is performed in homogeneous formations. It is known that BYM provides respectable ROP prediction in homogeneous formations (Kor and Altun, 2020). On the other hand, heterogeneous drilling environments make prediction accuracy of BYM decrease substantially and simply make the BYM impractical because of the formation heterogeneity. Besides, the studies performed in heterogeneous formations do not use the functional relationships of the BYM and do not compare the results with BYM as a benchmark. This study is primarily focused on BYM, which uses multiple linear regression as the predictive model. One of this diessertations's goals is to show the incapability of BYM in predicting the ROP in heterogeneous formations. Another goal is to propose an alternative machine learning method by using the functional relationships proposed by (Bourgoyne Jr. and Young Jr., 1974) with implementing several geomechanical parameters additional to the functions defined in BYM.



5. MACHINE LEARNING AND PREDICTIVE MODELS

It is not viable to represent the problem with only one independent variable in many situations in engineering. Several variables can have dependencies, of which those may affect each other depending on the problem's nature. Hence, it is not feasible to execute a simple regression analysis.

Computers can learn relationships, create models from those relationships, and make decisions based on the model. The perception of equipping computers with this capability is called artificial intelligence (AI) (Shi, 2011). The most common approach to make a computer learn from data is machine learning (ML). In ML, computers extract the information from the attributes via data mining, make decisions on the future circumstances, and engender solutions for congruous problems.

In ML, there exists a task to appoint the substances to different classes for binary classification. The idea is to generate a relationship between the features and the classes by building a model with a training data set (Flach, 2012, p. 11). The general concept of the ML is shown in Figure 5.1.

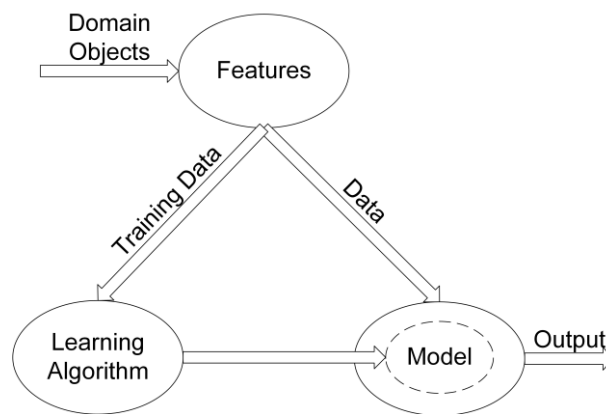


Figure 5.1 : General concept of ML.

As seen in Figure 5.1, an appropriate model built from data is necessary to generate a task. The model building process from the training data constitutes the learning problem. Tasks are undertaken by the models, while learning problems are solved by learning algorithms used to build the models (Flach, 2012, p. 12).

There are two types of ML models: Supervised and unsupervised learning models. Supervised learning involves building a model within a pre-labeled training data set (Zhang, 2001, p. 103). On the contrary, unsupervised learning generates the model by using an unlabeled data set (Ghahramani, 2004, p. 3). A simple illustration is shown in Figure 5.2.

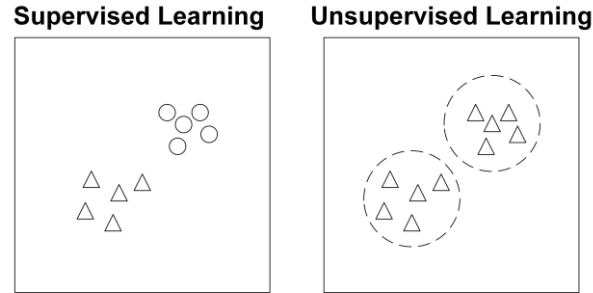


Figure 5.2 : A sample illustration of ML types.

ML models can be labeled under three subgroups. First is the geometric model, where attributes are expressed with geometrical terms such as distances, coordinate systems, lines, etc. (Mulmuley, 1994, p. 26)). An example of a geometrical expression is shown in Figure 5.3.

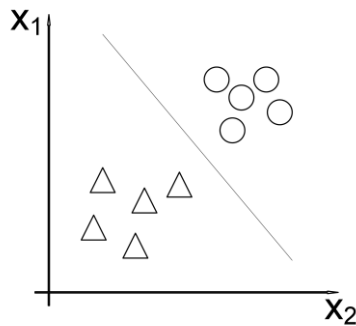


Figure 5.3 : An example of geometrical distribution in 2-D space.

The second is the probabilistic model and has the same concept as the Bayes classifier (Flach, 2012, p. 25). This model's basis estimates the posterior probabilities and the likelihoods with Bayes' rule via equation 5.1.

$$P(Y | X) = \frac{P(X | Y)P(Y)}{P(X)} \quad (5.1)$$

In equation 5.1, $P(X)$ is the probability and $P(Y)$ is the prior probability.

The third is the logical model, which makes the model been transformed into a tree structure that can be used to make decisions in an organized way (Flach, 2012, p. 32). A sample of a decision tree is given in Figure 5.4.

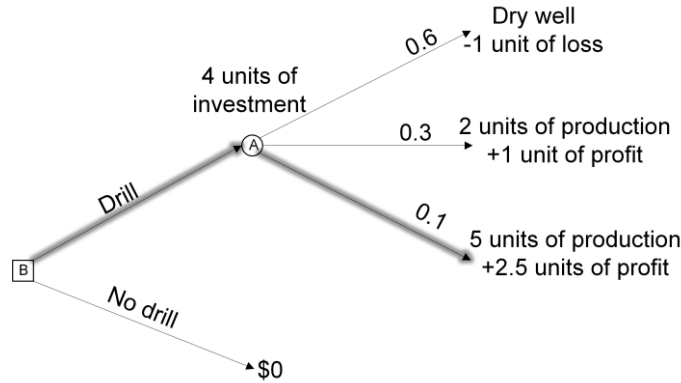


Figure 5.4 : An example of a decision tree.

5.1 Predictive Models

This section focuses on the theory and the mathematical background of the predictive models used in this study.

5.1.1 Multiple linear regression

The primary focus of the multiple linear regression (MLR) analysis is to assess the relationship between the dependent and independent variable(s) while estimating the effect of the independent variables on the dependent variable concerning change in response (Yan and Su, 2009, p. 41). Additionally, multiple regression analysis is much more problematic than a simple regression due to the complex interpretation of independent variables' interrelationships (Freund et al., 2006, p. 73). An MLR model can be expressed as in equation 5.2.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_n x_n + \Theta \quad (5.2)$$

In equation 5.2, β_i ($i=0,1,\dots,n$) are the regression coefficients, y is the dependent variable, x terms are independent variables, and Θ is the error term. Here, the model expresses a hyperplane in an n -dimensional space. β_0 is called intercept. β_j ($i=1,\dots,n$) represents the estimated change in the dependent variable when other independent variables stay constant (Montgomery and Runger, 2003, p. 411).

The vector-matrix notation of equation 5.2 is given in equation 5.3.

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\Theta} \quad (5.3)$$

In equation 5.3, the terms can be expressed as equation 5.4.

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{bmatrix}, \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}, \text{ and } \boldsymbol{\Theta} = \begin{bmatrix} \Theta_1 \\ \Theta_2 \\ \vdots \\ \Theta_n \end{bmatrix} \quad (5.4)$$

In the least-squares (LS) principle, the aim is to find the LS estimators that minimize equation 5.5.

$$L = \sum_{i=1}^n \Theta_i^2 = \boldsymbol{\Theta}'\boldsymbol{\Theta} = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \quad (5.5)$$

The solution for the $\boldsymbol{\beta}$ can be expressed as in equation 5.6.

$$\frac{\partial L}{\partial \boldsymbol{\Theta}} = \mathbf{0} \quad (5.6)$$

The system can be solved via equation 5.7.

$$\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}'\mathbf{y} \quad (5.7)$$

In equation 5.7, $\hat{\boldsymbol{\beta}}$ is the vector of LS estimators. Equation 5.7 can be solved via multiplication of both sides with the inverse of $\mathbf{X}'\mathbf{X}$. Hence, $\hat{\boldsymbol{\beta}}$ can be found with equation 5.8.

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{y} \quad (5.8)$$

The matrix form of the model is given in equation 5.9.

$$\begin{bmatrix} n & \sum_{i=1}^n x_{i1} & \sum_{i=1}^n x_{i2} & \cdots & \sum_{i=1}^n x_{ik} \\ \sum_{i=1}^n x_{i1} & \sum_{i=1}^n x_{i1}^2 & \sum_{i=1}^n x_{i1}x_{i2} & \cdots & \sum_{i=1}^n x_{i1}x_{ik} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \sum_{i=1}^n x_{ik} & \sum_{i=1}^n x_{i1}x_{ik} & \sum_{i=1}^n x_{ik}x_{i2} & \cdots & \sum_{i=1}^n x_{ik}^2 \end{bmatrix} \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_k \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^n y_i \\ \sum_{i=1}^n x_{i1}y_i \\ \vdots \\ \sum_{i=1}^n x_{ik}y_i \end{bmatrix} \quad (5.9)$$

The regression fit for equation 5.9 is given in equation 5.10.

$$\hat{y}_i = \hat{\beta}_0 + \sum_{j=1}^k \hat{\beta}_j x_{ij} \text{ where } i = 1, 2, \dots, n \quad (5.10)$$

The vector-matrix form of equation 5.10 is expressed as in equation 5.11.

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}} \quad (5.11)$$

The difference between the actual and the estimated value is called residual, defined in equation 5.12.

$$\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}} \quad (5.12)$$

5.1.2 Support vector machines

Support vector machines (SVM) is a learning method based on obtaining a hyperplane that separates multi-dimensional data into its classes (Boswell, 2002, p. 1). The aim is to generate a classifier appropriate for new data to be generalized ultimately, as shown in Figure 5.5.

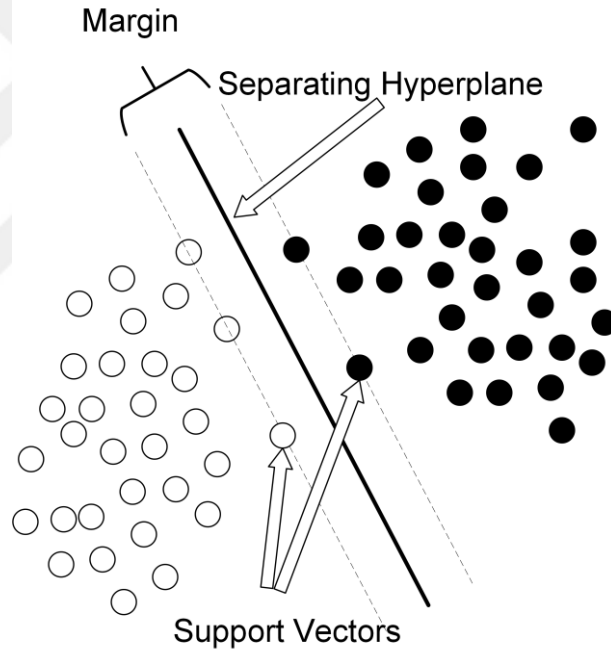


Figure 5.5 : Linear classification, adapted from (Meyer, 2014, p. 2).

In Figure 5.5, the nearby points of two classes are support vectors, the Euclidean distance between these support vectors is margin, and the classifier maximizing this margin is separating hyperplane (Gunn, 1998, p. 5).

5.1.2.1 Support vector machines for classification

The classification problem is to separate the dataset into its classes with a hyperplane, as expressed in equation 5.13.

$$f(x) = \langle \mathbf{w}, \mathbf{x} \rangle + b \quad (5.13)$$

$$D = \{(x^1, y^1), \dots, (x^l, y^l)\} \text{ and } x \in \mathbb{R}^n, y \in \{-1, 1\}$$

In (Eq 5.13), $\mathbf{w}$ is the weight vector, b is the bias and l is the number of training data. If b is assumed as zero, the data points corresponding to $\mathbf{w}^T \mathbf{x} = 0$ exist perpendicular to $\mathbf{w}$ while intersecting the origin. This condition can be described with a line in 2-D and a plane in 3-D, which are called a hyperplane. The bias term's geometrical implication is the interpretation of the distance between the hyperplane and the origin (Ben-Hur and Weston, 2009, p. 3). The hyperplane equation can be given as in equation 5.14.

$$\{x : f(\mathbf{x}) = \mathbf{w}^T \mathbf{x} + b = 0\} \quad (5.14)$$

Equation 5.14 divides the sample space into two parts. $f(\mathbf{x})$ is the linear discriminant function. The sign of $f(\mathbf{x})$ represents the part of the divided space. It is possible to state that the data points are optimally separated, and the margin is maximum while assuming that the error is zero. It is suitable to implement an optimal canonical separating hyperplane (OCSH) to eliminate any generality loss (Vapnik, 2000). Hence, the constraints can be described as in equation 5.15.

$$\min_i |\langle \mathbf{w}, \mathbf{x}_i \rangle + b| = 1 \quad (5.15)$$

The constraint in equation 5.15 makes the formulation simpler, which states that the norm or the measure of the weight vector should be equal to the inverse of the distance between the hyperplane and the closest data point(s) (Gunn, 1998, p. 6). A sample illustration of the OCSH is shown in Figure 5.6.

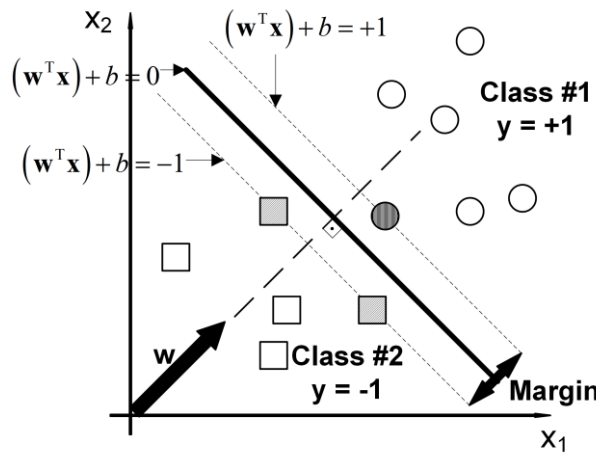


Figure 5.6 : The OCSH, adapted from (Kecman, 2001, p. 154).

The hyperplane shown in Figure 5.6 separates the data into two classes. Additionally, the nearest data points to the hyperplane, support vectors, are shown with the gray color in Figure 5.6.

In terms of the geometrical approach, an OCSH must compensate the constraints in equation 5.16.

$$y^i [\langle \mathbf{w}, \mathbf{x}_i \rangle + b] \geq 1 \quad (i = 1, \dots, l) \quad (5.16)$$

The distance from the hyperplane can be expressed as in equation 5.17.

$$d(w, b; x) = \frac{|\langle \mathbf{w}, \mathbf{x}_i \rangle + b|}{\|\mathbf{w}\|} \quad (5.17)$$

As stated previously, an OCSH maximizes the margin regarding the constraints in equation 5.16. The margin can be calculated via equations 5.18, 5.19, 5.20, and 5.21.

$$M = \min_{x_i: y_i = -1} d(w, b; x_i) + \min_{x_i: y_i = 1} d(w, b; x_i) \quad (5.18)$$

$$M = \min_{x_i: y_i = -1} \frac{|\langle \mathbf{w}, \mathbf{x}_i \rangle + b|}{\|\mathbf{w}\|} + \min_{x_i: y_i = 1} \frac{|\langle \mathbf{w}, \mathbf{x}_i \rangle + b|}{\|\mathbf{w}\|} \quad (5.19)$$

$$M = \frac{1}{\|\mathbf{w}\|} \left(\min_{x_i: y_i = -1} |\langle \mathbf{w}, \mathbf{x}_i \rangle + b| + \min_{x_i: y_i = 1} |\langle \mathbf{w}, \mathbf{x}_i \rangle + b| \right) \quad (5.20)$$

$$M = \frac{2}{\|\mathbf{w}\|} \quad (5.21)$$

Hence, the OCSH minimizes equation 5.22.

$$\Phi(w) = \frac{1}{2} \|\mathbf{w}\|^2 \quad (5.22)$$

This ideal situation is valid for the linearly-separable data. Otherwise, to determine the optimum hyperplane, noise in the data should be estimated (Gunn, 1998, p. 10). To make the hyperplane separate the data correctly, a penalty function with non-negative variables defined in equation 5.23 should be implemented in the training phase (Cortes and Vapnik, 1995).

$$F(\xi) = \sum_i \xi_i \quad (5.23)$$

In (Eq 5.23), ζ_i are called slack variables that represent errors due to misclassification. By implementing the slack variables, the optimization problem is transformed into minimizing both the bound of the Vapnik-Chervonenkis (VC) dimension and the misclassification error at the same time. The constraints in equation 5.16 should be updated as in equation 5.24 to perform this optimization.

$$\begin{aligned} y_i [\langle \mathbf{w}, \mathbf{x}_i \rangle + b] &\geq 1 - \zeta_i \\ (i = 1, \dots, l), \zeta_i &\geq 0 \end{aligned} \quad (5.24)$$

Thus, the generalized OCSH can be estimated via the weight vector that minimizes equation 5.25.

$$\Phi(\mathbf{w}, \zeta) = \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_i \zeta_i \quad (5.25)$$

In equation 5.25, the complexity, or the penalty parameter (C) is a user-defined scalar value used to modify the margin to minimize the slack variables. It should be noted that a high value of C may be resulted in a high penalty. Similarly, a low value of C may cause misclassification.

5.1.2.2 Kernel functions

In SVM, the OCSH is controlled to finesse generalization capability. If data is not suitable to be separated linearly, the classifier may not have the generalization ability. In this situation, the input space should be mapped into a high-dimensional feature space (Abe, 2010, p. 31). The mapping process can be performed by implementing kernel functions, which transform the data from the input space into the feature space by applying linear or non-linear mapping, as illustrated in Figure 5.7.

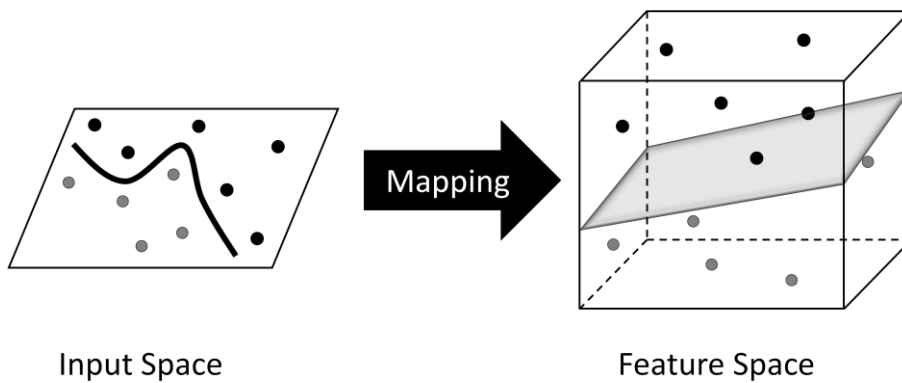


Figure 5.7 : Input and feature space mapping.

Kernel functions are based on Mercer's condition (Mercer, 1909) and Reproducing Kernel Hilbert Spaces (Aronszajn, 1950; Girosi, 1998; Heckman, 2012). Kernel functions can be expressed via equation 5.26 in terms of a dot product.

$$K(\mathbf{x}, \mathbf{x}') = \langle \phi(\mathbf{x}), \phi(\mathbf{x}') \rangle$$

$$\phi(\mathbf{x}) = (\phi_1(\mathbf{x}), \dots, \phi_l(\mathbf{x}))^T \quad (5.26)$$

In equation 5.26, K is the kernel function that maps the data linearly or non-linearly from input space to feature space.

Various kernel functions were introduced, such as linear, polynomial, radial basis function (RBF), splines, tensor products, etc. (Gunn, 1998, pp. 19-23). In this study, linear and RBF kernels are used for calculations given in equations 5.27 and 5.28, respectively.

$$K(\mathbf{x}, \mathbf{x}') = \mathbf{x}^T \mathbf{x}' \quad (5.27)$$

$$K(\mathbf{x}, \mathbf{x}') = \exp(-\gamma \|\mathbf{x} - \mathbf{x}'\|^2) \quad (5.28)$$

In equation 5.28, γ is a user-defined parameter to control the radius.

5.1.2.3 Support vector regression

SVM can also predict numbers instead of class labels. Support vector regression (SVR) is a type of SVM, which produces a number as the output (Flach, 2012). The basis of SVR is much similar to SVM with a few differences. Essentially, SVM can be performed as a regression by implementing a loss function (LF) (Smola, 1996). Several LFs are illustrated in Figure 5.8.

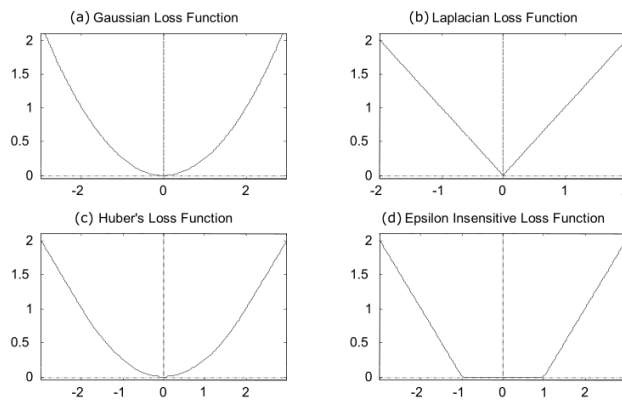


Figure 5.8 : Different loss functions, adapted from (Chu et al., 2001, p. 2).

In Figure 5.8, (a) is the quadratic LF based on the least-squares criterion. (b) is the Laplace type of LF that has minor deviation sensitivity when compared to quadratic LF. (c) is the Huber LF (Huber, 1964) considering the optimal conditions in an unknown distribution type. Nevertheless, (a), (b), and (c) do not represent the sparseness in the SVs (Gunn, 1998., p. 30). On the other hand, the LF (d) having a similar concept with (c) is suitable to be used to determine the support vectors (Vapnik, 2000).

ϵ -insensitive support vector regression

Like the SVM for classification, the optimal regression can be performed via minimizing equation 5.29.

$$\Phi(w, \zeta) = \frac{1}{2} \|w\|^2 + C \sum_i (\zeta_i^-, \zeta_i^+) \quad (5.29)$$

In equation 5.29, ζ^- and ζ^+ are lower and upper bounds of the slack variables. The ϵ -insensitive LF can be expressed as in equation 5.30.

$$L_\epsilon(y) = \begin{cases} 0, & |f(x) - y| < \epsilon \\ |f(x) - y| - \epsilon, & \text{otherwise} \end{cases} \quad (5.30)$$

The implementation of the ϵ -insensitive loss function and the slack variables that remain outside of the user-defined ϵ band are illustrated in Figure 5.9.

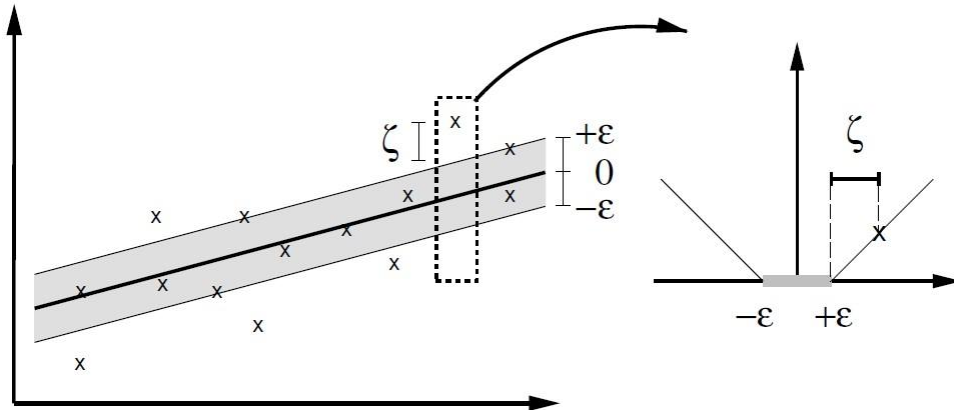


Figure 5.9 : ϵ -insensitive loss function and slack variables (Smola and Schölkopf, 2003).

ν support vector regression

In ϵ -SVR, each data point involves an error (ϵ) while the values outside of the ϵ envelope are slack variables (ζ) that are regulated by the penalty parameter (C).

Nevertheless, it is not simple to estimate the optimum ε (Abe, 2010, p. 431). Correspondingly, it is recommended to use a user-defined controller parameter (ν) to modify the training model by adjusting the amount of ε (Schölkopf et al., 1998). In ν -SVR, the optimization problem is expressed in equation 5.31 with the constraints in equation 5.32.

$$\Phi(\mathbf{w}, \zeta, \varepsilon) = \frac{1}{2} \|\mathbf{w}\|^2 + C \left(\nu \varepsilon + \frac{1}{l} \sum_{i=1}^l (\zeta_i + \zeta_i^*) \right) \quad (5.31)$$

$$\begin{aligned} (\langle \mathbf{w}, \mathbf{x}_i \rangle + b) - y_i &\leq \varepsilon + \zeta_i \\ y_i - (\langle \mathbf{w}, \mathbf{x}_i \rangle + b) &\leq \varepsilon + \zeta_i^* \\ \varepsilon &\geq 0, \zeta_i \geq 0 \end{aligned} \quad (5.32)$$

5.1.3 Artificial Neural Networks

Artificial neural networks (ANN) is a model based on neurons' nature in the human brain process. The fundamental unit in a neural network is a neuron. A neuron acquires input from the other neurons or an external source and produces an output. Each input is associated with weights (w), which is assigned to an input related by importance. A neuron applying a function to the weighted sum of inputs and generating output (perceptron) is shown in Figure 5.10.

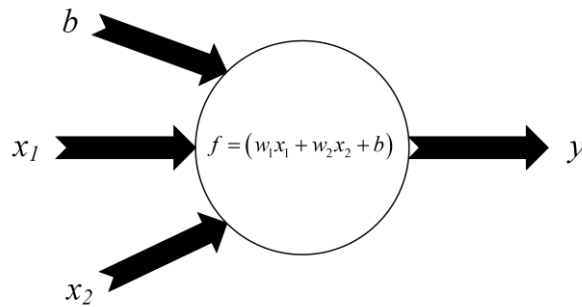


Figure 5.10 : A sample of a perceptron.

In Figure 5.10, numerical inputs x_1 and x_2 have associated weights w_1 and w_2 . The external input b is called bias. This variable affects the transmission of the information obtained due to the sum function to the next neuron. The prediction is computed via equation 5.33.

$$\hat{y} = \text{sgn} \left(\sum_{j=1}^l w_j x_j + b \right) \quad (5.33)$$

Like the SVM, perceptron has a linear model, where the separating hyperplane is defined as a line (Aggarwal, 2018, p.8). As a general concept, the goal of optimization is to minimize the prediction error. The perceptron algorithm was derived to minimize the misclassification errors. Hence, the perceptron algorithm can be expressed in terms of least-squares as in equation 5.34.

$$\min L_w = \sum_{(\mathbf{x}, y) \in D} (y - \hat{y})^2 \quad (5.34)$$

In equation 5.34, L_w is the loss function. Equation 5.34 is seen as in least-square form, which is valid for continuous target variables. However, the sign function is not differentiable at specific points. Moreover, the sign function is assumed zero over the domain's large scales, making the gradient zero at specific differentiable points (Aggarwal, 2018, p. 7). Therefore, a smoothing approximation in equation 5.35 is used to overcome this problem.

$$\nabla L_{smooth} = \sum_{(\mathbf{x}, y) \in D} (y - \hat{y}) \mathbf{x} \quad (5.35)$$

As mentioned before, a neural network feeds each input data into the network one at a time to perform a prediction. At the same time, the weights are updated gradually based on the prediction error. The weight updating is shown in equation 5.36.

$$\mathbf{w} \leftarrow \mathbf{w} + a(y - \hat{y}) \mathbf{x} \quad (5.36)$$

In equation 5.36, a represents the learning rate. The perceptron algorithm is run regularly until the convergence criteria are met. Hence, the gradient-descent updating in terms of error can be expressed as in equation 5.37.

$$\mathbf{w} \leftarrow \mathbf{w} + aE(\mathbf{x}) \mathbf{x} \quad (5.37)$$

One important assumption is that the neural network runs through the data points randomly in the training phase and updates the weights based on the corresponding data point's prediction error.

In ANN, the non-linear activation function f is used to implement non-linearity into the output by applying a fixed mathematical operation. The most frequently used activation functions are the sigmoid function and hyperbolic tangent function, shown in Figure 5.11.

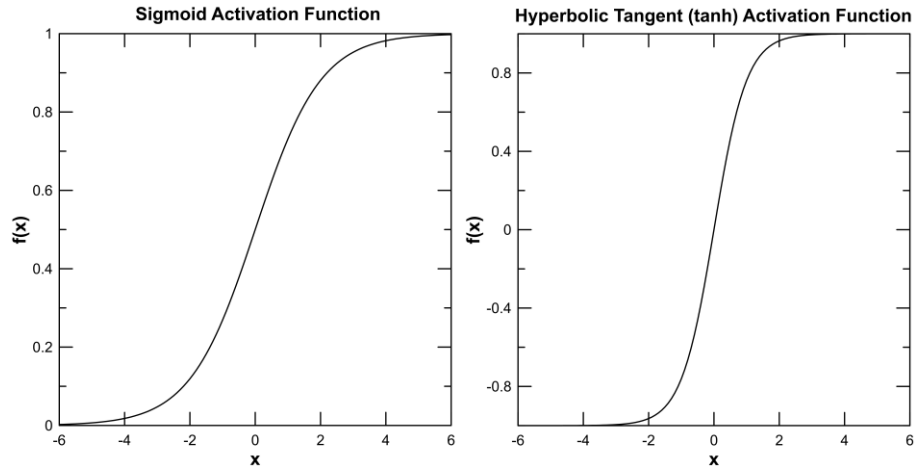


Figure 5.11 : Activation functions.

The activation functions can be defined as in equation 5.38.

$$\begin{aligned} f(x) &= \frac{1}{1+e^{-x}} \quad (\text{sigmoid function}) \\ f(x) &= \frac{e^{2x}-1}{e^{2x}+1} \quad (\text{tanh function}) \end{aligned} \quad (5.38)$$

A feedforward neural network is the primary type of ANN with multiple neurons formed in layers, connecting neighboring neurons. All the connections have associated weights. In a feedforward network, the output flows in the only forward direction through hidden layers. A Multi-Layer Perceptron (MLP) consists of one or more hidden layers. MLP can learn non-linear functions and make regression (and classification) using a given set of features and targets. Figure 5.12 shows an MLP with a single hidden layer.

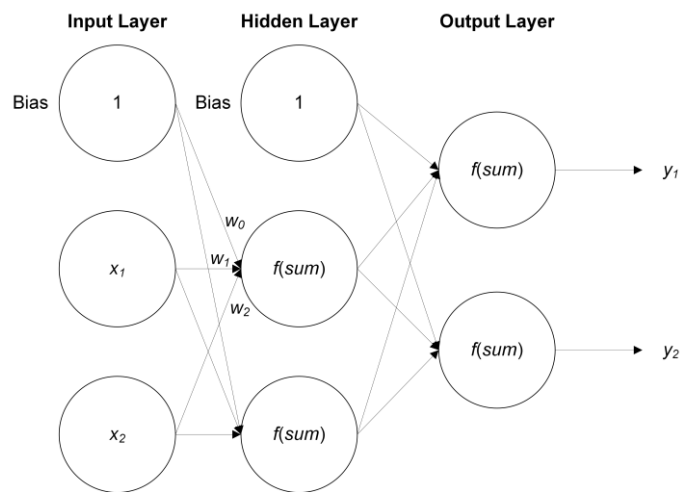


Figure 5.12 : An MLP with one hidden layer.

5.1.4 Random Forests

Bootstrap aggregation (or bagging) is a technique used to reduce the variance of estimation. The same regression tree is fitted through the bootstrapped training samples repeatedly, and the results are averaged (Hastie et al., 2009, p. 587).

Boosting is an ensemble algorithm to reduce both bias and variance (Breiman, 1996). Most boosting algorithms work by defining weak classifiers iteratively, adding to a final classifier, and assigning a weight based on the accuracy of the weak classifiers. Consequently, the data weights are re-weighted. While misclassified input points' weights are improved, correctly classified data points' weights are degraded. Hence, unseen weak learners engage in the misclassified data points.

Random forests (RF) algorithm, similar to boosting algorithms, is a bagging implementation that generates large-sized uncorrelated trees and averages the results (Breiman, 2001). For the identically distributed variables, the average variance can be expressed as in equation 5.39.

$$\sigma_{avg}^2 = \rho\sigma^2 + \frac{1-\rho}{l}\sigma^2 \quad (5.39)$$

In equation 5.39, ρ is the pairwise correlation, σ^2 is the sample variance, and l is the number of training data. Naturally, as the number of training data increases, the plus term in equation 5.39 converges to zero. The general idea in RF is to reduce the correlation to decrease the bagging variance (Hastie et al., 2009, p. 588). Hence, the input variables are chosen randomly to generate the trees. The RF regression predictor is expressed as in equation 5.40.

$$f_{RF}^l(x) = \frac{1}{l} \sum_{i=1}^l T(x, \xi_i) \quad (5.40)$$

In equation 5.40, T represents the tree and ξ_i characterizes the i^{th} RF.

Variable importance for RF can be estimated using out-of-bag (OOB) samples, which are the observations that generate its RF predictor by averaging designated bootstrapped samples in specific trees. In i^{th} tree, the OOB samples are delivered to the tree with designated prediction accuracy. Next, the values of j^{th} variable are randomly re-assigned within the OOB samples, which generates a new prediction accuracy. Hence, the importance of the variable j is computed by averaging the decline

in the prediction accuracy due to the permutation in all the trees, which can be defined as a percentage.

5.2 Overfitting and Cross-Validation

As stated, ML's general idea is to train a model with the training data set and make predictions on an unseen data set via a trained model with a maximum of prediction accuracy. In other words, maximizing training accuracy is not mandatory. Nevertheless, when the training accuracy is not maximum, the model may fit the noise to the learning curve by memorizing the data instead of building a bias (Dietterich, 1995, p. 326). This situation is called overfitting, which is shown in Figure 5.13.

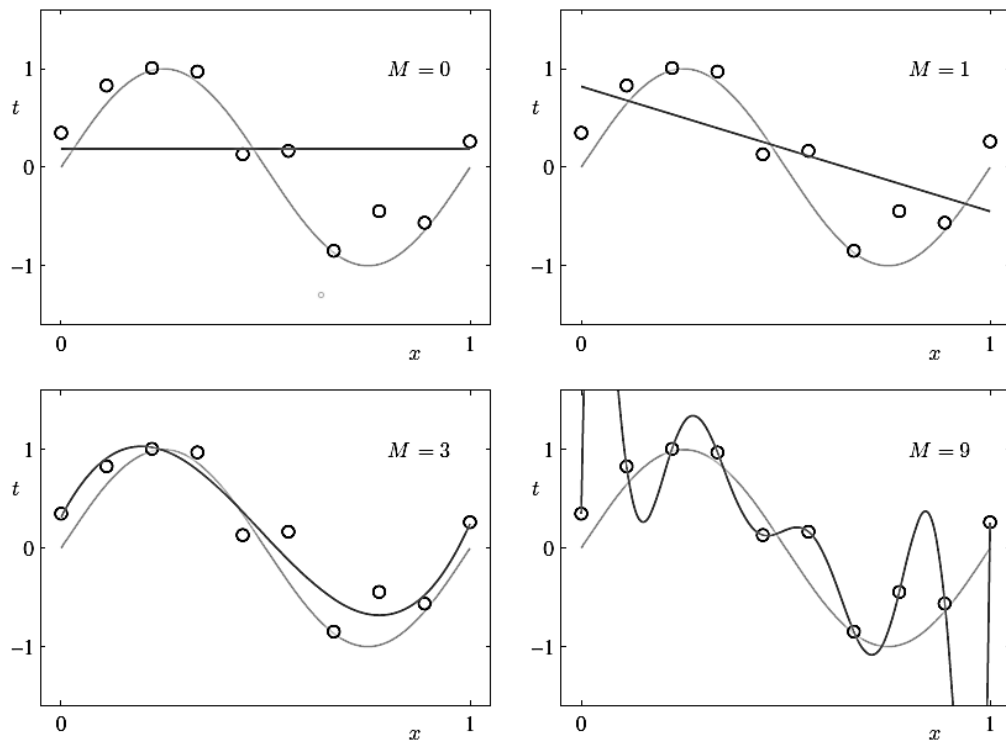


Figure 5.13 : Overfitting due to degree of the polynomial (Bishop, 2006, p. 7).

In Figure 5.13, the learning curve's fitting capability excels when the degree of the polynomial (M) is increased, even corresponds to data points with zero error at excessive degrees. This situation creates overfitting by making the learning curve represent the designated function in an inadequate way.

Overfitting may also occur due to the number of input data, which is shown in Figure 5.14.

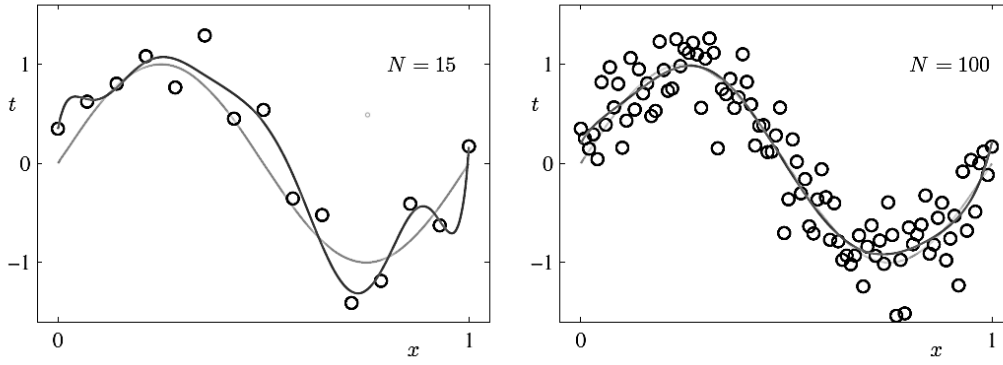


Figure 5.14 : Overfitting due to the number of input data (Bishop, 2006, p. 9).

In Figure 5.14, the learning curve's fitting performance is enhanced when the number of input data points (N) increases, which means the model complexity is improved. Hence, it is recommended to use a multiple (i.e., 5 or 10) of the feature number as the minimum of the input data amount (Bishop, 2006, p. 9).

To eliminate the overfitting problem, a statistical analysis method called cross-validation (CV) is used. CV is a technique to estimate the trained model's generalization ability on a validation data set independent from training and testing data sets (Flach, 2012, p. 349). One of the most common approaches in CV is the k-fold CV. In k-fold CV, the data is first divided randomly in k folds. The model is trained via $k-1$ folds while being tested in the k^{th} fold until each fold is tested. The most common application is to set k to 10, in other words, 10-fold CV.

5.3 Outlier Labeling and Standardization

In statistics, a data point that deviates substantially from the other samples in a distribution is called an outlier (Grubbs, 1969, p. 1). An outlier can arise from experimental errors, miscalculations, or measurement uncertainties, leading to crucial statistical analysis issues. An example of outliers is shown in Figure 5.15.

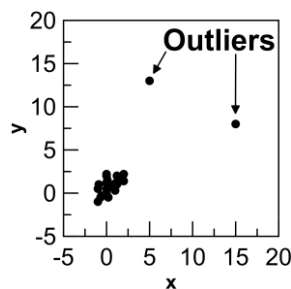


Figure 5.15 : Outlier example.

Statistical interpretation of the raw data sets, including outliers, may be misleading. Hence, outliers in the data must be determined and treated accordingly. Two of the most commonly used outliers labeling techniques are used in this study: Tukey's method and z-score analysis.

5.3.1 Tukey's method

It is possible to have some extreme values which can behave as outliers in a noisy dataset. Hence, outlier labeling must be performed on the data before processing. One of the best approaches is Tukey's method that takes the interquartile range (IQR) under consideration for labeling outliers, which applies equation 5.41.

$$IQR = Q_3 - Q_1 \quad (5.41)$$

In equation 5.41, Q_3 is the 3rd quartile, and Q_1 is the 1st quartile of the data. The upper and lower limits of the extreme boundaries are determined by multiplying IQR value with a factor, then subtracting from Q_1 and adding to Q_3 . The most commonly used value for this factor is 1.5. A sample process is given in Figure 5.16, which is also an example of the Box-Whisker plot.

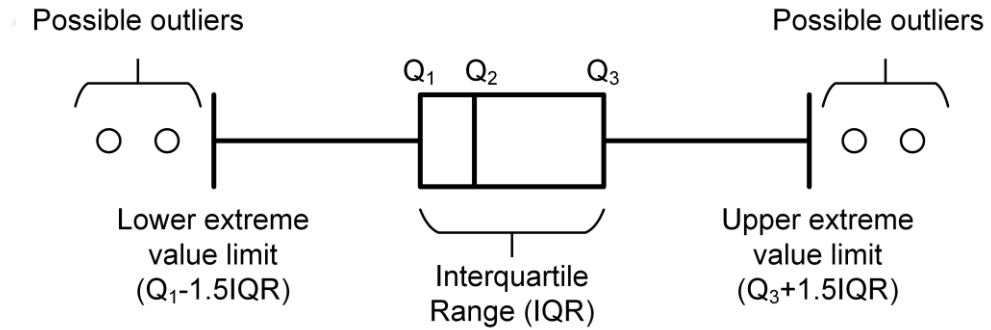


Figure 5.16 : Box-Whisker plot and outlier detection, adapted from (Bakar and Rosbi, 2017).

5.3.2 z-score analysis

The z-score analysis is a standardization method, which can be used for outlier labeling in one dimension. In this method, the data is assumed to be in Gaussian distribution. The data points placed in the extremes of the distribution are labeled as outliers. z-score for a data point is calculated with equation 5.42.

$$z_i = \frac{x_i - \mu}{\sigma} \quad (5.42)$$

In equation 5.42, z_i is the z-score of the data point x_i , μ and σ are the mean and the standard deviation, respectively. A data point is labeled as an outlier when the absolute value of its z-score is greater than a threshold. Commonly used threshold values are 2.5, 3.0, and 3.5. In this study, the threshold value is selected as 3.0. A sample illustration of standard normal distribution and the relationship between the z-scores and data centiles are shown in Figure 5.17.

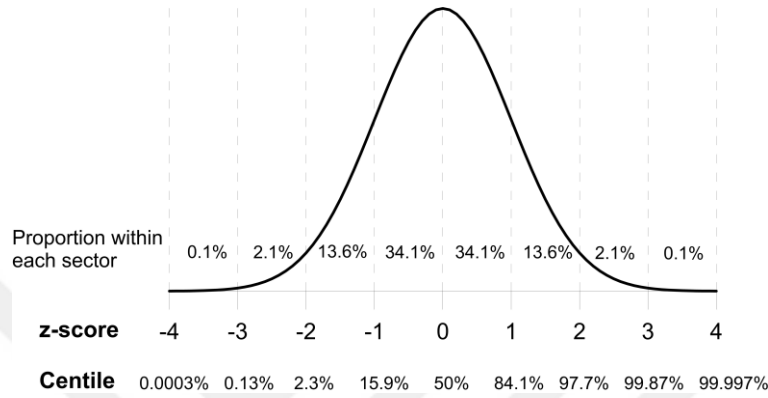


Figure 5.17 : Relationship between the z-scores and data centiles.

5.3.3 Data normalization

In many cases, samples may vary on a vast scale. The scaling problem can diverge any regression's accuracy by misestimating each variable's importance by the nature of regression analysis. Hence, minimum-maximum (min-max) normalization can be applied to the dataset before performing predictive analytics to overcome the scaling issue and reduce the scattering and total variance. In this way, the effect of scaling and variance is eliminated by normalizing each variable (or feature) within a pre-defined range (0 to 1) through equation 5.43.

$$x_{\text{norm}} = \frac{x - x_{\min}}{x_{\max} - x_{\min}} \quad (5.43)$$

In equation 5.43, x_{norm} is normalized value, $x_{\min}$ and $x_{\max}$ are the minimum and maximum data of the entire column.

5.4 Statistical Performance Measures

Various statistical performance measures are used to compare predictions from a predictive model and actual values. In this study, three statistical performance criteria

evaluate testing performance: root mean squared error (RMSE), correlation coefficient, and statistical significance test.

5.4.1 Root mean squared error

RMSE is a verification criterion that standardizes the results in terms of residuals. RMSE is calculated via equation 5.44.

$$\text{RMSE} = \sqrt{\frac{1}{l} \sum_i e_i^2} \quad (5.44)$$

In equation 5.44, l is the number of data and e is the residual.

5.4.2 Correlation coefficient

The correlation coefficient is the meaning of the statistical relationship between two variables. This relationship is expressed numerically between the range of -1 and 1 by equation 5.45.

$$r_{x,y} = \frac{\text{cov}(x,y)}{\sigma_x \sigma_y} \quad (5.45)$$

In equation 5.45, $r_{x,y}$ is the correlation coefficient, $\text{cov}(x,y)$ is the covariance, σ_x and σ_y are the standard deviation of the variables x and y , respectively. -1 means the negative correlation, 1 means the positive correlation, while 0 indicates no correlation.

5.4.3 Statistical significance test

The statistical significance level is expressed as a p-value, which is the probability of achieving the test results as same as the actual observed value assuming the null hypothesis is correct (LaerdStatistics, n.d.). One approach to obtaining the p-value is performing a t-test to determine whether the two test sets differ. The t-statistics can be calculated via equation 5.46.

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{l}} \quad (5.46)$$

In equation 5.46, t is the t-score, $\bar{x}$ is the mean of sample differences, μ_0 is the hypothesized mean, s is the standard deviation sample differences, and l is the sample size. After calculating the t-score, the corresponding p-value is found by using Student's t-table or software. A lower p-value means higher statistical significance.

5.5 Software and Toolboxes

It is impossible to solve the optimization problems mentioned in this study in an analytical way since the data sets include a massive input data amount. Hence, various software and toolboxes are used for calculations in ML applications. The software and toolboxes used in this study are either open-source (licensed under GNU General Public License) or licensed by Istanbul Technical University.

5.5.1 R statistical programming language

R is one of the most used statistical programming languages for statistics and ML applications. One of the best features of the open-source R language is the user-defined libraries (or packages) that make the programming language expandable. Most of the work performed in this dissertation is done via R programming language by using various valuable packages such as “PerformanceAnalytics” for data visualization, “e1071” for SVM, “neuralnet” for ANN, and “caret” for RF.

5.5.2 IBM SPSS statistics

SPSS initially stands for “Statistical Package for the Social Sciences” and later changed into “Statistical Product and Service Solutions”, an IBM software widely used for statistical analysis. Additionally, there are additional features such as data manipulation and documentation. The statistical toolboxes are embedded in the software accessible through the graphical user interface or the console via syntax. In this study, SPSS is used for statistical significance testing.

6. APPLICATION, RESULTS, AND DISCUSSIONS

The main goal of the applications in this study is to perform rate of penetration (ROP) predictions in heterogeneous formations. It was mentioned that SVR could make predictions with good approximation when the number of training data is limited (Kor and Altun, 2020). With a field data taken recently, it is possible to show the prediction performance of multiple linear regression (MLR), support vector regression (SVR), and artificial neural networks (ANN) on the dataset taken from heterogeneous lithology.

The applicability of MLR, SVR, and ANN is investigated by generating different case scenarios and using different datasets. The cases are constructed with the following structure:

1. Case 1: Data taken from the literature (homogeneous cases)
 - Case 1.1: Calculations with data from Bourgoyne Jr. and Young Jr. (1974)
 - Case 1.2: Calculations with data from Irawan et al., (2012)
 - Case 1.3: Calculations with data from Al-Betairi et al., (1988)
2. Case 2: Data taken from the Well C-1
 - Case 2.1: Calculations with data from formation F-4 (homogeneous case)
 - Case 2.2: Calculations with data from formation F-5 (homogeneous case)
 - Case 2.3: Calculations with data from formation F-6 (heterogeneous case)
3. Case 3: Data taken from the Well C-2
 - Case 3.1: Calculations with data from formation F-4 (heterogeneous case)
 - Case 3.2: Calculations with data from formation F-5 (homogeneous case)
 - Case 3.3: Calculations with data from formation F-6 (heterogeneous case)

6.1 Application

The general flow chart of application process is shown in Figure 6.1.

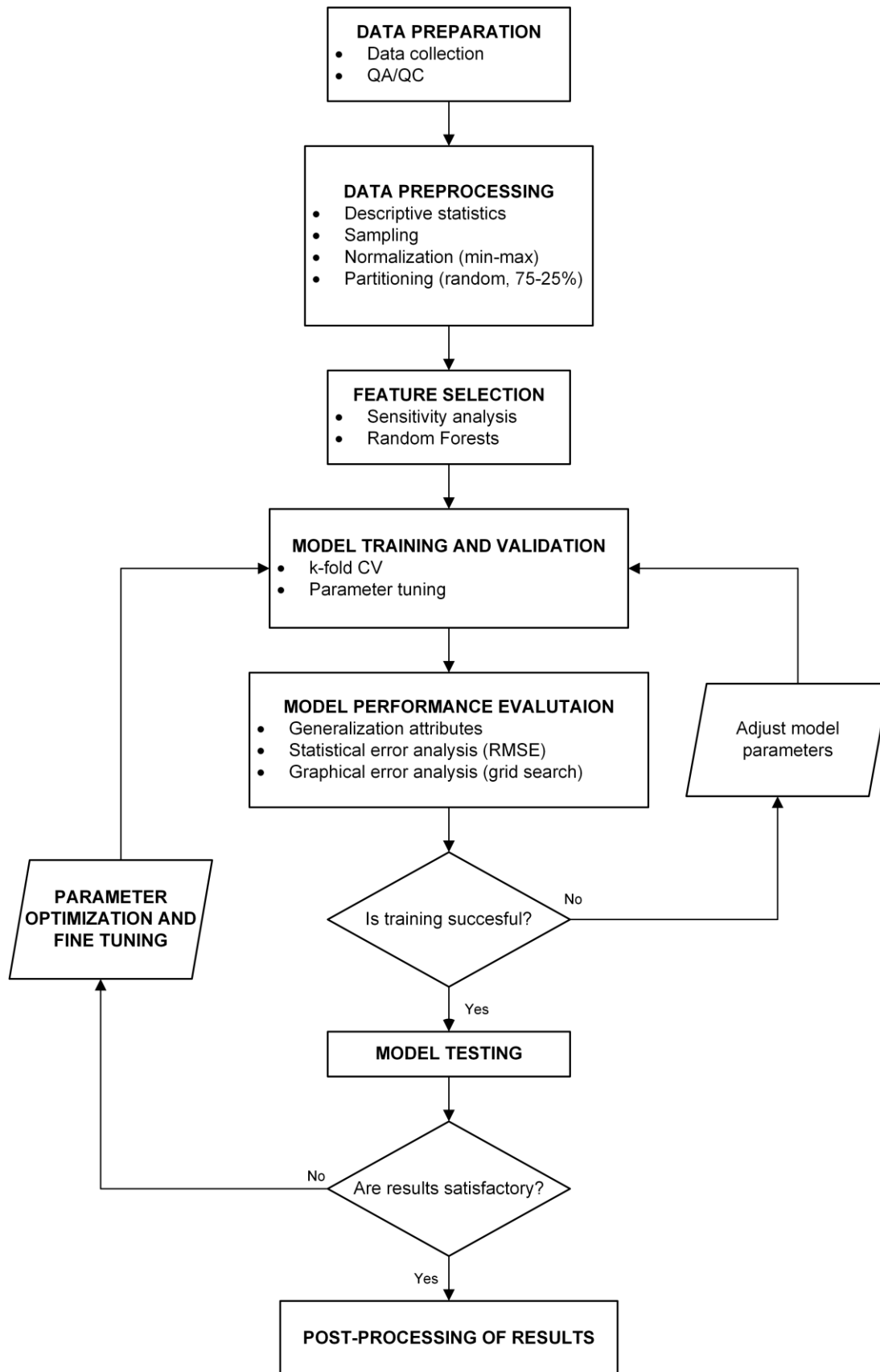


Figure 6.1 : General utilization flowchart.

Process steps in Figure 6.1 for each case and data set are explained in individual sections. The effect of additional geomechanical parameters along with the Bourgoyne & Young Model (BYM) is examined. The investigation sequence is constructed as following:

1. Default BYM parameters
2. Default BYM + Shale Index (SI)
3. Default BYM + Uniaxial Compressive Strength (UCS)
4. Default BYM + Brittleness Index (BI)
5. Default BYM + Shear Failure Gradient (SFG)
6. Default BYM + Torque (TRQ)
7. Default BYM + All additional parameters
8. Feature selection

6.1.1 Data preparation

In Case 1, data sets given in Appendix A are used for the calculations. ROP values for the data taken from the literature is given in Figure 6.2.

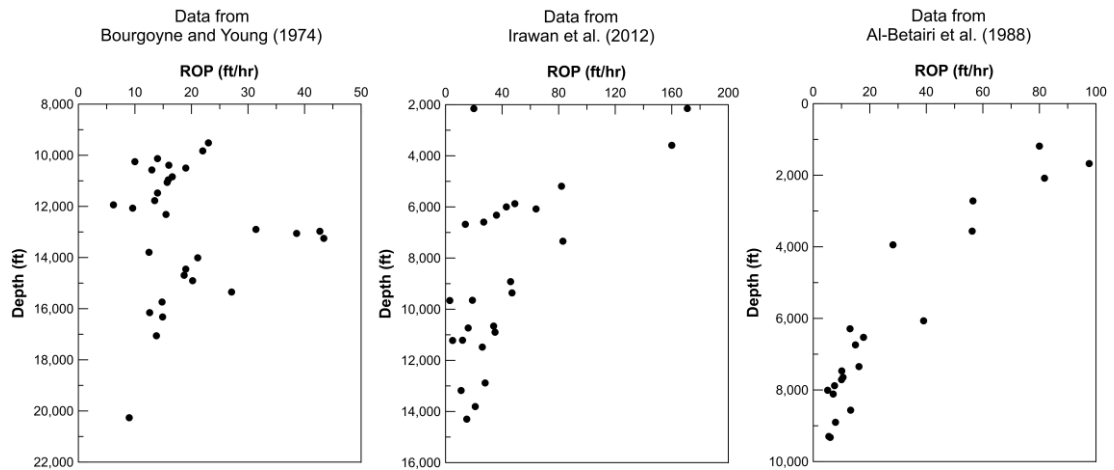


Figure 6.2 : ROP values of the datasets from the literature.

On the other hand, collection of real and recent field data with high quality constitutes the significant and essential part of this study. Data belonging to the wells drilled in one of the Turkish Petroleum Corporation fields are used in Case 2 and Case 3. Due to confidentiality, confidential information such as field name, well name, formation names, depth, and logs related to the well has been restricted. The wells are named C-

1 and C-2. The data was acquired from the ASCII version of mud logging unit (MLU) records, daily drilling reports, composite formation evaluation logs, drill bit records, and shear sonic logs. ROP values recorded in C-1 and C-2 wells are shown in Figure 6.3. Due to data confidentiality, the depth scale is restricted. ROP axis is scaled from 0 to 50 m/hr.

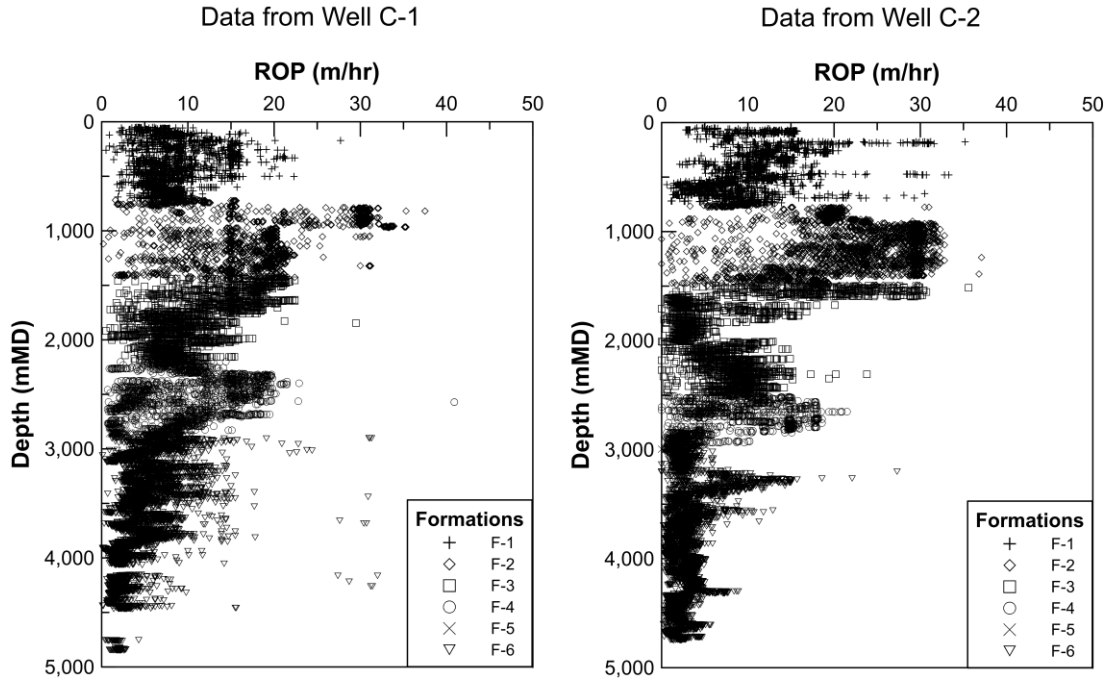


Figure 6.3 : ROP values of the datasets from the wells C-1 and C-2.

The sample rate of MLU is four records per meter. That means each ROP value represents the average penetration rate within the range of 0.25 m. Coring operations are excluded from the dataset since MLU was in standby mode. MLU records include the following drilling parameters: measured depth, average penetration rate, mud weight in and out, the flow rate in and out, hook load, weight on bit, hours on the bit, total pit volume, motor rotary speed, surface rotary speed, total rotary speed, and average standpipe pressure. Other BYM parameters such as jet impact force and equivalent circulating density are calculated via relevant formulas using MLU data for each data row. Fractional tooth wear values are granted from dull grading for each drill bit written in daily drilling reports and bit records. Geomechanical values are calculated by using gamma-ray and sonic log data. Finally, formation lithologies and heterogeneity are determined from composite formation evaluation logs. Heterogeneity is based on lithology volumes derived from gamma-ray readings. Bit torque is not available for both wells. Yet, since C-2 is a near vertical well, the

available top-drive (surface) torque data can be used instead of bit torque. The formation thickness and lithological composition of the wells C-1 and C-2 are given in Table 6.1.

Table 6.1 : Formation thickness and lithological composition for the wells C-1 and C-2.

Formation	Lithology	Thickness (m)	
		Well C-1	Well C-2
F-1	Sandstone* Claystone Conglomerate Dolomite Siltstone Shale	708	738
F-2	Siltstone* Sandstone* Claystone Shale	728	758
F-3	Shale* Siltstone* Sandstone Claystone	760	984
F-4	Marl Marl* Shale* Tuff Sandstone Siltstone Dolomite	634	478
F-5	Marl* Limestone w/ clay Tuff	36	45
F-6	Sandstone* Shale* Marl Siltstone Conglomerate	2,866	1,747

* Dominant lithology

Since ROP is the target function or dependent variable in regression analysis, it is important to estimate and analyze the noise in data. The reasons for noise can be time drilling, drilling in sliding mode, and ROP limitations by Geology and Geophysics Department for low ROP values. Time drilling is a process requiring an extremely low rate of penetration, sometimes less than 1 m/h. In directional or sidetracked wells, to kick off the well, the driller must time drill the first few meters to ease the well into its

new trajectory. Sliding is the predominant method to build and control or correct hole angle in modern directional drilling operations. Bit pointing is accomplished through the bent sub, which has a small angle offset from the axis of the drill string, and a measurement device to determine the direction of offset. Without turning the drill string, the bit is rotated with a mud motor, and drills in the direction it points. With steerable motors, when the desired wellbore direction is attained, the entire drill string is rotated and drills straight rather than at an angle. Additionally, drill breaks, MLU misleading, sensor malfunctions, and calibration errors can generate noise in accordance with high ROP values.

It is essential to perform quality assurance and quality control (QA/QC) on the dataset before using it in the regression analysis. The noise in the data should be estimated. The data points that generate noise caused by the reasons above and downscaling raw data resolution should be excluded from the main dataset. This process can be done by Tukey's Method and z-scores. First, data resolution is reduced to 1 m. In order to avoid data clutter and keep data in reserve for treating outliers, rows corresponding to integers on a depth basis have been selected for being used in the master dataset. Next, the data rows containing missing values, intervals drilled in sliding mode, and the values corresponding to one or more of the situations above are removed from the dataset. Then, outlier labeling is performed on the data before progressing to data normalization. This process is done via Tukey's method. Data rows containing one or more outlier are replaced with another data row representing the same characteristics within 1 m interval. After the data cleansing process, the raw data is ready for sampling.

6.1.2 Data preprocessing

The first process in this stage to select drilling parameters (or features) related to BYM and convert them into the functions of BYM as given in equations 4.3 through 4.9. It is essential to imply that BYM is designed for milled-tooth bits, and x_7 term can be neglected if insert or polycrystalline diamond compact (PDC) bits are used since insert bits are operated at moderate rotary speed and WOB (Bourgoyne Jr. and Young Jr., 1974). In F-4, F-5, and F-6 formations all drill bits were PDC bits for the both wells. Since tooth wearing tends to occur less in PDC bits and has a slight impact on ROP,

the x_7 term is not selected as a feature in this study. After the conversion process, geomechanical features are added to the datasets.

Then, z-score analysis is applied to the datasets to sample data correctly via equation 5.42. As mentioned previously, data that fall outside the threshold of -3 and +3 standard deviations are removed, and the datasets are finalized.

In the BYM, the functions vary on a vast scale. Hence, min-max normalization is applied to the datasets to subdue the scaling problem. This way, scaling and variance response on the output is discarded by normalizing each feature within a range of 0 to 1 via equation 5.43.

After the sampling and normalizing process, the datasets are partitioned into subsets of train and test randomly by the ratio of 75-25%, respectively.

6.1.3 Feature selection

The main goal of the calculations in this study is to perform ROP predictions in heterogeneous formations. As mentioned in the literature review, SVR could perform accurate predictions when the number of training data is limited (Kor and Altun, 2020). Since there is real field data taken recently with a sufficient number of inputs, it is possible to show MLR, SVR, and ANN's prediction performances on the dataset taken from heterogeneous lithology by implementing various geomechanical features to the BYM. Thus, to interpret the relationship between ROP and the heterogeneity, several subcases are constructed by selecting and investigating these geomechanical features individually.

Besides MLR and SVR, ANN is also used as an alternative predictive method in this study to create a benchmark. Before processing to the calculations, a sensitivity analysis to select an optimum number of hidden layers for ANN is performed. Since the initial conditions were chosen randomly, each model was trained 15 times, and the results were averaged to maintain the reliability of the results. The optimum number of hidden layers are selected by considering RMSE and correlation coefficient.

For the calculations of C-1 well, a sensitivity study was implemented to determine the most effective feature selection scenario to better map the ROP to a geomechanical feature. For the calculations of C-2 well, feature selection was performed via Random Forests (RF) method to estimate the feature importance case wise.

6.1.4 Model training and validation

After selecting features, cross-validation (CV) is applied to the training subset to avoid overfitting problems. One of the most commonly used CV methods is the k-fold CV. In this study, 10-fold CV is used, which means the model is trained on nine folds and evaluated on the remaining fold until each fold has been tested for once.

The hyperparameters (C , ε , ν , γ) of SVR are tuned via 10-fold CV and grid search. The combination of parameters that is resulted in minimum error within a given interval is selected as the optimum parameters for a specific case. After selecting the optimum parameters, the model is ready to be trained.

6.1.5 Model testing and post-processing of the results

After training and evaluating the training performance, the model is tested on a separate dataset that the model has not seen before. Three statistical performance criteria evaluate testing performance: root mean squared error (RMSE), correlation coefficient, and p-values. The null hypothesis for statistical significance test is determined as there is no difference between the means of the actual ROP values and the predicted ROP values.

6.2 Results

The applicability of different predictive models (MLR, SVR, and ANN) and the effect of additional geomechanical parameters are investigated by generating individual cases. Prediction results of training and testing steps for each case are given in graphical presentation since there are too many data to tabulate. Additionally, considering the data confidentiality, the predicted and actual ROP values are given in normalized form.

To construct a general idea about how the calculations are performed, the calculation steps of Case 1 are shown explicitly. All of the other cases are calculated with the same procedure and intermediate steps are given in Appendix D.

6.2.1 Case 1: data taken from the literature (homogeneous)

The applicability of different predictive models is investigated by generating individual cases for the drilling data taken from the literature. The data in each reference are taken from the wells drilled in homogeneous formations.

6.2.1.1 Case 1.1: calculations with data from Bourgoyne Jr. and Young Jr. (1974)

In Case 1.1, the original dataset of the Bourgoyne Jr. and Young Jr. (1974) with 30 rows is used. Since there are too few data rows, all data is used for both training and testing to evaluate the training performances. Descriptive statistics of the dataset is presented in Table 6.2.

Table 6.2 : Descriptive statistics of the dataset for Case 1.1.

	ROP (y)	Compaction (x ₂)	Under- compaction (x ₃)	Overbalance (x ₄)	WOB (x ₅)	RPM (x ₆)	Tooth Wear (x ₇)	Hydraulics (x ₈)
Number	30	30	30	30	30	30	30	30
Min	1.82E+0	-1.03E+4	0.00E+0	-3.46E+4	-1.60E+0	-5.45E-1	-9.10E-1	-1.60E+0
Max	3.77E+0	4.85E+2	7.12E+3	-3.32E+3	-6.19E-2	2.55E-1	-1.00E-2	2.55E-1
Range	1.95E+0	1.08E+4	7.12E+3	3.13E+4	1.54E+0	7.99E-1	9.00E-1	1.86E+0
Mean	2.83E+0	-2.95E+3	3.13E+3	-1.19E+4	-6.80E-1	-2.48E-1	-4.01E-1	-2.13E-1
Median	2.76E+0	-2.61E+3	4.16E+3	-8.34E+3	-6.47E-1	-2.61E-1	-3.80E-1	-9.52E-2
Q ₁	2.60E+0	-4.70E+3	1.83E+2	-1.47E+4	-8.98E-1	-3.71E-1	-5.70E-1	-2.90E-1
Q ₃	3.05E+0	-8.40E+2	5.40E+3	-6.30E+3	-3.50E-1	-2.11E-1	-2.00E-1	-1.61E-2
Variance	1.99E-1	6.53E+6	7.01E+6	6.66E+7	1.74E-1	3.46E-2	5.99E-2	1.81E-1
Std. Dev.	4.46E-1	2.56E+3	2.65E+3	8.16E+3	4.18E-1	1.86E-1	2.45E-1	4.25E-1
Skewness	3.43E-1	-8.76E-1	-1.12E-1	-1.48E+0	-5.02E-1	1.23E+0	-4.29E-1	-1.93E+0
Kurtosis	4.71E-1	6.64E-1	-1.83E+0	1.61E+0	-3.52E-1	2.09E+0	-8.19E-1	3.89E+0

The correlation plot of the dataset is represented in Figure 6.4. In Figure 6.4, scatter plots for each pair of variables are presented on the lower left, correlation coefficients and significances based on p-values are shown on the upper right as stars, histograms and density plots are placed in the middle.

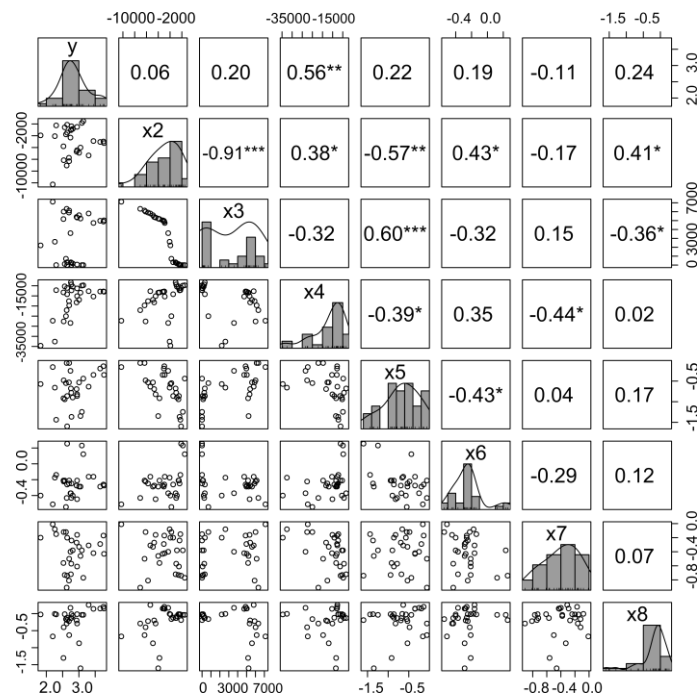


Figure 6.4 : Correlation plot of the dataset for Case 1.1.

It is seen in Figure 6.4 that x_2 and x_3 are highly negative correlated, which are the functions of depth and pore pressure gradient. Also, it can be seen in Figure 6.4 that there is a non-linear inter-relationship between the ROP and the other functions in the scatter plots.

Box-Whisker plots for each feature is given in Figure 6.5. The circles in Figure 6.5 indicate the outliers in the data.

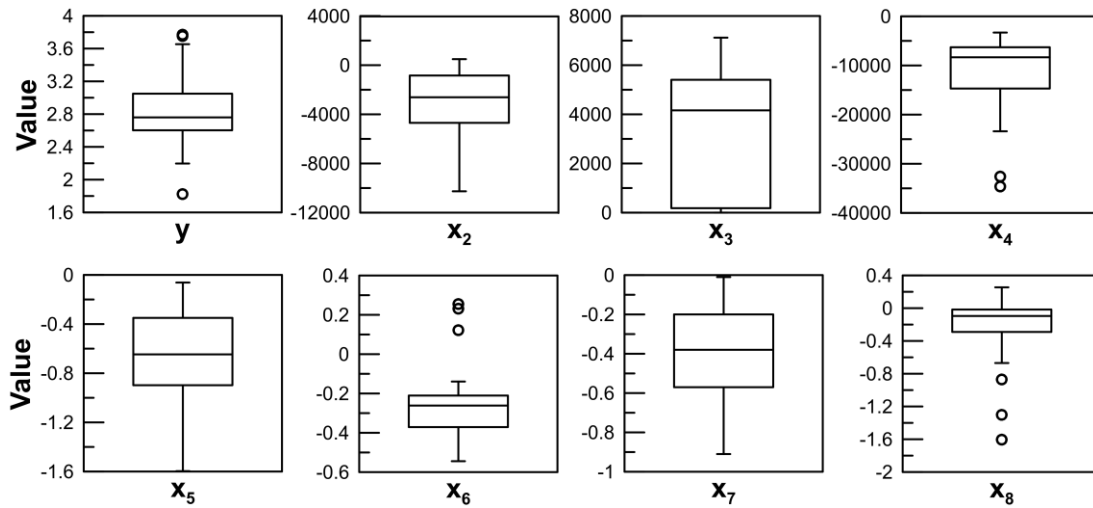


Figure 6.5 : Boxplots of the dataset for Case 1.1.

A sensitivity analysis to select a proper number of hidden layers for ANN is performed for each case individually. The number of nodes is determined case-wise through validation, which is between the sizes of the input and output layers. Since the initial conditions were chosen randomly, each model was trained 15 times, and the results were averaged to maintain the reliability of the results. The results of the sensitivity study for Case 1.1 are presented in Figure 6.6.

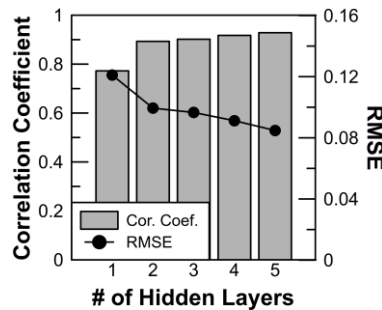


Figure 6.6 : ANN hidden layer sensitivity for Case 1.1.

According to Figure 6.6, the optimum number of hidden layers is selected as 5, since this setting has the highest correlation coefficient (0.93) and the lowest RMSE (0.085).

The SVR hyper-parameters are tuned via 5-fold CV, and grid search is plotted in Figure 6.7.

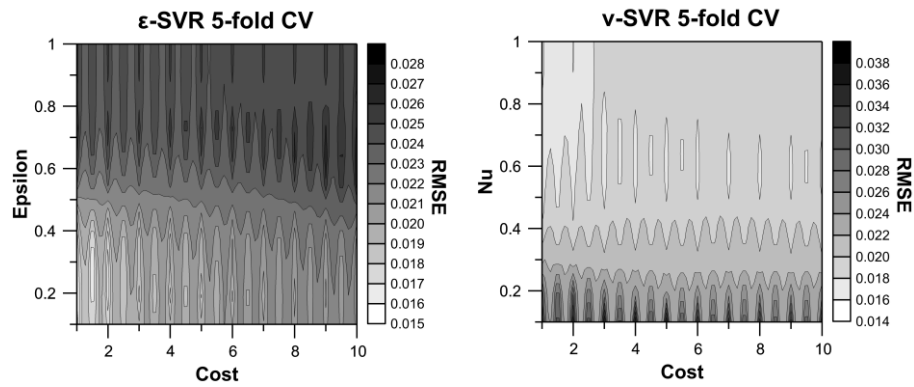


Figure 6.7 : 5-fold CV grid search plots for Case 1.1.

In Figure 6.7, hyper-parameters of SVR are selected according to the lowest RMSE corresponding to the combination of the parameters. The selected parameters for Case 1.1 are represented in Table 6.3.

Table 6.3 : SVR hyper-parameters for Case 1.1.

ϵ -SVR					ν -SVR				
Linear			RBF		Linear			RBF	
C	ϵ	C	γ	ϵ	C	ν	C	γ	ν
1.5	0.3	3.0	1.0	0.1	1.0	0.9	1.5	1.0	1.0

Normalized ROP predictions for each model are placed in Figure 6.8.

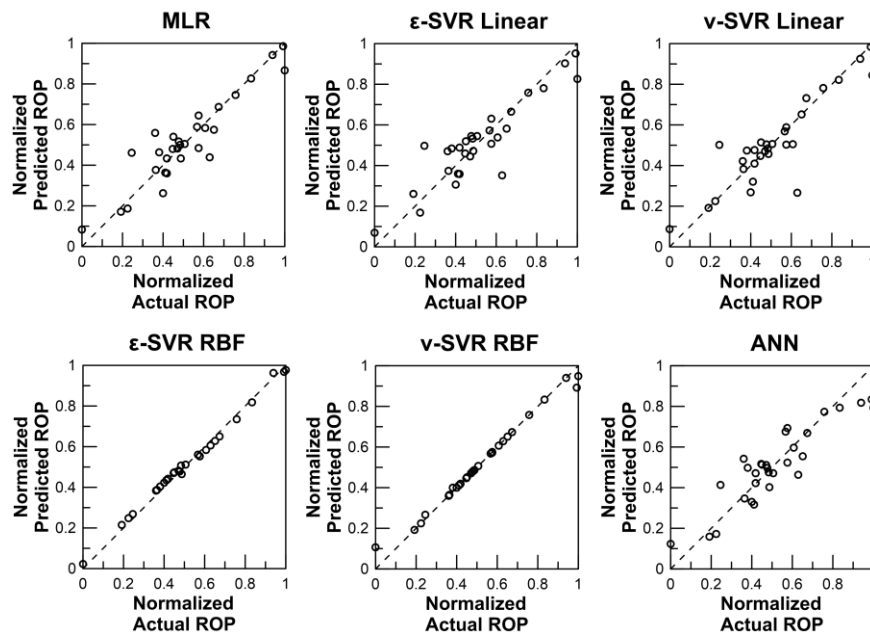


Figure 6.8 : ROP predictions for Case 1.1.

In Figure 6.8, real ROP values and predicted ROP values with each predictive model are shown in scatter plots in the normalized form. Residual histograms are given in Figure 6.9.

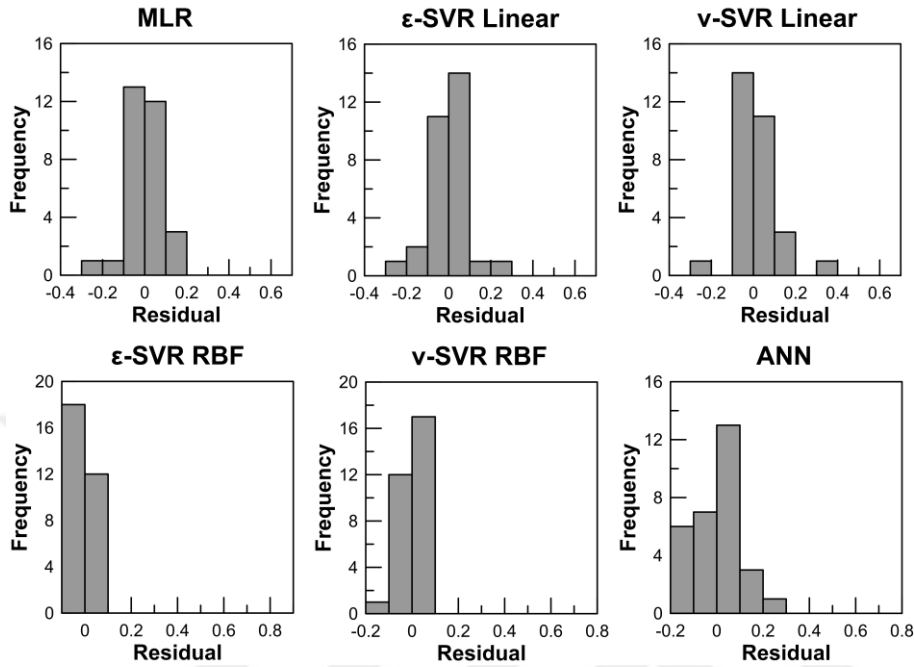


Figure 6.9 : Error histograms for Case 1.1.

Statistical comparison of the results is presented in Figure 6.10.

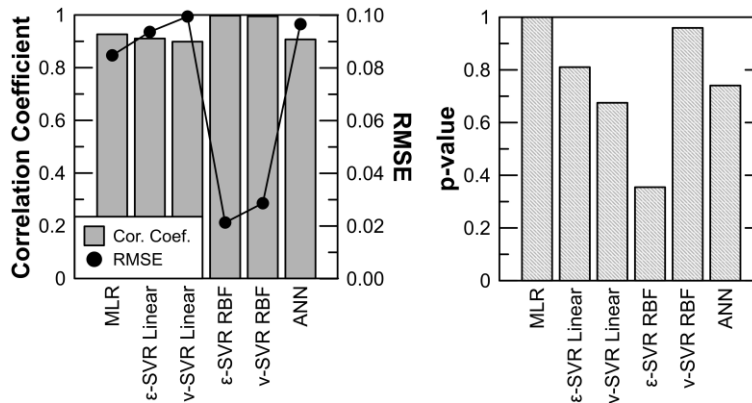


Figure 6.10 : Statistical comparison for Case 1.1.

Figure 6.10 shows that the correlation coefficient, RMSE and p-values are compared among predictive models. According to the p-values in Figure 6.10, it is failed to reject the null hypothesis. The best predictive models for Case 1.1 are SVR with RBF kernels, which have the highest correlation coefficients and the lowest RMSE values.

It can be also inferred that SVR is better model for predicting ROP rather than MLR (default regression method of the BYM) for the original data of the BYM.

6.2.1.2 Case 1.2: calculations with data from Irawan et al. (2012)

In Case 1.2, the dataset from the study of Irawan et al., (2012) with 25 rows is used. Since there are too few data rows, all data is used for both training and testing to evaluate the training performances. Descriptive statistics of the dataset is presented in Table 6.4.

Table 6.4 : Descriptive statistics of the dataset for Case 1.2.

	ROP (y)	Compaction (x ₂)	Under- compaction (x ₃)	Overbalance (x ₄)	WOB (x ₅)	RPM (x ₆)	Tooth Wear (x ₇)	Hydraulics (x ₈)
Number	25	25	25	25	25	25	25	25
Min	1.10E+0	-4.30E+3	-3.38E+2	-1.92E+4	-2.28E+0	-4.31E-1	-1.00E+0	-4.59E-1
Max	5.14E+0	7.85E+3	7.80E+2	-1.21E+3	-2.02E-1	5.88E-1	0.00E+0	2.91E-1
Range	4.04E+0	1.22E+4	1.12E+3	1.80E+4	2.08E+0	1.02E+0	1.00E+0	7.50E-1
Mean	3.35E+0	1.36E+3	-4.25E+0	-9.64E+3	-8.90E-1	2.17E-1	-3.35E-1	8.59E-3
Median	3.33E+0	6.37E+2	-1.27E+2	-9.73E+3	-8.21E-1	1.82E-1	-2.50E-1	6.02E-2
Q ₁	2.76E+0	-1.22E+3	-2.57E+2	-1.21E+4	-9.70E-1	1.61E-1	-6.25E-1	-1.54E-1
Q ₃	3.86E+0	3.94E+3	4.17E+1	-6.25E+3	-4.90E-1	4.70E-1	0.00E+0	1.81E-1
Variance	8.99E-1	1.22E+7	1.23E+5	2.28E+7	2.91E-1	8.91E-2	1.26E-1	4.98E-2
Std. Dev.	9.48E-1	3.49E+3	3.51E+2	4.78E+3	5.39E-1	2.98E-1	3.55E-1	2.23E-1
Skewness	-2.62E-1	2.37E-1	1.39E+0	-1.10E-2	-1.04E+0	-8.03E-1	-6.74E-1	-6.53E-1
Kurtosis	5.11E-1	-8.07E-1	6.65E-1	-1.69E-1	6.39E-1	3.65E-1	-9.66E-1	-5.00E-1

The correlation plot of the dataset is represented in Figure 6.11. In Figure 6.11, scatter plots for each pair of variables are presented on the lower left, correlation coefficients and significances based on p-values are shown on the upper right as stars, histograms and density plots are placed in the middle.

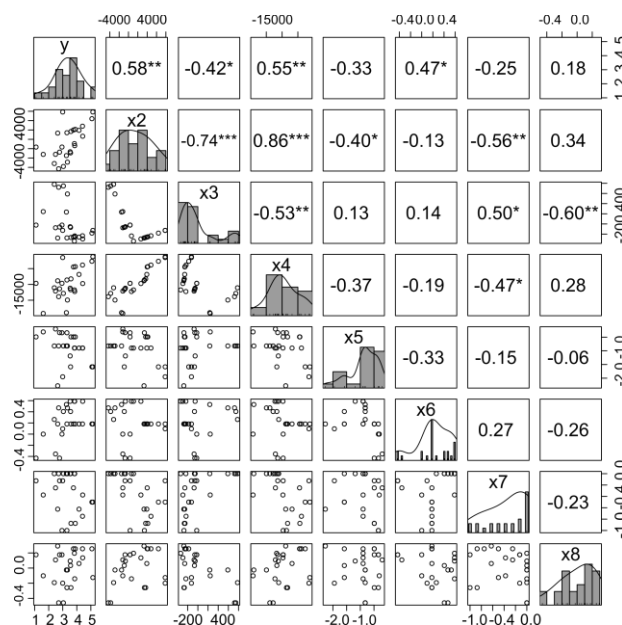


Figure 6.11 : Correlation plot of the dataset for Case 1.2

It is seen in Figure 6.11 that x_2 , x_3 , and x_4 are highly correlated, which are the functions of depth, pore pressure gradient, and ECD. Also, it can be seen in Figure 6.11 that there is a non-linear inter-relationship between the ROP and the other functions in the scatter plots.

Box-Whisker plots for each feature is given in Figure 6.12. The circles in Figure 6.12 indicate the outliers in the data.

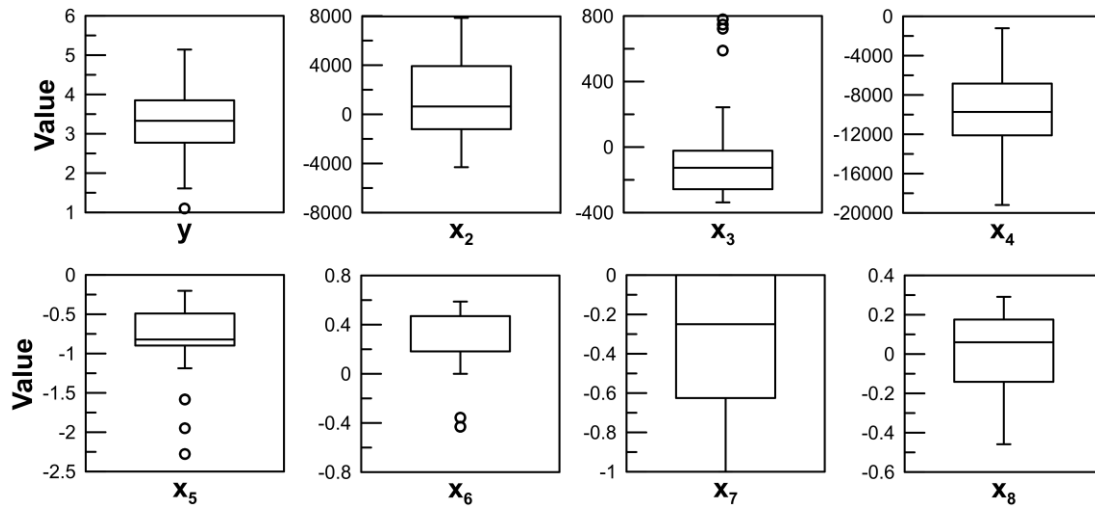


Figure 6.12 : Boxplots of the dataset for Case 1.2.

The results of the ANN sensitivity study for Case 1.2 are presented in Figure 6.13.

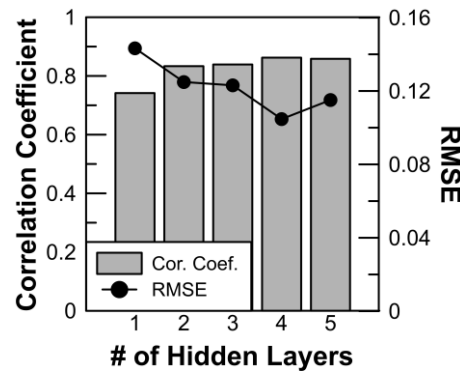


Figure 6.13 : ANN hidden layer sensitivity for Case 1.2.

According to Figure 6.13, the optimum number of hidden layers is selected as 4, since this setting has the highest correlation coefficient (0.86) and the lowest RMSE (0.1047).

The SVR hyper-parameters are tuned via 5-fold CV, and grid search is plotted in Figure 6.14.

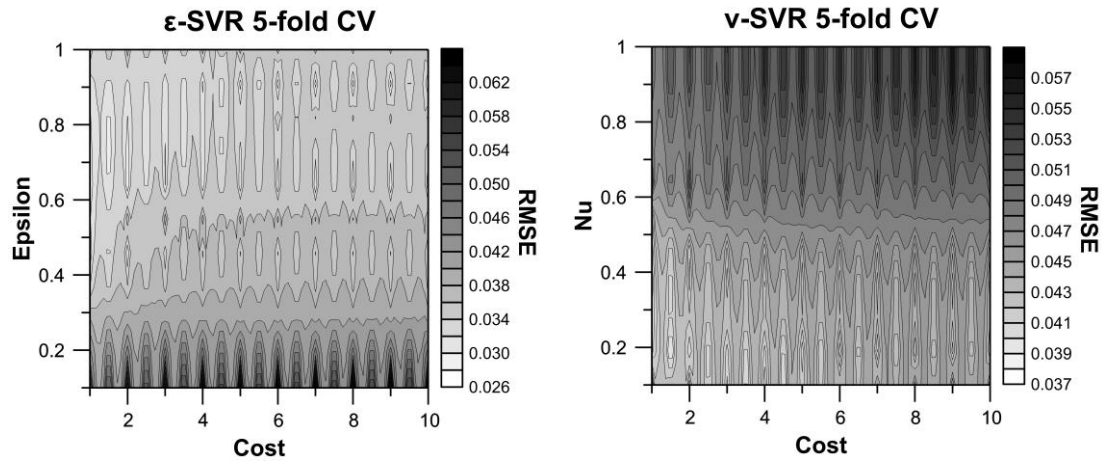


Figure 6.14 : 5-fold CV grid search plots for Case 1.2.

In Figure 6.14, hyper-parameters of SVR are selected according to the lowest RMSE corresponding to the combination of the parameters. The selected parameters for Case 1.2 are represented in Table 6.5.

Table 6.5 : SVR hyper-parameters for Case 1.2.

ε -SVR					ν -SVR				
Linear			RBF		Linear			RBF	
C	ε	C	γ	ε	C	ν	C	γ	ν
1.5	0.65	1.0	1.0	0.1	2.0	0.15	1.0	1.0	0.9

Normalized ROP predictions for each CV model are placed in Figure 6.15.

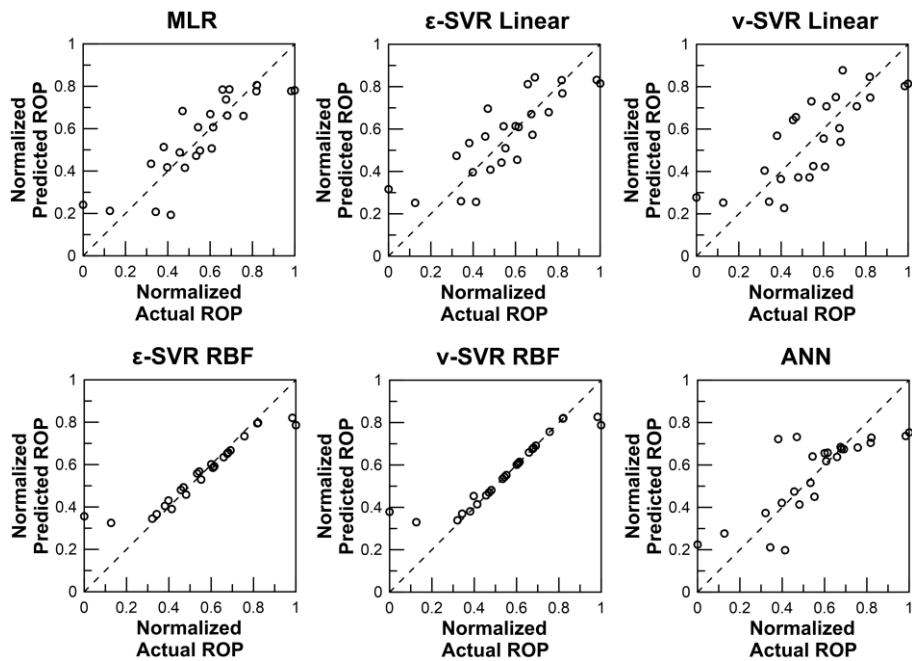


Figure 6.15 : ROP predictions for Case 1.2.

In Figure 6.15, real ROP values and predicted ROP values with each predictive model are shown in scatter plots in the normalized form. Error histograms are given in Figure 6.16.

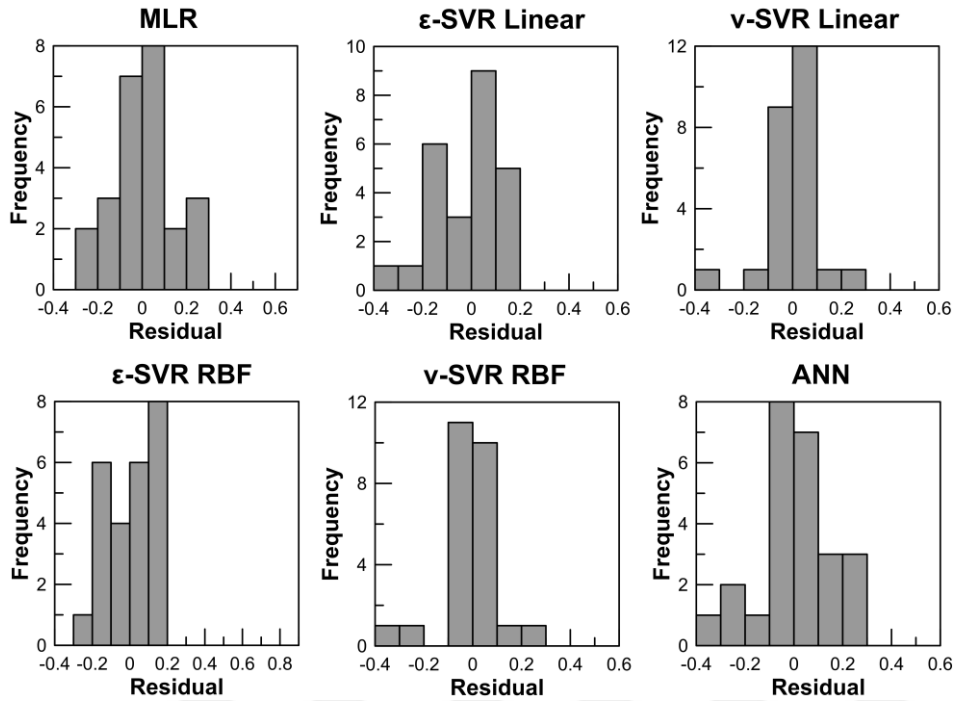


Figure 6.16 : Error histograms for Case 1.2.

Statistical comparison of the results is presented in Figure 6.17.

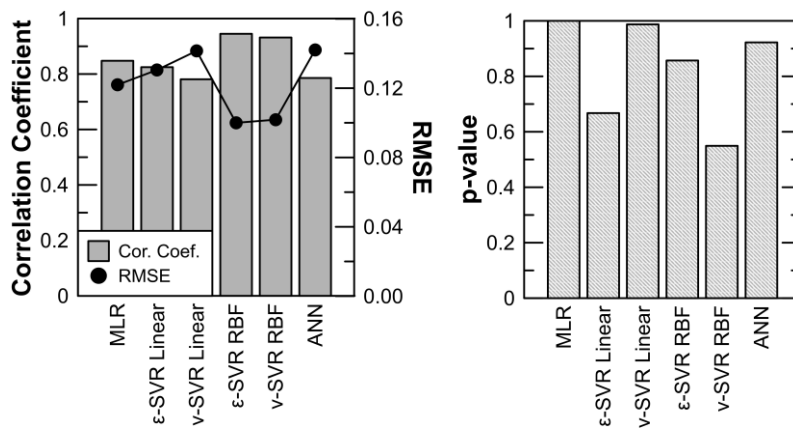


Figure 6.17 : Statistical comparison for Case 1.2.

According to the p-values in Figure 6.17, it is failed to reject the null hypothesis. The best predictive models for Case 1.2 are SVR with RBF kernels, which have the highest correlation coefficients and the lowest RMSE values. It can be also inferred that SVR is better model for predicting ROP rather than MLR for this dataset.

6.2.1.3 Case 1.3: calculations with data from Al-Betairi et al. (1988)

In Case 1.3, the dataset from the study of Al-Betairi et al., (1988) with 21 rows is used. Since there are too few data rows, all data is used for both training and testing to evaluate the training performances. Descriptive statistics of the dataset is presented in Table 6.6.

Table 6.6 : Descriptive statistics of the dataset for Case 1.3.

	ROP (y)	Compaction (x ₂)	Under- compaction (x ₃)	Overbalance (x ₄)	WOB (x ₅)	RPM (x ₆)	Tooth Wear (x ₇)	Hydraulics (x ₈)
Number	21	21	21	21	21	21	21	21
Min	1.63E+0	6.75E+2	-2.99E+2	-3.79E+4	-5.50E+0	-1.63E-1	-1.00E+0	1.64E+1
Max	4.58E+0	8.81E+3	2.74E+2	-1.10E+4	2.28E-1	1.82E-1	0.00E+0	4.89E+1
Range	2.96E+0	8.14E+3	5.73E+2	2.70E+4	5.73E+0	3.45E-1	1.00E+0	3.25E+1
Mean	2.86E+0	3.76E+3	1.06E+2	-2.76E+4	-3.96E-1	2.05E-2	-3.99E-1	4.00E+1
Median	2.58E+0	2.65E+3	1.55E+2	-3.01E+4	2.31E-2	0.00E+0	-5.00E-1	4.07E+1
Q ₁	2.05E+0	1.97E+3	6.69E+1	-3.18E+4	-1.78E-1	-2.64E-2	-6.25E-1	4.07E+1
Q ₃	3.76E+0	6.15E+3	2.21E+2	-2.61E+4	1.55E-1	9.53E-2	-1.25E-1	4.12E+1
Variance	9.25E-1	6.82E+6	3.30E+4	6.79E+7	1.78E+0	1.11E-2	9.32E-2	5.16E+1
Std. Dev.	9.62E-1	2.61E+3	1.82E+2	8.24E+3	1.33E+0	1.05E-1	3.05E-1	7.19E+0
Skewness	5.66E-1	7.95E-1	-1.55E+0	9.64E-1	-3.37E+0	1.89E-1	-2.57E-1	-2.22E+0
Kurtosis	-1.04E+0	-7.24E-1	1.37E+0	-1.57E-1	1.17E+1	-7.92E-1	-9.20E-1	6.16E+0

The correlation plot of the dataset is represented in Figure 6.18. In Figure 6.18, scatter plots for each pair of variables are presented on the lower left, correlation coefficients and significances based on p-values are shown on the upper right as stars, histograms and density plots are placed in the middle.

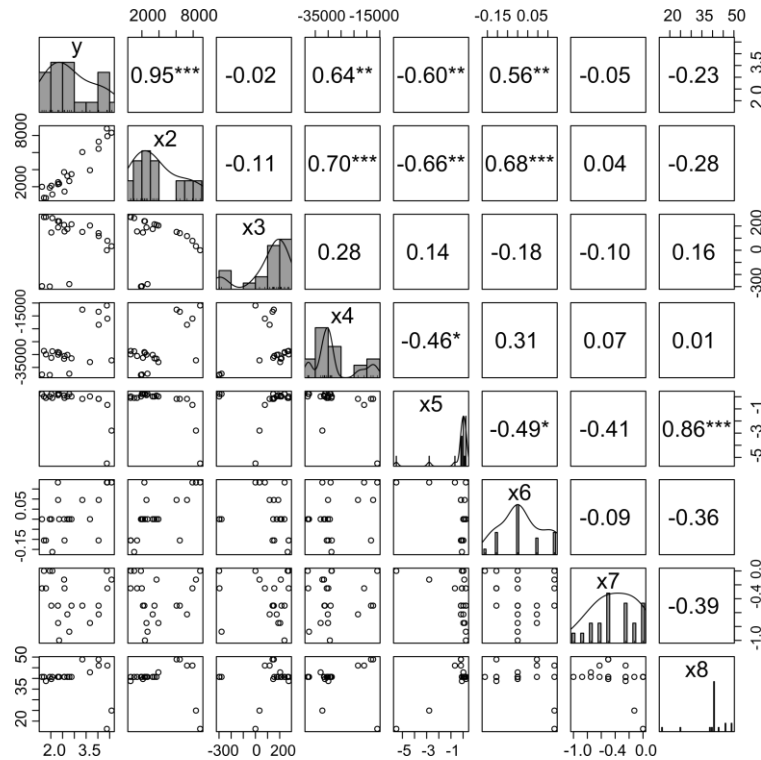


Figure 6.18 : Correlation plot of the dataset for Case 1.3.

It is seen in Figure 6.18 that y , x_2 , and x_4 are highly correlated, which are the functions of ROP, depth, pore pressure gradient, and ECD. Additionally, x_5 and x_8 are also high-correlated, which are the functions of WOB and hydraulics.

Box-Whisker plots for each feature is given in Figure 6.19. The circles in Figure 6.19 indicate the outliers in the data.

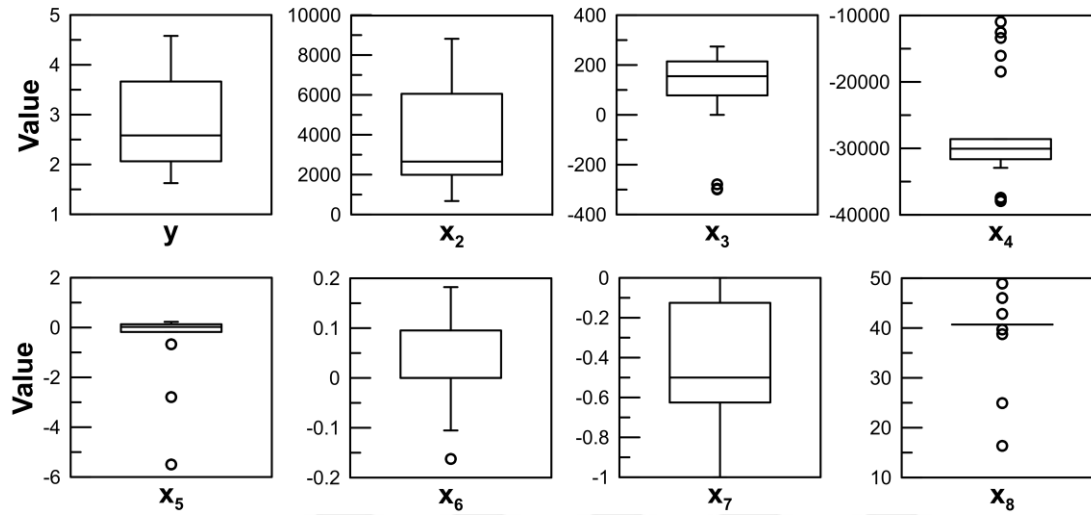


Figure 6.19 : Boxplots of the dataset for Case 1.3.

The results of the ANN sensitivity study for Case 1.3 are presented in Figure 6.20.

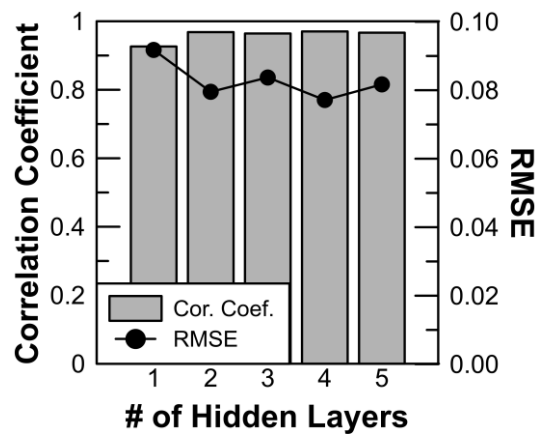


Figure 6.20 : ANN hidden layer sensitivity for Case 1.3.

According to Figure 6.20, the optimum number of hidden layers is selected as 4, since this setting has the highest correlation coefficient (0.97) and the lowest RMSE (0.0771).

The SVR hyper-parameters are tuned via 5-fold CV, and grid search is plotted in Figure 6.21.

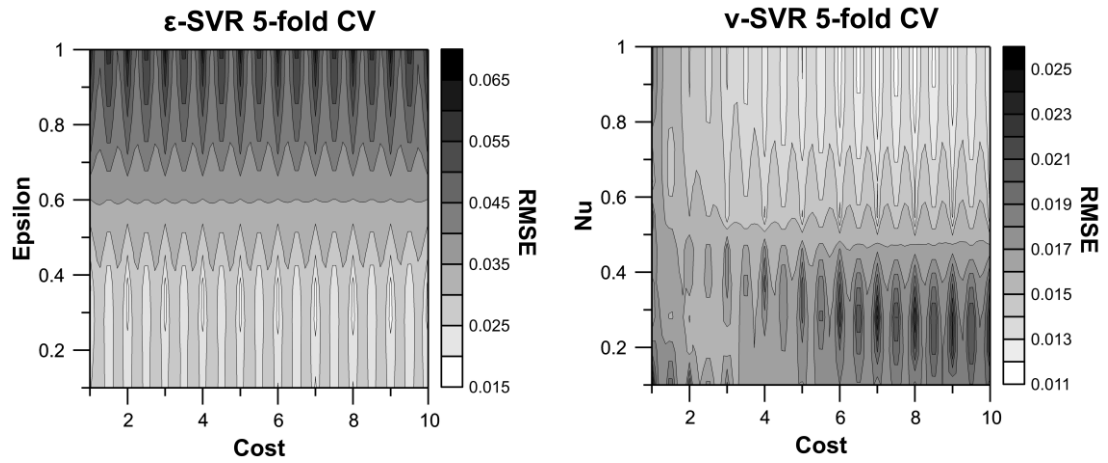


Figure 6.21 : 5-fold CV grid search plots for Case 1.3.

In Figure 6.21, hyper-parameters of SVR are selected according to the lowest RMSE corresponding to the combination of the parameters. The selected parameters for Case 1.3 are represented in Table 6.7.

Table 6.7 : SVR hyper-parameters for Case 1.3.

ε -SVR					ν -SVR				
Linear			RBF		Linear			RBF	
C	ε	C	γ	ε	C	ν	C	γ	ν
10.0	0.3	2.0	1.0	0.1	7.5	0.95	4.5	1.0	0.2

Normalized ROP predictions for each model are placed in Figure 6.22.

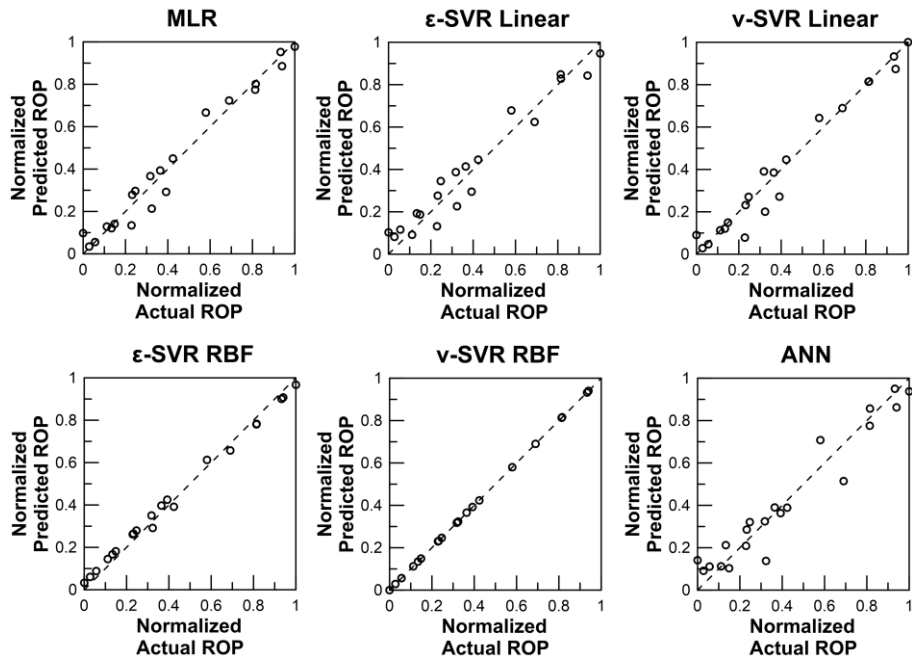


Figure 6.22 : ROP predictions for Case 1.3.

In Figure 6.22, real ROP values and predicted ROP values with each predictive model are shown in scatter plots in the normalized form. Error histograms are given in Figure 6.23.

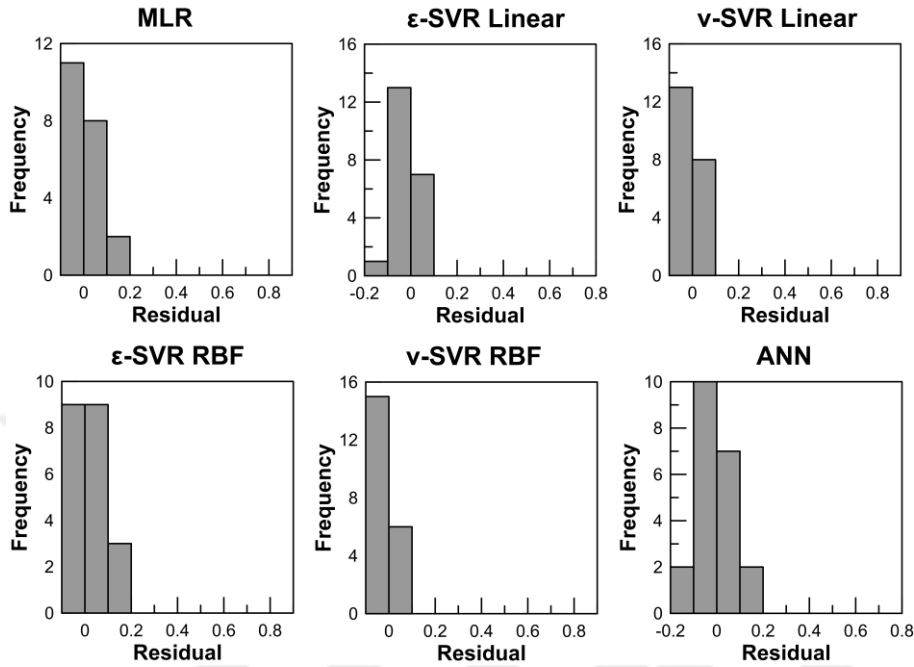


Figure 6.23 : Error histograms for Case 1.3.

Statistical comparison of the results is presented in Figure 6.24.

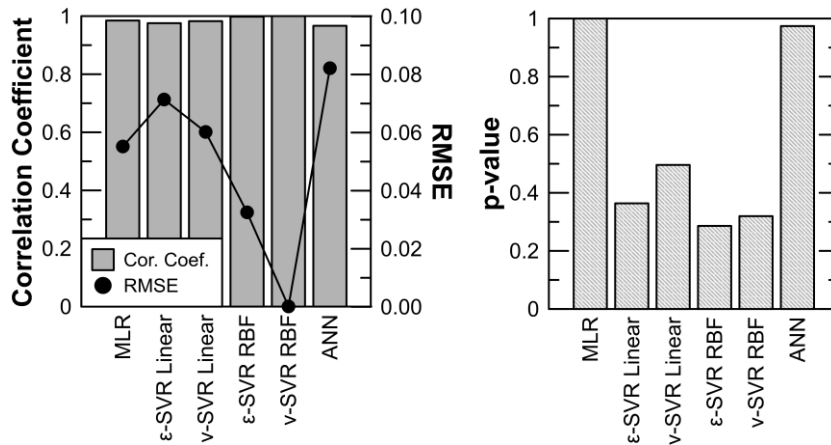


Figure 6.24 : Statistical comparison for Case 1.3.

According to the p-values in Figure 6.24, it is failed to reject the null hypothesis. The best predictive models for Case 1.3 are SVR with RBF kernels, which have the highest correlation coefficients and the lowest RMSE values. It can be also inferred that SVR is better model for predicting ROP rather than MLR and ANN for this dataset.

6.2.2 Case 2: data taken from the well C-1

The calculations performed in Case 2.1 and Case 2.2 (F-4 and F-5 formations) belong to the homogeneous formations. The effect of several geomechanical features with different predictive models on ROP prediction is investigated by generating individual cases for the drilling data taken from the heterogeneous F-6 formation in Well C-1.

6.2.2.1 Case 2.1: calculations with data from F-4

For calculations in Case 2.1, a dataset with 57 rows is used. Descriptive statistics of the dataset is presented in Table 6.8.

Table 6.8 : Descriptive statistics of the dataset for Case 2.1.

	ROP (y)	Compaction (x ₂)	Under- compaction (x ₃)	Overbalance (x ₄)	WOB (x ₅)	RPM (x ₆)	Tooth Wear (x ₇)
Number	57	57	57	57	57	57	57
Min	9.65E-1	7.21E+2	-3.55E+1	-1.90E+4	-1.67E+0	-1.63E-1	-1.51E-1
Max	3.90E+0	9.05E+2	3.39E+2	-7.82E+3	-2.24E-1	4.12E-1	4.43E-2
Range	2.94E+0	1.84E+2	3.75E+2	1.12E+4	1.44E+0	5.75E-1	1.95E-1
Mean	2.87E+0	8.13E+2	1.56E+2	-1.44E+4	-6.11E-1	1.81E-1	-9.92E-2
Median	3.30E+0	8.13E+2	1.80E+2	-1.45E+4	-5.38E-1	2.93E-1	-9.77E-2
Q ₁	1.98E+0	7.66E+2	7.32E+1	-1.63E+4	-8.21E-1	-8.92E-2	-1.21E-1
Q ₃	3.45E+0	8.59E+2	2.34E+2	-1.28E+4	-3.42E-1	3.15E-1	-8.58E-2
Variance	7.00E-1	2.96E+3	1.15E+4	7.07E+6	1.03E-1	4.29E-2	9.20E-4
Std. Dev.	8.36E-1	5.44E+1	1.07E+2	2.66E+3	3.22E-1	2.07E-1	3.03E-2
Skewness	-1.02E+0	0.00E+0	-1.03E-1	3.18E-1	-1.08E+0	-9.82E-1	1.82E+0
Kurtosis	-5.04E-1	-1.20E+0	-7.28E-1	-4.16E-1	1.02E+0	-9.66E-1	7.74E+0

The correlation plot of the dataset is represented in Figure 6.25. In Figure 6.25, scatter plots for each pair of variables are presented on the lower left, correlation coefficients and significances based on p-values are shown on the upper right as stars, histograms and density plots are placed in the middle.

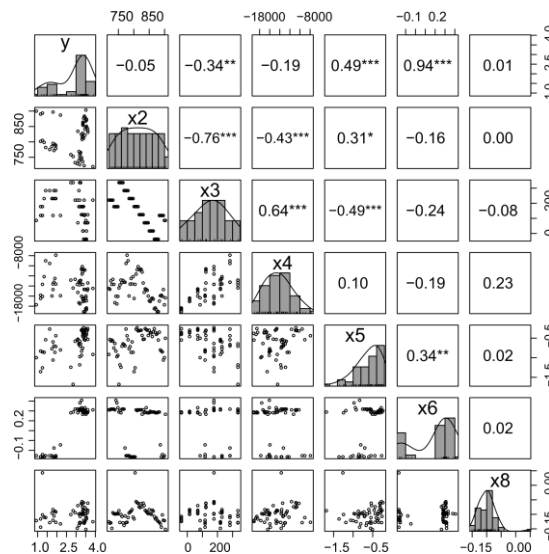


Figure 6.25 : Correlation plot of the dataset for Case 2.1.

It is seen in Figure 6.25 that y and x_6 are highly correlated, which are the functions of ROP and RPM. Additionally, x_2 and x_3 are also high-correlated, which are the functions of depth and pore pressure.

Box-Whisker plots for each feature is given in Figure 6.26. The circles in Figure 6.26 indicate the outliers in the data.

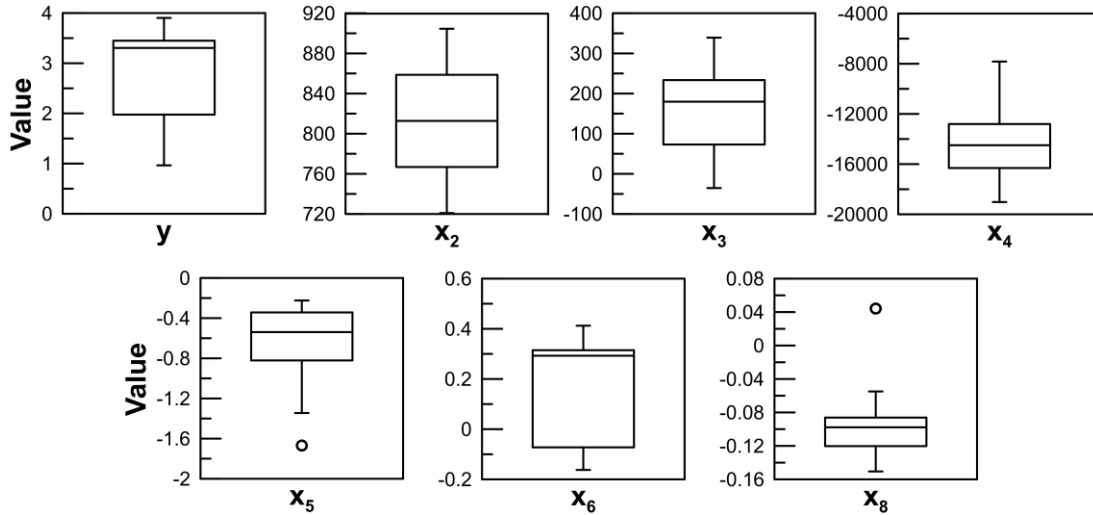


Figure 6.26 : Boxplots of the dataset for Case 2.1.

The results of the ANN sensitivity study for Case 2.1 are presented in Figure 6.27.

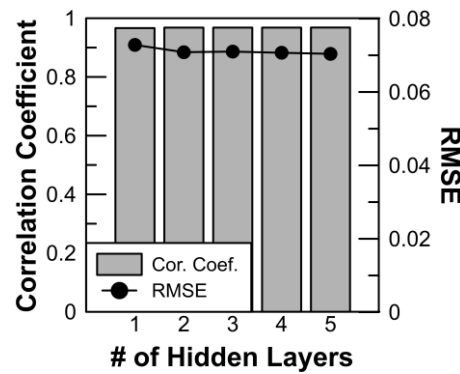


Figure 6.27 : ANN hidden layer sensitivity for Case 2.1.

According to Figure 6.27, the optimum number of hidden layers is selected as 5, since this setting has the highest correlation coefficient (0.97) and the lowest RMSE (0.0704).

The data is randomly split into training and testing subsets with the ratio of 75% and 25%, respectively. The SVR hyper-parameters are tuned via 10-fold CV, and grid search is plotted in Figure 6.28.

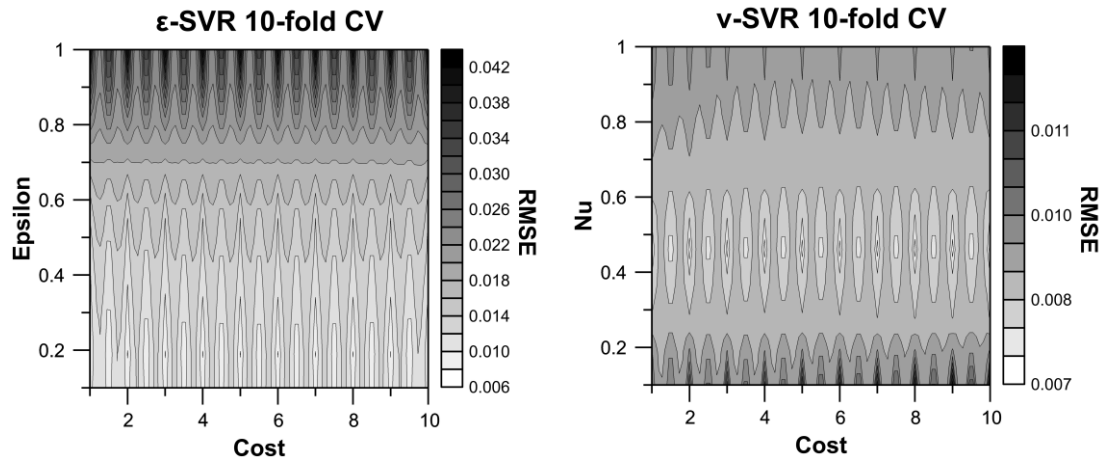


Figure 6.28 : 10-fold CV grid search plots for Case 2.1.

In Figure 6.28, hyper-parameters of SVR are selected according to the lowest RMSE corresponding to the combination of the parameters. The selected parameters for Case 2.1 are represented in Table 6.9.

Table 6.9 : SVR hyper-parameters for Case 2.1.

ϵ -SVR					ν -SVR				
Linear		RBF			Linear		RBF		
C	ϵ	C	γ	ϵ	C	ν	C	γ	ν
3.5	0.25	2.5	1.0	0.1	2.5	0.4	3.5	1.0	0.2

Scatterplots for the model training is given in Figure 6.29.

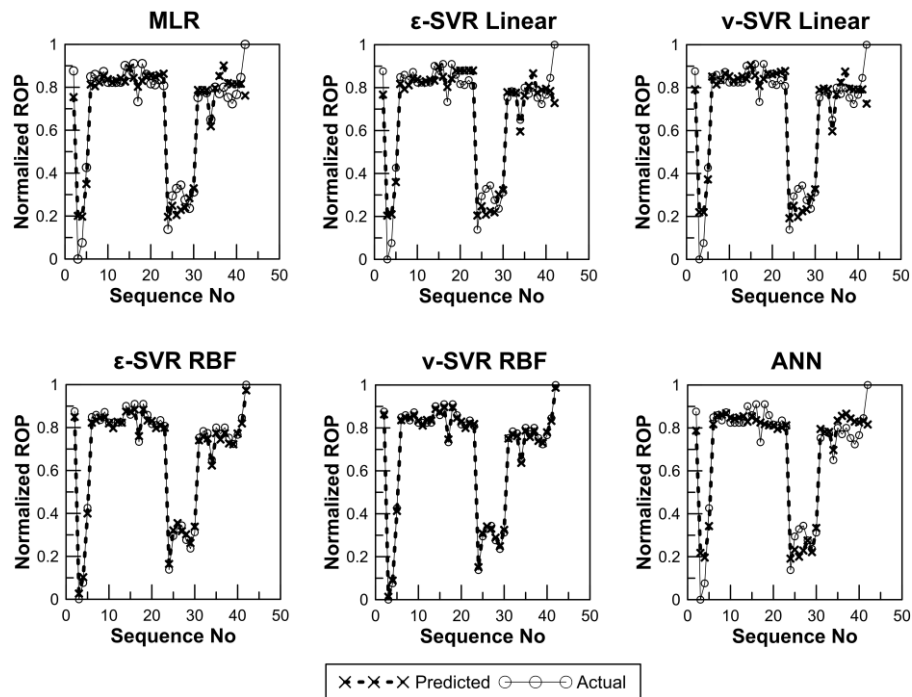


Figure 6.29 : Model training for Case 2.1.

Training errors for each predictive model are presented in Figure 6.30.

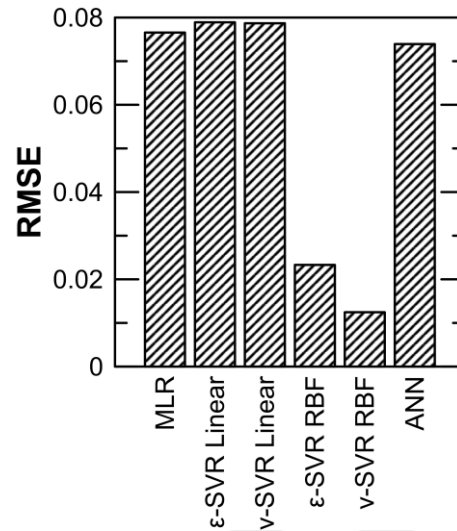


Figure 6.30 : Training errors for Case 2.1.

Normalized ROP predictions with testing subset for each model are placed in Figure 6.31.

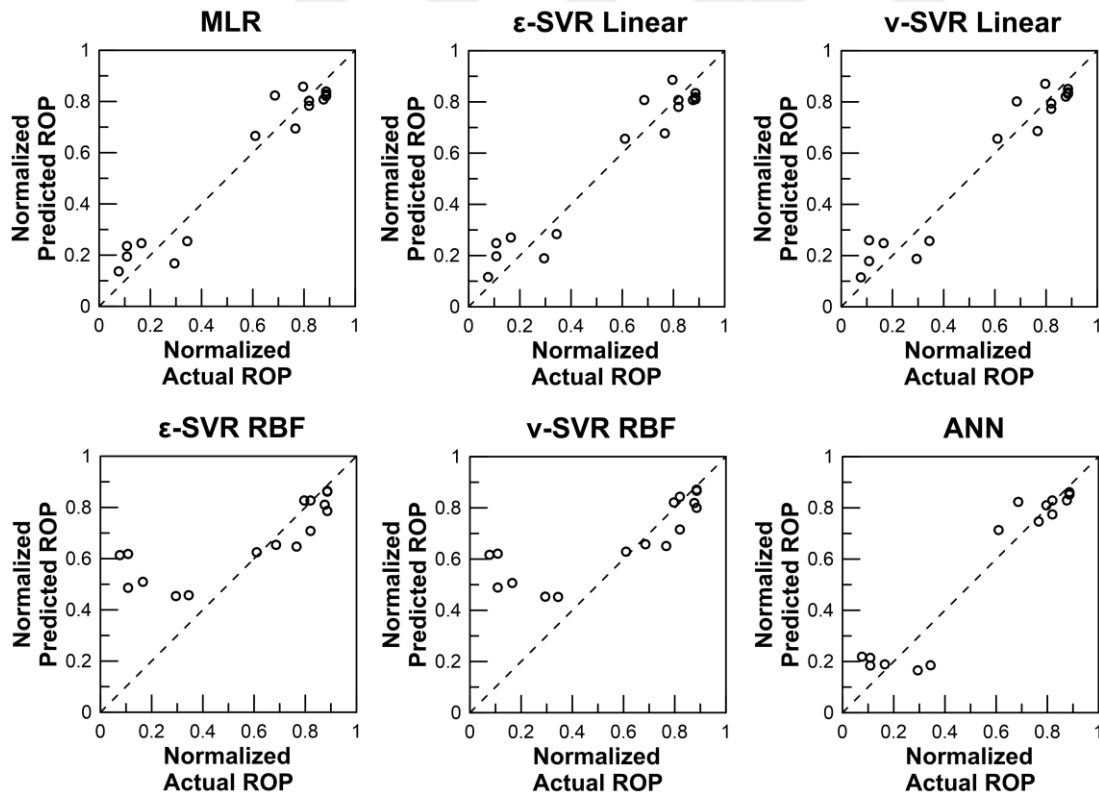


Figure 6.31 : ROP predictions with testing subset for Case 2.1.

Error histograms are given in Figure 6.32.

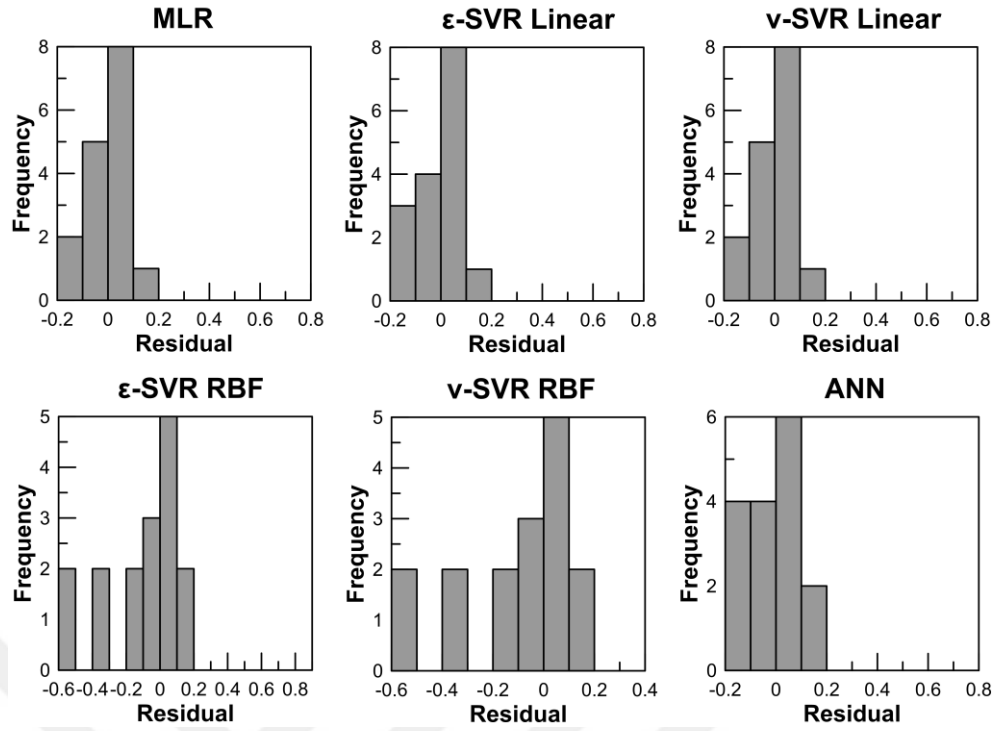


Figure 6.32 : Error histograms for Case 2.1.

Statistical comparison of the results is presented in Figure 6.33.

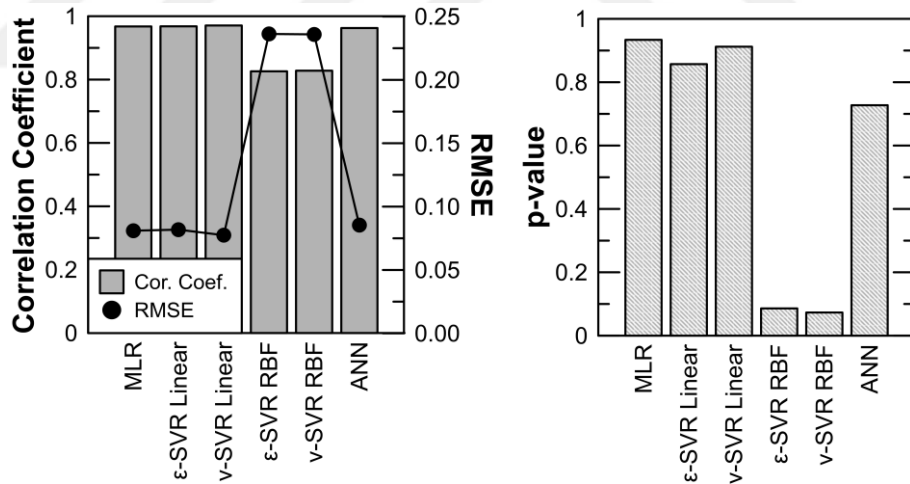


Figure 6.33 : Statistical comparison for Case 2.1.

According to the p-values in Figure 6.33, it is failed to reject the null hypothesis. The best predictive models for Case 2.1 are MLR, SVR with linear kernels, and ANN, which have the highest correlation coefficients with high p-values and the lowest RMSE values. It can be also inferred from Figures 6.30, 6.32, and 6.33 that linear models give better testing results rather than RBF kernels, although RBF kernels give the best training performance.

6.2.2.2 Case 2.2: calculations with data from F-5

For calculations in Case 2.2, a dataset with 34 rows is used. Descriptive statistics of the dataset is presented in Table 6.10.

Table 6.10 : Descriptive statistics of the dataset for Case 2.2.

	ROP (y)	Compaction (x ₂)	Under- compaction (x ₃)	Overbalance (x ₄)	WOB (x ₅)	RPM (x ₆)	Hydraulics (x ₈)
Number	34	34	34	34	34	34	34
Min	1.37E+0	6.03E+2	2.48E+2	-1.82E+4	-1.16E+0	-1.74E-1	-6.02E-1
Max	3.36E+0	7.18E+2	5.39E+2	-2.80E+3	-7.81E-2	4.19E-1	-2.35E-2
Range	1.99E+0	1.15E+2	2.91E+2	1.54E+4	1.08E+0	5.93E-1	5.79E-1
Mean	2.48E+0	6.60E+2	3.61E+2	-1.00E+4	-6.52E-1	1.04E-1	-1.59E-1
Median	2.75E+0	6.57E+2	3.46E+2	-1.03E+4	-6.61E-1	2.66E-1	-1.28E-1
Q ₁	1.66E+0	6.29E+2	2.66E+2	-1.18E+4	-8.84E-1	-1.63E-1	-1.87E-1
Q ₃	3.18E+0	6.91E+2	4.43E+2	-8.33E+3	-5.31E-1	3.36E-1	-1.12E-1
Variance	5.82E-1	1.27E+3	8.66E+3	1.04E+7	8.82E-2	5.94E-2	1.17E-2
Std. Dev.	7.63E-1	3.56E+1	9.31E+1	3.23E+3	2.97E-1	2.44E-1	1.08E-1
Skewness	-2.41E-1	1.70E-2	2.68E-1	1.50E-1	3.96E-1	-1.82E-1	-2.58E+0
Kurtosis	-1.77E+0	-1.31E+0	-1.37E+0	5.78E-1	-5.16E-1	-2.01E+0	8.35E+0

The correlation plot of the dataset is represented in Figure 6.34. In Figure 6.34, scatter plots for each pair of variables are presented on the lower left, correlation coefficients and significances based on p-values are shown on the upper right as stars, histograms and density plots are placed in the middle.

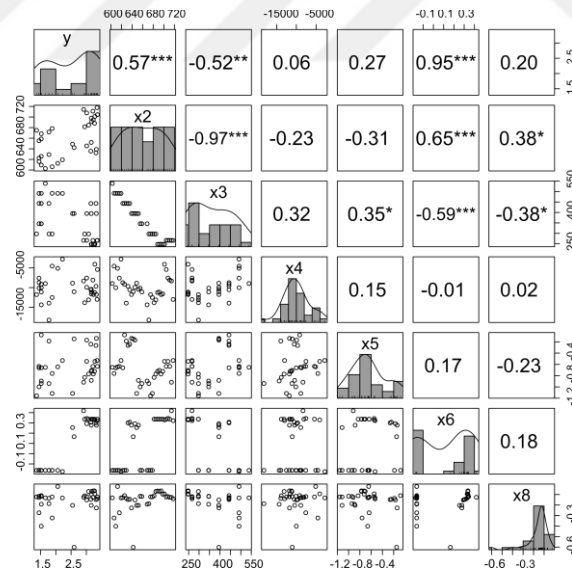


Figure 6.34 : Correlation plot of the dataset for Case 2.2.

It is seen in Figure 6.34 that y and x₆ are highly correlated, which are the functions of ROP and RPM. Additionally, x₂ and x₃ are also high-correlated, which are the functions of depth and pore pressure.

Box-Whisker plots for each feature is given in Figure 6.35. The circles in Figure 6.35 indicate the outliers in the data.

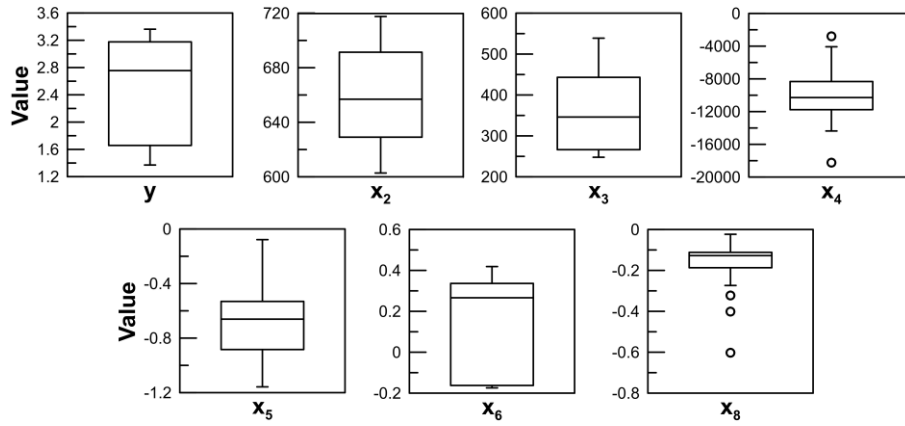


Figure 6.35 : Boxplots of the dataset for Case 2.2.

The results of the ANN sensitivity study for Case 2.2 are presented in Figure 6.36.

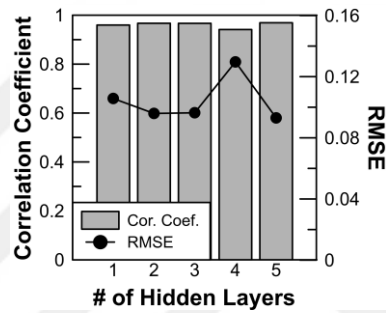


Figure 6.36 : ANN hidden layer sensitivity for Case 2.2.

According to Figure 6.36, the optimum number of hidden layers is selected as 5, since this setting has the highest correlation coefficient (0.97) and the lowest RMSE (0.0931).

The data is randomly split into training and testing subsets with the ratio of 75% and 25%, respectively. The SVR hyper-parameters are tuned via 5-fold CV, and grid search is plotted in Figure 6.37.

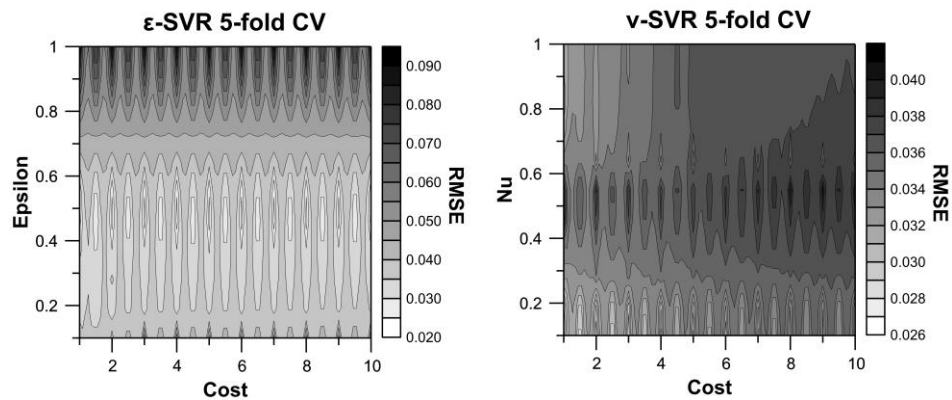


Figure 6.37 : 5-fold CV grid search plots for Case 2.2.

In Figure 6.37, hyper-parameters of SVR are selected according to the lowest RMSE corresponding to the combination of the parameters. The selected parameters for Case 2.2 are represented in Table 6.11.

Table 6.11 : SVR hyper-parameters for Case 2.2.

ε -SVR					ν -SVR				
Linear			RBF		Linear			RBF	
C	ε	C	γ	ε	C	ν	C	γ	ν
1.0	0.5	1.5	1.0	0.1	1.0	0.1	1.5	1.0	1.0

Scatterplots for the model training is given in Figure 6.38.

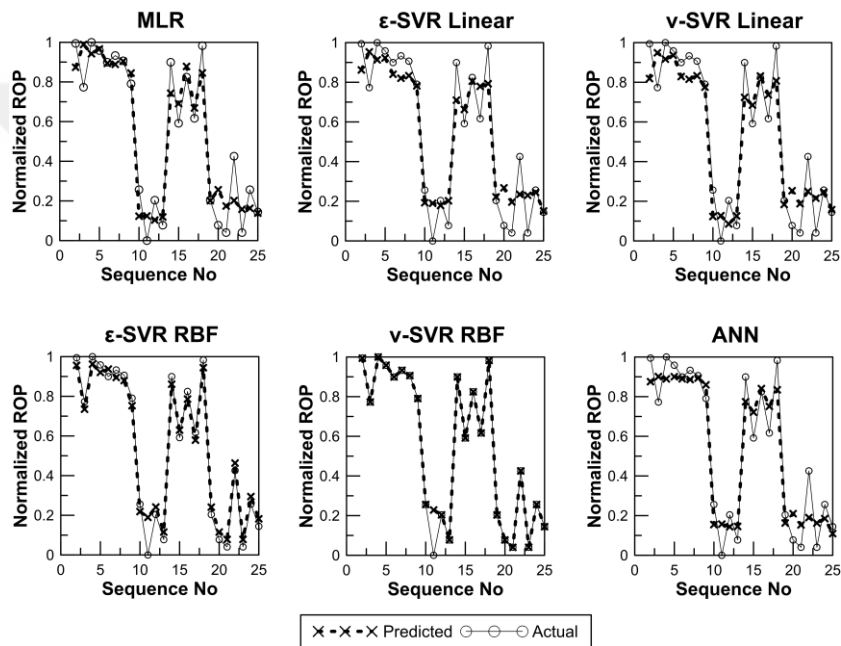


Figure 6.38 : Model training for Case 2.2.

Training errors for each predictive model are presented in Figure 6.39.

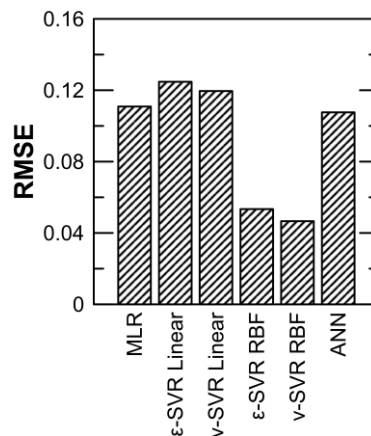


Figure 6.39 : Training errors for Case 2.2.

Normalized ROP predictions with testing subset for each model are placed in Figure 6.40.

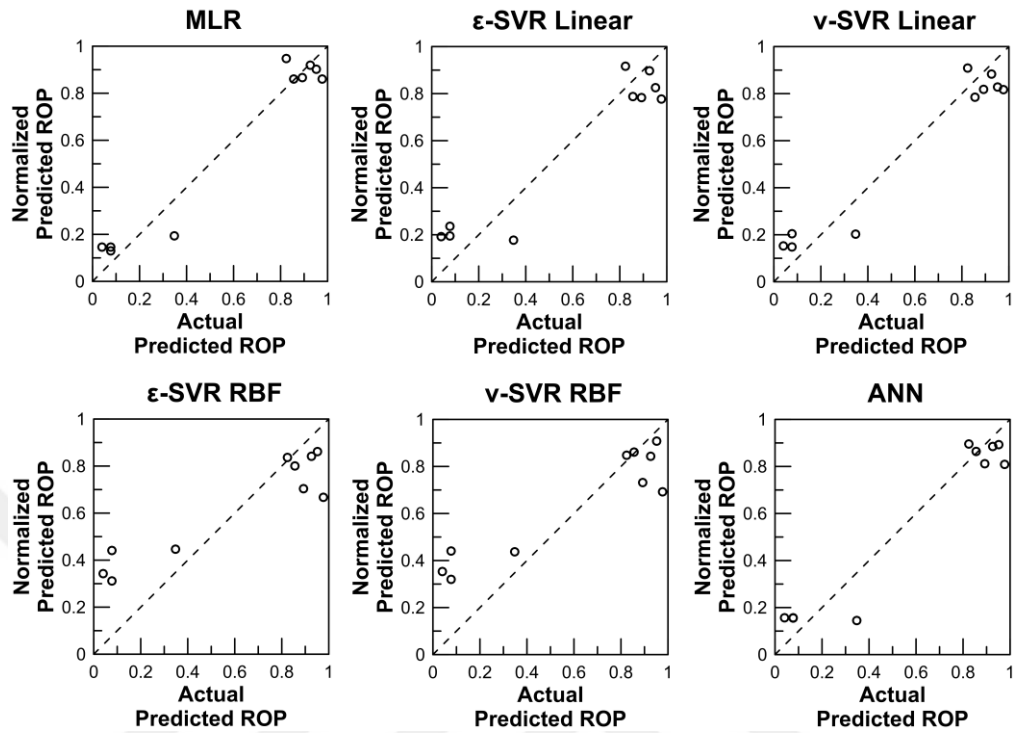


Figure 6.40 : ROP predictions with testing subset for Case 2.2.

Error histograms are given in Figure 6.41.

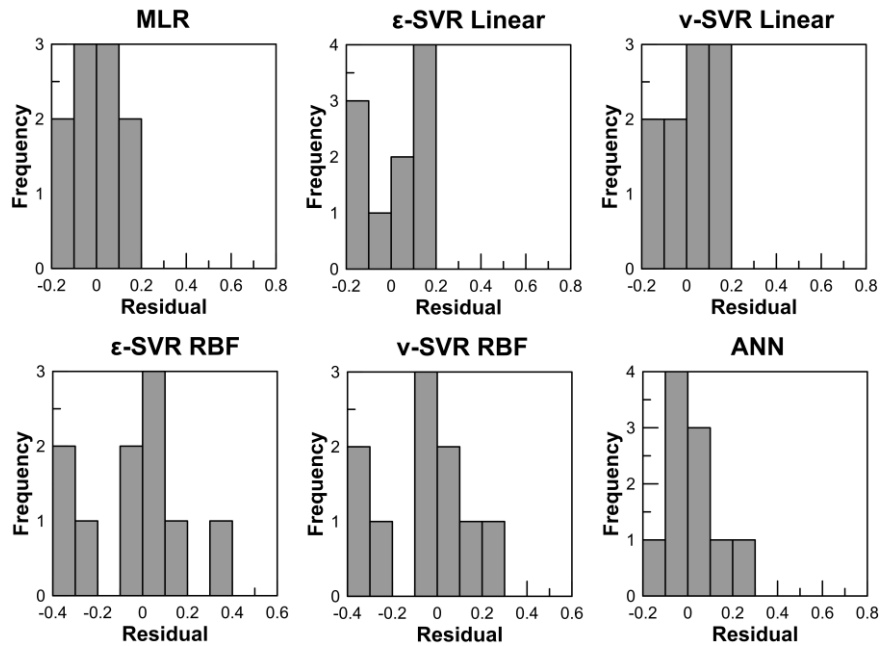


Figure 6.41 : Error histograms for Case 2.2.

Statistical comparison of the results is presented in Figure 6.42.

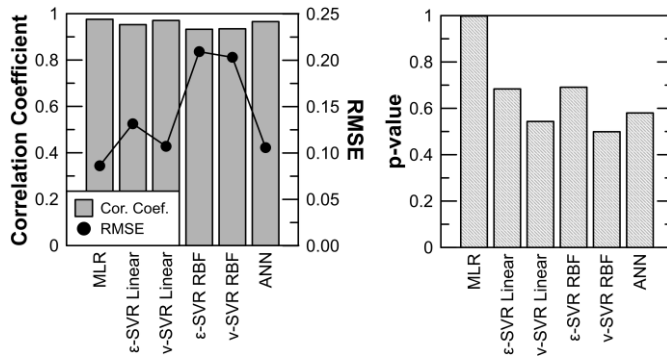


Figure 6.42 : Statistical comparison for Case 2.2.

According to the p-values in Figure 6.42, it is failed to reject the null hypothesis. The best predictive models for Case 2.2 are MLR, SVR with linear kernels, and ANN, which have the highest correlation coefficients (0.97) and the lowest RMSE values. It can be also inferred from Figures 6.39, 6.41, and 6.42 that linear models give better testing results rather than RBF kernels, although RBF kernels give the best training performance.

6.2.2.3 Case 2.3: calculations with data from F-6

For calculations in Case 2.3, a dataset with 919 rows is used. Descriptive statistics of the dataset is presented in Table 6.12. The correlation plot of the dataset is represented in Figure 6.43.

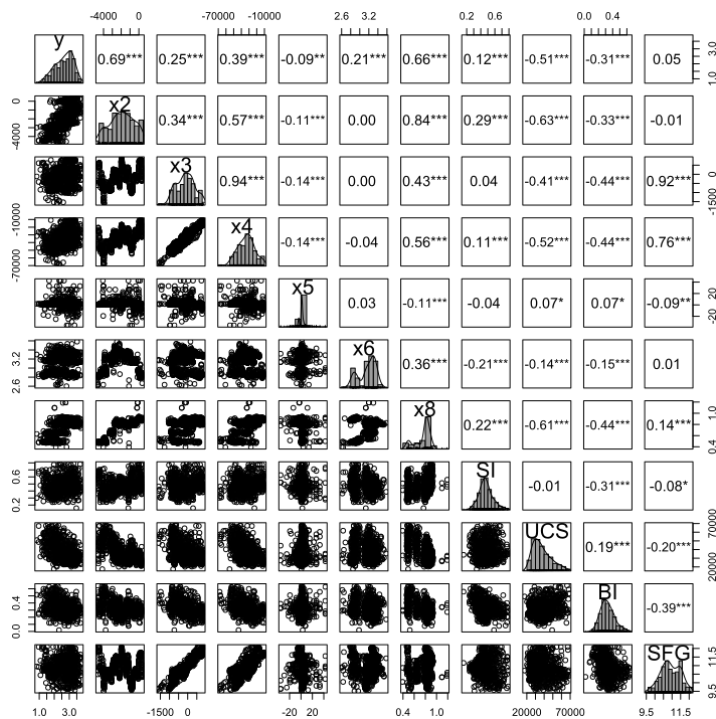


Figure 6.43 : Correlation plot of the dataset for Case 2.3.

Table 6.12 : Descriptive statistics of the dataset for Case 2.3.

	ROP (y)	Compaction (x ₂)	Under- compaction (x ₃)	Overbalance (x ₄)	WOB (x ₅)	RPM (x ₆)	Hydraulics (x ₈)	SI	UCS	BI	SFG
Number	919	919	919	919	919	919	919	919	919	919	919
Min	8.31E-01	-4.66E+03	-1.59E+03	-6.83E+04	-3.57E+01	2.60E+00	3.92E-01	1.55E-01	1.74E+04	1.74E-02	9.51E+00
Max	3.81E+00	3.96E+02	8.56E+02	-1.02E+04	4.26E+01	3.58E+00	1.19E+00	7.90E-01	6.88E+04	6.46E-01	1.21E+01
Range	2.97E+00	5.06E+03	2.45E+03	5.81E+04	7.83E+01	9.83E-01	7.99E-01	6.34E-01	5.14E+04	6.29E-01	2.61E+00
Mean	2.64E+00	-1.94E+03	-1.30E+02	-3.39E+04	1.46E+00	3.16E+00	7.59E-01	4.73E-01	3.74E+04	3.27E-01	1.09E+01
Median	2.74E+00	-1.92E+03	-1.24E+02	-3.32E+04	1.39E+00	3.22E+00	8.13E-01	4.62E-01	3.50E+04	3.17E-01	1.09E+01
Q ₁	2.18E+00	-2.90E+03	-5.20E+02	-4.16E+04	3.04E-01	2.98E+00	6.63E-01	4.10E-01	2.92E+04	2.56E-01	1.05E+01
Q ₃	3.13E+00	-8.93E+02	2.03E+02	-2.66E+04	2.91E+00	3.30E+00	8.36E-01	5.25E-01	4.34E+04	3.91E-01	1.14E+01
Variance	3.63E-01	1.84E+06	2.51E+05	1.22E+08	5.54E+01	3.67E-02	1.85E-02	1.01E-02	1.09E+08	9.95E-03	3.22E-01
Std. Dev.	6.02E-01	1.35E+03	5.01E+02	1.10E+04	7.44E+00	1.92E-01	1.36E-01	1.01E-01	1.05E+04	9.98E-02	5.68E-01
Skewness	-4.58E-01	-1.36E-01	-5.90E-02	-1.24E-01	1.03E+00	-6.12E-01	-9.08E-01	4.76E-01	8.20E-01	4.98E-01	-6.40E-02
Kurtosis	-5.56E-01	-9.01E-01	-5.24E-01	-1.27E-01	9.82E+00	-7.44E-01	3.50E-02	4.90E-01	4.00E-03	2.64E-01	-6.24E-01

It is seen in Figure 6.43 that x_3 , x_4 , and SFG are highly correlated, which are the functions of various drilling parameters such as depth, pore pressure, ECD, and shear failure gradient. Box-Whisker plots for each feature is given in Figure 6.44. The circles in Figure 6.35 indicate the outliers in the data.

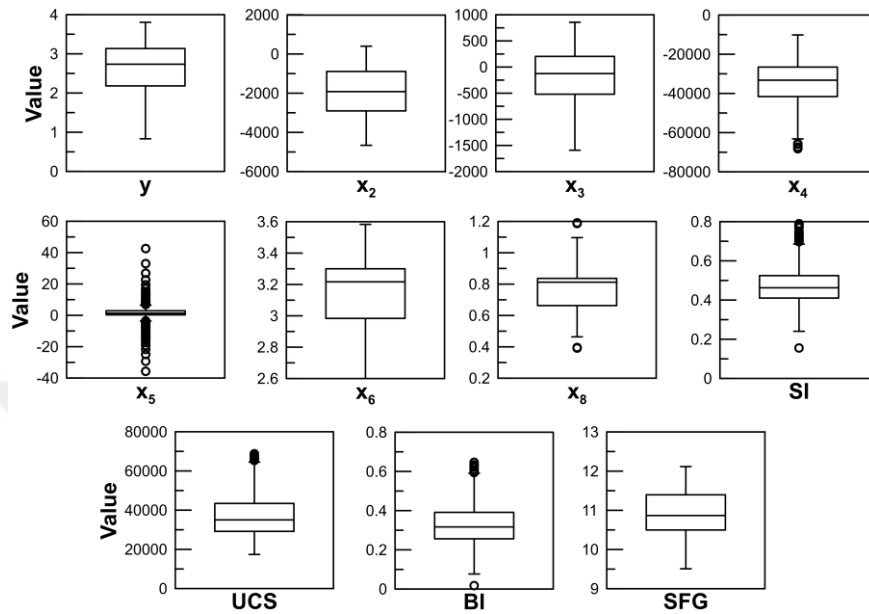


Figure 6.44 : Boxplots of the dataset for Case 2.3.

The data is randomly split into training and testing subsets with the ratio of 75% and 25%, respectively. The results and graphical presentations of the intermediate steps for Case 2.3 is given in Appendix D. The results of the ANN sensitivity study, ANN hidden-layer settings, optimum SVR hyper-parameters and grid search plots, scatterplots for the model training and training errors, normalized ROP predictions with testing subset for each model, and error histograms can be found in the corresponding section of the Appendix D. Cumulative statistical comparison of the results for Case 2.3 is presented in Figure 6.45.

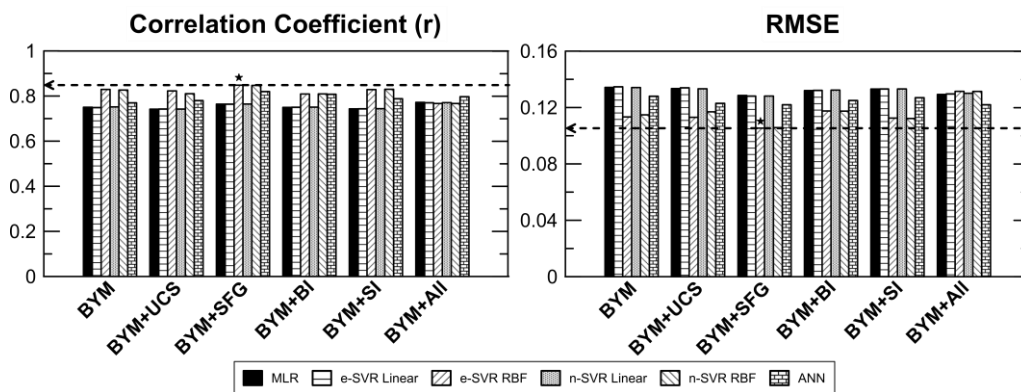


Figure 6.45 : Cumulative statistical comparison for Case 2.3.

As seen in Figure 6.45, prediction accuracies in Case 2.3.2, Case 2.3.4, and Case 2.3.6 are not better than the default BYM case (Case 2.3.1). On the other hand, adding the SFG parameter to default BYM (Case #2.3.3) and predicting with ε -SVR RBF gives the best result in terms of correlation coefficient (0.848). Additionally, in Figure 6.43, SFG is strongly correlated with Functions x_3 and x_4 . Hence, to find a possibility of increasing the prediction accuracy, a feature selection study is performed. Figure 6.43 clearly shows that the Function x_5 is not correlated to any other variable in the master dataset. Additionally, the Function x_5 has a narrow data range when normalized, meaning that WOB can be effective in ROP prediction when the parameter range is wider. However, in this study, the effect of WOB in terms of ROP prediction with BYM is negligible. As a result of the feature selection study, extracting the Function x_5 from the SFG model increases prediction accuracy. Training errors for feature selection is presented in Figure 6.46.

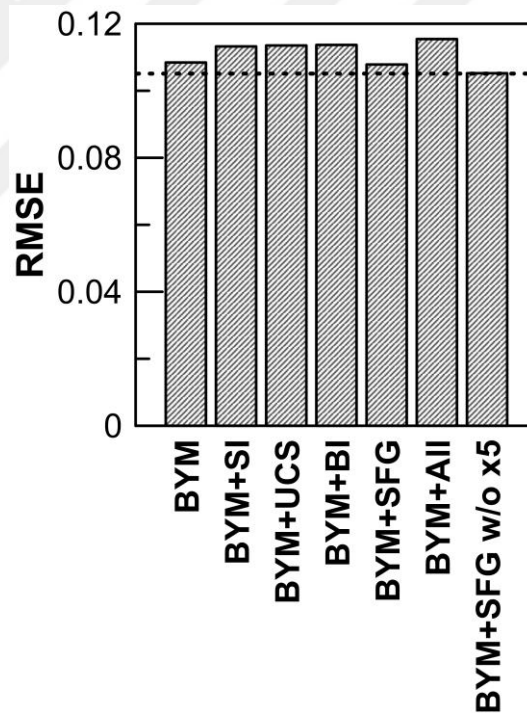


Figure 6.46 : Training errors of different models with ε -SVR RBF for Case 2.3.

As seen from Figure 6.46, excluding the Function x_5 from the model increases training accuracy. Another case (Case 2.3.7) is created by adding SFG and excluding the Function x_5 from the default BYM to determine testing accuracy. Performance comparison charts are given in Figure 6.47 in terms of correlation coefficient and RMSE.

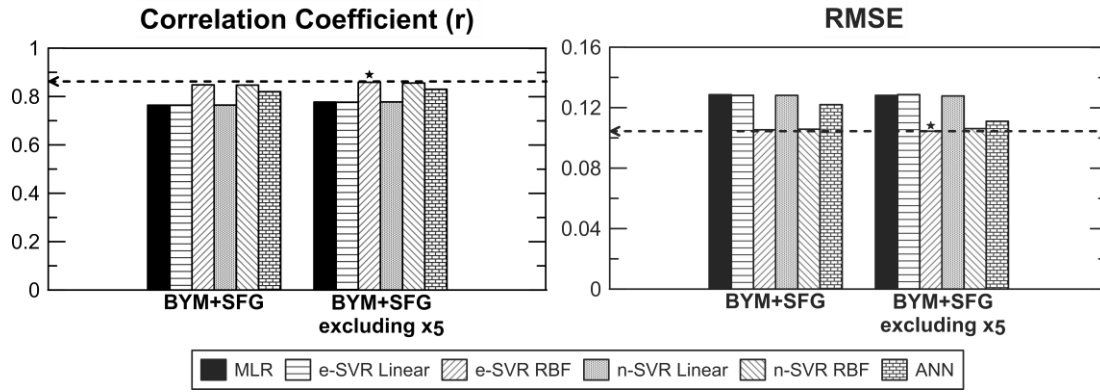


Figure 6.47 : Statistical performance comparison of Function x_5 for Case 2.3.

Error box-plots are presented in Figure 6.48.

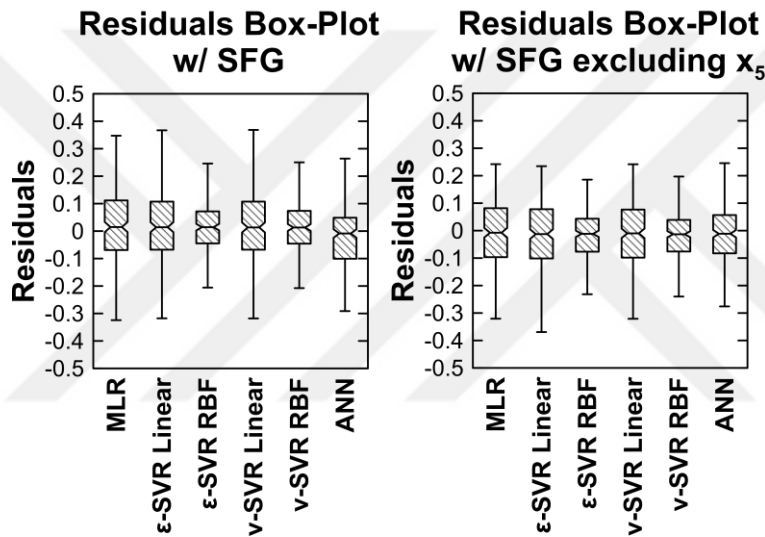


Figure 6.48 : Residual box-plots for SFG model with and without Function x_5 for Case 2.3.

Linear models including MLR and SVR do not give satisfactory results in terms of high RMSE values, except the case of BYM and all other geomechanical features (Case 2.3.6). ANN gives better results rather than linear models as seen in Figures 6.45, 6.47, and 6.48. For the most of the cases (Cases 2.3.1-2.3.5 and Case 2.3.7), RBF kernels give the best results based on RMSE and correlation coefficient. This outcome can also be inferred from Figure 6.48, which indicates that RBF kernels have narrow error (or residuals) distribution among other predictive models. The best result is in the case with ϵ -SVR RBF, which includes BYM with SFG and excluding Function x_5 (Case 2.3.7) with a higher correlation coefficient (0.859) than that of including Function x_5 (Case 2.3.3).

6.2.3 Case 3: data taken from the well C-2

The calculations performed in Case 3.1 and Case 3.2 (F-4 and F-5 formations) belong to the homogeneous formations. The effect of several geomechanical features with different predictive models on ROP prediction is investigated by generating individual cases for the drilling data taken from the heterogeneous F-6 formation in Well C-2.

6.2.3.1 Case 3.1: calculations with data from F-4

For calculations in Case 3.1, a dataset with 218 rows is used. Descriptive statistics of the dataset is presented in Table 6.13.

The correlation plot of the dataset is represented in Figure 6.49. In Figure 6.49, scatter plots for each pair of variables are presented on the lower left, correlation coefficients and significances based on p-values are shown on the upper right as stars, histograms and density plots are placed in the middle.

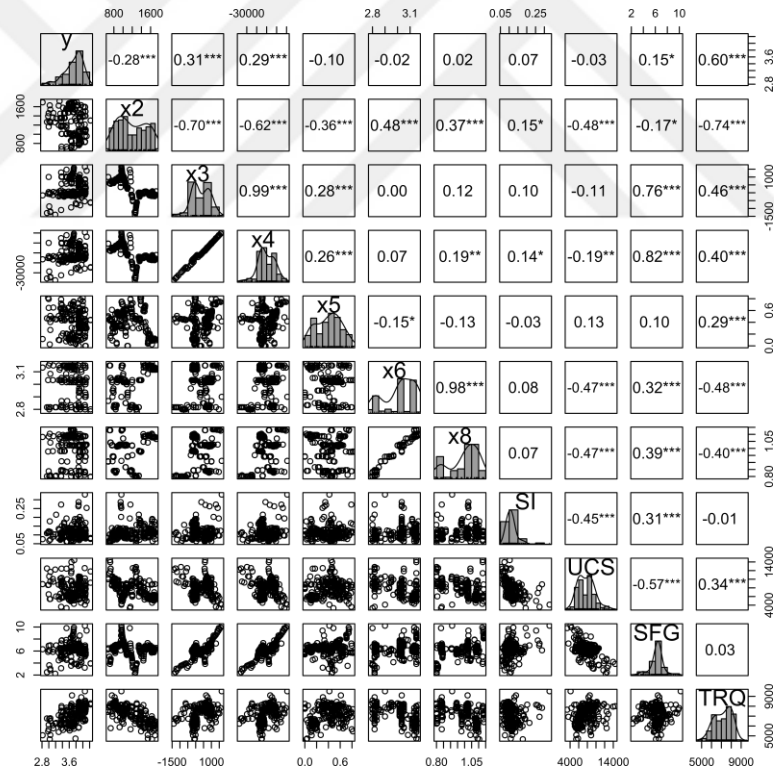


Figure 6.49 : Correlation plot of the dataset for Case 3.1.

It is seen in Figure 6.49 that x_3 , x_4 , and SFG are highly correlated, which are the functions of depth, pore pressure gradient, ECD, and shear failure gradient. Additionally, x_6 and x_8 are also high-correlated, which are the functions of RPM and jet impact force.

Table 6.13 : Descriptive statistics of the dataset for Case 3.1.

	ROP (y)	Compaction (x ₂)	Under- compaction (x ₃)	Overbalance (x ₄)	WOB (x ₅)	RPM (x ₆)	Hydraulics (x ₈)	SI	UCS	SFG	TRQ
Number	218	218	218	218	218	218	218	218	218	218	218
Min	2.82E+00	6.75E+02	-1.40E+03	-3.75E+04	0.00E+00	2.78E+00	7.96E-01	5.65E-02	3.31E+03	2.22E+00	4.66E+03
Max	4.23E+00	1.72E+03	1.66E+03	1.07E+04	8.15E-01	3.17E+00	1.14E+00	3.30E-01	1.44E+04	1.02E+01	9.52E+03
Range	1.41E+00	1.05E+03	3.06E+03	4.82E+04	8.15E-01	3.83E-01	3.42E-01	2.73E-01	1.11E+04	8.00E+00	4.86E+03
Mean	3.73E+00	1.19E+03	3.22E+02	-1.04E+04	3.99E-01	3.02E+00	9.97E-01	1.16E-01	7.76E+03	6.27E+00	7.16E+03
Median	3.80E+00	1.16E+03	2.81E+02	-1.21E+04	4.35E-01	3.03E+00	1.02E+00	1.10E-01	7.74E+03	6.35E+00	7.26E+03
Q ₁	3.59E+00	9.21E+02	-9.69E+01	-1.56E+04	2.10E-01	2.93E+00	9.41E-01	9.20E-02	6.20E+03	5.80E+00	6.40E+03
Q ₃	3.90E+00	1.48E+03	8.04E+02	-3.83E+03	5.61E-01	3.13E+00	1.08E+00	1.27E-01	8.87E+03	6.84E+00	7.90E+03
Variance	7.67E-02	9.43E+04	3.36E+05	7.23E+07	4.37E-02	1.45E-02	1.01E-02	1.56E-03	3.74E+06	1.55E+00	7.82E+05
Std. Dev.	2.77E-01	3.07E+02	5.79E+02	8.50E+03	2.09E-01	1.20E-01	1.01E-01	3.95E-02	1.93E+03	1.24E+00	8.84E+02
Skewness	-9.54E-01	1.50E-01	-2.05E-01	-2.90E-01	-9.30E-02	-6.09E-01	-7.17E-01	2.28E+00	6.21E-01	-2.31E-01	-3.04E-01
Kurtosis	6.09E-01	-1.33E+00	-1.36E-01	5.30E-01	-9.70E-01	-9.54E-01	-7.50E-01	7.51E+00	5.24E-01	2.00E+00	-5.05E-01

Box-Whisker plots for each feature is given in Figure 6.50. The circles in Figure 6.50 indicate the outliers in the data.

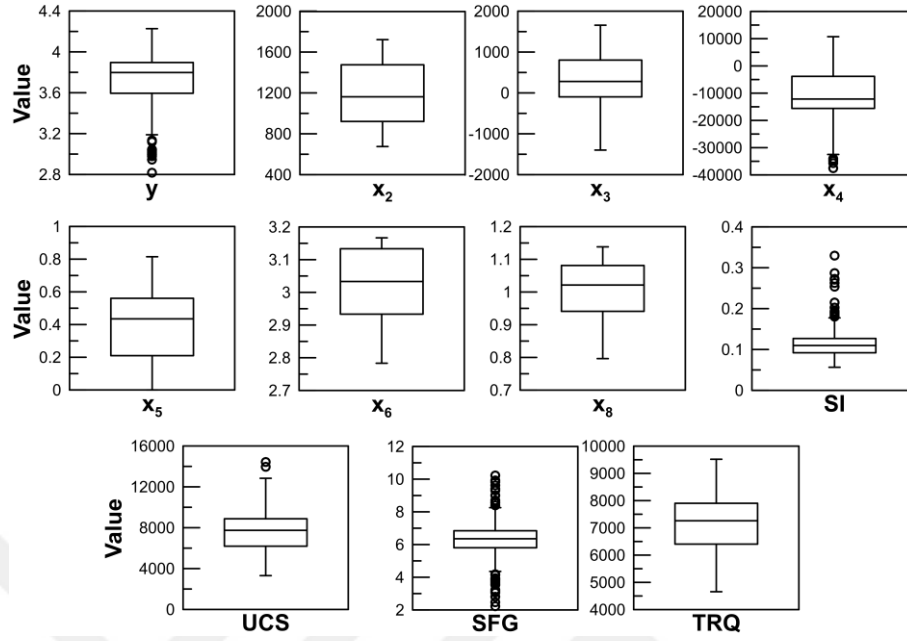


Figure 6.50 : Boxplots of the dataset for Case 3.1.

The data is randomly split into training and testing subsets with the ratio of 75% and 25%, respectively. The results and graphical presentations of the intermediate steps for Case 3.1 is given in Appendix D. The results of the ANN sensitivity study, ANN hidden-layer settings, optimum SVR hyper-parameters and grid search plots, scatterplots for the model training and training errors, normalized ROP predictions with testing subset for each model, and error histograms can be found in the corresponding section of the Appendix D.

Cumulative statistical comparison of the results for Case 3.1 is presented in Figure 6.51.

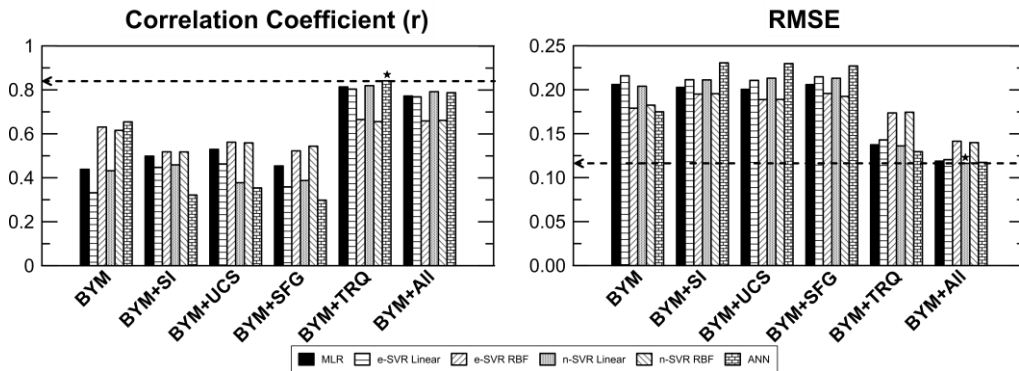


Figure 6.51 : Cumulative statistical comparison for Case 3.1.

As seen in Figure 6.51, adding torque parameter to default BYM (Case 3.1.5) and using ANN give the best result in terms of the correlation coefficient (0.819). On the other hand, adding all additional parameters to default BYM (Case 3.1.6) and using ν -SVR with RBF kernel yield the minimum RMSE value (0.1162). Hence, feature selection is performed via RF algorithm to find a possibility of increasing the prediction accuracy. Feature importance are presented in Figure 6.52.

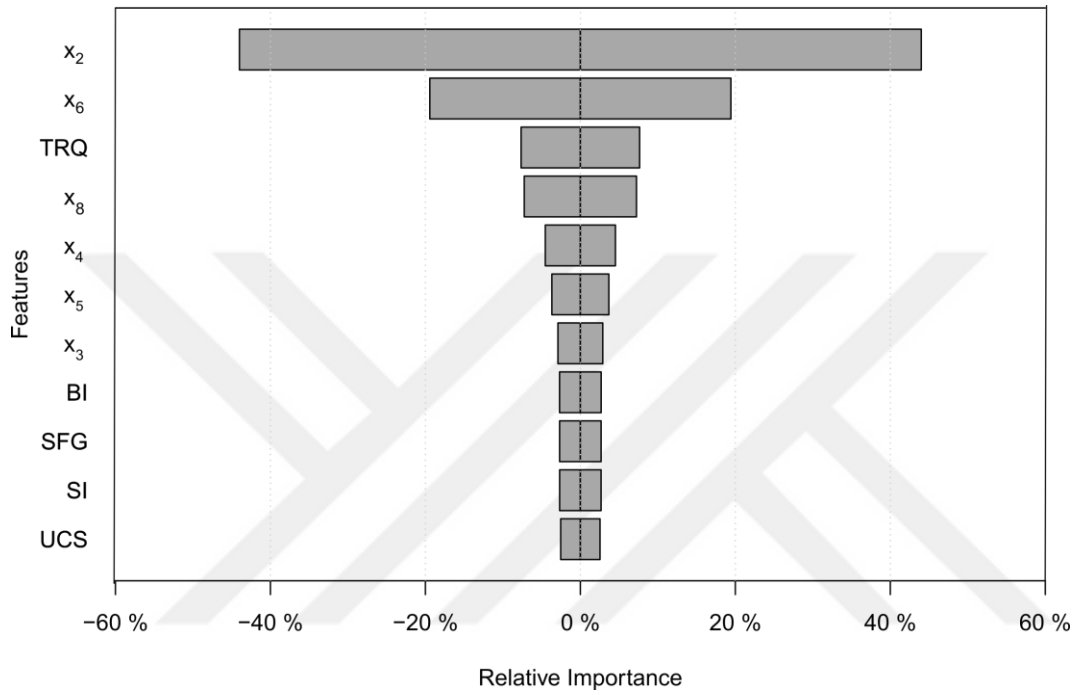


Figure 6.52 : Relative feature importance for Case 3.1.

As a performance indicator, RMSE values of the sensitivity analysis for the feature selection algorithm regarding relative feature importance showed in Figure 6.52 is presented in Table 6.14.

Table 6.14 : RMSE values of the feature selection for Case 3.1.

	MLR	ϵ -SVR Linear	ϵ -SVR RBF	ν -SVR Linear	ν -SVR RBF	ANN
TRQ	0.1524	0.1479	0.1546	0.1531	0.1588	0.1546
+ x ₂	0.1740	0.1757	0.1633	0.1751	0.1618	0.1591
+ x ₅	0.1230	0.1233	0.1066	0.1230	0.1083	0.1246
+ x ₄	0.1420	0.1420	0.1509	0.1431	0.1504	0.1472
+ x ₈	0.1515	0.1532	0.1335	0.1510	0.1498	0.1331
+ x ₃	0.1420	0.1481	0.1625	0.1405	0.1665	0.1305
+ UCS	0.1364	0.1457	0.1662	0.1356	0.1662	0.1317
+ SFG	0.1471	0.1523	0.1399	0.1539	0.1385	0.1233
+ SI	0.1400	0.1379	0.1671	0.1380	0.1668	0.1503
+ x ₆	0.1188	0.1205	0.1414	0.1162	0.1399	0.1426

As seen from Table 6.14, the minimum error (0.1066) is found when torque, x_2 , and x_5 parameters are used in the regression. Another case (Case 3.1.7) is created by using torque, x_2 , and x_5 parameters as features to determine testing accuracy. Performance comparison charts are given in Figure 6.53 in terms of correlation coefficient and RMSE.

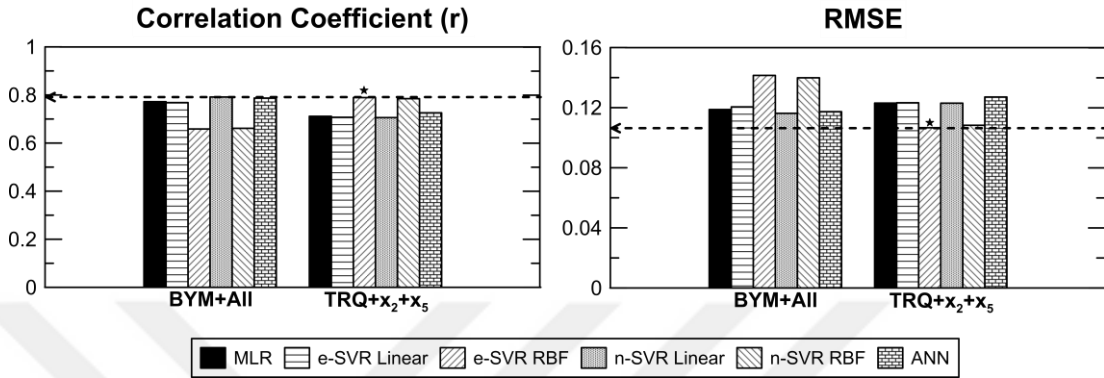


Figure 6.53 : Statistical performance comparison of feature selection for Case 3.1.

Error box-plots are presented in Figure 6.54.

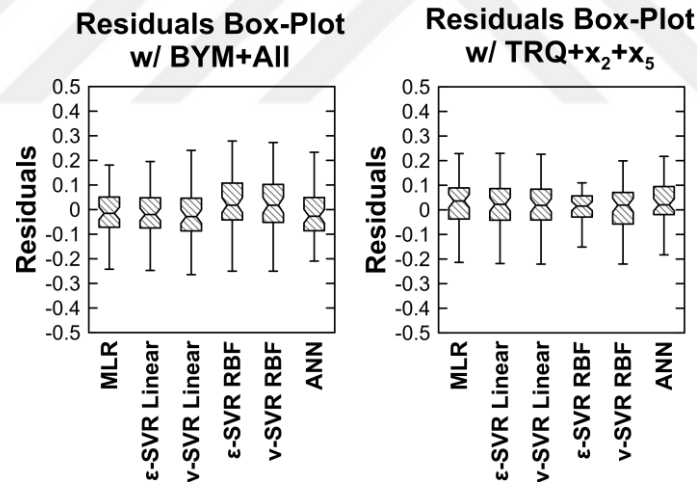


Figure 6.54 : Residual box-plots of feature selection for Case 3.1.

As seen in Figure 6.51, linear models, including MLR and SVR, do not give satisfactory results for default BYM (Case 3.1.1) and the cases that additional features used individually, except torque parameter (Case 3.1.5). By using torque, ANN gives a better correlation coefficient when compared to other cases and models. However, as seen in Figure 6.53, the minimum error is in Case 3.1.7 with ϵ -SVR RBF, including torque, x_2 , and x_5 functions, which can also be seen from the error distribution in Figure 6.54.

6.2.3.2 Case 3.2: calculations with data from F-5

For calculations in Case 3.2, a dataset with 31 rows is used. Descriptive statistics of the dataset is presented in Table 6.15.

The correlation plot of the dataset is represented in Figure 6.55. In Figure 6.55, scatter plots for each pair of variables are presented on the lower left, correlation coefficients and significances based on p-values are shown on the upper right as stars, histograms and density plots are placed in the middle.



Figure 6.55 : Correlation plot of the dataset for Case 3.2.

It is seen in Figure 6.55 that y , x_2 , x_3 , x_4 , SI, and SFG are highly correlated, which are the functions of depth, pore pressure gradient, ECD, shale index and shear failure gradient. Additionally, x_6 and x_8 are also high-correlated, which are the functions of RPM and jet impact force.

Box-Whisker plots for each feature is given in Figure 6.56.

Table 6.15 : Descriptive statistics of the dataset for Case 3.2.

	ROP (y)	Compaction (x ₂)	Under- compaction (x ₃)	Overbalance (x ₄)	WOB (x ₅)	RPM (x ₆)	Hydraulics (x ₈)	SI	UCS	SFG	TRQ
Number	31	31	31	31	31	31	31	31	31	31	31
Min	9.65E-01	1.53E+02	8.76E+01	-1.71E+04	1.50E-01	2.83E+00	8.66E-01	1.02E-02	1.33E+04	9.84E-01	3.97E+03
Max	2.65E+00	2.98E+02	1.19E+03	1.69E+03	6.50E-01	3.10E+00	9.77E-01	2.18E-01	3.05E+04	4.50E+00	8.58E+03
Range	1.68E+00	1.44E+02	1.10E+03	1.88E+04	5.00E-01	2.67E-01	1.10E-01	2.08E-01	1.72E+04	3.52E+00	4.61E+03
Mean	1.82E+00	2.19E+02	5.37E+02	-9.42E+03	4.25E-01	2.98E+00	9.00E-01	1.27E-01	2.16E+04	2.95E+00	6.08E+03
Median	1.78E+00	2.09E+02	4.54E+02	-1.09E+04	4.04E-01	2.95E+00	8.88E-01	1.35E-01	2.27E+04	3.38E+00	6.12E+03
Q ₁	1.30E+00	1.81E+02	2.24E+02	-1.47E+04	3.35E-01	2.93E+00	8.79E-01	8.97E-02	1.74E+04	1.34E+00	5.24E+03
Q ₃	2.25E+00	2.74E+02	8.30E+02	-4.45E+03	5.35E-01	3.03E+00	9.23E-01	1.65E-01	2.51E+04	3.83E+00	6.73E+03
Variance	2.74E-01	2.28E+03	1.27E+05	3.70E+07	1.43E-02	4.40E-03	8.54E-04	2.72E-03	2.14E+07	1.46E+00	1.19E+06
Std. Dev.	5.23E-01	4.78E+01	3.57E+02	6.08E+03	1.19E-01	6.64E-02	2.92E-02	5.21E-02	4.62E+03	1.21E+00	1.09E+03
Skewness	8.40E-02	3.99E-01	4.01E-01	4.08E-01	-1.12E-01	4.67E-01	1.04E+00	-2.37E-01	-1.73E-01	-6.30E-01	2.42E-01
Kurtosis	-1.16E+00	-1.25E+00	-1.19E+00	-1.18E+00	-5.98E-01	-3.62E-01	3.66E-01	-4.31E-01	-9.54E-01	-1.22E+00	-1.10E-01

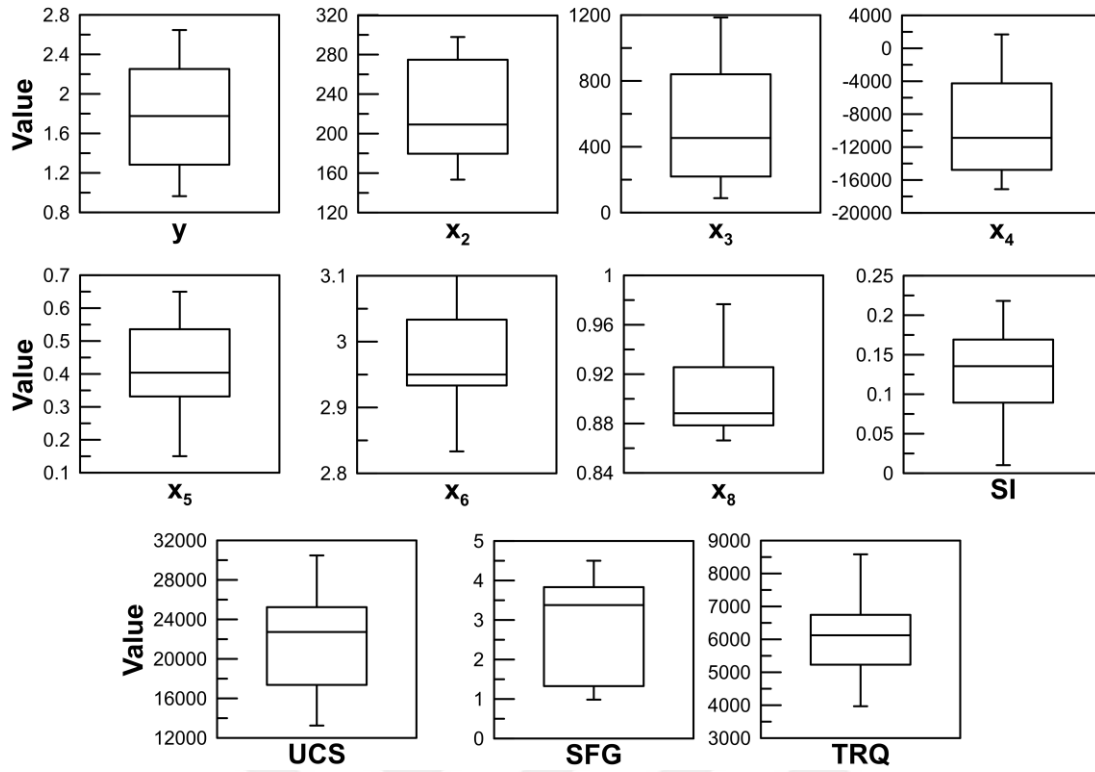


Figure 6.56 : Boxplots of the dataset for Case 3.2.

The data is randomly split into training and testing subsets with the ratio of 75% and 25%, respectively. The results and graphical presentations of the intermediate steps for Case 3.2 is given in Appendix D. The results of the ANN sensitivity study, ANN hidden-layer settings, optimum SVR hyper-parameters and grid search plots, scatterplots for the model training and training errors, normalized ROP predictions with testing subset for each model, and error histograms can be found in the corresponding section of the Appendix D.

Cumulative statistical comparison of the results for Case 3.2 is presented in Figure 6.57.

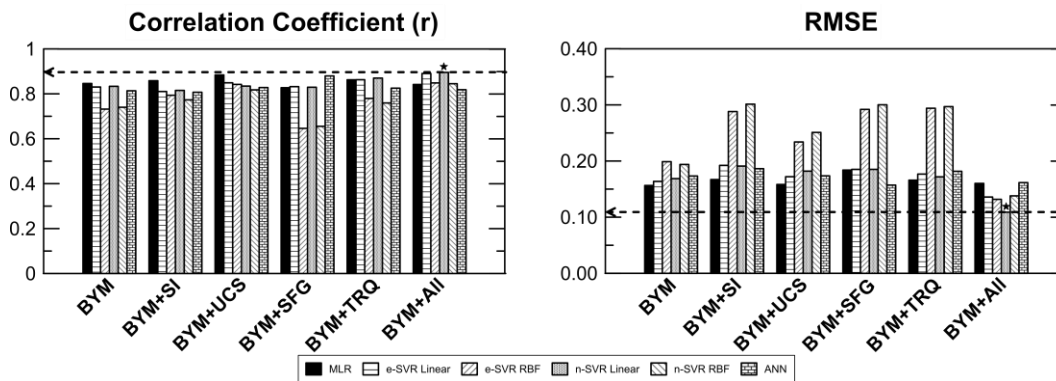


Figure 6.57 : Cumulative statistical comparison for Case 3.2.

As seen in Figure 6.57, adding all additional parameters to default BYM (Case 3.2.6) and using v-SVR with linear kernel give the best result in terms of the maximum correlation coefficient (0.895) and the minimum RMSE (0.1089). Feature selection is performed via RF algorithm to find a possibility of increasing the prediction accuracy. Feature importance are presented in Figure 6.58.

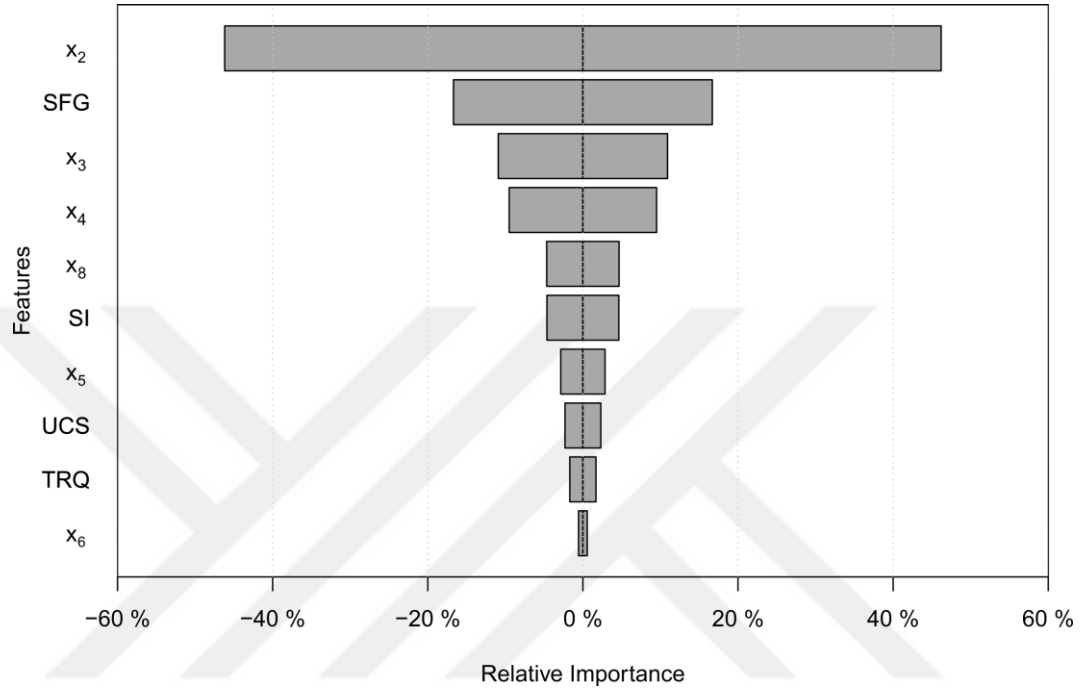


Figure 6.58 : Relative feature importance for Case 3.2.

As a performance indicator, RMSE values of the sensitivity analysis for the feature selection algorithm regarding relative feature importance showed in Figure 6.58 is presented in Table 6.16.

Table 6.16 : RMSE values of the feature selection for Case 3.2.

	MLR	ϵ -SVR Linear	ϵ -SVR RBF	v-SVR Linear	v-SVR RBF	ANN
x_2	0.1759	0.1680	0.1599	0.1835	0.1589	0.1766
+ SFG	0.1533	0.1658	0.1860	0.1548	0.1847	0.1495
+ x_3	0.1469	0.1515	0.1777	0.1522	0.1876	0.1367
+ x_4	0.1416	0.1455	0.1979	0.1394	0.2001	0.1254
+ x_8	0.1261	0.1645	0.1536	0.1517	0.1770	0.1306
+ SI	0.2001	0.1958	0.2083	0.1916	0.2072	0.1580
+ x_5	0.1766	0.1838	0.2603	0.1802	0.2745	0.1614
+ UCS	0.1963	0.1233	0.2237	0.1185	0.2243	0.1732
+ TRQ	0.0813	0.1054	0.2532	0.1079	0.2544	0.1397
+ x_6	0.1601	0.1359	0.1318	0.1089	0.1380	0.1637

As seen from Table 6.16, the minimum error (0.0813) is found when all the parameters except x_6 are used in the regression. Another case (Case 3.2.7) is created by using all the parameters except x_6 as features to determine testing accuracy. Performance comparison charts are given in Figure 6.59 in terms of correlation coefficient and RMSE.

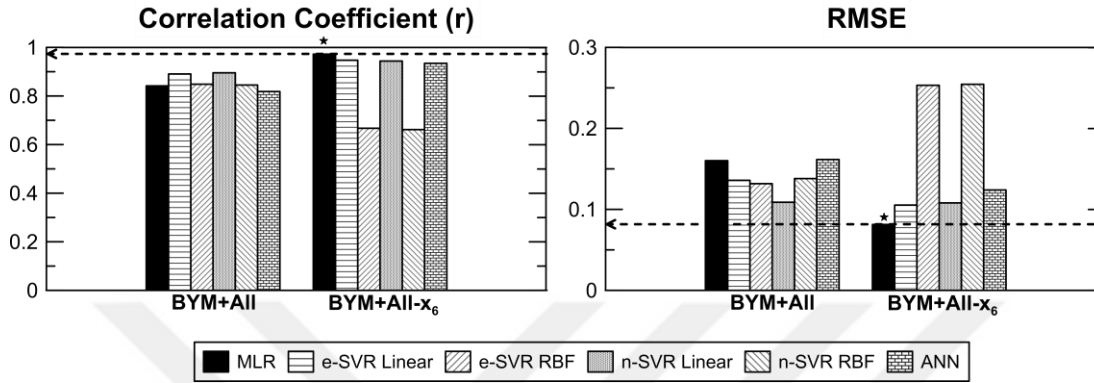


Figure 6.59 : Statistical performance comparison of feature selection for Case 3.2.

Error box-plots are presented in Figure 6.60.

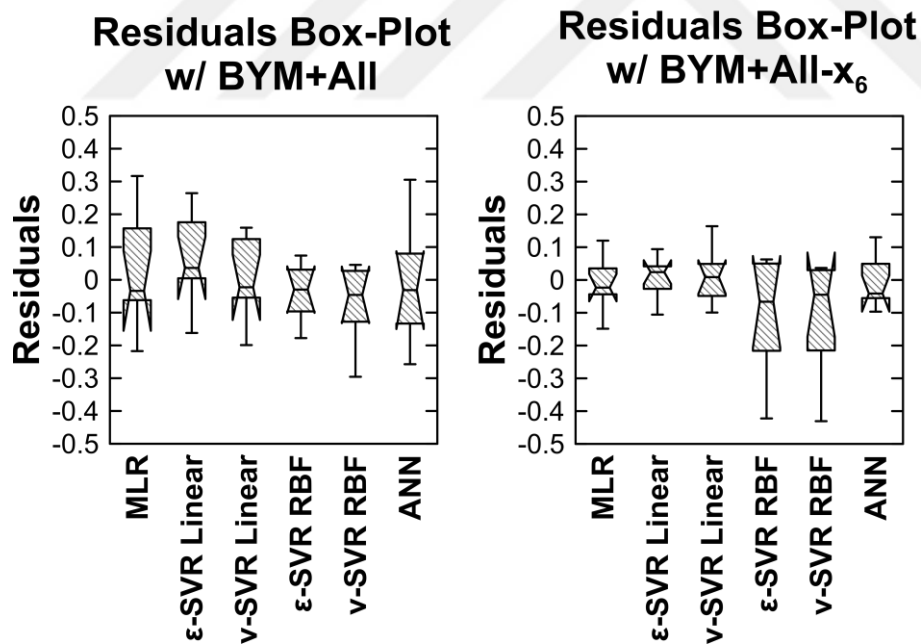


Figure 6.60 : Residual box-plots of feature selection for Case 3.2.

As seen in Figure 6.57, SVR with RBF kernels do not give satisfactory results for entire cases. Furthermore, as seen in Figure 6.59, the maximum correlation coefficient and the minimum error is in Case 3.2.7 with MLR, including all the parameters except Function x_6 , which can also be seen from the error distribution in Figure 6.60.

6.2.3.3 Case 3.3: calculations with data from F-6

For calculations in Case 3.3, a dataset with 967 rows is used. Descriptive statistics of the dataset is presented in Table 6.17.

The correlation plot of the dataset is represented in Figure 6.61. In Figure 6.61, scatter plots for each pair of variables are presented on the lower left, correlation coefficients and significances based on p-values are shown on the upper right as stars, histograms and density plots are placed in the middle.

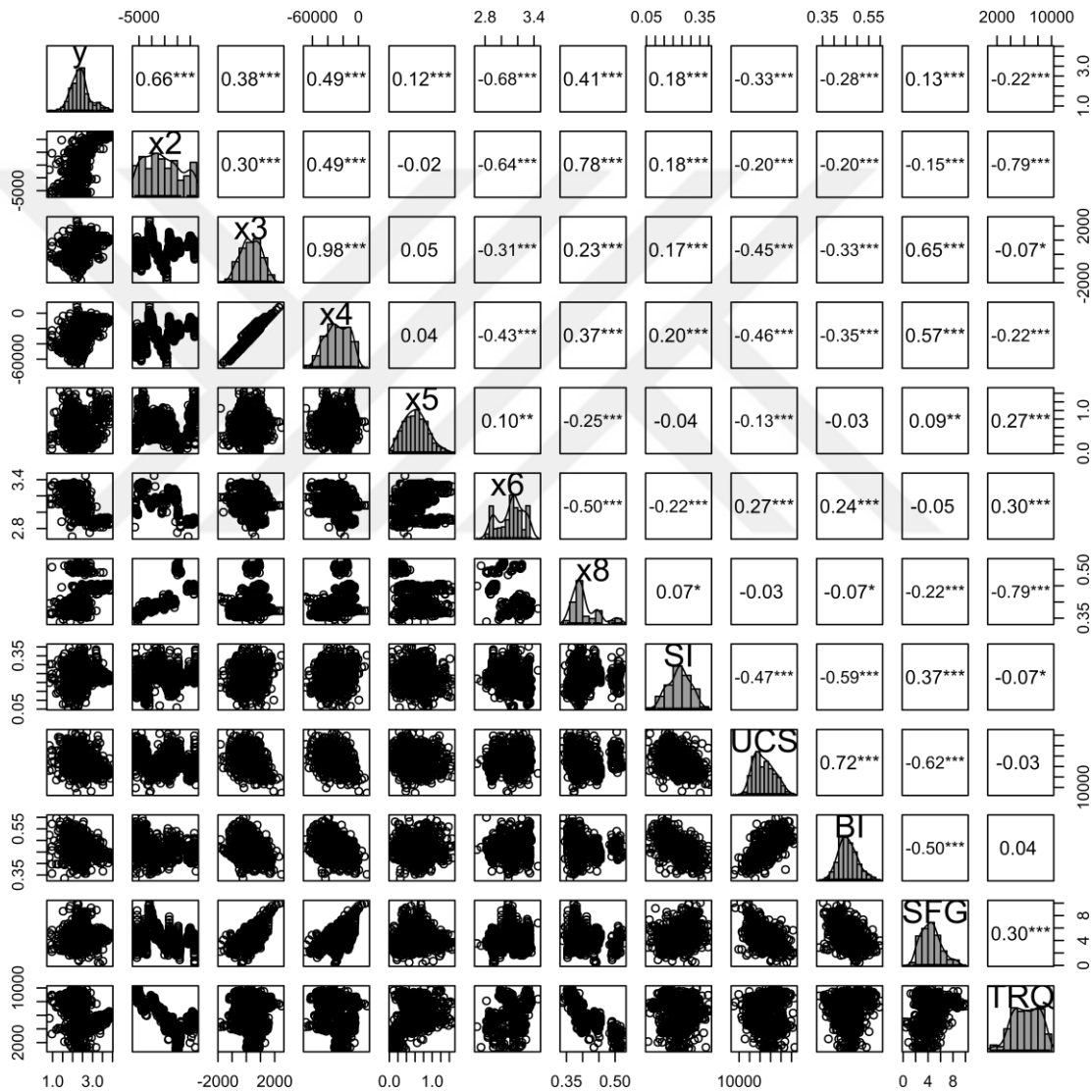


Figure 6.61 : Correlation plot of the dataset for Case 3.3.

It is seen in Figure 6.61 that x_3 and x_4 are highly correlated, which are the functions of depth, pore pressure gradient, and ECD. Additionally, x_2 , x_8 , and TRQ are also high-correlated, which are the functions of depth, jet impact force, and torque. UCS and BI can be named as high-correlated.

Table 6.17 : Descriptive statistics of the dataset for Case 3.3.

	ROP (y)	Compaction (x ₂)	Under- compaction (x ₃)	Overbalance (x ₄)	WOB (x ₅)	RPM (x ₆)	Hydraulics (x ₈)	SI	UCS	BI	SFG	TRQ
Number	967	967	967	967	967	967	967	967	967	967	967	967
Min	8.31E-01	-5.33E+03	-1.83E+03	-6.85E+04	4.93E-02	2.70E+00	3.38E-01	5.90E-02	7.36E+03	3.35E-01	2.33E-01	1.00E+03
Max	3.90E+00	-5.55E+02	2.48E+03	1.11E+04	1.47E+00	3.45E+00	5.26E-01	4.02E-01	3.65E+04	6.00E-01	1.00E+01	9.95E+03
Range	3.06E+00	4.77E+03	4.31E+03	7.96E+04	1.43E+00	7.50E-01	1.89E-01	3.43E-01	2.92E+04	2.65E-01	9.82E+00	8.95E+03
Mean	2.40E+00	-3.15E+03	3.26E+02	-2.77E+04	6.20E-01	3.12E+00	4.07E-01	2.35E-01	2.19E+04	4.60E-01	4.50E+00	6.23E+03
Median	2.35E+00	-3.30E+03	3.18E+02	-2.81E+04	6.12E-01	3.15E+00	3.94E-01	2.38E-01	2.13E+04	4.57E-01	4.40E+00	6.36E+03
Q ₁	2.06E+00	-4.40E+03	-1.73E+02	-3.90E+04	4.08E-01	3.00E+00	3.81E-01	1.90E-01	1.80E+04	4.30E-01	3.29E+00	4.71E+03
Q ₃	2.62E+00	-2.08E+03	8.85E+02	-1.42E+04	8.08E-01	3.23E+00	4.35E-01	2.80E-01	2.54E+04	4.87E-01	5.45E+00	7.89E+03
Variance	2.66E-01	1.98E+06	5.14E+05	2.24E+08	7.87E-02	2.23E-02	1.66E-03	3.98E-03	2.49E+07	1.81E-03	2.82E+00	3.87E+06
Std. Dev.	5.16E-01	1.41E+03	7.17E+02	1.50E+04	2.81E-01	1.49E-01	4.07E-02	6.31E-02	4.99E+03	4.25E-02	1.68E+00	1.97E+03
Skewness	4.62E-01	2.47E-01	-6.80E-02	-1.74E-01	3.56E-01	-3.25E-01	1.19E+00	-7.70E-02	3.75E-01	3.78E-01	5.64E-01	-2.90E-01
Kurtosis	6.70E-01	-1.07E+00	-4.72E-01	-8.33E-01	-2.04E-01	-9.12E-01	6.92E-01	-3.91E-01	-4.36E-01	1.64E-01	1.33E-01	-6.07E-01

Box-Whisker plots for each feature is given in Figure 6.62.

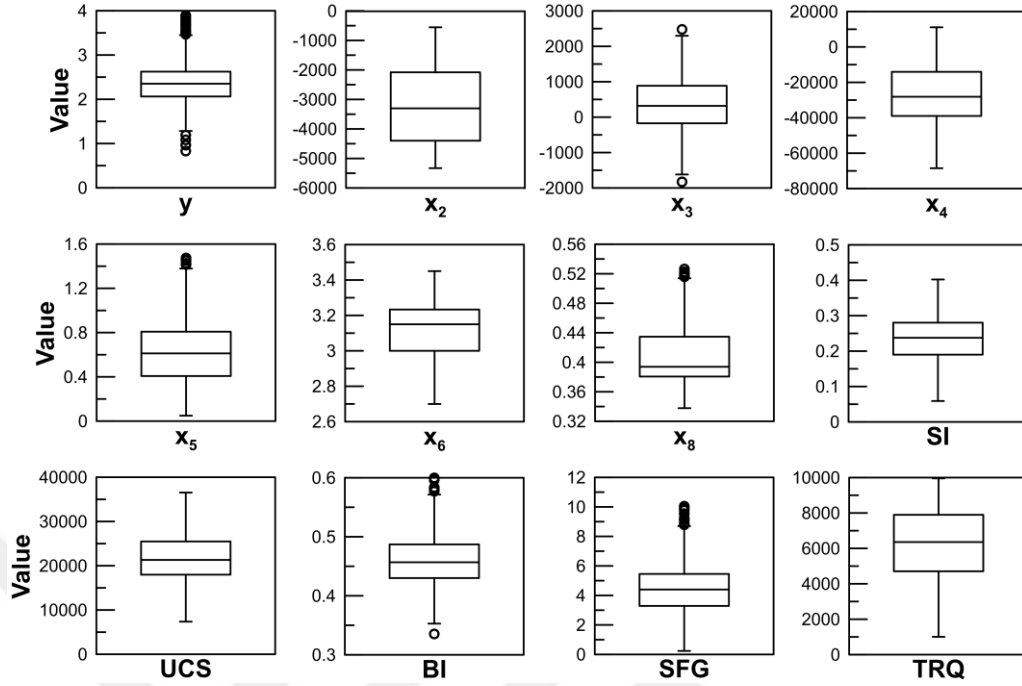


Figure 6.62 : Boxplots of the dataset for Case 3.3.

The data is randomly split into training and testing subsets with the ratio of 75% and 25%, respectively. The results and graphical presentations of the intermediate steps for Case 3.3 is given in Appendix D. The results of the ANN sensitivity study, ANN hidden-layer settings, optimum SVR hyper-parameters and grid search plots, scatterplots for the model training and training errors, normalized ROP predictions with testing subset for each model, and error histograms can be found in the corresponding section of the Appendix D.

Cumulative statistical comparison of the results for Case 3.3 is presented in Figure 6.63.

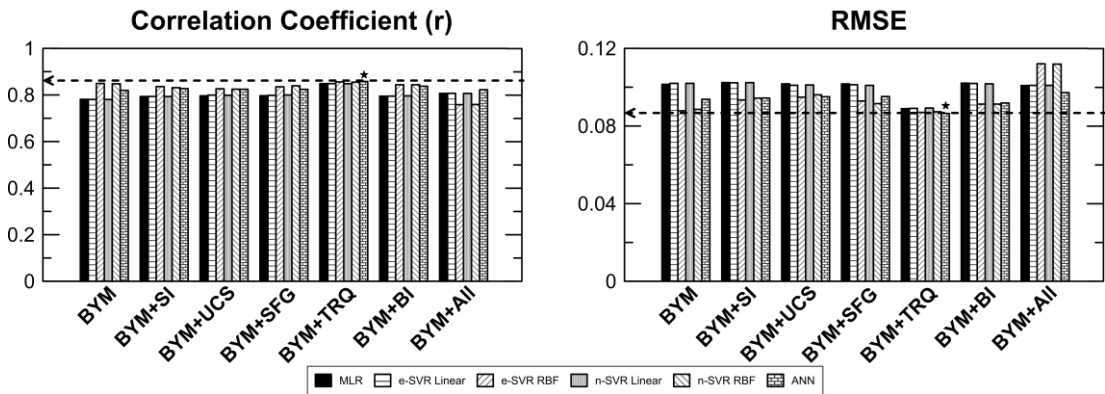


Figure 6.63 : Cumulative statistical comparison for Case 3.3.

As seen in Figure 6.63, adding torque parameter to default BYM (Case 3.3.5) and using ANN give the best result in terms of the maximum correlation coefficient (0.858) and the minimum RMSE (0.0865). Feature selection is performed via RF algorithm to find a possibility of increasing the prediction accuracy. Feature importance are presented in Figure 6.64.

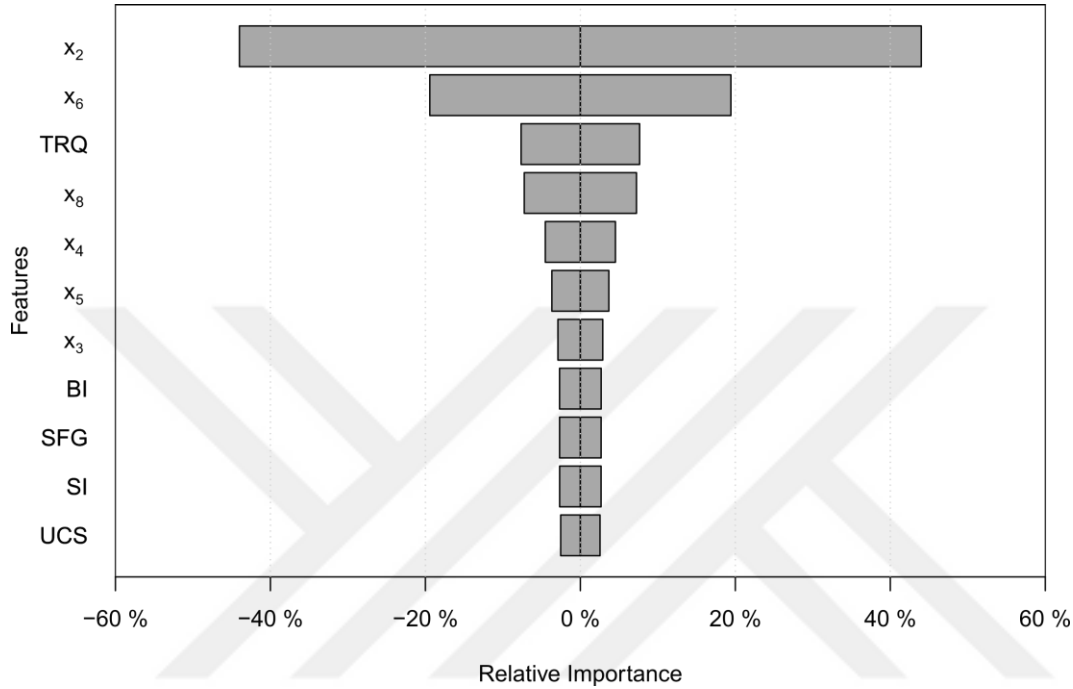


Figure 6.64 : Relative feature importance for Case 3.3.

As a performance indicator, RMSE values of the sensitivity analysis for the feature selection algorithm regarding relative feature importance showed in Figure 6.64 is presented in Table 6.18.

Table 6.18 : RMSE values of the feature selection for Case 3.3.

	MLR	ϵ -SVR Linear	ϵ -SVR RBF	ν -SVR Linear	ν -SVR RBF	ANN
x ₂	0.1294	0.1295	0.0990	0.1294	0.0994	0.1069
+ x ₆	0.1185	0.1186	0.0938	0.1191	0.0938	0.1042
+ TRQ	0.1014	0.1019	0.0949	0.1017	0.0948	0.0993
+ x ₈	0.0974	0.0976	0.0826	0.0977	0.0831	0.0931
+ x ₄	0.0962	0.0961	0.0930	0.0960	0.0895	0.0909
+ x ₅	0.0913	0.0916	0.0798	0.0917	0.0815	0.0862
+ x ₃	0.0889	0.0891	0.0870	0.0893	0.0873	0.0841
+ BI	0.0999	0.1008	0.1029	0.1008	0.1026	0.0974
+ SFG	0.0978	0.0978	0.0964	0.0975	0.0975	0.0936
+ SI	0.0988	0.0980	0.1109	0.0981	0.1108	0.0995

As seen from Table 6.18, the minimum error (0.0798) is found when default BYM except x_6 and torque parameters are used in the regression. Another case (Case 3.3.8) is created by default BYM+TRQ except x_6 as features to determine testing accuracy. Performance comparison charts are given in Figure 6.65 in terms of correlation coefficient and RMSE.

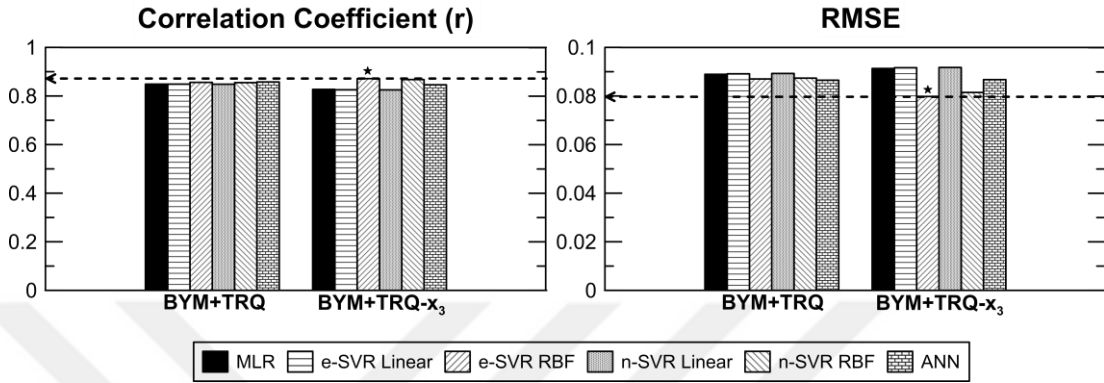


Figure 6.65 : Statistical performance comparison of feature selection for Case 3.3.

Error box-plots are presented in Figure 6.66.

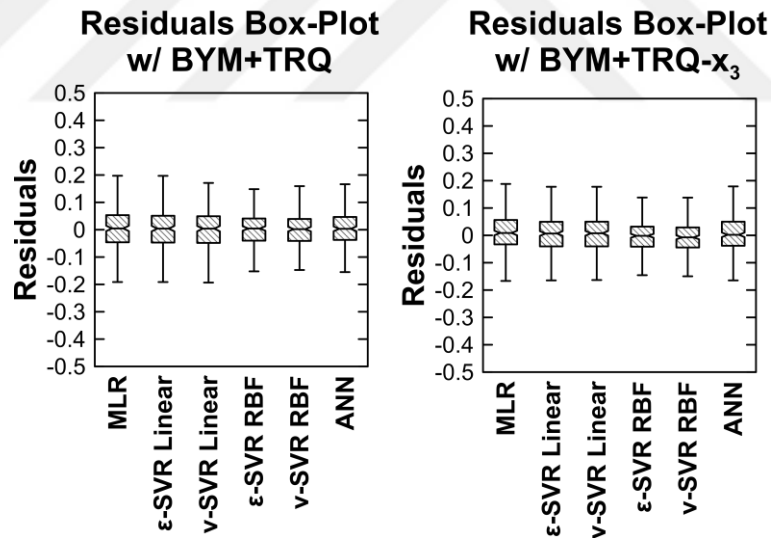


Figure 6.66 : Residual box-plots of feature selection for Case 3.3.

As seen in Figure 6.63, linear models do not give satisfactory results for most of the cases, except Case 3.3.7, which is the case that all additional parameters alongside default BYM are used for regression. Furthermore, as seen in Figure 6.65, the maximum correlation coefficient and the minimum error is in Case 3.3.8 with ϵ -SVR RBF kernel, including all the parameters with torque and except Function x_6 , which can also be seen from the error distribution in Figure 6.66.

6.3 Discussions and General Evaluations

As mentioned previously, the primary focus of this dissertation is to apply a different regression model as an alternative to MLR and to modify BYM by implementing additional geomechanical features to generate a relationship between the heterogeneity and the ROP. The concept is limited to a single type of homogeneous or heterogeneous formation, and the study is performed from a well programming stage perspective. Several predictive models (MLR, SVR, and ANN) are used to build an ROP prediction model. Feature selection is performed with sensitivity analysis and RF algorithm. Prediction accuracies for each predictive model are compared with performance metrics, including RMSE, correlation coefficient, and p-value. The summary of the overall results is presented in Table 6.19.

Table 6.19 : Summary of the overall results.

Case No	Formation Type	No of Input Data	Feature Selection	Best Model	Effective Features
1.1	Homogeneous	30	No	ϵ -SVR RBF	BYM
1.2	Homogeneous	25	No	ϵ -SVR RBF	BYM
1.3	Homogeneous	21	No	ϵ -SVR RBF	BYM
2.1	Homogeneous	57	No	ν -SVR RBF	BYM
2.2	Homogeneous	34	No	MLR	BYM
2.3	Heterogeneous	919	Sensitivity	ϵ -SVR RBF	BYM+SFG- x_5
3.1	Heterogeneous	218	RF	ϵ -SVR RBF	TRQ+ x_2 + x_5
3.2	Homogeneous	31	RF	MLR	BYM+All- x_6
3.3	Heterogeneous	967	RF	ϵ -SVR RBF	BYM+TRQ- x_3

As seen in Table 6.19, it is shown that ϵ -SVR with RBF kernel can predict ROP with high accuracy for the cases, including data taken from the literature (Cases 1.1, 1.2, and 1.3). However, these cases contain a limited number of training data (being less than 30). On the other hand, it is seen that ν -SVR with RBF kernel and MLR give better predictions in remaining homogeneous cases, including data taken from the wells C-1 and C-2 (Cases 2.1, 2.2, and 3.2). Finally, ϵ -SVR with RBF kernel with RF feature selection algorithm is the best predictive model for the heterogeneous cases (Cases 2.3, 3.1, and 3.3).

The results are related to multicollinearity between the features, which can be directly seen in pairwise correlations in the designated correlograms. Moreover, the number of training data also has an essential effect on the prediction accuracies. In homogeneous cases, some overfitting problems are observed in the training phases since the number

of training data is relatively low (less than the square of the number of features). This problem is overcome by fine-tuning the hyper-parameters of SVR and performing cross-validation techniques.

As mentioned in Chapter 2: Literature Review, 51 studies in the literature consider ML applications for ROP optimization and prediction. However, only 4 of these studies include formation heterogeneity (Ahmed et al., 2019; Gan et al., 2019c; Yuswandari et al., 2019; Elkatatny, 2021). Moreover, these studies take the complexity in the formation lithology as drilling through various types of formations. On the other hand, lithology complexity and formation heterogeneity are defined as containing different volumes of lithologies for a unit interval in this dissertation. Additionally, none of these four studies mentioned above is neither compared with nor based on BYM. Furthermore, one particular study (Elkatatny, 2021) has its limitation by constructing the testing subset within the range of the training subset. Moreover, the other studies for heterogeneous formations, which consider PDC and hybrid bit designing, do not include any functional relationships of the drilling parameters defined in BYM.

In line with the explanations above, different regression models are implemented as an alternative to MLR by modifying BYM to predict ROP in heterogeneous formations. This process is carried out by introducing several geomechanical features and surface torque data to the BYM to generate a relationship between the ROP and the heterogeneity, which creates the unique aspect of this dissertation.



7. CONCLUSIONS

The primary purpose of this study is to apply a different type of regression model and modify BYM by implementing geomechanical parameters to construct a relationship between the mechanical rock properties and the heterogeneity. These geomechanical parameters are shale index, uniaxial compressive strength, brittleness index, shear failure gradient, and torque. In this study, several predictive methods are used to build a model for ROP prediction: multiple linear regression (MLR), support vector regression (SVR), and artificial neural networks (ANN).

Several predictive methods are applied to the datasets taken from the literature, which belong to homogeneous formations. Finally, the results are compared from a statistical point of view to determine the best predictor in each case. The findings of these cases are presented in an SCI-e indexed journal paper (Kor and Altun, 2020). It was mentioned that SVR could make predictions with good approximation when the number of training data is limited (Kor and Altun, 2020).

As expressed in Chapter 6, the main goal of the applications in this study is to perform rate of penetration (ROP) predictions in heterogeneous formations. For this purpose, data belonging to the wells drilled in one of the Turkish Petroleum Corporation fields are used. The applicability of predicting ROP in heterogeneous formations by modifying BYM is shown and presented for the first time by the findings of this dissertation, which are also published in an SCI-e indexed journal paper (Kor et al., 2021). The results are compared within statistical aspects, and the main conclusions regarding the particular datasets are given below.

- It is essential to perform quality assurance and quality control (QA/QC) on the dataset before using it in the regression analysis. The noise in the data should be estimated. The data points that generate noise should be excluded from the main dataset.
- Different implementations of SVR are applied to the dataset taken from the studies of other researchers. It is showed that when the number of training data

is insufficient (being less than 30 datasets), the accuracy of MLR is relatively low in predicting ROP.

- For the cases including a limited number of training data (Case 1.2 and 1.3), SVR with RBF kernels generates better predictions.
- Formations F-4 and F-5 of the well C-1 contain homogeneous lithology. Therefore, the cases belong to these formations (Case 2.1 and 2.2) are related to the homogeneous formations. For each case, linear models (MLR and SVR with the linear kernel) give the best prediction performances among other predictive models.
- Formation F-6 is highly heterogeneous. Linear models (MLR and SVR with the linear kernel) do not perform well in ROP prediction by yielding higher RMSE and lower correlation coefficient. On the other hand, SVR with RBF kernels gives better prediction results for most cases (Cases 2.3.1-2.3.5).
- Furthermore, the prediction accuracies of SVR with RBF kernels are higher than MLR, which is the default predictor method of BYM. ANN also performs better than MLR in terms of ROP prediction.
- Additional geomechanical parameters related to heterogeneity have different effects on prediction performances for all models. Adding SFG parameter to default BYM (Case 2.3.3) gives maximum increment in correlation coefficient between actual and predicted ROP values.
- Feature selection is highly informative. Excluding the WOB effect (Function x_5) from Case 2.3.3 resulted in an increment in prediction accuracy (Case 2.3.7). WOB can be effective in ROP prediction when the parameter range is wider. The effect of WOB for ROP prediction is negligible for this particular dataset due to the narrow range of the variable.
- For the C-2 well, in heterogeneous formations (Case 3.1 and 3.3), SVR with RBF kernels gives better prediction results in RMSE. On the contrary, for the homogeneous formations (Case 3.2), linear models produce lower errors.
- Adding some extra features related to the rock mechanics has different effects on prediction performances for all models. Torque is found as the key

parameter to define the relationship between ROP and heterogeneity. Depth is another critical parameter on ROP prediction in heterogeneous formations.

- Feature selection based on the Random Forests algorithm is applied. However, the cut-off value is formation-specific. Therefore, feature selection should be performed for each data set exclusively.
- This study considers ROP prediction in well programming phase by using offset drilling data. However, the approach provided in this study can be modified for real-time applications. Further investigation on predicting ROP in heterogeneous formations in real-time operations is required.





REFERENCES

- Abbas, A. K., Assi, A. H., Abbas, H., Almubarak, H., and Al Saba, M.** (2019). Drill bit selection optimization based on rate of penetration: application of artificial neural networks and genetic algorithms, *Abu Dhabi International Petroleum Exhibition & Conference*. Abu Dhabi, UAE, November 11-14. doi:10.2118/197241-MS.
- Abbas, A. K., Rushdi, S., Alsaba, M., and Al Dushaishi, M. F.** (2019). Drilling rate of penetration prediction of high-angled wells using artificial neural networks, *Journal of Energy Resources Technology*, 141(11), 112904. doi:10.1115/1.4043699.
- Abdulmalek, A. S., Elkatatny, S., Abdulraheem, A., ... Mohamed, I. M.** (2018). Prediction of rate of penetration of deep and tight formation using support vector machine, *SPE Kingdom of Saudi Arabia Annual Technical Symposium and Exhibition*, Dammam, Saudi Arabia, April 23-26. doi:10.2118/192316-MS.
- Abe, S.** (2010). *Support Vector Machines for Pattern Classification*. 2nd Ed. London: Springer.
- Aggarwal, C. C.** (2018). *Neural Networks and Deep Learning*. Yorktown Heights, NY: Springer.
- Agustawijaya, D. S.** (2007). The uniaxial compressive strength of soft rock, *Civil Engineering Dimension*, 9(1), 9–14.
- Ahmed, O. S., Adeniran, A. A., and Samsuri, A.** (2019). Computational intelligence based prediction of drilling rate of penetration: a comparative study, *Journal of Petroleum Science and Engineering*, 172, 1–12. doi:10.1016/j.petrol.2018.09.027.
- Al-AbdulJabbar, A., Elkatatny, S., Mahmoud, M., and Abdulraheem, A.** (2018). Predicting rate of penetration using artificial intelligence techniques, *SPE Kingdom of Saudi Arabia Annual Technical Symposium and Exhibition*. Dammam, Saudi Arabia, April 23-26. doi:10.2118/192343-MS.
- Al-AbdulJabbar, A., Elkatatny, S., Mahmoud, M., and Abdelgawad, K.** (2019). A robust rate of penetration model for carbonate formation, *Journal of Energy Resources Technology*, 141(4), 042903. doi:10.1115/1.4041840.
- Al-Baiyat, I. and Heinze, L.** (2012). Implementing artificial neural networks and support vector machines in stuck pipe prediction, *SPE Kuwait International Petroleum Conference and Exhibition*, Kuwait City, Kuwait, December 10-12. doi:10.2118/163370-MS.

- Al-Betairi, E. A., Moussa, M. M., and Al-Otaibi, S.** (1988). Multiple regression approach to optimize drilling operations in the arabian gulf area, *SPE Drilling & Completion*, 3(01), 83–88. doi:10.2118/13694-PA.
- Al-Saleh, A.** (1999). Bit optimization in kuwait, *SPE/IADC Middle East Drilling Technology Conference*, Abu Dhabi, UAE, November 8-10. doi:10.2118/57552-MS.
- Al Dabyah, A. A., Elsherbeny, W. I., and Taylor, R. H.** (2018). Impact of real-time optimisation and geomechanics to enhance performance and optimise cost, *Offshore Technology Conference Asia*, Kuala Lumpur, Malaysia, March 20-23. doi:10.4043/28614-MS.
- Al Rubaii, M. M., Sehsah, O. R., and Omini, E.** (2018). Approach to improve well drilling performance, *SPE Kingdom of Saudi Arabia Annual Technical Symposium and Exhibition*, Dammam, Saudi Arabia, April 23-26. doi:10.2118/194223-ms.
- Alkinani, H. H., Al-Hameedi, A. T. T., Dunn-Norman, S., ... Amer, A. S.** (2018) Examination of the relationship between rate of penetration and mud weight based on unconfined compressive strength of the rock, *Journal of King Saud University - Science*, 31, 966–972. doi:10.1016/j.jksus.2018.07.020.
- Alsubaih, A., Albadran, F., and Alkanaani, N.** (2018). Mechanical specific energy and statistical techniques to maximizing the drilling rates for production section of mishrif wells in southern iraq fields, *SPE/IADC Middle East Drilling Technology Conference and Exhibition*, Abu Dhabi, UAE, January, 29-31. doi:10.2118/189354-MS.
- Amar, K. and Ibrahim, A.** (2012). Rate of penetration prediction and optimization using advances in artificial neural networks, a comparative study. In *International Joint Conference on Computational Intelligence. Proceedings of the 4th International Joint Conference on Computational Intelligence*, Barcelona, Spain, October 5-7. doi:10.5220/0004172506470652.
- Amer, M. M., Dahab, A. S., and El-Sayed, A. H.** (2017). An rop predictive model in nile delta area using artificial neural networks, *SPE Kingdom of Saudi Arabia Annual Technical Symposium and Exhibition*, Dammam, Saudi Arabia, April 24-27. doi:10.2118/187969-MS.
- Amjad, Q. B., Waheed, S., Jadoon, M. S. K., and Tariq, F.** (2015). Drilling optimization of kohat/potohar region by mathematical model (using matlab) and comparative method-a case study, *SPE/PAPG Pakistan section Annual Technical Conference*, Islamabad, Pakistan, November 24-25. doi: 0.2118/181124-MS.
- Anemangely, M., Ramezanzadeh, A., Tokhmechi, B., ... Mohammadian, A.** (2018). Development of a new rock drillability index for oil and gas reservoir rocks using punch penetration test, *Journal of Petroleum Science and Engineering*, 166, 131–145. doi:10.1016/j.petrol.2018.03.024.

- Arabjamaloei, R. and Karimi Dehkordi, B.** (2012). Investigation of the most efficient approach of the prediction of the rate of penetration, *Energy Sources, Part A: Recovery, Utilization, and Environmental Effects*, **34**(7), 581–590. doi:10.1080/15567036.2010.493925.
- Arabjamaloei, R. and Shadizadeh, S.** (2011). Modeling and optimizing rate of penetration using intelligent systems in an Iranian southern oil field (ahwaz oil field), *Petroleum Science and Technology*, **29**(16), 1637–1648. doi:10.1080/10916460902882818.
- Aronszajn, N.** (1950). Theory of reproducing kernels, *Transactions of the American Mathematical Society*, **68**(3), 337–404. doi:10.2307/1990404.
- Ashrafi, S. B., Anemangely, M., Sabah, M., and Ameri, M. J.** (2019). Application of hybrid artificial neural networks for predicting rate of penetration (rop): a case study from marun oil field, *Journal of Petroleum Science and Engineering*, **175**, 604–623. doi:10.1016/j.petrol.2018.12.013.
- Bahari, A. and Seyed, A. B.** (2007). Trust-region approach to find constants of bourgoyne and young penetration rate model in Khangiran Iranian gas field, *Latin American and Caribbean Petroleum Engineering Conference*, Buenos Aires, Argentina, April 15-18. doi:10.2118/107520-MS.
- Bahari, A. and Seyed, A. B.** (2009). Drilling cost optimization in a hydrocarbon field by combination of comparative and mathematical methods, *Petroleum Science*, **6**(4), 451–463. doi:10.1007/s12182-009-0069-x.
- Bahari, M. H., Bahari, A., Moharrami, F. N., and Sistani, M. B. N.** (2008). Determining bourgoyne and young model coefficients using genetic algorithm to predict drilling rate, *Journal of Applied Sciences*, **8**(17), 3050–3054. doi:10.3923/jas.2008.3050.3054.
- Bakar, N. A. and Rosbi, S.** (2017). Identification of non-equilibrium growth for bitcoin rxchange rate: mathematical derivation method in Islamic financial engineering, *International Journal of Scientific Research and Management*, **5**(12), 7772–7781. doi:10.18535/ijrm/v5i12.25.
- Barragan, R. V., Santos, O. L. A., and Maidla, E. E.** (1997). Optimization of multiple bit runs, in *SPE/IADC Drilling Conference*. Amsterdam, The Netherlands, March 4-6. doi:10.2118/37644-MS.
- Bassiouni, Z.** (1994). *Theory, Measurement, And Interpretation of Well Logs*. Richardson, TX: Society of Petroleum Engineers.
- Bataee, M., Kamyab, M., and Ashena, R.** (2010). Investigation of various rop models and optimization of drilling parameters for pdc and roller-cone bits in Shadegan oil field, *International Oil and Gas Conference and Exhibition in China*. Beijing, China, June 8-10. doi:10.2118/130932-MS.
- Bataee, M. and Mohseni, S.** (2011). Application of artificial intelligent systems in rop optimization: a case study, *SPE Middle East Unconventional Gas Conference and Exhibition*, Muscat, Oman, January 31-February 2. doi:10.2118/140029-MS.

- Bezminabadi, S. N., Ramezanzadeh, A., Jalali, S.E., ... Roustaei, A.** (2017). Effect of rock properties on rop modeling using statistical and intelligent methods: a case study of an oil well in southwest of iran, *Archives of Mining Sciences*, 62(1), 131–144. doi:10.1515/amsc-2017-0010.
- Bilgesu, H. I., Tetrick, L. T., Altmis, U., Mohaghegh, S., and Ameri, S.** (1997). New approach for the prediction of rate of penetration (rop) values, *SPE Eastern Regional Meeting*, Lexington, KY, October 22-24. doi:10.2523/39231-MS.
- Bishop, C. M.** (2006). *Pattern Recognition and Machine Learning*. Singapore: Springer.
- Bjornsson, E., Hucik, B., Szutiak, G., Brown Jr., L. A., ... Perry, P.** (2004). Drilling optimization using bit selection expert system and rop prediction algorithm improves drilling performance and enhances operational decision making by reducing performance uncertainties, *SPE Annual Technical Conference and Exhibition*, Houston, TX, September 26-29. doi:10.2118/90752-MS.
- Boswell, D.** (2002). Introduction to Support Vector Machines.
- Bourgoyne Jr., A. T., Millheim, K. K., Chenevery, M. E., and Young Jr., F. S.** (1991). *Applied Drilling Engineering*. Richardson, TX: Society of Petroleum Engineers.
- Bourgoyne Jr., A. T. and Young Jr., F. S.** (1974). A multiple regression approach to optimal drilling and abnormal pressure detection, *Society of Petroleum Engineers Journal*, 14(4), 371–384. doi:10.2118/4238-PA.
- Breiman, L.** (1996). Bias, Variance, and Arcing Classifier (Report No. 460). Berkeley, CA : Technical Report.
- Breiman, L.** (2001). Random forests, *Machine Learning*, 45, 5–32. doi:10.1023/A:1010933404324.
- Briševac, Z., Hrženjak, P., and Buljan, R.** (2016). Models for estimating uniaxial compressive strength and elastic modulus, *Građevinar*, 68(1), 19–28. doi:10.14256/JCE.1431.2015.
- Caicedo, H. U., Calhoun, W. M., and Ewy, R. T.** (2005). Unique rop predictor using bit-specific coefficient of sliding friction and mechanical efficiency as a function of confined compressive strength impacts drilling performance, *SPE/IADC Drilling Conference*, Amsterdam, The Netherlands, February 23-25. doi: 10.2118/92576-MS.
- Chu, W., Keerthi, S., and Ong, C.** (2001). A unified loss function in bayesian framework for support vector regression. In *ICML 2001. Proceedings of the Eighteenth International Conference on Machine Learning* Williamstown, MA, June 28-July 1.
- Combs, G. D.** (1968). Prediction of pore pressure from penetration rate, *Fall Meeting of the Society of Petroleum Engineers of AIME*, Houston, TX, September 29-October 2. doi: 10.2118/2162-MS.
- Cortes, C. and Vapnik, V.** (1995). support-vector networks, *Machine Learning*, 20, 273–297.

- Cunningham, R. A. and Eenink, J. G.** (1959). Laboratory study of effect of overburden, formation and mud column pressures on drilling rate of permeable formations, *Transactions of the AIME*, 217(01), 9–17. doi:10.2118/1094-g.
- Çağlayan, B. K.** (2014). *Torque and drag applications for deviated and horizontal wells: a case study* (Master's thesis). Middle East Technical University, The Graduate School of Natural and Applied Sciences, Ankara.
- Diaz, M. B., Kim, K. Y., Kang, T., and Shin, H.** (2018). Drilling data from an enhanced geothermal project and its pre-processing for rop forecasting improvement, *Geothermics*, 72, 348–357. doi:10.1016/j.geothermics.2017.12.007.
- Dietterich, T.** (1995). Overfitting and undercomputing in machine learning, *ACM Computing Surveys*, 27(3), 326–327. doi:10.1145/212094.212114.
- Doiron, H. H. and Deane, J. D.** (1982). Effects of hydraulic parameter cleaning variations on rate of penetration of soft formation insert bits, *SPE Annual Technical Conference and Exhibition*, New Orleans, LA, September 26-29. doi:10.2523/11058-MS.
- Dupriest, F. E. and Koederitz, W. L.** (2005). Maximizing drill rates with real-time surveillance of mechanical specific energy, *SPE/IADC Drilling Conference*, Amsterdam, The Netherlands, February 23-25. doi:10.2118/92194-MS.
- Eckel, J. R.** (1967). Microbit studies of the effect of fluid properties and hydraulics on drilling rate, *Journal of Petroleum Technology*, 19(4), 541–546. doi:10.2118/1520-PA.
- Edwards, J. H.** (1964). Engineering design of drilling operations, *Drilling and Production Practice*, 39–55.
- Elkatatny, S.** (2018). Rate of penetration prediction using self-adaptive differential evolution-artificial neural network, *SPE Kingdom of Saudi Arabia Annual Technical Symposium and Exhibition*, Dammam, Saudi Arabia, April 23-26. doi: 10.2118/192186-MS.
- Elkatatny, S.** (2020). Real-time prediction of rate of penetration in s-shape well profile using artificial intelligence models, *Sensors*, 20(12), 3506. doi:10.3390/s20123506.
- Elkatatny, S.** (2021). Real-time prediction of rate of penetration while drilling complex lithologies using artificial intelligence techniques, *Ain Shams Engineering Journal*, 12(1), 917–926. doi:10.1016/j.asej.2020.05.014.
- Elkatatny, S. M., Tariq, Z., Mahmoud, M. A., and Al-AbdulJabbar, A.** (2017). Optimization of rate of penetration using artificial intelligent techniques, *51st US Rock Mechanics/Geomechanics Symposium*, San Francisco, CA, June 25-28.
- Eren, T. and Kok, M. V.** (2018). A new drilling performance benchmarking: rop indexing methodology, *Journal of Petroleum Science and Engineering*, 163, 387–398. doi:10.1016/j.petrol.2018.01.002.

- Eren, T. and Ozbayoglu, M. E.** (2010). Real time optimization of drilling parameters during drilling operations, *SPE Oil and Gas India Conference and Exhibition* Mumbai, India, January 20-22. doi:10.2118/129126-MS.
- Eskandarian, S., Bahrami, P., and Kazemi, P.** (2017). A comprehensive data mining approach to estimate the rate of penetration: application of neural network, rule based models and feature ranking, *Journal of Petroleum Science and Engineering*, 156(6), 605–615. doi:10.1016/j.petrol.2017.06.039.
- Esmaili, A., Elahifar, B., Fruhwirth, R. K., and Thonhauser, G.** (2012). ROP modeling using neural network and drill string vibration data, *SPE Kuwait International Petroleum Conference and Exhibition*, Kuwait City, Kuwait, December 10-12. doi:10.2118/163330-MS.
- Evangelatos, G. I. and Payne, M. L.** (2016). Advanced bha-rop modeling including neural network analysis of drilling performance data, *IADC/SPE Drilling Conference and Exhibition*, Fort Worth, TX, March 1-3. doi:10.2118/178852-MS.
- Fear, M. J.** (1999). How to improve rate of penetration in field operations, *SPE Drilling and Completion*, 14(1), 42–49. doi:10.2118/55050-PA.
- Flach, P.** (2012) *Machine Learning: The Art and Science of Algorithms that Make Sense of Data*. Cambridge: Cambridge University Press.
- Freund, R. J., Wilson, W. J., and Sa, P.** (2006). *Regression Analysis: Statistical Modeling of a Response Variable*. 2nd Edition. USA: Elsevier.
- Galle, E. M. and Woods, H. B.** (1963). Best constant weight and rotary speed for rotary rock bit, *Drilling and Production Practice*, 48–73.
- Gan, C., Cao, W., Liu, K., Wu, M., Wang, F., and Zhang, S.** (2019). A new hybrid bat algorithm and its application to the rop optimization in drilling processes, *IEEE Transactions on Industrial Informatics*, 16(12), 7338–7348. doi:10.1109/tii.2019.2943165.
- Gan, C., Cao, W., Wu, M., Chen, X., Hu, Y., Liu, K., Wang, F., and Zhang, S.** (2019). Prediction of drilling rate of penetration (rop) using hybrid support vector regression: a case study on the shennongjia area, central china, *Journal of Petroleum Science and Engineering*, 181, 106200. doi:10.1016/j.petrol.2019.106200.
- Gan, C., Cao, W., Min, W., Liu, K., Chen, X., Hu, Y., and Ning, F.** (2019). Two-level intelligent modeling method for the rate of penetration in complex geological drilling process, *Applied Soft Computing*, 80, 592–602. doi:10.1016/j.asoc.2019.04.020.
- Garnier, A. J. and van Lingen, N. H.** (1959). Phenomena affecting drilling rates at depth, *Transactions of the AIME*, 216(01), 232–239. doi:10.2118/1097-g.
- Ghahramani, Z.** (2004). Unsupervised Learning. In O. Bousquet, G. Raetsch, and U. von Luxburg (Eds.), *Advanced Lectures on Machine Learning LNAI 3176* (pp. 72-113), New York : Springer.

- Gidh, Y., Purwanto, A., and Ibrahim, H.** (2012). Artificial neural network drilling parameter optimization system improves rop by predicting/managing bit wear, *SPE Intelligent Energy International*, Utrecht, The Netherlands, March 27-29. doi:10.2118/149801-MS.
- Girosi, F.** (1998). An equivalence between sparse approximation and support vector machines, *Neural Computation*, 10(6), 1455–1480. doi:10.1162/089976698300017269.
- Graham, J. W. and Muench, N. L.** (1959). Analytical determination of optimum bit weight and rotary speed combinations, *Fall Meeting of the Society of Petroleum Engineers of AIME*, Dallas, TX, October 4-7. doi:10.2118/1349-g.
- Grubbs, F. E.** (1969). Procedures for detecting outlying observations in samples, *Technometrics*, 11(1), 1–21. doi:10.1080/00401706.1969.10490657.
- Gunn, S. R.** (1998). Support vector machines for classification and regression (Technical Report). Southampton : University of Southampton.
- Guo, Y., Guan, S., and Liu, Z.** (2017). Rop optimization technology for poor drillability formation in Western South China Sea,” in *SPE Middle East Oil & Gas Show and Conference*. Manama: Society of Petroleum Engineers. doi: 10.2118/183912-ms.
- Hamad-Allah, S. M. and Ismael, A. A.** (2008). Application of mathematical drilling model on southern iraqi oil fields, *Journal of Engineering*, 14(3), 2763–2784.
- Han, Y., Li, X. and Feng, Y.** (2020). A new approach to evaluate rock drillability of polycrystalline diamond compact bits using scratch test data, *Energy Exploration & Exploitation*, 38(4), 884–904. doi:10.1177/0144598719901091.
- Hareland, G., Wu, A., Rashidi, B., and James, J. A.** (2010). A new drilling rate model for tricone bits and its application to predict rock compressive strength, *44th US Rock Mechanics Symposium & 5th US/Canada Rock Mechanics Symposium*, Salt Lake City, UT, June 27-30.
- Hassan, A., Al-Majed, A., Elkatatny, S., Mahmoud, M., ... Khamis, M.** (2018). Developing an efficient drilling system by coupling torque modelling with mechanical specific energy, in *SPE Kingdom of Saudi Arabia Annual Technical Symposium and Exhibition*, Dammam, Saudi Arabia, April 23-26. doi:10.2118/192251-MS.
- Hastie, T., Tibshirani, R., and Freidman, J.** (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2nd Edition. New York, NY : Springer.
- He, J., Chen, Y., Zhengchun, L., and Samuel, R.** (2016). Global correlation of rock brittleness indices with petrophysical and geomechanical properties and its application to the prediction of rate of penetration rop, *IADC/SPE Asia Pacific Drilling Technology Conference*, Singapore, August 22-24. doi:10.2118/180518-MS.

- Heckman, N.** (2012). The theory and application of penalized methods or reproducing kernel hilbert spaces made easy, *Statistics Surveys*, 6(1), 113–141. doi:10.1214/12-SS101.
- Hegde, C., Daigle, H., Millwater, H., and Gray, K.** (2017). Analysis of rate of penetration (rop) prediction in drilling using physics-based and data-driven models, *Journal of Petroleum Science and Engineering*, 159, 295–306. doi:10.1016/j.petrol.2017.09.020.
- Hegde, C., Millwater, H., Pyrcz, M., Daigle, H., and Gray, K.** (2019). Rate of penetration (rop) optimization in drilling with vibration control, *Journal of Natural Gas Science and Engineering*, 67, 71–81. doi:10.1016/j.jngse.2019.04.017.
- Hegde, C., Daigle, H., and Gray, K. E.** (2018). Performance comparison of algorithms for real-time rate-of-penetration optimization in drilling using data-driven models, *SPE Journal*, 23(5), 1706–1722. doi:10.2118/191141-pa.
- Hegde, C. and Gray, K.** (2018). Evaluation of coupled machine learning models for drilling optimization, *Journal of Natural Gas Science and Engineering*, 56, 397–407. doi:10.1016/j.jngse.2018.06.006.
- Hegde, C., Soares, C., and Gray, K.** (2018). Rate of penetration (rop) modeling using hybrid models: deterministic and machine learning, *SPE/AAPG/SEG Unconventional Resources Technology Conference*, Houston, TX, July 23–25. doi: 10.15530/urtec-2018-2896522.
- Huber, P. J.** (1964). Robust estimation of a location parameter, *The Annals of Mathematical Statistics*, 35(1), 73–101. doi:10.1214/aoms/1177703732.
- Iqbal, F.** (2008). Drilling optimization technique - using real time parameters, *SPE Russian Oil and Gas Technical Conference and Exhibition*, Moscow, Russia, October 28–30. doi:10.2118/114543-MS.
- Irawan, S., Rahman, A. M. A., and Tunio, S. Q.** (2012). Optimization of weight on bit during drilling operation based on rate of penetration model, *Research Journal of Applied Sciences, Engineering and Technology*, 4(12), 1690–1695.
- Jahanbakhshi, R., Keshavarzi, R., Shoorehdeli, M. A., and Emamzadeh, A.** (2012). Intelligent prediction of differential pipe sticking by support vector machine compared with conventional artificial neural networks: an example of iranian offshore oil fields, *SPE Drilling and Completion*, 27(4), 586–595. doi:10.2118/163062-PA.
- Jamshidi, E. and Mostafavi, H.** (2013). Soft computation application to optimize drilling bit selection utilizing virtual intelligence and genetic algorithms, *International Petroleum Technology Conference*, Beijing, China, March 26–28. doi:10.2523/16446-MS.
- Jiang, W. and Samuel, R.** (2016). Optimization of rate of penetration in a convoluted drilling framework using ant colony optimization, *IADC/SPE Drilling Conference and Exhibition*, Fort Worth, TX, March 1–3. doi: 10.2118/178847-MS.

- John, Z., Ahsan, A., and Reid, I.** (2002). Optimized decision making through real time access to drilling and geological data from remote wellsites, *SPE Asia Pacific Oil and Gas Conference and Exhibition*, Melbourne, Australia, October 8-10. doi: 10.2118/77855-MS.
- Judzis, A., Bland, R. G., Curry, D. A., Black, A. D., ... Grant, T. C.** (2009). Optimization of deep-drilling performance - benchmark testing drives rop improvements for bits and drilling fluids, *SPE/IADC Drilling Conference*, Amsterdam, The Netherlands, February 20-22. doi:10.2118/105885-MS.
- Kecman, V.** (2001). *Learning and Soft Computing: Support Vector Machines, Neural Networks, and Fuzzy Logic Models*. USA: The MIT Press.
- Kor, K. and Altun, G.** (2020). Is support vector regression method suitable for predicting rate of penetration?, *Journal of Petroleum Science and Engineering*, 194, 107542. doi:10.1016/j.petrol.2020.107542.
- Kor, K., Ertekin, S., Yamanlar, S., and Altun, G.** (2021). Penetration rate prediction in heterogeneous formations: a geomechanical approach through machine learning, *Journal of Petroleum Science and Engineering*, 207, 109138. doi:10.1016/j.petrol.2021.109138.
- Kristjansson, S. D., Neudfeldt, A., Lai, S. W., Wang, J., and Tremaine, D.** (2016). Use of historic data to improve drilling efficiency: a pattern recognition method and trial results, *IADC/SPE Drilling Conference and Exhibition*, Fort Worth, TX, March 1-3. doi:10.2118/178901-MS.
- Kutas, D. T., Nascimento, A., Elmergi, A. M., ... Mathias, M. H.** (2015). A study of the applicability of bourgoyne & young rop model and fitting reliability through regression, *International Petroleum Technology Conference*, Doha, Qatar, December 6-9. doi:10.2523/iptc-18521-MS.
- Laerd Statistics.** (n.d.). *Hypothesis Testing*. Retrieved April 9, 2021, from <https://statistics.laerd.com/statistical-guides/hypothesis-testing-3.php>.
- Lee, S. S., Kim, K. Y., and Seo, J. W.** (2018). Development of a trip time for bit exchange simulator for drilling time estimation, *Geothermics*, 71, 24–33. doi:10.1016/j.geothermics.2017.07.006.
- Lummus, J. L.** (1971). Acquisition and analysis of data for optimized drilling, *Journal of Petroleum Technology*, 23(11), 1285–1293. doi:10.2118/3716-PA.
- Lyons, W. C.** (2010). *Working Guide to Drilling Equipment and Operations*. 1st Edition. Burlington, MA : Elsevier.
- Maehs, J., Renne, S., Logan, B., and Diaz, N.** (2010). Proven methods and techniques to reduce torque and drag in the pre-planning and drilling execution of oil and gas wells, *IADC/SPE Drilling Conference and Exhibition*, New Orleans, LA, February 2-4. doi:10.2118/128329-MS.
- Mantha, B. and Samuel, R.** (2016). Rop optimization using artificial intelligence techniques with statistical regression coupling, *SPE Annual Technical Conference and Exhibition*, Dubai, UAE, September 26-28. doi:10.2118/181382-MS.

- Mathia, E., Ratcliffe, K., and Wright, M.** (2016). Brittleness index - a parameter to embrace or avoid?, *SPE/AAPG/SEG Unconventional Resources Technology Conference*, San Antonio, TX, August 1-3. doi:10.15530/urtec-2016-2448745.
- Maurer, W. C.** (1962). The 'perfect - cleaning' theory of rotary drilling, *Journal of Petroleum Technology*, 14(11), 1270–1274. doi:10.2118/408-PA.
- Mazen, A. Z., Mujtaba, I. M., Hassanpour, A., and Rahmanian, N.** (2020). Mathematical modelling of performance and wear prediction of pdc drill bits: impact of bit profile, bit hydraulic, and rock strength, *Journal of Petroleum Science and Engineering*, 188, 106849. doi:10.1016/j.petrol.2019.106849.
- McLean, M. R. and Addis, M. A.** (1990). Wellbore stability analysis: a review of current methods of analysis and their field application,” 1990 *IADC/SPE Drilling Conference*, Houston, TX, February 27-March 2.
- Mehrad, M., Bajolvand, M., Ramezanzadeh, A., and Neycharan, J. G.** (2020). Developing a new rigorous drilling rate prediction model using a machine learning technique, *Journal of Petroleum Science and Engineering*, 192, 107338. doi:10.1016/j.petrol.2020.107338.
- Mensa-Wilmot, G., Douglas, C., and Schell, E.** (2004). New pdc bit technology, improved drillability analysis, and operational practices improve drilling performance in hard and highly heterogeneous applications, *SPE Eastern Regional Meeting*, Charleston, WV, September 15-17. doi:10.2523/91423-MS.
- Mercer, J.** (1909). Functions of positive and negative type, and their connection with the theory of integral equations, *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character.*, 209, 415–446.
- Meyer, D.** (2014). Support vector machines: the interface to libsvm in package e1071 [Software documentation]. doi:10.1007/978-0-387-78189-1_11.
- Miao, S., Blackman, M., Luo J., and Luo, L.** (2016). Hybrid bit produces breakthrough performance in heterogeneous formations in china, *Offshore Technology Conference Asia*, Kuala Lumpur, Malaysia, March 22-25. doi:10.4043/26357-MS.
- Millan, E. and Ringer, M.** (2018). A new workflow for estimating bit wear and monitoring drilling efficiency in real time during drilling operations, *IADC/SPE Drilling Conference and Exhibition*, Fort Worth, TX, March 6-8. doi:10.2118/189582-MS.
- Millheim, K. K. and Gaebler, T.** (1999). Virtual experience simulation for drilling - the concept, *SPE/IADC Drilling Conference*, Amsterdam, The Netherlands, March 9-11. doi:10.2523/52803-MS.
- Monazami, M., Hashemi, A., and Shahbazian, M.** (2012). Drilling rate of penetration prediction using artificial neural network: a case study of one of iranian southern oil fields, *Oil and Gas Business*, 6, pp. 21–31.

- Monden, T. and Chia, C. R.** (2007). Operation support centers - real time drilling optimization and risk mitigation, *SPE Saudi Arabia Section Technical Symposium*, Dhahran, Saudi Arabia, May 7-8. doi:10.2523/110950-MS.
- Montgomery, D. C. and Runger, G. C.** (2003). *Applied Statistics and Probability for Engineers*. 3rd Edition. USA: John Wiley & Sons.
- Moradi, H., Bahari, M. H., Sistani, M. B. N., and Bahari, A.** (2010). Drilling rate prediction using an innovative soft computing approach, *Scientific Research and Essays*, 5(13), 1583–1588.
- Moraveji, M. K. and Naderi, M.** (2016). Drilling rate of penetration prediction and optimization using response surface methodology and bat algorithm, *Journal of Natural Gas Science and Engineering*, 31, 829–841. doi:10.1016/j.jngse.2016.03.057.
- Mulmuley, K.** (1994). *Computational Geometry: An Introduction Through Randomized Algorithms*. NJ : Prentice-Hall.
- Murray, A. S. and Cunningham, R. A.** (1955). Effect of mud column pressure on drilling rates, *Transactions of the AIME*, 204(01), 196–204. doi:10.2118/505-g.
- Nadhan, D., Mayani, M. G., and Rommetveit, R.** (2018). Drilling with digital twins, *IADC/SPE Asia Pacific Drilling Technology Conference and Exhibition*, Bangkok, Thailand, August 27-29. doi: 10.2118/191388-MS.
- Nasir, E. and Rickabaugh, C.** (2018). Optimizing drilling parameters using a random forests rop model in the permian basin, *SPE Liquids-Rich Basins Conference - North America*, Midland, TX, September 5-6. doi:10.2118/191796-MS.
- NPD.** (2020). *Resource Report Exploration 2020* (Report). Stavanger : Norwegian Petroleum Directorate.
- Okay, A. I.** (2008). Geology of turkey: a synopsis, *Anschnitt*, 14, 19–42.
- Okutucu, A., Di Pasquale, M., Calvaresi, E., ... Guerrini, G.** (2013). Hybrid bit produces breakthrough performance and savings for inland application in Turkey, *Offshore Mediterranean Conference and Exhibition*, Ravenna, Italy, March 20-23.
- OnePetro.** (2021). Retrieved February 7, 2021, from <https://onepetro.org/>.
- Osgouei, R. E.** (2007). *Rate of penetration estimation model for directional and horizontal wells* (Master's thesis). Middle East Technical University, The Graduate School of Natural and Applied Sciences, Ankara.
- Pessier, R. C. and Fear, M. J.** (1992). Quantifying common drilling problems with mechanical specific energy and a bit-specific coefficient of sliding friction, *SPE Annual Technical Conference and Exhibition*, Washington, DC, October 4-7. doi:10.2118/24584-MS.

- Pinheiro, M., Carneiro, R., and Lima, A.** (2015). Successful application of drilling optimization methodology and integrated solutions in a pre-salt well, *OTC Brasil*, Rio de Janeiro, Brazil, October 27-29. doi:10.4043/26284-MS.
- Rahimi, R. and Nygaard, R.** (2015). Comparison of rock failure criteria in predicting borehole shear failure, *International Journal of Rock Mechanics & Mining Sciences*, 79, 29–40. doi:10.1016/j.ijrmms.2015.08.006.
- Rahimzadeh, H., Mostofi, M., and Hashemi, A.** (2011). A new method for determining bourgoyne and young penetration rate model constants, *Petroleum Science and Technology*, 29(9), 886–897. doi:10.1080/10916460903452009.
- Rao, S. S.** (2009). *Engineering Optimization: Theory and Practice*. 4th Ed. Hoboken, NJ : John Wiley & Sons.
- Rastegar, M., Hareland, G., Nygaard, R., and Bashari, A.** (2008). Optimization of multiple bit runs based on rop models and cost equation: a new methodology applied for one of the persian gulf carbonate fields, *IADC/SPE Asia Pacific Drilling Technology Conference and Exhibition*, Jakarta, Indonesia, August 25-27. doi:10.2118/114665-ms.
- Reed, R. L.** (1972). A monte carlo approach to optimal drilling, *Society of Petroleum Engineers Journal*, 12(05), 423–438. doi:10.2118/3513-PA.
- Rickman, R., Mullen, M., Petre, E., Grieser, B., and Kundert, D.** (2008). A practical use of shale petrophysics for stimulation design optimization: all shale plays are not clones of the barnett shale, *SPE Annual Technical Conference and Exhibition*, Denver, CO, September 21-24. doi:10.2118/115258-MS.
- Rommetveit, R., Bjørkevoll, K. S., Halsey, G. W., ... Ødegaard, S. I.** (2004). Drilltronics: an integrated system for real-time optimization of the drilling process, *IADC/SPE Drilling Conference*, Dallas, TX, March 2-4. doi:10.2118/87124-MS.
- Sabah, M., Talebkeikhah, M., Wood, D. A., ... Younesi, A.** (2019). A machine learning approach to predict drilling rate using petrophysical and mud logging data, *Earth Science Informatics*, 12(3), 319–339. doi:10.1007/s12145-019-00381-4.
- Schölkopf, B., Bartlett, P., Smola, A., and Williamson, R.** (1998). Support Vector Regression with Automatic Accuracy Control. In Niklasson L., Bodén M., Ziemke T. (Eds.), *ICANN 98*. London : Springer.
- Sevinç, S.** (2020). *Gözenek Basıncı, In Situ Stresler ve Kuyu Stabilitesi* (Rapor). Ankara : TPAO Arama Daire Başkanlığı.
- Shi, X., Liu, G., Gong, X., Zhang, J., Wang, J., and Zhang, H.** (2016). An efficient approach for real-time prediction of rate of penetration in offshore drilling, *Mathematical Problems in Engineering*, 2016, 3575380. doi:10.1155/2016/3575380.
- Shi, Z.** (2011). *Advanced Artificial Intelligence*. Singapore : World Scientific Publishing.

- Simmons, E. L.** (1986). A technique for accurate bit programming and drilling performance optimization, *IADC/SPE Drilling Conference*, Dallas, TX, February 10-12. doi:10.2118/14784-MS.
- Smith.** (2008). *Roller Cone Dull Grading Manual*. Houston : Smith International.
- Smola, A.** (1996). *Regression estimation with support vector learning machines* (Master's Thesis). Technische Universität München, Physik Department, München.
- Smola, A. J. and Schölkopf, B.** (2003). A tutorial on support vector regression (Technical report). Australia : RSISE
- Soares, C. and Gray, K.** (2019). Real-time predictive capabilities of analytical and machine learning rate of penetration (rop) models, *Journal of Petroleum Science and Engineering*, 172, 934–959. doi:10.1016/j.petrol.2018.08.083.
- Speer, J. W.** (1959). A method for determining optimum drilling techniques, *Gulf Coast Drilling and Production Meeting*, Lafayette, LA, February. doi:10.2118/1242-G.
- Stassi-D'Alia, F.** (1967). Flow and fracture of materials according to a new limiting condition of yielding, *Meccanica*, 2(3), 178–195. doi:10.1007/BF02128173.
- Stevenson, N., Kvevik, G., Mañalac, E., Persson, C., and Dorr, J.** (2013). Reducing brownfield development costs via improved drilling practice and enhanced understanding of wellbore instability in weak formations, *SPE Annual Technical Conference and Exhibition*, New Orleans, LA, September 30-October 2. doi:10.2118/166300-MS.
- Şahin, F.** (1995). *A study on kastel and karadut marl drilling problems* (Master's thesis). Middle East Technical University, The Graduate School of Natural and Applied Sciences, Ankara.
- Tansev, E.** (1975). A heuristic approach to drilling optimization, *Fall Meeting of the Society of Petroleum Engineers of AIME*, Dallas, TX, September 28-October 1. doi:10.2523/5546-MS.
- Taylor, A., Wesley, A., Adari, R., and Kurian, L.** (2016). Employing process, benchmark automation, and real-time analysis to improve drilling efficiency, *IADC/SPE Asia Pacific Drilling Technology Conference*, Singapore, August 22-24. doi:10.2118/180638-MS.
- Thomson, I., Krasuk, R., Silva, N., and Romero, K.** (2011). Hybrid drill bit improves drilling performance in heterogeneous formations in brazil, *Brazil Offshore*, Macaé, Brazil, June 14-17. doi:10.2118/143686-MS.
- Tingay, M. R. P., Hillis, R. R., Morley, C. K., ... Drake S. J.** (2005). Present-day stress orientation in brunei: a snapshot of 'prograding tectonics' in a tertiary delta, *Journal of the Geological Society*, 162(1), 39–49. doi:10.1144/0016-764904-017.
- Ursem, L., Williams, J. H., Pellerin, N. M., and Kaminski, D. H.** (2003). Real time operations centers; the people aspects of drilling decision making, *SPE/IADC Drilling Conference*, Amsterdam, The Netherlands, February 19-21. doi:10.2118/79893-MS.

- Valisevich, A., Ruzhnikov, A., Bebeshko, I., ... Zhentichka, M.** (2015). Drillbit optimization system: real-time approach to enhance rate of penetration and bit wear monitoring main phases of implementation, *SPE Russian Petroleum Technology Conference*, Moscow, Russia, October 26-28. doi:10.2118/176517-MS.
- Vapnik, V. N.** (2000). *The Nature of Statistical Learning Theory*. 2nd Edition. New York : Springer.
- Vidrine, D. J. and Benit, E. J.** (1968). Field verification of the effect of differential pressure on drilling rate, *Journal of Petroleum Technology*, 20(07), 676–682. doi:10.2118/1859-PA.
- Wallace, S. P., Hegde, C. M., and Gray, K. E.** (2015). A system for real-time drilling performance optimization and automation based on statistical learning methods, *SPE Middle East Intelligent Oil and Gas Conference and Exhibition*, Abu Dhabi, UAE, September 15-16. doi:10.2118/176804-MS.
- Warren, T. M.** (1987). Penetration-rate performance of roller-cone bits, *SPE Drilling Engineering*, 2(1), 9–18. doi:10.2118/13259-PA.
- Wiktorski, E., Kuznetsov, A., and Sui, D.** (2017). Rop optimization and modeling in directional drilling process, *SPE Bergen One Day Seminar*, Bergen, Norway, April 5. doi:10.2118/185909-MS.
- Wright, J., Chukwu, G. A., Khataniar, S., and Patil, S.** (2003). An economic appraisal of hole cleaning using hydraulic horsepower and jet impact force, *SPE Western Regional/AAPG Pacific Section Joint Meeting*, Long Beach, CA, May 19-24. doi:10.2118/83496-MS.
- Wu, A., Hareland, G., and Fazelizadeh, M.** (2011). Torque & drag analysis using finite element method, *Modern Applied Science*, 5(6), 13–27. doi:10.5539/mas.v5n6p13.
- Yan, X. and Su, X. G.** (2009). *Linear Regression Analysis: Theory and Computing*. Singapore : World Scientific Publishing.
- Yang, X. S.** (2010). *Engineering Optimization: An Introduction with Metaheuristic Applications*. New Jersey : John Wiley & Sons.
- Yi, P., Kumar, A., and Samuel, R.** (2014). Real-time rate of penetration optimization using the shuffled frog leaping algorithm (sfla), *SPE Intelligent Energy Conference & Exhibition*, Utrecht, The Netherlands, April 1-3. doi:10.2118/167824-MS.
- Young Jr., F. S.** (1969). Computerized drilling control, *Journal of Petroleum Technology*, 21(04), 483–496. doi:10.2118/2241-PA.
- Yuswandari, A., Prayoga, A., and Purba, D.** (2019). Rate of penetration (rop) prediction using artificial neural network to predict rop for nearby well in a geothermal field, *44th Workshop on Geothermal Reservoir Engineering*, Stanford, CA, February 11-13.
- Zhang, T.** (2001). An introduction to support vector machines and other kernel-based learning methods, *AI Magazine*, 22(2), 103–115. doi:10.1108/k.2001.30.1.103.6.

- Zhao, Y., Noorbakhsh, A., Koopialipoor, M., Azizi, A., and Tahir, M. M.** (2020). A new methodology for optimization and prediction of rate of penetration during drilling operations, *Engineering with Computers*, 36(2), 587–595. doi:10.1007/s00366-019-00715-2.
- Zhu, H., Deng J., Xie, Y., Huang, K., Zhao, J., and Yu, B.** (2012). Rock mechanics characteristic of complex formation and faster drilling techniques in western south china sea oilfields, *Ocean Engineering*, 44, 33–45. doi:10.1016/j.oceaneng.2012.01.031.





APPENDICES

APPENDIX A: Datasets Taken from the Literature

APPENDIX B: Derivation of Support Vector Machines

APPENDIX C: R Script for the Calculations

APPENDIX D: Graphical Results of Intermediate Steps





APPENDIX A: Datasets Taken from the Literature

Table A.1 : Data from Bourgoyne Jr. and Young Jr. (1974).

Data Entry	Depth [ft]	ROP [ft/hr]	Bit Weight [1000 lb/in]	Rotary Speed [rpm]	Tooth Wear	Reynolds Number Function	ECD [lb/gal]	Pore Gradient [lb/gal]
1	9515	23.0	2.58	113	0.77	0.964	9.5	9.0
2	9830	22.0	1.15	126	0.38	0.964	9.5	9.0
3	10130	14.0	0.81	129	0.74	0.827	9.6	9.0
4	10250	10.0	0.95	87	0.15	0.976	9.7	9.0
5	10390	16.0	1.02	78	0.24	0.984	9.7	9.0
6	10500	19.0	1.69	81	0.61	0.984	9.7	9.1
7	10575	13.0	1.56	81	0.73	0.984	9.7	9.2
8	10840	16.6	1.63	67	0.38	0.932	9.8	9.3
9	10960	15.9	1.83	65	0.57	0.878	9.8	9.4
10	11060	15.7	2.03	69	0.72	0.878	9.8	9.5
11	11475	14.0	1.69	77	0.20	0.887	10.3	9.5
12	11775	13.5	2.31	58	0.12	0.852	11.8	10.1
13	11940	6.2	2.26	67	0.2	0.976	15.3	12.4
14	12070	9.6	2.07	84	0.08	0.993	15.7	13.0
15	12315	15.5	3.11	69	0.40	1.185	16.3	14.4
16	12900	31.4	2.82	85	0.42	1.150	16.7	15.9
17	12975	42.7	3.48	77	0.17	1.221	16.7	16.1
18	13055	38.6	3.29	75	0.29	1.161	16.8	16.2
19	13250	43.4	2.82	76	0.43	1.161	16.8	16.2
20	13795	12.5	1.60	81	0.56	0.272	16.8	16.2
21	14010	21.1	1.04	75	0.46	0.201	16.8	16.2
22	14455	19.0	1.76	64	0.16	0.748	16.9	16.2
23	14695	18.7	2.00	76	0.27	0.819	17.1	16.2
24	14905	20.2	2.35	75	0.33	0.419	17.2	16.4
25	15350	27.1	2.12	85	0.31	1.29	17.0	16.5
26	15740	14.8	2.35	78	0.81	0.802	17.3	16.5
27	16155	12.6	2.47	80	0.12	0.670	17.9	16.5
28	16325	14.9	3.76	81	0.50	0.532	17.5	16.6
29	17060	13.8	3.76	65	0.91	0.748	17.6	16.6
30	20265	9.0	3.41	60	0.01	0.512	17.7	16.6

Table A.2 : Data from Irawan et al. (2012).

Data Entry	Depth [ft]	ROP [ft/hr]	Bit Weight [1000 lb/in]	Rotary Speed [rpm]	Tooth Wear	Reynolds Number Function	ECD [lb/gal]	Pore Gradient [lb/gal]
1	2150	171	0.82	120	0.500	0.882	8.93	8.365
2	2155	20	0.57	110	0.125	0.819	9.06	8.365
3	3591	160	0.82	120	0.500	1.290	9.11	8.365
4	5190	82	1.63	120	0.750	1.290	9.11	8.365
5	5872	49	2.45	120	0.875	1.290	9.11	8.365
6	6000	43	2.45	120	0.250	1.290	9.11	8.365
7	6080	64	1.63	120	0.625	1.062	9.49	8.365
8	6322	36	2.45	120	0.875	0.772	9.67	8.365
9	6592	27	2.85	120	1.000	0.772	9.67	8.365
10	6679	14	0.41	120	0.625	1.338	9.69	8.365
11	7341	83	1.63	180	0.375	1.145	9.69	8.365
12	8921	46	1.63	180	0.000	1.216	9.68	8.365
13	9363	47	1.63	180	0.000	0.868	9.88	8.571
14	9652	19	2.85	100	1.000	1.192	9.96	8.960
15	9660	3	2.45	65	0.125	1.192	9.96	8.960
16	10662	34	1.22	180	0.000	1.097	9.96	8.910
17	10735	16	2.86	65	0.125	1.192	9.96	8.900
18	10900	35	0.82	150	0.000	1.034	9.96	8.890
19	11214	12	3.27	70	0.250	1.114	9.96	8.880
20	11224	5	2.94	100	0.375	0.903	11.10	9.390
21	11481	26	1.76	170	0.000	0.975	11.02	9.370
22	12885	28	1.76	160	0.000	0.975	11.02	9.860
23	13180	11	1.76	130	0.000	0.825	10.96	10.120
24	13810	21	1.76	150	0.000	0.632	10.97	10.040
25	14300	15	1.76	160	0.000	0.632	10.95	9.980

Table A.3 : Data from Al-Betairi et al. (1988).

Data Entry	Depth [ft]	ROP [ft/hr]	Bit Weight [1000 lb/in]	Rotary Speed [rpm]	Tooth Wear	Jet Impact Force (lbf)	ECD [lb/gal]	Pore Gradient [lb/gal]
1	1187	80.0	0.02	120	0.000	16.362	18.2	9.0
2	1675	97.6	0.24	120	0.125	24.933	28.5	9.2
3	2084	81.8	2.04	120	0.250	46.031	17.1	9.4
4	2720	56.5	3.35	110	0.625	46.031	16.3	9.5
5	3563	56.2	3.35	90	0.500	48.907	13.3	9.5
6	3944	28.2	3.35	110	0.500	48.907	12.7	9.5
7	6069	39.0	4.09	100	0.750	42.842	14.9	9.5
8	6290	13.0	4.41	100	0.125	40.700	14.5	9.5
9	6530	17.8	4.09	100	0.500	40.700	14.3	9.5
10	6739	14.9	4.09	100	0.625	40.700	13.9	9.4
11	7346	16.2	5.03	100	0.875	40.700	13.5	8.4
12	7467	10.1	4.56	110	0.750	40.700	13.4	9.4
13	7646	10.5	5.03	100	1.000	40.700	13.3	9.5
14	7711	10.0	5.03	120	0.500	40.700	13.3	9.5
15	7881	7.5	5.03	100	0.500	39.707	13.3	9.3
16	8008	5.1	5.03	100	0.250	40.700	13.1	8.4
17	8114	7.1	4.09	100	0.000	40.700	13.1	8.4
18	8564	13.2	3.63	90	0.250	40.700	12.9	9.3
19	8903	7.9	3.63	85	0.000	40.700	12.7	9.5
20	9298	5.5	4.09	90	0.000	40.700	12.6	9.5
21	9325	6.0	3.63	90	0.250	38.762	12.7	9.5

APPENDIX B: Derivation of Support Vector Machines

The classification problem is to separate the dataset into its classes with a hyperplane, as expressed in equation B.1.

$$f(x) = \langle \mathbf{w}, \mathbf{x} \rangle + b \quad (B.1)$$

$$D = \{(x^1, y^1), \dots, (x^l, y^l)\} \text{ and } x \in \mathbb{R}^n, y \in \{-1, 1\}$$

In equation B.1, $\mathbf{w}$ is the weight vector, b is the bias and l is the number of training data. If b is assumed as zero, the data points corresponding to $\mathbf{w}^T \mathbf{x} = 0$ exist perpendicular to $\mathbf{w}$ while intersecting the origin. This condition can be described with a line in 2-D and a plane in 3-D, which are called a hyperplane. The bias term's geometrical implication is the interpretation of the distance between the hyperplane and the origin. The hyperplane equation can be given as in equation B.2.

$$\{x: f(\mathbf{x}) = \mathbf{w}^T \mathbf{x} + b = 0\} \quad (B.2)$$

Equation B.2 divides the sample space into two parts. $f(\mathbf{x})$ is the linear discriminant function. The sign of $f(\mathbf{x})$ represents the part of the divided space. It is possible to state that the data points are optimally separated, and the margin is maximum while assuming that the error is zero. It is suitable to implement an optimal canonical separating hyperplane (OCSH) to eliminate any generality loss. Hence, the constraints can be described as in equation B.3.

$$\min_i |\langle \mathbf{w}, \mathbf{x}_i \rangle + b| = 1 \quad (B.3)$$

The constraint in equation B.3 makes the formulation simpler, which states that the norm or the measure of the weight vector should be equal to the inverse of the distance between the hyperplane and the closest data point(s).

In terms of the geometrical approach, an OCSH must compensate the constraints in equation B.4.

$$y^i [\langle \mathbf{w}, \mathbf{x}_i \rangle + b] \geq 1 \quad (B.4)$$

$$(i = 1, \dots, l)$$

The distance from the hyperplane can be expressed as in equation B.5.

$$d(\mathbf{w}, b; x) = \frac{|\langle \mathbf{w}, \mathbf{x}_i \rangle + b|}{\|\mathbf{w}\|} \quad (B.5)$$

As stated previously, an OCSH maximizes the margin regarding the constraints in equation B.4. The margin can be calculated via equations B.6, B.7, B.8, and B.9.

$$M = \min_{x_i: y_i = -1} d(\mathbf{w}, b; x_i) + \min_{x_i: y_i = 1} d(\mathbf{w}, b; x_i) \quad (B.6)$$

$$M = \min_{x_i: y_i = -1} \frac{|\langle \mathbf{w}, \mathbf{x}_i \rangle + b|}{\|\mathbf{w}\|} + \min_{x_i: y_i = 1} \frac{|\langle \mathbf{w}, \mathbf{x}_i \rangle + b|}{\|\mathbf{w}\|} \quad (B.7)$$

$$M = \frac{1}{\|\mathbf{w}\|} \left(\min_{x_i: y_i = -1} |\langle \mathbf{w}, \mathbf{x}_i \rangle + b| + \min_{x_i: y_i = 1} |\langle \mathbf{w}, \mathbf{x}_i \rangle + b| \right) \quad (\text{B.8})$$

$$M = \frac{2}{\|\mathbf{w}\|} \quad (\text{B.9})$$

Hence, the OCSH minimizes equation B.10.

$$\Phi(\mathbf{w}) = \frac{1}{2} \|\mathbf{w}\|^2 \quad (\text{B.10})$$

Equation B.10 does not involve the bias term since the bias in equation B.4 does not change the value of margin but only affects the hyperplane by shifting it within the normal direction. Equation B.10 can be minimized within the constraints in equation B.4 by implementing Lagrangian multipliers as in equation B.11.

$$\Phi(\mathbf{w}, b, \alpha) = \frac{1}{2} \|\mathbf{w}\|^2 - \sum_{i=1}^l \alpha_i (y_i [\langle \mathbf{w}, \mathbf{x}_i \rangle + b] - 1) \quad (\text{B.11})$$

In equation B.11, α is the Lagrangian multiplier, minimized for $\mathbf{w}$ and b while being maximized with $\alpha \geq 0$. Equation B.11 can be transformed into a dual problem for simplicity as in equation B.12.

$$\max_{\alpha} W(\alpha) = \max_{\alpha} \left(\min_{\mathbf{w}, b} \Phi(\mathbf{w}, b, \alpha) \right) \quad (\text{B.12})$$

The minimization process with respect to $\mathbf{w}$ and b is shown in equation B.13.

$$\begin{aligned} \frac{\partial \Phi}{\partial b} = 0 &\rightarrow \sum_{i=1}^l \alpha_i y_i = 0 \\ \frac{\partial \Phi}{\partial \mathbf{w}} = 0 &\rightarrow \sum_{i=1}^l \alpha_i y_i \mathbf{x}_i = \mathbf{w} \end{aligned} \quad (\text{B.13})$$

By merging equations B.12 and B.13, the primal problem is converged into the dual problem as in equation B.14.

$$\max_{\alpha} W(\alpha) = \max_{\alpha} -\frac{1}{2} \sum_{i=1}^l \sum_{j=1}^l \alpha_i \alpha_j y_i y_j \langle \mathbf{x}_i, \mathbf{x}_j \rangle + \sum_{k=1}^l \alpha_k \quad (\text{B.14})$$

The solution for equation B.14 is given in equation B.15 with the constraints in equation B.16.

$$\alpha^* = \arg \min_a \frac{1}{2} \sum_{i=1}^l \sum_{j=1}^l \alpha_i \alpha_j y_i y_j \langle \mathbf{x}_i, \mathbf{x}_j \rangle - \sum_{k=1}^l \alpha_k \quad (\text{B.15})$$

$$\begin{aligned} \sum_{j=1}^l \alpha_j y_j &= 0 \\ \alpha_i &\geq 0 \quad (i = 1, \dots, l) \end{aligned} \quad (\text{B.16})$$

Equations B.13 and B.16 are called Karush-Kuhn-Tucker (KKT) conditions. Applying the conditions in equation B.16 to equation B.15 and solving the problem provide the OCSH as given in equations B.17 and B.18.

$$\mathbf{w}^* = \sum_{i=1}^l \alpha_i y_i \mathbf{x}_i \quad (\text{B.17})$$

$$b^* = -\frac{1}{2} \langle \mathbf{w}^*, \mathbf{x}_{\text{class1}} + \mathbf{x}_{\text{class2}} \rangle \quad (\text{B.18})$$

In equation B.18, $\mathbf{x}_{\text{class1}}$ and $\mathbf{x}_{\text{class2}}$ are the support vector for designated classes.

Generally, classifiers can be defined in two subgroups: Hard margin classifiers and soft margin classifiers, which are given in equations B.19 and B.20, respectively.

$$f(\mathbf{x}) = \text{sgn}(\langle \mathbf{w}^*, \mathbf{x} \rangle + b) \quad (\text{B.19})$$

$$f(\mathbf{x}) = h_{\text{VC}}(\langle \mathbf{w}^*, \mathbf{x} \rangle + b) \text{ where } h(z) = \begin{cases} -1, & z < -1 \\ z, & -1 \leq z \leq 1 \\ +1, & z > 1 \end{cases} \quad (\text{B.20})$$

In equation B.20, h_{VC} is Vapnik-Chervonenkis (VC) dimension, which is a scalar value that can be used the bounds of any hyperplane in machine learning. VC dimension can be expressed via equation B.21.

$$h_{\text{VC}} \leq \min(R^2 A^2, n) + 1 \quad (\text{B.21})$$

In equation B.21, R is the radius of the sphere encapsulating the data points, A is the holding bound and n is the degree of dimension.

Soft margin classifiers generate an output of a real number within the range of -1 and 1 when the hyperplane is inquired within the margin, where there is not any training data. Therefore, it is recommended to use soft margin classifiers rather than hard margin classifiers.

An additional KKT condition is given in equation B.22.

$$\alpha_i [y_i (\langle \mathbf{w}, \mathbf{x}_i \rangle + b) - 1] = 0 \quad (i = 1, \dots, l) \quad (\text{B.22})$$

The only data points $\mathbf{x}_i$ having non-zero Lagrangian multipliers are called support vectors (SVs), which satisfy equation B.23.

$$y_i (\langle \mathbf{w}, \mathbf{x}_i \rangle + b) = 1 \quad (\text{B.23})$$

SVs are placed through the margin if the data is linearly separable. Additionally, the hyperplane is generated via a small number of the training data, while other data points are ignored. That situation yields equation B.24 valid.

$$\|\mathbf{w}\|^2 = \sum_{i=1}^l \alpha_i = \sum_{i \in \text{SV}} \alpha_i = \sum_{i \in \text{SV}} \sum_{j \in \text{SV}} \alpha_i \alpha_j y_i y_j \langle \mathbf{x}_i, \mathbf{x}_j \rangle \quad (\text{B.24})$$

This ideal situation is valid for the linearly-separable data. Otherwise, to determine the optimum hyperplane, noise in the data should be estimated. To make the hyperplane separate the data correctly, a penalty function with non-negative variables defined in equation B.25 should be implemented in the training phase.

$$F(\xi) = \sum_i \xi_i \quad (\text{B.25})$$

In equation B.25, ξ_i are called slack variables that represent errors due to misclassification. By implementing the slack variables, the optimization problem is transformed into minimizing both the bound of the VC dimension and the misclassification error at the same time. The constraints in equation B.4 should be updated as in equation B.26 to perform this optimization.

$$\begin{aligned} y_i [\langle \mathbf{w}, \mathbf{x}_i \rangle + b] &\geq 1 - \xi_i \\ (i = 1, \dots, l), \xi_i &\geq 0 \end{aligned} \quad (\text{B.26})$$

Thus, the generalized OCSH can be estimated via the weight vector that minimizes equation B.27.

$$\Phi(\mathbf{w}, \xi) = \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_i \xi_i \quad (\text{B.27})$$

In equation B.27, the complexity, or the penalty parameter (C) is a user-defined scalar value used to modify the margin to minimize the slack variables. It should be noted that a high value of C may be resulted in a high penalty. Similarly, a low value of C may cause misclassification.

The goal is finding a solution to equation B.27 with the constraints in equation B.26. The solution is given in equation B.28.

$$\begin{aligned} \Phi(\mathbf{w}, b, \alpha, \xi, \beta) &= \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_i \xi_i \\ &\quad - \sum_{i=1}^l \alpha_i (y_i [\langle \mathbf{w}, \mathbf{x}_i \rangle + b] - 1 + \xi_i) - \sum_{i=1}^l \beta_i \xi_i \end{aligned} \quad (\text{B.28})$$

In equation B.28, α and β are the Lagrangian multipliers. The aim is given in equation B.29 in terms of the dual problem.

$$\max_{\alpha} W(\alpha, \beta) = \max_{\alpha, \beta} \left(\min_{\mathbf{w}, b, \xi} \Phi(\mathbf{w}, b, \alpha, \xi, \beta) \right) \quad (\text{B.29})$$

The minimization process with respect to $\mathbf{w}$, b , and ξ is shown in equation B.30.

$$\begin{aligned} \frac{\partial \Phi}{\partial b} &= 0 \rightarrow \sum_{i=1}^l \alpha_i y_i = 0 \\ \frac{\partial \Phi}{\partial \mathbf{w}} &= 0 \rightarrow \sum_{i=1}^l \alpha_i y_i \mathbf{x}_i = \mathbf{w} \\ \frac{\partial \Phi}{\partial \xi} &= 0 \rightarrow \alpha_i + \beta_i = C \end{aligned} \quad (\text{B.30})$$

By merging equations B.30, B.29, and B.28, the primal problem is converged into the dual problem as in equation B.31.

$$\max_{\alpha} W(\alpha) = \max_{\alpha} -\frac{1}{2} \sum_{i=1}^l \sum_{j=1}^l \alpha_i \alpha_j y_i y_j \langle \mathbf{x}_i, \mathbf{x}_j \rangle + \sum_{k=1}^l \alpha_k \quad (\text{B.31})$$

The solution for equation B.31 is given in equation B.32 with the constraints in equation B.33.

$$\alpha^* = \arg \min_{\alpha} \frac{1}{2} \sum_{i=1}^l \sum_{j=1}^l \alpha_i \alpha_j y_i y_j \langle \mathbf{x}_i, \mathbf{x}_j \rangle - \sum_{k=1}^l \alpha_k \quad (\text{B.32})$$

$$\begin{aligned} \sum_{j=1}^l \alpha_j y_j &= 0 \\ 0 \leq \alpha_i &\leq C \quad (i=1, \dots, l) \end{aligned} \quad (\text{B.33})$$

Like the SVM for classification, the optimal regression can be performed via minimizing equation B.34.

$$\Phi(\mathbf{w}, \xi) = \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_i (\xi_i^-, \xi_i^+) \quad (\text{B.34})$$

In equation B.34, ξ^- and ξ^+ are lower and upper bounds of the slack variables. The ε -insensitive LF can be expressed as in equation B.35.

$$L_{\varepsilon}(y) = \begin{cases} 0, & |f(\mathbf{x}) - y| < \varepsilon \\ |f(\mathbf{x}) - y| - \varepsilon, & \text{otherwise} \end{cases} \quad (\text{B.35})$$

Equation B.35 can be implemented to the solution via equation B.36.

$$\begin{aligned} \bar{\alpha}, \bar{\alpha}^* &= \arg \min_{\alpha, \alpha^*} \frac{1}{2} \sum_{i=1}^l \sum_{j=1}^l (\alpha_i - \alpha_i^*) (\alpha_j - \alpha_j^*) \langle \mathbf{x}_i, \mathbf{x}_j \rangle - \sum_{i=1}^l (\alpha_i - \alpha_i^*) y_i \\ &\quad + \sum_{i=1}^l (\alpha_i - \alpha_i^*) \varepsilon \end{aligned} \quad (\text{B.36})$$

Equation B.36 is solved within the constraints in equation B.37.

$$\begin{aligned} \sum_{i=1}^l (\alpha_i - \alpha_i^*) &= 0 \\ 0 \leq \alpha_i, \alpha_i^* &\leq C \quad (i=1, \dots, l) \end{aligned} \quad (\text{B.37})$$

Hence, the solution becomes as in equations B.38 and B.39.

$$\bar{\mathbf{w}} = \sum_{i=1}^l (\alpha_i - \alpha_i^*) \mathbf{x}_i \quad (\text{B.38})$$

$$\bar{b} = -\frac{1}{2} \langle \bar{\mathbf{w}}, \mathbf{x}_{\text{class1}} + \mathbf{x}_{\text{class2}} \rangle \quad (\text{B.39})$$

The solution must meet the KKT conditions in equation B.40.

$$\bar{\alpha}_i \bar{\alpha}_i^* = 0 \quad (i=1, \dots, l) \quad (\text{B.40})$$

As mentioned, the SVs are non-zero Lagrangian multipliers. Therefore, ε is set to zero. The optimization problem is simplified into equation B.41 with the constraints in equation B.42.

$$\min_{\beta} \frac{1}{2} \sum_{i=1}^l \sum_{j=1}^l \beta_i \beta_j \langle \mathbf{x}_i, \mathbf{x}_j \rangle - \sum_{i=1}^l \beta_i y_i \quad (\text{B.41})$$

$$\sum_{i=1}^l \beta_i = 0 \quad (B.42)$$

$$-C \leq \beta_i \leq C \quad (i=1, \dots, l)$$

Finally, the regression function becomes as in equations B.43 and B.44.

$$\bar{\mathbf{w}} = \sum_{i=1}^l \beta_i \mathbf{x}_i \quad (B.43)$$

$$\bar{b} = -\frac{1}{2} \langle \bar{\mathbf{w}}, \mathbf{x}_{\text{class1}} + \mathbf{x}_{\text{class2}} \rangle \quad (B.44)$$

In ε -SVR, each data point involves an error while the values outside of the ε envelope are regulated by the penalty parameter. Nevertheless, it is not simple to estimate the optimum ε . It is recommended to use a user-defined controller parameter (ν) to modify the training model by adjusting the amount of ε . In ν -SVR, the optimization problem is expressed in equation B.45 with the constraints in equation B.46.

$$\Phi(\mathbf{w}, \xi, \varepsilon) = \frac{1}{2} \|\mathbf{w}\|^2 + C \left(\nu \varepsilon + \frac{1}{l} \sum_{i=1}^l (\xi_i + \xi_i^*) \right) \quad (B.45)$$

$$\begin{aligned} (\langle \mathbf{w}, \mathbf{x}_i \rangle + b) - y_i &\leq \varepsilon + \xi_i \\ y_i - (\langle \mathbf{w}, \mathbf{x}_i \rangle + b) &\leq \varepsilon + \xi_i^* \\ \varepsilon &\geq 0, \xi_i \geq 0 \end{aligned} \quad (B.46)$$

The solution can be expressed explicitly as in equation B.47.

$$\begin{aligned} \Phi(\mathbf{w}, b, \alpha, \beta, \eta, \xi, \varepsilon) &= \frac{1}{2} \|\mathbf{w}\|^2 + C \nu \varepsilon + \frac{C}{l} \sum_{i=1}^l (\xi_i + \xi_i^*) \\ &\quad - \sum_{i=1}^l \alpha_i (\xi_i + y_i - \langle \mathbf{w}, \mathbf{x}_i \rangle - b + \varepsilon) \\ &\quad - \sum_{i=1}^l \alpha_i^* (\xi_i^* + y_i - \langle \mathbf{w}, \mathbf{x}_i \rangle - b + \varepsilon) - \beta \varepsilon \\ &\quad - \sum_{i=1}^l (\eta_i \xi_i + \eta_i^* \xi_i^*) \end{aligned} \quad (B.47)$$

In equation B.47, α , β , and η are the Lagrangian multipliers. If the derivatives of the primal variables ($\mathbf{w}$, ε , b , ξ) are set to zero, the dual problem yields to equation B.48.

$$\begin{aligned} \mathbf{w} &= \sum_{i=1}^l (\alpha_i^* - \alpha_i) \mathbf{x}_i \\ C \nu - \sum_{i=1}^l (\alpha_i^* + \alpha_i) - \beta &= 0 \end{aligned} \quad (B.48)$$

Finally, ν -SVR optimization can be performed via maximizing equation B.49 under the constraints in equation B.50.

$$W(\alpha) = -\frac{1}{2} \sum_{i,j=1}^l (\alpha_i - \alpha_i^*) \langle \mathbf{x}_i, \mathbf{x}_j \rangle (\alpha_j - \alpha_j^*) - \sum_{i=1}^l (\alpha_i - \alpha_i^*) y_i \quad (B.49)$$

$$\begin{aligned}
\sum_{i=1}^l (\alpha_i - \alpha_i^*) &= 0 \\
\sum_{i=1}^l (\alpha_i + \alpha_i^*) &= C\nu \\
0 \leq \alpha &\leq \frac{C}{l}
\end{aligned}
\tag{B.50}$$





APPENDIX C: R Script for the Calculations

R Script

```
#####
#R Script Source Code for PhD Dissertation#
#Author: Korhan Kor                               #
#Version: v2.1                                     #
#####

#####
#Plotting Correlogram#
#####

#Loading libraries
library(readxl)
library(psych)

#Loading dataset
input <- read_excel("Data.xlsx")
dataset <- data.frame(input)
summary(dataset)

#Correlogram settings
pairs.panels(dataset,smooth = FALSE,scale = FALSE,density = TRUE,ellipses =
FALSE,method = "pearson",pch = 21,lm = FALSE,cor = TRUE,jiggle = FALSE,factor
= 2,hist.col = 8,stars = TRUE,ci = FALSE)

#####
#Random Forests Feature Importance#
#####

#Loading libraries
library(readxl)
library(mlbench)
library(caret)

#Loading master dataset
setwd("E:/Kovan/PhD/Progress_Reports/6/Data")
input <- read_excel("Data.xlsx")
dataset <- data.frame(input)
summary(dataset)

#Training the random forest model
control <- trainControl(method="repeatedcv", number=10, repeats=3)
model <- train(y~., data=dataset, method="rf", preProcess="scale",
trControl=control)

#Estimating variable importance
importance <- varImp(model, scale=FALSE)
print(importance)
plot(importance)
results <- as.vector(importance$importance)
pp <- results/sum(results)
write.csv(results$importance, file="importance.csv")
write.csv(pp, file="percentage.csv")

#####
#Main Script for the Calculations#
#####

#Loading libraries
library(readxl)
library(e1071)
library(hydroGOF)
library(PerformanceAnalytics)
library(caTools)
library(grDevices)
library(neuralnet)
library(RColorBrewer)
```

```

#Loading master dataset
input <- read_excel("Data.xlsx", sheet = 2)
dataset <- data.frame(input)
summary(dataset)

#Dividing dataset into train and test subsets
set.seed(101)
sample = sample.split(dataset, SplitRatio = 0.75)
dataset.train = subset(dataset, sample == TRUE)
dataset.test = subset(dataset, sample == FALSE)
summary(dataset.train)
summary(dataset.test)

#Saving data to file
write.csv(dataset.train, file = "data_train.csv")
write.csv(dataset.test, file = "data_test.csv")

#Calculating total sum of squares
tsstrain <- sum((dataset.train$y-mean(dataset.train$y))^2)
tsstest <- sum((dataset.test$y-mean(dataset.test$y))^2)

#Multiple Regression
reg <- lm(y ~., data = dataset.train)
print(summary(reg))

#Training performance for MLR
reg.predtrain <- predict(reg,newdata = dataset.train)
write.csv(reg.predtrain, file = "reg_predtrain.csv")
reg.rsstrain <- sum((dataset.train$y-reg.predtrain)^2)
reg.r2train <- 1-(reg.rsstrain/tsstrain)
reg.rmsetrain <- rmse(reg.predtrain,dataset.train$y)
reg.cortrain <- cor(reg.predtrain,dataset.train$y)
reg.traincortext <- signif(reg.cortrain, digits = 2)

#Testing performance for MLR
reg.predtest <- predict(reg,newdata = dataset.test)
write.csv(reg.predtest, file = "reg_predtest.csv")
reg.rsstest <- sum((dataset.test$y-reg.predtest)^2)
reg.rmsetest <- rmse(reg.predtest,dataset.test$y)
reg.cortest <- cor(reg.predtest,dataset.test$y)
reg.r2test <- 1-(reg.rsstest/tsstest)
reg.cortext <- signif(reg.cortest, digits = 2)

#Setting up k-fold cross validation
tc <- tune.control(cross = 10)

#e-SVR w/ linear kernel
obj.esvr1 <- tune.svm(y ~.,data = dataset.train, type = "eps-regression",
kernel = "linear", cost = seq(from=1.0,to=10.0,by=0.5), epsilon =
seq(from=0.1,to=1.0,by=0.05), tolerance=0.001, tunecontrol = tc)
print(obj.esvr1)
k.esvr1 <- which.min(obj.esvr1$performances[,3])
c.esvr1 <- obj.esvr1$performances[,1][k.esvr1]
e.esvr1 <- obj.esvr1$performances[,2][k.esvr1]
kk.esvr1 <- min(obj.esvr1$performances[,3])
aa.esvr1 <- paste("Best performance: ", signif(kk.esvr1, digits = 4))
esvr1 <- svm(y ~.,data = dataset.train, scale = T, type = "eps-regression",
kernel = "linear", cost = c.esvr1, epsilon =
e.esvr1,tolerance=0.001,shrinking=T,fitted=T)
print(esvr1)
write.csv(obj.esvr1$performances, file = "eSVR_CV.csv")

#Training performance for e-SVR linear
esvr1.predtrain <- predict(esvr1,newdata = dataset.train)
write.csv(esvr1.predtrain, file = "esvr1_predtrain.csv")
esvr1.rsstrain <- sum((dataset.train$y-esvr1.predtrain)^2)
esvr1.r2train <- 1-(esvr1.rsstrain/tsstrain)
esvr1.rmsetrain <- rmse(esvr1.predtrain,dataset.train$y)
esvr1.cortrain <- cor(esvr1.predtrain,dataset.train$y)
esvr1.traincortext <- signif(esvr1.cortrain, digits = 2)

#Testing performance for e-SVR linear

```

```

esvr1.predtest <- predict(esvr1,newdata = dataset.test)
write.csv(esvr1.predtest, file = "esvr1_predtest.csv")
esvr1.rsstest <- sum((dataset.test$y-esvr1.predtest)^2)
esvr1.rmsetest <- rmse(esvr1.predtest,dataset.test$y)
esvr1.cortest <- cor(esvr1.predtest,dataset.test$y)
esvr1.r2test <- 1-(esvr1.rsstest/tsstest)
esvr1.cortext <- signif(esvr1.cortest, digits = 2)

#e-SVR w/ RBF kernel
obj.esvr2 <- tune.svm(y ~.,data = dataset.train, type = "eps-regression",
kernel = "radial", cost = seq(from=1.0,to=10.0,by=0.5), epsilon =
seq(from=0.1,to=1.0,by=0.1), gamma = seq(from=0,to=10,by=1) ,tolerance=0.001,
tunecontrol = tc)
print(obj.esvr2)
k.esvr2 <- which.min(obj.esvr2$performances[,4])
g.esvr2 <- obj.esvr2$performances[,1][k.esvr2]
c.esvr2 <- obj.esvr2$performances[,2][k.esvr2]
e.esvr2 <- obj.esvr2$performances[,3][k.esvr2]
esvr2 <- svm(y ~., data = dataset.train, scale = T, type = "eps-regression",
kernel = "radial", cost = c.esvr2, epsilon = e.esvr2, gamma = g.esvr2,
tolerance=0.001,shrinking=T,fitted=T)
print(esvr2)

#Training performance for e-SVR RBF
esvr2.predtrain <- predict(esvr2,newdata = dataset.train)
write.csv(esvr2.predtrain, file = "esvr2_predtrain.csv")
esvr2.rsstrain <- sum((dataset.train$y-esvr2.predtrain)^2)
esvr2.r2train <- 1-(esvr2.rsstrain/tsstrain)
esvr2.rmsetrain <- rmse(esvr2.predtrain,dataset.train$y)
esvr2.cortrain <- cor(esvr2.predtrain,dataset.train$y)
esvr2.traincortext <- signif(esvr2.cortrain, digits = 2)

#Testing performance for e-SVR RBF
esvr2.predtest <- predict(esvr2,newdata = dataset.test)
write.csv(esvr2.predtest, file = "esvr2_predtest.csv")
esvr2.rsstest <- sum((dataset.test$y-esvr2.predtest)^2)
esvr2.rmsetest <- rmse(esvr2.predtest,dataset.test$y)
esvr2.cortest <- cor(esvr2.predtest,dataset.test$y)
esvr2.r2test <- 1-(esvr2.rsstest/tsstest)
esvr2.cortext <- signif(esvr2.cortest, digits = 2)

#n-SVR with linear kernel
obj.nsvr1 <- tune.svm(y ~.,data = dataset.train, type = "nu-regression",
kernel = "linear", cost = seq(from=1.0,to=10.0,by=0.5), nu =
seq(from=0.1,to=1.0,by=0.05), tolerance=0.001, tunecontrol = tc)
print(obj.nsvr1)
k.nsvr1 <- which.min(obj.nsvr1$performances[,3])
c.nsvr1 <- obj.nsvr1$performances[,1][k.nsvr1]
n.nsvr1 <- obj.nsvr1$performances[,2][k.nsvr1]
kk.nsvr1 <- min(obj.nsvr1$performances[,3])
aa.nsvr1 <- paste("Best performance: ", signif(kk.nsvr1, digits = 4))
nsvr1 <- svm(y ~.,data = dataset.train, scale = T, type = "nu-regression",
kernel = "linear", cost = c.nsvr1, nu =
n.nsvr1,tolerance=0.001,shrinking=T,fitted=T)
print(nsvr1)
write.csv(obj.nsvr1$performances, file = "nSVR_CV.csv")

#Training performance for n-SVR linear
nsvr1.predtrain <- predict(nsvr1,newdata = dataset.train)
write.csv(nsvr1.predtrain, file = "nsvr1_predtrain.csv")
nsvr1.rsstrain <- sum((dataset.train$y-nsvr1.predtrain)^2)
nsvr1.r2train <- 1-(nsvr1.rsstrain/tsstrain)
nsvr1.rmsetrain <- rmse(nsvr1.predtrain,dataset.train$y)
nsvr1.cortrain <- cor(nsvr1.predtrain,dataset.train$y)
nsvr1.traincortext <- signif(nsvr1.cortrain, digits = 2)

#Testing performance for n-SVR linear
nsvr1.predtest <- predict(nsvr1,newdata = dataset.test)
write.csv(nsvr1.predtest, file = "nsvr1_predtest.csv")
nsvr1.rsstest <- sum((dataset.test$y-nsvr1.predtest)^2)
nsvr1.rmsetest <- rmse(nsvr1.predtest,dataset.test$y)
nsvr1.cortest <- cor(nsvr1.predtest,dataset.test$y)

```

```

nsvr1.r2test <- 1-(nsvr1.rsstest/tsstest)
nsvr1.cortext <- signif(nsvr1.cortest, digits = 2)

#n-SVR with RBF kernel
obj.nsvr2 <- tune.svm(y ~.,data = dataset.train, type = "nu-regression",
kernel = "radial", cost = seq(from=1,to=10.0,by=0.5), nu =
seq(from=0.1,to=1.0,by=0.1), gamma = seq(from=0,to=10,by=1) ,tolerance=0.001,
tunecontrol = tc)
print(obj.nsvr2)
k.nsvr2 <- which.min(obj.nsvr2$performances[,4])
g.nsvr2 <- obj.nsvr2$performances[,1][k.nsvr2]
c.nsvr2 <- obj.nsvr2$performances[,2][k.nsvr2]
n.nsvr2 <- obj.nsvr2$performances[,3][k.nsvr2]
nsvr2 <- svm(y ~.,data = dataset.train, scale = T, type = "nu-regression",
kernel = "radial", cost = c.nsvr2, nu = n.nsvr2, gamma = g.nsvr2,
tolerance=0.001,shrinking=T,fitted=T)
print(nsvr2)

#Training performance for n-SVR RBF
nsvr2.predtrain <- predict(nsvr2,newdata = dataset.train)
write.csv(nsvr2.predtrain, file = "nsvr2_predtrain.csv")
nsvr2.rsstrain <- sum((dataset.train$y-nsvr2.predtrain)^2)
nsvr2.r2train <- 1-(nsvr2.rsstrain/tsstrain)
nsvr2.rmsetrain <- rmse(nsvr2.predtrain,dataset.train$y)
nsvr2.cortrain <- cor(nsvr2.predtrain,dataset.train$y)
nsvr2.traincortext <- signif(nsvr2.cortrain, digits = 2)

#Testing performance for n-SVR RBF
nsvr2.predtest <- predict(nsvr2,newdata = dataset.test)
write.csv(nsvr2.predtest, file = "nsvr2_predtest.csv")
nsvr2.rsstest <- sum((dataset.test$y-nsvr2.predtest)^2)
nsvr2.rmsetest <- rmse(nsvr2.predtest,dataset.test$y)
nsvr2.cortest <- cor(nsvr2.predtest,dataset.test$y)
nsvr2.r2test <- 1-(nsvr2.rsstest/tsstest)
nsvr2.cortext <- signif(nsvr2.cortest, digits = 2)

#Neural Network
ff <- ncol(dataset)
feats <- names(dataset[2:ff])
f <- paste(feats,collapse=' + ')
f <- paste('y ~',f)
f <- as.formula(f)
nn <- neuralnet(f,dataset.train,hidden=5,linear.output=FALSE)

#Training performance for ANN
ann.train <- compute(nn,dataset.train[2:ff])
ann.predict1 <- ann.train$net.result
ann.predtrain <- data.frame(ann.predict1)
write.csv(ann.predict1, file = "ann_predtrain.csv")
ann.rsstrain <- sum((dataset.train$y-ann.predtrain)^2)
ann.r2train <- 1-(ann.rsstrain/tsstrain)
ann.cortrain <- cor(ann.predtrain,dataset.train$y)
ann.traincortext <- signif(ann.cortrain, digits = 2)
ann.rmsetrain <- rmse(ann.predtrain,dataset.train$y)

#Testing performance for ANN
ann.test <- compute(nn,dataset.test[2:ff])
ann.predict2 <- ann.test$net.result
ann.predtest <- data.frame(ann.predict2)
write.csv(ann.predict2, file = "ann_predtest.csv")
ann.rsstest <- sum((dataset.test$y-ann.predtest)^2)
ann.r2test <- 1-(ann.rsstest/tsstest)
ann.cortest <- cor(ann.predtest,dataset.test$y)
ann.cortext <- signif(ann.cortest, digits = 2)
ann.rmsetest <- rmse(ann.predtest,dataset.test$y)

#Statistical comparison
plot.rss <- c(reg.rsstest, esvr1.rsstest, esvr2.rsstest, nsvr1.rsstest,
nsvr2.rsstest, ann.rsstest)
plot.rmse <- c(reg.rmsetest, esvr1.rmsetest, esvr2.rmsetest, nsvr1.rmsetest,
nsvr2.rmsetest, ann.rmsetest)

```

```
plot.cor <- c(reg.cortest, esvr1.cortest, esvr2.cortest, nsvr1.cortest,  
             nsvr2.cortest, ann.cortest)  
plot.r2 <- c(reg.r2test, esvr1.r2test, esvr2.r2test, nsvr1.r2test,  
            nsvr2.r2test, ann.r2test)  
plot.train <- c(reg.rmsetrain, esvr1.rmsetrain, esvr2.rmsetrain,  
              nsvr1.rmsetrain, nsvr2.rmsetrain, ann.rmsetrain)
```





APPENDIX D: Graphical Results of Intermediate Steps

D.1. Intermediate Steps for Case 2.3

D.1.1. Case 2.3.1: Default BYM

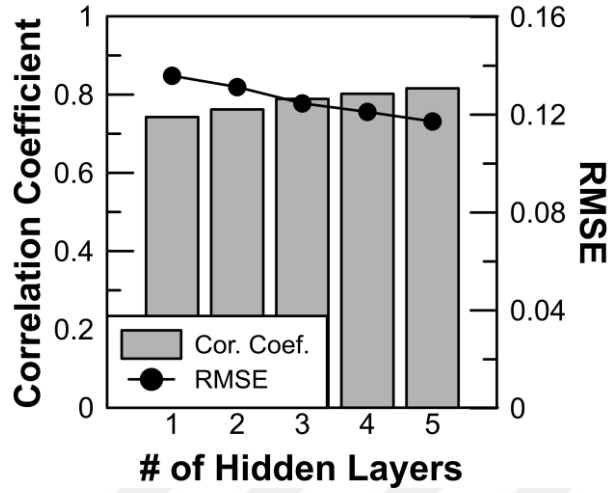


Figure D.1 : ANN hidden layer sensitivity for Case 2.3.1.

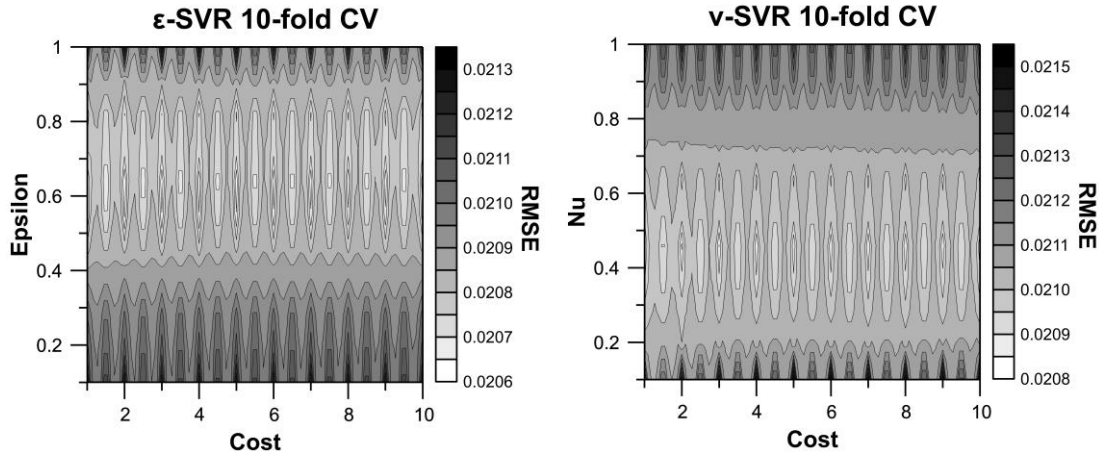


Figure D.2 : 10-fold CV grid search plots for Case 2.3.1.

Table D.1 : SVR hyper-parameters for Case 2.3.1.

ϵ -SVR					ν -SVR				
Linear		RBF			Linear		RBF		
C	ϵ	C	γ	ϵ	C	ν	C	γ	ν
2.5	0.65	3.0	1.0	0.15	1.0	0.45	3.0	1.0	1.0

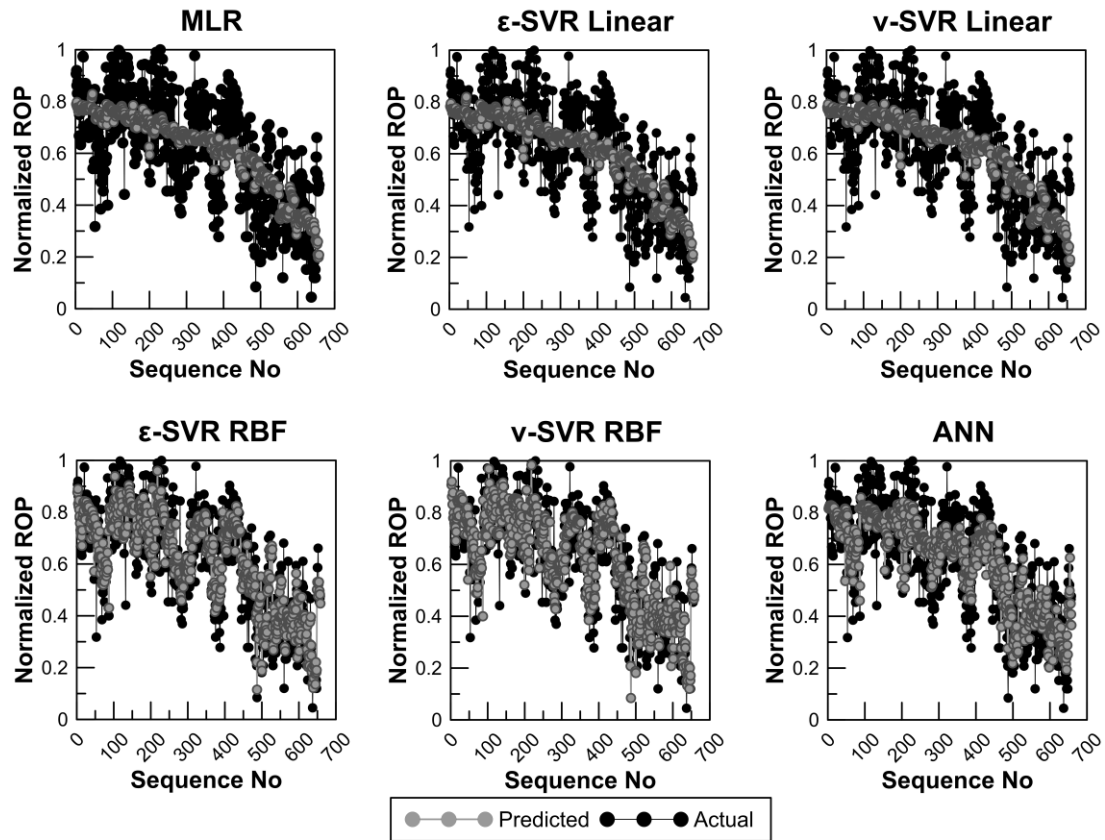


Figure D.3 : Model training for Case 2.3.1.

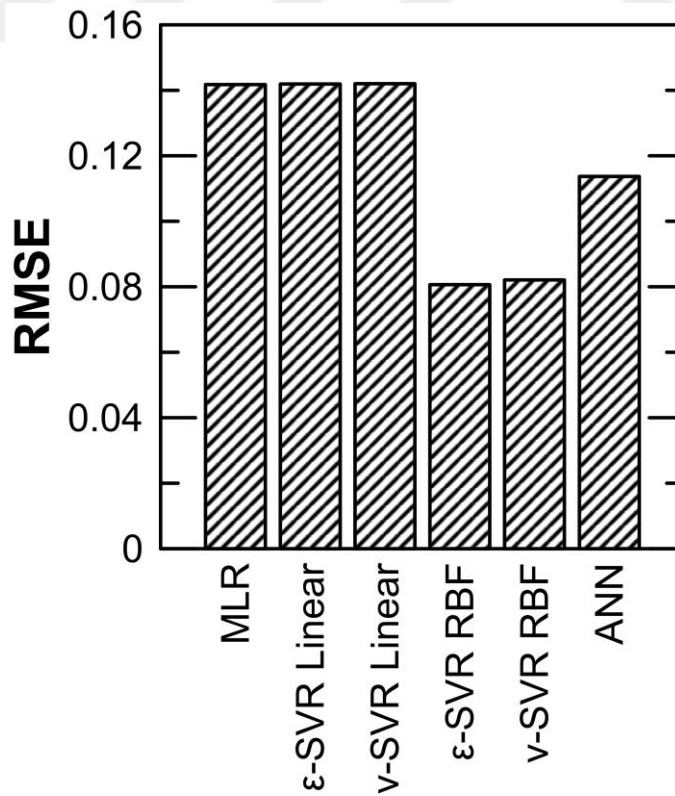


Figure D.4 : Training errors for Case 2.3.1.

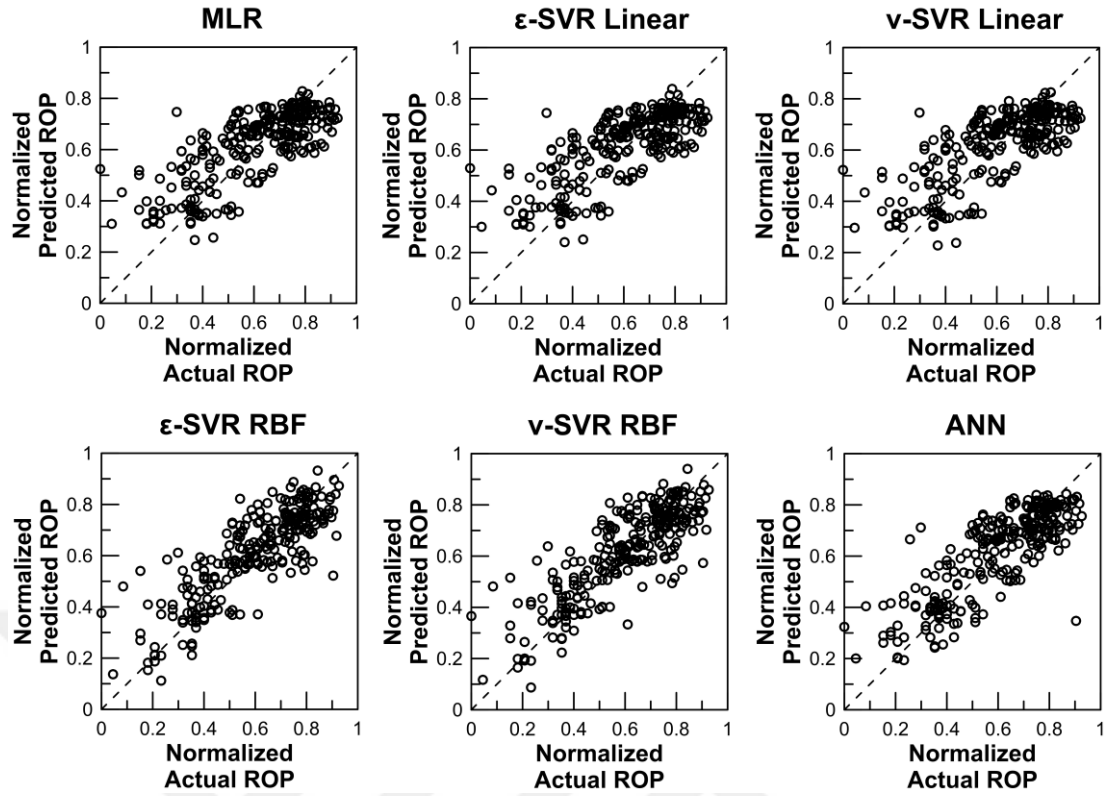


Figure D.5 : ROP predictions with testing subset for Case 2.3.1.

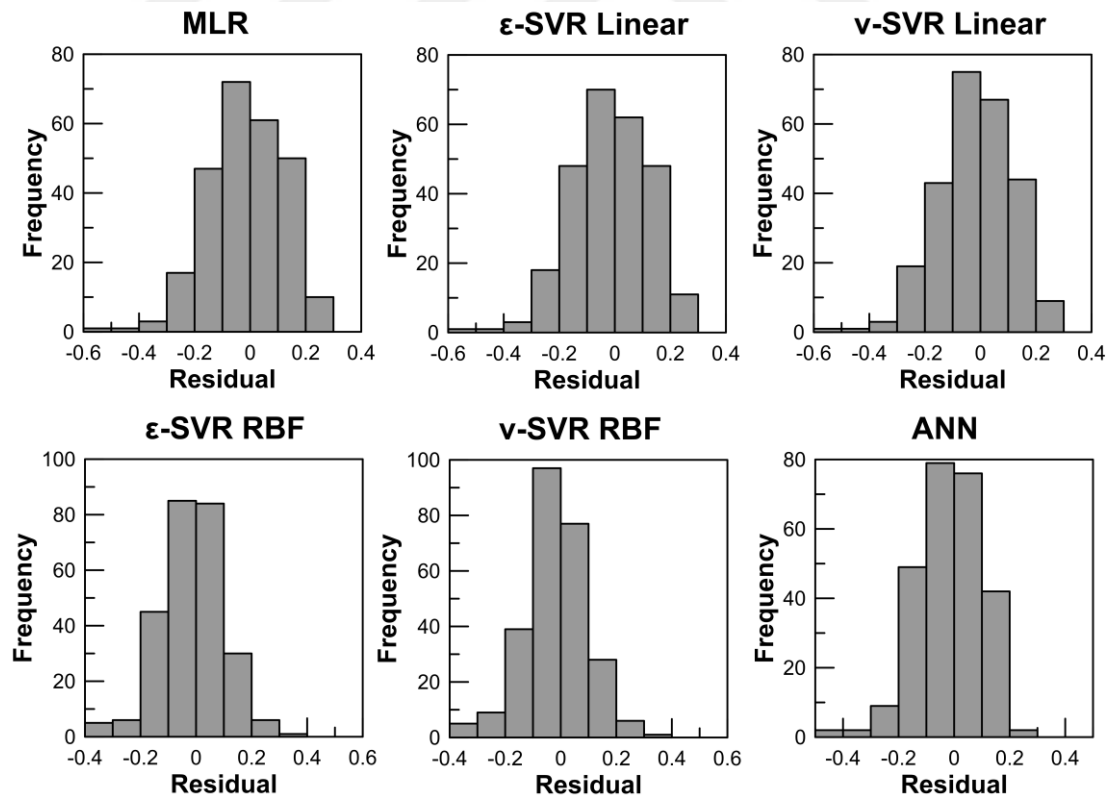


Figure D.6 : Error histograms for Case 2.3.1.

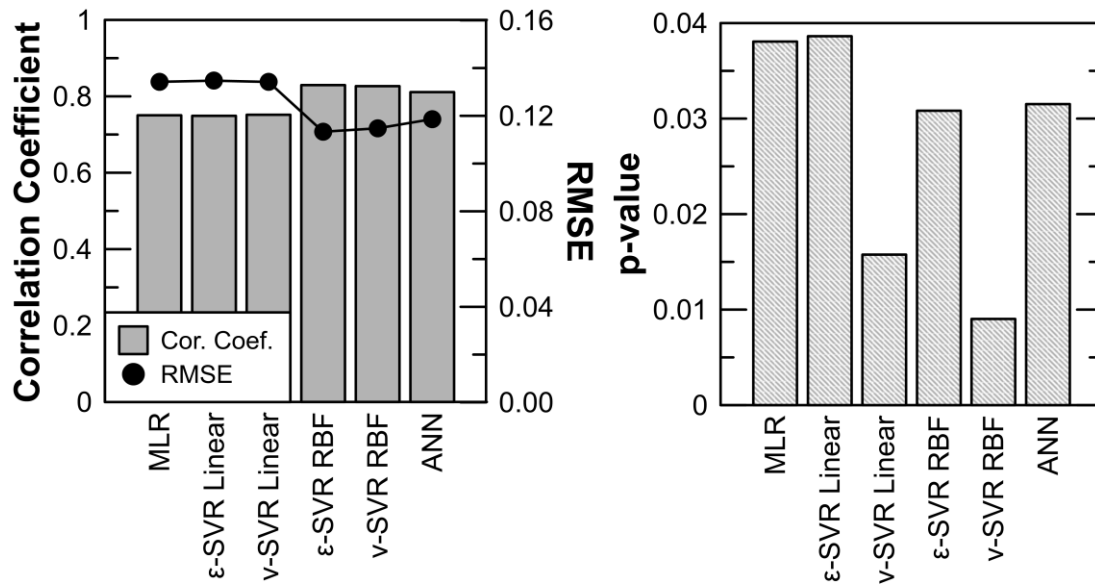


Figure D.7 : Statistical comparison for Case 2.3.1.

D.1.2. Case 2.3.2: BYM+SI

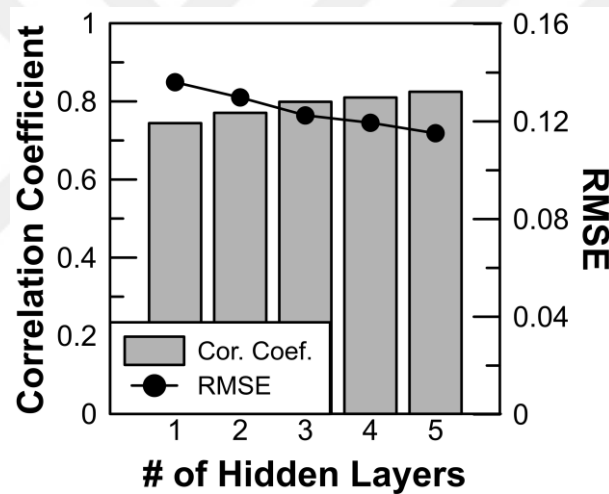


Figure D.8 : ANN hidden layer sensitivity for Case 2.3.2.

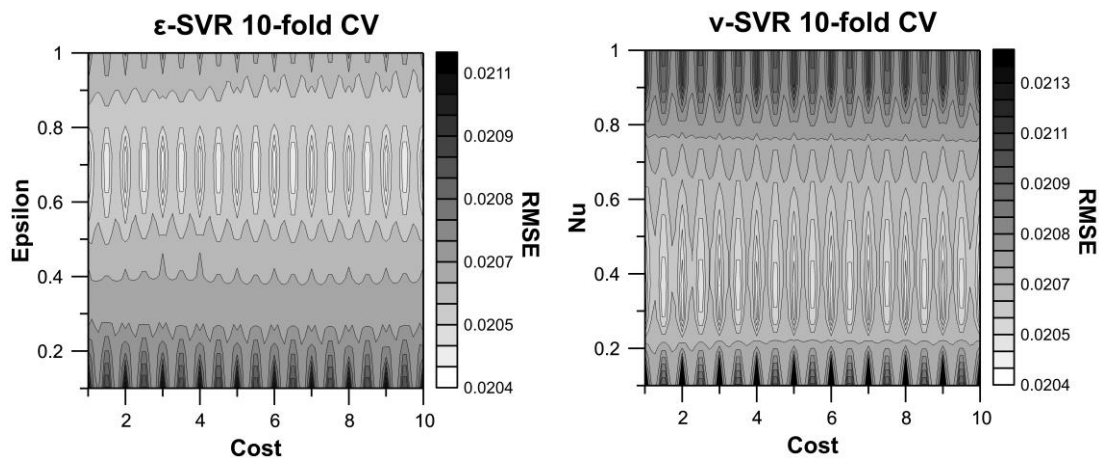


Figure D.9 : 10-fold CV grid search plots for Case 2.3.2.

Table D.2 : SVR hyper-parameters for Case 2.3.2.

ε -SVR					ν -SVR				
Linear			RBF		Linear			RBF	
C	ε	C	γ	ε	C	ν	C	γ	ν
1.0	0.7	2.0	1.0	0.1	6.5	0.3	1.5	1.0	0.5

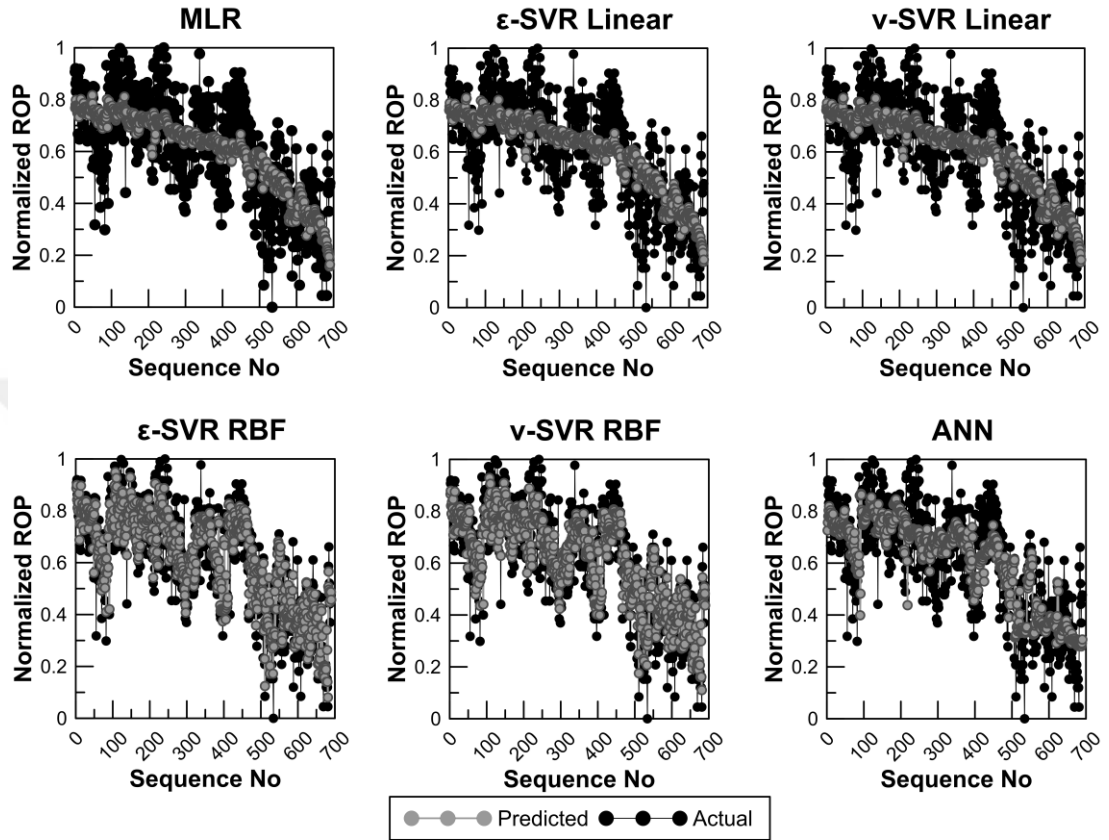


Figure D.10 : Model training for Case 2.3.2.

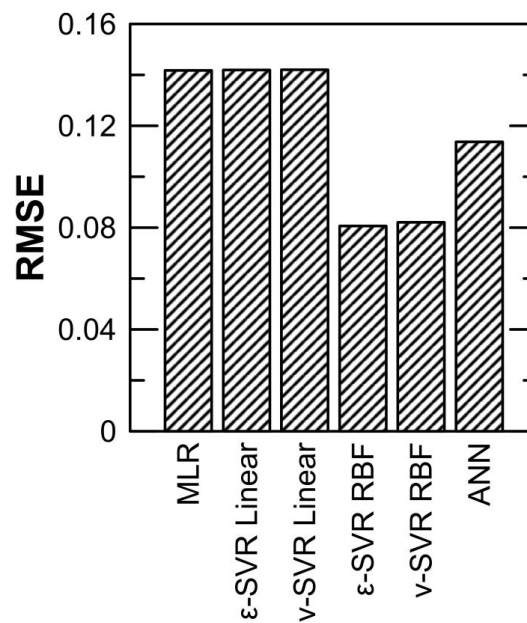


Figure D.11 : Training errors for Case 2.3.2.

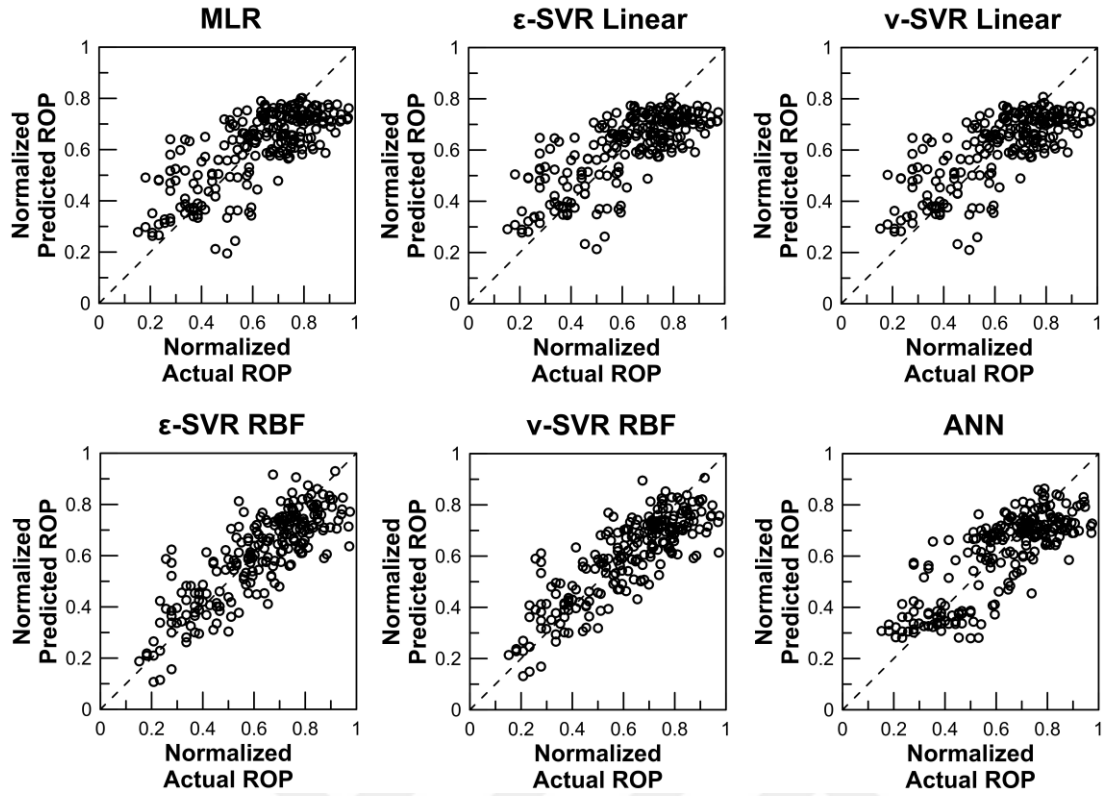


Figure D.12 : ROP predictions with testing subset for Case 2.3.2.

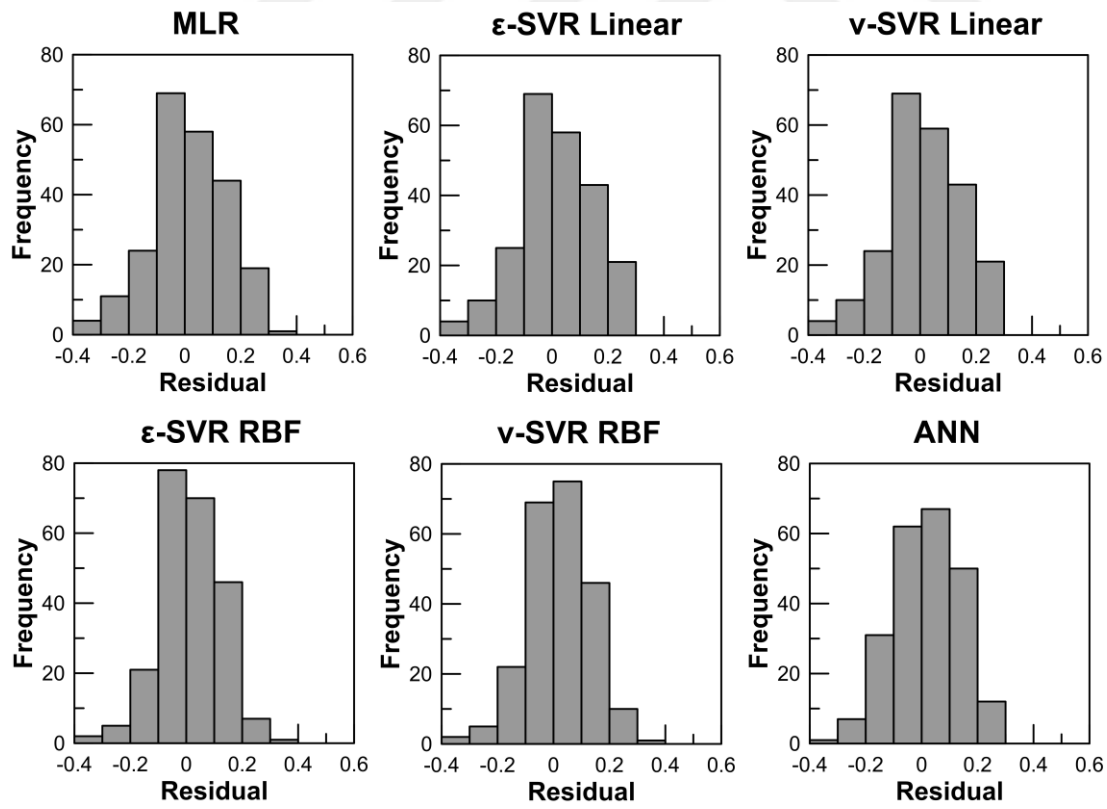


Figure D.13 : Error histograms for Case 2.3.2.

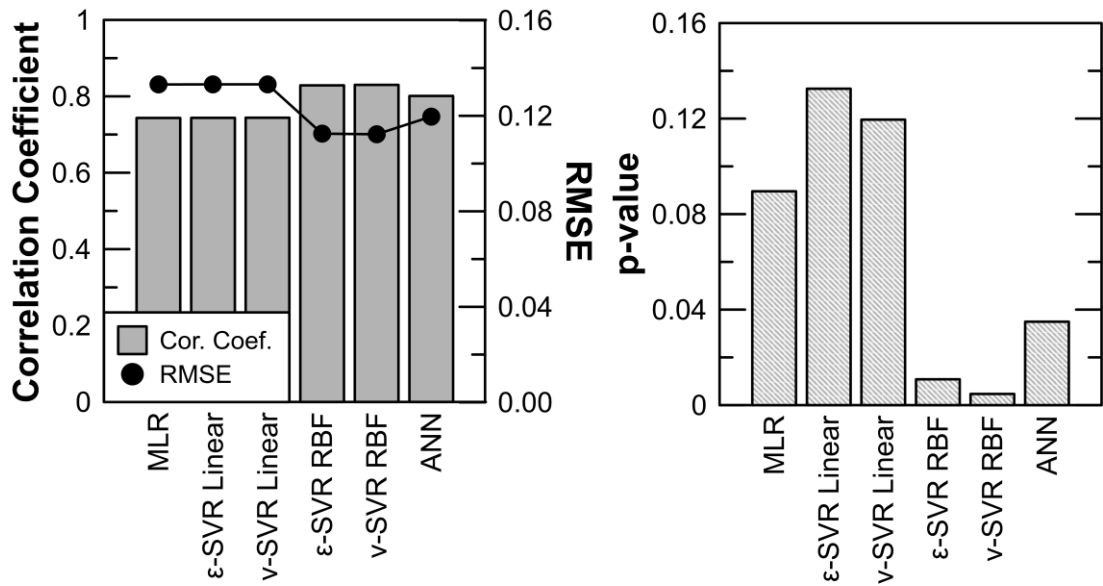


Figure D.14 : Statistical comparison for Case 2.3.2.

D.1.3. Case 2.3.3: BYM+UCS

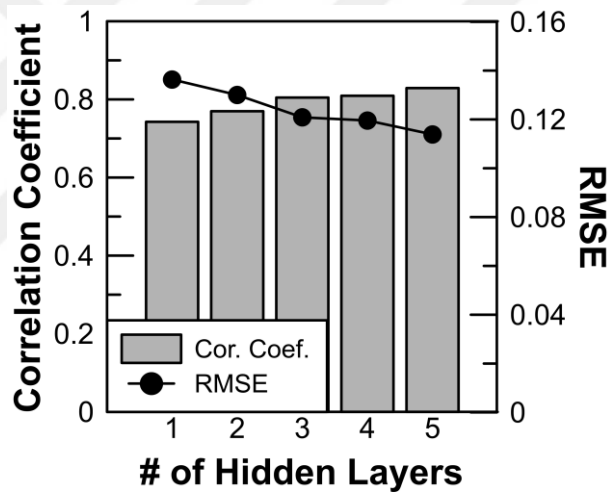


Figure D.15 : ANN hidden layer sensitivity for Case 2.3.3.

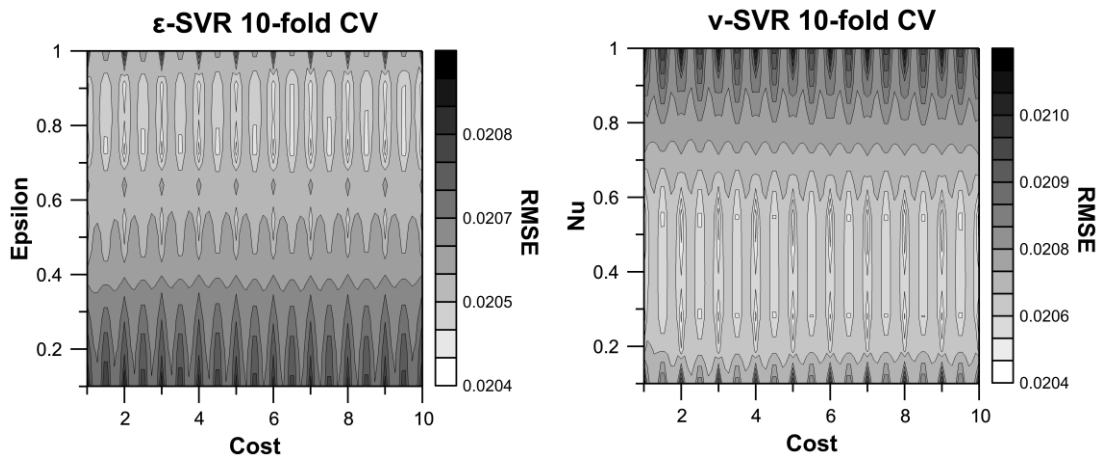


Figure D.16 : 10-fold CV grid search plots for Case 2.3.3.

Table D.3 : SVR hyper-parameters for Case 2.3.3.

ε -SVR					ν -SVR				
Linear			RBF		Linear			RBF	
C	ε	C	γ	ε	C	ν	C	γ	ν
7.5	0.75	1.5	1.0	0.2	1.5	0.6	1.5	1.0	0.4

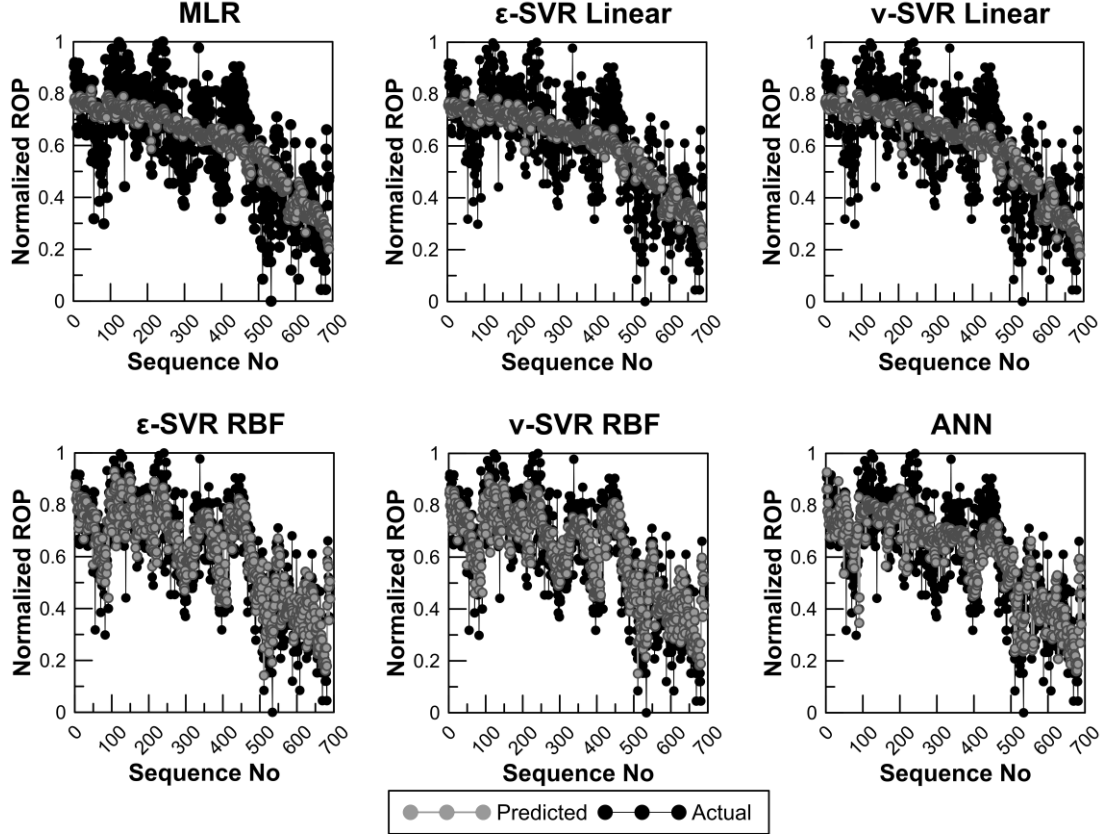


Figure D.17 : Model training for Case 2.3.3.

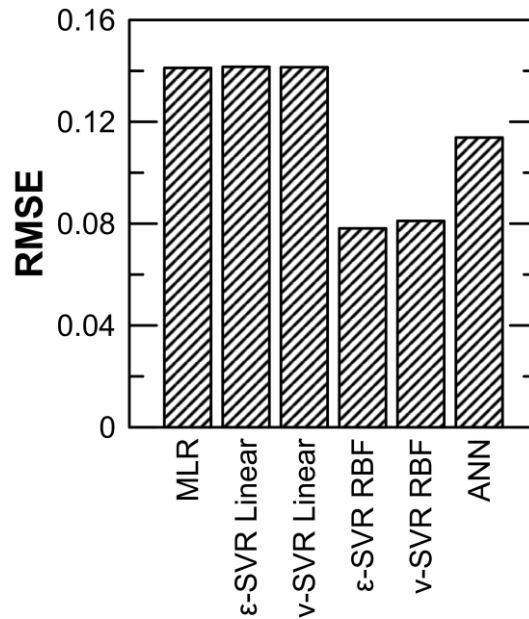


Figure D.18 : Training errors for Case 2.3.3.

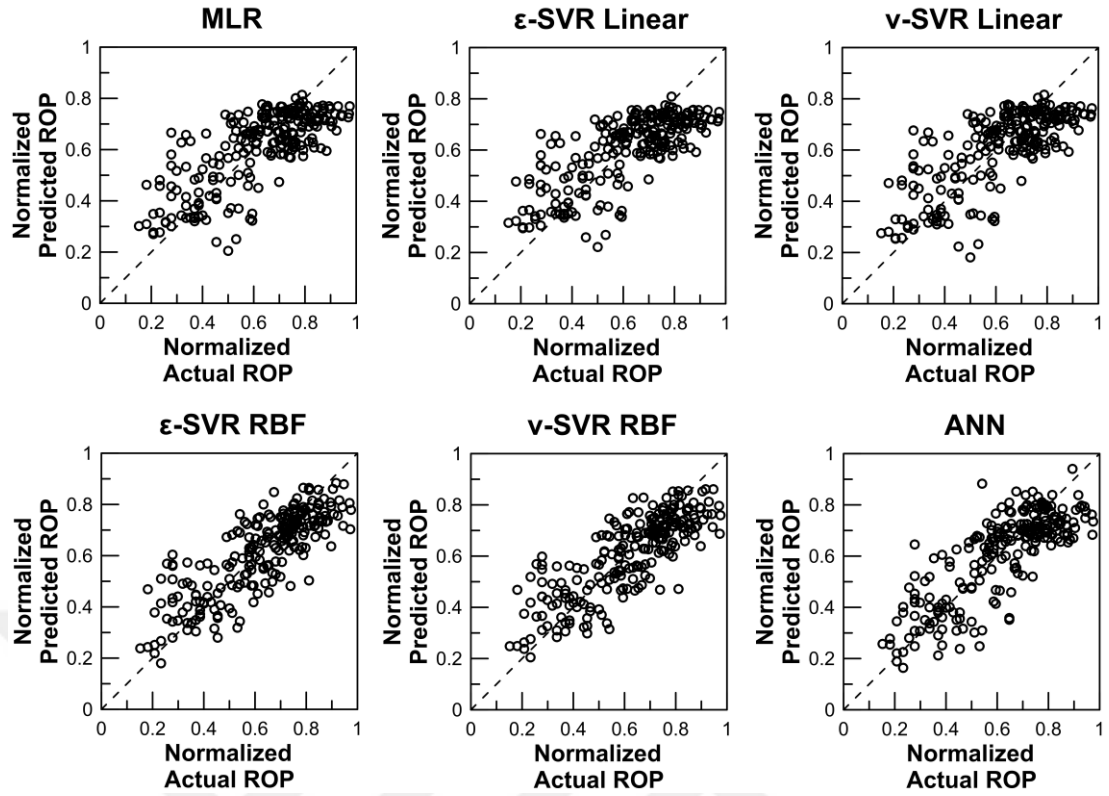


Figure D.19 : ROP predictions with testing subset for Case 2.3.3.

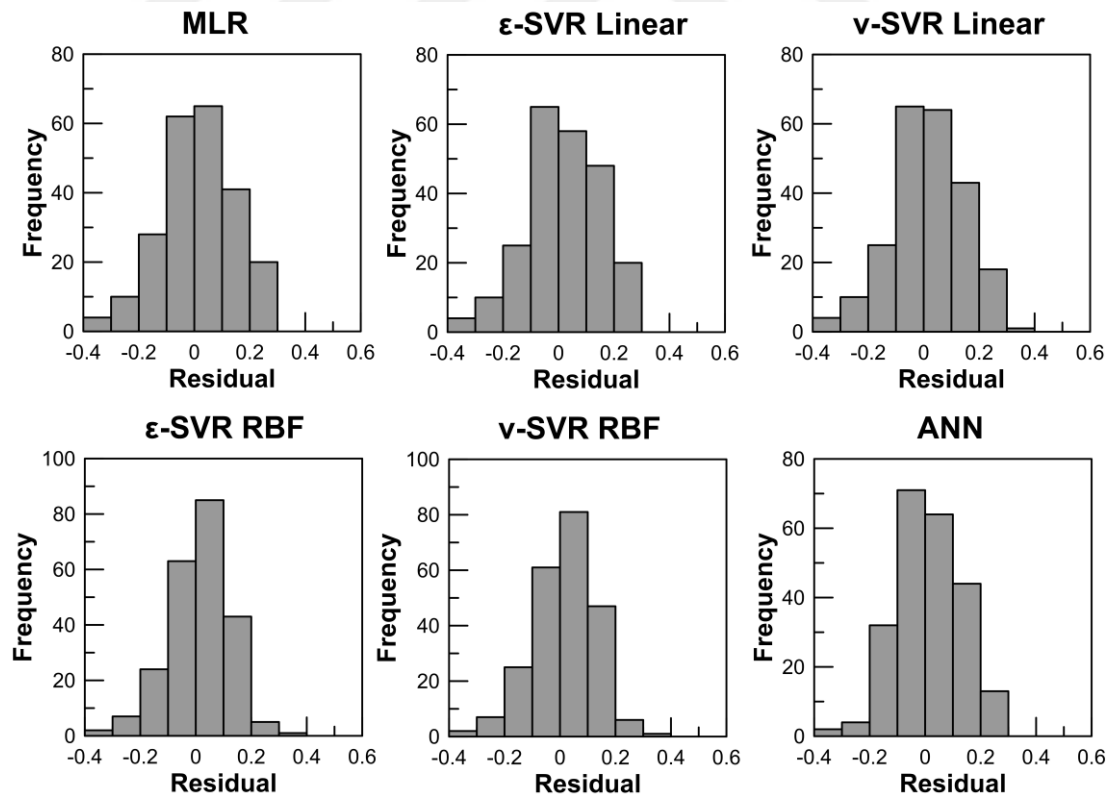


Figure D.20 : Error histograms for Case 2.3.3.

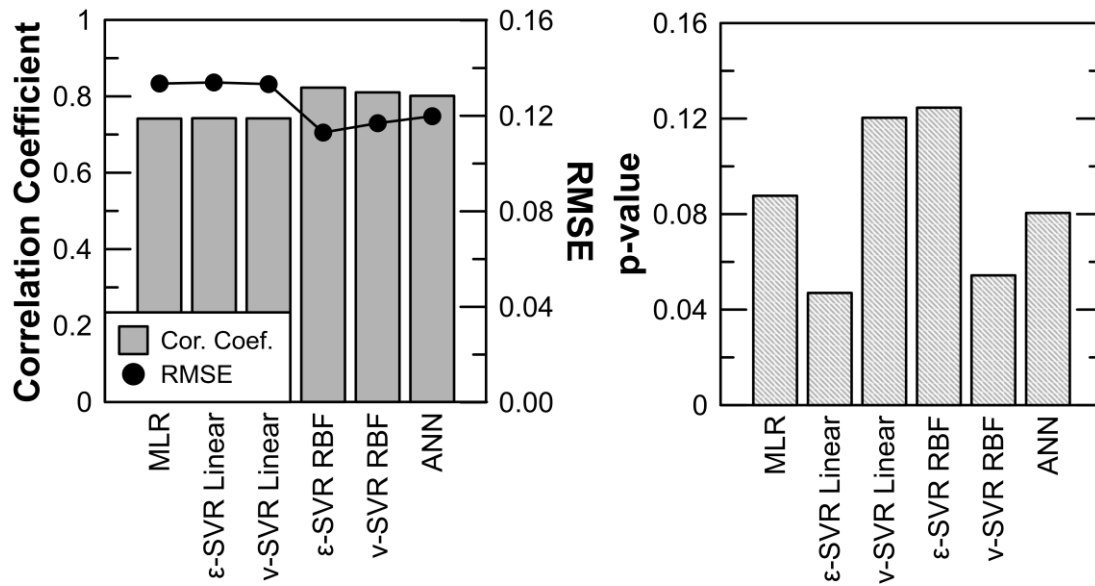


Figure D.21 : Statistical comparison for Case 2.3.3.

D.1.4. Case 2.3.4: BYM+BI

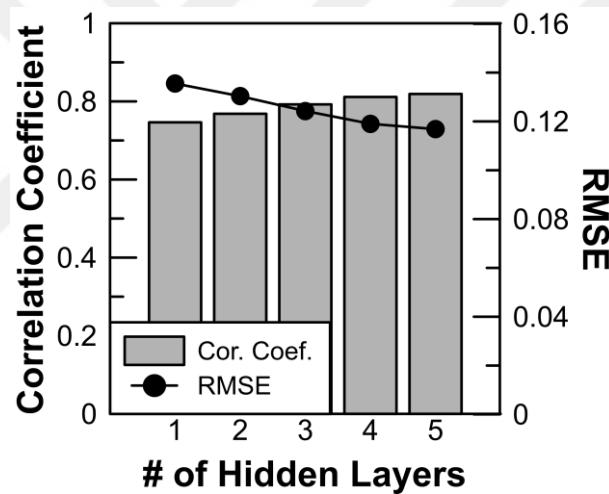


Figure D.22 : ANN hidden layer sensitivity for Case 2.3.4.

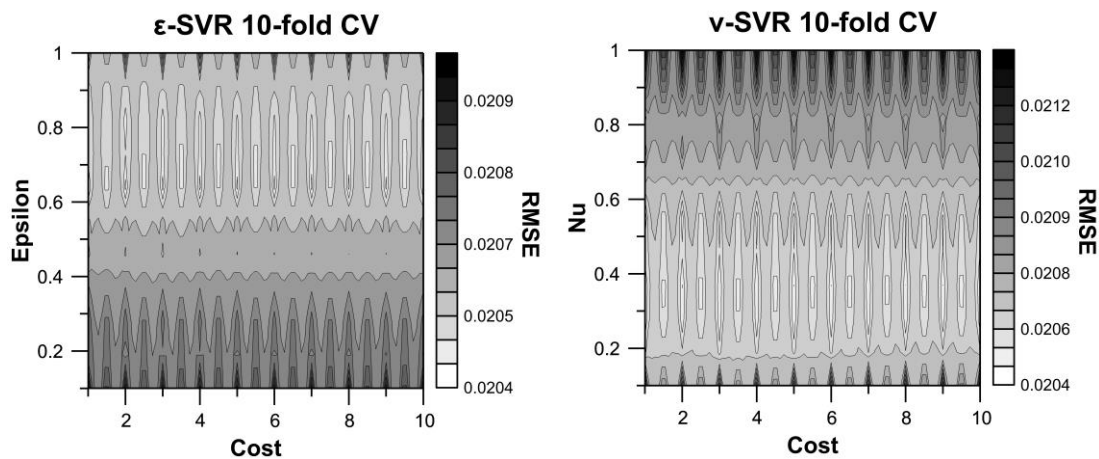


Figure D.23 : 10-fold CV grid search plots for Case 2.3.4.

Table D.4 : SVR hyper-parameters for Case 2.3.4.

ε -SVR					ν -SVR				
Linear			RBF		Linear			RBF	
C	ε	C	γ	ε	C	ν	C	γ	ν
2.0	0.7	1.5	1.0	0.1	2.0	0.3	1.5	1.0	0.6

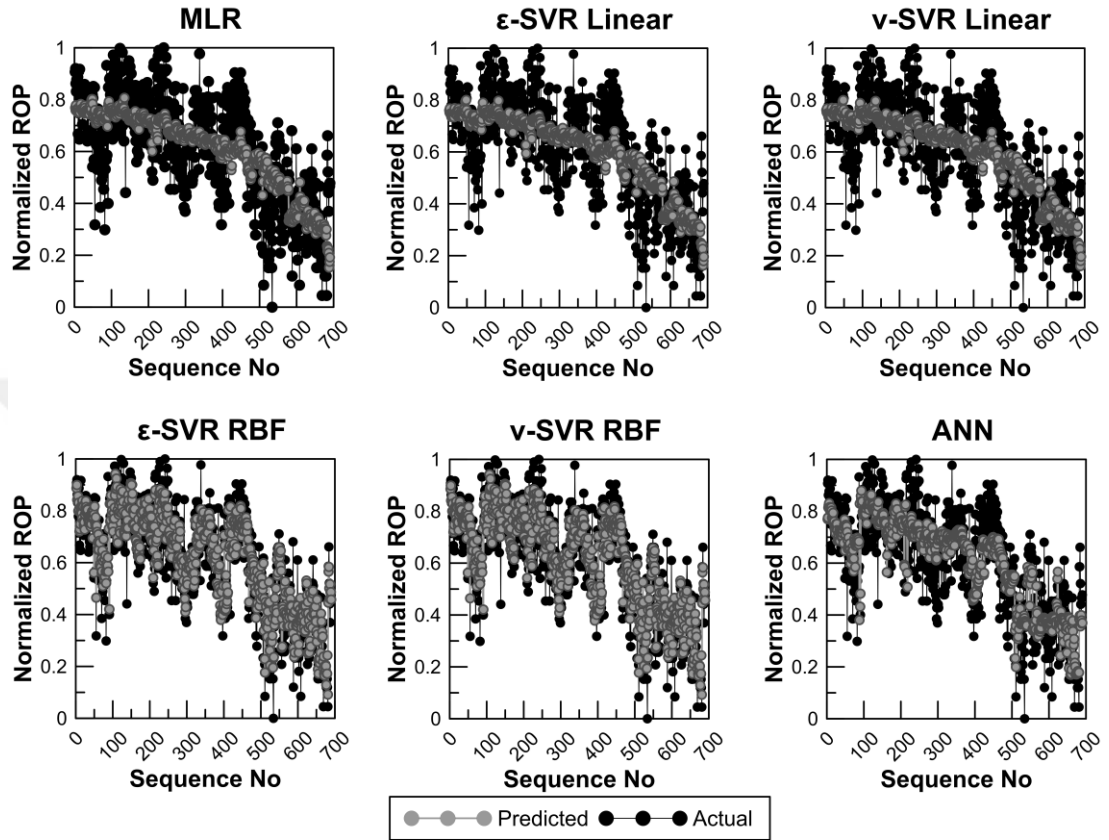


Figure D.24 : Model training for Case 2.3.4.

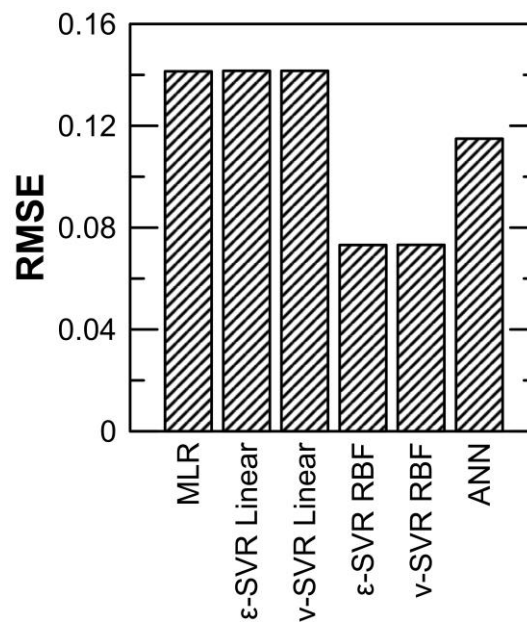


Figure D.25 : Training errors for Case 2.3.4.

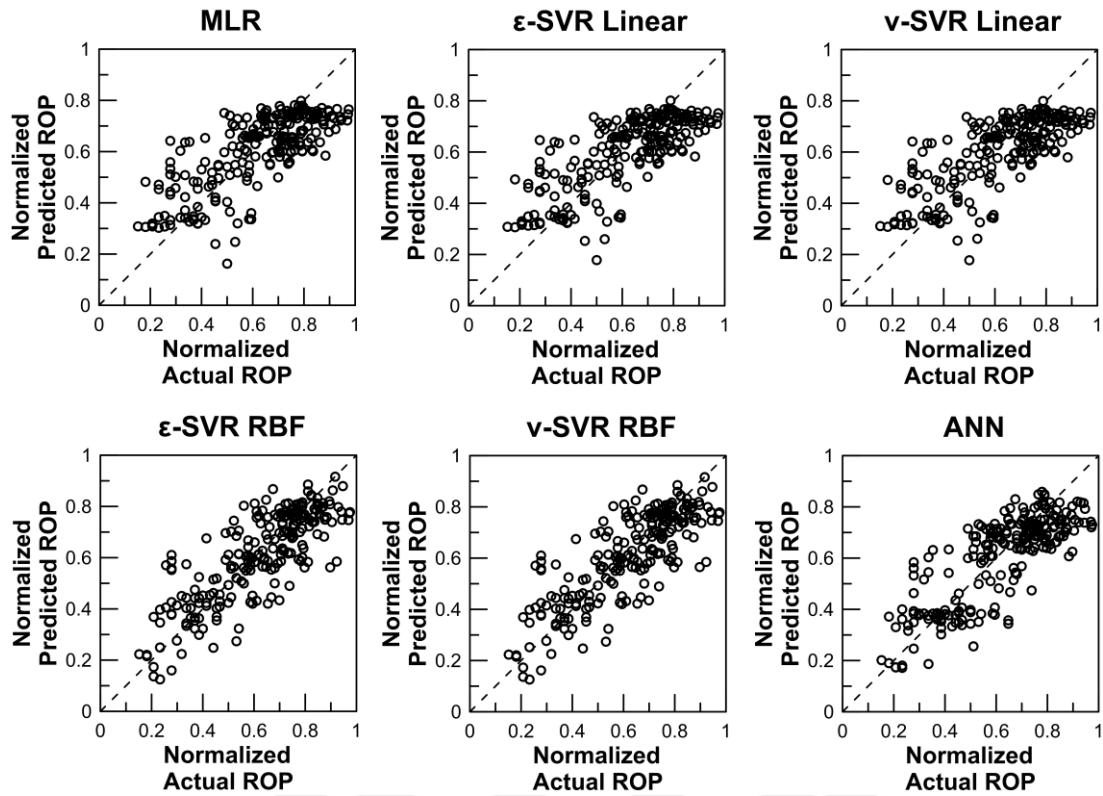


Figure D.26 : ROP predictions with testing subset for Case 2.3.4.

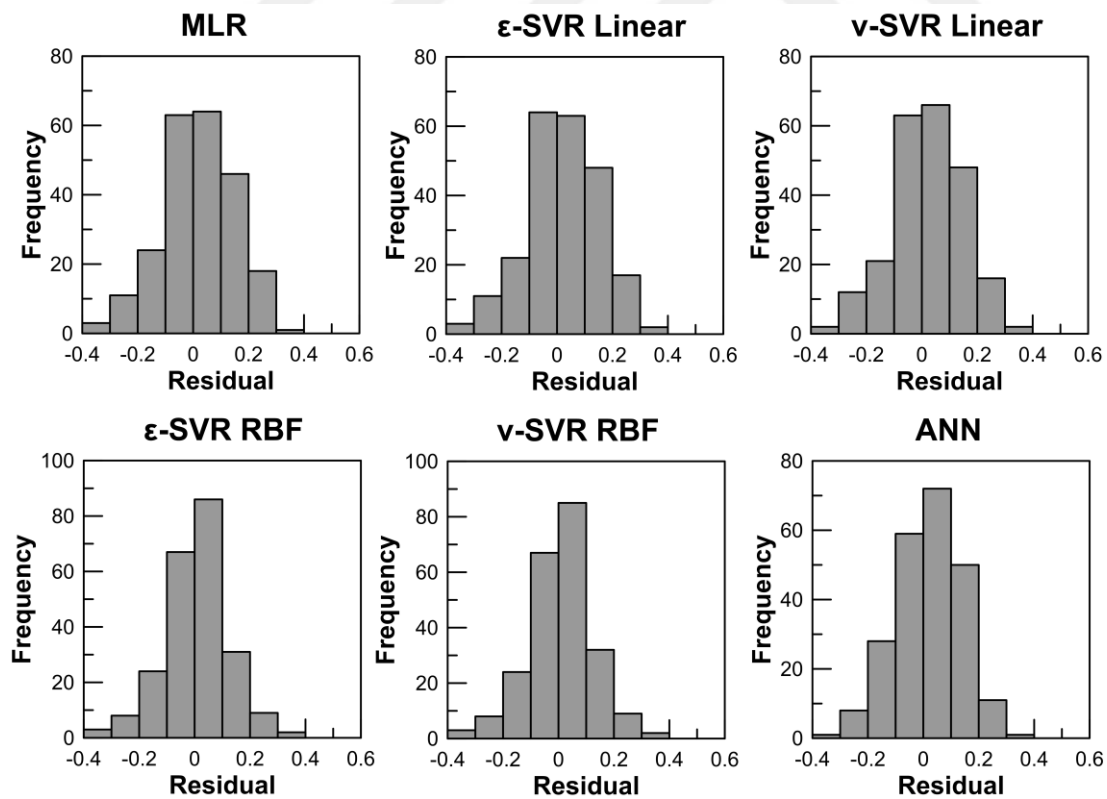


Figure D.27 : Error histograms for Case 2.3.4.

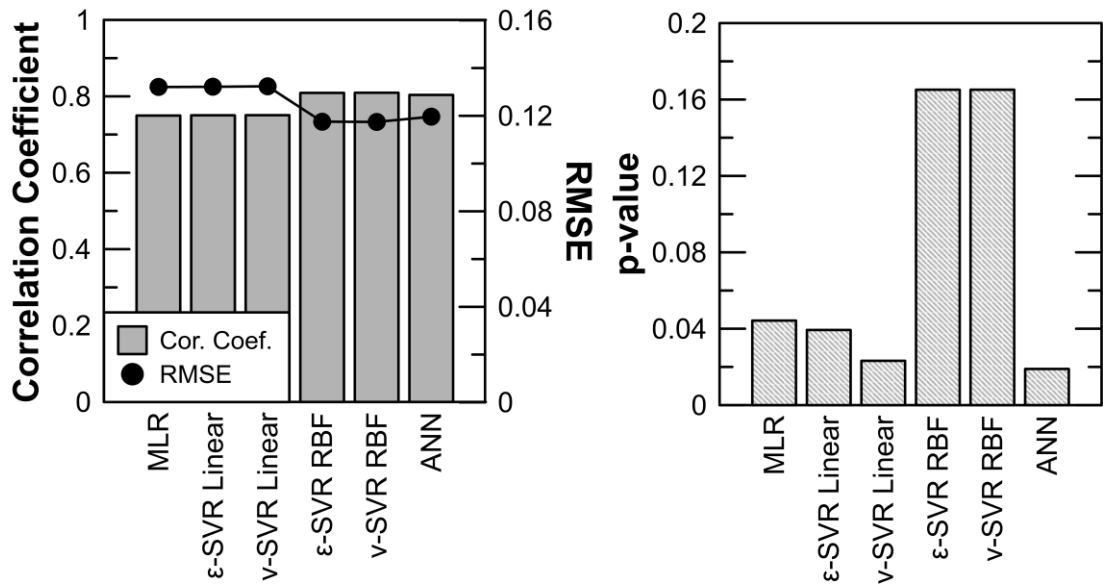


Figure D.28 : Statistical comparison for Case 2.3.4.

D.1.5. Case 2.3.5: BYM+SFG

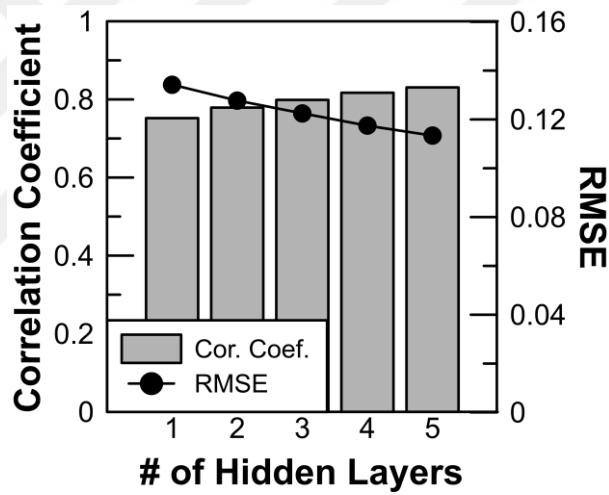


Figure D.29 : ANN hidden layer sensitivity for Case 2.3.5.

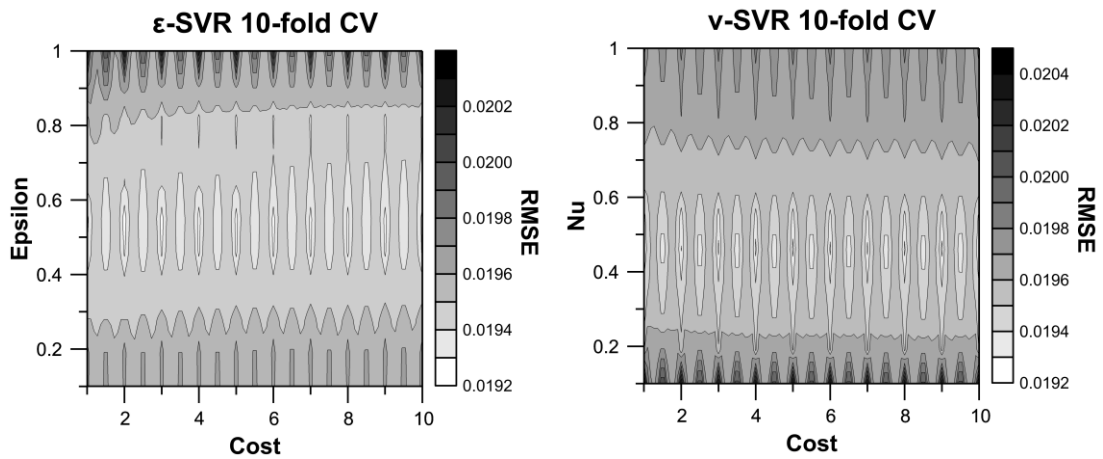


Figure D.30 : 10-fold CV grid search plots for Case 2.3.5.

Table D.5 : SVR hyper-parameters for Case 2.3.5.

ε -SVR					ν -SVR				
Linear			RBF		Linear			RBF	
C	ε	C	γ	ε	C	ν	C	γ	ν
8.5	0.5	1.5	1.0	0.1	7.5	0.45	1.5	1.0	0.7

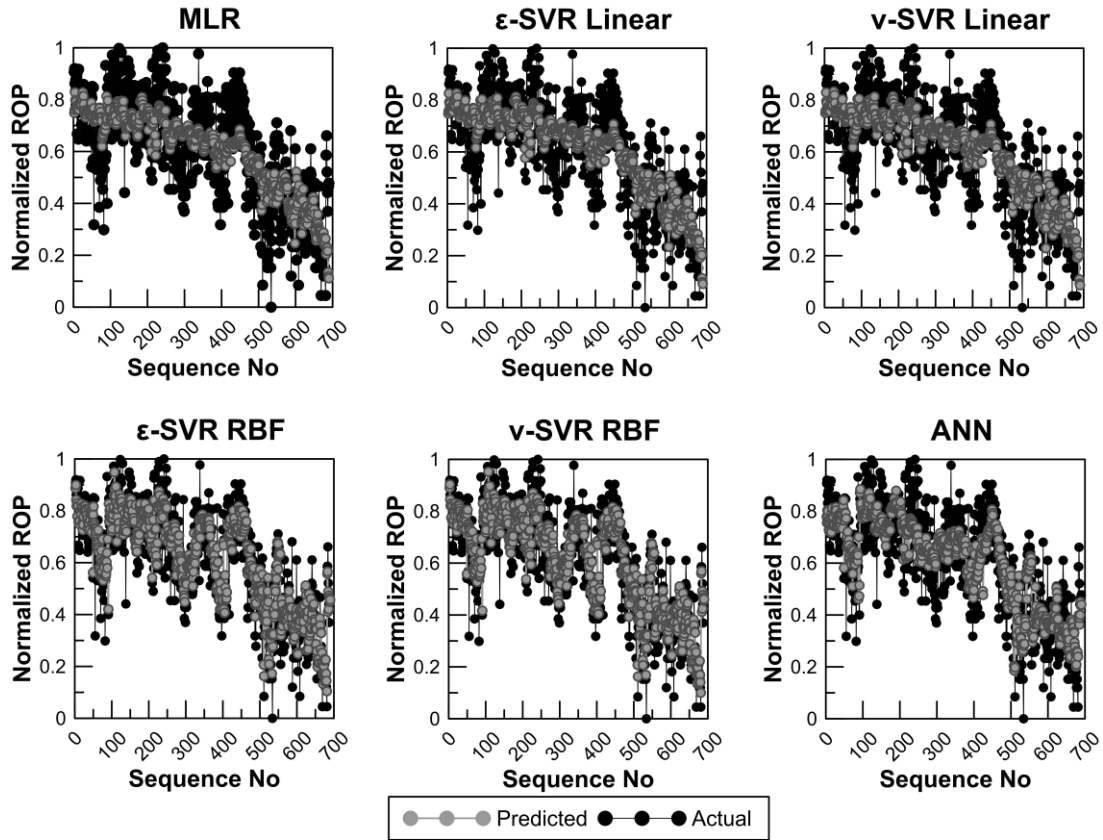


Figure D.31 : Model training for Case 2.3.5.

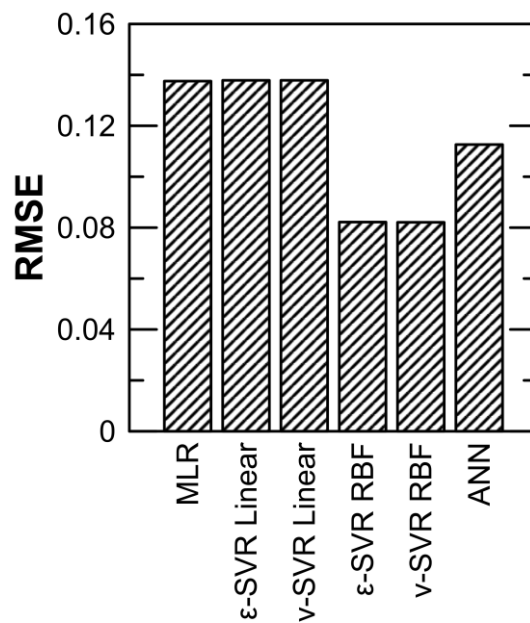


Figure D.32 : Training errors for Case 2.3.5.

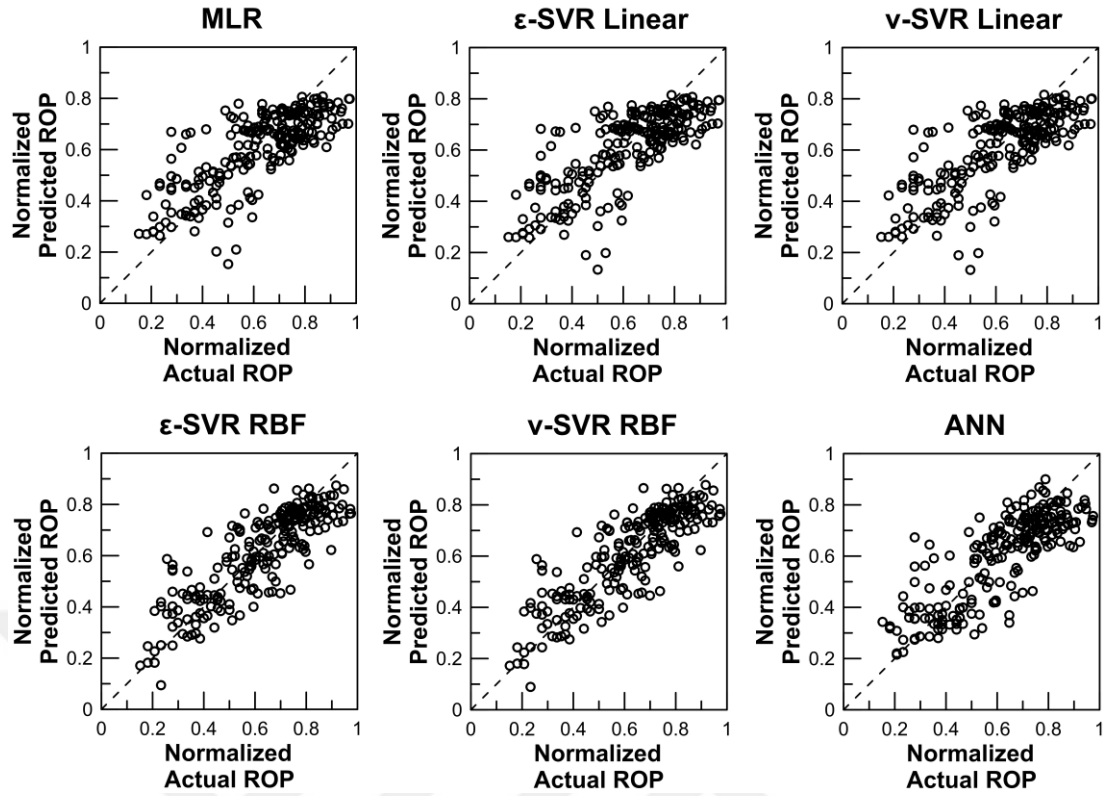


Figure D.33 : ROP predictions with testing subset for Case 2.3.5.

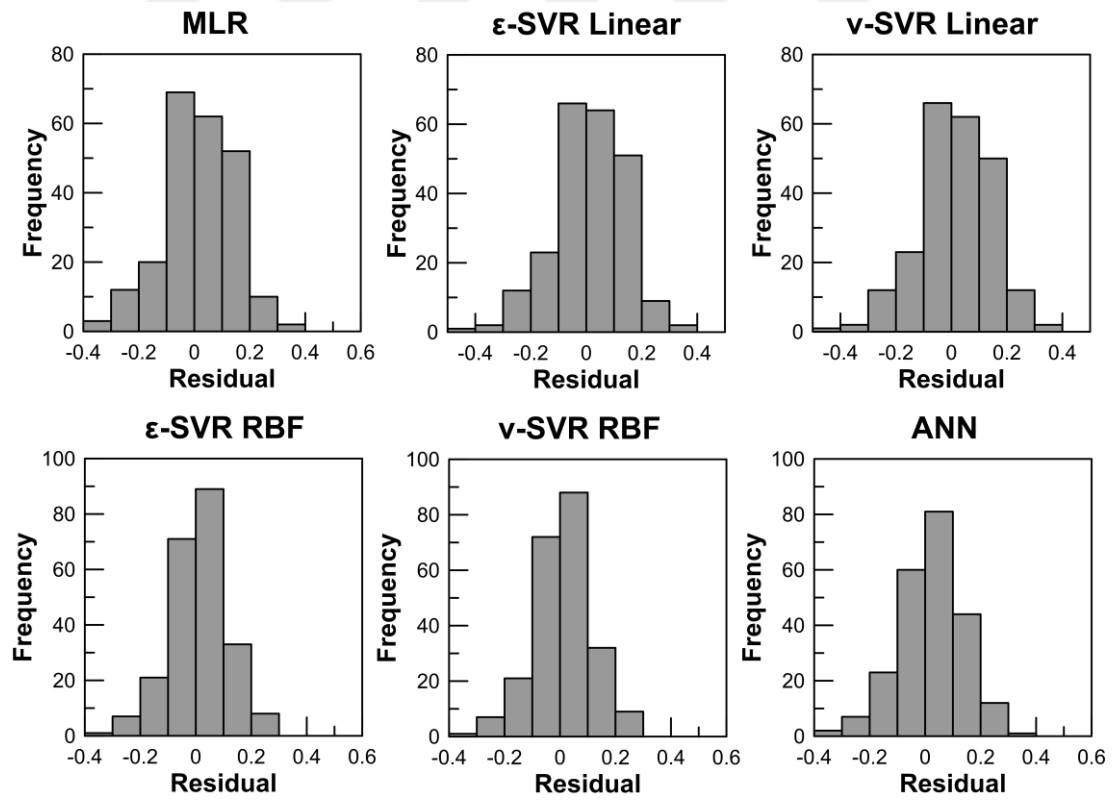


Figure D.34 : Error histograms for Case 2.3.5.

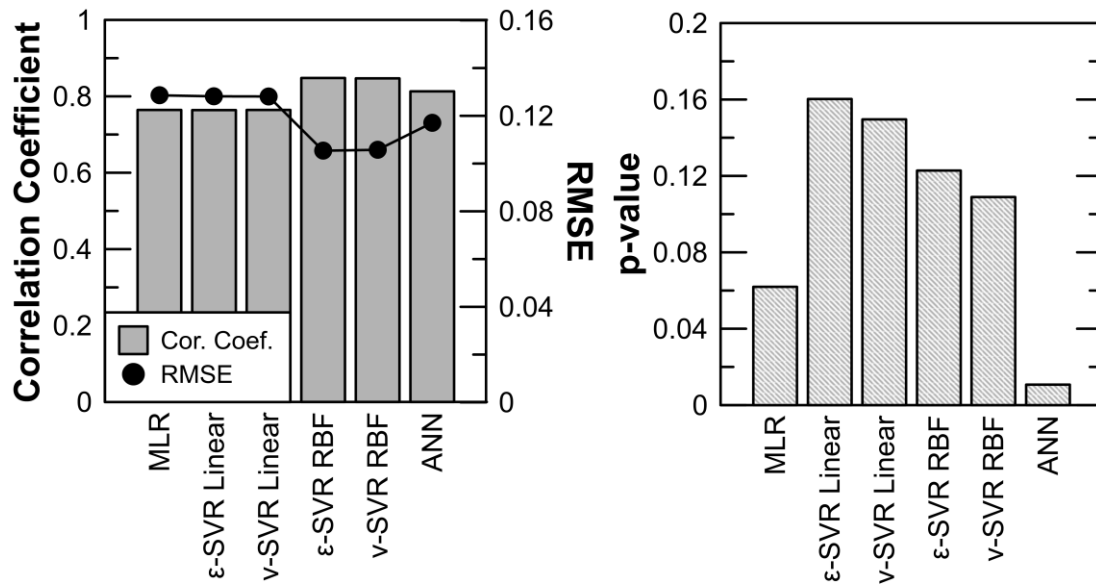


Figure D.35 : Statistical comparison for Case 2.3.5.

D.1.6. Case 2.3.6: BYM+All

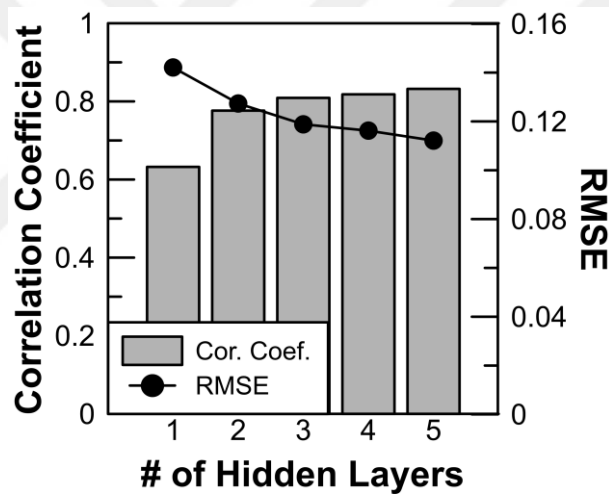


Figure D.36 : ANN hidden layer sensitivity for Case 2.3.6.

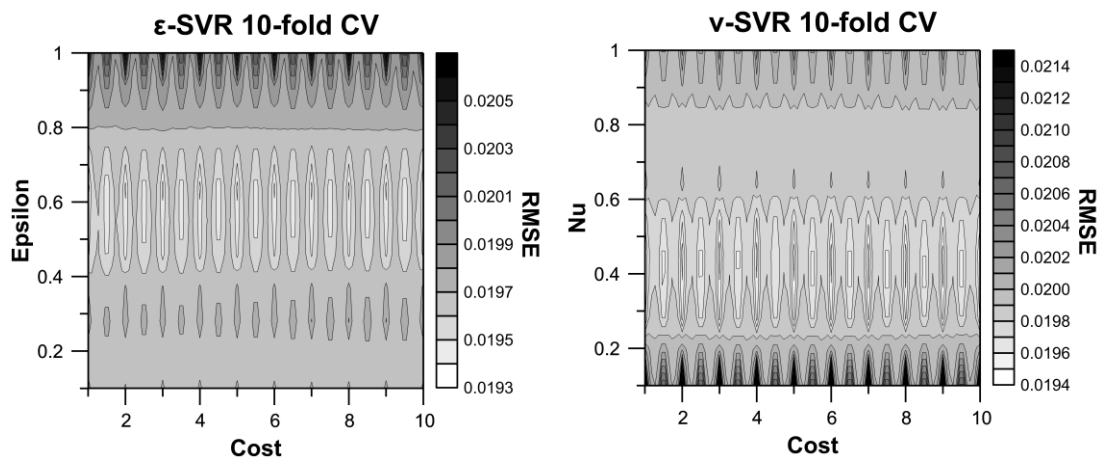


Figure D.37 : 10-fold CV grid search plots for Case 2.3.6.

Table D.6 : SVR hyper-parameters for Case 2.3.6.

ε -SVR					ν -SVR				
Linear			RBF		Linear			RBF	
C	ε	C	γ	ε	C	ν	C	γ	ν
1.5	0.5	2.0	1.0	0.1	1.5	0.3	1.5	1.0	0.7

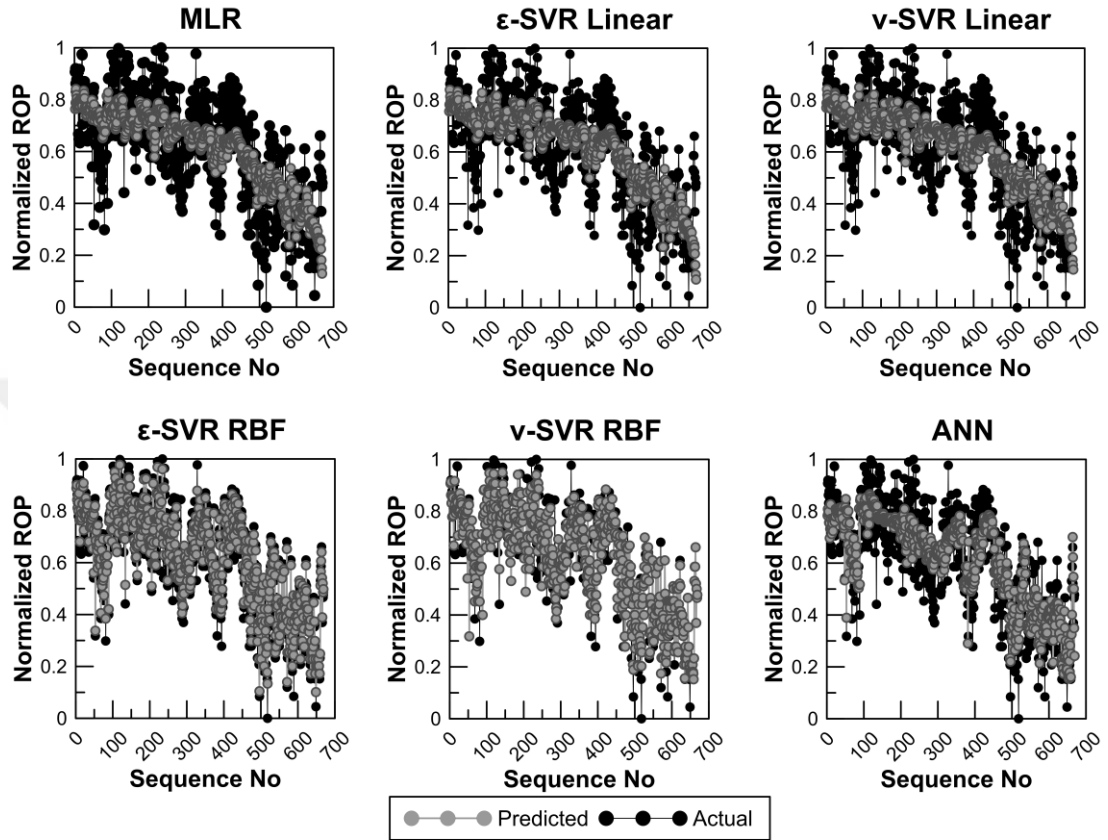


Figure D.38 : Model training for Case 2.3.6.

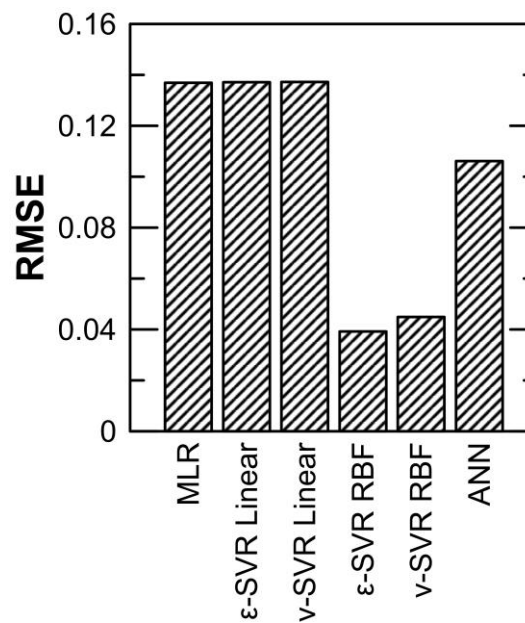


Figure D.39 : Training errors for Case 2.3.6.

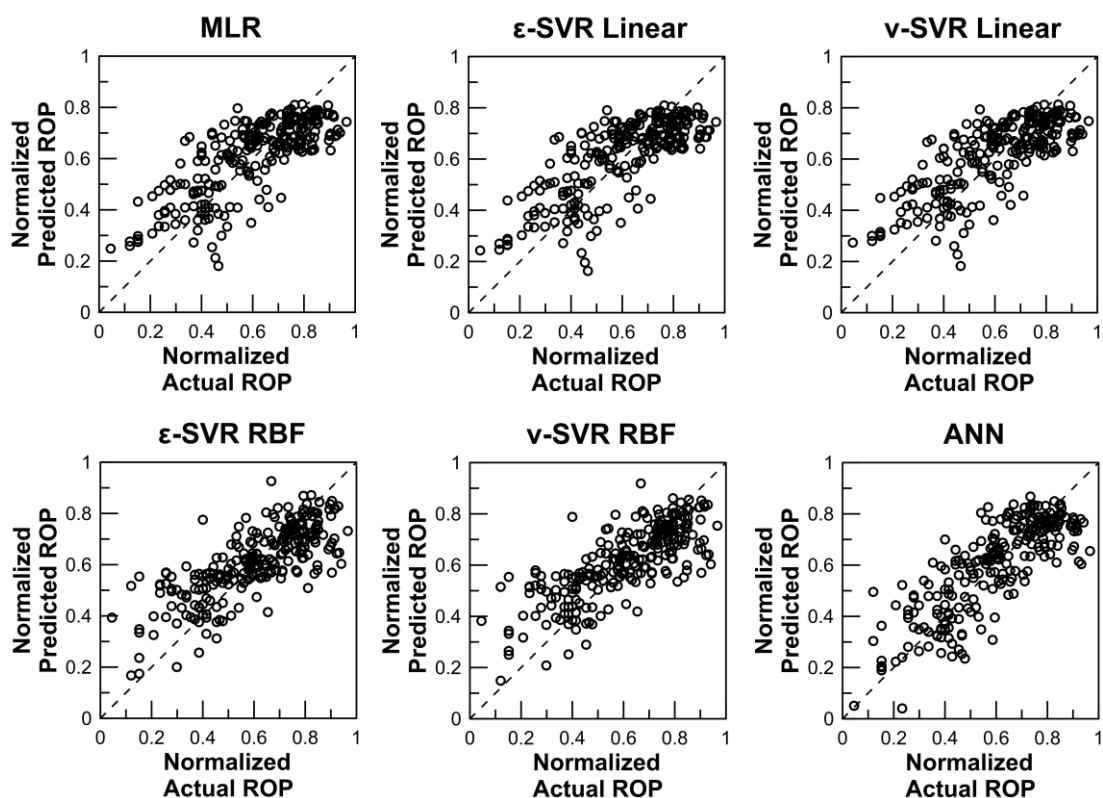


Figure D.40 : ROP predictions with testing subset for Case 2.3.6.

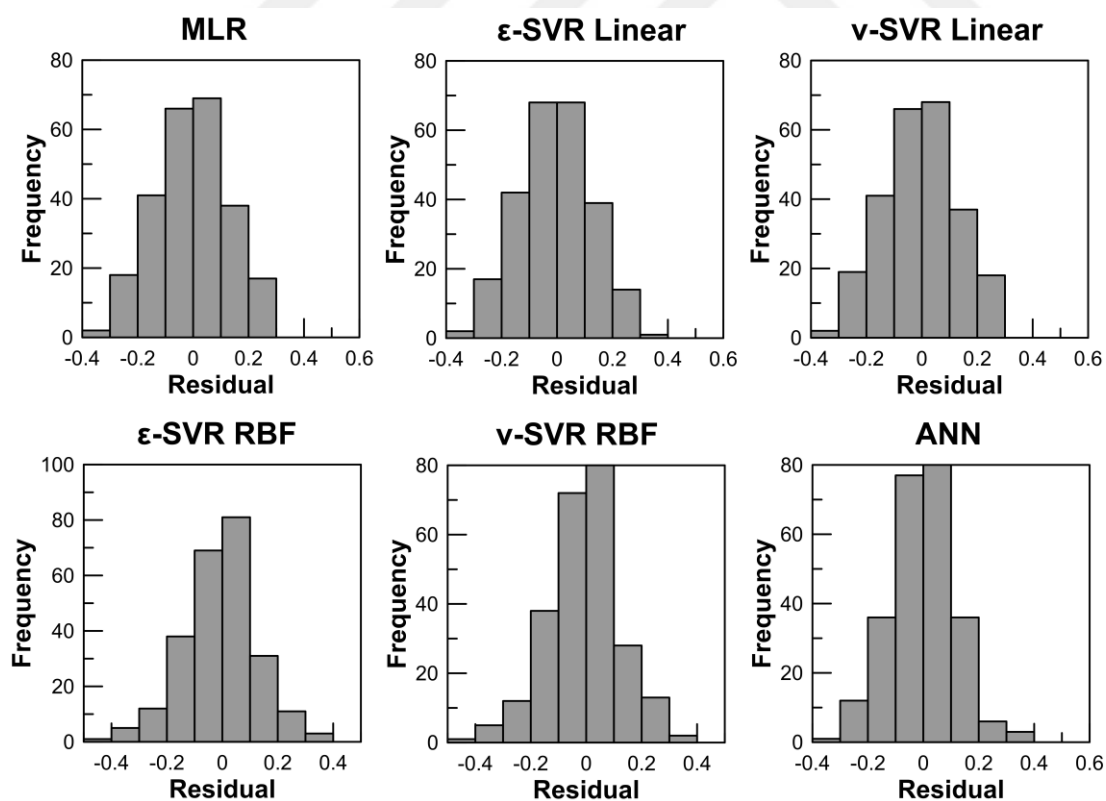


Figure D.41 : Error histograms for Case 2.3.6.

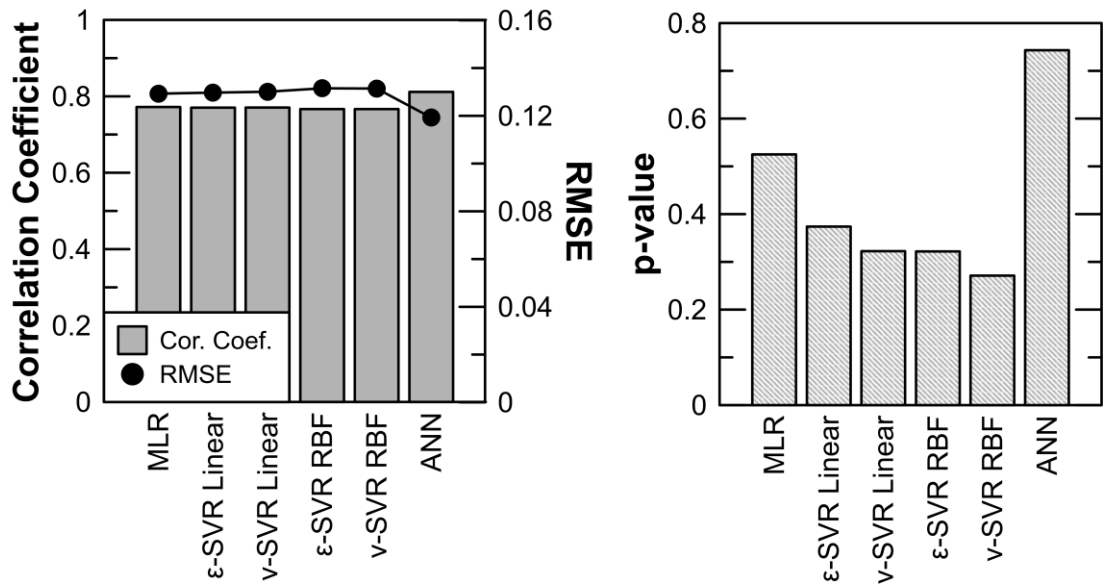


Figure D.42 : Statistical comparison for Case 2.3.6.

D.1.7. Case 2.3.7: Feature Selection (BYM+SFG excluding x_5)

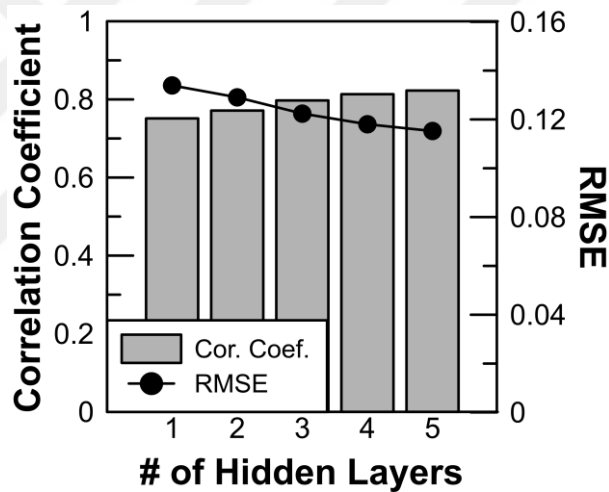


Figure D.43 : ANN hidden layer sensitivity for Case 2.3.7.

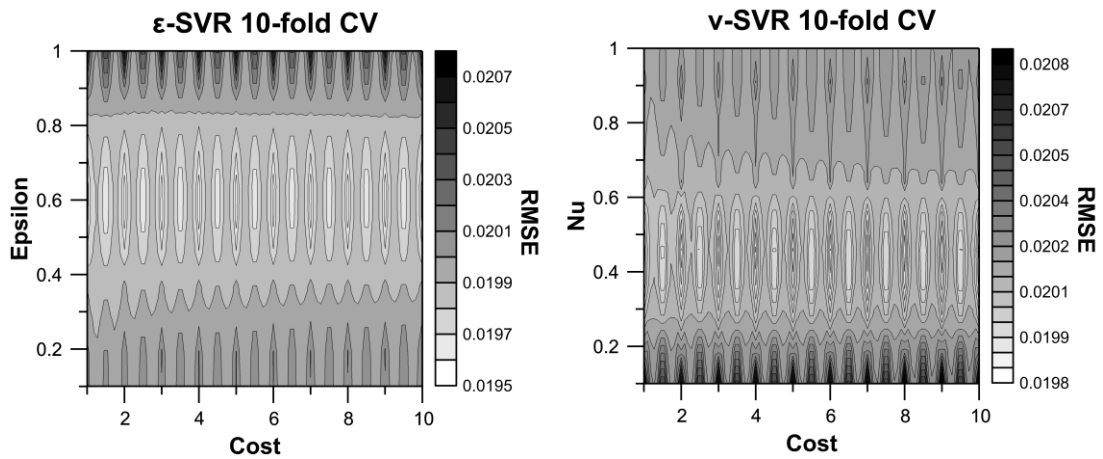


Figure D.44 : 10-fold CV grid search plots for Case 2.3.7.

Table D.7 : SVR hyper-parameters for Case 2.3.7.

ε -SVR					ν -SVR				
Linear			RBF		Linear			RBF	
C	ε	C	γ	ε	C	ν	C	γ	ν
7.5	0.65	3.0	1.0	0.2	6.0	0.4	3.5	1.0	0.6

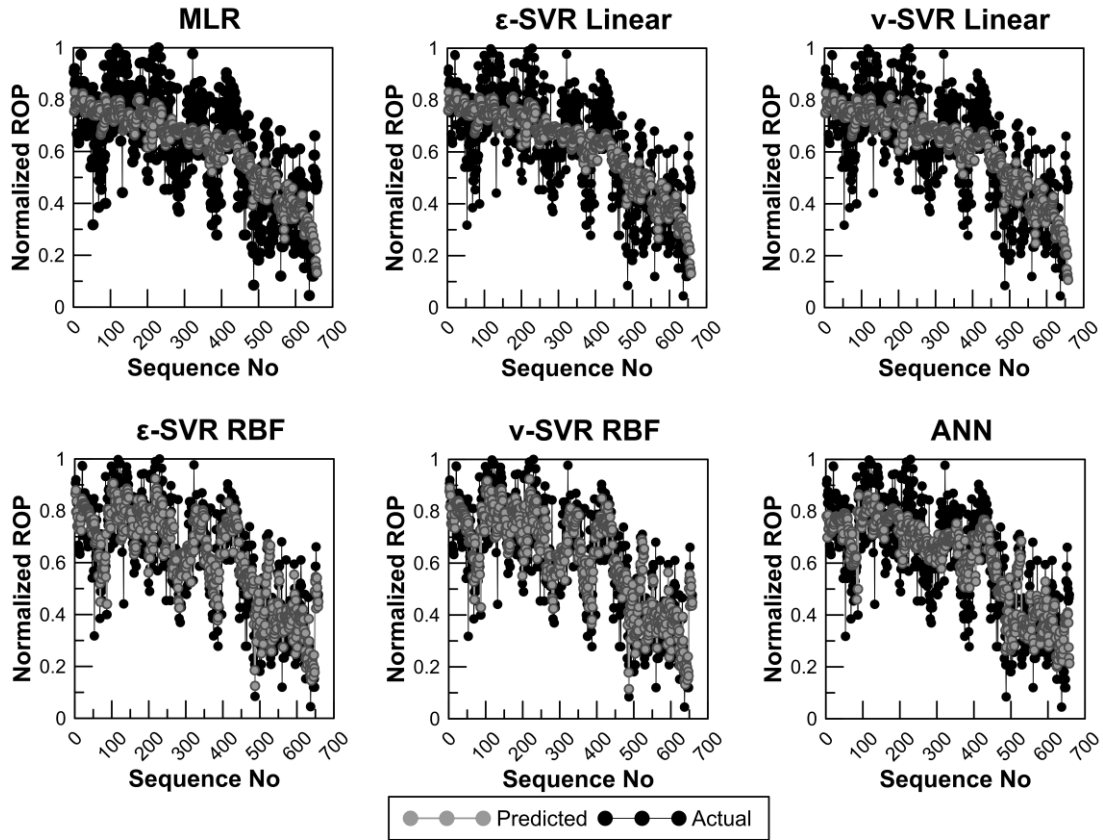


Figure D.45 : Model training for Case 2.3.7.

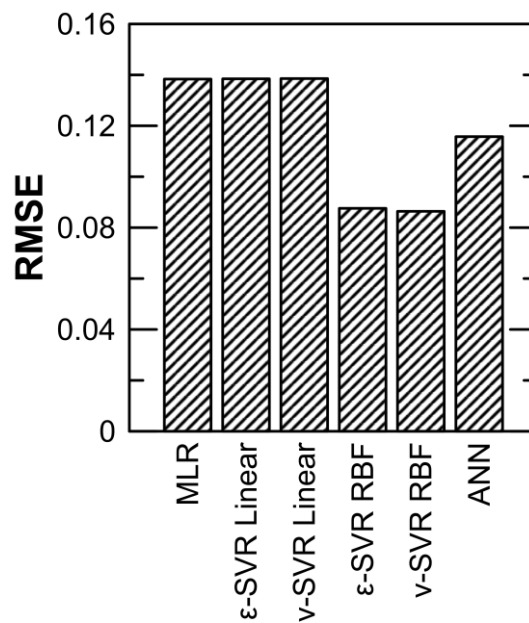


Figure D.46 : Training errors for Case 2.3.7.

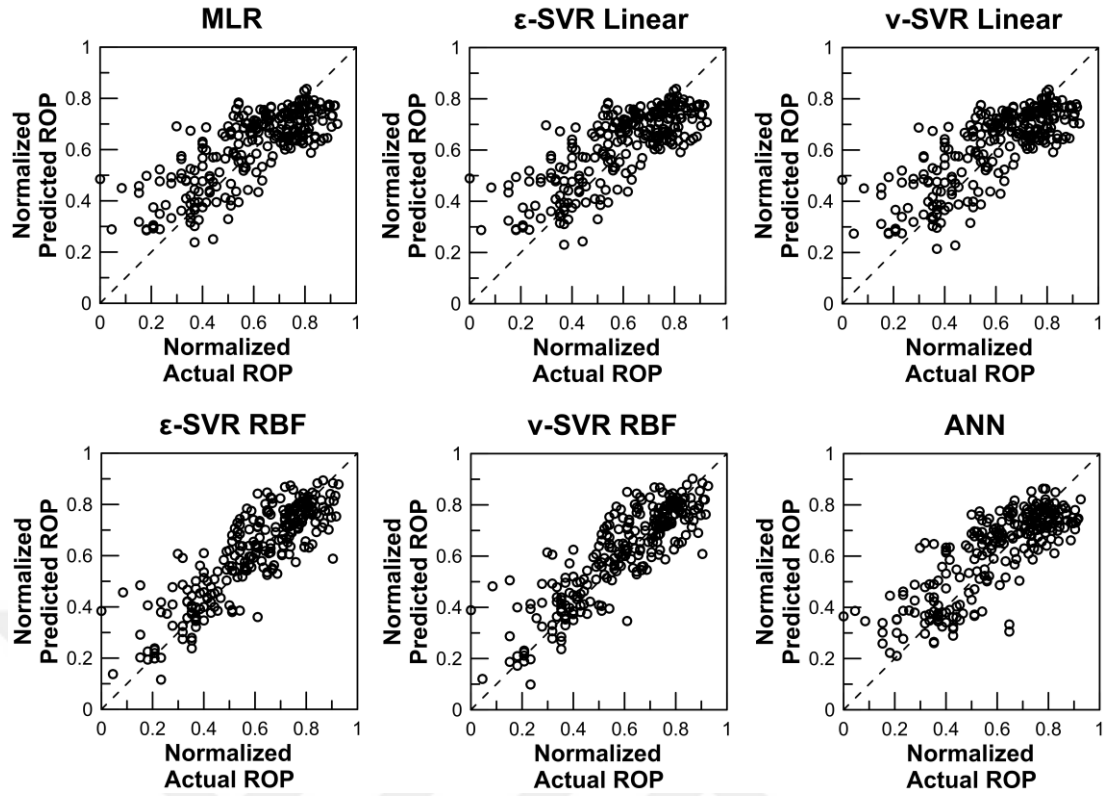


Figure D.47 : ROP predictions with testing subset for Case 2.3.7.

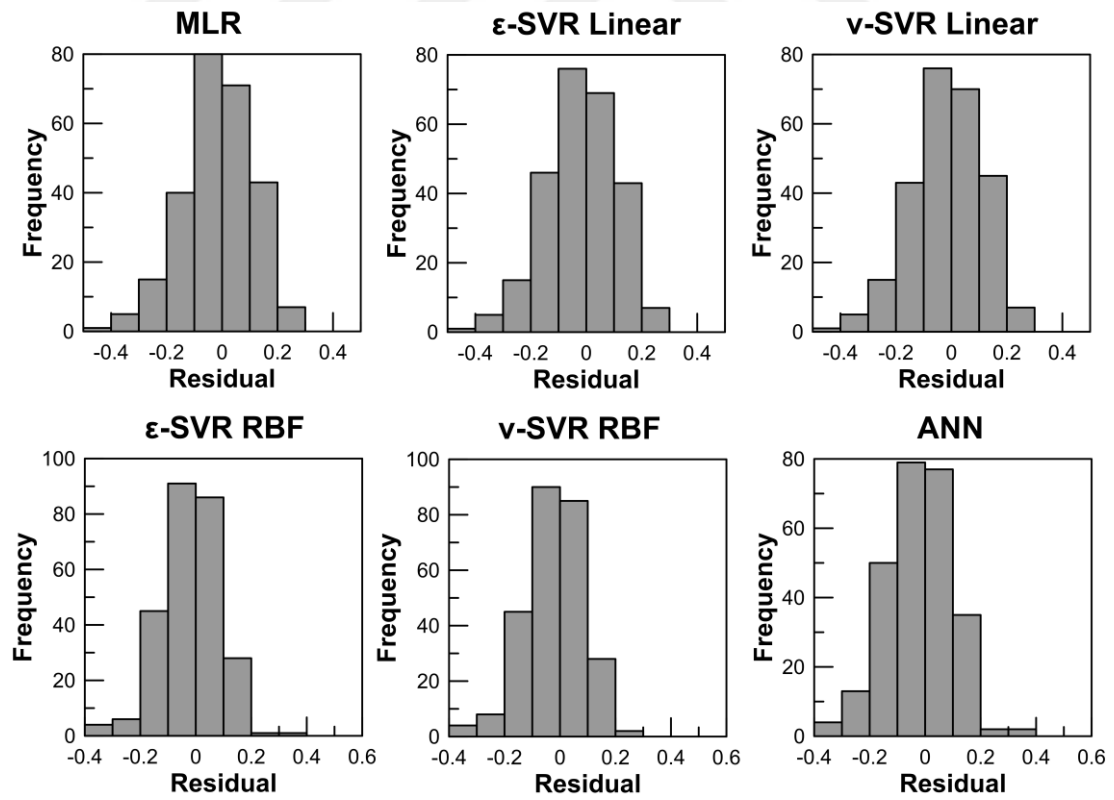


Figure D.48 : Error histograms for Case 2.3.7.

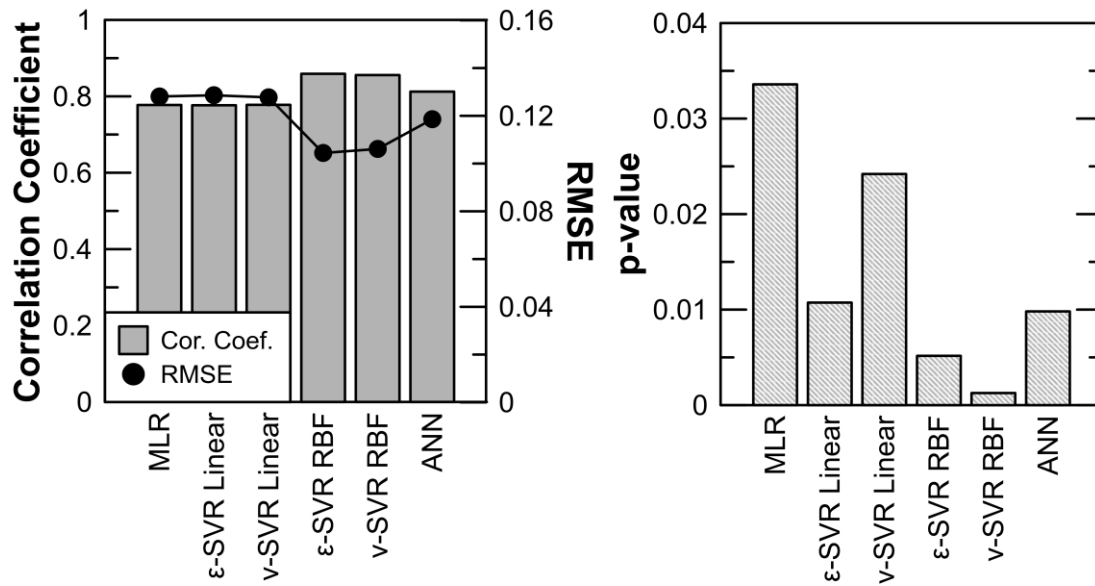


Figure D.49 : Statistical comparison for Case 2.3.7.

D.2. Intermediate Steps for Case 3.1

D.2.1. Case 3.1.1: Default BYM

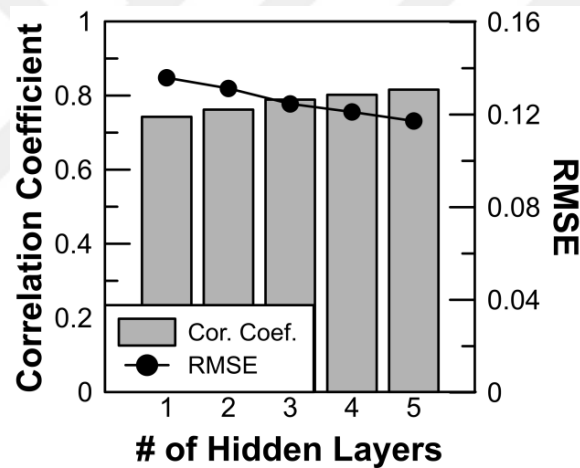


Figure D.50 : ANN hidden layer sensitivity for Case 3.1.1.

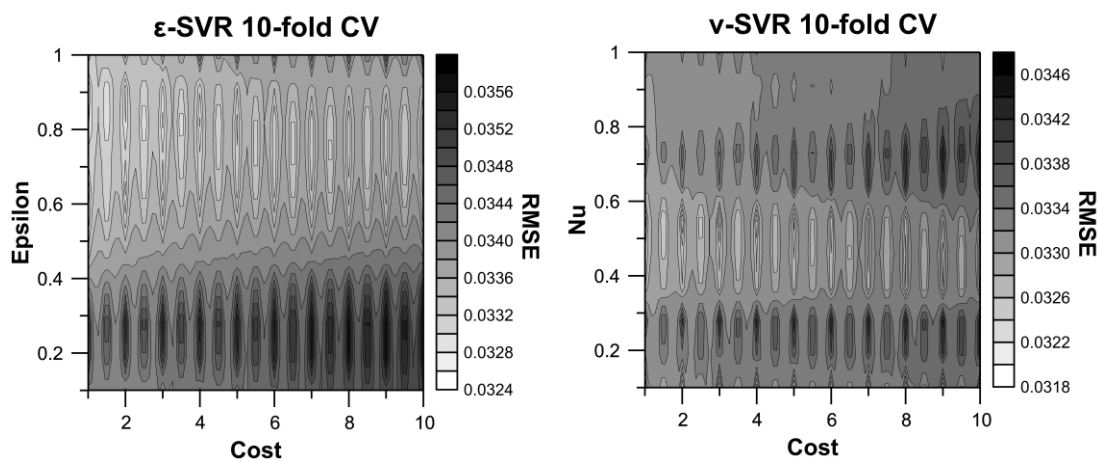


Figure D.51 : 10-fold CV grid search plots for Case 3.1.1.

Table D.8 : SVR hyper-parameters for Case 3.1.1.

ε -SVR					ν -SVR				
Linear			RBF		Linear			RBF	
C	ε	C	γ	ε	C	ν	C	γ	ν
2.5	0.85	1.5	1.0	0.1	1.0	0.55	1.5	3.0	0.9

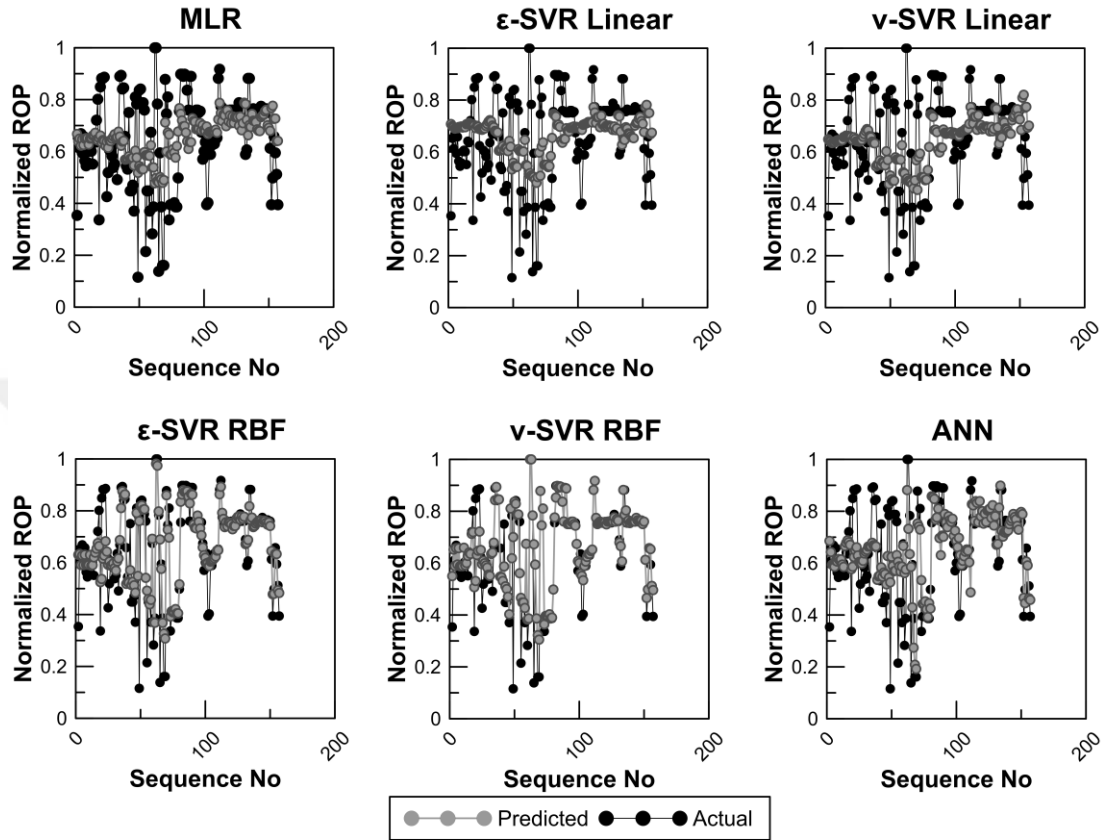


Figure D.52 : Model training for Case 3.1.1.

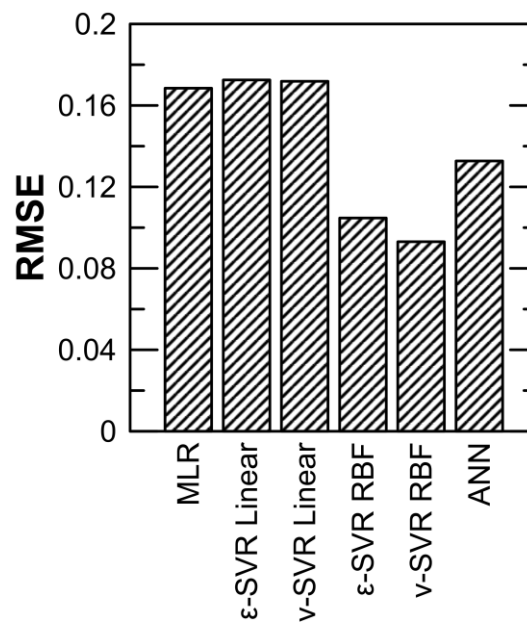


Figure D.53 : Training errors for Case 3.1.1.

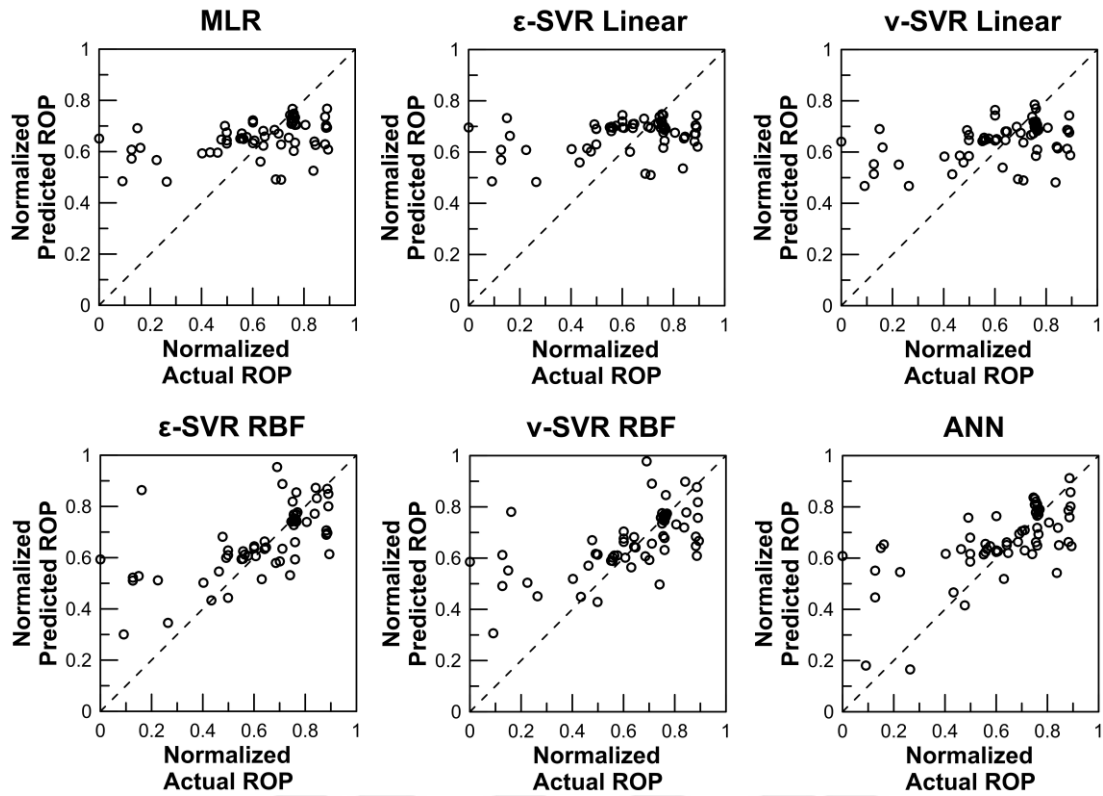


Figure D.54 : ROP predictions with testing subset for Case 3.1.1.

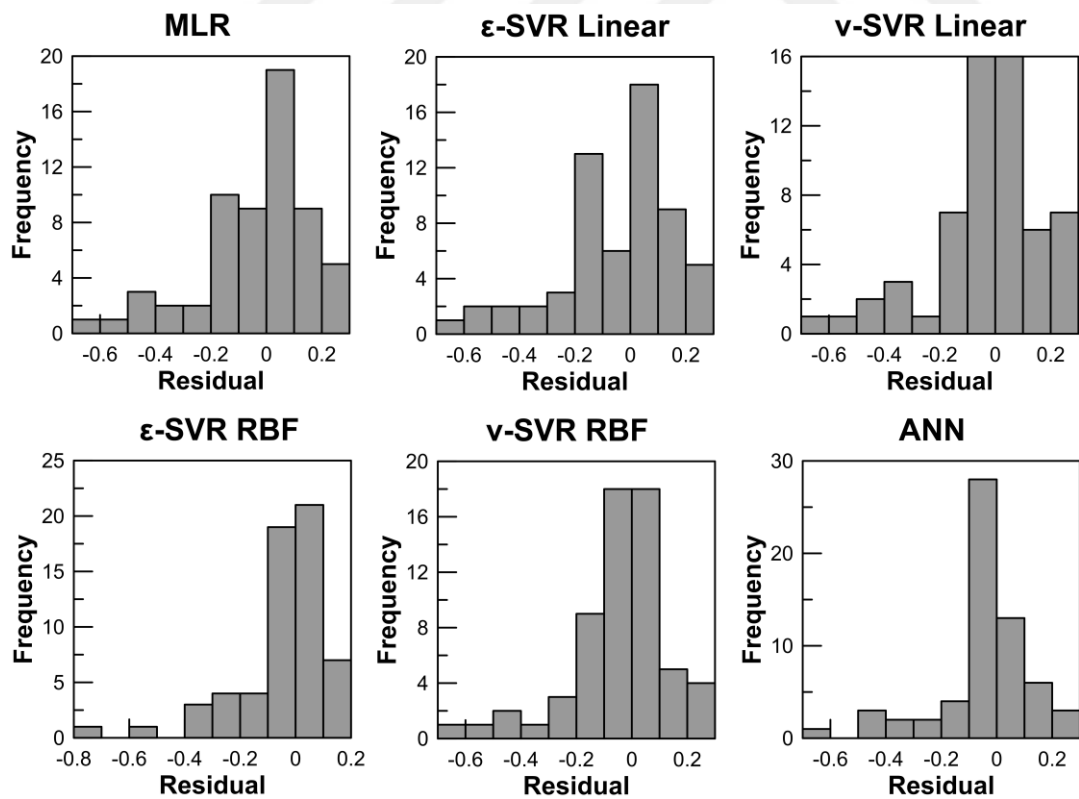


Figure D.55 : Error histograms for Case 3.1.1.

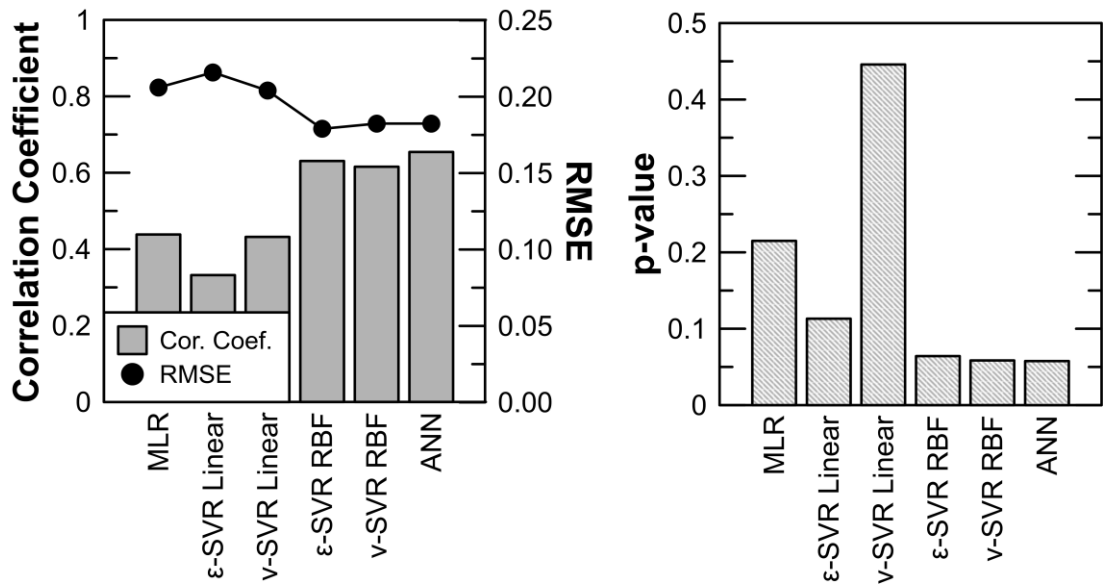


Figure D.56 : Statistical comparison for Case 3.1.1.

D.2.2. Case 3.1.2: BYM+SI

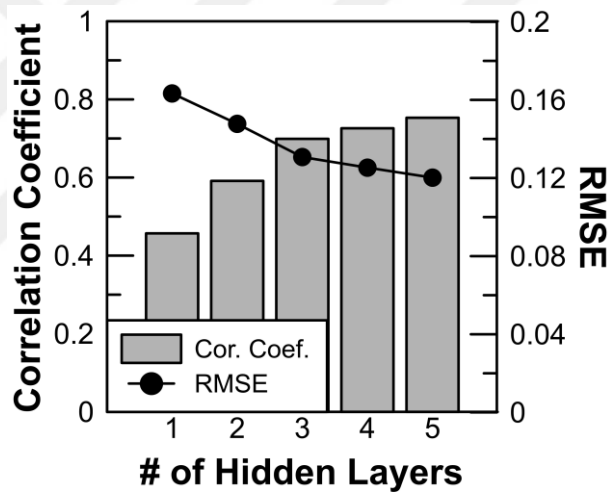


Figure D.57 : ANN hidden layer sensitivity for Case 3.1.2.

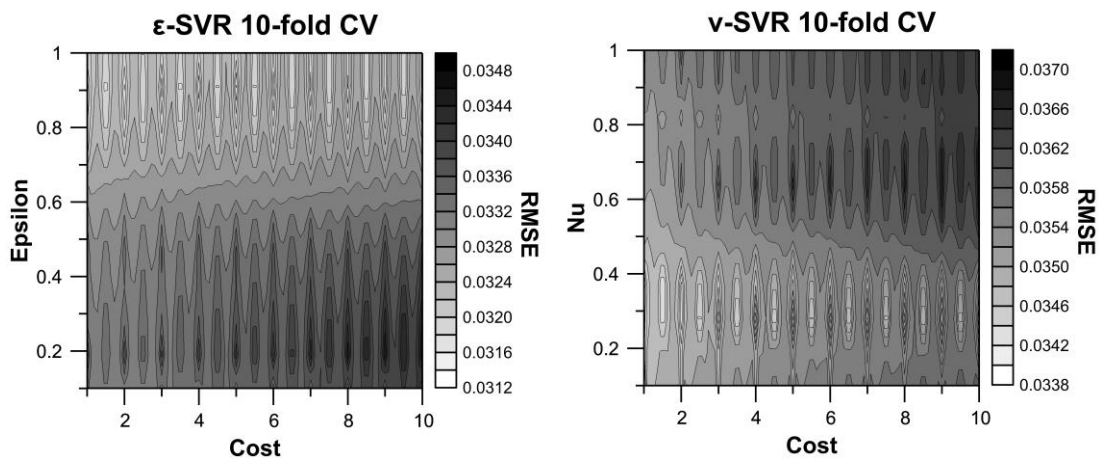


Figure D.58 : 10-fold CV grid search plots for Case 3.1.2.

Table D.9 : SVR hyper-parameters for Case 3.1.2.

ε -SVR					ν -SVR				
Linear		RBF			Linear		RBF		
C	ε	C	γ	ε	C	ν	C	γ	ν
5.0	0.9	1.5	1.0	0.1	4.0	0.3	1.5	1.0	0.6

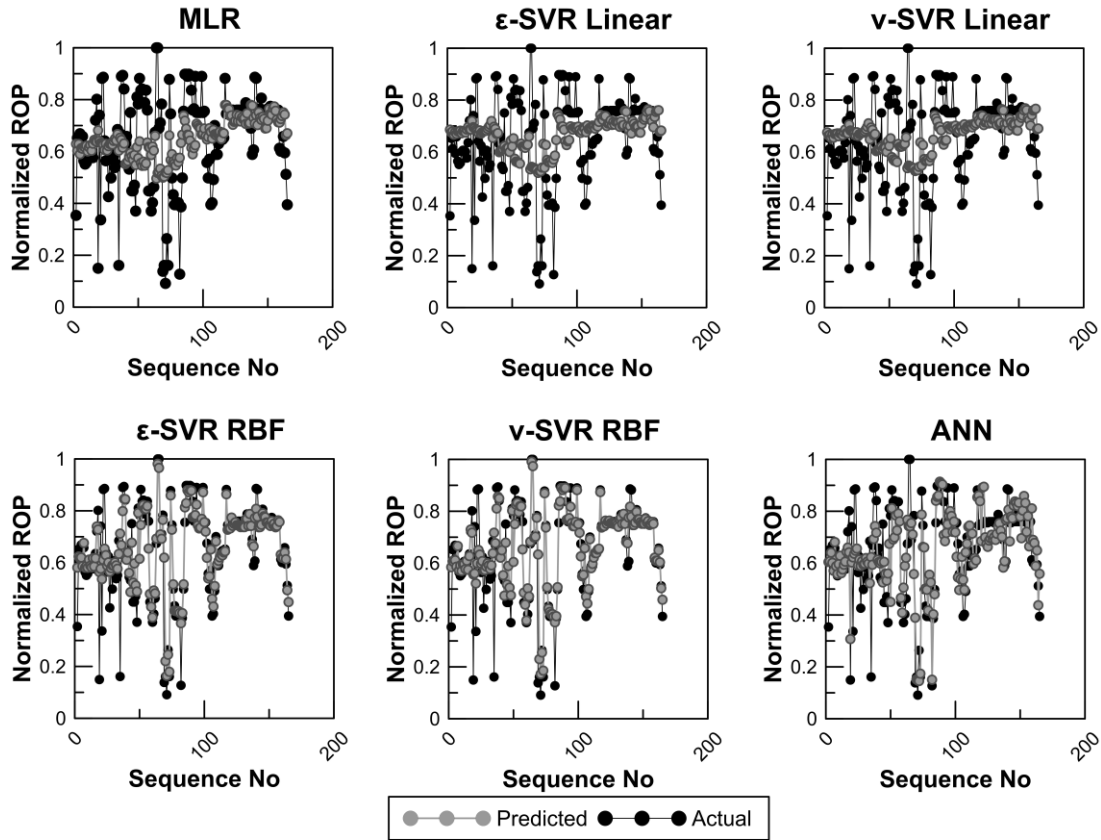


Figure D.59 : Model training for Case 3.1.2.

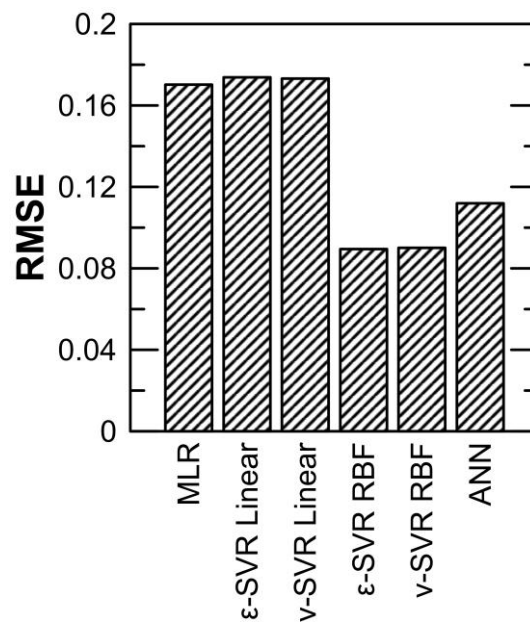


Figure D.60 : Training errors for Case 3.1.2.

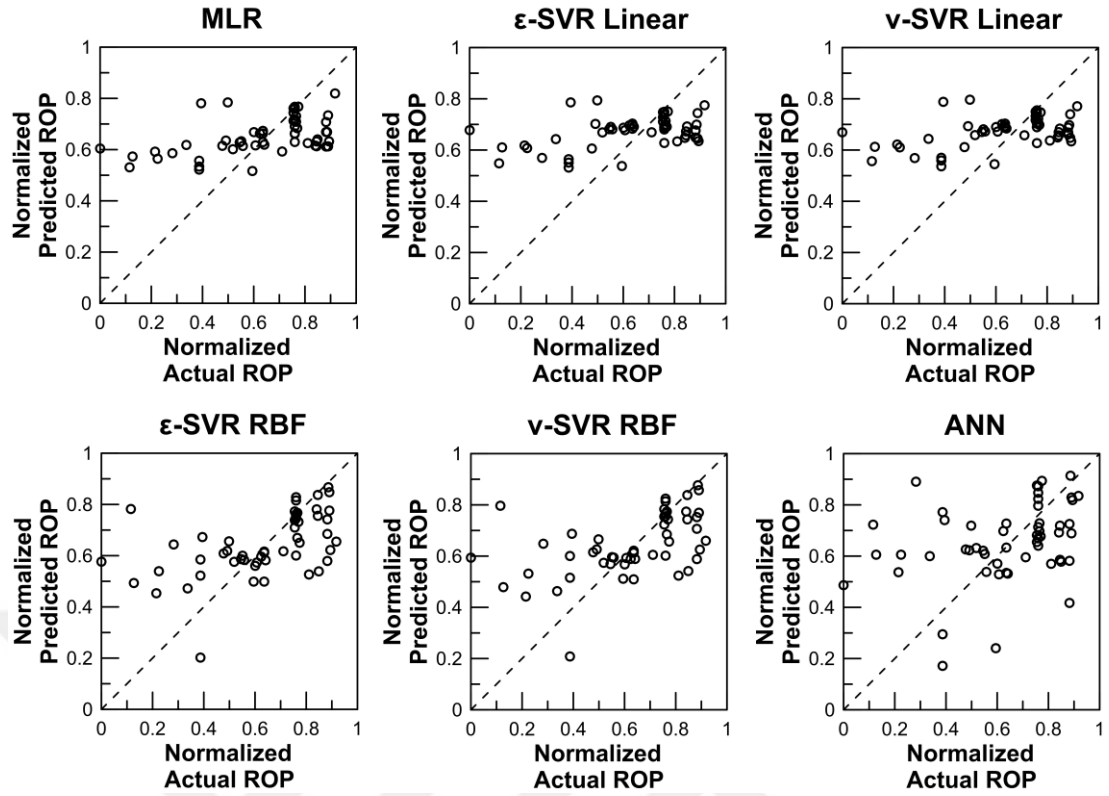


Figure D.61 : ROP predictions with testing subset for Case 3.1.2.

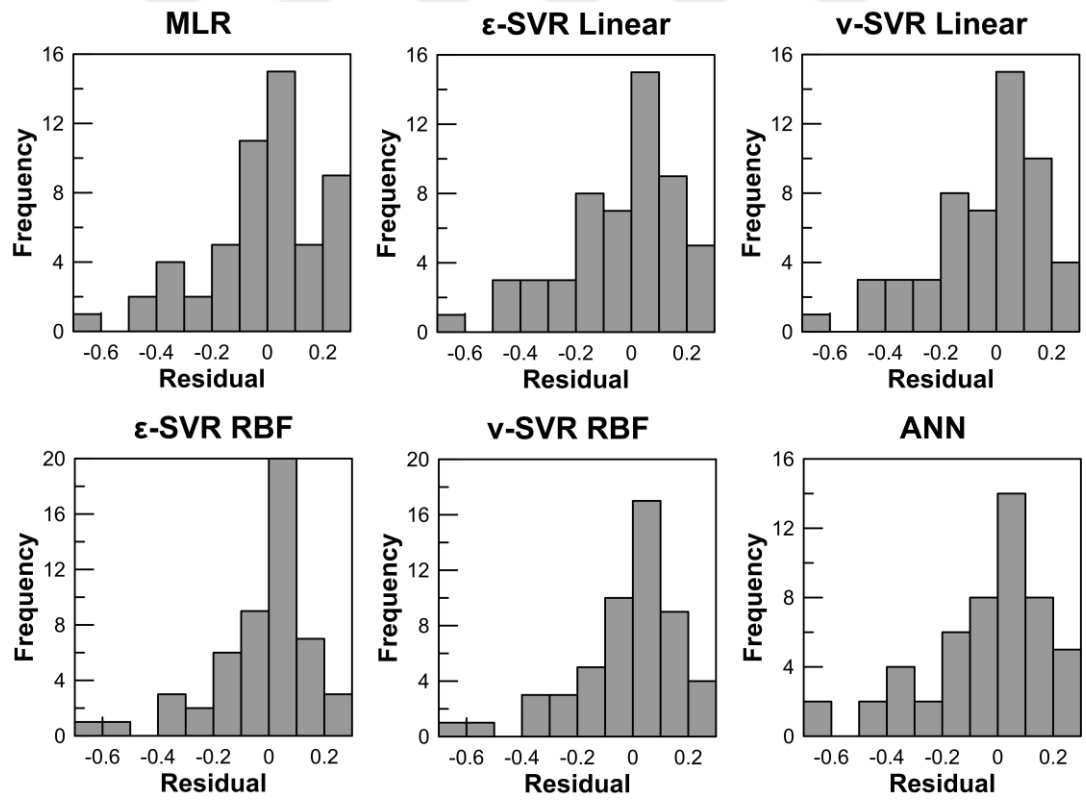


Figure D.62 : Error histograms for Case 3.1.2.

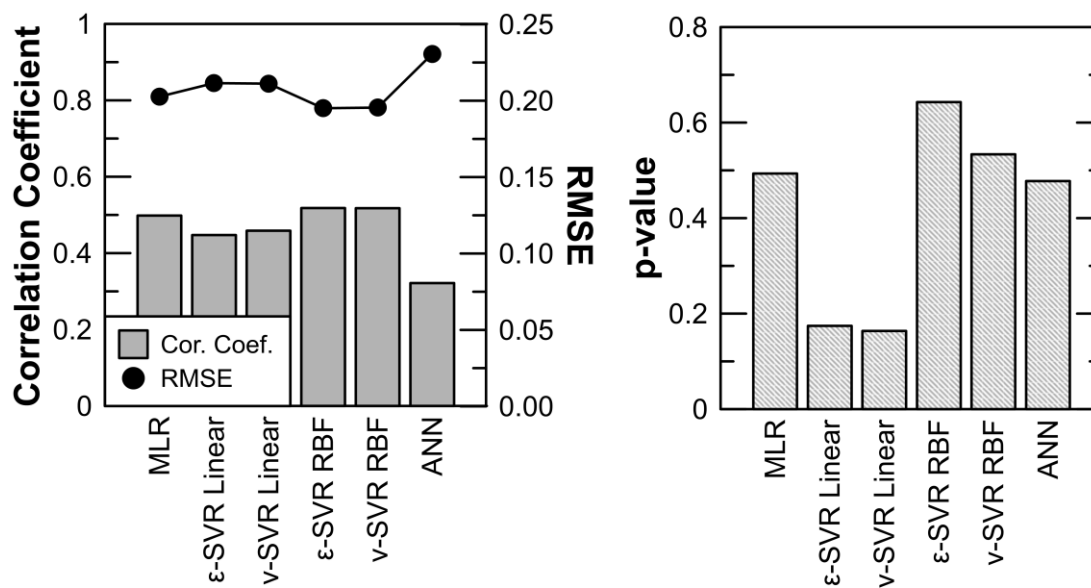


Figure D.63 : Statistical comparison for Case 3.1.2.

D.2.3. Case 3.1.3: BYM+UCS

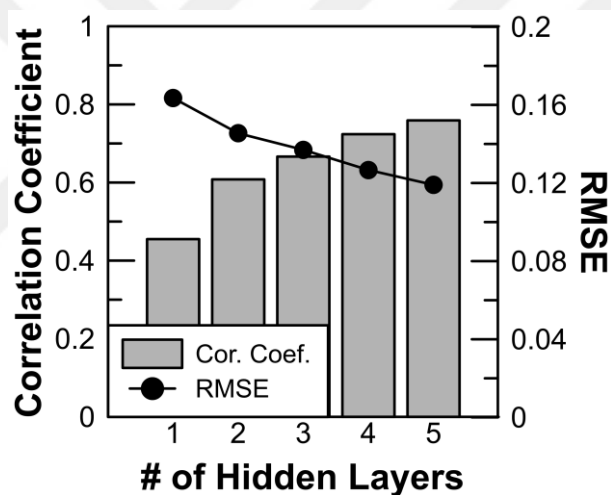


Figure D.64 : ANN hidden layer sensitivity for Case 3.1.3.

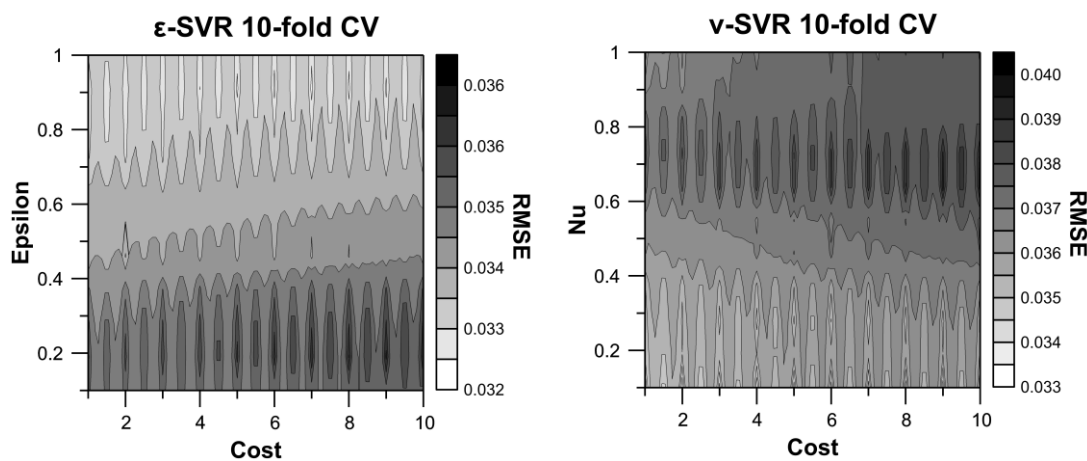


Figure D.65 : 10-fold CV grid search plots for Case 3.1.3.

Table D.10 : SVR hyper-parameters for Case 3.1.3.

ε -SVR					ν -SVR				
Linear			RBF		Linear			RBF	
C	ε	C	γ	ε	C	ν	C	γ	ν
7.0	0.9	1.5	1.0	0.2	1.0	0.1	2.0	1.0	0.4

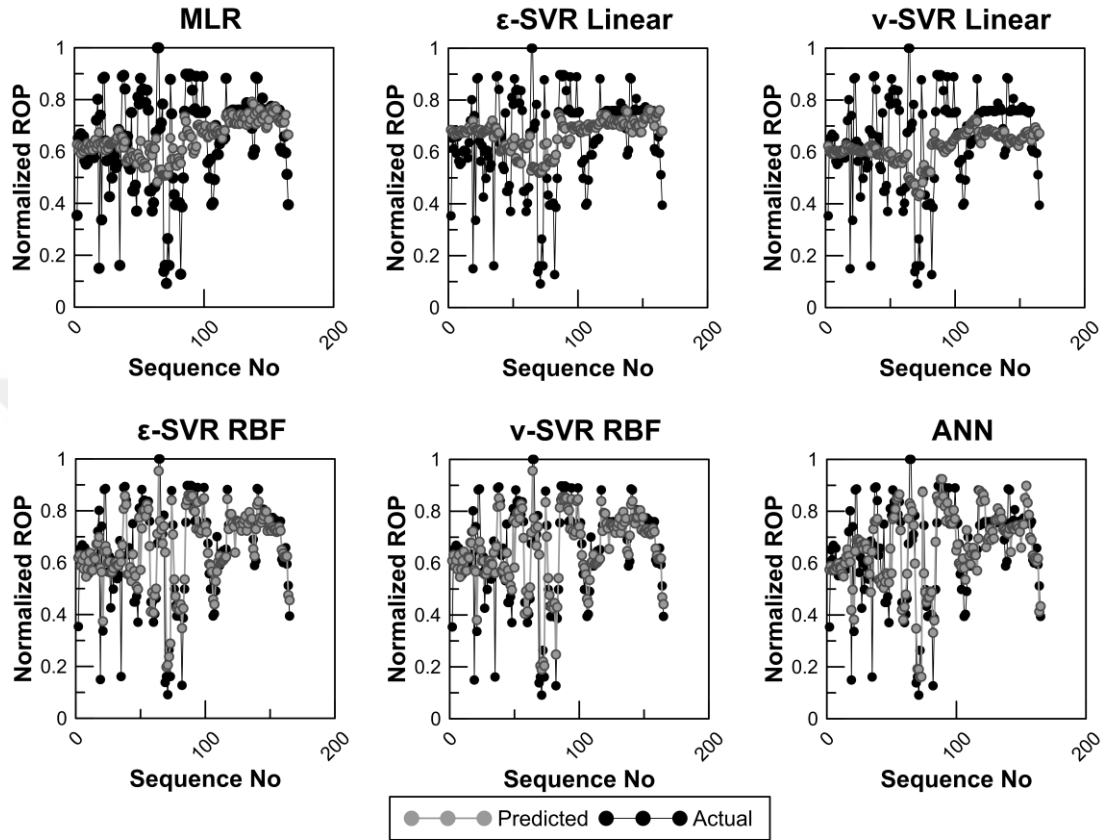


Figure D.66 : Model training for Case 3.1.3.

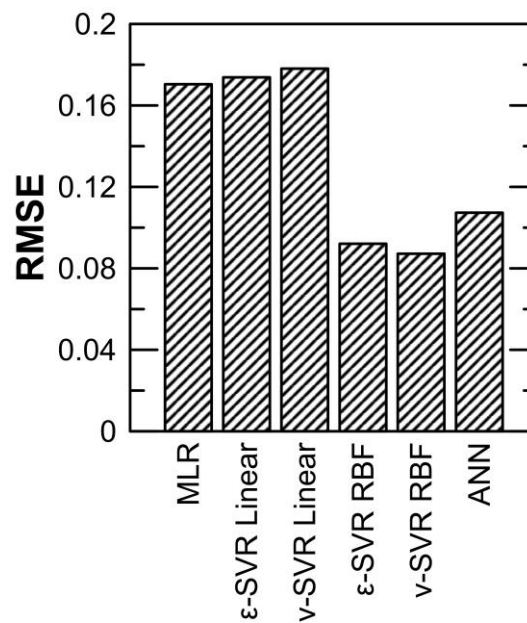


Figure D.67 : Training errors for Case 3.1.3.

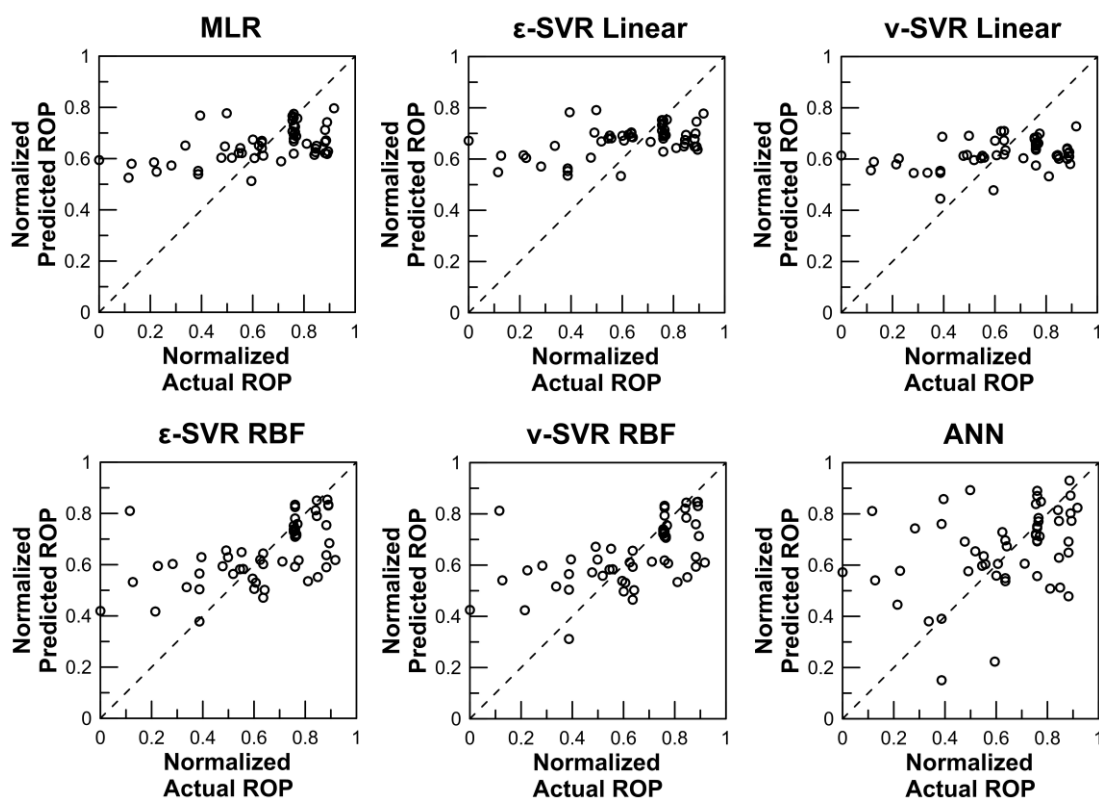


Figure D.68 : ROP predictions with testing subset for Case 3.1.3.

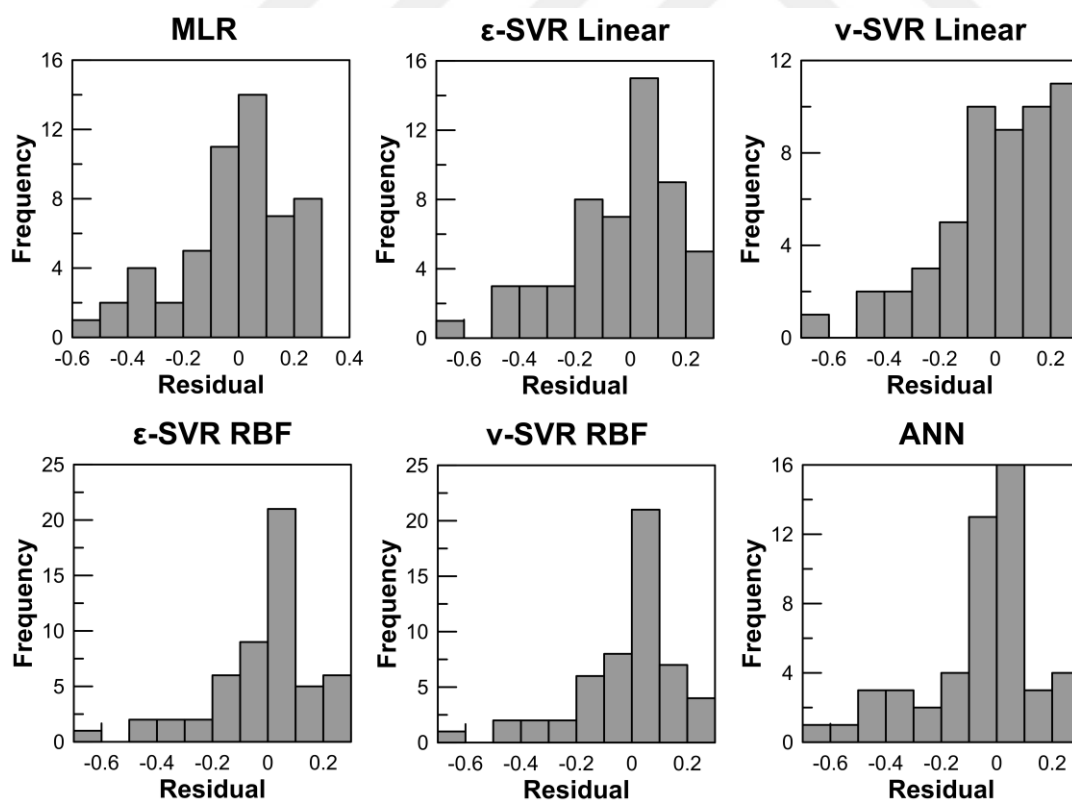


Figure D.69 : Error histograms for Case 3.1.3.

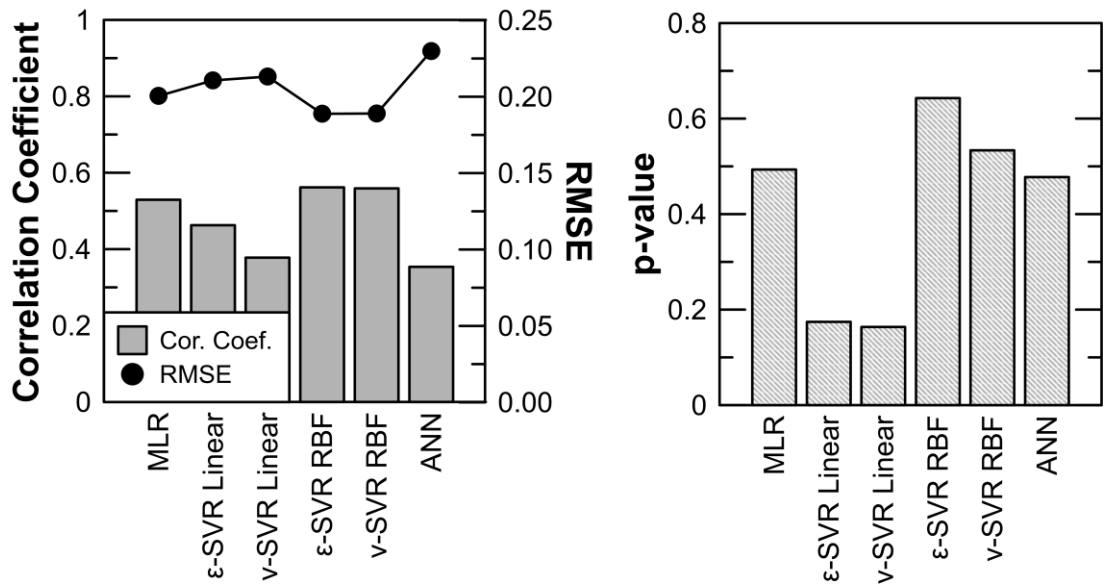


Figure D.70 : Statistical comparison for Case 3.1.3.

D.2.4. Case 3.1.4: BYM+SFG

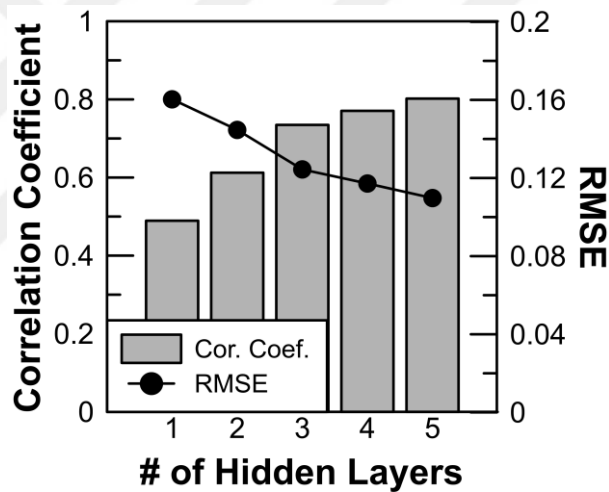


Figure D.71 : ANN hidden layer sensitivity for Case 3.1.4.

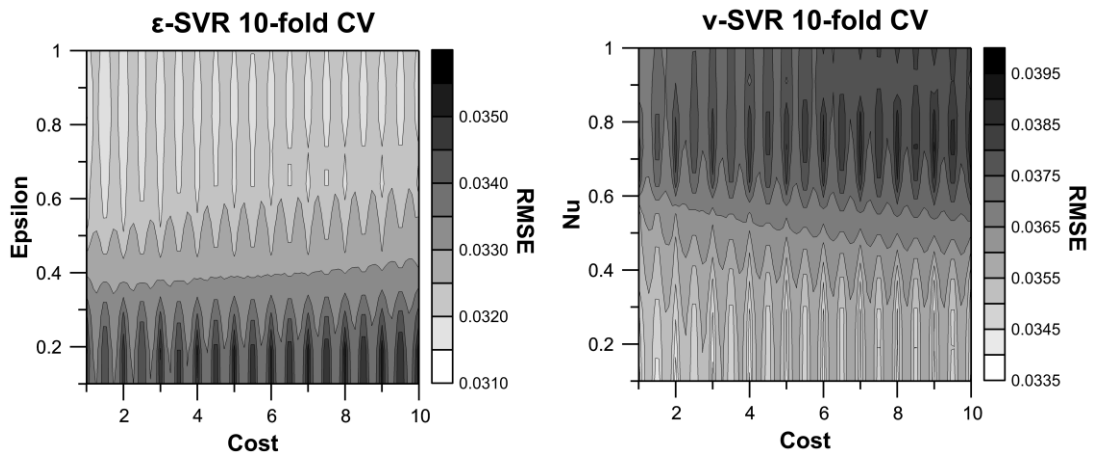


Figure D.72 : 10-fold CV grid search plots for Case 3.1.4.

Table D.11 : SVR hyper-parameters for Case 3.1.4.

ε -SVR					ν -SVR				
Linear		RBF			Linear		RBF		
C	ε	C	γ	ε	C	ν	C	γ	ν
4.5	0.85	1.0	1.0	0.2	1.0	0.1	1.5	1.0	0.6

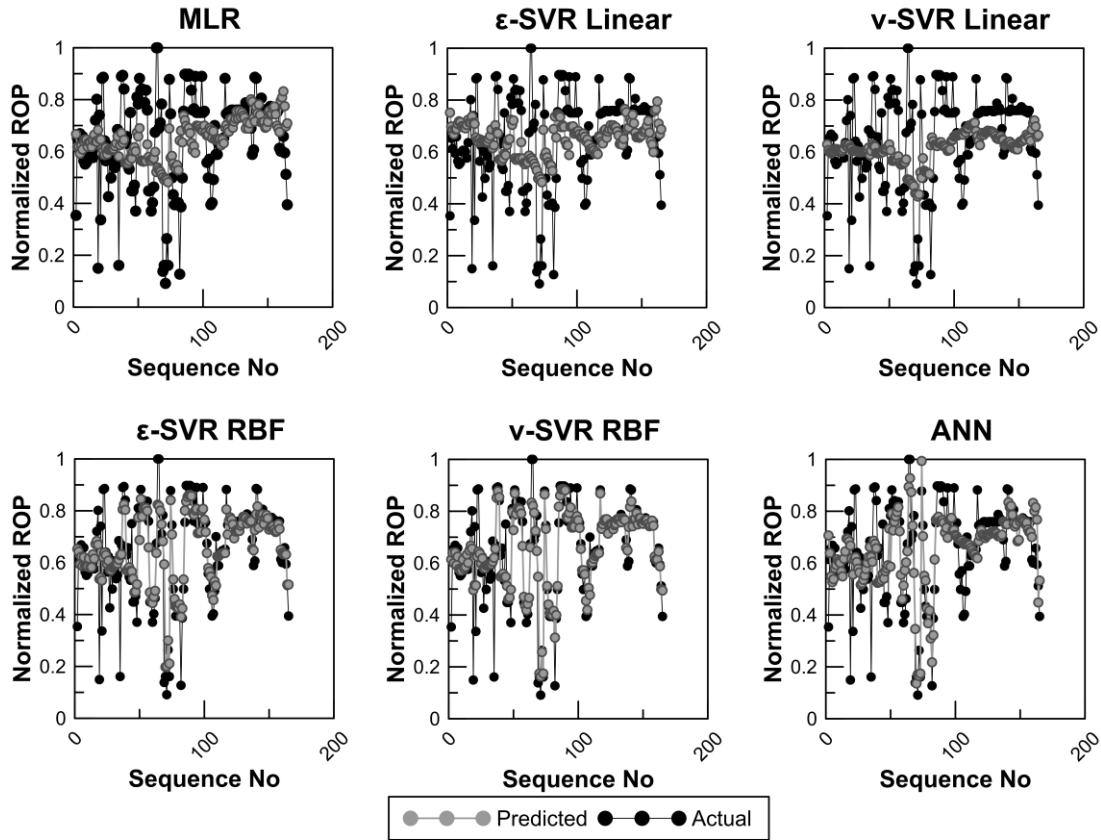


Figure D.73 : Model training for Case 3.1.4.

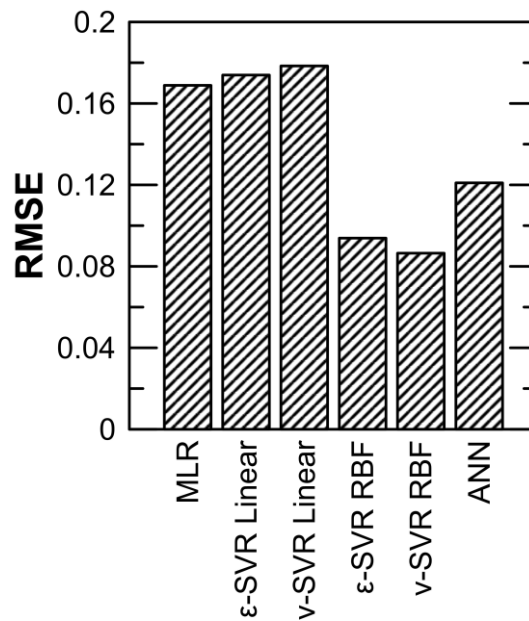


Figure D.74 : Training errors for Case 3.1.4.

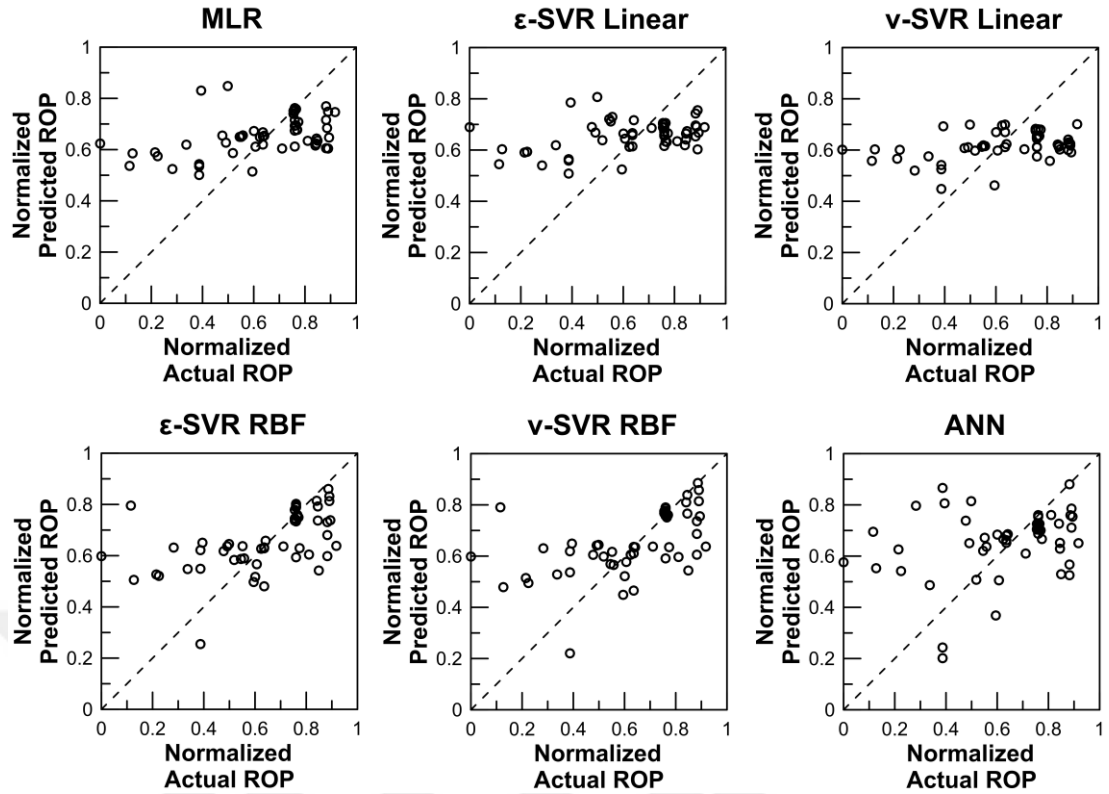


Figure D.75 : ROP predictions with testing subset for Case 3.1.4.

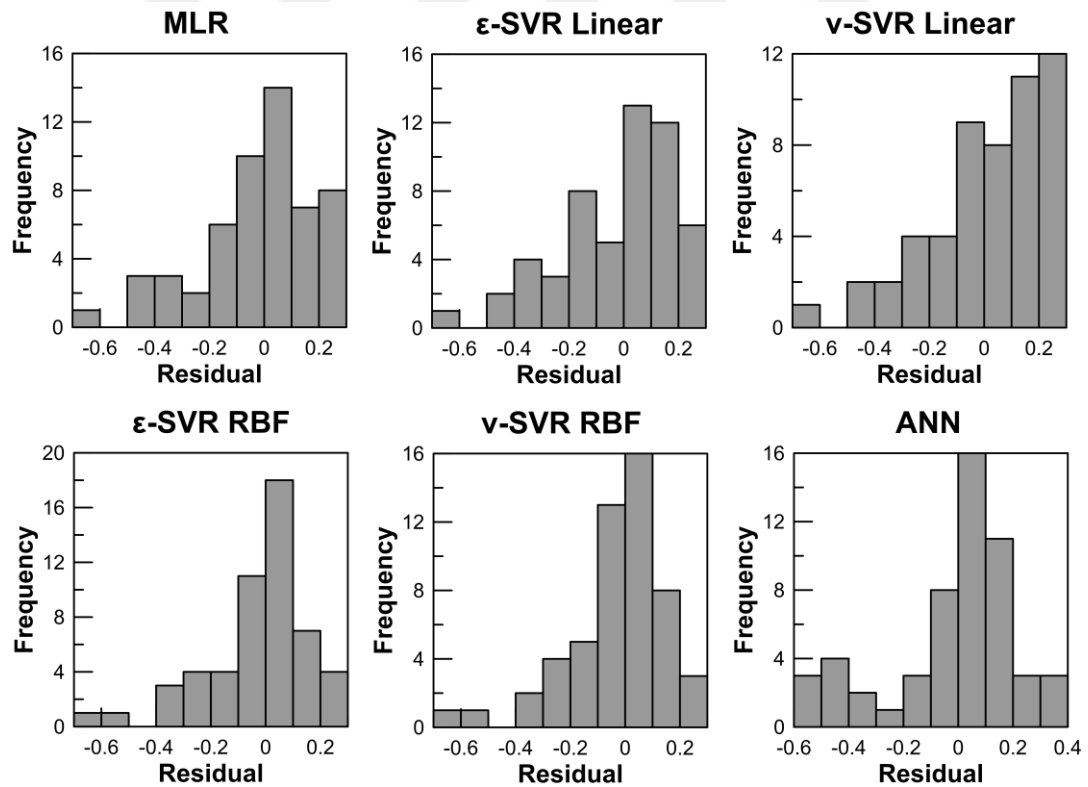


Figure D.76 : Error histograms for Case 3.1.4.

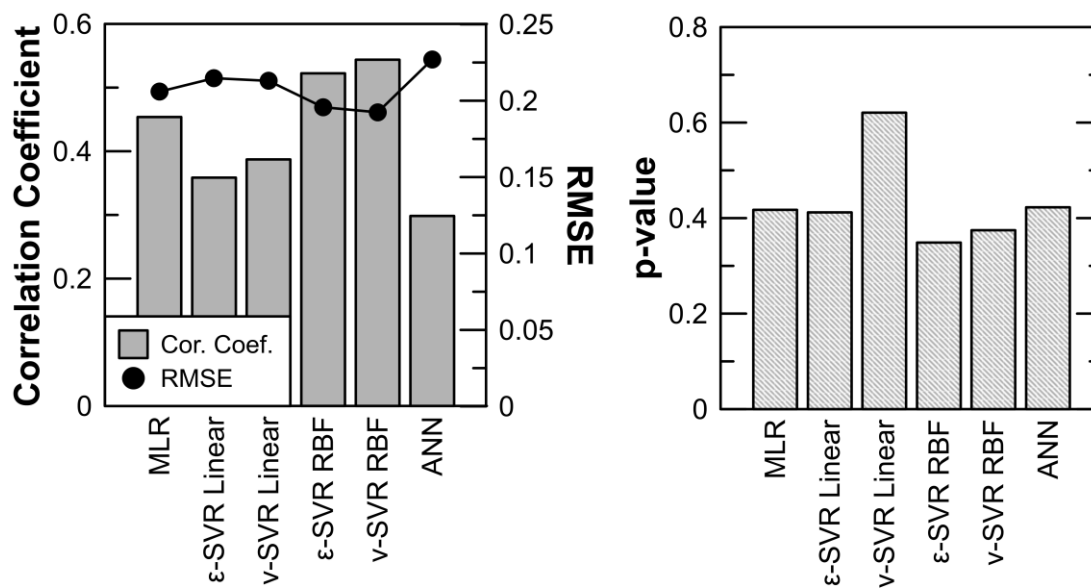


Figure D.77 : Statistical comparison for Case 3.1.4.

D.2.5. Case 3.1.5: BYM+TRQ

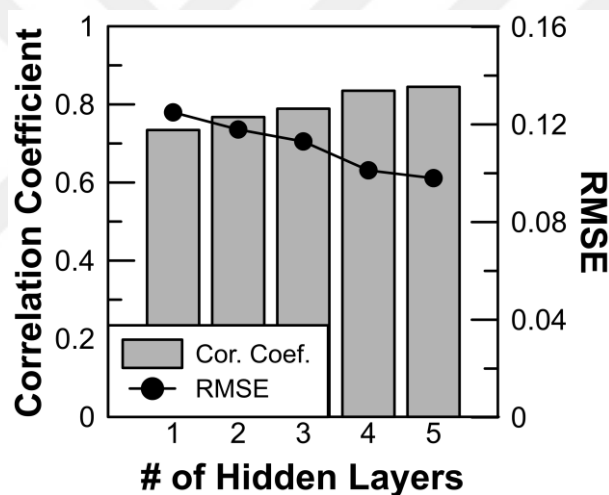


Figure D.78 : ANN hidden layer sensitivity for Case 3.1.5.

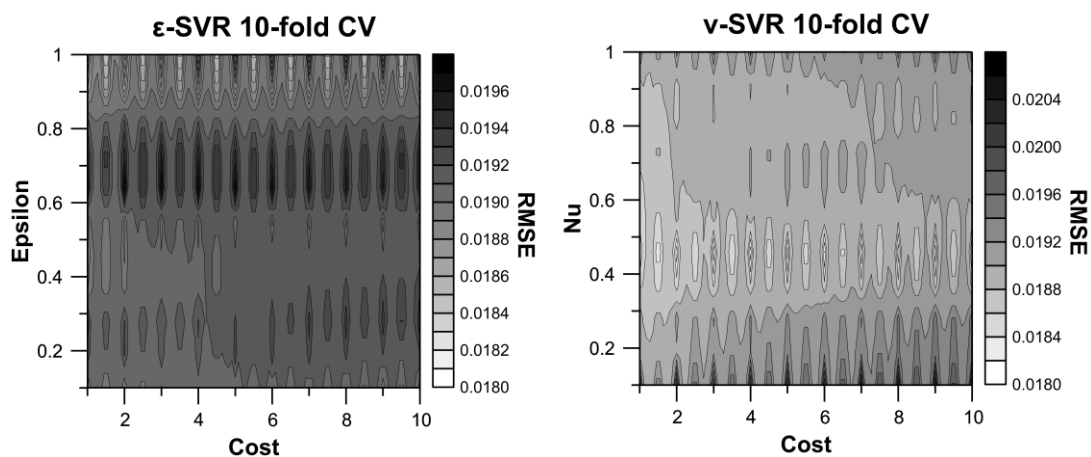


Figure D.79 : 10-fold CV grid search plots for Case 3.1.5.

Table D.12 : SVR hyper-parameters for Case 3.1.5.

ε -SVR					ν -SVR				
Linear			RBF		Linear		RBF		
C	ε	C	γ	ε	C	ν	C	γ	ν
7.5	1.0	1.0	1.0	0.1	1.5	0.45	1.5	1.0	1.0

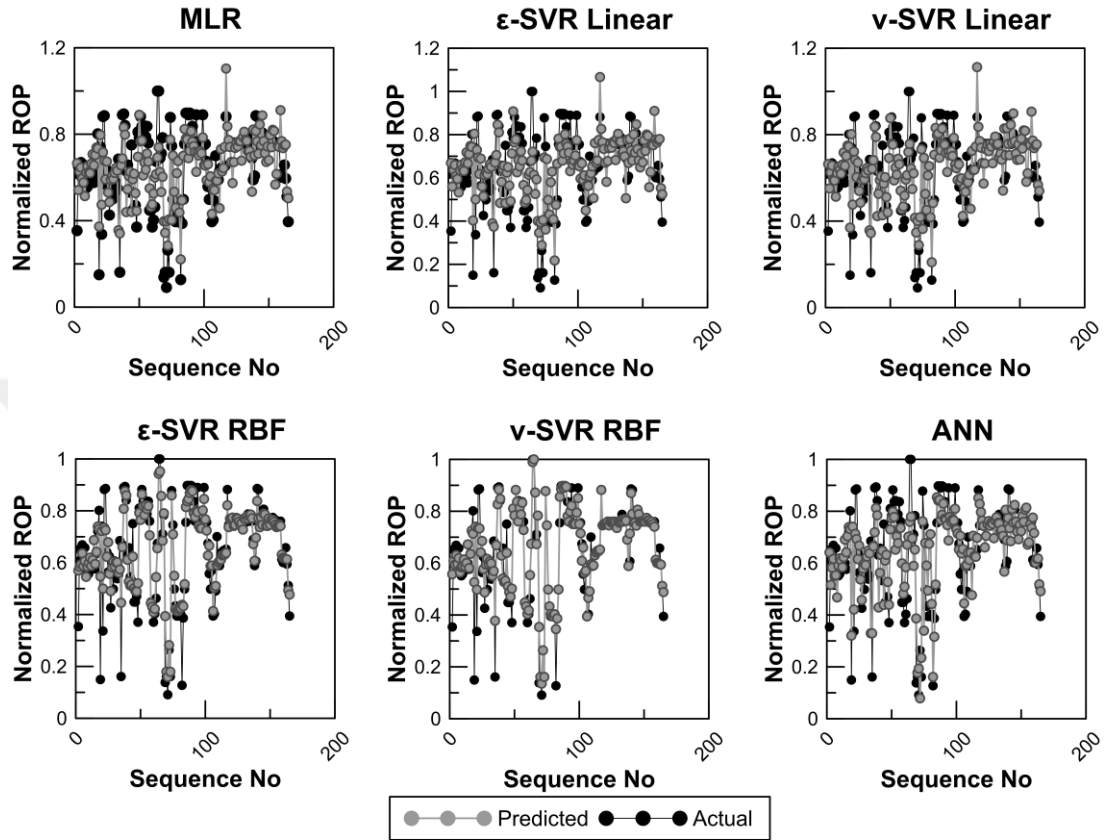


Figure D.80 : Model training for Case 3.1.5.

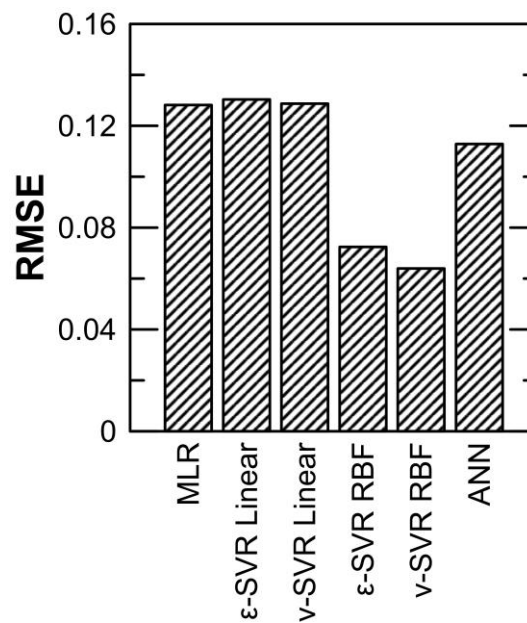


Figure D.81 : Training errors for Case 3.1.5.

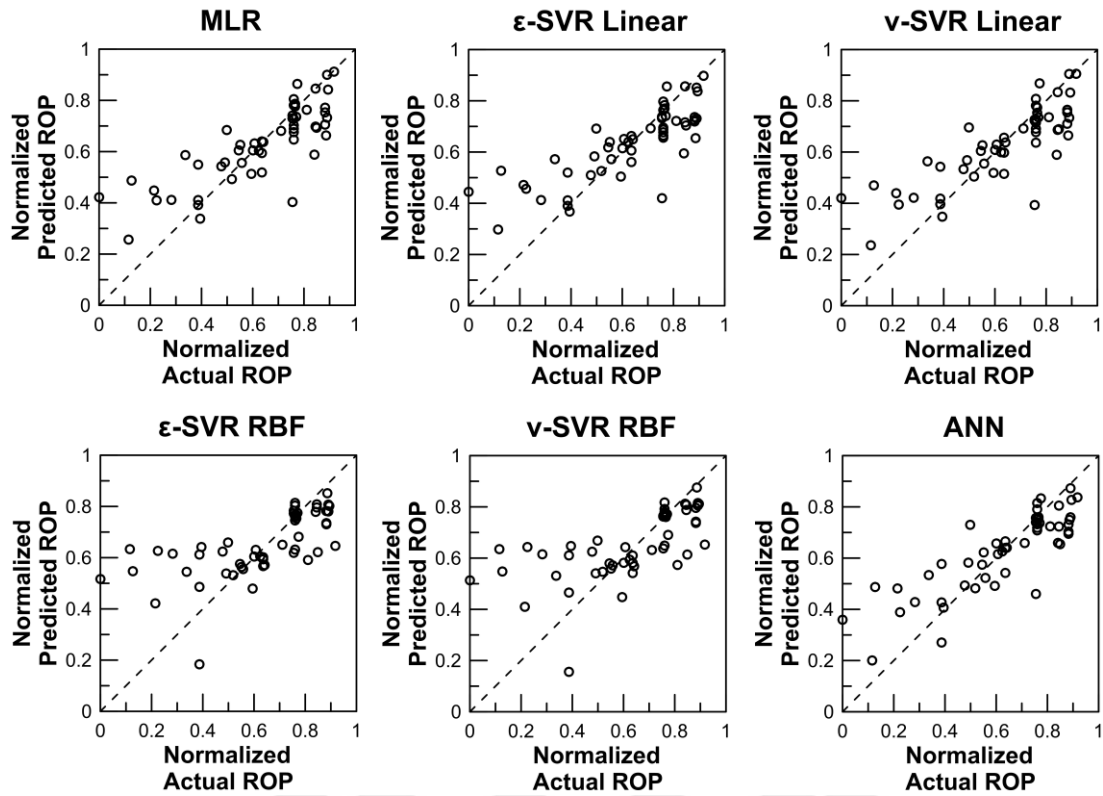


Figure D.82 : ROP predictions with testing subset for Case 3.1.5.

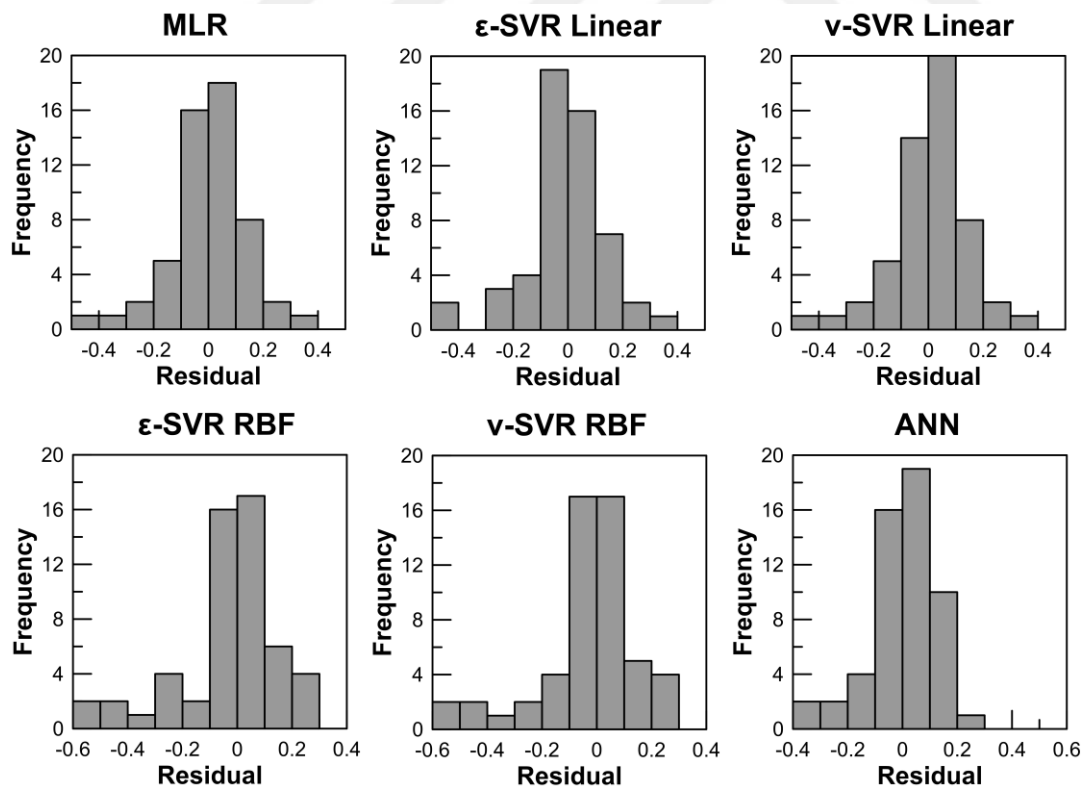


Figure D.83 : Error histograms for Case 3.1.5.

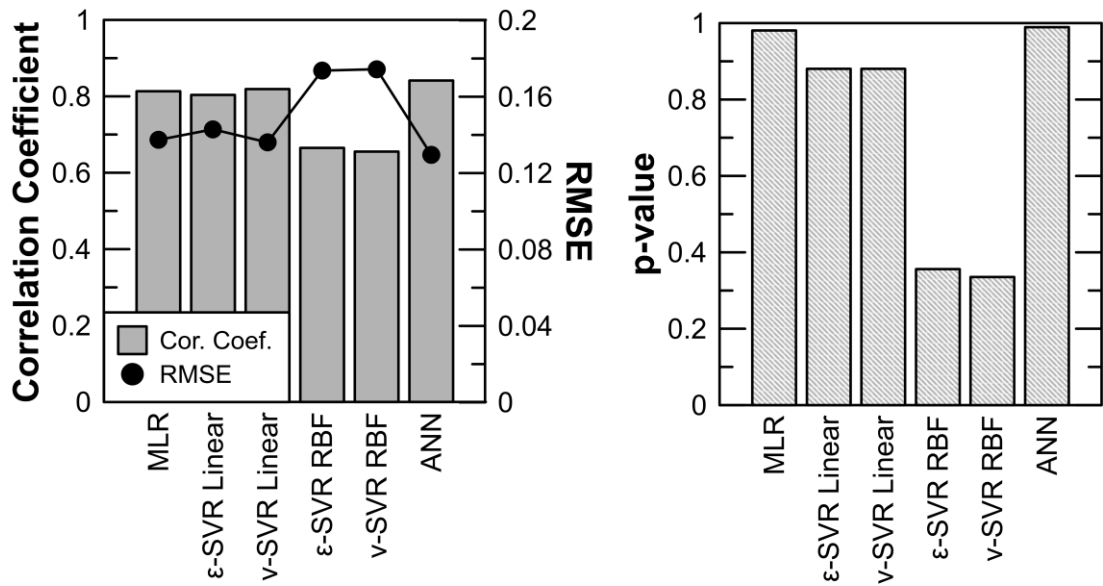


Figure D.84 : Statistical comparison for Case 3.1.5.

D.2.6. Case 3.1.6: BYM+All

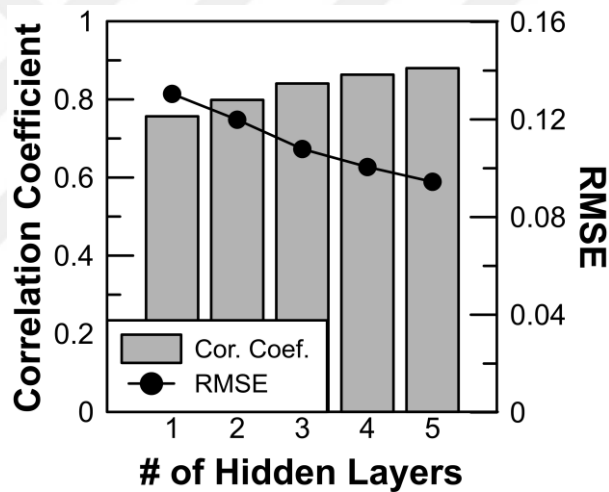


Figure D.85 : ANN hidden layer sensitivity for Case 3.1.6.

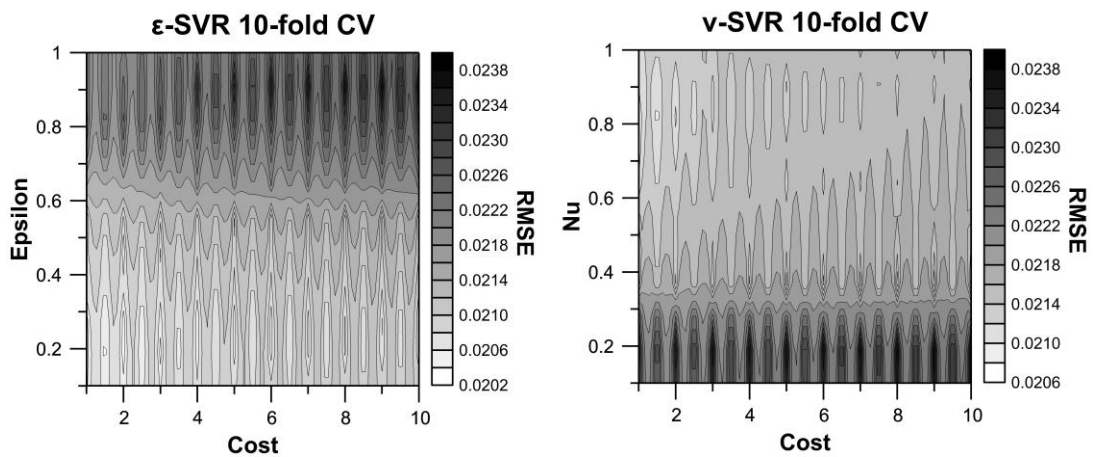


Figure D.86 : 10-fold CV grid search plots for Case 3.1.6.

Table D.13 : SVR hyper-parameters for Case 3.1.6.

ε -SVR					ν -SVR				
Linear			RBF		Linear			RBF	
C	ε	C	γ	ε	C	ν	C	γ	ν
1.5	0.25	2.5	1.0	0.1	1.0	0.85	2.5	1.0	0.7

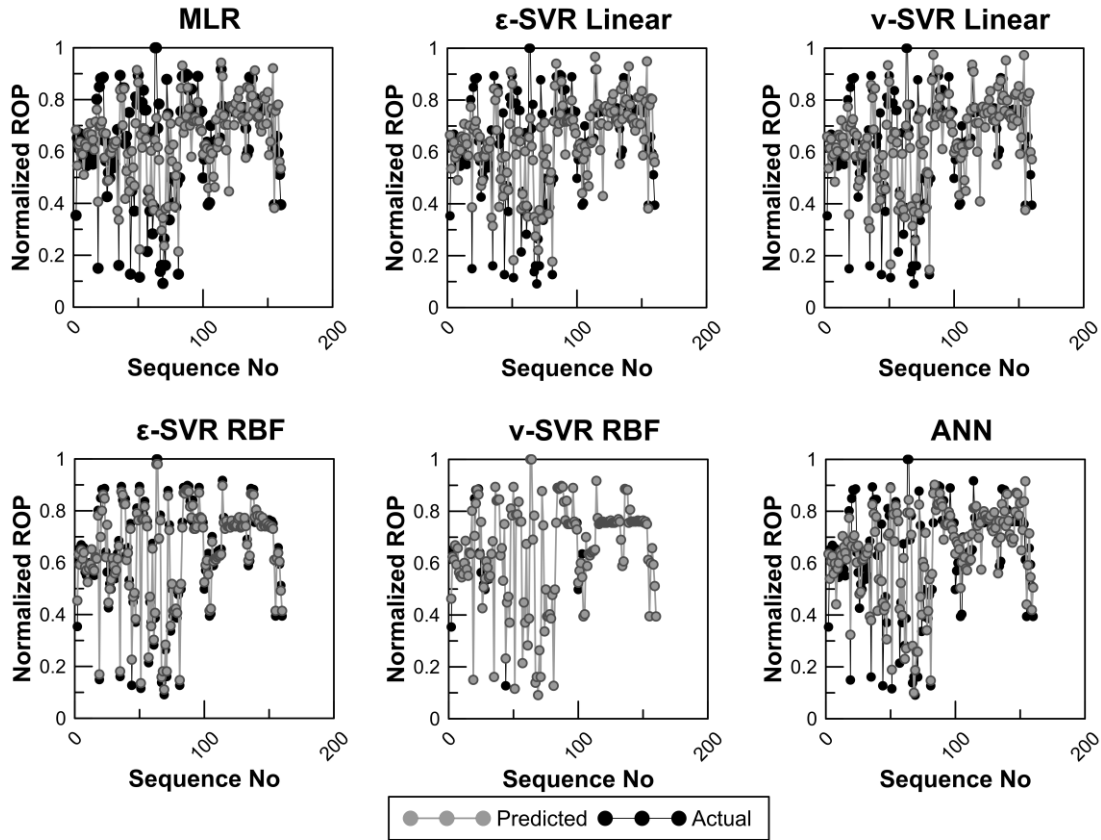


Figure D.87 : Model training for Case 3.1.6.

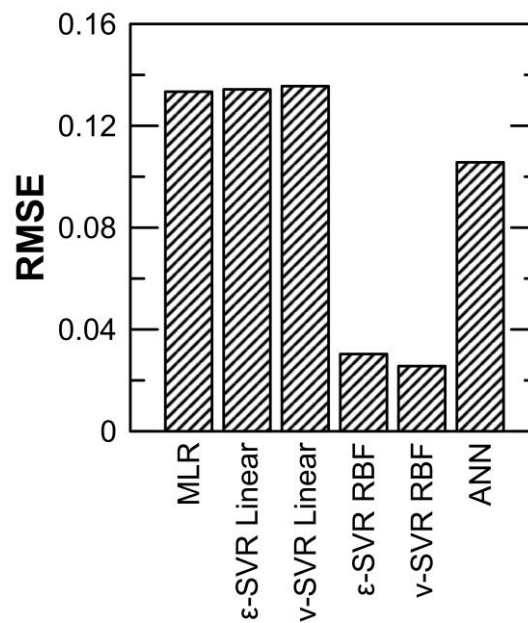


Figure D.88 : Training errors for Case 3.1.6.

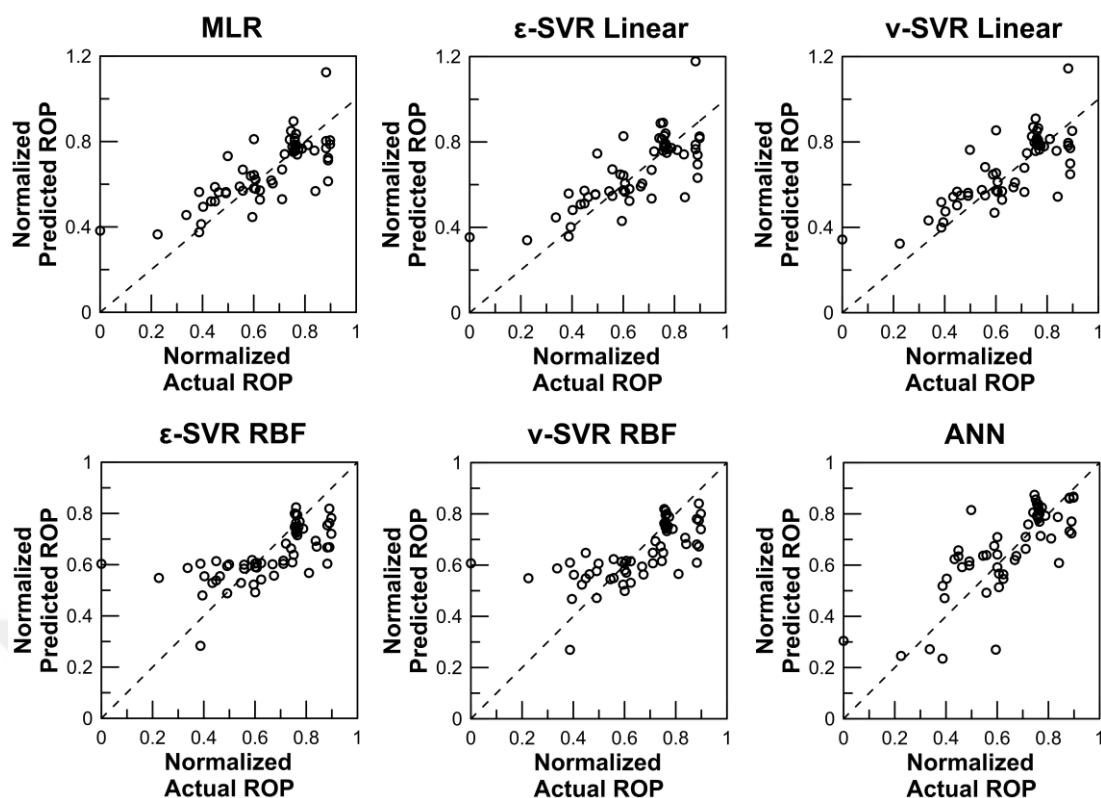


Figure D.89 : ROP predictions with testing subset for Case 3.1.6.

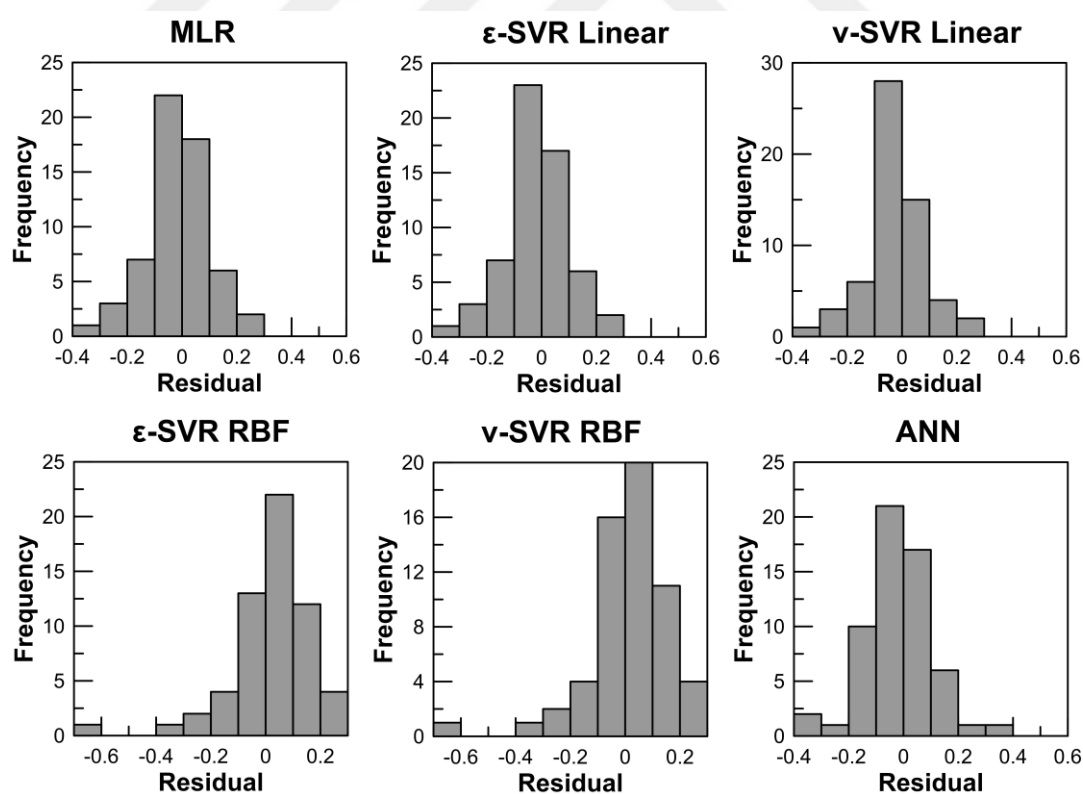


Figure D.90 : Error histograms for Case 3.1.6.

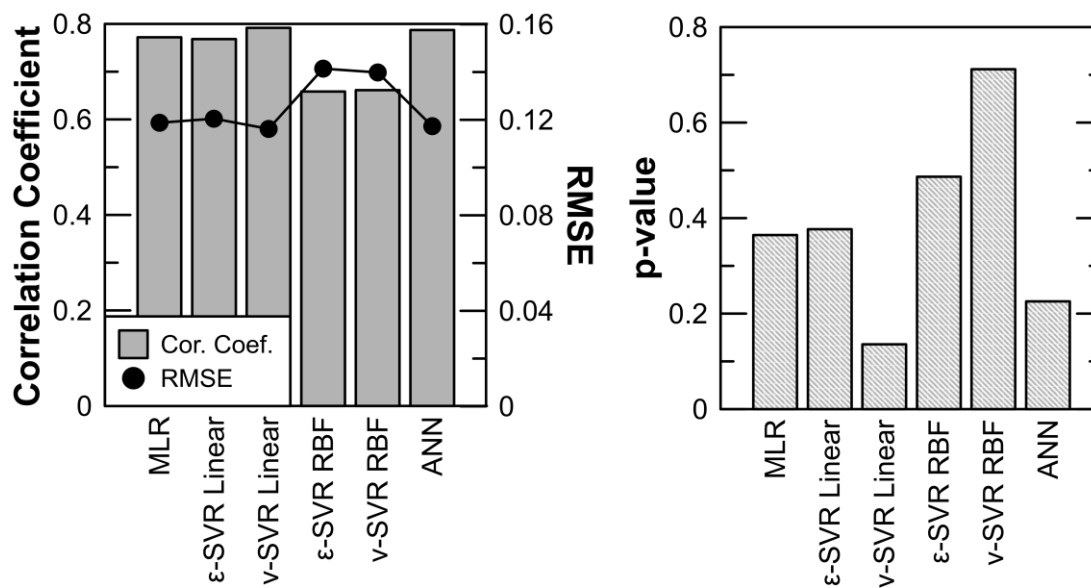


Figure D.91 : Statistical comparison for Case 3.1.6.

D.2.7. Case 3.1.7: Feature Selection (x_2+x_5+TRQ)

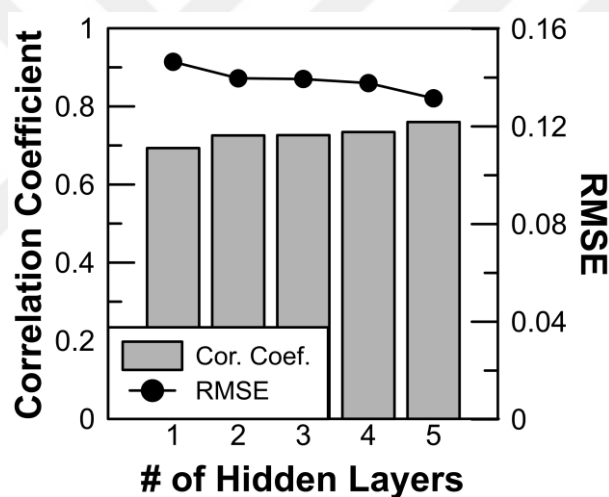


Figure D.92 : ANN hidden layer sensitivity for Case 3.1.7.

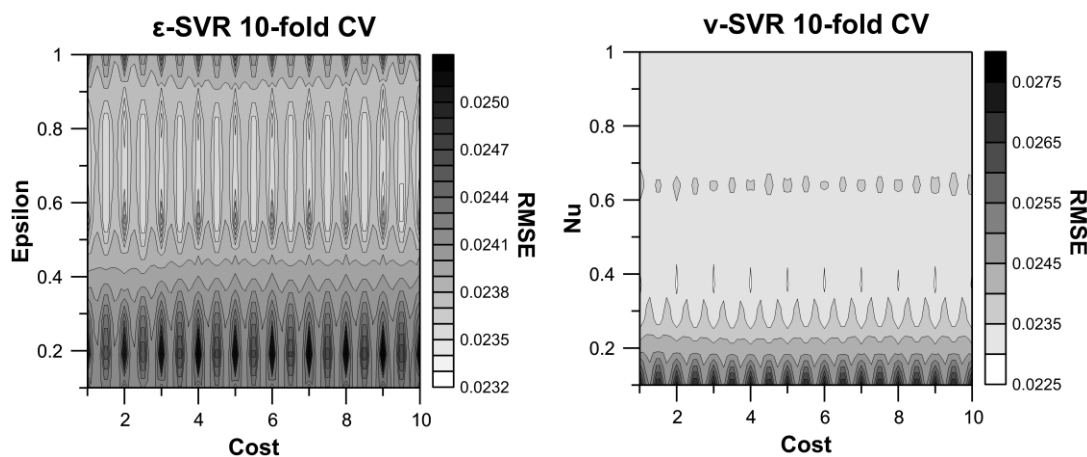


Figure D.93 : 10-fold CV grid search plots for Case 3.1.7.

Table D.14 : SVR hyper-parameters for Case 3.1.7.

ε -SVR					v-SVR				
Linear			RBF		Linear			RBF	
C	ε	C	γ	ε	C	ν	C	γ	ν
1.5	0.55	5.0	1.0	0.4	1.0	0.35	4.0	1.0	0.2

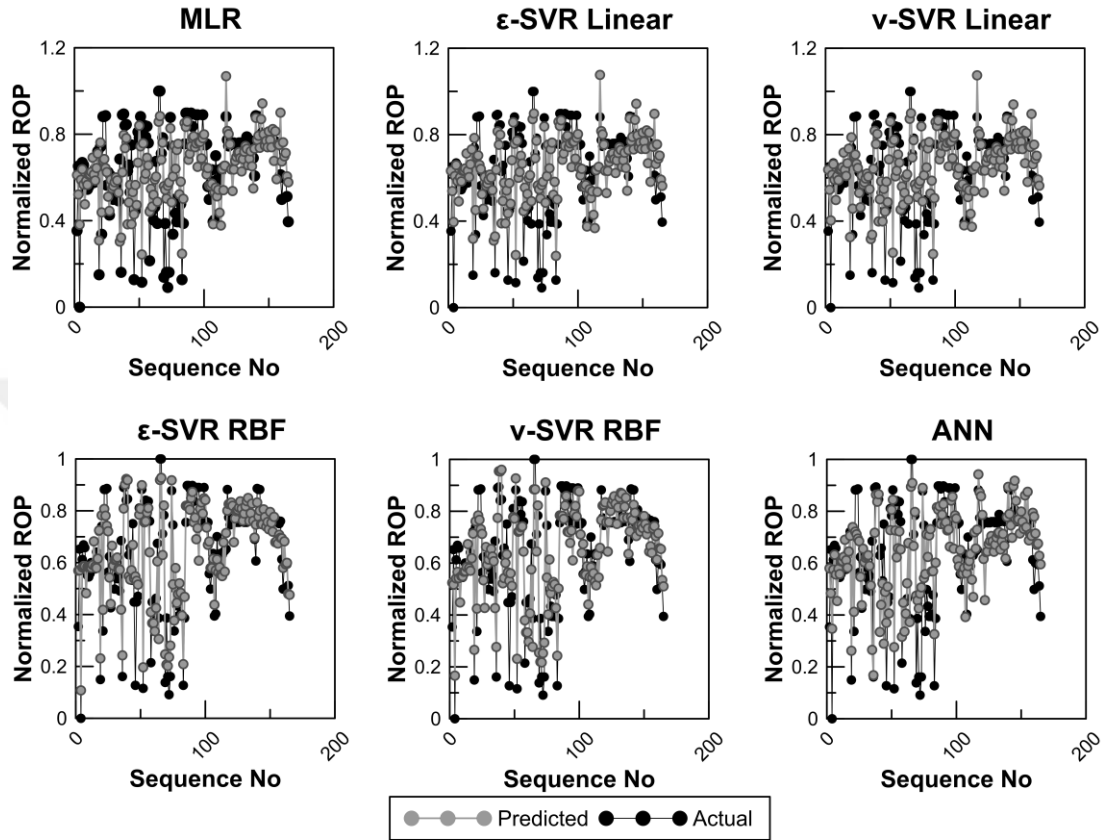


Figure D.94 : Model training for Case 3.1.7.

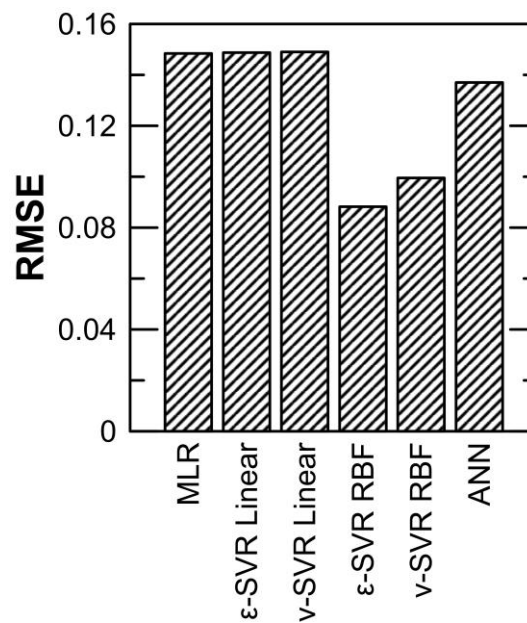


Figure D.95 : Training errors for Case 3.1.7.

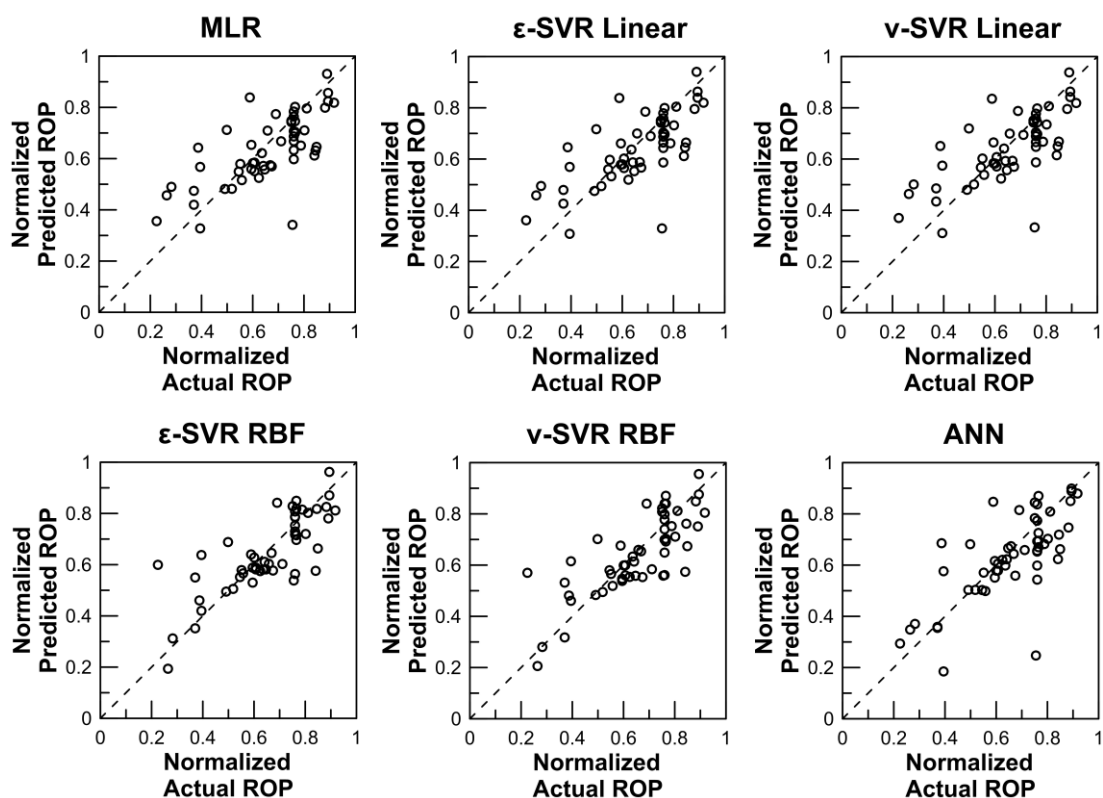


Figure D.96 : ROP predictions with testing subset for Case 3.1.7.

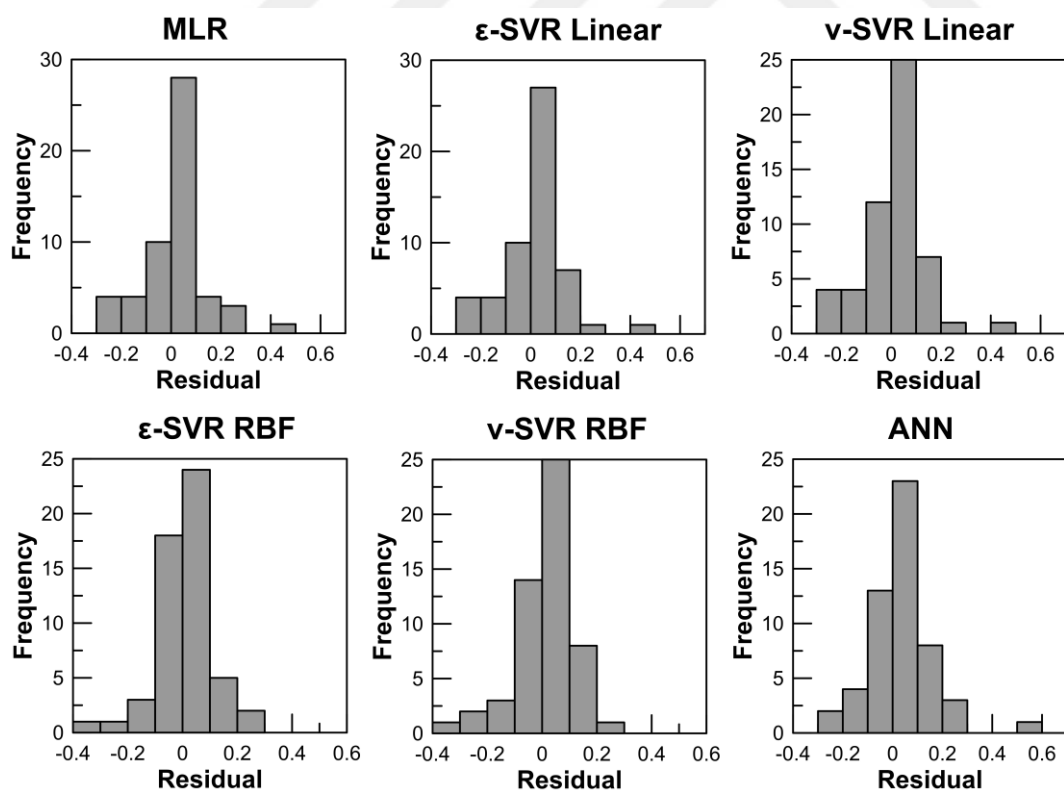


Figure D.97 : Error histograms for Case 3.1.7.

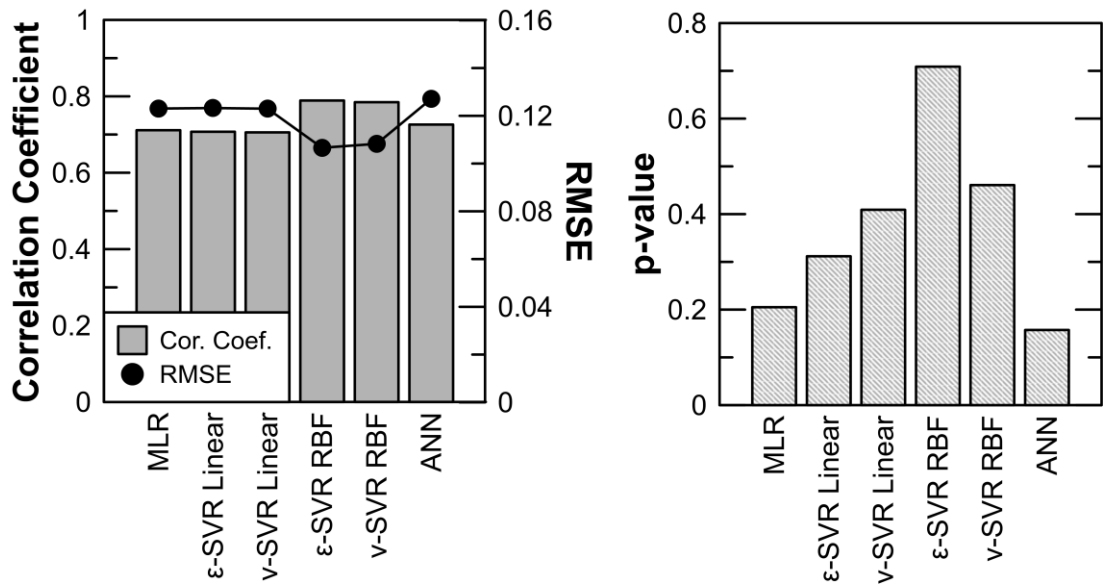


Figure D.98 : Statistical comparison for Case 3.1.7.

D.3. Intermediate Steps for Case 3.2

D.3.1. Case 3.2.1: Default BYM

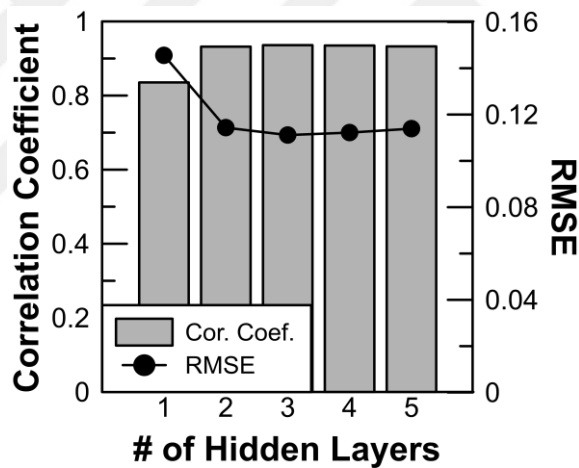


Figure D.99 : ANN hidden layer sensitivity for Case 3.2.1.

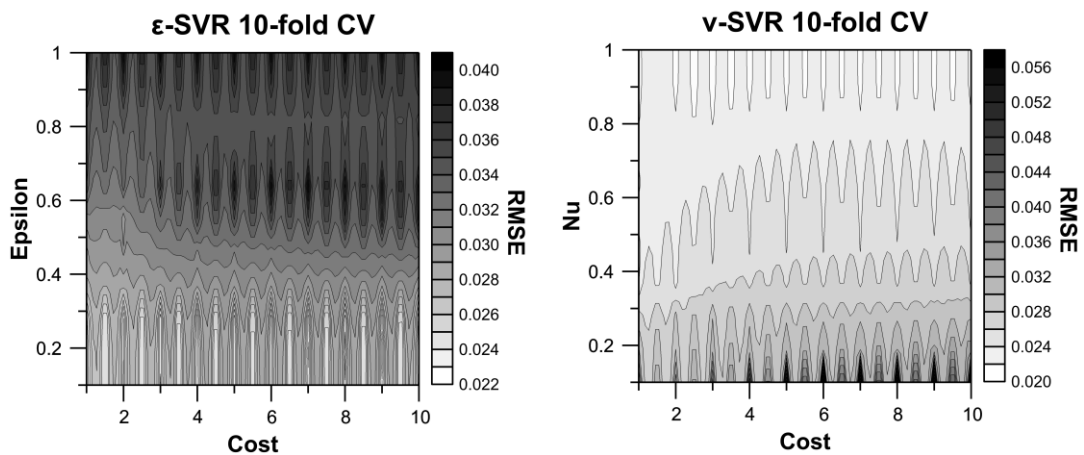


Figure D.100 : 10-fold CV grid search plots for Case 3.2.1.

Table D.15 : SVR hyper-parameters for Case 3.2.1.

ε -SVR					ν -SVR				
Linear		RBF			Linear		RBF		
C	ε	C	γ	ε	C	ν	C	γ	ν
1.5	0.25	3.0	1.0	0.2	3.0	0.95	2.5	1.0	0.2

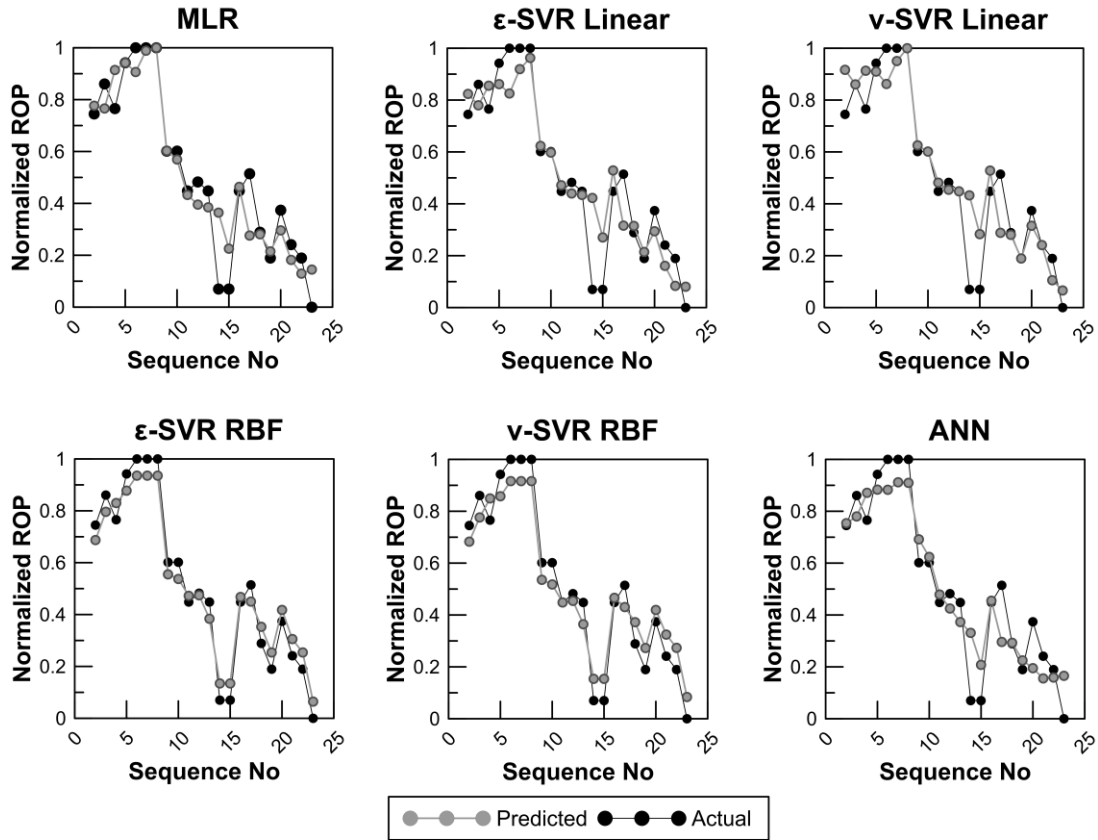


Figure D.101 : Model training for Case 3.2.1.

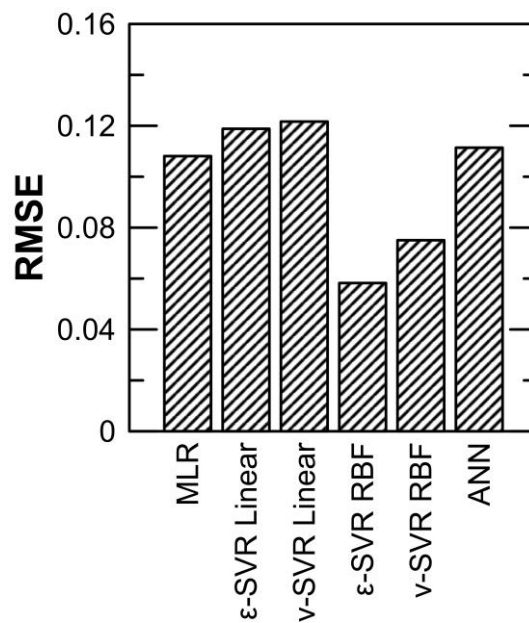


Figure D.102 : Training errors for Case 3.2.1.

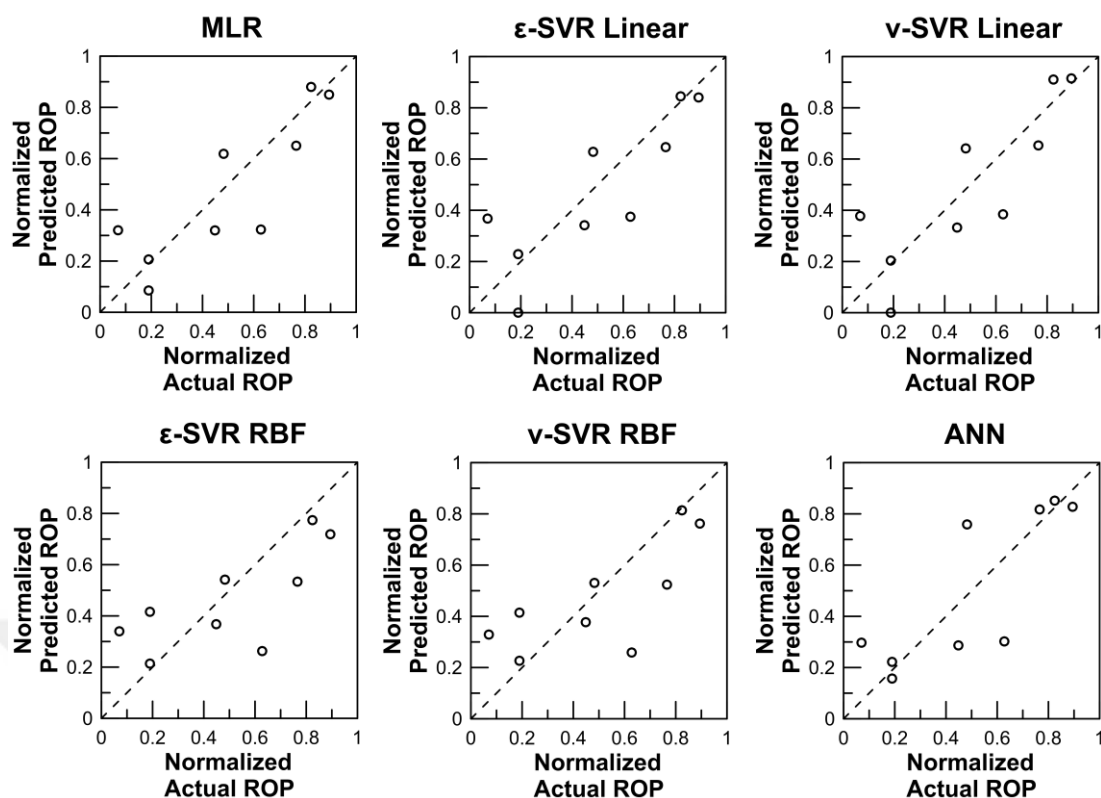


Figure D.103 : ROP predictions with testing subset for Case 3.2.1.

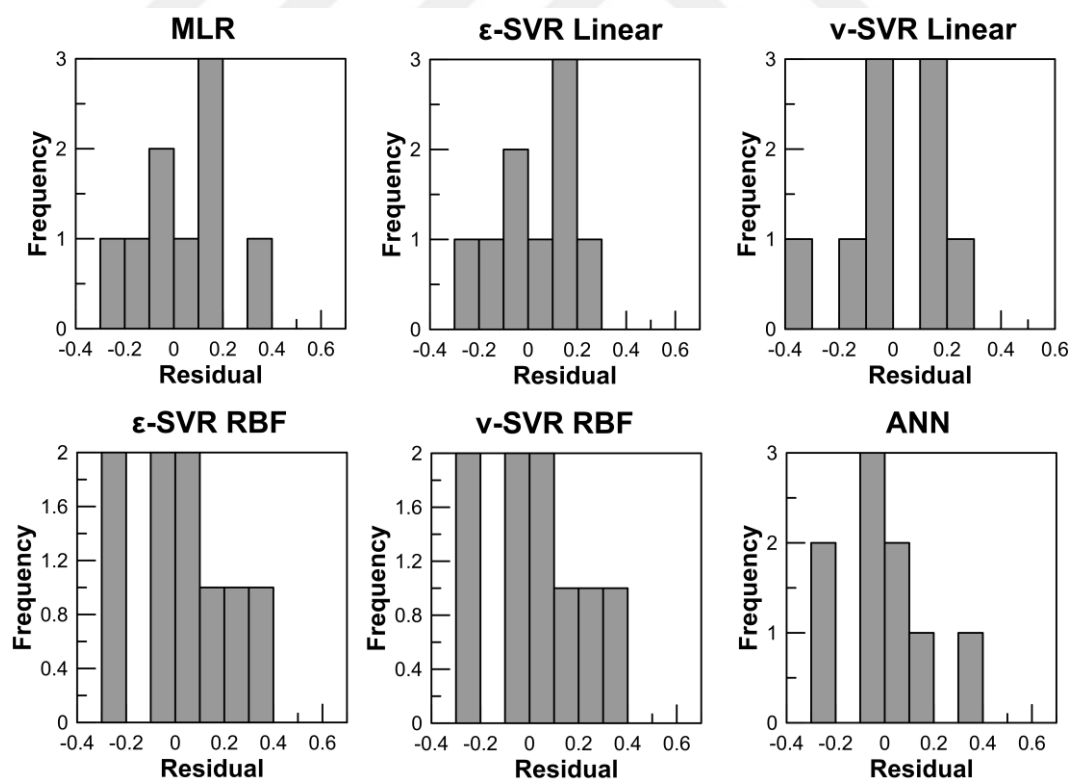


Figure D.104 : Error histograms for Case 3.2.1.

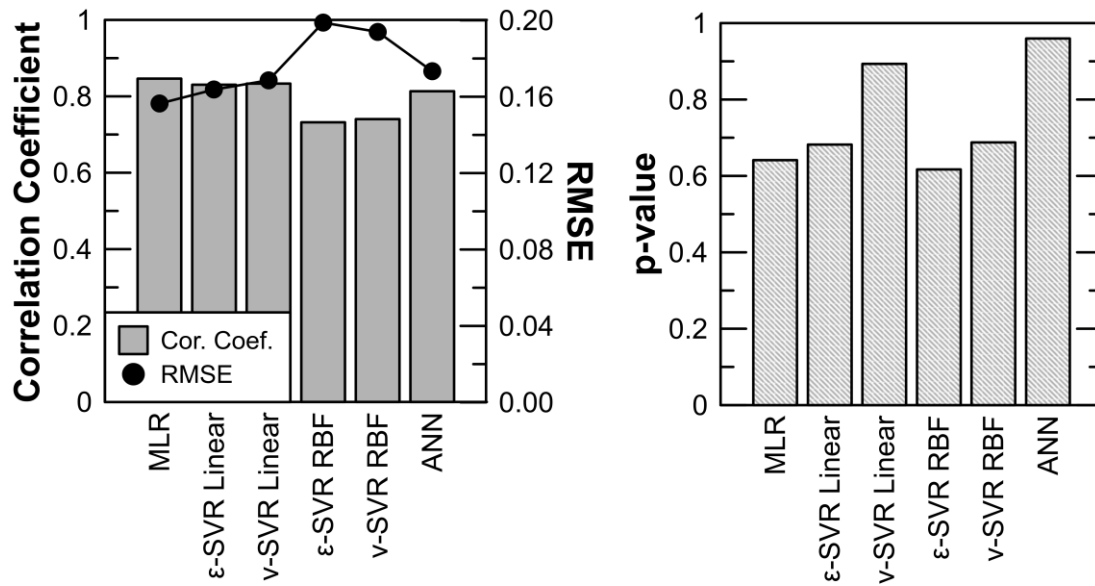


Figure D.105 : Statistical comparison for Case 3.2.1.

D.3.2. Case 3.2.2: BYM+SI

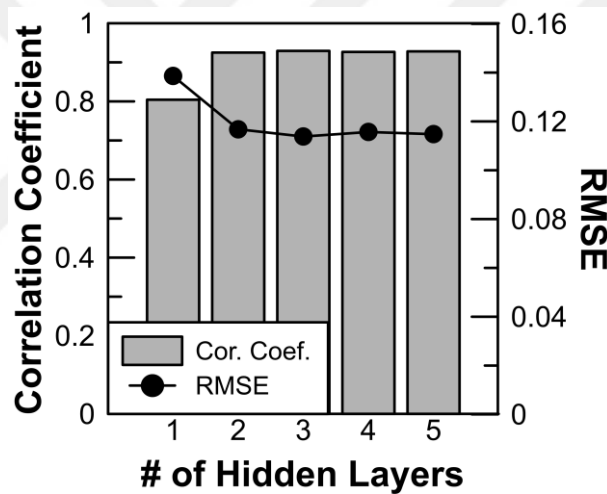


Figure D.106 : ANN hidden layer sensitivity for Case 3.2.2.

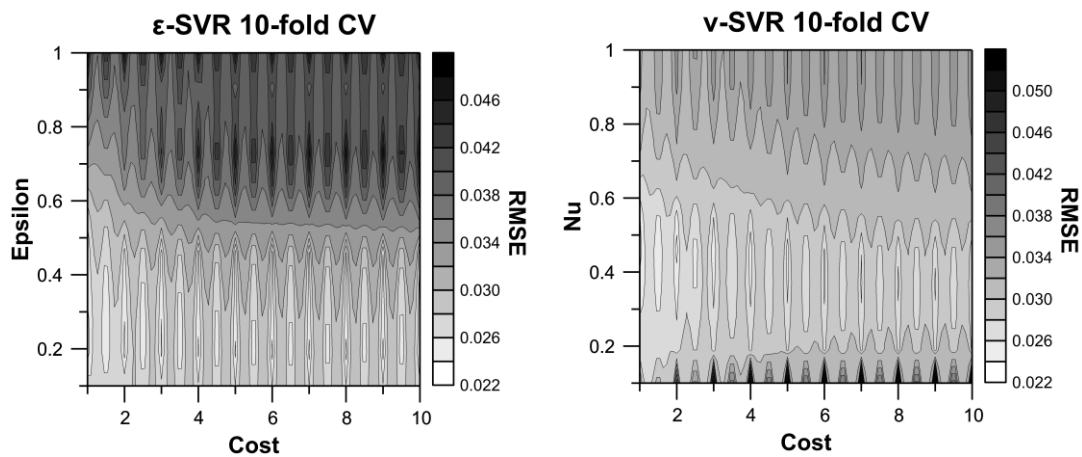


Figure D.107 : 10-fold CV grid search plots for Case 3.2.2.

Table D.16 : SVR hyper-parameters for Case 3.2.2.

ϵ -SVR					ν -SVR				
Linear		RBF			Linear		RBF		
C	ϵ	C	γ	ϵ	C	ν	C	γ	ν
1.0	0.25	1.5	1.0	0.1	1.0	0.5	1.5	1.0	0.6

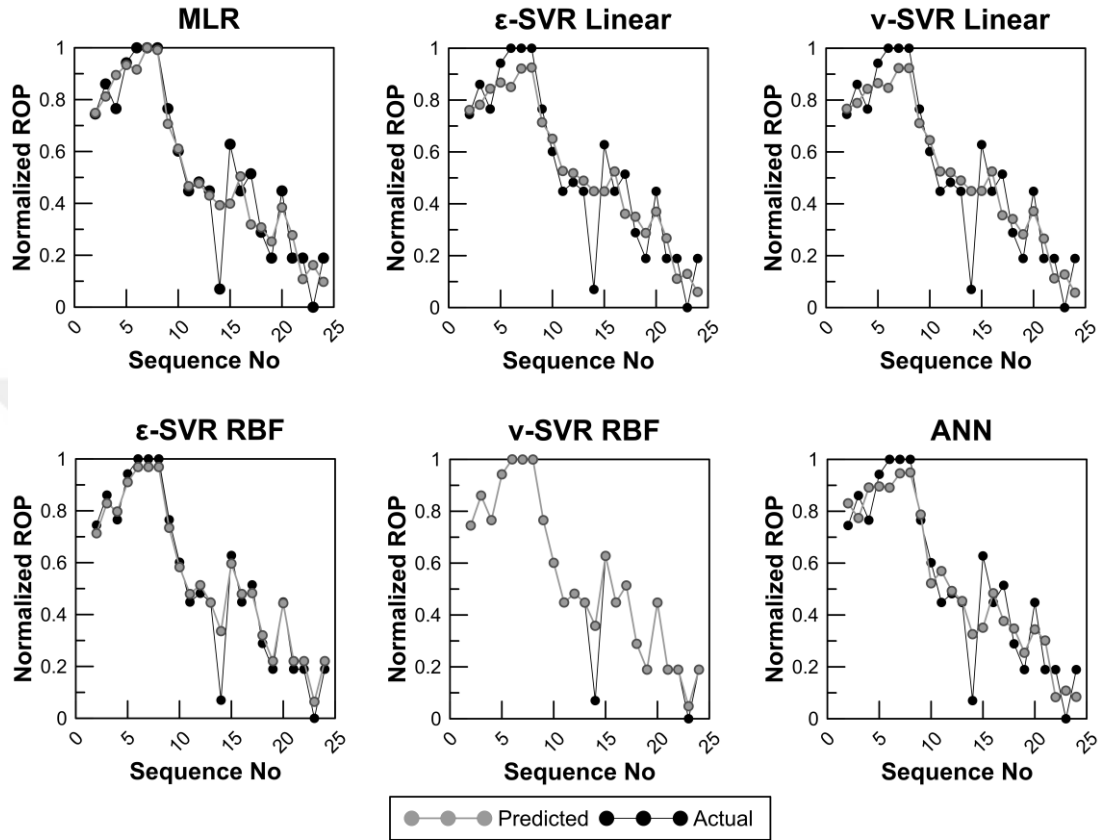


Figure D.108 : Model training for Case 3.2.2.

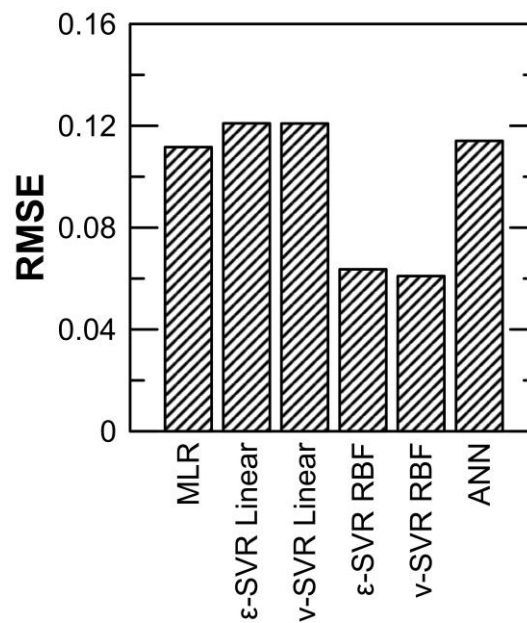


Figure D.109 : Training errors for Case 3.2.2.

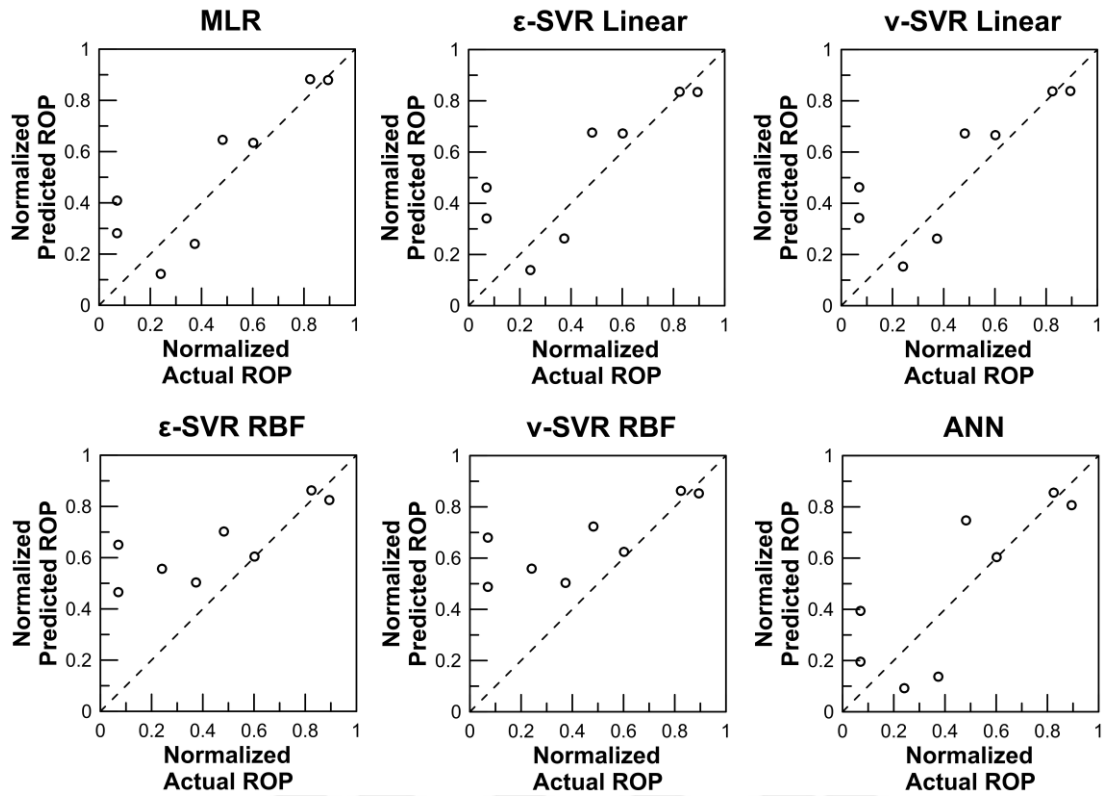


Figure D.110 : ROP predictions with testing subset for Case 3.2.2.

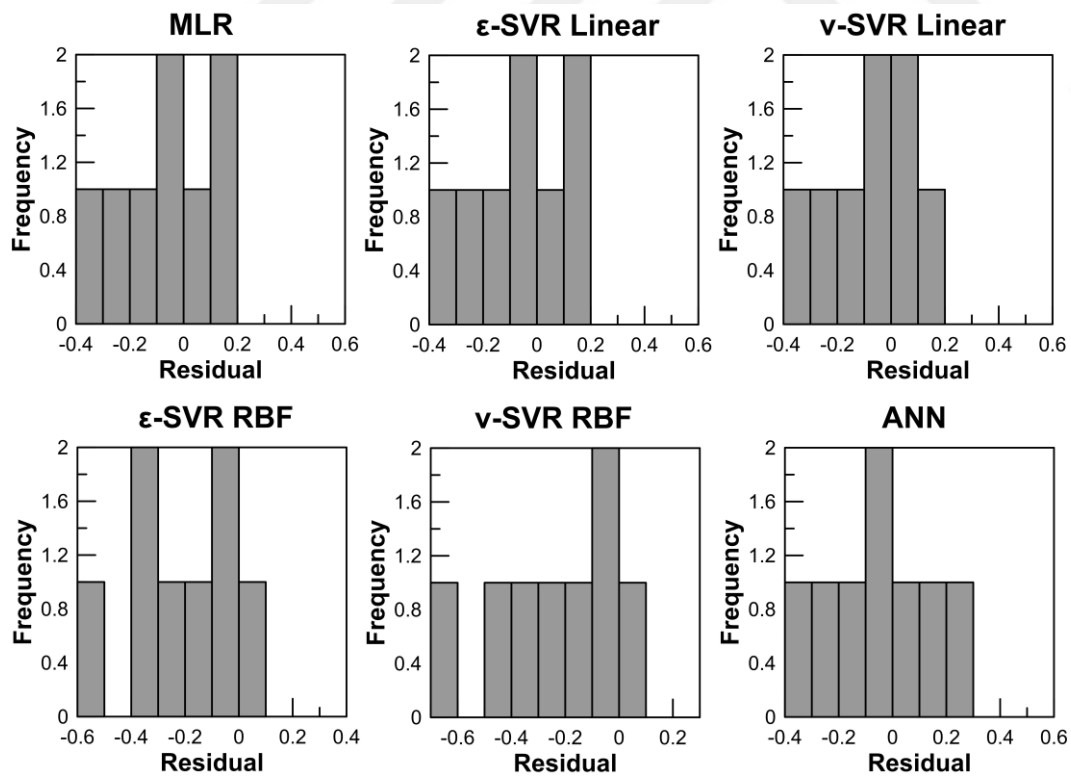


Figure D.111 : Error histograms for Case 3.2.2.

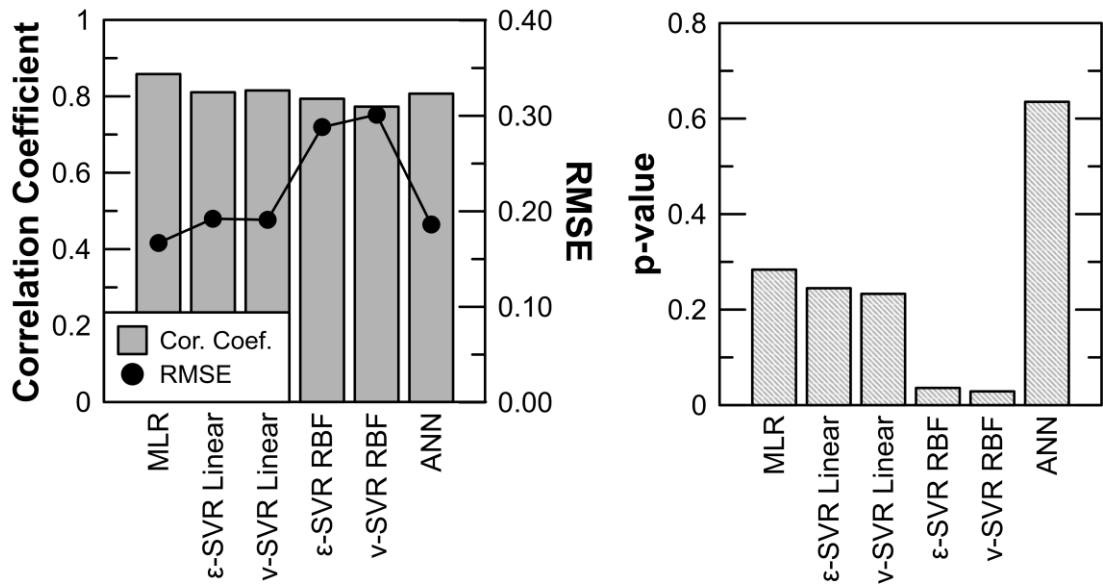


Figure D.112 : Statistical comparison for Case 3.2.2.

D.3.3. Case 3.2.3: BYM+UCS

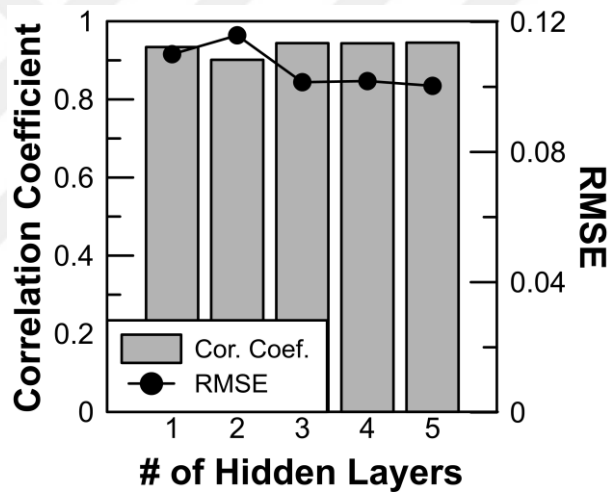


Figure D.113 : ANN hidden layer sensitivity for Case 3.2.3.

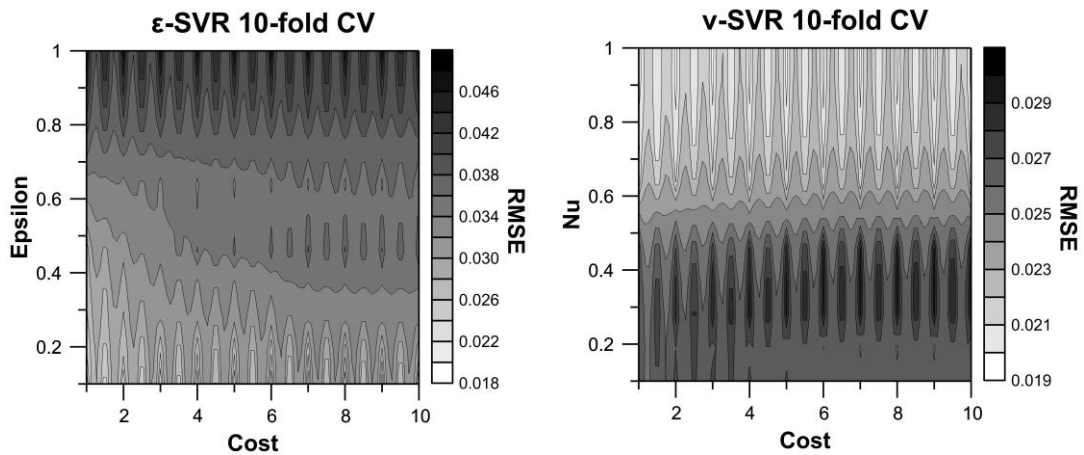


Figure D.114 : 10-fold CV grid search plots for Case 3.2.3.

Table D.17 : SVR hyper-parameters for Case 3.2.3.

ε -SVR					ν -SVR				
Linear		RBF			Linear		RBF		
C	ε	C	γ	ε	C	ν	C	γ	ν
1.0	0.1	3.5	1.0	0.1	2.0	0.7	8.0	1.0	0.5

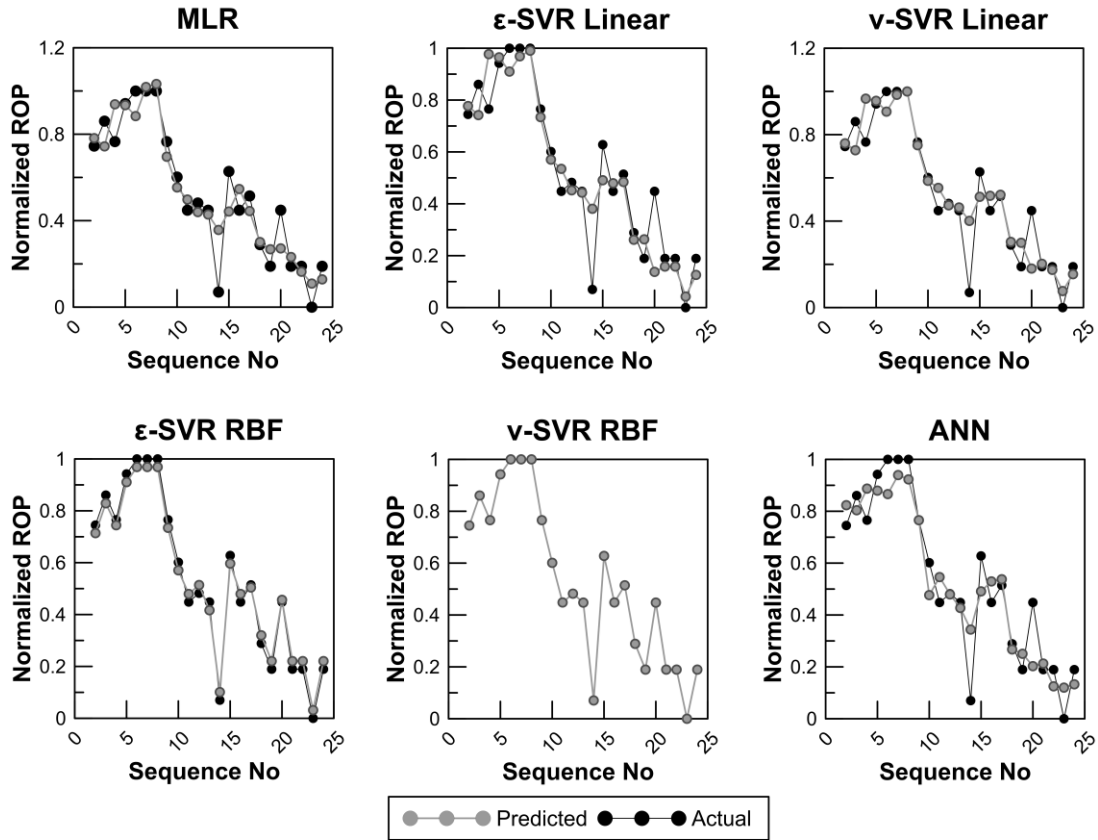


Figure D.115 : Model training for Case 3.2.3.

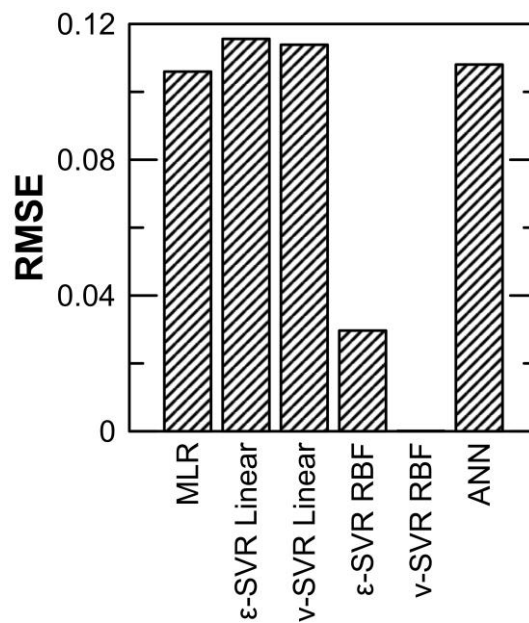


Figure D.116 : Training errors for Case 3.2.3.

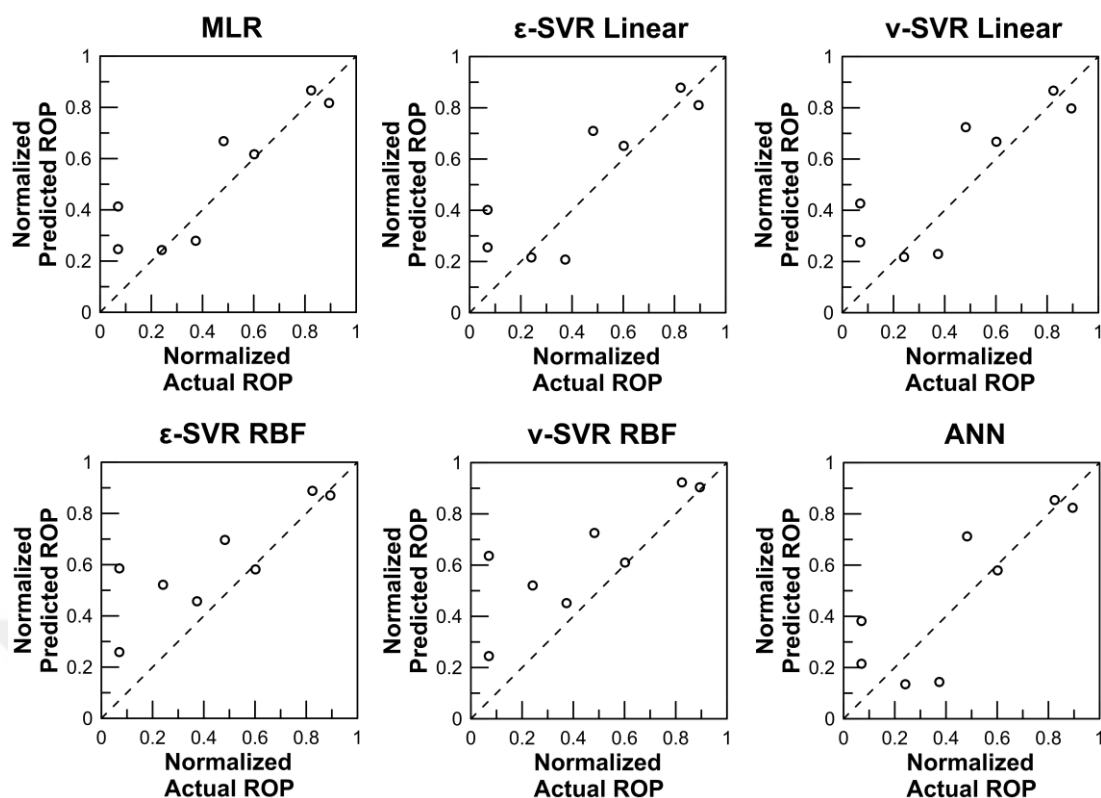


Figure D.117 : ROP predictions with testing subset for Case 3.2.3.

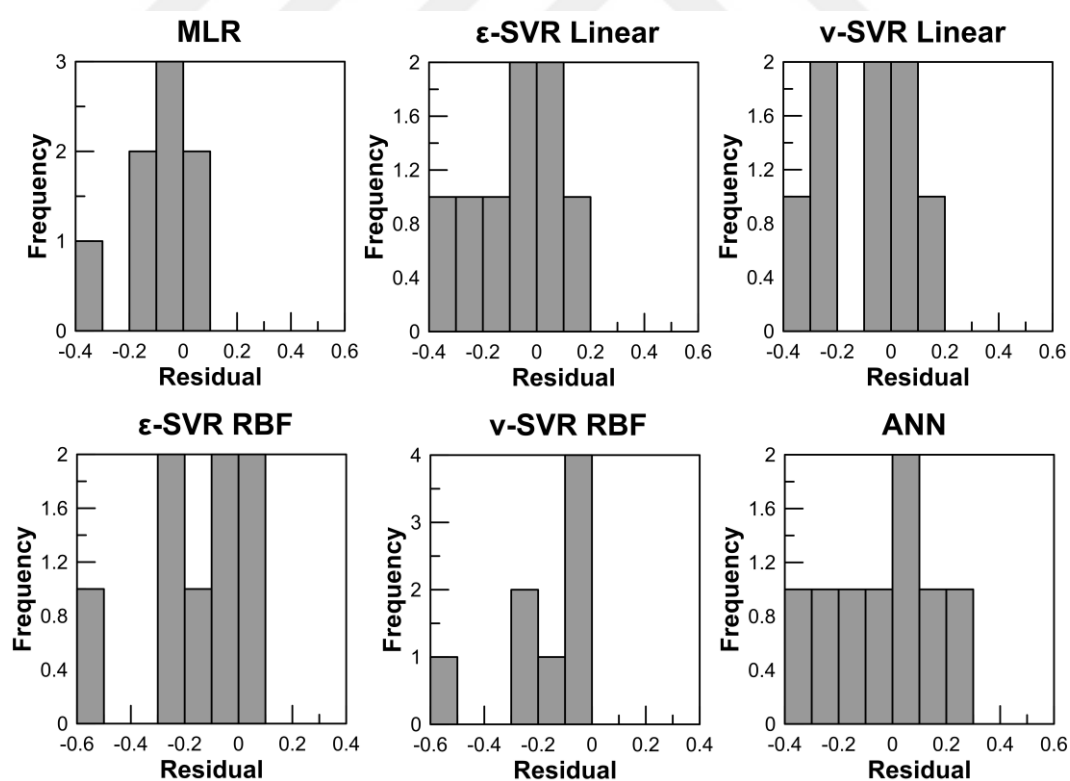


Figure D.118 : Error histograms for Case 3.2.3.

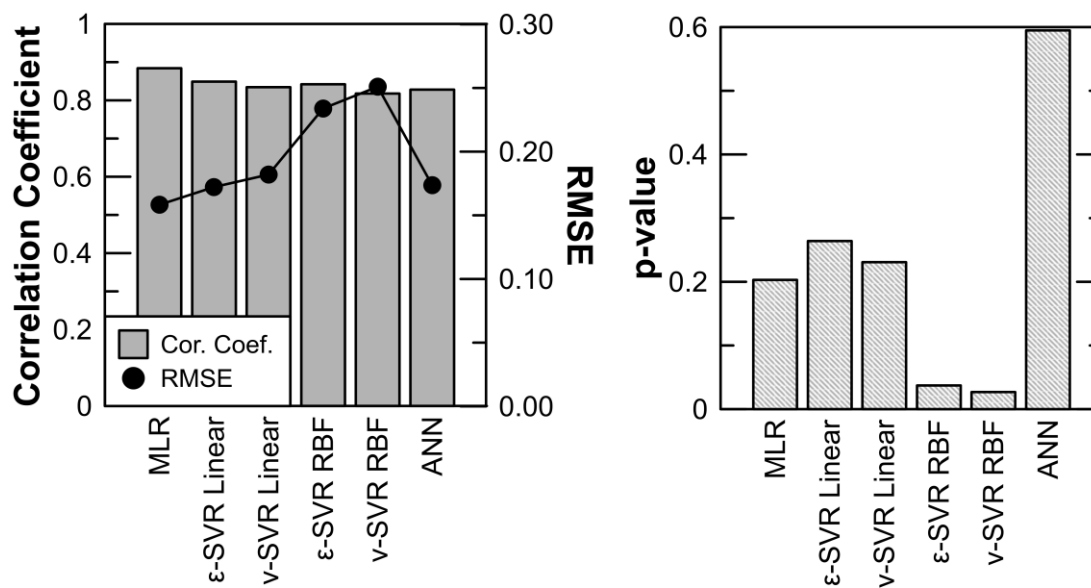


Figure D.119 : Statistical comparison for Case 3.2.3.

D.3.4. Case 3.2.4: BYM+SFG

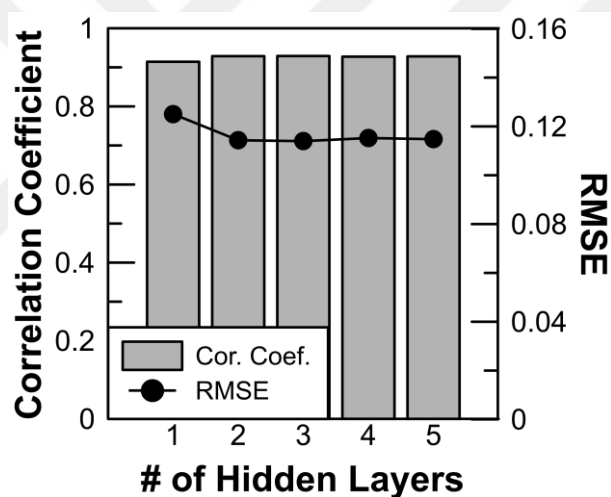


Figure D.120 : ANN hidden layer sensitivity for Case 3.2.4.

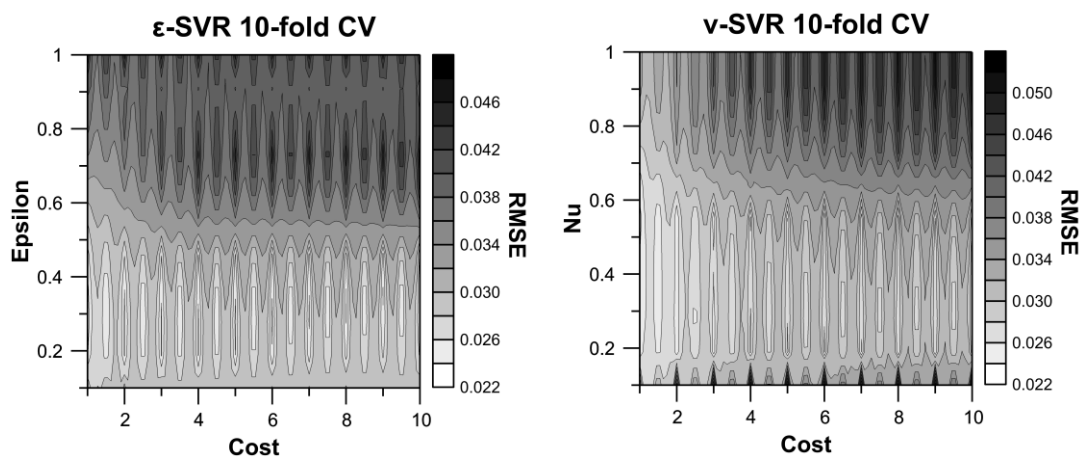


Figure D.121 : 10-fold CV grid search plots for Case 3.2.4.

Table D.18 : SVR hyper-parameters for Case 3.2.4.

ε -SVR					ν -SVR				
Linear		RBF			Linear		RBF		
C	ε	C	γ	ε	C	ν	C	γ	ν
1.0	0.25	2.0	1.0	0.1	1.0	0.5	2.0	1.0	1.0

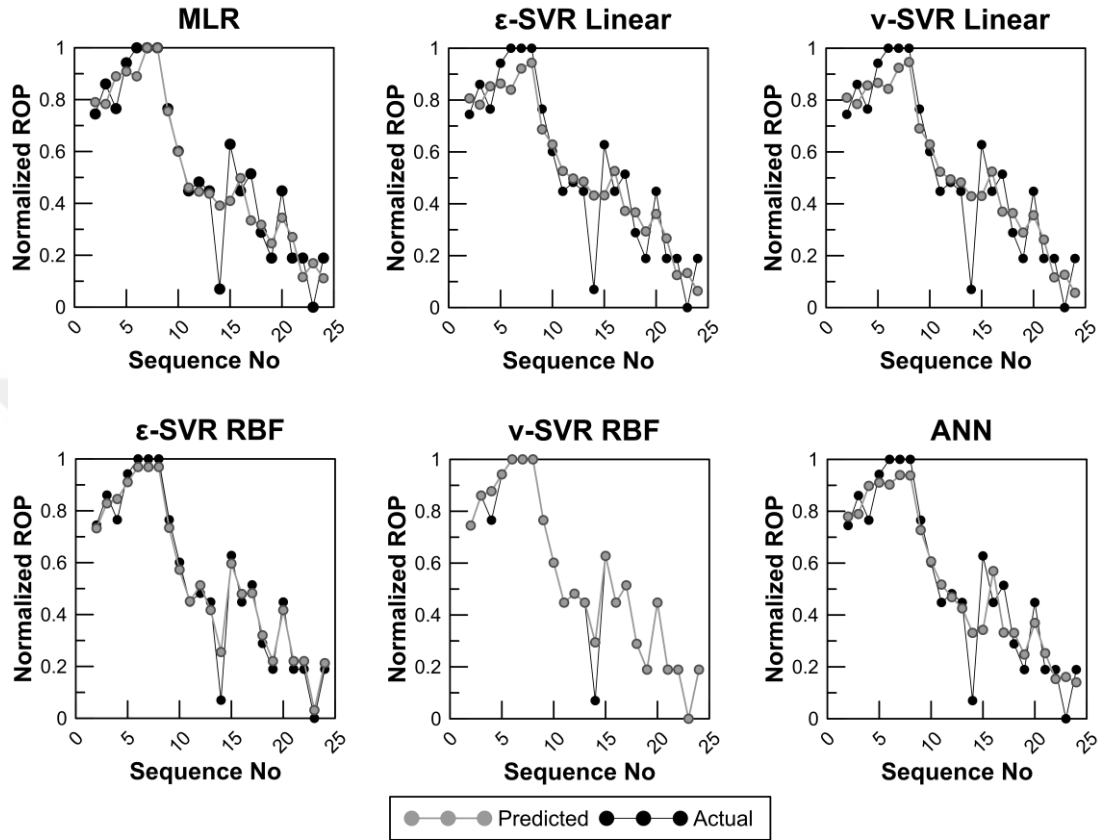


Figure D.122 : Model training for Case 3.2.4.

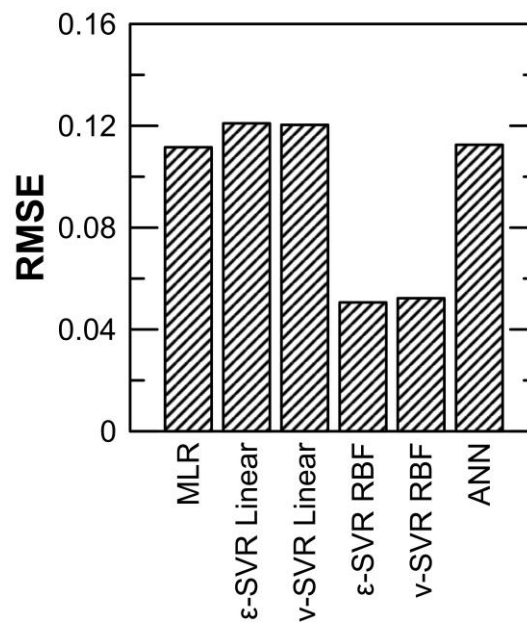


Figure D.123 : Training errors for Case 3.2.4.

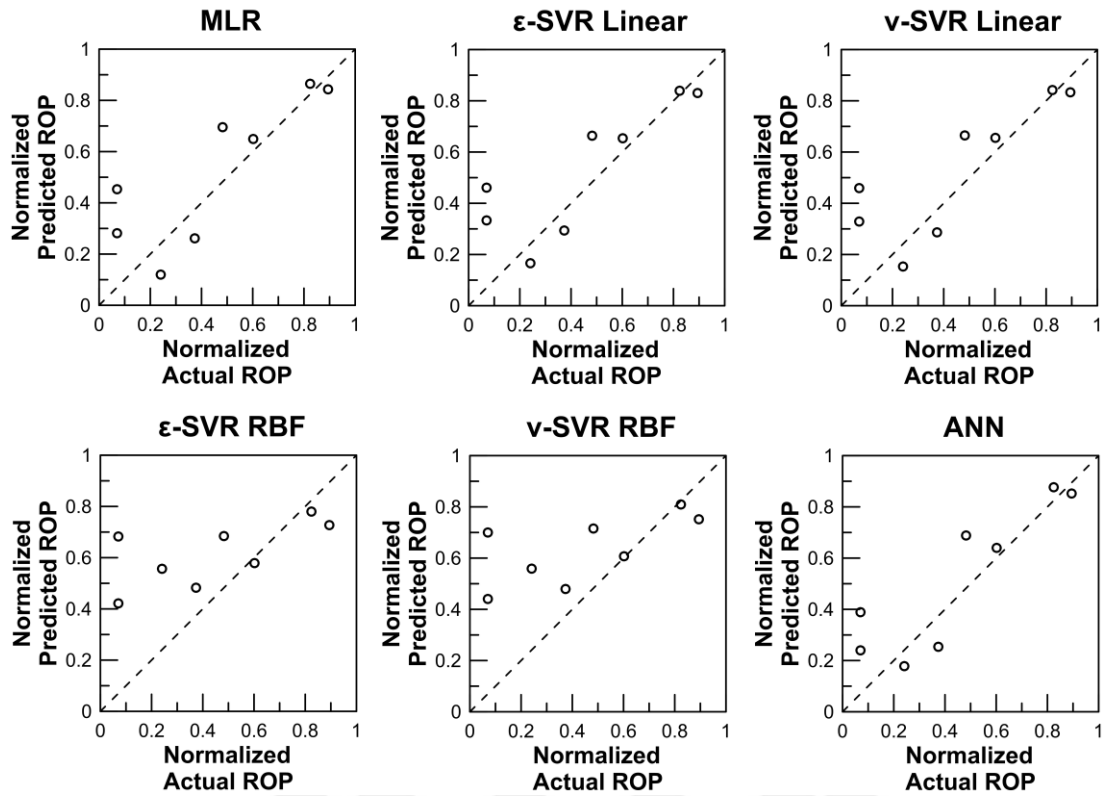


Figure D.124 : ROP predictions with testing subset for Case 3.2.4.

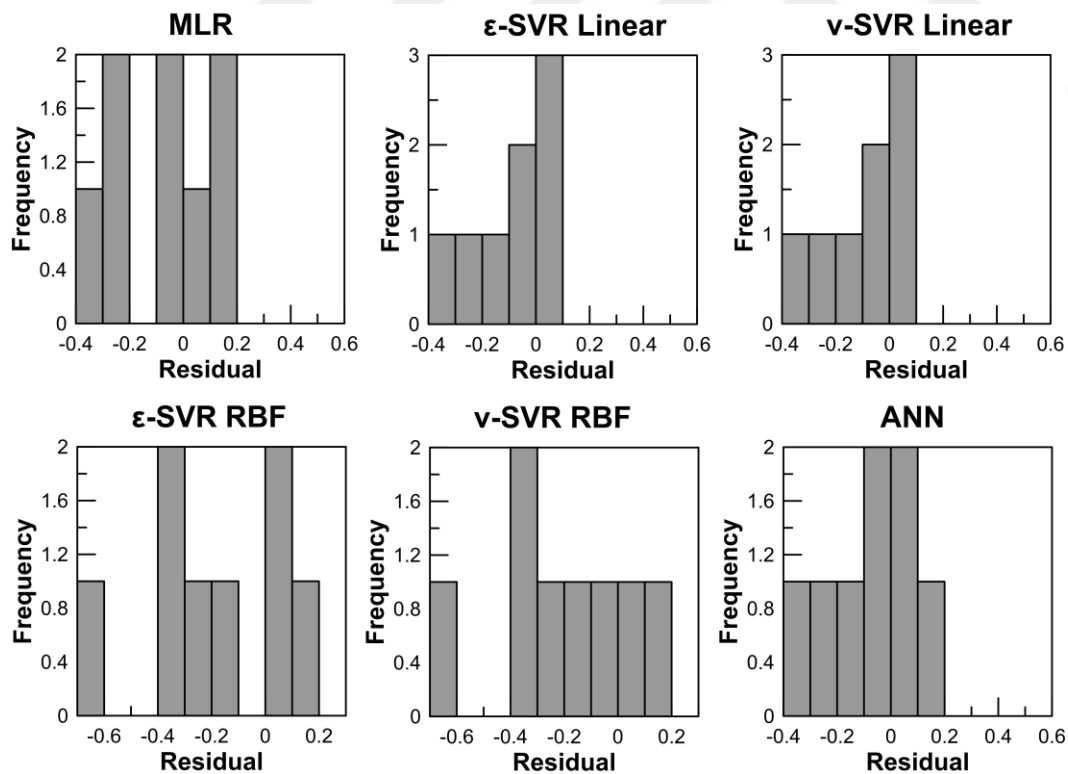


Figure D.125 : Error histograms for Case 3.2.4.

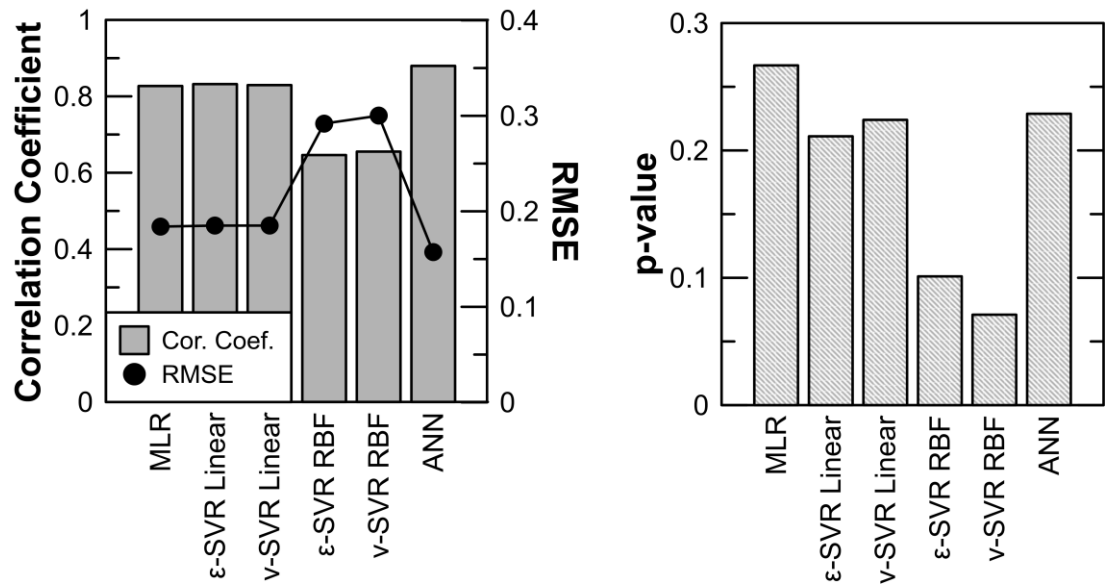


Figure D.126 : Statistical comparison for Case 3.2.4.

D.3.5. Case 3.2.5: BYM+TRQ

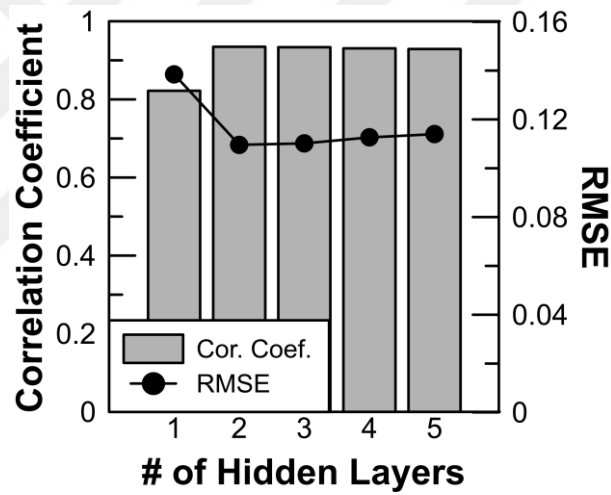


Figure D.127 : ANN hidden layer sensitivity for Case 3.2.5.

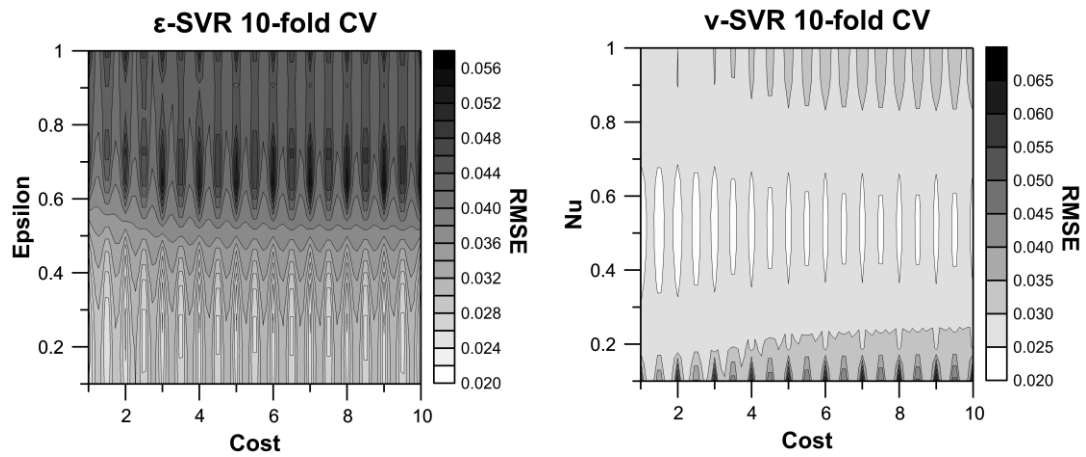


Figure D.128 : 10-fold CV grid search plots for Case 3.2.5.

Table D.19 : SVR hyper-parameters for Case 3.2.5.

ε -SVR					ν -SVR				
Linear		RBF			Linear		RBF		
C	ε	C	γ	ε	C	ν	C	γ	ν
1.0	0.25	2.0	1.0	0.1	3.0	0.55	10.0	1.0	1.0

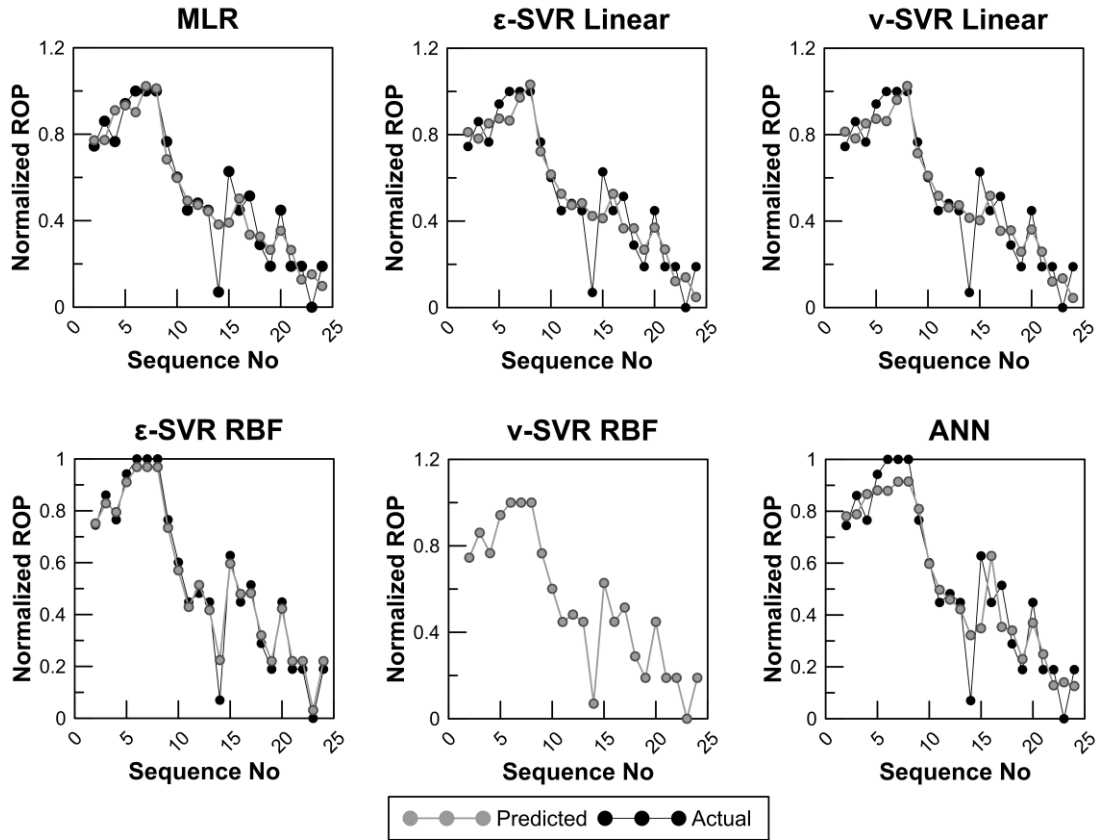


Figure D.129 : Model training for Case 3.2.5.

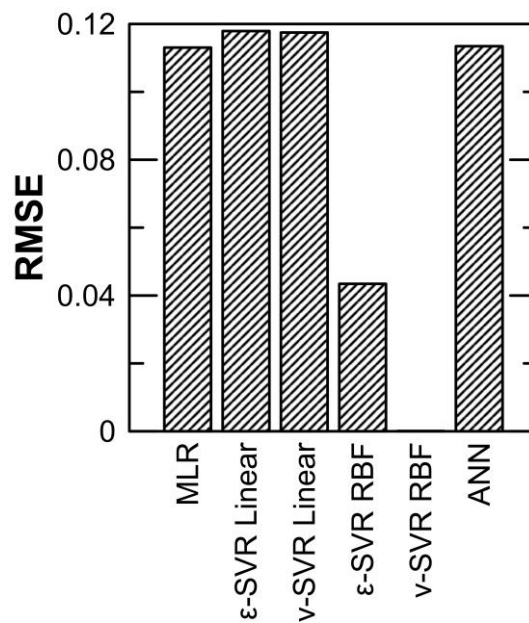


Figure D.130 : Training errors for Case 3.2.5.

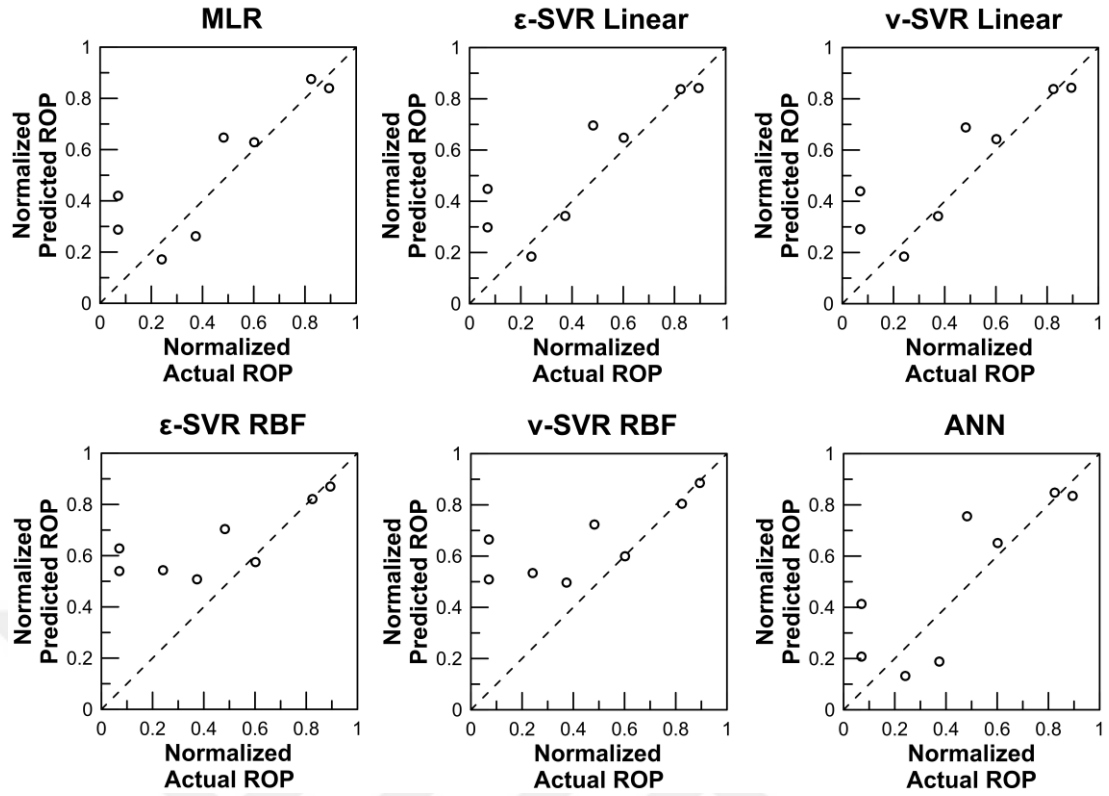


Figure D.131 : ROP predictions with testing subset for Case 3.2.5.

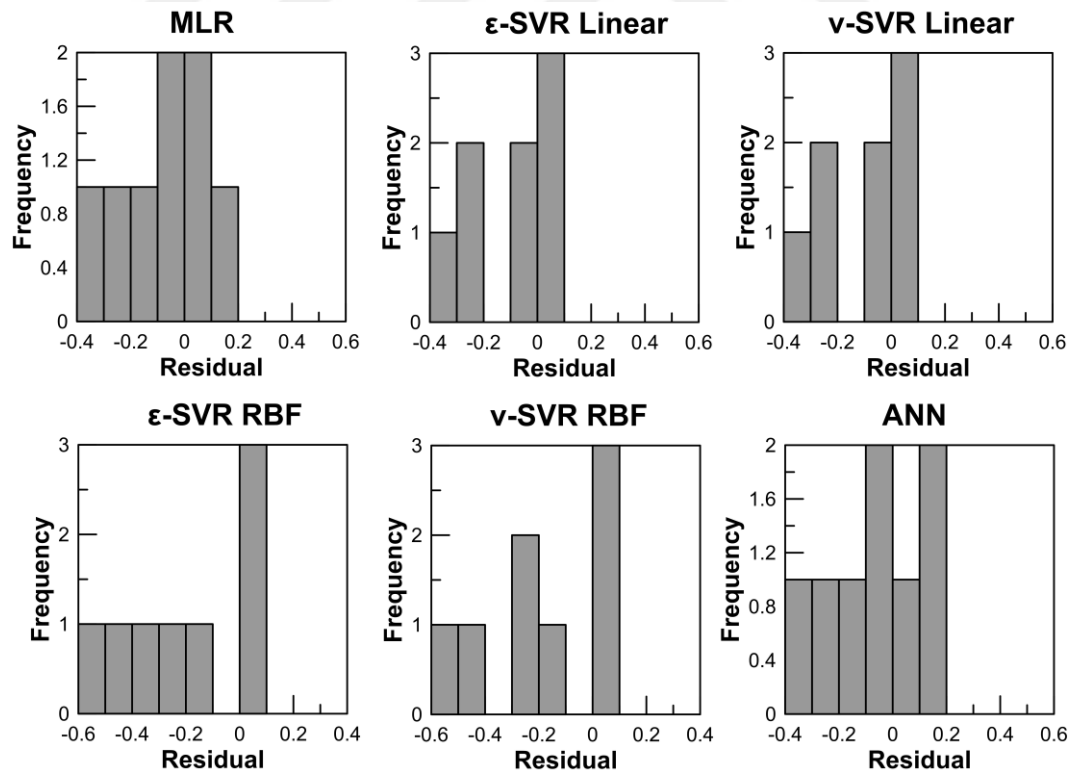


Figure D.132 : Error histograms for Case 3.2.5.

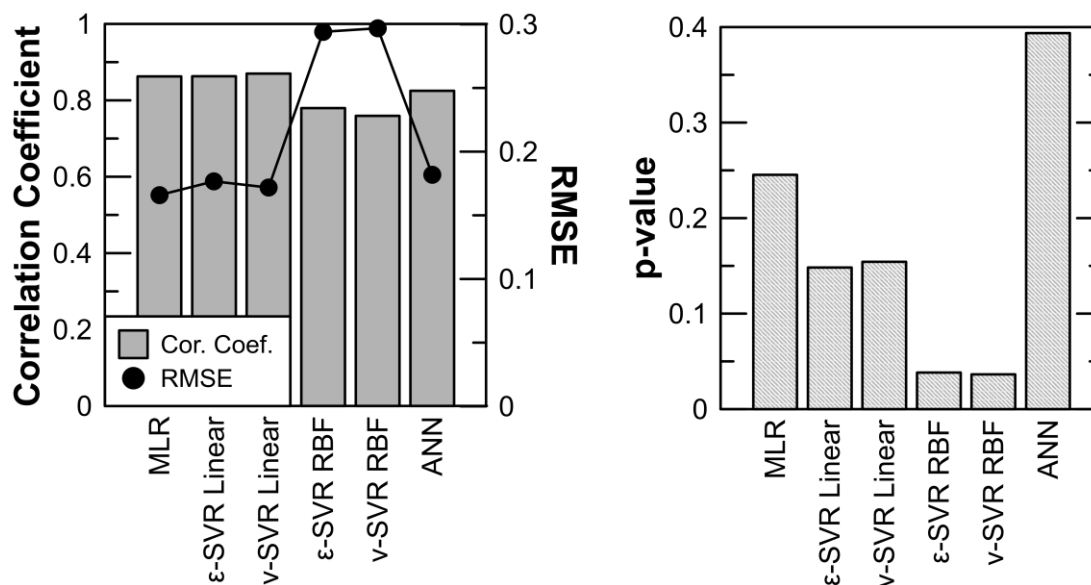


Figure D.133 : Statistical comparison for Case 3.2.5.

D.3.6. Case 3.2.6: BYM+All

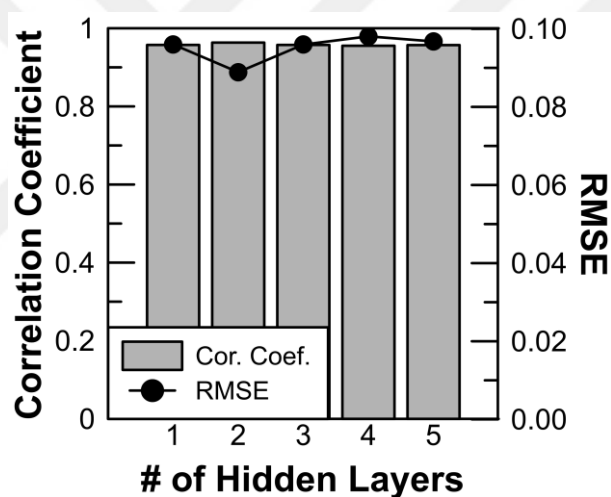


Figure D.134 : ANN hidden layer sensitivity for Case 3.2.6.

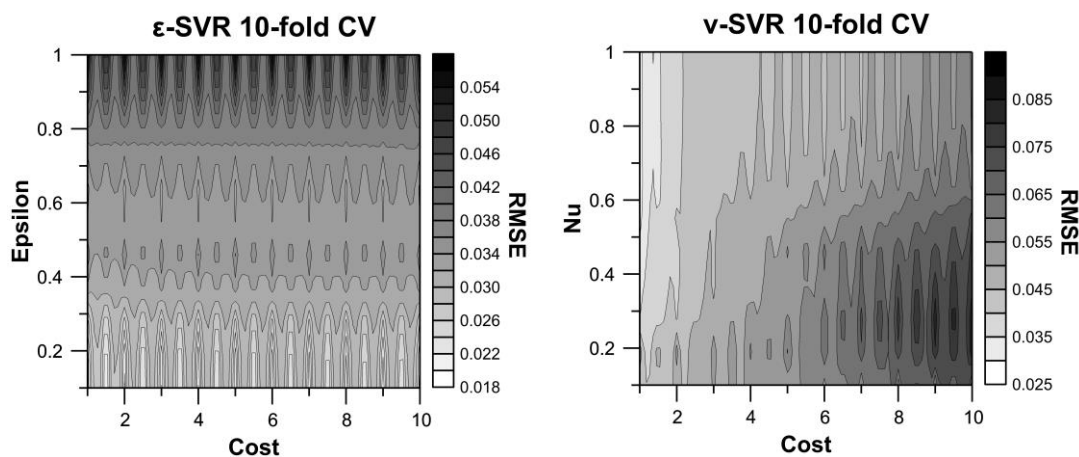


Figure D.135 : 10-fold CV grid search plots for Case 3.2.6.

Table D.20 : SVR hyper-parameters for Case 3.2.6.

ε -SVR					ν -SVR				
Linear			RBF		Linear			RBF	
C	ε	C	γ	ε	C	ν	C	γ	ν
1.0	0.25	2.0	1.0	0.1	1.0	0.65	5.5	1.0	0.9

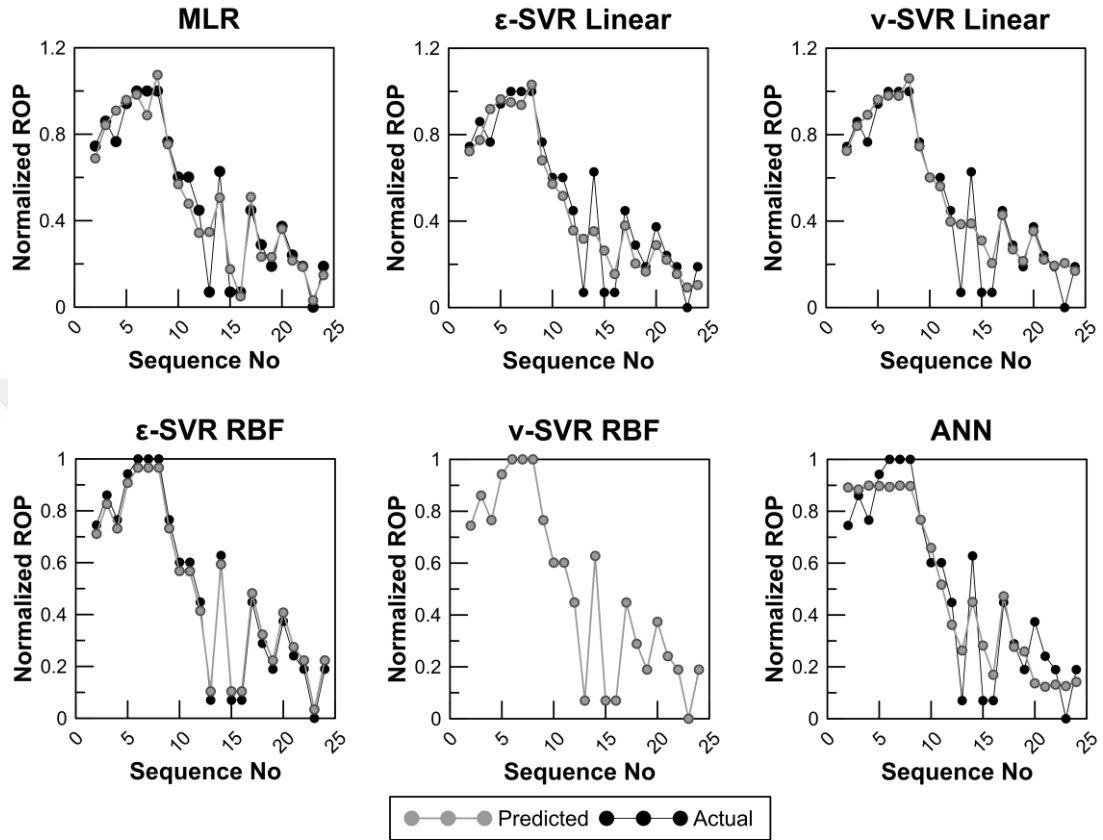


Figure D.136 : Model training for Case 3.2.6.

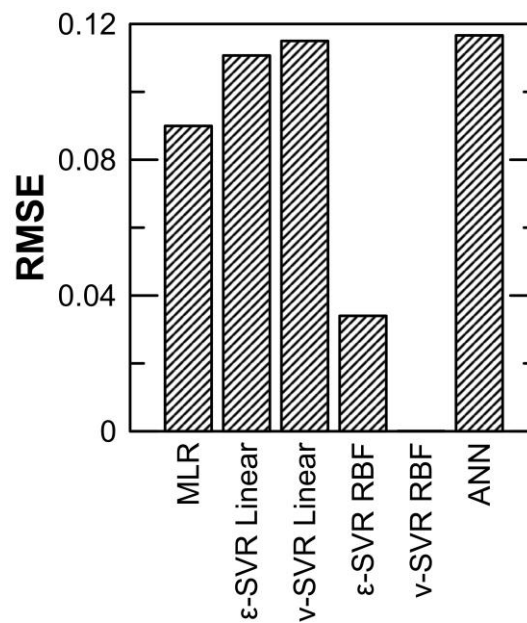


Figure D.137 : Training errors for Case 3.2.6.

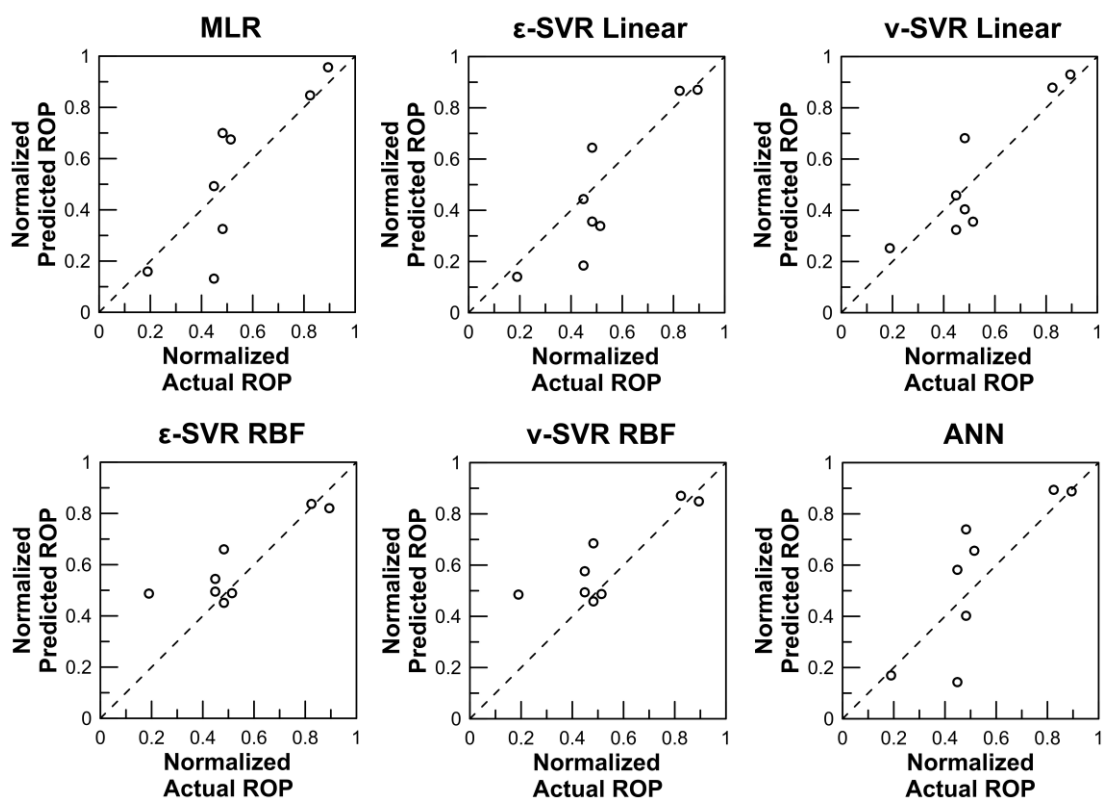


Figure D.138 : ROP predictions with testing subset for Case 3.2.6.

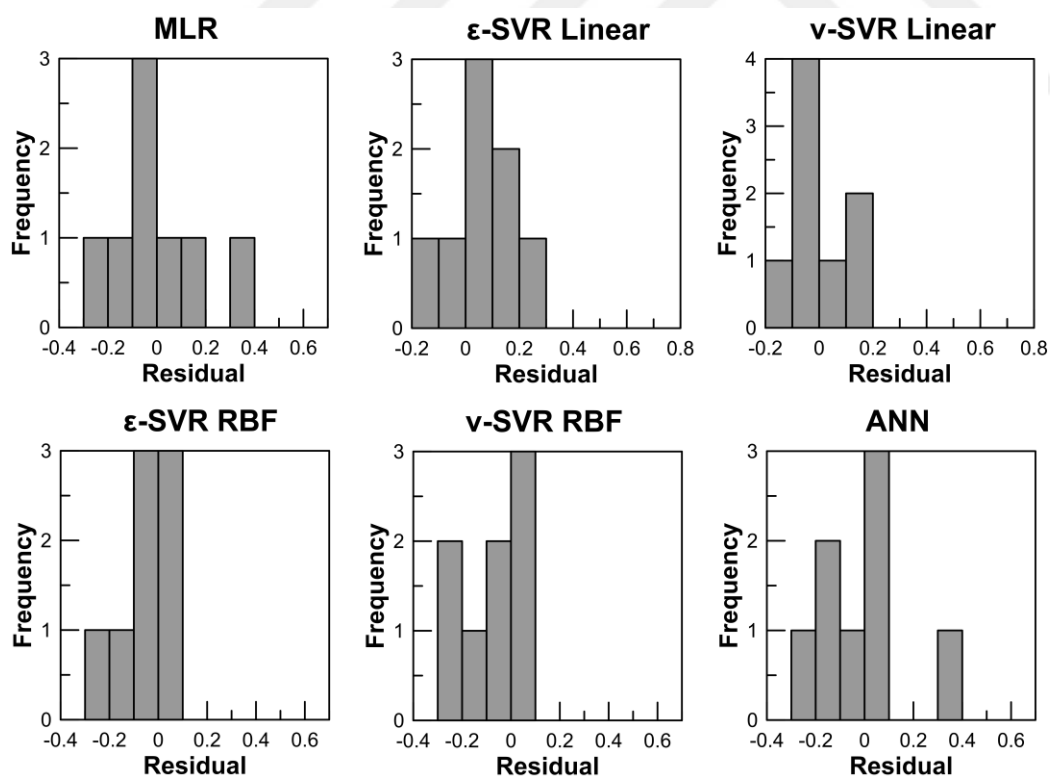


Figure D.139 : Error histograms for Case 3.2.6.

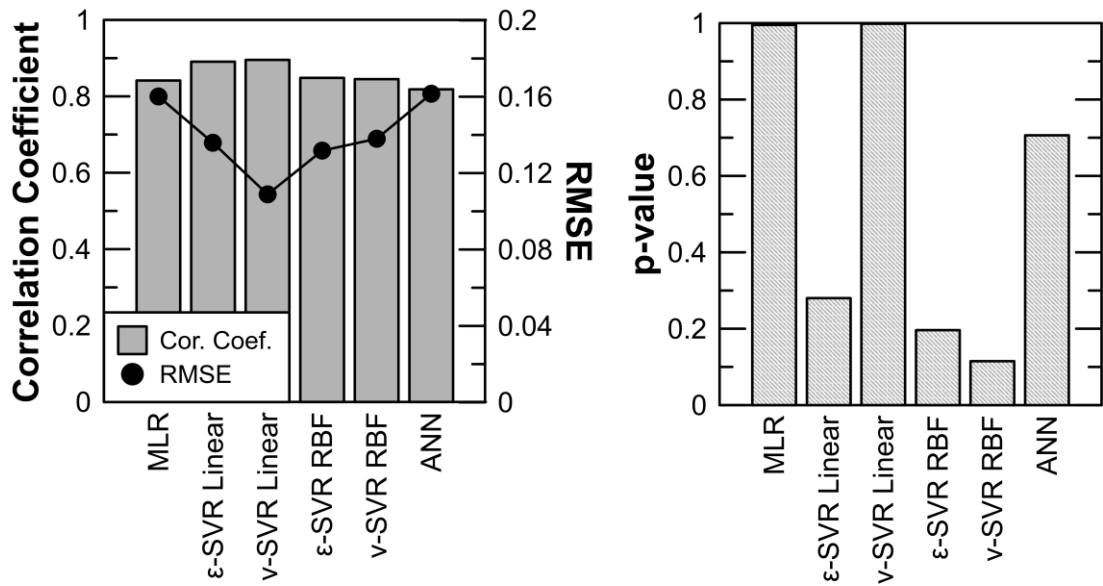


Figure D.140 : Statistical comparison for Case 3.2.6.

D.3.7. Case 3.2.7: Feature Selection (BYM+All excluding x_6)

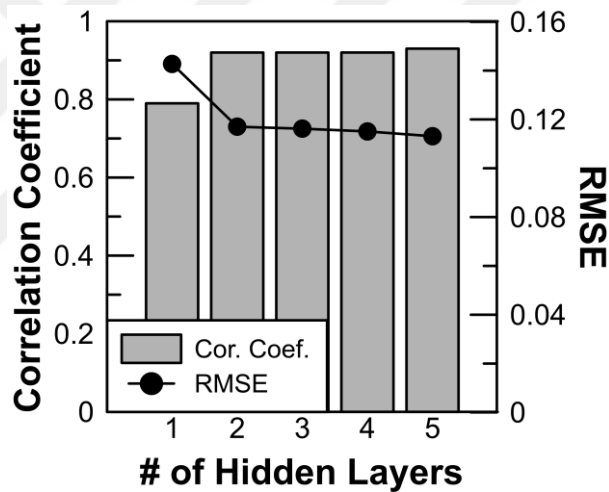


Figure D.141 : ANN hidden layer sensitivity for Case 3.2.7.

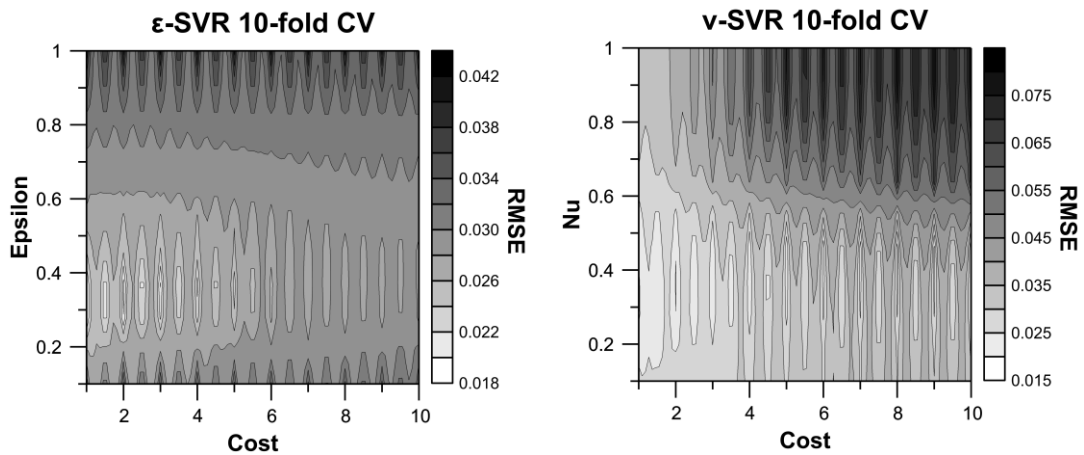


Figure D.142 : 10-fold CV grid search plots for Case 3.2.7.

Table D.21 : SVR hyper-parameters for Case 3.2.7.

ε -SVR					ν -SVR				
Linear		RBF			Linear		RBF		
C	ε	C	γ	ε	C	ν	C	γ	ν
2.0	0.35	2.0	1.0	0.1	1.0	0.25	9.5	1.0	0.6

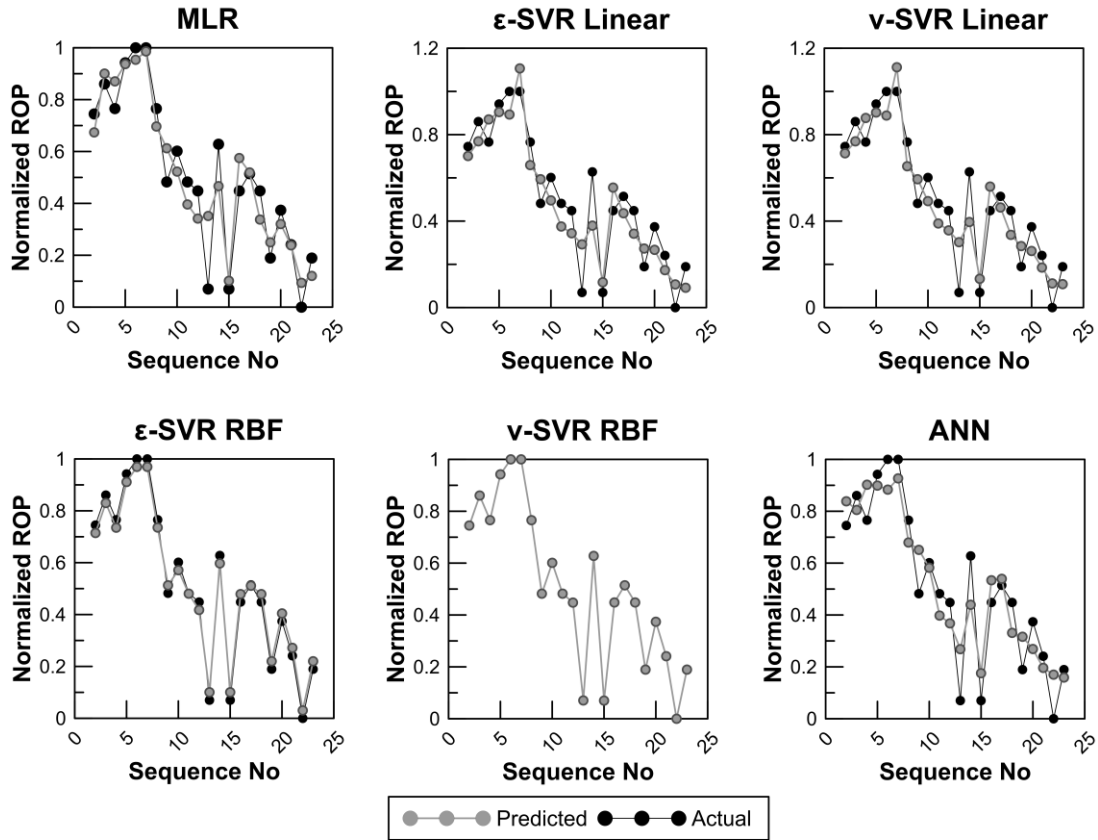


Figure D.143 : Model training for Case 3.2.7.

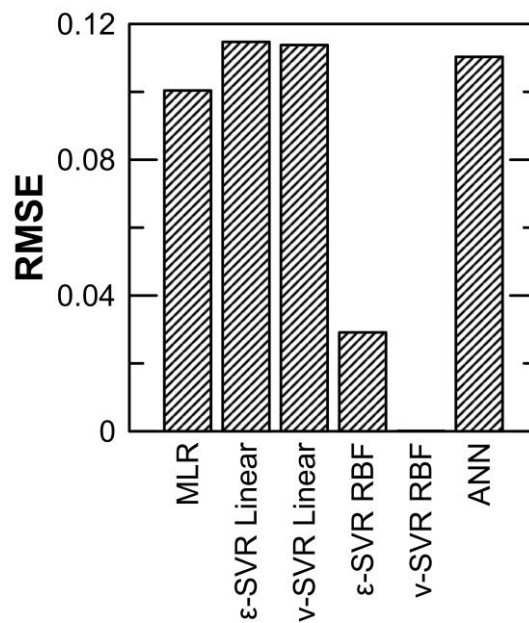


Figure D.144 : Training errors for Case 3.2.7.

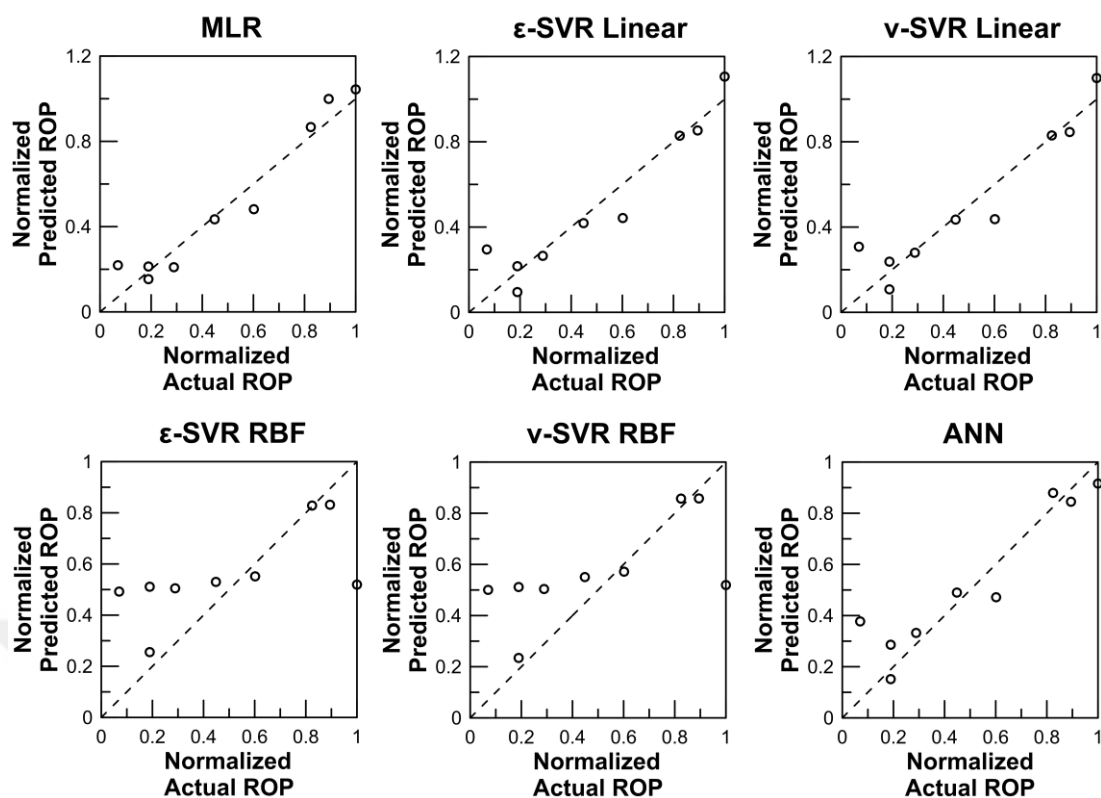


Figure D.145 : ROP predictions with testing subset for Case 3.2.7.

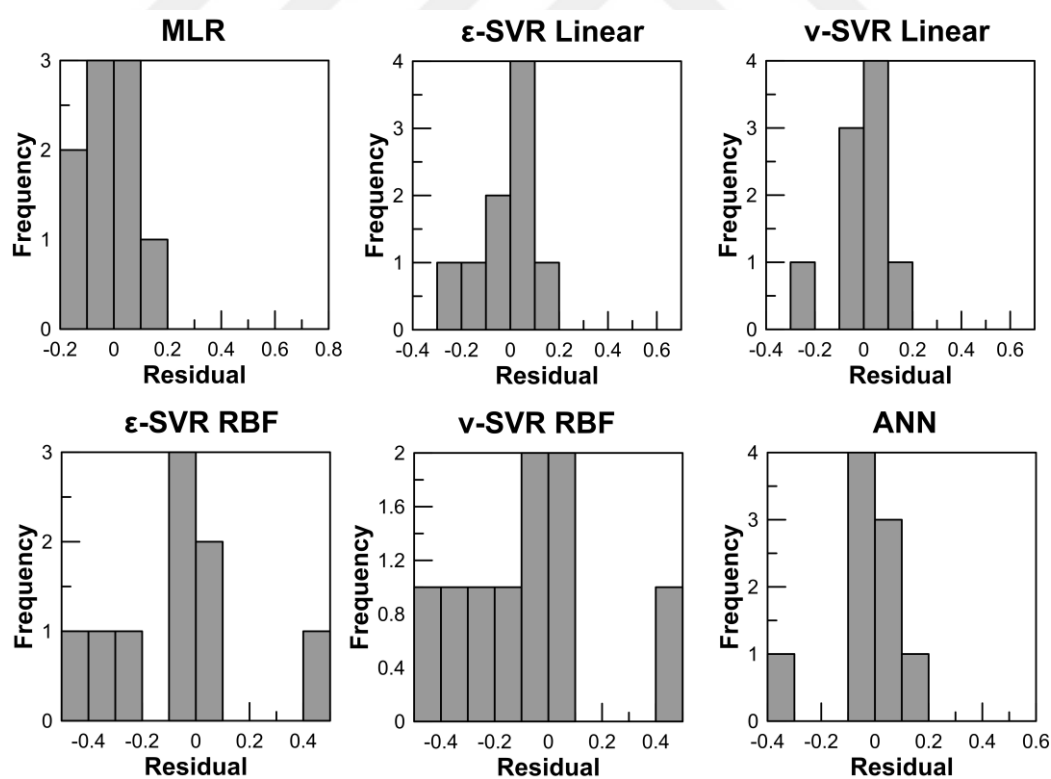


Figure D.146 : Error histograms for Case 3.2.7.

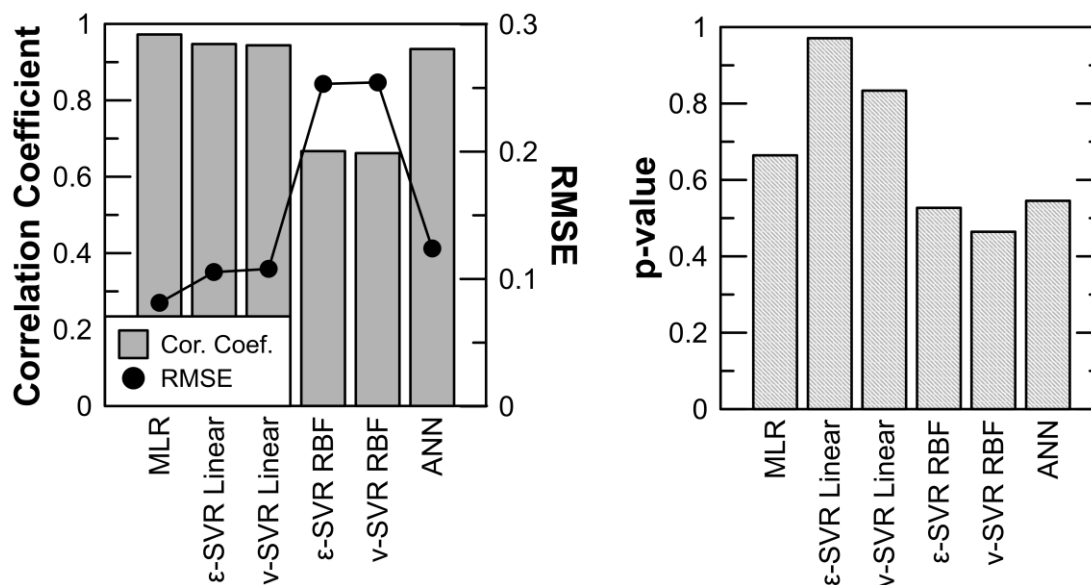


Figure D.147 : Statistical comparison for Case 3.2.7.

D.4. Intermediate Steps for Case 3.3

D.4.1. Case 3.3.1: Default BYM

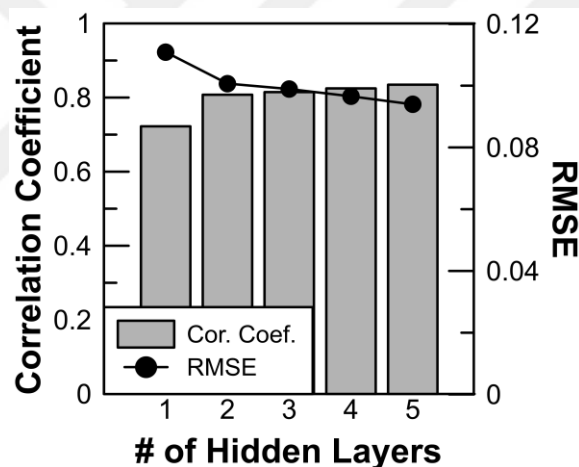


Figure D.148 : ANN hidden layer sensitivity for Case 3.3.1.

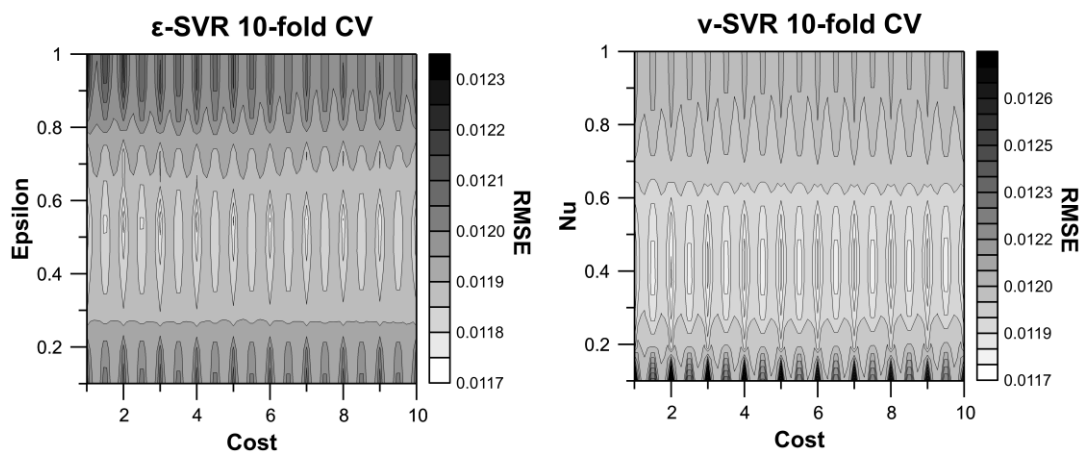


Figure D.149 : 10-fold CV grid search plots for Case 3.3.1.

Table D.22 : SVR hyper-parameters for Case 3.3.1.

ε -SVR					ν -SVR				
Linear			RBF		Linear		RBF		
C	ε	C	γ	ε	C	ν	C	γ	ν
3.0	0.5	2.0	1.0	0.1	7.0	0.35	2.5	1.0	0.6

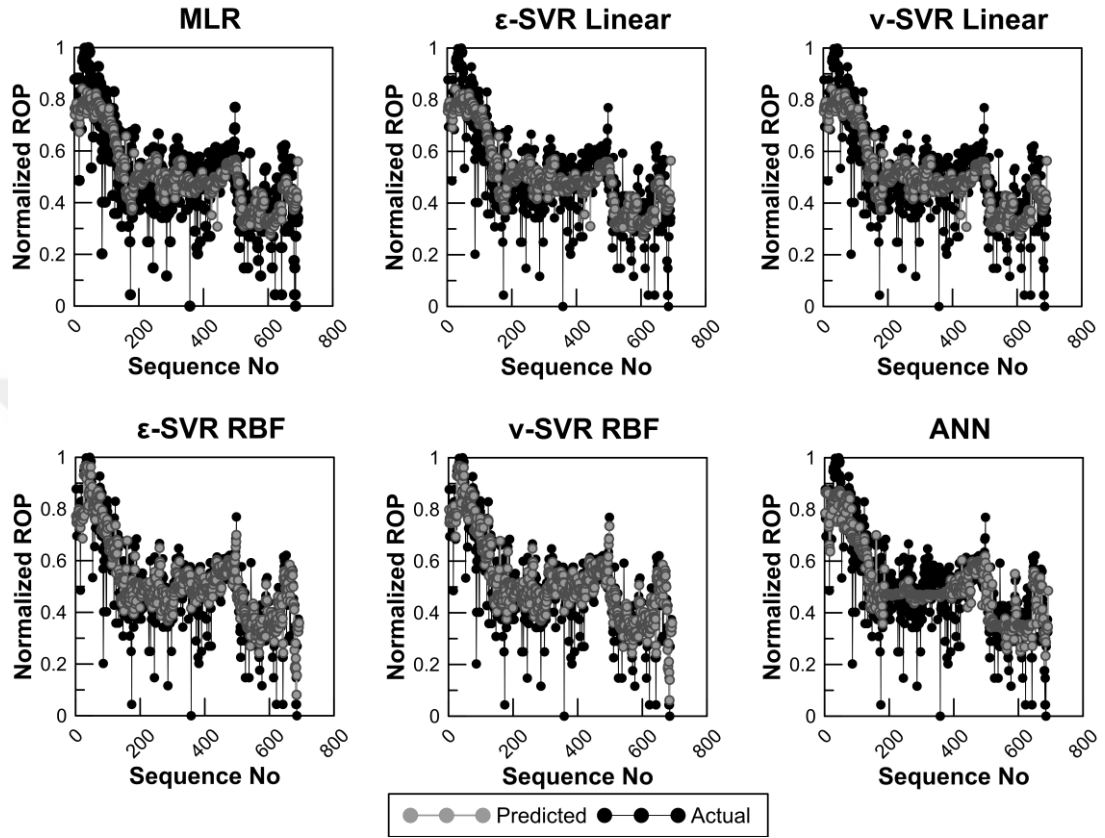


Figure D.150 : Model training for Case 3.3.1.

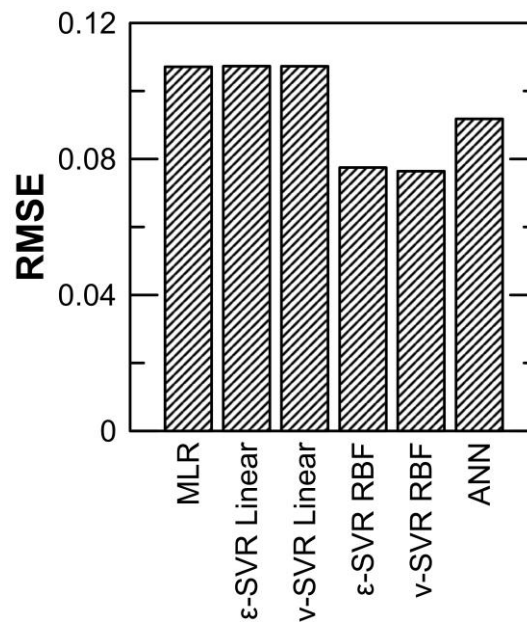


Figure D.151 : Training errors for Case 3.3.1.

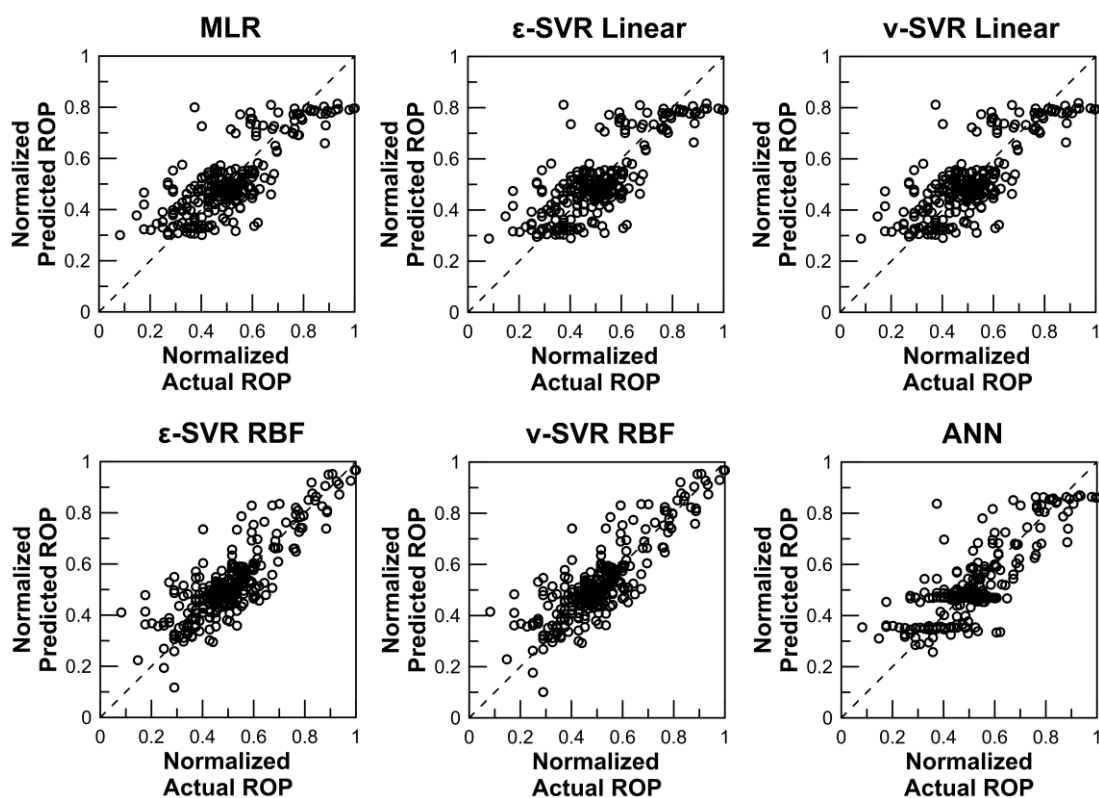


Figure D.152 : ROP predictions with testing subset for Case 3.3.1.

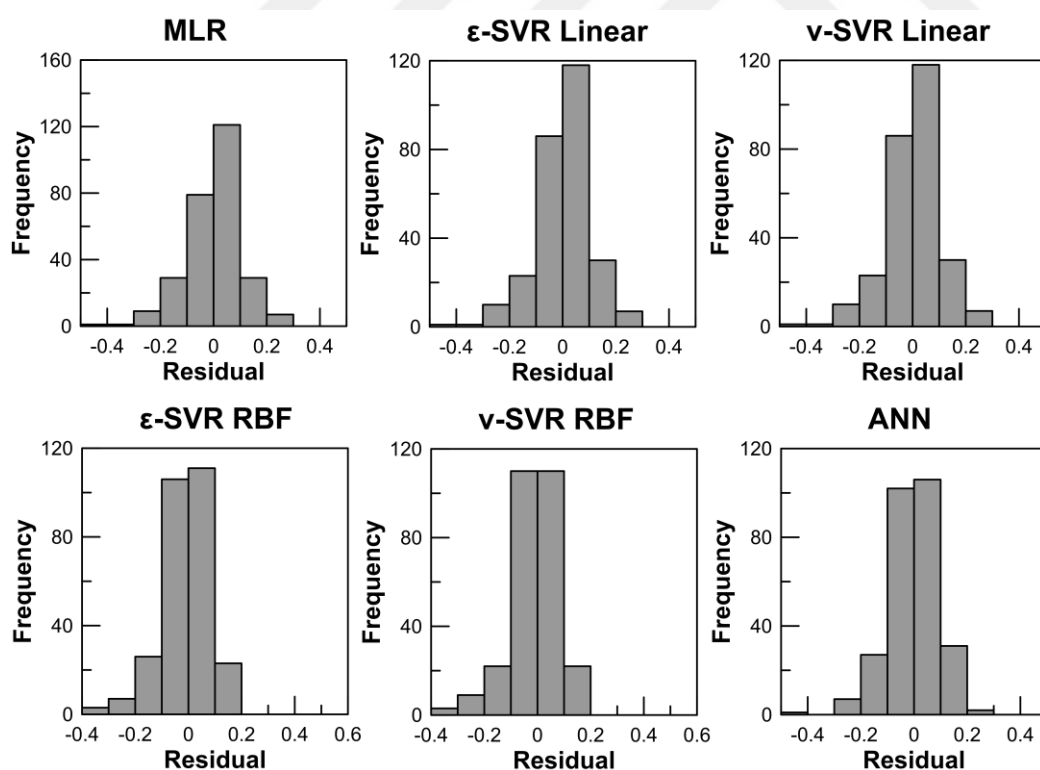


Figure D.153 : Error histograms for Case 3.3.1.

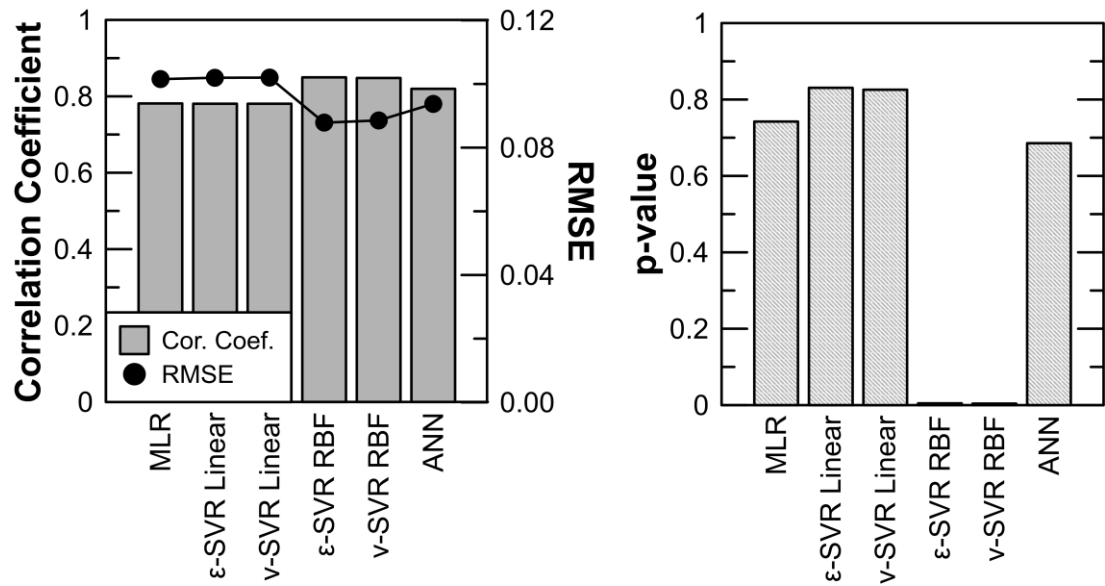


Figure D.154 : Statistical comparison for Case 3.3.1.

D.4.2. Case 3.3.2: BYM+SI

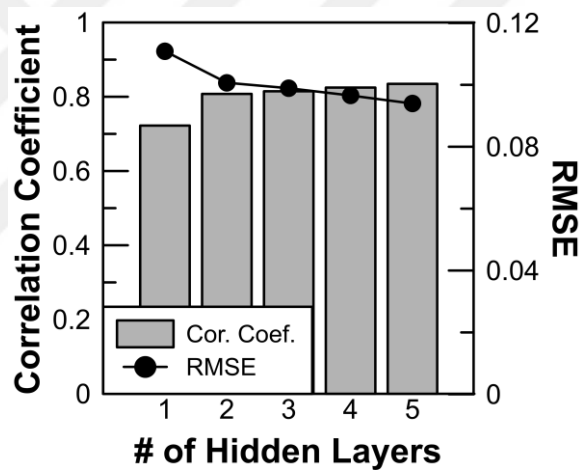


Figure D.155 : ANN hidden layer sensitivity for Case 3.3.2.

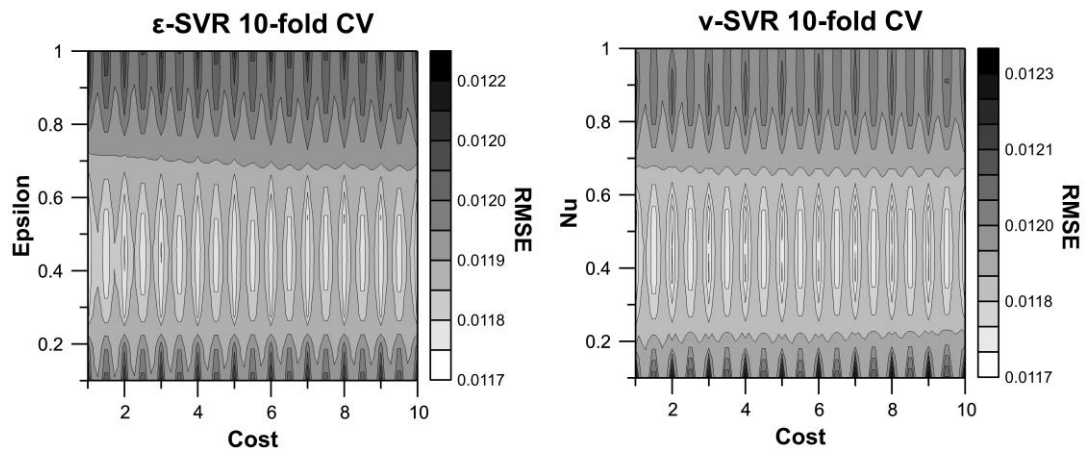


Figure D.156 : 10-fold CV grid search plots for Case 3.3.2.

Table D.23 : SVR hyper-parameters for Case 3.3.2.

ε -SVR					ν -SVR				
Linear			RBF		Linear			RBF	
C	ε	C	γ	ε	C	ν	C	γ	ν
2.5	0.5	1.0	1.0	0.1	1.5	0.35	1.0	1.0	0.4

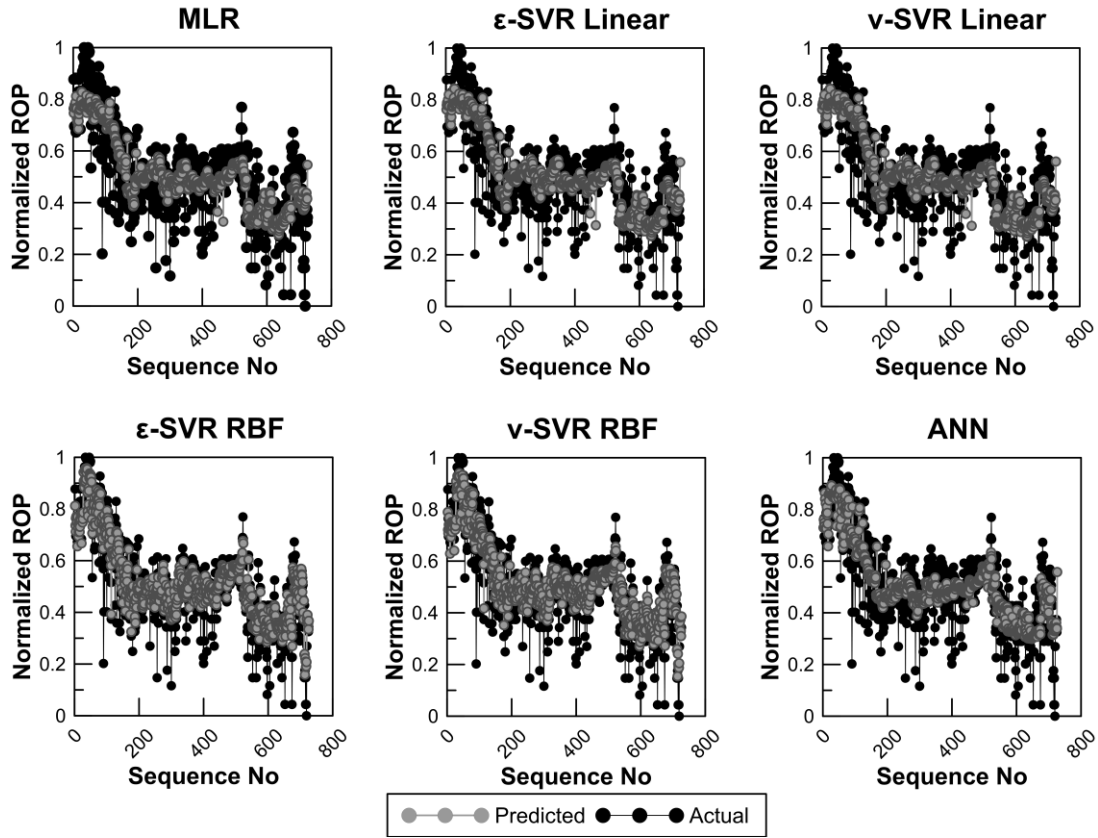


Figure D.157 : Model training for Case 3.3.2.

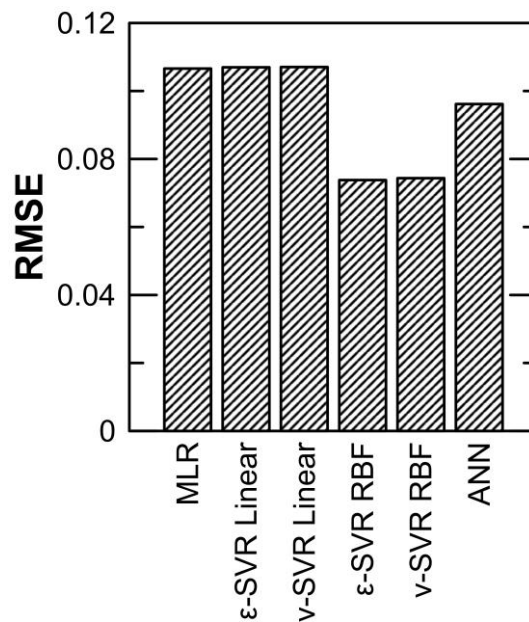


Figure D.158 : Training errors for Case 3.3.2.

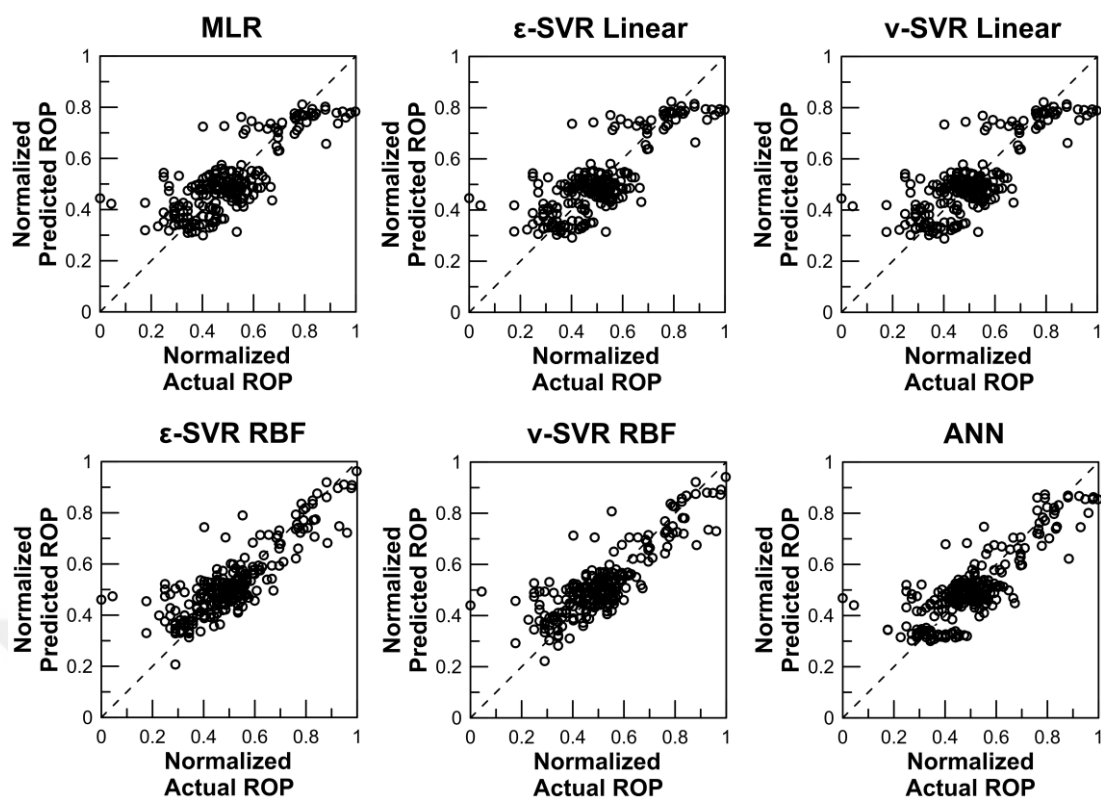


Figure D.159 : ROP predictions with testing subset for Case 3.3.2.

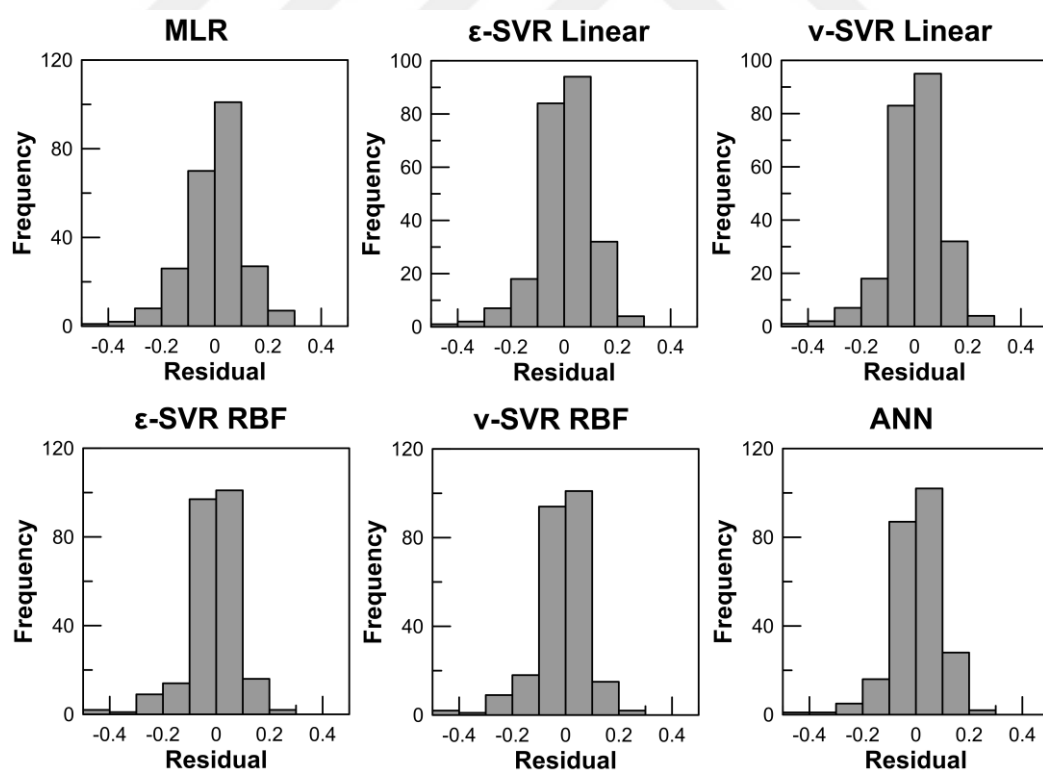


Figure D.160 : Error histograms for Case 3.3.2.

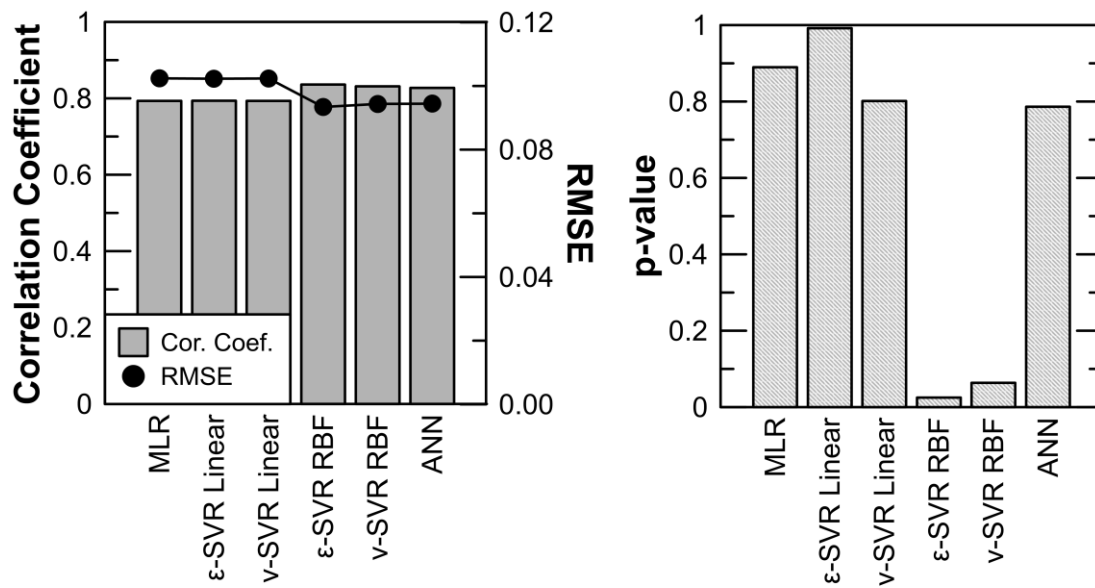


Figure D.161 : Statistical comparison for Case 3.3.2.

D.4.3. Case 3.3.3: BYM+UCS

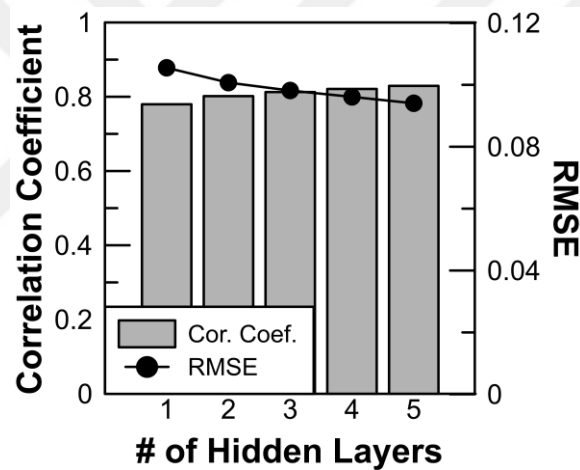


Figure D.162 : ANN hidden layer sensitivity for Case 3.3.3.

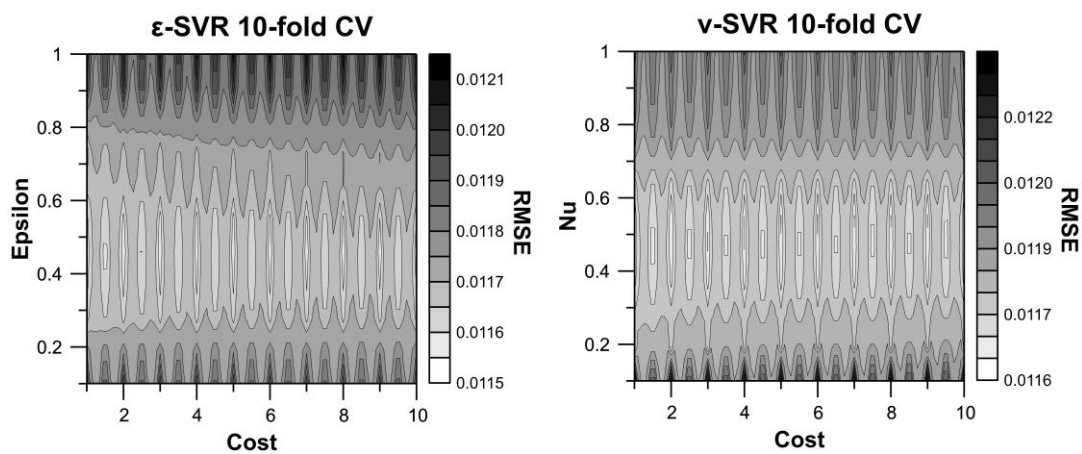


Figure D.163 : 10-fold CV grid search plots for Case 3.3.3.

Table D.24 : SVR hyper-parameters for Case 3.3.3.

ε -SVR					v-SVR				
Linear			RBF		Linear			RBF	
C	ε	C	γ	ε	C	ν	C	γ	ν
9.5	0.45	1.5	1.0	0.3	2.5	0.45	1.5	1.0	0.7

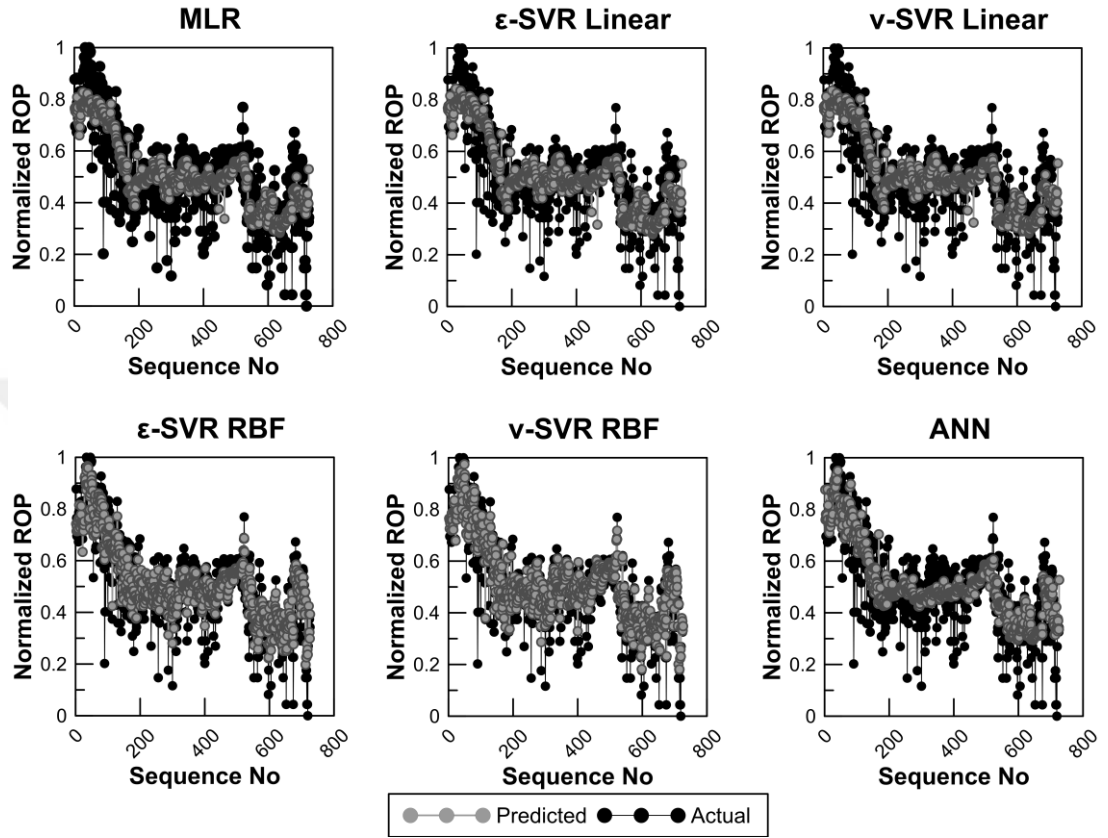


Figure D.164 : Model training for Case 3.3.3.

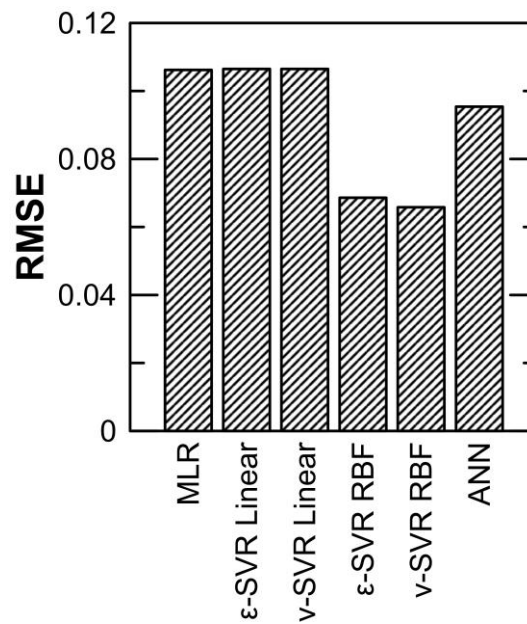


Figure D.165 : Training errors for Case 3.3.3.

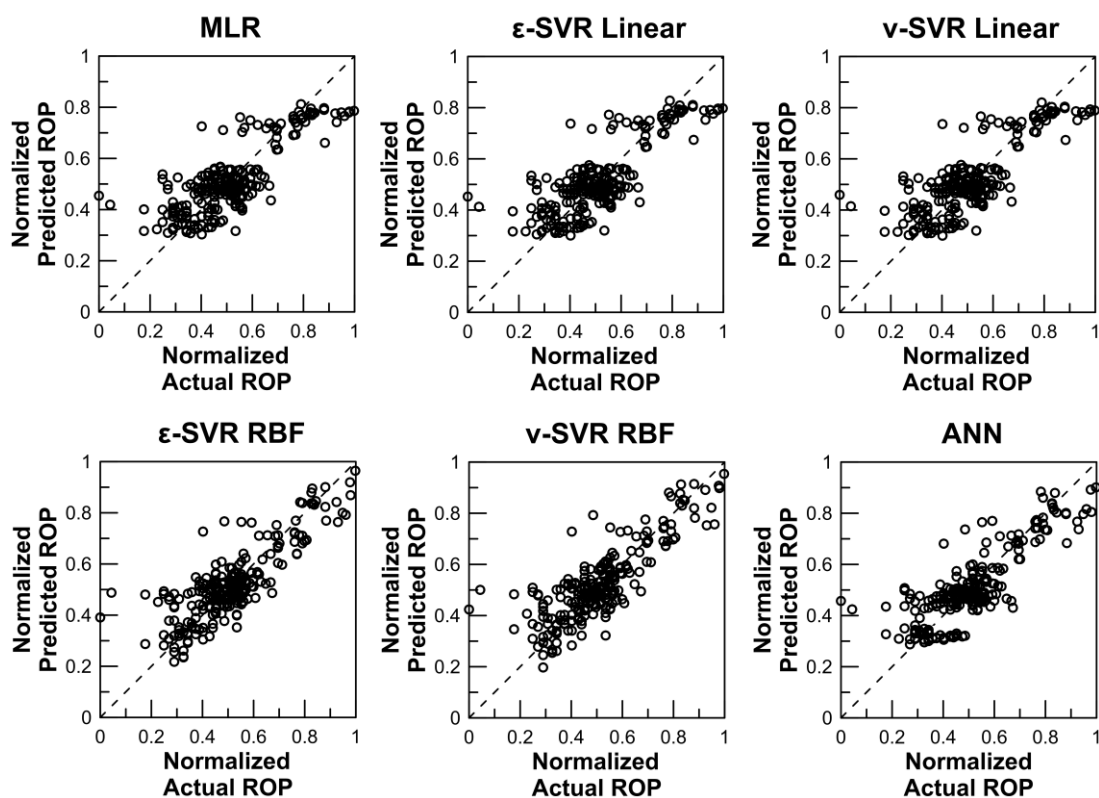


Figure D.166 : ROP predictions with testing subset for Case 3.3.3.

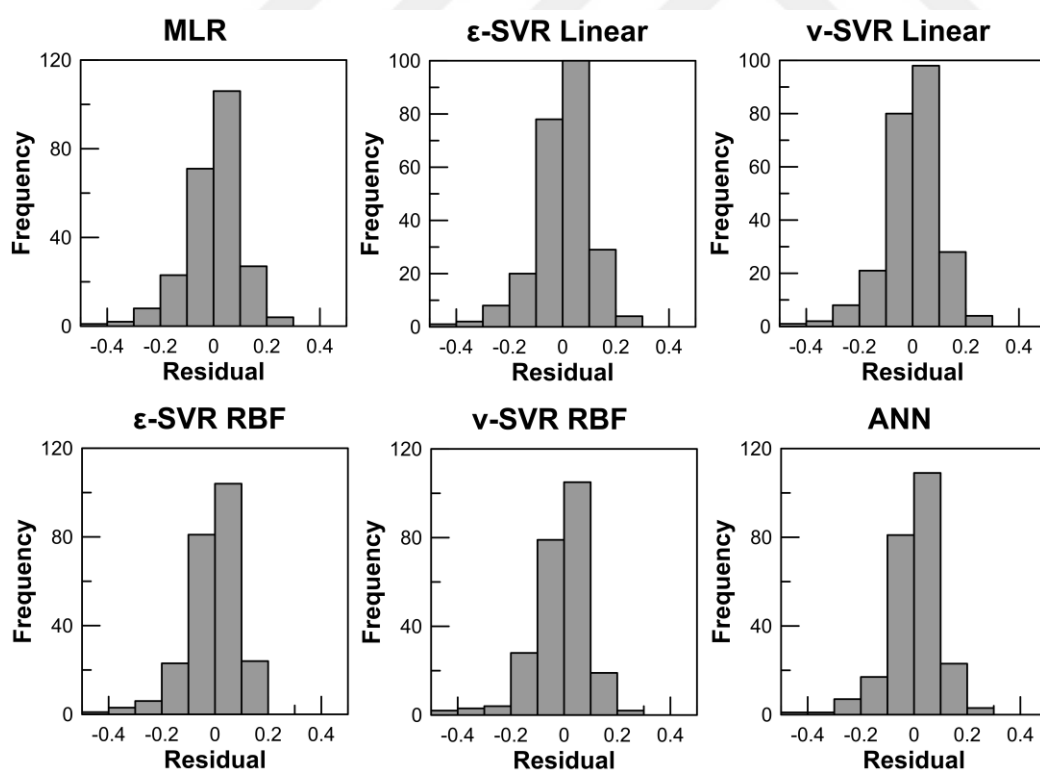


Figure D.167 : Error histograms for Case 3.3.3.

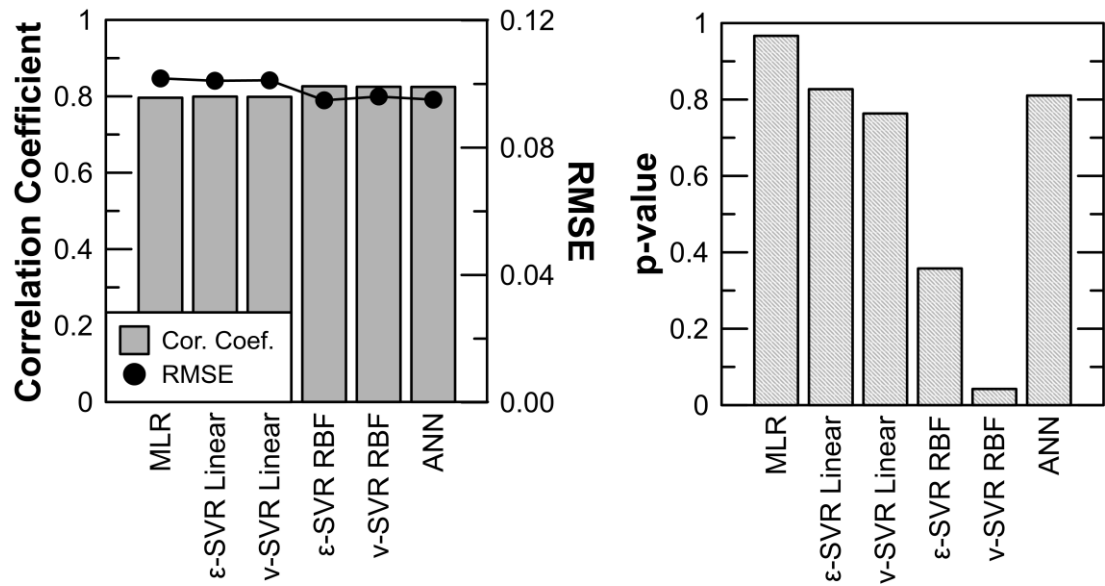


Figure D.168 : Statistical comparison for Case 3.3.3.

D.4.4. Case 3.3.4: BYM+SFG

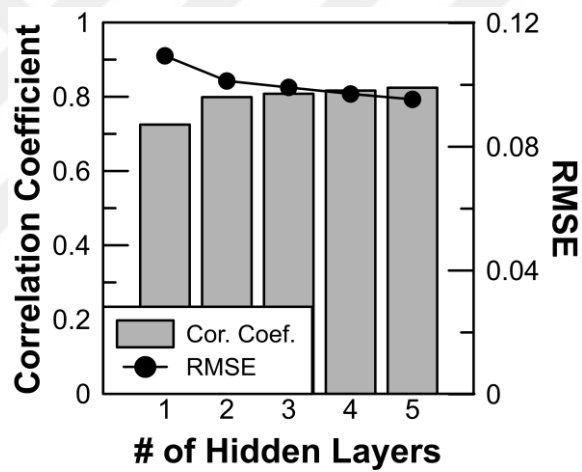


Figure D.169 : ANN hidden layer sensitivity for Case 3.3.4.

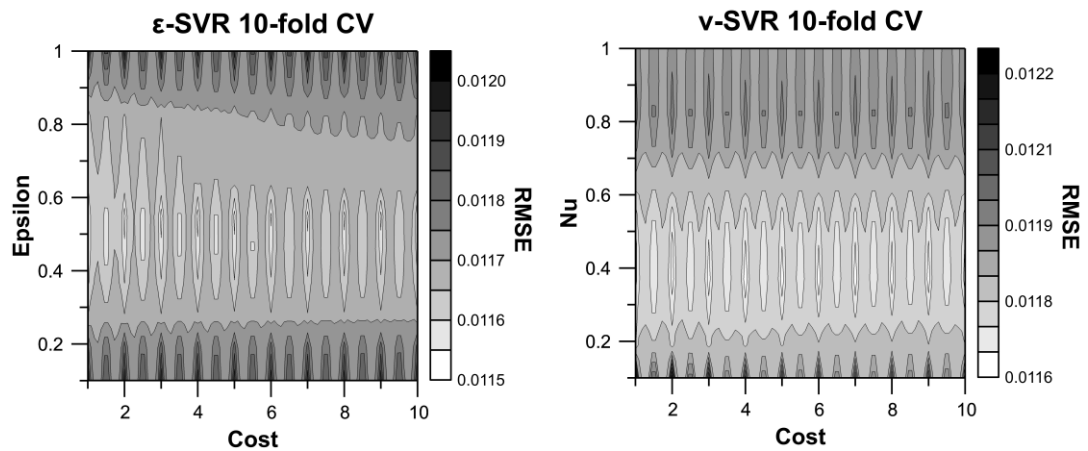


Figure D.170 : 10-fold CV grid search plots for Case 3.3.4.

Table D.25 : SVR hyper-parameters for Case 3.3.4.

ε -SVR					ν -SVR				
Linear			RBF		Linear			RBF	
C	ε	C	γ	ε	C	ν	C	γ	ν
2.5	0.55	2.0	1.0	0.2	4.0	0.35	1.5	1.0	0.4

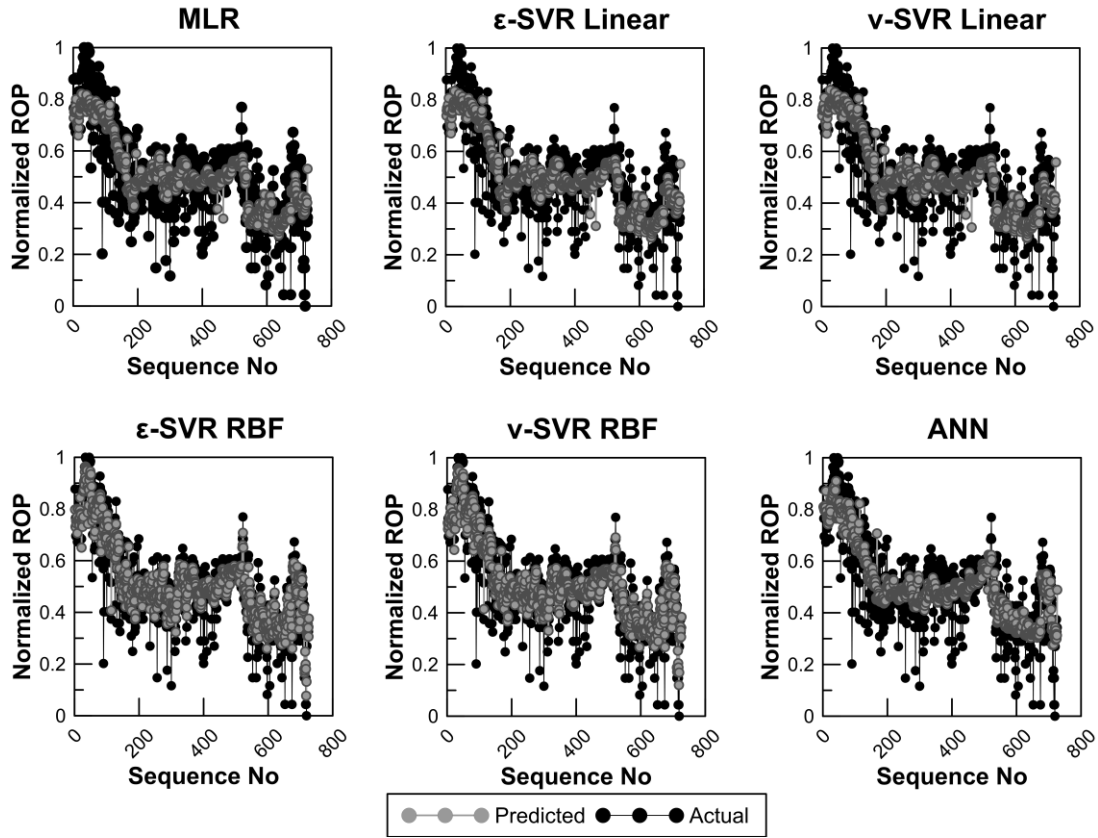


Figure D.171 : Model training for Case 3.3.4.

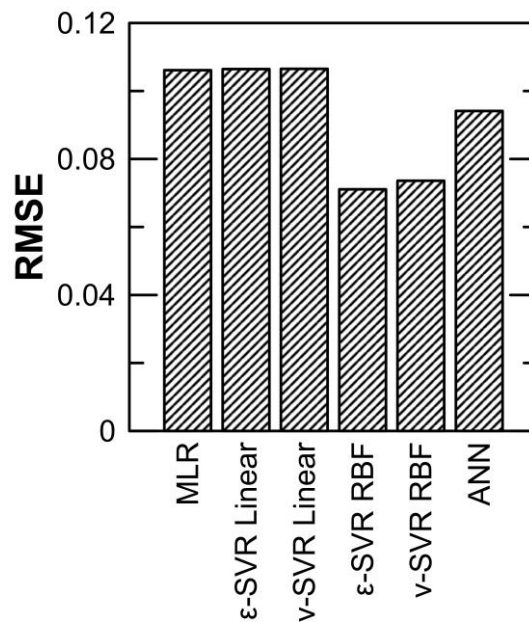


Figure D.172 : Training errors for Case 3.3.4.

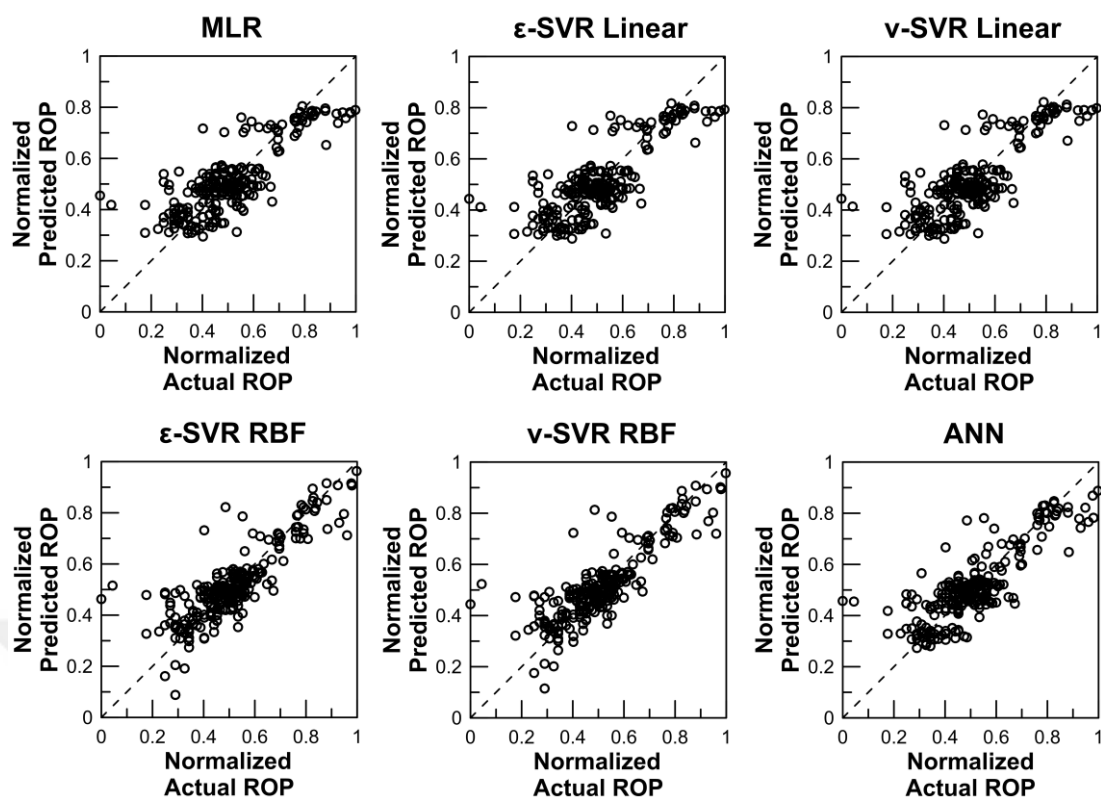


Figure D.173 : ROP predictions with testing subset for Case 3.3.4.

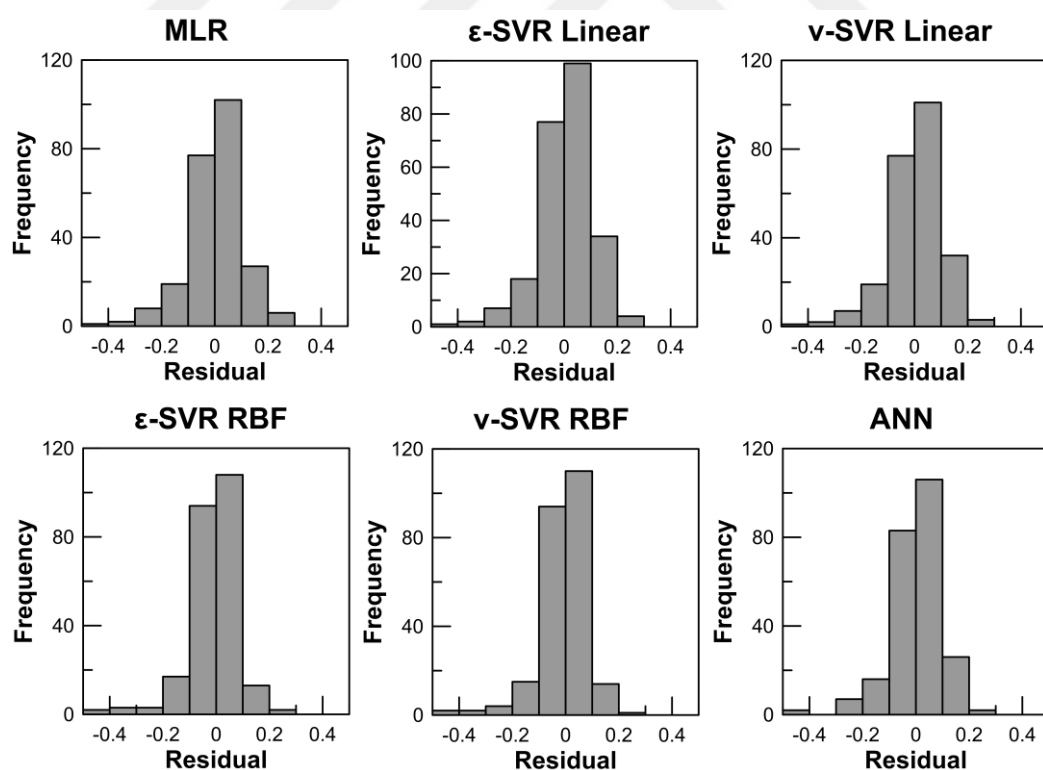


Figure D.174 : Error histograms for Case 3.3.4.

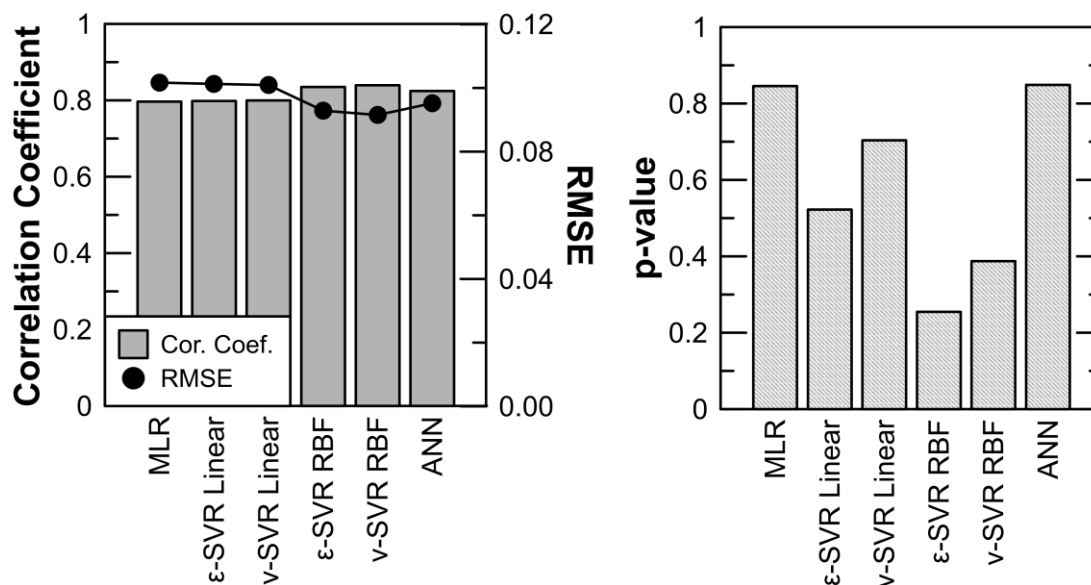


Figure D.175 : Statistical comparison for Case 3.3.4.

D.4.5. Case 3.3.5: BYM+TRQ

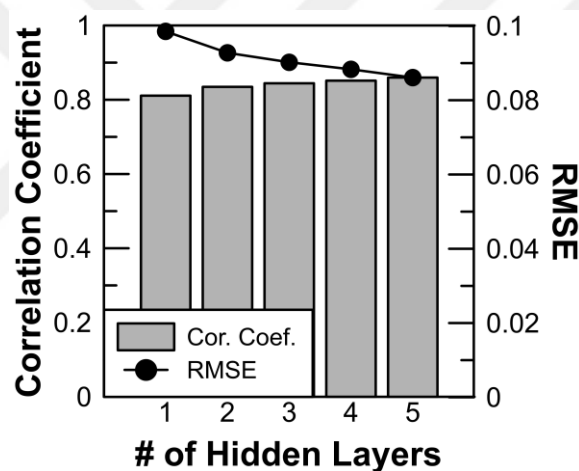


Figure D.176 : ANN hidden layer sensitivity for Case 3.3.5.

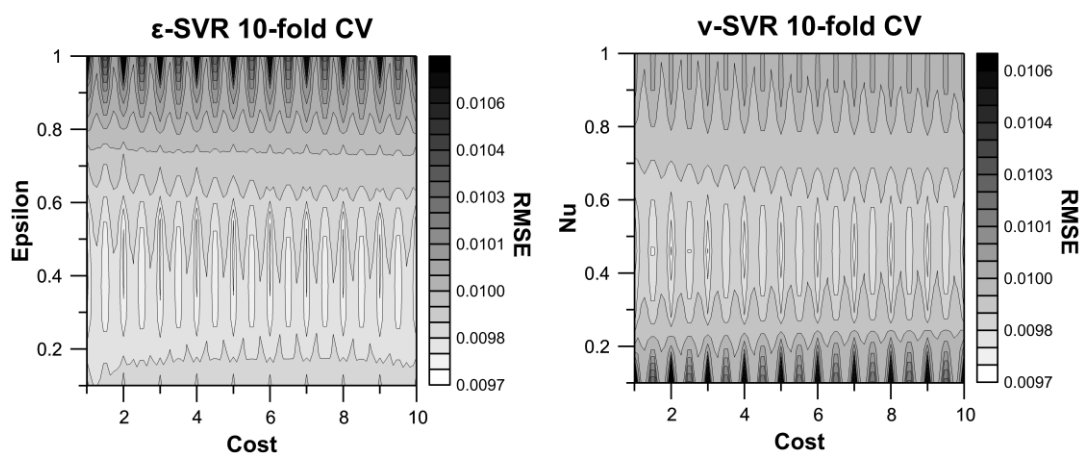


Figure D.177 : 10-fold CV grid search plots for Case 3.3.5.

Table D.26 : SVR hyper-parameters for Case 3.3.5.

ε -SVR					ν -SVR				
Linear			RBF		Linear			RBF	
C	ε	C	γ	ε	C	ν	C	γ	ν
1.0	0.3	1.5	1.0	0.2	1.0	0.45	1.5	1.0	0.5

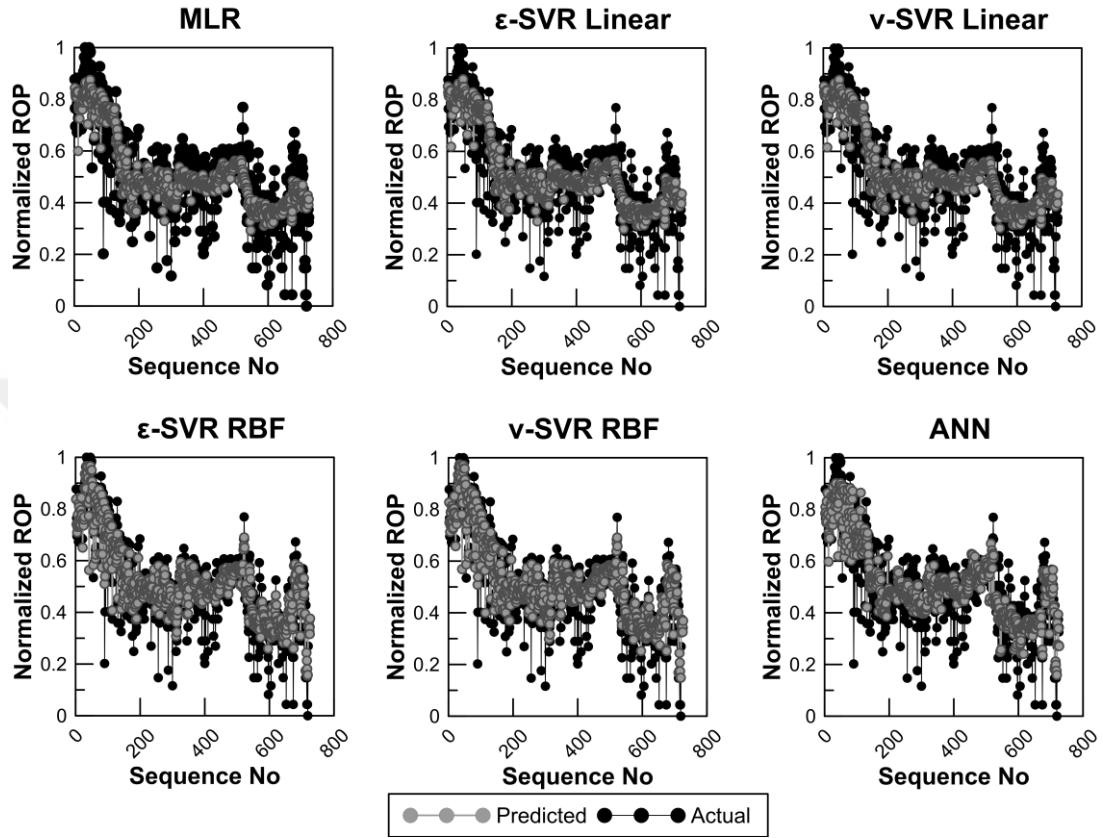


Figure D.178 : Model training for Case 3.3.5.

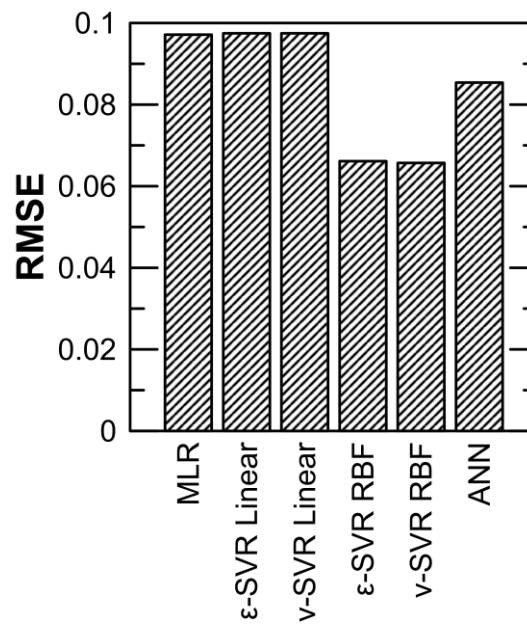


Figure D.179 : Training errors for Case 3.3.5.

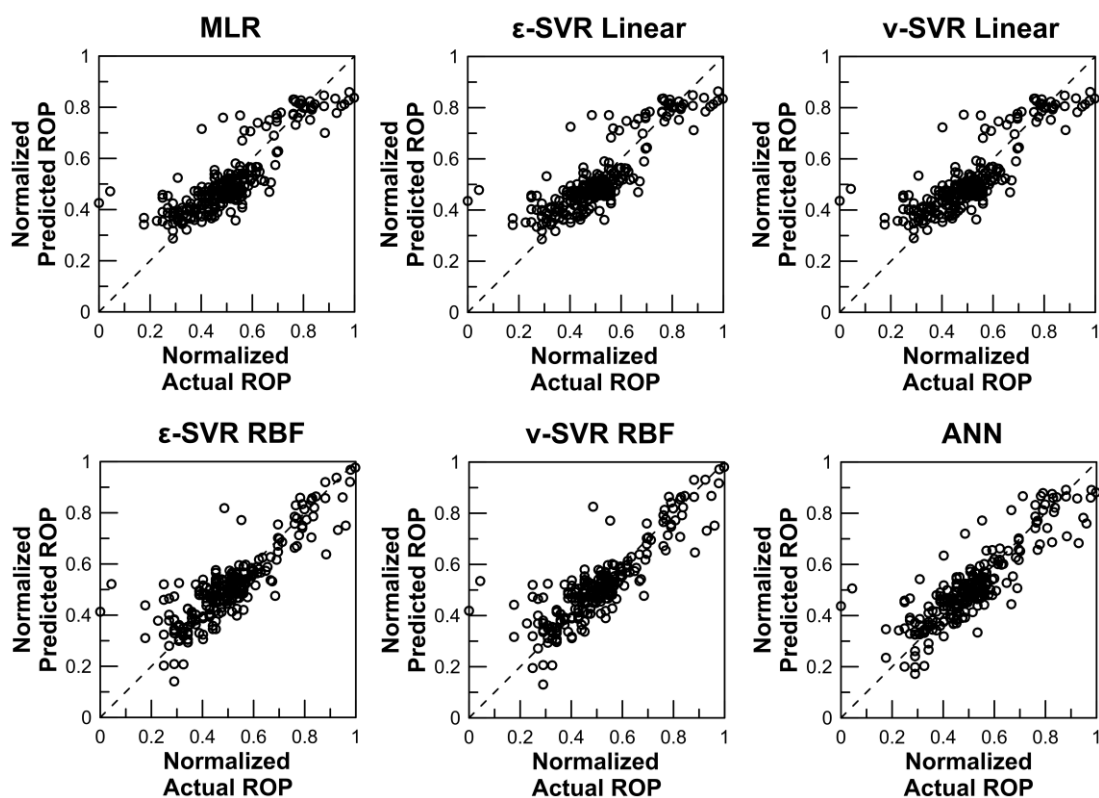


Figure D.180 : ROP predictions with testing subset for Case 3.3.5.

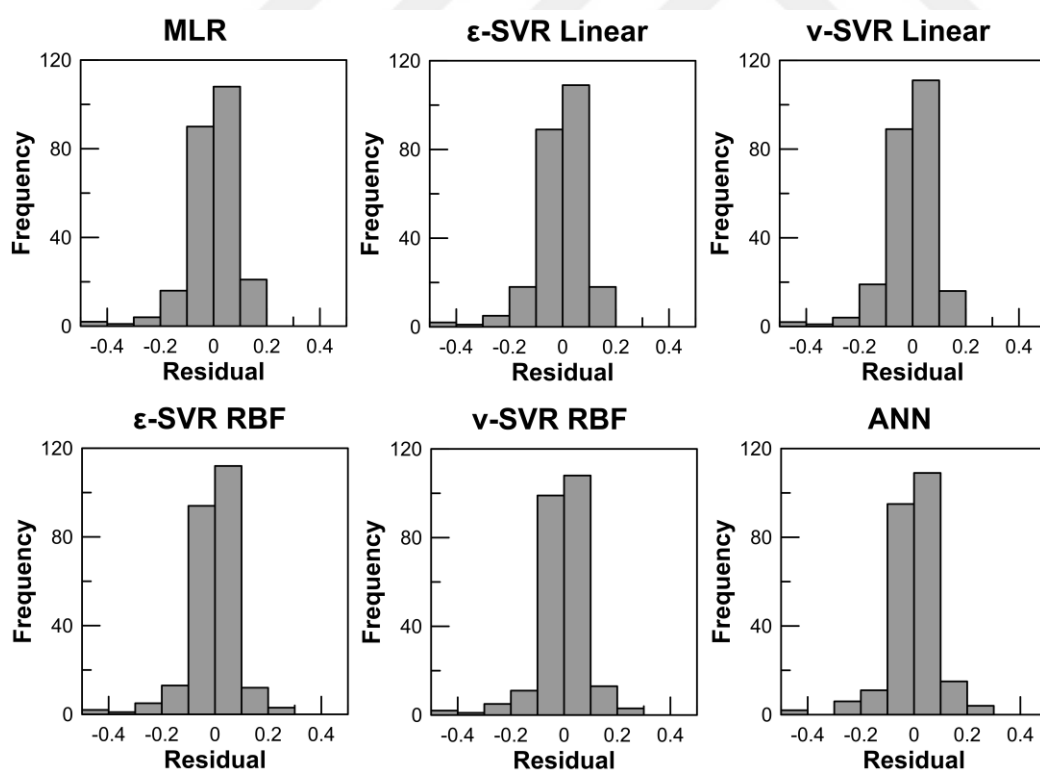


Figure D.181 : Error histograms for Case 3.3.5.

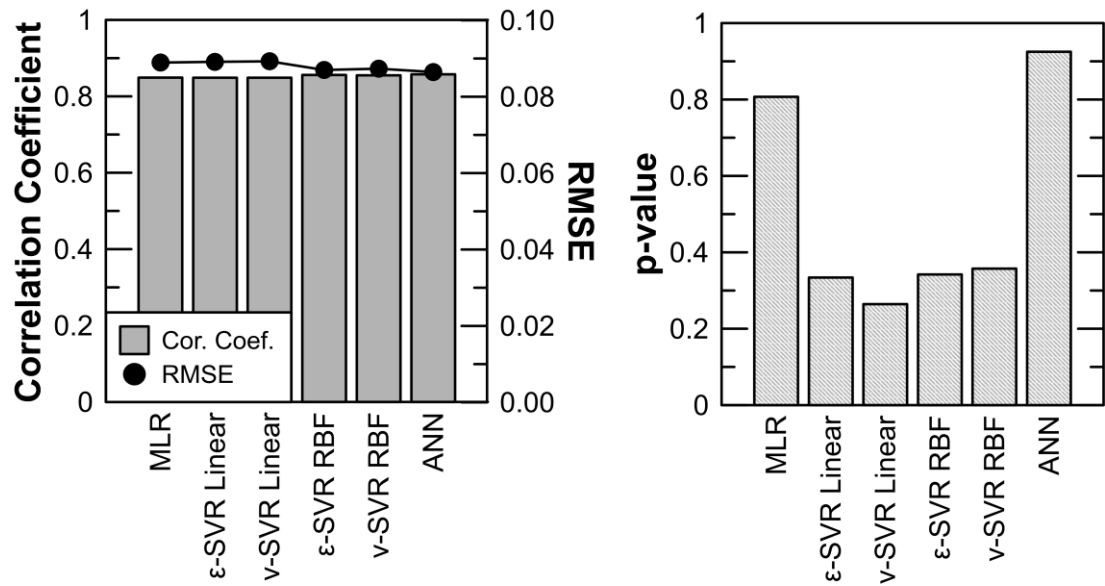


Figure D.182 : Statistical comparison for Case 3.3.5.

D.4.6. Case 3.3.6: BYM+BI

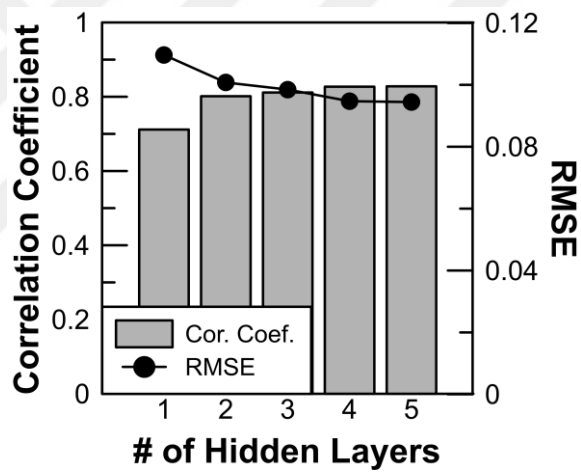


Figure D.183 : ANN hidden layer sensitivity for Case 3.3.6.

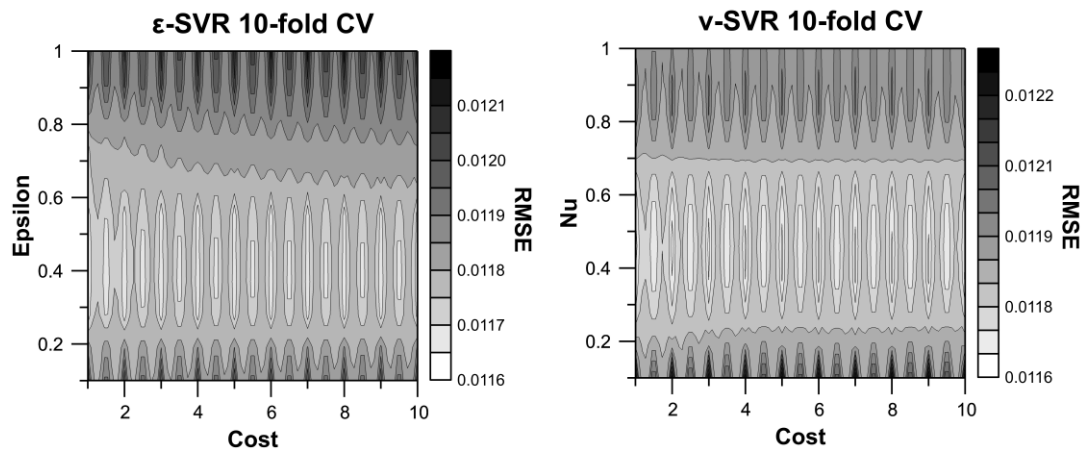


Figure D.184 : 10-fold CV grid search plots for Case 3.3.6.

Table D.27 : SVR hyper-parameters for Case 3.3.6.

ε -SVR					ν -SVR				
Linear			RBF		Linear			RBF	
C	ε	C	γ	ε	C	ν	C	γ	ν
8.0	0.5	1.0	1.0	0.1	6.5	0.35	1.0	1.0	0.6

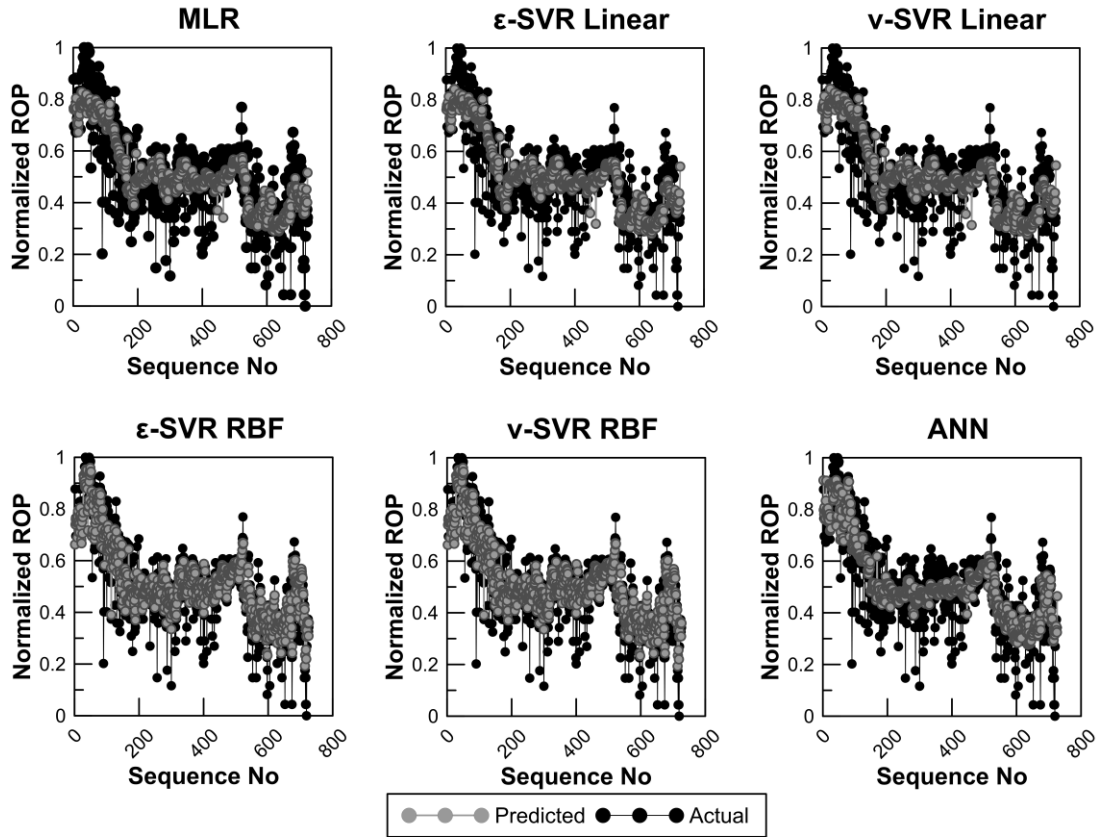


Figure D.185 : Model training for Case 3.3.6.

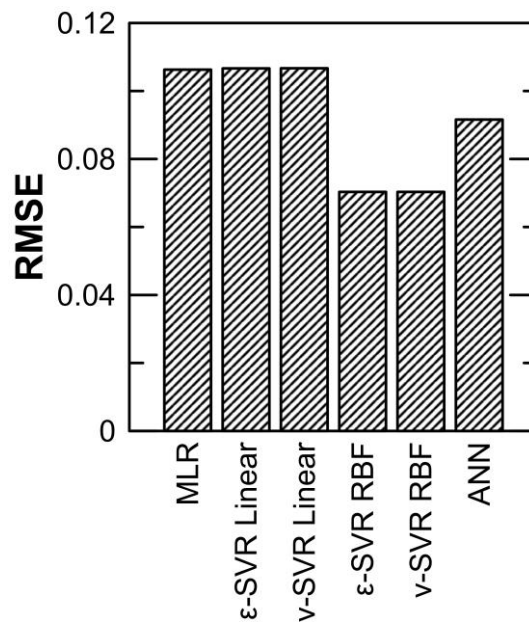


Figure D.186 : Training errors for Case 3.3.6.

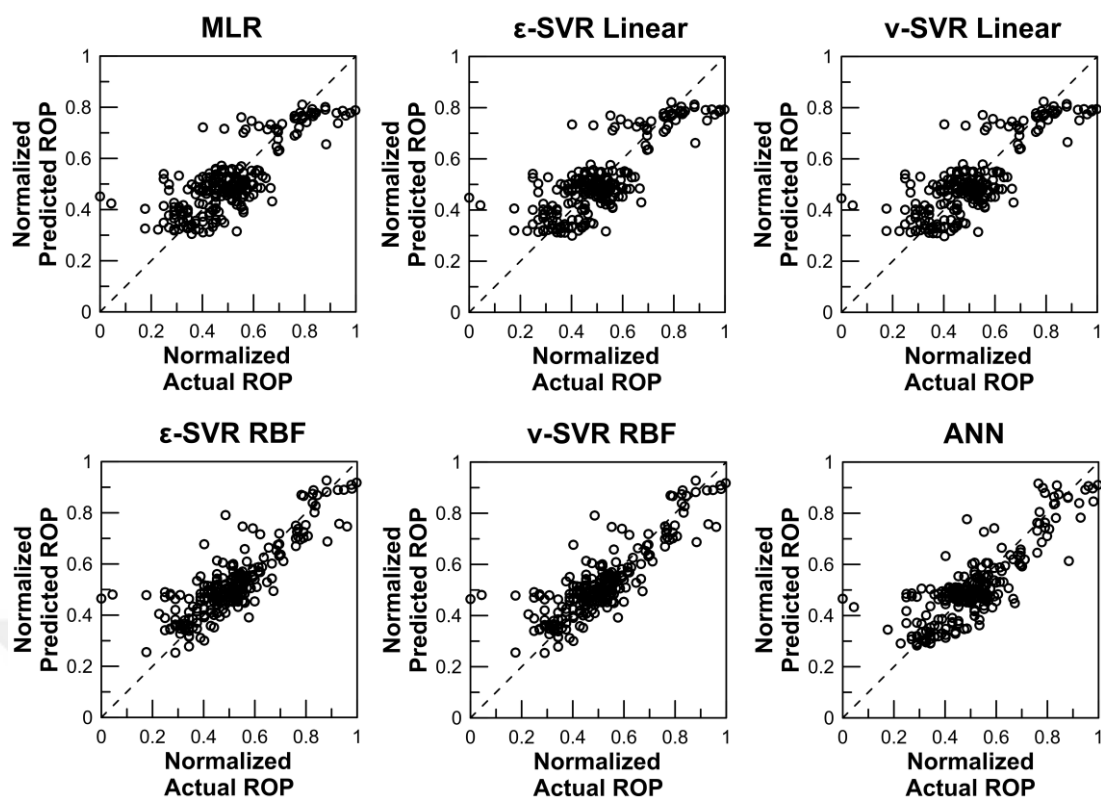


Figure D.187 : ROP predictions with testing subset for Case 3.3.6.

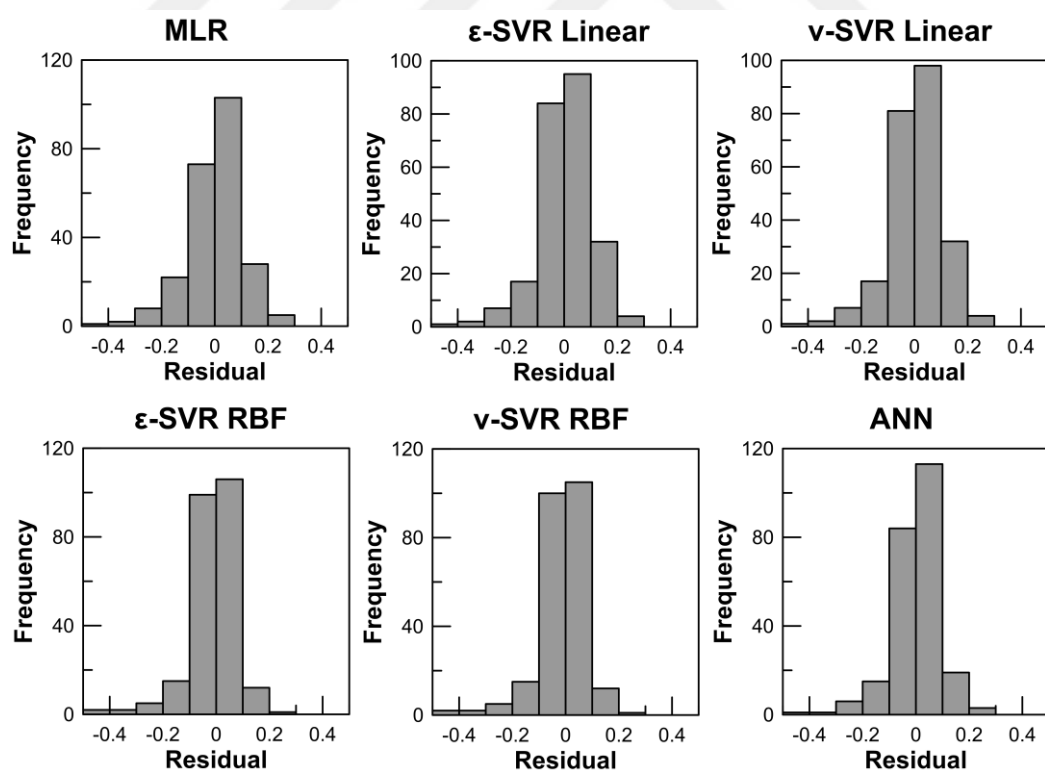


Figure D.188 : Error histograms for Case 3.3.6.

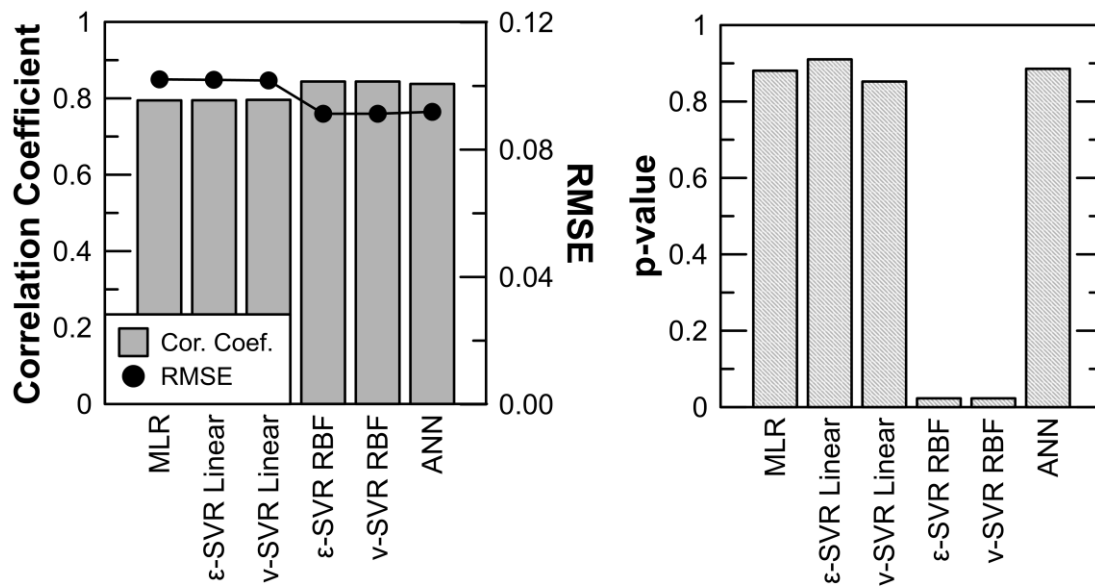


Figure D.189 : Statistical comparison for Case 3.3.6.

D.4.7. Case 3.3.7: BYM+All

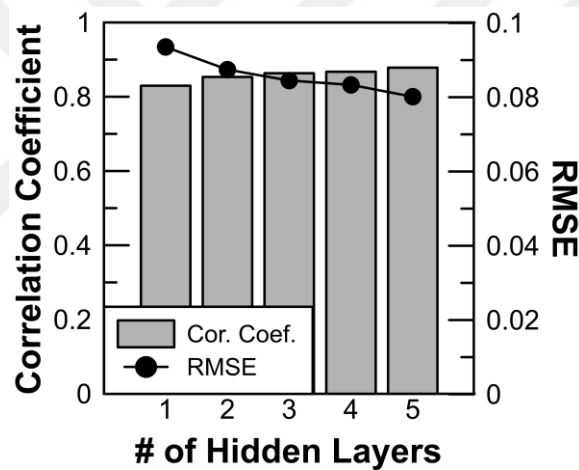


Figure D.190 : ANN hidden layer sensitivity for Case 3.3.7.

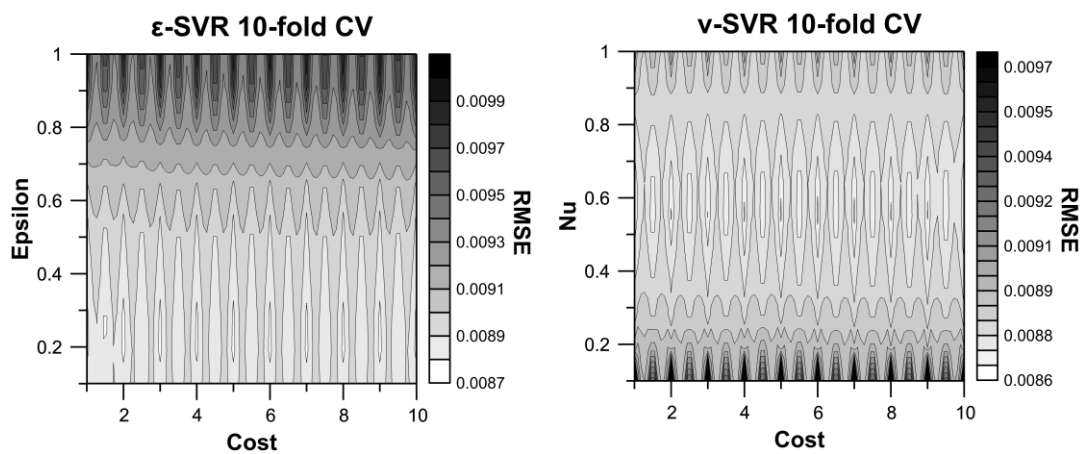


Figure D.191 : 10-fold CV grid search plots for Case 3.3.7.

Table D.28 : SVR hyper-parameters for Case 3.3.7.

ε -SVR					v-SVR				
Linear			RBF		Linear			RBF	
C	ε	C	γ	ε	C	ν	C	γ	ν
1.0	0.25	1.5	1.0	0.1	2.0	0.5	1.5	1.0	0.4

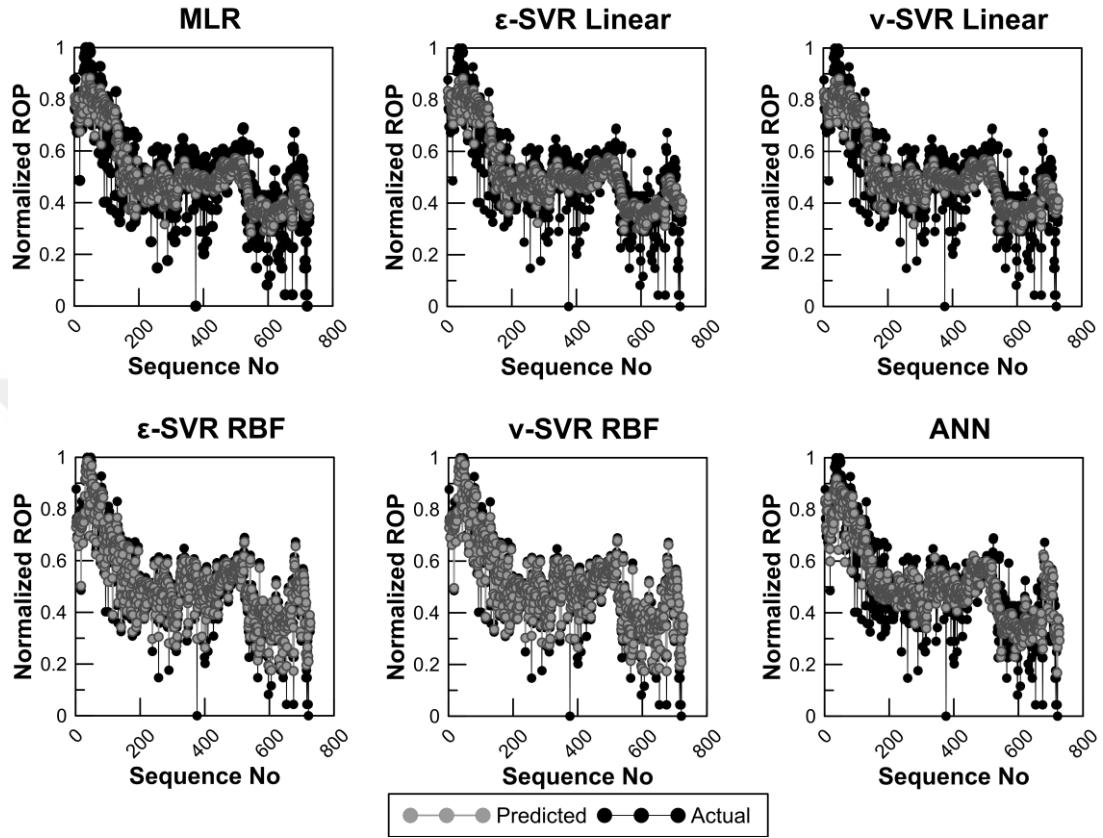


Figure D.192 : Model training for Case 3.3.7.

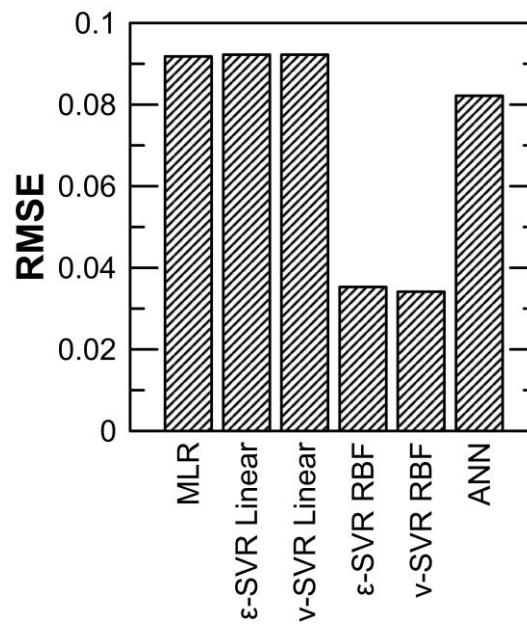


Figure D.193 : Training errors for Case 3.3.7.

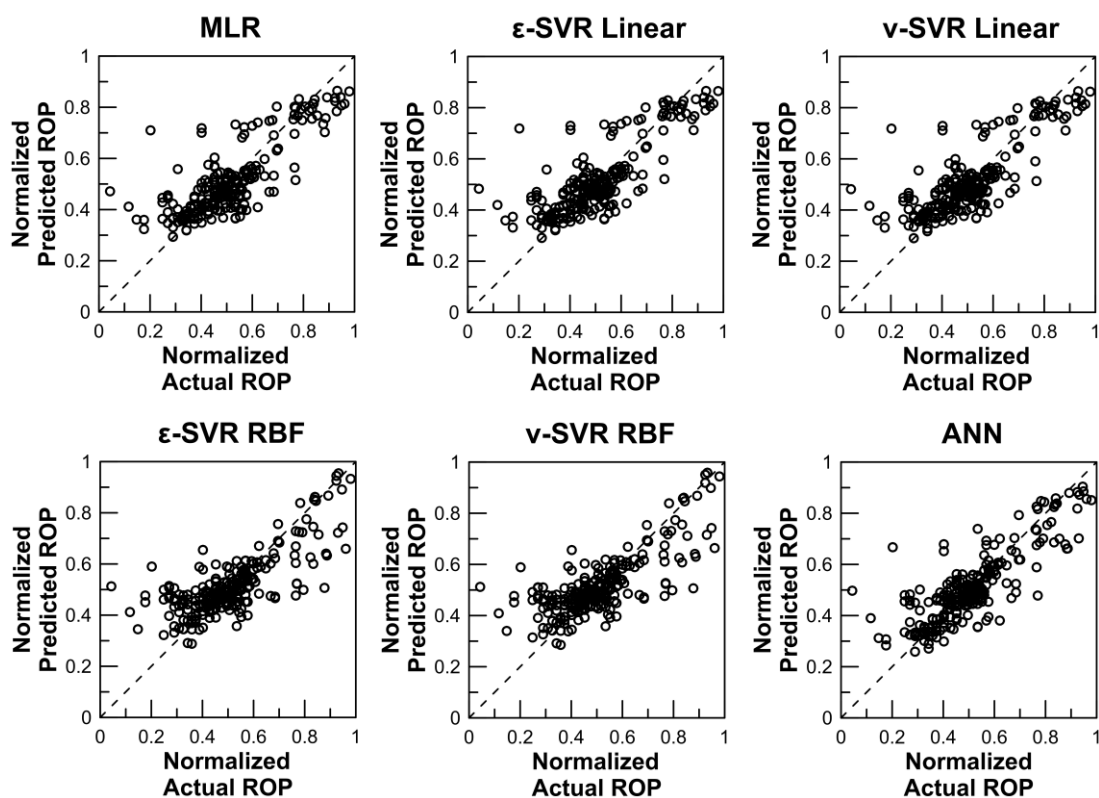


Figure D.194 : ROP predictions with testing subset for Case 3.3.7.

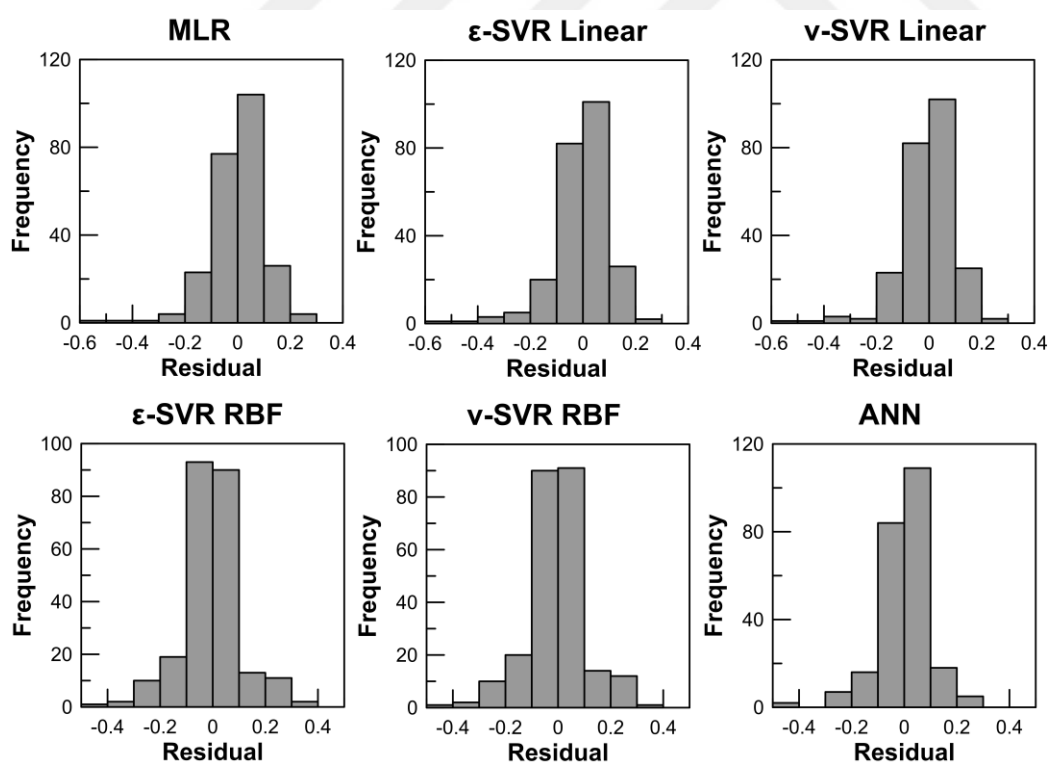


Figure D.195 : Error histograms for Case 3.3.7.

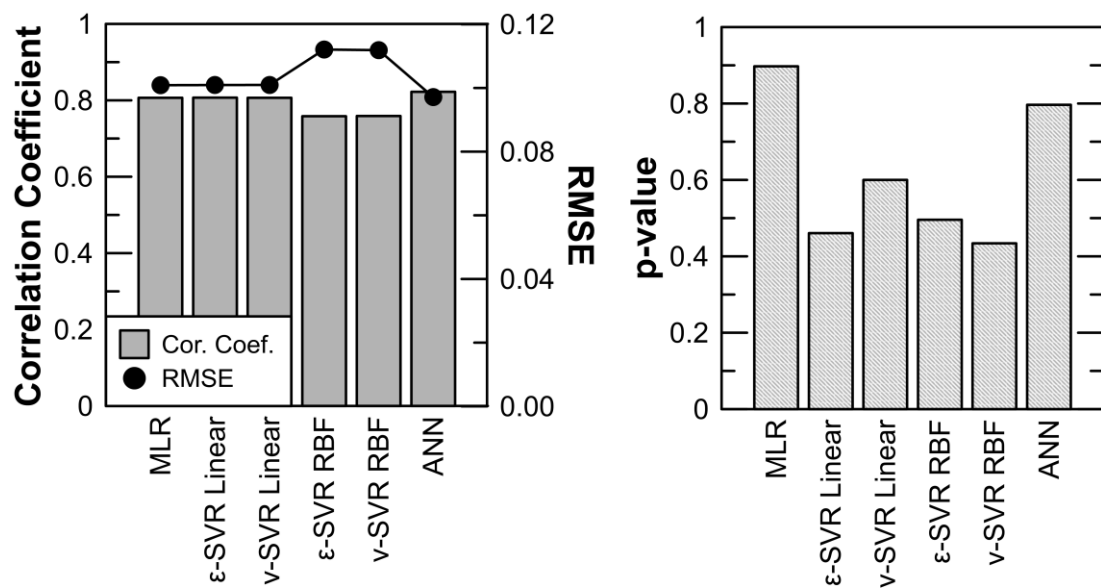


Figure D.196 : Statistical comparison for Case 3.3.7.

D.4.8. Case 3.3.8: Feature Selection (BYM+TRQ excluding x_3)

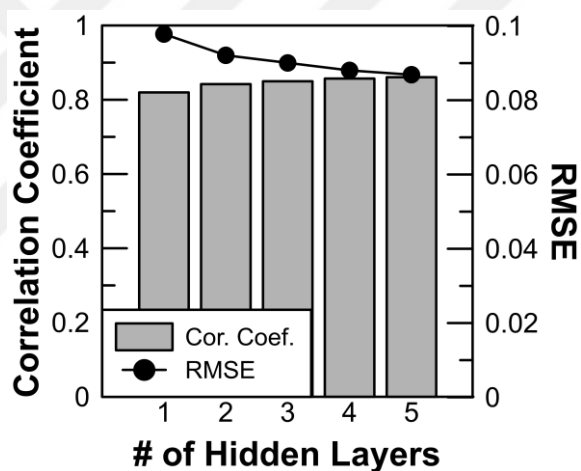


Figure D.197 : ANN hidden layer sensitivity for Case 3.3.8.

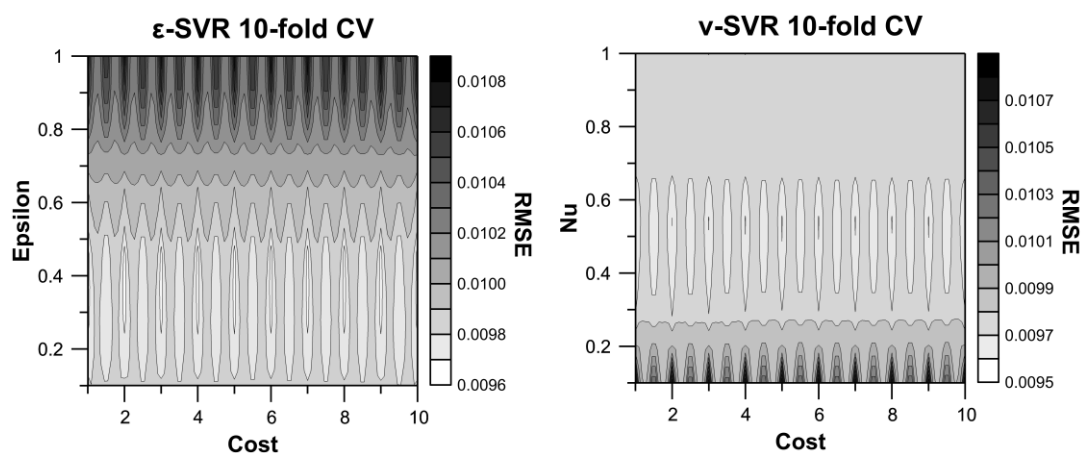


Figure D.198 : 10-fold CV grid search plots for Case 3.3.8.

Table D.29 : SVR hyper-parameters for Case 3.3.8.

ε -SVR					ν -SVR				
Linear			RBF		Linear			RBF	
C	ε	C	γ	ε	C	ν	C	γ	ν
1.0	0.35	1.5	1.0	0.2	1.5	0.45	2.0	1.0	0.6

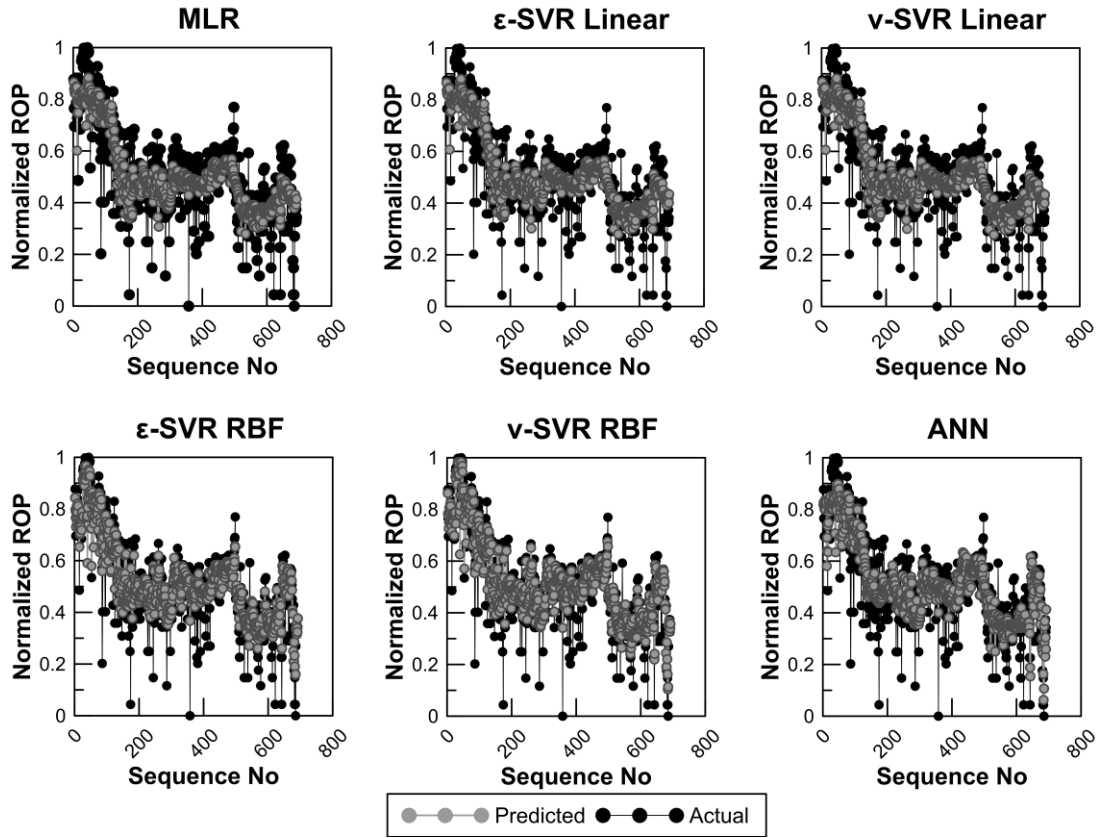


Figure D.199 : Model training for Case 3.3.8.

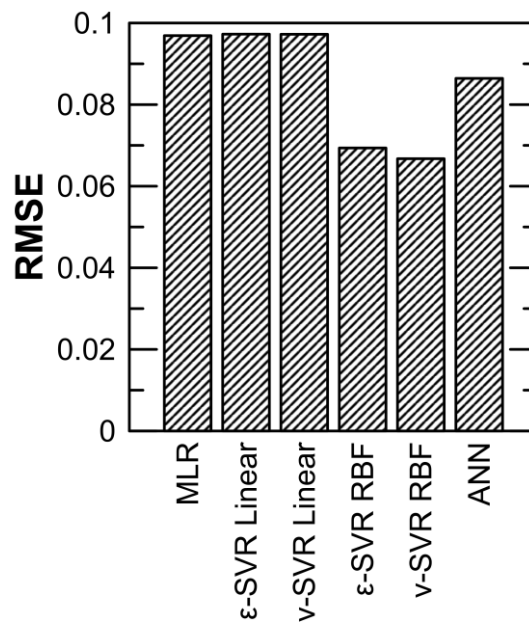


Figure D.200 : Training errors for Case 3.3.8.

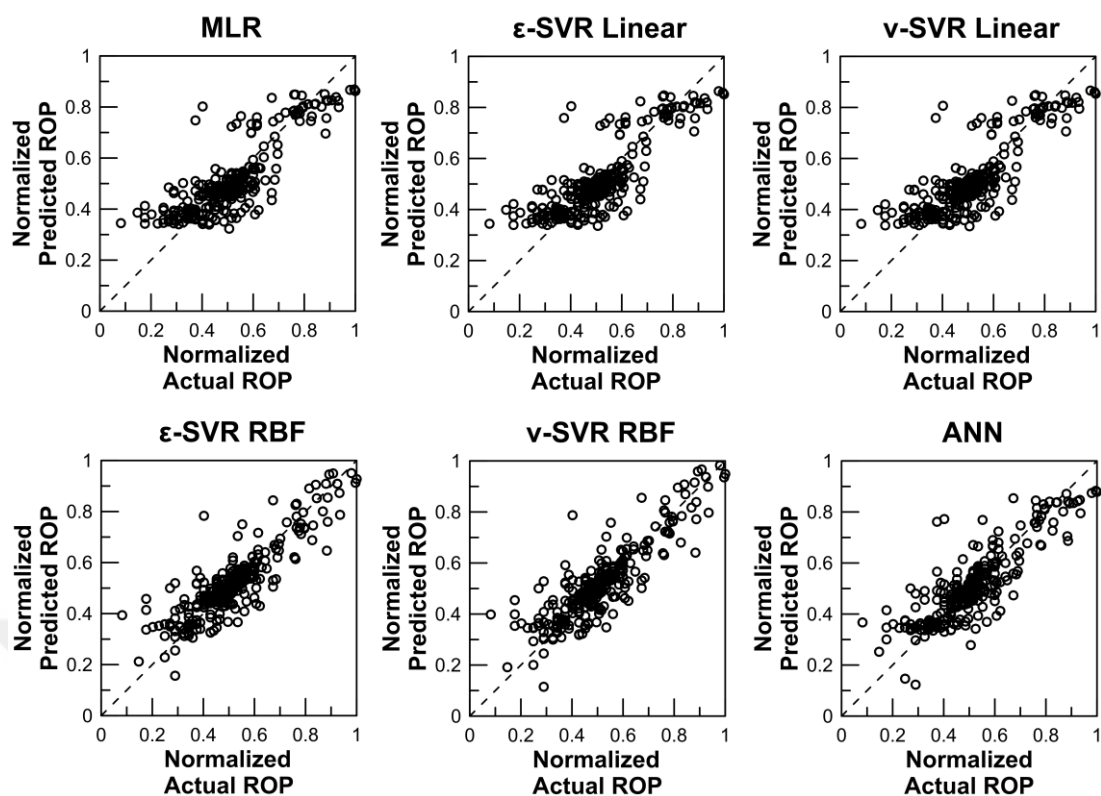


Figure D.201 : ROP predictions with testing subset for Case 3.3.8.

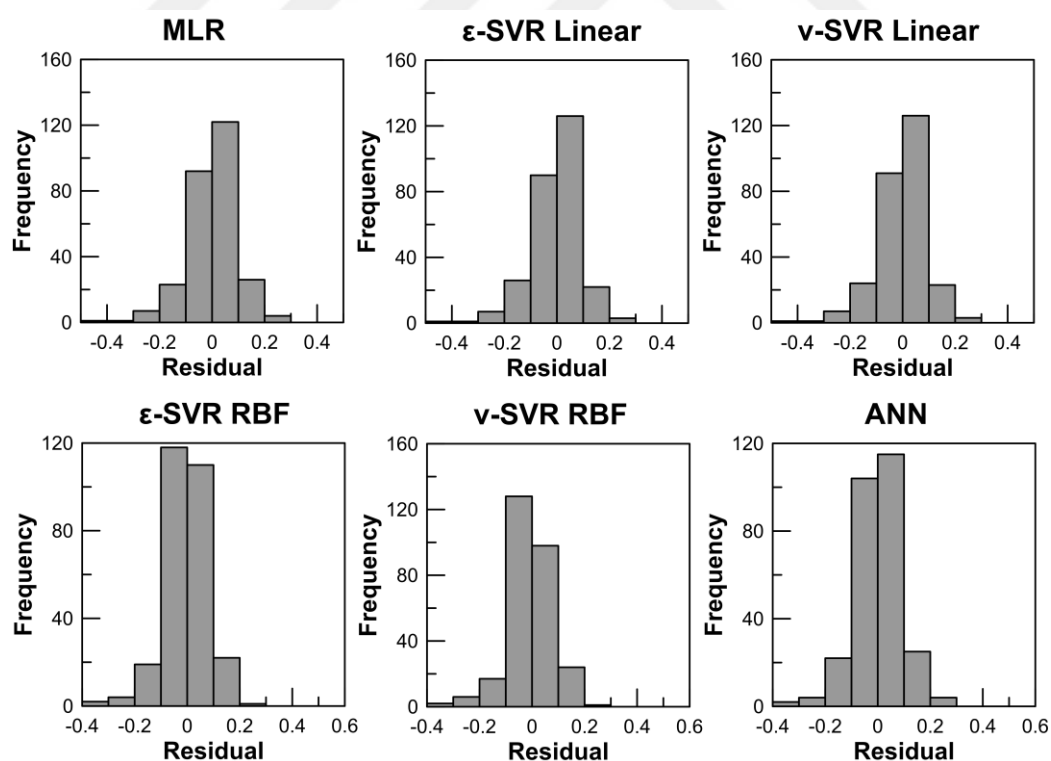


Figure D.202 : Error histograms for Case 3.3.8.

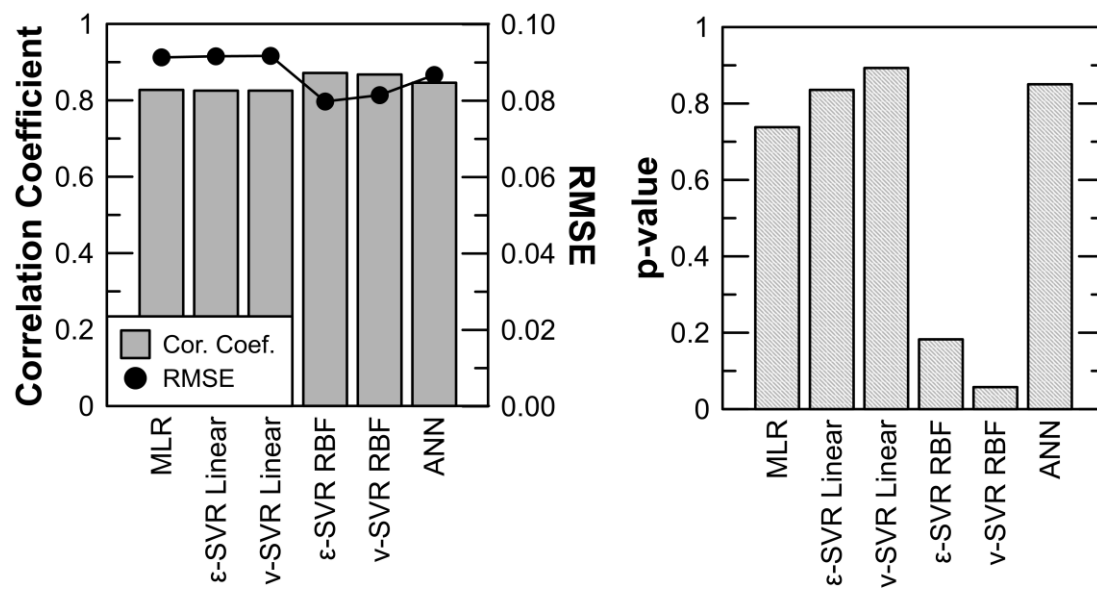


Figure D.203 : Statistical comparison for Case 3.3.8.

CURRICULUM VITAE

Name Surname : Korhan KOR

EDUCATION :

- **B.Sc.** : 2012, ITU, Faculty of Mines, Petroleum and Natural Gas Engineering
- **M.Sc.** : 2015, ITU, Grad. School of Sci. Eng. and Tech., Petroleum and Natural Gas Engineering

PROFESSIONAL EXPERIENCE AND REWARDS:

- 2015- Drilling Engineer at Turkish Petroleum Corp. (TPAO)
- 2013-2015 Research Assistant at Istanbul Technical University

PUBLICATIONS, PRESENTATIONS AND PATENTS ON THE THESIS:

- **Kor, K.**, Ertekin, Ş., Yamanlar, Ş., and Altun, G. 2021. Penetration Rate Prediction in Heterogeneous Formations: A Geomechanical Approach Through Machine Learning, *Journal of Petroleum Science and Engineering*, 207, 109138.
- **Kor, K.** and Altun, G. 2020. Is Support Vector Regression Method Auitable for Predicting Rate of Penetration?, *Journal of Petroleum Science and Engineering*, 194, 107542.

OTHER PUBLICATIONS, PRESENTATIONS AND PATENTS:

- Canbaz, C.H., Temizel, C., Palabıyık, Y., **Kor, K.**, Hendrandingrat, L., Hoşgör, F.B., and Akşahan, F. 2021. Environmentally-Friendly Production and Recovery Processes for Heavy Oils. SPE Trinidad and Tobago Section Energy Resources Conference, June, 2021, Online.
- **Kor, K.** 2020. Petrol ve Doğal Gaz Mühendisliği'nde Yapay Zeka Uygulamaları. In M.F. Tuğan, O.Yardımcı (Eds.), *21. Yüzyılda Petrol ve Doğal Gaz Mühendisliği* (2nd ed., pp.231-250), Ankara : Gazi Kitabevi.
- **Kor, K.** and Altun, G. 2020. ROP Prediction with Bourgoyne and Young Method: A Statistical Point of View. Turkey IV. Scientific and Technical Petroleum Congress, April 13-15, Ankara.

- **Kor, K.** and Altun, G. 2015. Destek Vektör Makinesi Yöntemi ile Sondaj İlerleme Hızı Optimizasyonu. IPETGAS 2015 – 20th International Petroleum and Natural Gas Congress and Exhibition of Turkey, May 27-29, Ankara.

