**ÇUKUROVA UNIVERSITY**
**INSTITUTE OF NATURAL AND APPLIED SCIENCES**

**PhD THESIS**

**Mustafa Mikail ÖZÇİLOĞLU**

# PREDICTON OF UPPER BODY POWER OF CROSS-COUNTRY SKIERS USING MACHINE LEARING METHODS COMBINED WITH FEATURE SELECTION ALGORITHMS

**DEPARTMENT OF ELECTRICAL AND ELECTRONICS ENGINEERING**

**ADANA, 2016**

ÇUKUROVA UNIVERSITY
INSTITUTE OF NATURAL AND APPLIED SCIENCES

**PREDICTON OF UPPER BODY POWER OF CROSS-COUNTRY SKIERS USING MACHINE LEARNING METHODS COMBINED WITH FEATURE SELECTION ALGORITHMS**

**Mustafa Mikail ÖZÇİLOĞLU**

**PhD THESIS**

**DEPARTMENT OF ELECTRICAL AND ELECTRONICS ENGINEERING**

We certify that the thesis titled above was reviewed and approved for the award of degree of the Doctor of Philosophy by the board of jury on 22/06/2016.

| ................................................ | .................................. | .................................. |
| --- | --- | --- |
| Assoc. Prof. Dr. M. Fatih AKAY | Prof. Dr. Ulus ÇEVİK | Prof. Dr. Bülent TAVLI |
| SUPERVISOR | MEMBER | MEMBER |

| ........................................................ | ........................................ |
| --- | --- |
| Assoc. Prof. Dr. Zekeriya TÜFEKÇİ | Asst. Prof. Dr. Şule ÇOLAK |
| MEMBER | MEMBER |

This PhD Thesis is written at the Department of Institute of Natural And Applied Sciences of Çukurova University.
**Registration Number**:

**Prof. Dr. Mustafa GÖK**
**Director**
**Institute of Natural and Applied Sciences**

**This thesis was supported by the Scientific Research Projects center of Çukurova University. (Project Number: FDK-2015-5093).**

# ABSTRACT

## PhD THESIS

**PREDICTION OF UPPER BODY POWER OF CROSS-COUNTRY SKIERS USING MACHINE LEARNING METHODS COMBINED WITH FEATURE SELECTION ALGORITHMS**


**Mustafa Mikail ÖZÇİLOĞLU**

**ÇUKUROVA UNIVERSITY**
**INSTITUTE OF NATURAL AND APPLIED SCIENCES**
**DEPARTMENT OF ELECTRICAL AND ELECTRONICS ENGINEERING**

Supervisor : Assoc. Prof. Dr. M. Fatih AKAY
Year: 2016, Pages: 95
Jury : Prof. Dr. Ulus ÇEVİK
: Prof. Dr. Bülent TAVLI
: Assoc. Prof. Dr. Zekeriya TÜFEKÇİ
: Asst. Prof. Dr. Şule ÇOLAK


Upper body power (UBP) is one of the most important factors affecting the performance of cross-country skiers during races. Although some initial studies have already attempted to predict UBP, until now, no study has attempted to apply machine learning methods combined with various feature selection algorithms to identify the discriminative features for prediction of UBP. The purpose of this study is to develop new prediction models for predicting the 10-second UBP ($UBP_{10}$) and 60-second UBP ($UBP_{60}$) of cross-country skiers by using General Regression Neural Networks (GRNN), Radial-Basis Function Network (RBF), Multilayer Perceptron (MLP), Support Vector Machine (SVM), Single Decision Tree (SDT) and Tree Boost (TB) along with the Relief-F feature selection algorithm, minimum redundancy maximum relevance (mRMR) feature selection algorithm and the Correlation-based Feature Subset Selection (CFS). Several models have been developed to predict $UBP_{10}$ and $UBP_{60}$ of cross-country skiers using two datasets. 10-fold cross validation has been performed for model testing. The efficiency of the prediction models has been calculated with their multiple correlation coefficients (*R’s*), standard error of estimates (*SEE’s*) and mean absolute percentage errors (*MAPE’s*). The results emphasize that GRNN-based prediction models show higher performance than the other regression methods. Also, using less number of predictor variables than the full set of predictor variables can be useful for prediction of $UBP_{10}$ and $UBP_{60}$ with comparable error rates.



**Key Words:** Upper Body Power, Feature Selection Algorithms, Machine Learning Methods

## ÖZ

## DOKTORA TEZİ

# NİTELİK SEÇME ALGORİTMALARIYLA BİRLEŞTİRİLMİŞ MAKİNE ÖĞRENME YÖNTEMLERİ KULLANILARAK KROS KAYAKÇILARIN ÜST VÜCUT GÜÇ TÜKETİMİNİN TAHMİN EDİLMESİ

**Mustafa Mikail ÖZÇİLOĞLU**

**ÇUKUROVA ÜNİVERSİTESİ**
**FEN BİLİMLERİ ENSTİTÜSÜ**
**ELEKTRİK ELEKTRONİK MÜHENDİSLİĞİ ANABİLİM DALI**

Danışman : Doç. Dr. M. Fatih AKAY
Yıl: 2016, Sayfa: 95
Jüri : Prof. Dr. Ulus ÇEVİK
: Prof. Dr. Bülent TAVLI
: Doc. Dr. Zekeriya TÜFEKÇİ
: Yrd. Doc. Dr. Şule ÇOLAK

Üst Vücut Gücü (ÜVG), kros kayakçıların yarış esnasındaki performanslarını etkileyen en önemli faktörlerden biridir. ÜVG tahmini için bazı çalışmalar bulunsa da, şimdiye kadar, makine öğrenme metotlarıyla çeşitli özellik seçimi algoritmaları birlikte kullanılarak UBP tahmininde ayrımcı özellikleri tanımlamak için hiçbir çalışma bulunmamaktadır. Bu tezin amacı, Relief-F, mRMR ve CFS nitelik seçimi algoritmalarıyla Genelleştirilmiş Regresyon Sinir Ağı (General Regression Neural Network –GRNN), Ağaç Destekleme (Tree Boost), Çok Katmanlı Algılayıcı (Multilayer Perceptron- MLP), Destek Vektör Makinesi (Support Vector Machine - SVM), Radyal Tabanlı Fonksiyon Sinir Ağı (Radial Basis Function Neural Network – RBF) ve Tekli Karar Ağacı (Single Decision Tree - SDT) kullanılarak kros kayak yarışçılarının 10 saniye ÜVG ($ÜVG_{10}$) ve 60 saniye ÜVG ($ÜVG_{60}$) tahmini için yeni tahmin modelleri geliştirmektir. Söz konusu regresyon yöntemleri kullanılarak, kros kayakçılarının $ÜVG_{10}$ ve $ÜVG_{60}$ tahmini için birçok çeşitli tahmin modeli geliştirilmiştir. Model testleri için 10 katlı çapraz doğrulama uygulanmıştır. Tahmin modellerinin verimliliği, onların çoklu korelasyon katsayıları (*R*), standart tahmin hataları (*SEE*) ve mutlak yüzdelik hata ortalamaları (*MAPE*) ile hesaplanmıştır. Sonuçlar göstermektedir ki, GRNN tabanlı tahmin modelleri diğer regresyon yöntemleriyle üretilmiş olan tahmin modellerinden çok daha iyi performans (en yüksek *R* ve en düşük *SEE*) vermektedir. Ayrıca, tahmin değişkenlerinin tamamını kullanmak yerine daha az miktarda tahmin değişkeni kullanmak, benzer hata oranları ile $ÜVG_{10}$ ve $ÜVG_{60}$ tahmini için yararlı olabilir.

**Anahtar Kelimeler:** Üst vücut güç tüketimi, Nitelik seçimi algoritmaları, Makine öğrenme yöntemleri

## ACKNOWLEDGMENTS

Foremost, I wish to express my sincere gratitude to my supervisor Assoc. Prof. Dr. M. Fatih AKAY, for his supervision guidance, patience, motivation and useful suggestions for this work.

I would like to thank Assoc. Prof. Dr. Dan HEIL for providing us with the datasets.

I would like to thank members of the PhD thesis jury, Prof. Dr. Ulus ÇEVİK, Prof. Dr. Bülent TAVLI, Assoc. Prof. Dr. Zekeriya TÜFEKÇİ and Asst. Prof. Dr. Şule ÇOLAK, for their suggestions and corrections.

Also, I would like to thank Cukurova University Scientific Research Projects Center for supporting this work (Project no: FDK-2015-5093).

Last but not the least, I would like to thank my wife, my family and my dear all friends for their endless support and encouragements for my life and career.

**CONTENTS** **PAGE**

ABSTRACT ........................................................................................................I
ÖZ……….. ........................................................................................................ II
ACKNOWLEDGMENTS ........................................................................................ III
CONTENTS ........................................................................................................IV
LIST OF TABLES…. ............................................................................................VI
LIST OF FIGURES…. ............................................................................................ X
LIST OF ABBREVIATIONS ................................................................................ XIV
1. INTRODUCTION.................................................................................................. 1
1.1. Overview of Cross-Country Skiing and Upper Body Power ............................ 1
1.2. Motivation, Purpose and Contributions of the Thesis....................................... 4
1.3. Overview of the Dataset...................................................................................... 5
1.4. Literature Review................................................................................................ 6
1.4.1. Measurement of UBP ..................................................................................... 6
1.4.2. Prediction of UBP without Feature Selection ............................................... 7
1.4.3. Prediction of UBP with Feature Selection .................................................... 9
2. OVERVIEW OF METHODS ................................................................................. 13
2.1. Overview of Regression Methods ..................................................................... 13
2.1.1. Generalized Regression Neural Network ...................................................... 13
2.1.2. Support Vector Machines ............................................................................... 15
A. Linear SVM ........................................................................................................ 15
B. Nonlinear SVM ................................................................................................... 17
2.1.3. Tree Boost ........................................................................................................ 18
2.1.4. Multilayer Perceptron...................................................................................... 19
2.1.5. Radial Basis Function Network....................................................................... 21
2.1.6. Single Decision Trees....................................................................................... 22
2.2. Overview of Feature Selection Algorithms ........................................................ 23
2.2.1. Relief-F.............................................................................................................. 23
2.2.2. Minimum Redundancy Maximum Relevance................................................. 24
2.2.3. Correlation-Based Feature Selection............................................................... 25
3. DEVELOPMENT OF PREDICTION MODELS ................................................... 27
3.1. GRNN-based Prediction Models......................................................................... 34

3.2. SVM- based Prediction Models ..................................................................... 37
3.3. TB- based Prediction Models.......................................................................... 40
3.4. MLP- based Prediction Models....................................................................... 41
3.5. RBF- based Prediction Models ....................................................................... 43
3.6. SDT- based Prediction Models ....................................................................... 45
4. RESULTS AND DISCUSSION ............................................................................ 47
4.1. Results and Discussion for Prediction Models with Relief-F ......................... 47
4.2. Results and Discussion for Prediction Models with mRMR .......................... 61
4.3. Results and Discussion for Prediction Models with CFS ............................... 73
4.4. General Discussion .......................................................................................... 83
5. CONCLUSION ...................................................................................................... 85
REFERENCES.......................................................................................................... 87
CURRICULUM VITAE ........................................................................................... 95

**LIST OF TABLES** **PAGE**


Table 1.1. Statistical information about the dataset ................................................. 6
Table 3.1. Features' Relief-F scores regarding $UBP_{10}$ in UBP-set1 ......................... 27
Table 3.2. Features' Relief-F scores regarding $UBP_{60}$ in UBP-set1 ......................... 28
Table 3.3. Features' Relief-F scores regarding $UBP_{10}$ in UBP-set2 ......................... 28
Table 3.4. Features' Relief-F scores regarding $UBP_{60}$ in UBP-set2 ......................... 28
Table 3.5. mRMR scores of the features regarding $UBP_{10}$ in UBP-set1 ................. 29
Table 3.6. mRMR scores of the features regarding $UBP_{60}$ in UBP-set1 ................. 29
Table 3.7. mRMR scores of the features regarding $UBP_{10}$ in UBP-set2 ................. 29
Table 3.8. mRMR scores of the features regarding $UBP_{60}$ in UBP-set2................. 30
Table 3.9. Overview of $UBP_{10}$ models produced by Relief-F scores for UBP-set1 31
Table 3.10. Overview of $UBP_{60}$ models produced by Relief-F scores for UBP-set1 31
Table 3.11. Overview of $UBP_{10}$ models produced by Relief-F scores for UBP-set2 32
Table 3.12. Overview of $UBP_{60}$ models produced by Relief-F scores for UBP-set2 32
Table 3.13. Overview of $UBP_{10}$ models produced by mRMR scores for UBP-set1 . 32
Table 3.14. Overview of $UBP_{60}$ models produced by mRMR scores for UBP-set1 . 33
Table 3.15. Overview of $UBP_{10}$ models produced by mRMR scores for UBP-set2 . 33
Table 3.16. Overview of $UBP_{60}$ models produced by mRMR scores for UBP-set2 . 33
Table 3.17. Overview of $UBP_{10}$ and $UBP_{60}$ models produced by CFS for UBP-set1 34
Table 3.18. Overview of $UBP_{10}$ and $UBP_{60}$ models produced by CFS for UBP-set2 34
Table 3.19. Values of the utilized parameters for the GRNN models ....................... 35
Table 3.20. The intervals for values of the parameters for the SVM model.............. 38
Table 3.21. Values of the utilized parameters for the TB models............................. 40
Table 3.22. Values of the utilized parameters for the MLP models .......................... 41
Table 3.23. Values of the utilized parameters for the RBF models ........................... 43
Table 3.24. Values of the utilized parameters for the SDT models ........................... 45
Table 4.1. *SEE*'s for Relief-F-based $UBP_{10}$ prediction models (UBP-set1) ............ 47
Table 4.2. *SEE*'s for Relief-F-based $UBP_{60}$ prediction models (UBP-set1) ............ 48
Table 4.3. *R*'s for Relief-F-based $UBP_{10}$ prediction models (UBP-set1)................. 48
Table 4.4. *R*'s for Relief-F-based $UBP_{60}$ prediction models (UBP-set1)................. 49

Table 4.5. *MAPE*'s for Relief-F-based $UBP_{10}$ prediction models (UBP-set1) ........ 49
Table 4.6. *MAPE*'s for Relief-F-based $UBP_{60}$ prediction models (UBP-set1) ........ 50
Table 4.7. *SEE*'s for Relief-F-based $UBP_{10}$ prediction models (UBP-set2) ............ 54
Table 4.8. *SEE*'s for Relief-F-based $UBP_{60}$ prediction models (UBP-set2) ............ 54
Table 4.9. *R*'s for Relief-F-based $UBP_{10}$ prediction models (UBP-set2)................ 55
Table 4.10. *R*'s for Relief-F-based $UBP_{60}$ prediction models (UBP-set2)................. 55
Table 4.11. *MAPE*'s for Relief-F-based $UBP_{10}$ prediction models (UBP-set2) ....... 56
Table 4.12. *MAPE*'s for Relief-F-based $UBP_{60}$ prediction models (UBP-set2) ....... 56
Table 4.13. *SEE*'s for mRMR-based $UBP_{10}$ prediction models (UBP-set1) ............. 61
Table 4.14. *SEE*'s for mRMR-based $UBP_{60}$ prediction models (UBP-set1) ............. 61
Table 4.15. *R*'s for mRMR-based $UBP_{10}$ prediction models (UBP-set1).................. 62
Table 4.16. *R*'s for mRMR-based $UBP_{60}$ prediction models (UBP-set1).................. 62
Table 4.17. *MAPE*'s for mRMR-based $UBP_{10}$ prediction models (UBP-set1).......... 63
Table 4.18. *MAPE*'s for mRMR-based $UBP_{60}$ prediction models (UBP-set1).......... 63
Table 4.19. *SEE*'s for mRMR-based $UBP_{10}$ prediction models (UBP-set2) ............. 68
Table 4.20. *SEE*'s for mRMR-based $UBP_{60}$ prediction models (UBP-set2) ............. 68
Table 4.21. *R*'s for mRMR-based $UBP_{10}$ prediction models (UBP-set2).................. 69
Table 4.22. *R*'s for mRMR-based $UBP_{60}$ prediction models (UBP-set2).................. 69
Table 4.23. *MAPE*'s for mRMR-based $UBP_{10}$ prediction models (UBP-set2).......... 70
Table 4.24. *MAPE*'s for mRMR-based $UBP_{60}$ prediction models (UBP-set2).......... 70
Table 4.25. *SEE*'s for CFS-based $UBP_{10}$ prediction models (UBP-set1) .................. 73
Table 4.26. *SEE*'s for CFS-based $UBP_{60}$ prediction models (UBP-set1) .................. 73
Table 4.27. *R*'s for CFS-based $UBP_{10}$ prediction models (UBP-set1)....................... 74
Table 4.28. *R*'s for CFS-based $UBP_{60}$ prediction models (UBP-set1)....................... 74
Table 4.29. *MAPE*'s for CFS-based $UBP_{10}$ prediction models (UBP-set1) .............. 74
Table 4.30. *MAPE*'s for CFS-based $UBP_{60}$ prediction models (UBP-set1) .............. 74
Table 4.31. *SEE*'s for CFS-based $UBP_{10}$ prediction models (UBP-set2) .................. 78
Table 4.32. *SEE*'s for CFS-based $UBP_{60}$ prediction models (UBP-set2) .................. 78
Table 4.33. *R*'s for CFS-based $UBP_{10}$ prediction models (UBP-set2)....................... 79
Table 4.34. *R*'s for CFS-based $UBP_{60}$ prediction models (UBP-set2)....................... 79
Table 4.35. *MAPE*'s for CFS-based $UBP_{10}$ prediction models (UBP-set2) .............. 79

Table 4.36. *MAPE*'s for CFS-based $UBP_{60}$ prediction models (UBP-set2) .............. 79

**LIST OF FIGURES** **PAGE**


Figure 1.1. The example for classic style of cross-country skiing ............................ 1
Figure 1.2. The example for skate skiing of cross-country skiing ............................ 2
Figure 1.3. The laboratory of measuring UBP ........................................................ 3
Figure 2.1. A typical GRNN architecture ............................................................. 13
Figure 2.2. A typical MLP structure. ..................................................................... 20
Figure 2.3. Pseudo code of Relief-F algorithm ...................................................... 23
Figure 3.1. Flow chart of GRNN-based model ...................................................... 36
Figure 3.2. Flow chart of SVM-based model ......................................................... 39
Figure 3.3. Flow chart of RBF-based model .......................................................... 44
Figure 4.1. Percentage *decrease* rates in *SEE* values of $UBP_{10}$ for GRNN-based models compared to *SEE* values obtained by SVM, MLP, TB, RBF and SDT-based models using Relief-F on UBP-set1 ............................ 51
Figure 4.2. Percentage decrease *rates* in *SEE* values of $UBP_{60}$ for GRNN-based models compared to *SEE* values obtained by SVM, TB, MLP, RBF and SDT-based models using Relief-F on UBP-set1 ............................ 52
Figure 4.3. Average *SEE*'s of *prediction* models for $UBP_{10}$ using Relief-F on UBP-set1 ........................................................................................ 52
Figure 4.4. Average *SEE*'s of *prediction* models for $UBP_{60}$ using Relief-F on UBP-set1 ........................................................................................ 53
Figure 4.5. Percentage decrease rates in *SEE* values of $UBP_{10}$ for GRNN-based models compared to *SEE* values obtained by SVM, MLP, TB RBF and SDT-based models using Relief-F on UBP-set2 ............................ 58
Figure 4.6. Percentage decrease rates in *SEE* values of $UBP_{60}$ for GRNN-based models compared to *SEE* values obtained by SVM, TB, MLP RBF and SDT-based models using Relief-F on UBP-set2 ............................ 58
Figure 4.7. Average *SEE*'s of prediction models for $UBP_{10}$ using Relief-F on UBP-set2 ........................................................................................ 59
Figure 4.8. Average *SEE*'s of prediction models for $UBP_{60}$ using Relief-F on UBP-set2 ........................................................................................ 59

Figure 4.9. Percentage decrease rates in *SEE* values of $UBP_{10}$ for GRNN-based models compared to *SEE* values obtained by SVM, MLP, TB, RBF and SDT-based models using mRMR on UBP-set1 ............................ 65
Figure 4.10. Percentage decrease rates in *SEE* values of $UBP_{60}$ for GRNN-based models compared to *SEE* values obtained by SVM, TB, MLP, RBF and SDT-based models using mRMR on UBP-set1 ............................ 65
Figure 4.11. Average *SEE*‘s of prediction models for $UBP_{10}$ using mRMR on UBP-set1 ............................................................................ 66
Figure 4.12. Average *SEE*‘s of prediction models for $UBP_{60}$ using mRMR on UBP-set1 ............................................................................ 66
Figure 4.13. Percentage decrease rates in *SEE* values of $UBP_{10}$ for GRNN-based models compared to *SEE* values obtained by SVM, MLP, TB, RBF and SDT-based models using mRMR on UBP-set2 ............................ 71
Figure 4.15. Average *SEE*‘s of prediction models for $UBP_{10}$ using mRMR on UBP-set2 ............................................................................ 72
Figure 4.16. Average *SEE*‘s of prediction models for $UBP_{60}$ using mRMR on UBP-set2 ............................................................................ 72
Figure 4.17. Percentage decrease rates in *SEE* values of $UBP_{10}$ GRNN-based models compared to *SEE* values obtained by for SVM, MLP, TB, RBF and SDT-based models using CFS on UBP-set1 ......................... 76
Figure 4.18. Percentage decrease rates in *SEE* values of $UBP_{60}$ for GRNN-based models compared to *SEE* values obtained by SVM, TB, MLP, RBF and SDT-based models using CFS on UBP-set1 ................................. 76
Figure 4.19. Average *SEE*‘s of prediction models for $UBP_{10}$ using CFS on UBP-set1 ............................................................................ 77
Figure 4.20. Average *SEE*‘s of prediction models for $UBP_{60}$ using CFS on UBP-set1 ............................................................................ 77
Figure 4.21. Percentage decrease rates in *SEE* values of $UBP_{10}$ for GRNN-based models compared to *SEE* values obtained by SVM, MLP, TB, RBF and SDT-based models using CFS on UBP-set2 ................................. 81

Figure 4.22. Percentage decrease rates in *SEE* values of $UBP_{60}$ for GRNN-based models compared to *SEE* values obtained by SVM, TB, MLP, RBF and SDT-based models using CFS on UBP-set2 ................................ 81
Figure 4.23. Average *SEE*'s of prediction models for $UBP_{10}$ using CFS on UBP-set2 ................................ 82
Figure 4.24. Average *SEE*'s of prediction models for $UBP_{60}$ using CFS on UBP-set2 ................................ 82

## LIST OF ABBREVIATIONS

| | | |
|---|---|---|
| BMI | : | Body Mass Index |
| CCN | : | Cascade Correlation Network |
| CFS | : | Correlation-Based Feature Selection |
| DTF | : | Decision Tree Forest |
| ERM | : | Empirical Risk Minimization |
| GEP | : | Gene Expression Programming |
| GRNN | : | Generalized Regression Neural Networks |
| HR | : | Heart Rate |
| HRLT | : | Heart Rate at Lactate Threshold |
| HRmax | : | Maximal Heart Rate |
| MAPE | : | Mean Absolute Percentage Error |
| MFANN | : | Multi-layer Feed-forward artificial Neural Network |
| MLP | : | Multilayer Perceptron |
| MLR | : | Multi Linear Regression |
| MRMR | : | Minimum Redundancy Maximum Relevance |
| MSU | : | Montana State University |
| R | : | Multiple Correlation Coefficient |
| RBF | : | Radial Basis Function |
| RS | : | Race Speed |
| $RVO_2max$ | : | Relative Maximum Oxygen Uptake |
| SDT | : | Single Decision Tree |
| SEE | : | Standard Error of Estimation |
| SRM | : | Structural Risk Minimization |
| SVM | : | Support Vector Machine |
| SVM-Linear | : | Support Vector Machine with the linear kernel function |
| SVM-Poly | : | Support Vector Machine with the polynomial kernel function |

| | | |
|---|---|---|
| SVM-RBF | : | Support Vector Machine with the radial basis kernel function |
| SVM- Sigmoid | : | Support Vector Machine with the sigmoid kernel function |
| TB | : | Tree Boost |
| UBP | : | Upper Body Power |
| $UBP_{10}$ | : | 10-second Upper Body Power |
| $UBP_{60}$ | : | 60-second Upper Body Power |
| $UBP_{peak}$ | : | Peak Power Output |
| UBP-set1 | : | Upper Body Power – Data Set 1 |
| UBP-set2 | : | Upper Body Power – Data Set 2 |
| $VO_2max$ | : | Maximum Oxygen Uptake |
| $VO_{2peak}$ | : | Peak Oxygen Uptake |
| W | : | Watt |

# 1. INTRODUCTION

## 1.1. Overview of Cross-Country Skiing and Upper Body Power

In cross-country skiing, skiers utilize their own condition to move across snow-covered terrain, instead of utilizing ski lifts or other kind of special equipment. Aided by arms pushing on ski poles against the snow, skiers propel themselves either by striding forward (classic style) or side-to-side in a skating motion (skate skiing) (Hindman, 2005). Figure 1 and Figure 2 show typical examples of classic style and skate skiing, respectively.


Figure 1.1. The example for classic style of cross-country skiing

Cross-country skiing is a very strenuous sport because the cross-country skiers intensively use all of upper and lower body musculature (Heil and Camenisc, 2014). Regardless of the style of skiing, the most important component that determines the performance of a cross-country skier in a race is his/her ability to utilize the upper body power (UBP) (Marsland et al., 2012).

Figure 1.2. The example for skate skiing of cross-country skiing

UBP is the rate at which work can be performed using the arm, shoulder, and trunk muscles. Power generated by the upper body during cross-country skiing is transmitted through the poles and assists in forward motion. For instance, the upper body has been appeared to contribute as much as half to the aggregate propelling force during uphill skating and 15% to 30% during uphill classical skiing (Alsobrook and Heil, 2009b).

A custom-built ergometer is used to simulate a skier's double-poling motion on snow and measure UBP. Double-poling has recently become a popular technique due to the improved preparation of tracks and technological advances in cross-country ski equipment (Heil and Camenisc, 2014). Figure 1.3 shows an experimental setup to measure UBP.

Although measurement using experimental setup is the direct and most accurate method to determine UBP, this method involves several limitations. First of all, the equipment for measuring UBP is not yet usually accessible due to its high cost. The tests of UBP have all been based upon custom-designed ergometers which can only be found in specific sports research laboratories (Alsobrook and Heil, 2009a). In addition, the measurement of UBP hasn't got standardized as it is still a relatively new physiological construct. Also, measuring UBP is a time-consuming process and it requires the presence of a qualified and experienced staff.

Figure 1.3. The laboratory of measuring UBP

## 1.2. Motivation, Purpose and Contributions of the Thesis

As mentioned in Section 1.1, determination of UBP through measurements comes with several drawbacks. On the other hand, standard sports science and research laboratories have been used by many cross-country skiers and most of these laboratories are capable of measuring maximum oxygen uptake ($VO_2$max), maximal heart rate (HRmax), heart rate at lactate threshold (HRLT) and several other physiological parameters (height, weight, BMI etc.) which are related to UBP. As a result, it may be advantageous to predict rather than measure UBP using actual training data previously collected by tests on ergometers specifically designed for UBP measurements.

Daneshvar et al. (2013) was the first study in literature that proposed to predict UBP of cross-country skiers by utilizing promising machine learning methods. Although, since then, the problem of predicting UBP of cross-country skiers has been studied in details, the identification of relevant and irrelevant features affecting UBP is still an open area to research.

The objective of the current PhD thesis is to construct new prediction models and determine the discriminative features for prediction of 10-second UBP ($UBP_{10}$) and 60-second UBP ($UBP_{60}$) by using General Regression Neural Networks (GRNN), Radial-Basis Function Network (RBF), Multilayer Perceptron (MLP), Support Vector Machine (SVM), Single Decision Tree (SDT) and Tree Boost (TB) combined with several feature selection algorithms. The feature selection algorithms include Relief-F, minimum redundancy maximum relevance (mRMR) and the Correlation-based Feature Subset Selection (CFS). Two versions of the same dataset have been utilized in order to develop UBP prediction models. The first version of the dataset (UBP-set1) involves the features protocol, gender, age, weight, height, $VO_2$max, HRmax, exercise time and HRLT. In the second version of the dataset (UBP-set2), height and weight are replaced with BMI keeping the rest of features as they are in UBP-set1. By using Relief-F and mRMR, the ranking of the features is obtained. The least effective feature is removed at a time until the model includes a single feature. With CFS, a subset of the important features has been obtained.

Several models have been developed to predict $UBP_{10}$ and $UBP_{60}$ using the mentioned regression methods. 10-fold cross validation has been performed for model testing. Performance metrics including *R*, *SEE* and *MAPE* have been utilized to measure the efficiency of the prediction models.

To the best of our knowledge, this is the first comprehensive study in literature to determine relevant and irrelevant features and then predict $UBP_{10}$ and $UBP_{60}$. The results show that utilizing smaller number of features instead of the full set of features yields the most accurate predictions for $UBP_{10}$ and $UBP_{60}$. The models involving height and weight in place of BMI give lower error rates for the prediction of $UBP_{10}$ and $UBP_{60}$ regardless of which regression methods have been used. Also, GRNN-based models always outperform the other models for predicting $UBP_{10}$ and $UBP_{60}$.

### 1.3. Overview of the Dataset

The dataset was obtained from the Montana State University. The dataset includes data of cross-country skiers ranging in age from 15 to 25 years. UBP-set1 includes 75 (38 females and 37 males) healthy skiers. UBP-set1 includes Protocol, Gender, Age, Weight, Height, $VO_2$max, HRmax, Time, HRLT and also $UBP_{10}$, $UBP_{60}$ which are target variables. UBP-set2 includes the same subjects. In UBP-set2, height and weight were replaced with BMI.

Brief statistical measures (minimum, maximum, median, mean and standard deviation) of the dataset is given in Table 1.1.

Table 1.1. Statistical information about the dataset

| **Feature** | **Minimum** | **Maximum** | **Median** | **Mean** | **Standard Deviation** |
|---|---|---|---|---|---|
| Protocol | 0.00 | 1.00 | 1.00 | 0.77 | 0.43 |
| Gender (0: F, 1: M) | 0.00 | 1.00 | 0.00 | 0.49 | 0.50 |
| Age | 15.00 | 25.00 | 18.00 | 18.52 | 2.29 |
| BMI (kg/m2) | 17.83 | 27.90 | 21.73 | 22.08 | 1.94 |
| Weight (kg) | 48.30 | 87.00 | 67.00 | 67.09 | 6.97 |
| Height (cm) | 158.30 | 196.60 | 173.00 | 174.40 | 8.04 |
| $VO_2max$ ($ml.kg^{-1}.min^{-1}$) | 46.00 | 79.01 | 61.70 | 62.33 | 8.38 |
| HRmax (bpm) | 180.00 | 213.00 | 197.00 | 196.70 | 7.34 |
| Time (s) | 5.10 | 13.70 | 12.00 | 11.55 | 1.62 |
| HRLT (bpm) | 161.00 | 210.00 | 182.00 | 181.00 | 9.92 |
| $UBP_{10}$ (W) | 110.00 | 350.00 | 207.00 | 225.40 | 71.08 |
| $UBP_{60}$ (W) | 92.00 | 285.00 | 170.00 | 172.40 | 54.03 |

### 1.4. Literature Review

#### 1.4.1. Measurement of UBP

For two different groups of skiers, Alsobrook (2005) investigated the relationship between UBP and classical cross-country ski race performance. The first group included 7 males and 3 females, whereas the second group included 10 males and 5 females. The results show that classical race performance is closely related to both long and short term UBP and that short and long term UBP's are affected by the same factors.

Alsobrok and Heil (2009a) studied the relationship between short and long duration measures of UBP and mass start classical cross-country ski performance. Three different tests of UBP were completed by a group of skiers on a double poling ergometer. The authors concluded that a strong correlation exists between short and long duration UBP tests for mass start classical ski race performance.

Heil et al. (2004) evaluated the influence of the three ski pole grip systems in common use by competitive cross-country skiers on UBP. Nine men who experienced about cross-country ski racing in these systems performed three successive UBP tests on a modified double-poling ergometer. All participants finished 15-s tests of UBP using stiff cross-country ski poles and a resistance corresponding to 3% of body mass. During the last 10-seconds for every test, the highest 5-seconds average power output was assessed as Peak UBP. Peak UBP with the integrated system yielded higher value than the ones produced by both the modern and traditional systems.

### 1.4.2. Prediction of UBP without Feature Selection

Daneshvar et al. (2013) predicted $UBP_{10}$ using a multi-layer feed-forward artificial neural network (MFANN). Twelve prediction models were developed by using combinations of different features. This study showed that the model including the features age, gender, body mass and maximum heart rate yielded the lowest *SEE*'s (24.74 W) whereas the model including the variables age, gender, body mass and time yielded the highest *SEE*'s (33.06 W) for prediction of UBP.

Sanli et al. (2014) investigated the effects of $VO_2$max and HRmax on $UBP_{10}$ and $UBP_{60}$ using SVM and compared the results with those obtained by multiple linear regression (MLR). Two groups of common features combined with $VO_2$max and HRmax were used to create eight different models. First common group included protocol, gender, age, height, weight, exercise time and second common group was comprised of gender, age, BMI, exercise time. The results showed that the combination of the common variables from both groups with $VO_2$max comparatively yielded the lowest *SEE*'s, while their combination with HRmax gave the highest *SEE*'s. In more details, the model in the first group including the features protocol, age, gender, weight, height, exercise time and $VO_2$max yielded the lowest *SEE*'s and similarly, the model in the second group including age, gender, exercise time, $VO_2$max and BMI yields the lowest *SEE*'s for UBP prediction.

Isoglu et al. (2014) aimed to predict $UBP_{10}$ and $UBP_{60}$ using SVM and MLR. The data set which included features (gender, age, body mass, height, HRmax, relative maximum oxygen uptake ($RVO_2$max), exercise time) and target variables ($UBP_{10}$ and $UBP_{60}$) consisted of 57 healthy volunteers. Seven different models had been developed to predict $UBP_{10}$ and $UBP_{60}$. The performance of SVM models were much better than the ones obtained by MLR for prediction of $UBP_{10}$ and $UBP_{60}$. Also, the performance of $UBP_{60}$ prediction models was worse than the performance of $UBP_{10}$ prediction models.

Akgol et al. (2015a) used TB, Gene Expression Programming (GEP) and SDT for prediction of $UBP_{10}$ and $UBP_{60}$. By using the combination of the variables time, $VO_2$max and HRmax with protocol, age, gender, weight and height seven different prediction models were produced. As a result, TB-based models, in which $VO_2$max instead of time were used with age, protocol, gender, height and weight performed better than the models developed by the other methods while GEP-based models were more accurate than SDT-based models.

Akay et al. (2015a) developed prediction models for estimation of $UBP_{10}$ and $UBP_{60}$ using SVM, MLP and MLR. The first dataset consisted of gender, age, weight, height, $VO_2$max, HRmax and exercise time. The second dataset included age, gender, BMI, HRmax, $VO_2$max, exercise time and also, both datasets consisted of the target variables $UBP_{10}$ and $UBP_{60}$. SVM-RBF-based models yielded the lowest *SEE*'s and the highest *R's* for prediction of $UBP_{10}$ and $UBP_{60}$. On the other hand, MLR-based models yielded higher *SEE's* but produced faster results for prediction. In addition, adding BMI to the models significantly increased the performance of models for prediction of $UBP_{10}$ and $UBP_{60}$. Error rates related to the prediction of $UBP_{60}$ were always lower than that of $UBP_{10}$ due to high standard deviation of $UBP_{10}$.

Daneshvar et al. (2015) used Cascade Correlation Network (CCN), Decision Tree Forest (DTF) and RBF to create new models for prediction of $UBP_{10}$ and $UBP_{60}$. The dataset consisted of the features gender, age, BMI, heart rate (HR), exercise time, $VO_2$max and the target variables were $UBP_{10}$ and $UBP_{60}$. 14 different models had been created by using these features to predict $UBP_{10}$ and $UBP_{60}$. According to the results, the model consisting of the features gender, age, BMI and $VO_2$max gave the highest *SEE*'s for both predictions of $UBP_{10}$ and $UBP_{60,}$ regardless of whether CCN, DTF or RBF had been used. Also, CCN-based models consistently showed the best performance in terms of the achieved *SEE*'s and *R*'s; whereas SDT-based models outperform DTF-based models.

### 1.4.3. Prediction of UBP with Feature Selection

Akgol et al. (2014) used SVM and MLR combined with feature selection to develop several different models for prediction $UBP_{10}$ and $UBP_{60}$. The dataset included age, gender, height, weight, exercise time, HRmax and $RVO_2max$. The ranks of each variable had been calculated by the Relief-F feature selection algorithm combined with a ranker search method. The ranking (from highest to the lowest score) of the features for $UBP_{10}$ is weight, height, age, gender, exercise time, $RVO_2max$, HRmax and the ranking of the features for $UBP_{60}$ is weight, height, age, $RVO_2max$, gender, HRmax and exercise time. The models were created by eliminating one variable at a time having the lowest rank score. According to the results, SVM models were better than MLR models for prediction of $UBP_{10}$ and $UBP_{60}$. The model which included weight, height, age, $RVO_2max$ and gender yielded the lowest *SEE*'s and the highest *R*'s for prediction of $UBP_{60,}$ independent of which regression methods had been used. Additionally, the model comprising of weight, height, age, gender, exercise time and $RVO_2max$ gave the lowest *SEE*'s and the highest *R*'s for prediction of $UBP_{10}$, regardless of whether SVM or MLR had been used.

Ozciloglu et al. (2015a) built new feature selection based models for prediction of $UBP_{10}$ and $UBP_{60}$ using CCN, MLP and SDT combined with feature selection algorithm. The dataset consisted of the features age, gender, exercise time, BMI, $HR_{max}$, $VO_2max$ and the target variables $UBP_{10}$ and $UBP_{60}$. Relief-F feature selection algorithm was used to calculate the ranks of features. According to these calculations, the features exercise time and BMI had been evaluated with the highest score for $UBP_{10}$ and $UBP_{60}$, respectively. Based on these ranking scores of features, twelve different models had been developed for prediction of $UBP_{10}$ and $UBP_{60}$. The model involving age, gender, $VO_2max$, BMI and exercise time gave the lowest *SEE*'s and the highest *R*'s for prediction of $UBP_{10}$, on the other hand for prediction of $UBP_{60,}$ the model comprised of age, gender, $VO_2max$ and BMI yielded the lowest *SEE*'s and the highest *R*'s, regardless of which regression methods had been used.

CCN-based prediction models yielded better results than the ones obtained by MLP and SDT regression methods.

Akgol et al. (2015b) used GRNN, RBF, DTF regression methods combined with a feature selection algorithm to produce prediction models for predicting $UBP_{10}$ and $UBP_{60}$ of cross-country skiers. The rank of each feature has been calculated by using the Relief-F feature selection algorithm. The dataset was produced by 57 skiers whose ages between 16 and 24. Totally, fourteen different prediction models had been created by removing the feature with the lowest rank at a time. The performance of the prediction models had been evaluated by using 10-fold cross validation. The results show that gender and $VO_2max$ were the most effective variables for prediction of $UBP_{10}$ and $UBP_{60}$. GRNN-based prediction models performed better than the ones obtained by other methods. On the other hand, RBF-based prediction models were more accurate than DTF-based models. In particular, GRNN-based models yielded in average 9.69% and 24.17% lower *SEE*'s than the *SEE*'s of RBF and DTF-based models for $UBP_{10}$ and $UBP_{60}$. Also, $UBP_{60}$ prediction models gave lower *SEE*'s than the ones of the models for $UBP_{10}$ due to high standard deviation of $UBP_{10}$.

Akay et al. (2015b) aimed to develop new prediction models to find out the relevant predictors of $UBP_{10}$ and $UBP_{60}$. $UBP_{10}$ and $UBP_{60}$ were predicted using SVM, MLP and MLR combined with Relief-F. The results showed that UBP was highly correlated with age and gender, therefore the prediction models including these features gave much lower *SEE*'s and higher *R*'s than other models. In addition, the SVM-RBF-based model comprising of the features age, weight, gender, exercise time and height gave the lowest *SEE*'s (27.81 W) and highest *R*'s (0.92) for prediction of $UBP_{10}$. The model including the features age, gender, weight, height and $VO_2max$ yielded the lowest *SEE*'s (18.73 W) and highest *R*'s (0.94) for prediction of $UBP_{60}$.

Ozciloglu et al. (2015b) developed new prediction models for predicting $UBP_{10}$ and $UBP_{60}$ by using GRNN, SDT and Tree Boost along with the mRMR feature selection algorithm. 18 different models had been developed for prediction of $UBP_{10}$ and $UBP_{60}$ by the ranking scores of features assigned with the mRMR. The results showed that the model including the features gender, weight, height, $VO_2$max, time, age and HRmax yielded the most reduced *SEE*'s for prediction of $UBP_{10}$, while the model including the features gender, age, weight, $VO_2$max, HRLT, height and HRmax led to the most reduced *SEE*'s for prediction of $UBP_{60}$. Using these two models instead of the full set of features yielded up to 8.80%, 5.20% and 8.26% decrement rates on the average in *SEE*'s for GRNN, Tree Boost and SDT, respectively. Additionally, the feature time provided a significant improvement for prediction of $UBP_{10}$. A same case for $UBP_{60}$, the feature $VO_2$max improved significantly about prediction of $UBP_{60}$.

Ozciloglu et al. (2016a) determined the effect of features and developed new prediction models on UBP by using different tree based methods including DTF, SDT and TB combined with the Relief-F feature selection algorithm. 16 prediction models had been developed by using Relief-F which calculated the ranks of features. The results emphasized that the features gender, age, BMI and time play an important role for predicting $UBP_{10}$ and $UBP_{60}$, regardless of which tree-based method has been used. As opposed to the model included full set of features; the model comprising of only 3 features (gender, BMI and time) yielded 5.68%, 3.12% and 7.41% lower *SEE*'s for TB, DTF and SDT-based $UBP_{10}$ prediction models, respectively. As for $UBP_{60}$ prediction models, the model consisted of less number of features (gender, age and BMI) instead of full set of features led to 6.48%, 12.94% and 6.83% lower *SEE*'s for TB, DTF and SDT, respectively. So, the feature BMI and age significantly improved the performance of models for prediction of $UBP_{10}$ and $UBP_{60}$. Additionally, $UBP_{60}$ prediction models gave lower *SEE*'s than the ones of the models for $UBP_{10}$ due to high standard deviation of $UBP_{10}$.

In Ozciloglu et al. (2016b), MLP and SDT combined with mRMR feature selection algorithm were used to produce sixteen prediction models for predicting $UBP_{10}$ and $UBP_{60}$. The results showed that MLP-based models gave in average 16.00% and 14.49% lower *SEE's* than the ones given by the SDT-based models, for predicting $UBP_{10}$ and $UBP_{60}$, respectively. The prediction model consisting of gender, BMI, $VO_2$max, HRLT and time led to the lowest *SEE*'s and highest *R's* for prediction of $UBP_{10}$, while the model comprising of gender, age and BMI yielded the lowest *SEE*'s and highest *R's* for prediction of $UBP_{60}$, regardless of which artificial intelligence methods had been used. Applying these two models instead of the full set of features yielded up to 4.95% and 6.83% decrement rates in *SEE*'s for MLP and SDT based UBP prediction models, respectively.

# 2. OVERVIEW OF METHODS

## 2.1. Overview of Regression Methods

### 2.1.1. Generalized Regression Neural Network

A GRNN includes four layers, the names of which are given in Figure 2.1. A typical GRNN architecture is illustrated in Fig. 2.1.

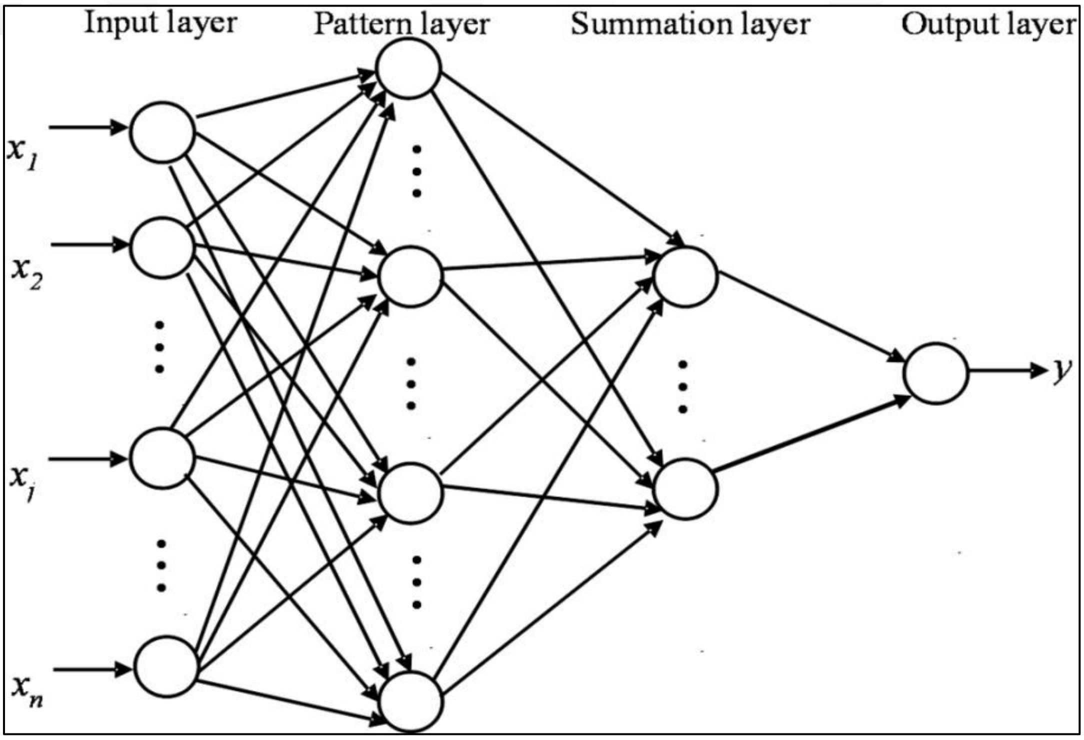


Figure 2.1. A typical GRNN architecture

Total number of features specifies the number of input structures. As shown in Figure 2.1, each layer is directly connected to next layer without having any feedback loops. The summation layer is comprised of a single division unit and summation units. A normalization of the output set is performed by both the summation and output layer.

GRNN can be considered as a technique to predict the joint pdf of *x* and *y*, when only a training set is given. Let *f (x, y)* represent the joint pdf of a vector random variable, *x* and a scalar random variable, *y*. In this case, the conditional expected value of *y* given X, is calculated by (2.1.).

$$E[y|X] = \frac{\int_{-\infty}^{\infty} y f(X,y) dy}{\int_{-\infty}^{\infty} f(X,y) dy} \tag{2.1.}$$

In case *f (x, y)* is not known, a sample of observations of x and *y* has to be used to predict *f (x, y)*. Let *f'(X, Y)* (given by (2.2.)) be the probability estimator. *f'(X, Y)* includes sample values $X^i$ and $Y^i$ of the random variables x and *y*. In (2.2.), *n* is the number of sample observations, $\sigma$ is the width and *p* is the dimension of the vector variable x.

$$f'(X,Y) = \frac{1}{(2\pi)^{(p+1)/2}\sigma^{(+1)}} \frac{1}{n} \sum_{i=1}^{n} \exp[-\frac{(X-X^i)^T(X-X^i)}{2\sigma^2}] \exp[-\frac{(Y-Y^i)^2}{2\sigma^2}] \tag{2.2.}$$

The scalar function $D_i^2$ *is* given by (2.3.) and (2.4.) is obtained after carrying out the necessary integrations.

$$D_i^2 = (X - X^i)^T (X - X^i) \tag{2.3.}$$

$$Y'(X) = \frac{\sum_{i=1}^{n} Y^i \exp(-\frac{D_i^2}{2\sigma^2})}{\sum_{i=1}^{n} EXP(-\frac{D_i^2}{2\sigma^2})} \tag{2.4.}$$

It should be noted that the final equation given by (2.4.) is applicable to problems involving numerical data.

### 2.1.2. Support Vector Machines

SVM is one of the most promising statistical learning methods for classification and regression. (Vapnik, 1995) developed the principles of SVM. Since then, SVM have gained enormous popularity owing to the fact that SVM having many effective features, and providing good performance. Instead of the ERM principle, which is used by artificial neural networks, SRM has been adopted in SVM (Gunn et al., 1997). The error in the training set is minimized in the ERM, whereas the expected risk is minimized in SRM. The generalization ability, which is one of the main aims of statistical learning, of SVM comes from this difference. Although SVM were originally designed to solve classification problems, they have been lately restructured to solve regression problems (Vapnik et al., 1997).

#### A. Linear SVM

We are given the training data $(x_i, y_i), (i = 1, ..., \ell)$, where $x$ is a $d$-dimensional input vector with $x \in \Re^d$ and the output vector is $y \in \Re$ . (2.5.) shows the linear regression model (Vapnik, 2000):

$$f(x) = \langle \omega, x \rangle + b, \qquad \omega, x \in \Re^d, \ b \in \Re, \tag{2.5.}$$

In (2.5.), the target function is represented by $f(x)$ and $\langle .,. \rangle$ gives the dot product in $\Re^d$ .

To measure the empirical risk, some sort of loss function definition is required. The ε-insensitive loss function, which is proposed by Vapnik (Vapnik, 2000), is the most frequently used function. (2.6.) defines the ε-insensitive loss function:

$$L_\varepsilon(y) = \begin{cases} 0 & \text{for } |f(x) - y| \le \varepsilon \\ |f(x) - y| - \varepsilon & \text{otherwise} \end{cases} \tag{2.6.}$$

The optimization problem given in (2.7.) (Gunn, 1998) should be solved to find out the optimal $\bar{\omega}$ and $\bar{b}$ values

$$\min \ \frac{1}{2}\|\omega\|^2 + C\sum_{i=1}^{\ell}(\xi_i^- + \xi_i^+) \tag{2.7.}$$

with constraints:

$$\begin{aligned} &\mathrm{y_i} - \langle \omega, x_i \rangle - b \le \varepsilon + \xi_i^+, \\ &\langle \omega, x_i \rangle + b - \mathrm{y_i} \le \varepsilon + \xi_i^+, \\ &\xi_i^+, \xi_i^- \ge 0, \quad \mathrm{i}=1,........,\ell \end{aligned} \tag{2.8.}$$

In (2.8.), there exists toleration for the deviations larger than $\varepsilon$ and the tradeoff between the flatness of $f(x)$ is determined by *C*. The deviations from the $\varepsilon$-tube are represented by the variables $\xi^-$ and $\xi^+$.

The dual optimization problem given in (2.9.)

$$\max_{\alpha,\alpha^*} -\frac{1}{2}\sum_{i=1}^{\ell}\sum_{j=1}^{\ell}(\alpha_i^* - \alpha_i)(\alpha_j^* - \alpha_j)\langle x_i, x_j \rangle - \sum_{i=1}^{\ell} y_i(\alpha_i^* - \alpha_i) - \varepsilon\sum_{i=1}^{\ell}(\alpha_i^* + \alpha_i) \tag{2.9.}$$

with constraints:

$$\begin{aligned} &0 \le \alpha_i, \alpha_i^* \le C, \quad i = 1,...,\ell\ , \\ &\sum_{\mathrm{i}=1}^{\ell}(\alpha_i - \alpha_i^*) = 0 \end{aligned} \tag{2.10.}$$

has to be solved, which in turn gives the optimum values of the Lagrange multipliers $\alpha$ and $\alpha^*$, while $\bar{\omega}$ and $\bar{b}$ are given by

$$\begin{aligned} &\bar{\omega} = \sum_{i=1}^{\ell}(\alpha_i^* - \alpha_i)x_i \quad , \\ &\bar{b} = -\frac{1}{2}\left\langle \bar{\omega}, (x_r + x_s) \right\rangle , \end{aligned} \tag{2.11.}$$

where $x_r$ and $x_s$ are support vectors (Gunn, 1998).

**B. Nonlinear SVM**

Nonlinear SVM can be constructed using a nonlinear mapping $\phi$ of the input space onto a higher dimension feature space. (2.12.) shows the nonlinear regression model

$$f(x) = \langle \omega, \phi(x) \rangle + b, \quad \omega, x \in \Re^{d}\ ,\ b \in \Re\ , \tag{2.12.}$$

Where

$$\begin{aligned} \overline{\omega} &= \sum_{i=1}^{\ell} (\alpha_i - \alpha_i^*) \phi(x_i), \\ \left\langle \omega, \overline{\phi}(x) \right\rangle &= \sum_{i=1}^{\ell} (\alpha_i - \alpha_i^*) \langle \phi(x_i), \phi(x) \rangle = \sum_{i=1}^{\ell} (\alpha_i - \alpha_i^*) K(x_i, x), \\ \overline{b} &= -\frac{1}{2} \sum_{i=1}^{\ell} (\alpha_i - \alpha_i^*)(K(x_i, x_r) + K(x_i, x_s)) \end{aligned} \tag{2.13.}$$

the support vectors are represented by $x_r$ and $x_s$ . A Kernel function $K$ satisfying Mercer's conditions has been used to explain dot products (Vapnik, 2000).

After $\overline{b}$ is integrated into the kernel function, (2.12.) becomes:

$$\sum_{i=1}^{\ell} (\alpha_i - \alpha_i^*) K(x_i, x) \tag{2.14.}$$

Many different kernel functions including the radial basis function (RBF), the polynomial function and the sigmoid function exist. However; RBF, as defined in (2.15.) is the most commonly used Kernel function:

$$K(x, x') = \exp\left(-\frac{\|x - x'\|^2}{2\rho^2}\right). \tag{2.15.}$$

In (2.15.) the width of the RBF function is defined by $\rho$ .

### 2.1.3. Tree Boost

Tree Boost is some sort of a machine learning algorithm used for the solution of both regression and classification problems, which creates an estimation model that has the shape of a group of low estimation models. Boosting Tree algorithm is calculated by using boosting methods for regression trees. The common idea is to evaluate an order of simple trees, where every progressive tree is constructed for the estimation leftover of the former tree. This method will construct binary trees. It is assumed that the complexities of the trees are limited with only 3 nodes. These are two child nodes and a root node. Therefore, at each step of the boosting, a basic dividing of the data is specified, and the deviations of the noted values from the individual means are calculated. The following 3-node tree will be fitted to those residuals, in order to find alternate partition that will further decrease the residual variance for the data, given the former a group of trees (Hastie et al., 2009).

A boosted classifier (or regressor) has the form

$$H(x) = \sum_t \alpha_t h_t(x) \tag{2.16.}$$

This form can be trained by optimizing scalar $\alpha_t$ and weak learner $h_t(x)$ at each iteration *t*. Before training begins, each data sample $x_i$ is assigned a non-negative weight $w_i$ (which is derived from L). After each iteration, misclassified samples are weighted more heavily thereby increasing the intensity of misclassifying them in next iterations. Regardless of the type of Boosting used (i.e. AdaBoost, LogitBoost, L2Boost, etc.), each iteration requires training a new weak learner given the sample weights (Appel et al., 2013).

### 2.1.4. Multilayer Perceptron

Several problems in different fields including control, function approximation, pattern recognition and associative memory (Haykin, 1994) can be solved by the biologically inspired intelligent systems called Artificial Neural Networks (ANN’s). ANN’s are comprised of many simple processors interconnected by many interconnection networks. The use of ANN’s provides the following advantages: learning ability, generalization ability, fault tolerance, input–output mapping, low energy consumption, nonlinearity, distributed representation and computation (Huang, 2009).

Multi-layer feed-forward ANN’s (i.e. MLP) including one input layer, one output layer, and some hidden layers are one of the most useful and popular structures. The back propagation algorithm that relies on minimization of a convenient cost function is used for training the MLP.

A typical MLP is shown Figure 2.2. A general L-layer FNN includes input hidden and output layers. The layers are connected in a feed-forward style. There are no connections between the neurons of the same layer. Also, there is no feedback in the network structure. In Figure 2.2., the inputs are donated by $x_1, \dots, x_I$, the neurons of the hidden layer are donated by $y_1, \dots, y_J$, and the outputs are denoted by $o_1, \dots, o_K$. The weights in the input layer are shown by $V$ whereas the weights in the hidden layer are shown by $W$.

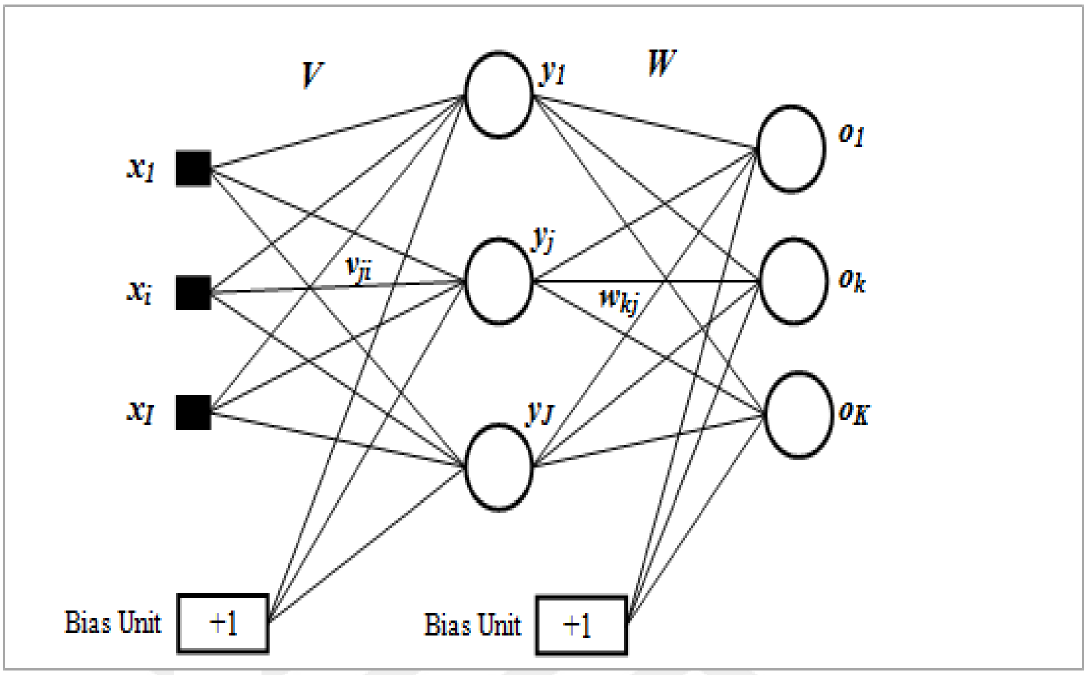


Figure 2.2. A typical MLP structure.

The most commonly used training method for the MLP is the back-propagation, which is basically used to determine the final weights of network. (Li et al., 2012). In the back-propagation algorithm, the error calculated in the output layer neurons is back-propagated to the hidden layer using the gradient-descent method for minimizing the squared-error cost function in neurons according to the formula given in (2.17.)

$$E_p = \frac{1}{2}\sum_{k=1}^{K}(d_{kp} + o_{kp})^2 + E_p \qquad (2.17.)$$

In (2.17.), the count of training patterns is given by *p*, *o* is the real output, the term $E_p$ corresponds to the sum of the squared errors of the cost function and *d* represents the desired output matrix with size of *K*, o is an actual output. Different training functions including Quasi-Newton, Fletcher–Reeves and Levenberg–Marquardt can be used for back propagation.

### 2.1.5. Radial Basis Function Network

RBF's appeared as a variant of artificial neural network in the last years of 1980. Nonetheless, their roots are settled in much more pattern recognition techniques. These techniques can be clustering, mixture models, potential functions, spline interpolation and functional approximation.

RBF network includes several layers and the first layer has input neurons. These neurons feed the feature vectors into the network. The second one is the hidden layer that calculates the result of the main functions. The last layer is the output layer that calculates a linear combination of the main functions (Park and Sanberg, 1991). Simple structures of these networks give a decreasing for training time and make possible learning in stages.

RBF has the feed-forward structure. It has one layer that consists of hidden units. These units are completely adjusted with the output units. (2.18) indicates the output units ($\psi_j$) from a linear combination of the main functions (Ghosh-Dastidar et al., 2008).

$$\psi_j(x) = K\left(\frac{\left\|x - c_j\right\|}{\sigma_j^2}\right) \tag{2.18.}$$

$f : R^n \to R^1$ is estimated by using an RBF network. $x \in R^n$ is an input, $\psi(x, c_j, \sigma_j)$ is the $j^{th}$ function with centre $c_j \in R^n$, width $\sigma_j$, and $w = (w_1, w_2, ..., w_m) \in R^M$ is the vector of linear output weights. M represents the main number function. The M centres $c_j \in R^n$ are connected to obtain $c = (c_1, c_2, ..., c_m) \in R^{nM}$. Lastly, the widths are connected to obtain $\sigma = (\sigma_1, \sigma_2, ..., \sigma_M) \in R^M$. The output of the network for $x \in R^n$ and $\sigma \in R^M$ is shown by (2.19).

$$F(x,c,\sigma,\omega) = \sum_{j=1}^{M} (\omega_j \psi(x,c_j,\sigma_j)) \tag{2.19.}$$

Suppose that $y = (y_1, y_2, ..., y_n)$ is the weighted output vector and $(x_i, y_i): i = 1, 2, ..., N$ is a series of training pairs. For each $c \in R^{nM}$ , $\omega \in R^M, \sigma \in R^M$ and for random weights $\lambda_i, i = 1, 2, ..., N$ , which are taken as positive numbers to point out importance of definite domains of the input space, set (2.20) (Wright et al., 2013).

$$E(c,\sigma,\omega) = \frac{1}{2} \sum_{i=1}^{N} \left[ \lambda_i (y_i - F(x_i, c, \sigma, \omega)) \right]^2 \tag{2.20.}$$

### 2.1.6. Single Decision Trees

Classification and prediction problems can be solved by decision trees, which are powerful data mining methodologies (Kantardzic, 2003). In general, a decision tree can be seen as a predictive model, whereas a tree structure represents a similar behavior to that of a flowchart. Target values can be obtained by decision trees by utilizing classification and analysis (Han et al., 2006).

The classification made by decision trees is based on sorting instances. The sorting is carried out by traversing the tree starting from the root node on top and ending at some intermediate leaf node. Classification of the instance is acheived by this process. There exist several popular tree-growing and tree-pruning algorithms. Quinlan's ID3 algorithm, its extended version C4.5 and C5.0 algorithms all utilize Greedy search methods for tree-growing and tree-pruning.

## 2.2. Overview of Feature Selection Algorithms

### 2.2.1. Relief-F

Relief-F is an instance-based algorithm: Instances are tried randomly and checked whether to have neighborhood of the same or different classes by this algorithm. The algorithm works on continuous and discrete class data. The parameters determine the number of neighbors to check as well as the number of instances to sample. (Witten et al., 2011).

The Pseudo code of the Relief-F algorithm is given in Figure 2.3. Initially, all weights are set to zero (line 1). Then, at each iteration, an instance $R_i$ is randomly selected (line 3). k nearest neighbors of $R_i$, which are called nearest hits Hj and having the same class, are searched (line 4). On the other hand, k nearest neighbors of $R_i$, which are called nearest misses $M_j$ ($C$) and having different class, are searched (lines 5 and 6). $W[A]$ is updated for all attributes with respect to the values of $R_i$ and $H_j$ (Robnik-Sikonja et. al, 2003).

*Input:* for each training instance a vector of attribute values and the class value
*Output:* the vector W of estimations of the qualities of attributes

1. set all weights $W[A] := 0.0$;
2. **for** i := 1 **to** $m$ **do begin**
3. randomly select an instance $R_i$;
4. find $k$ nearest hits $H_j$;
5. **for each class** $C \neq class(R_i)$ **do**
6. from class $C$ find $k$ nearest misses $M_j(C)$;
7. **for** $A$ := 1 **to** a **do**
8. $W[A] := W[A] - \sum_{j=1}^{k} \text{diff}(A, R_i, H_j)/(m \cdot k) +$
9. $\sum_{C \neq class(R_i)} [\frac{P(C)}{1-P(class(R_i))} \sum_{j=1}^{k} \text{diff}(A, R_i, M_j(C))]/(m \cdot k)$;
10. **end**;

Figure 2.3. Pseudo code of Relief-F algorithm

### 2.2.2. Minimum Redundancy Maximum Relevance

The first idea behind mRMR is simple: Highly correlated features should not be taken into account. In other words, the redundancy among the selected features should be as large as possible (Acid, et. al, 2011). In order to achieve this goal, first of all, redundancy among features must be measured somehow.

Assume that $\{X_1, X_2,..., X_n\}$ are discrete random variables and U consists of a set of these random variables. Also, assume that $C$ is a class variable that takes the values $\{c_1, c_2,..., c_k\}$. Any subset of U is represented by $S \subseteq$ U. The redundancy among features is determined by (2.21.)

$$W_I(S) = \frac{1}{|S|^2} \sum_{X_i, X_j \epsilon S} MI(X_i, X_j) \tag{2.21.}$$

In (2.21.), the measure of mutual information between the variables $X_i$ and $X_j$ is given by $(Xi, \ )$.

The second idea behind mRMR is as follows: Maximum relevance criterion of the features with respect to the class variable must be considered. The relevance of the features regarding the class variable C is given by

$$V_I(S) = \frac{1}{|S|} \sum_{X_i \epsilon S} MI(C, X_i) \tag{2.22.}$$

(2.23.) shows a simple but efficient formula to obtain the combination of redundant and relevant features.

$$S^* = argmax_{S \subseteq U} \ (V_I(S) - W_I(S)) \tag{2.23.}$$

It should be noted that the best subset cannot be found by utilizing an exhaustive search. Therefore, an incremental technique, which starts with the feature having maximum value of $(C; X_i)$ $(S_0 = \{X_{i0}\})$ and adds to the current subset $S_{m-1}$ the feature, is required for this purpose. The related formula for this process is given by

$$\max_{X_j \epsilon U \backslash S_{m-1}} \left( MI(C, X_j) - \frac{1}{m-1} \sum_{X_i \epsilon S_{m-1}} MI(X_j, X_i) \right) \tag{2.24.}$$

### 2.2.3. Correlation-Based Feature Selection

In this algorithm, a subset of features is selected depending on a heuristic given by (2.25.),

$$Merit_s = \frac{k\overline{r_{cf}}}{\sqrt{k + k(k-1)\overline{r_{ff}}}}. \tag{2.25.}$$

In (2.25), $Merit_s$ shows how much value a subset containing $k$ features has. The average feature class correlation is designated by $\overline{r_{cf}}$ whereas the average feature-feature intercorrelation is given by $\overline{r_{ff}}$. The numerator gives a sign of the class a group of features and determines the redundance among them is determined by the denominator. The heuristic processes non-relevant features since they are weak estimators of the class. Excess attributes are differentiated as they are the most related with one feature (Hall et al., 1997).

# 3. DEVELOPMENT OF PREDICTION MODELS

The Relief-F and mRMR feature selection algorithms have been utilized to find out the ranking of each feature affecting $UBP_{10}$ and $UBP_{60}$ for all datasets. More specifically, a score has been assigned to each feature with the help of Relief-F and mRMR and these scores have been used to sort the features in descending order. The ranking of the features obtained by applying Relief-F and mRMR for UBP-set1 and UBP-set2 is given in Table 3.1 through Table 3.8. Sixty-eight UBP prediction models have been created by keeping the features with higher scores and removing the ones with lower scores. In contrast to Relief-F and mRMR algorithms, the CFS algorithm which is combined with diffirent search methods such as BestFirst, GeneticSearch, GreedyStepwise and RandomSearch gives a set of selected features to develop four different models for UBP-set1 and five different models for UBP-set2.

Table 3.1. Features' Relief-F scores regarding $UBP_{10}$ in UBP-set1

| **Feature** | **Relief-F Score** |
|---|---|
| Time | 0.04982 |
| Weight | 0.03119 |
| Gender | 0.02179 |
| Height | 0.01217 |
| Age | 0.00982 |
| HRLT | 0.00530 |
| $VO_2max$ | -0.00192 |
| HRmax | -0.00342 |
| Protocol | -0.01735 |

Table 3.2. Features' Relief-F scores regarding $UBP_{60}$ in UBP-set1

| Feature | Relief-F Score |
|---|---|
| Weight | 0.0314 |
| Age | 0.0208 |
| Gender | 0.0182 |
| Height | 0.0071 |
| $VO_2$max | 0.0062 |
| HRmax | 0.0016 |
| Time | 0.0012 |
| HRLT | -0.00112 |
| Protocol | -0.02521 |

Table 3.3. Features' Relief-F scores regarding $UBP_{10}$ in UBP-set2

| Feature | Relief-F Score |
|---|---|
| Gender | 0.0391 |
| Time | 0.0377 |
| BMI | 0.0189 |
| HRLT | 0.0070 |
| Age | 0.0048 |
| $VO_2$max | 0.0047 |
| HRmax | -0.0035 |
| Protocol | -0.0176 |

Table 3.4. Features' Relief-F scores regarding $UBP_{60}$ in UBP-set2

| Feature | Relief-F Score |
|---|---|
| Gender | 0.0300 |
| BMI | 0.0217 |
| Age | 0.0180 |
| $VO_2$max | 0.0071 |
| Time | 0.0039 |
| Hrmax | 0.00004 |
| HRLT | -0.0008012 |
| Protocol | -0.0116566 |

Table 3.5. mRMR scores of the features regarding $UBP_{10}$ in UBP-set1

| **Feature** | **mRMR Score** |
|---|---|
| Gender | 0.79 |
| Weight | 0.40 |
| Height | 0.39 |
| $VO_2$max | 0.34 |
| Time | 0.28 |
| Age | 0.24 |
| HRmax | 0.23 |
| Protocol | 0.19 |
| HRLT | 0.09 |

Table 3.6. mRMR scores of the features regarding $UBP_{60}$ in UBP-set1

| **Feature** | **mRMR Score** |
|---|---|
| Gender | 0.94 |
| Age | 0.89 |
| Weight | 0.87 |
| $VO_2$max | 0.74 |
| HRLT | 0.73 |
| Height | 0.69 |
| HRmax | 0.68 |
| Protocol | 0.54 |
| Time | 0.49 |

Table 3.7. mRMR scores of the features regarding $UBP_{10}$ in UBP-set2

| **Feature** | **mRMR Score** |
|---|---|
| Gender | 0.79 |
| BMI | 0.44 |
| $VO_2$max | 0.38 |
| HRLT | 0.37 |
| Time | 0.26 |
| Age | 0.22 |
| HRmax | 0.21 |
| Protocol | 0.16 |

Table 3.8. mRMR scores of the features regarding $UBP_{60}$ in UBP-set2

| Feature | mRMR Score |
|---|---|
| Gender | 0.93 |
| Age | 0.89 |
| BMI | 0.80 |
| $VO_2$max | 0.75 |
| HRLT | 0.75 |
| HRmax | 0.68 |
| Protocol | 0.53 |
| Time | 0.48 |

For each model, $UBP_{10}$ and $UBP_{60}$ have been predicted using GRNN, SVM, MLP, TB, RBF and SDT. *SEE*, *R* and *MAPE*, the formulas of which are given in (3.1), (3.2) and (3.3), respectively, have been used to evaluate the performance of the developed prediction models. 10-fold cross validation has been utilized to reduce the generalization errors of the prediction models.

$$SEE = \sqrt{\frac{\sum\left(Y - Y'\right)^2}{N}} \tag{3.1.}$$

$$R = \sqrt{1 - \frac{\sum\left(Y - Y'\right)^2}{\sum\left(Y - \bar{Y}\right)^2}} \tag{3.2.}$$

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{Y - Y'}{Y}\right| \tag{3.3.}$$

In (3.1.), (3.2.) and (3.3), $Y$ is the measured UBP value, $Y'$ is the predicted UBP value, $\bar{Y}$ is the mean of the measured UBP values and $N$ is the number of samples in a test subset.

Table 3.9 through Table 3.16 show the prediction models for $UBP_{10}$, $UBP_{60}$ and the features that each model contains.

Table 3.9. Overview of $UBP_{10}$ models produced by Relief-F scores for UBP-set1

| **Models** | **Features** |
|---|---|
| Model 1 | Time, Weight, Gender, Height, Age, HRLT, $VO_2$max, HRmax, Protocol |
| Model 2 | Time, Weight, Gender, Height, Age, HRLT, $VO_2$max, HRmax |
| Model 3 | Time, Weight, Gender, Height, Age, HRLT, $VO_2$max |
| Model 4 | Time, Weight, Gender, Height, Age, HRLT |
| Model 5 | Time, Weight, Gender, Height, Age |
| Model 6 | Time, Weight, Gender, Height |
| Model 7 | Time, Weight, Gender |
| Model 8 | Time, Weight |
| Model 9 | Time |

Table 3.10. Overview of $UBP_{60}$ models produced by Relief-F scores for UBP-set1

| **Models** | **Features** |
|---|---|
| Model 10 | Weight, Age, Gender, Height, $VO_2$max, HRmax, Time, HRLT, Protocol |
| Model 11 | Weight, Age, Gender, Height, $VO_2$max, HRmax, Time, HRLT |
| Model 12 | Weight, Age, Gender, Height, $VO_2$max, HRmax, Time |
| Model 13 | Weight, Age, Gender, Height, $VO_2$max, HRmax |
| Model 14 | Weight, Age, Gender, Height, $VO_2$max |
| Model 15 | Weight, Age, Gender, Height |
| Model 16 | Weight, Age, Gender |
| Model 17 | Weight, Age |
| Model 18 | Weight |

Table 3.11. Overview of $UBP_{10}$ models produced by Relief-F scores for UBP-set2

| **Models** | **Features** |
|---|---|
| Model 19 | Gender, Time, BMI, HRLT, Age, $VO_2$max, HRmax, Protocol |
| Model 20 | Gender, Time, BMI, HRLT, Age, $VO_2$max, HRmax |
| Model 21 | Gender, Time, BMI, HRLT, Age, $VO_2$max |
| Model 22 | Gender, Time, BMI, HRLT, Age |
| Model 23 | Gender, Time, BMI, HRLT |
| Model 24 | Gender, Time, BMI |
| Model 25 | Gender, Time |
| Model 26 | Gender |

Table 3.12. Overview of $UBP_{60}$ models produced by Relief-F scores for UBP-set2

| **Models** | **Features** |
|---|---|
| Model 27 | Gender, BMI, Age, $VO_2$max, Time, HRmax, HRLT, Protocol |
| Model 28 | Gender, BMI, Age, $VO_2$max, Time, HRmax, HRLT |
| Model 29 | Gender, BMI, Age, $VO_2$max, Time, HRmax |
| Model 30 | Gender, BMI, Age, $VO_2$max, Time |
| Model 31 | Gender, BMI, Age, $VO_2$max |
| Model 32 | Gender, BMI, Age |
| Model 33 | Gender, BMI |
| Model 34 | Gender |

Table 3.13. Overview of $UBP_{10}$ models produced by mRMR scores for UBP-set1

| **Models** | **Features** |
|---|---|
| Model 35 | Gender, Weight, Height, $VO_2$max, Time, Age, HRmax, Protocol, HRLT |
| Model 36 | Gender, Weight, Height, $VO_2$max, Time, Age, HRmax, Protocol |
| Model 37 | Gender, Weight, Height, $VO_2$max, Time, Age, HRmax |
| Model 38 | Gender, Weight, Height, $VO_2$max, Time, Age |
| Model 39 | Gender, Weight, Height, $VO_2$max, Time |
| Model 40 | Gender, Weight, Height, $VO_2$max |
| Model 41 | Gender, Weight, Height |
| Model 42 | Gender, Weight |
| Model 43 | Gender |

Table 3.14. Overview of $UBP_{60}$ models produced by mRMR scores for UBP-set1

| **Models** | **Features** |
|---|---|
| Model 44 | Gender, Age, Weight, $VO_2$max, HRLT, Height, HRmax, Protocol, Time |
| Model 45 | Gender, Age, Weight, $VO_2$max, HRLT, Height, HRmax, Protocol |
| Model 46 | Gender, Age, Weight, $VO_2$max, HRLT, Height, HRmax |
| Model 47 | Gender, Age, Weight, $VO_2$max, HRLT, Height |
| Model 48 | Gender, Age, Weight, $VO_2$max, HRLT |
| Model 49 | Gender, Age, Weight, $VO_2$max |
| Model 50 | Gender, Age, Weight |
| Model 51 | Gender, Age |
| Model 52 | Gender |

Table 3.15. Overview of $UBP_{10}$ models produced by mRMR scores for UBP-set2

| **Models** | **Features** |
|---|---|
| Model 53 | Gender, BMI, $VO_2$max, HRLT, Time, Age, HRmax, Protocol |
| Model 54 | Gender, BMI, $VO_2$max, HRLT, Time, Age, HRmax |
| Model 55 | Gender, BMI, $VO_2$max, HRLT, Time, Age |
| Model 56 | Gender, BMI, $VO_2$max, HRLT, Time |
| Model 57 | Gender, BMI, $VO_2$max, HRLT |
| Model 58 | Gender, BMI, $VO_2$max |
| Model 59 | Gender, BMI |
| Model 60 | Gender |

Table 3.16. Overview of $UBP_{60}$ models produced by mRMR scores for UBP-set2

| **Models** | **Features** |
|---|---|
| Model 61 | Gender, Age, BMI, $VO_2$max, HRLT, HRmax, Protocol, Time |
| Model 62 | Gender, Age, BMI, $VO_2$max, HRLT, HRmax, Protocol |
| Model 63 | Gender, Age, BMI, $VO_2$max, HRLT, HRmax |
| Model 64 | Gender, Age, BMI, $VO_2$max, HRLT |
| Model 65 | Gender, Age, BMI, $VO_2$max |
| Model 66 | Gender, Age, BMI |
| Model 67 | Gender, Age |
| Model 68 | Gender |

Table 3.17. Overview of $UBP_{10}$ and $UBP_{60}$ models produced by CFS for UBP-set1

| | **Models** | **Features** |
|---|---|---|
| **$UBP_{10}$** | Model 69 | Gender, Weight, Height, $VO_2$max, Time |
| | Model 70 | Gender, Weight, Height, HRmax, Time, HRLT |
| **$UBP_{60}$** | Model 71 | Protocol, Gender, Weight, Height, $VO_2$max, Time |
| | Model 72 | Gender, Weight, Height, HRmax, Time, HRLT |

Table 3.18. Overview of $UBP_{10}$ and $UBP_{60}$ models produced by CFS for UBP-set2

| | **Models** | **Features** |
|---|---|---|
| **$UBP_{10}$** | Model 73 | Gender, BMI, $VO_2$max, Time, HRLT |
| | Model 74 | Gender, BMI, Time, HRLT |
| **$UBP_{60}$** | Model 75 | Protocol, Gender, $VO_2$max, HRmax, Time, HRLT |
| | Model 76 | Protocol, Gender, $VO_2$max, Time, HRLT |
| | Model 77 | Gender, HRmax, Time, HRLT |

### 3.1. GRNN-based Prediction Models

The flow chart of the GRNN-based UBP prediction model is shown in Figure 3.1. The first step includes the assignment of the scores to each feature and ranking of the features based on the assigned scores. In the GRNN part, the total number of features directly determines the number of input units in the input layer. The input layer is followed by the pattern layer, in which each neuron presents a training pattern and its output. The summation layer comes right after the pattern layer. The summation layer includes a single division unit and summation units. A normalization of the output set is performed by both the summation and output layer. In training of network, radial basis and linear activation functions are used in hidden and output layers. With respect to the cross validation process, UBP is predicted for each fold and also performance measures including *SEE* and *R* are calculated. This process is repeated ten times. Finally, the feature having the minimum score has been removed. At the end of the flow, if the number of features equals zero, the process is terminated, otherwise, the flow returns to the part of dataset partitioning.

The list of intervals for values of the utilized parameters for GRNN-based prediction models is given in Table 3.19.

Table 3.19. Values of the utilized parameters for the GRNN models

| **Parameter** | **Value** |
|---|---|
| Number of neurons | [0, 100] |
| Hidden layer kernel function | Gaussian |
| Weight range | [1, 5] |

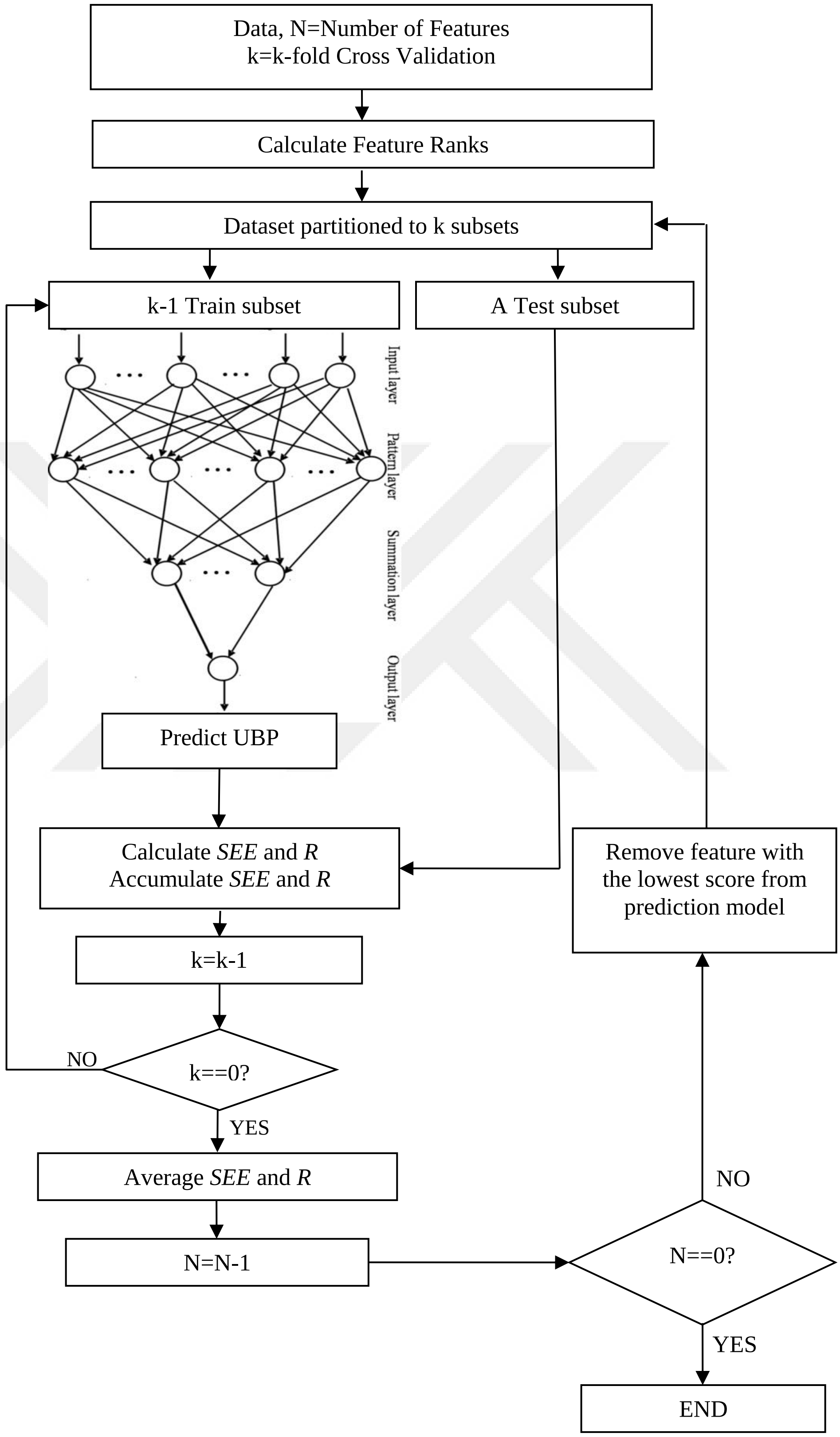


Figure 3.1. Flow chart of GRNN-based model

### 3.2. SVM- based Prediction Models

Building an efficient SVM regression model requires to get the best values of the parameters cost (*C*), epsilon ($\varepsilon$) and $\gamma$, which is the parameter of the RBF kernel. Hence, one needs an effective search algorithm to find the best values of these parameters. There exist many techniques, which include genetic algorithm (Friedrichs and Igel, 2005), particle swarm optimization (Guo et al., 2006), grid search (Hsu et al., 2003), grid search and cross validation, that can be used to obtain the best values of the mentioned parameters. In grid search, the values of the parameters are varied within a pre-defined range in grid search and the triple yielding the maximum performance is selected. The performance can be represented using several metrics including the minimization of root mean square error, the maximization of correlation coefficient or mean absolute error. The problem of generalization can be overcome using cross validation. The original data set is randomly split into k subsets in the case of *k* fold cross validation. The training data is formed using *k - 1* subsets whereas the rest of the data is used as the validation data. This procedure is carried out exactly *k* times (known as the folds) so that the validation data includes every *k* subsets. The final performance measure is obtained by averaging or combining the results of each fold.

The flow chart of the SVM-based UBP prediction model is shown in Figure 3.2. The first step includes the assignment of the scores to each feature and ranking of the features based on the assigned scores. The data set is then preprocessed using standardization so that the predictor variables and the target variable have zero mean and unit variance. This process created new train and test sets. The advantage of applying the standardization process is that the predictor variables with high values get scaled hence the computational power regarding creating the SVM prediction model is reduced. 10-fold cross validation has been applied on the data set to overcome the problem of generalization and improve the reliability of the presented results. The optimal values of the *C*, $\varepsilon$ and $\gamma$ have been found implementing grid search along with. The root mean square error is used as the performance metric for 10-fold cross validation. The triple yielding the lowest value of the root mean square error has been selected for building the prediction model, which in turn is utilized for predicting UBP values in the test set.

Finally, the feature having the minimum score has been removed. At the end of the flow, if the number of features equals zero, the process is terminated, otherwise, the flow returns to the part of dataset partitioning.

Table 3.20. The intervals for values of the parameters for the SVM model

| **Parameter** | **Value** |
|---|---|
| Cost (*C*) | [0.1, 5000] |
| Gamma ($\gamma$) | [0.0001, 100] |
| Epsilon (ε) | [0.001, 50] |

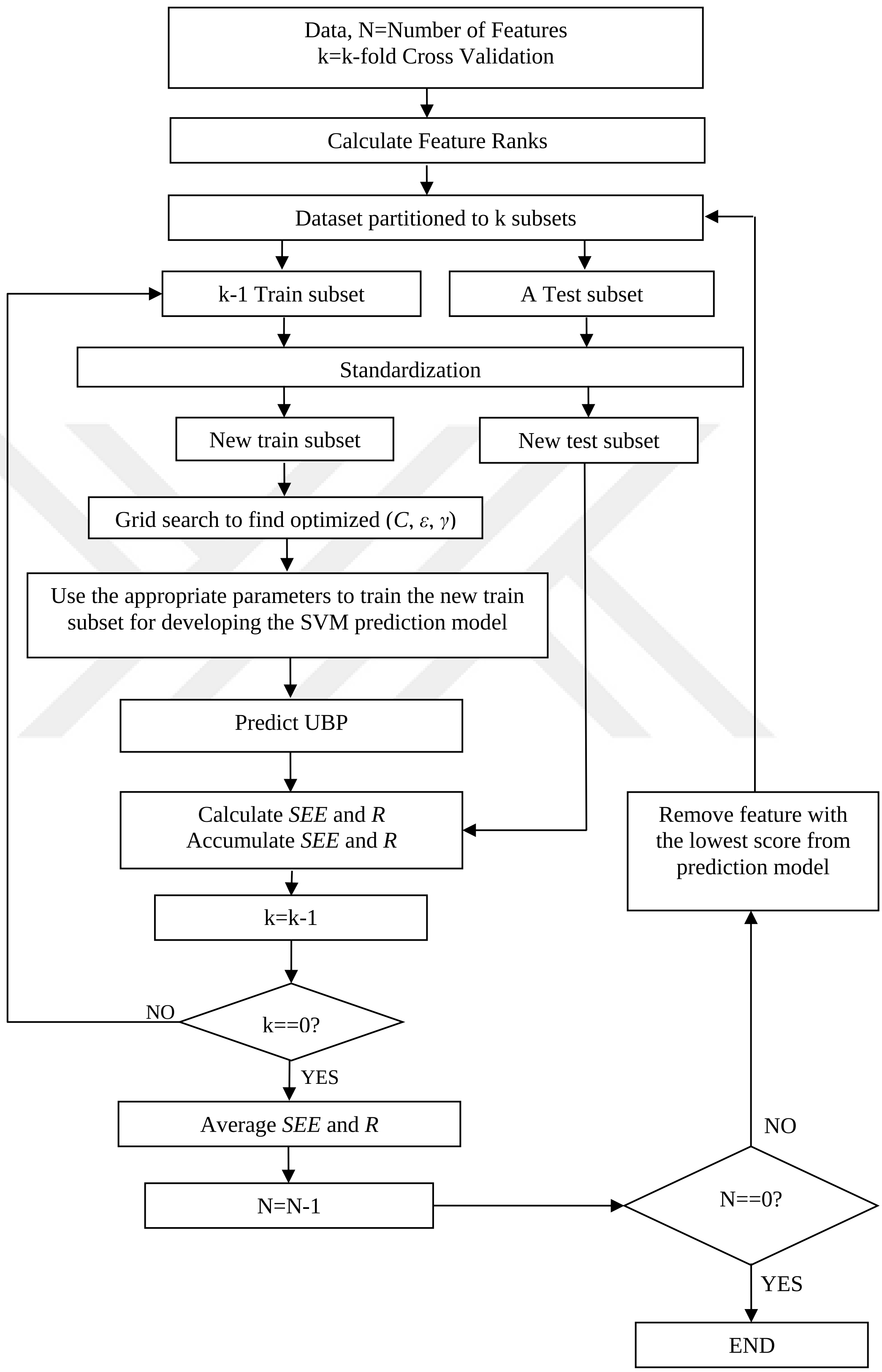


Figure 3.2. Flow chart of SVM-based model

### 3.3. TB- based Prediction Models

The Tree Boost algorithm is given by (3.4.).

$$F(x) = F_0 + A1 * T1(x) + A2 * T2(x) + \cdots + AM * TM(x) \quad (3.4.)$$

In (3.4.), F(x) represents the predicted target value, F0 is the initial value, x is a vector representing the residuals, the first tree of the series that fits the residuals is represented by T1(x) and the coefficients of the tree nodes are A1, A2, etc. The following rules have been applied to build TB-based UBP prediction models:

1. The mean of the target variable (i.e. the coefficient of $F_0$) has been found.
2. The rows feeding the next tree have been selected.
3. Huber's Quantile Cutoff factor has been used to transform the residuals (i.e. pseudo-residuals) that have been sorted.
4. The first tree (T1) has been fitted to pseudo-residuals.
5. The mean of the pseudo-residuals in each of the terminal nodes (i.e. predicted variable) has been calculated.
6. The residuals between the predicted variable and the pseudo-residuals that fed the tree were calculated and Huber's Quantile Cutoff factor has been applied again to obtain the mean of these residuals.
7. The boost coefficient (A1) has been calculated.
8. The boost coefficient has been multiplied by the shrink value to terminate the learning process.

The search intervals for values of the parameters for the TB models are given in Table 3.21.

Table 3.21. Values of the utilized parameters for the TB models

| **Parameter** | **Value** |
|---|---|
| Maximum number of trees used in series | [200, 550] |
| Depth of individual trees | [5, 7] |
| Minimum size node to split | [8, 23] |

### 3.4. MLP- based Prediction Models

UBP prediction models have been developed with the help of MLP utilizing the back propagation algorithm. The performance of an MLP based prediction model is affected by many factors including preprocessing, the number of hidden layers, the type of activation functions used in the hidden and output layers and the number of hidden layer neurons.

The inputs and outputs are normalized so that they have zero mean and unity variance. The tansigmoid activation function is used in the hidden layers whereas the pure-linear activation function is used in the output layers. The Levenberg–Marquardt algorithm is utilized for training the network. There is no specific rule for determining the number of hidden layers and the number of neurons in each hidden layer. The numbers can change from one problem to another. For developing UBP prediction models, multilayer feed-forward neural networks with 1 to 4 layers have been implemented. The number of hidden layer neurons has been varied within specific intervals and the number yielding the lowest *SEE* in the training phase has been chosen to develop UBP prediction models.

The search intervals for values of the parameters for the MLP models are given in Table 3.22.

Table 3.22. Values of the utilized parameters for the MLP models

| **Parameter** | **Value** |
|---|---|
| Number of hidden layers | [1, 4] |
| Number of neurons in the hidden layer | [2, 20] |
| Hidden layer activation function | Tansigmoid |
| Output layer activation function | Linear |

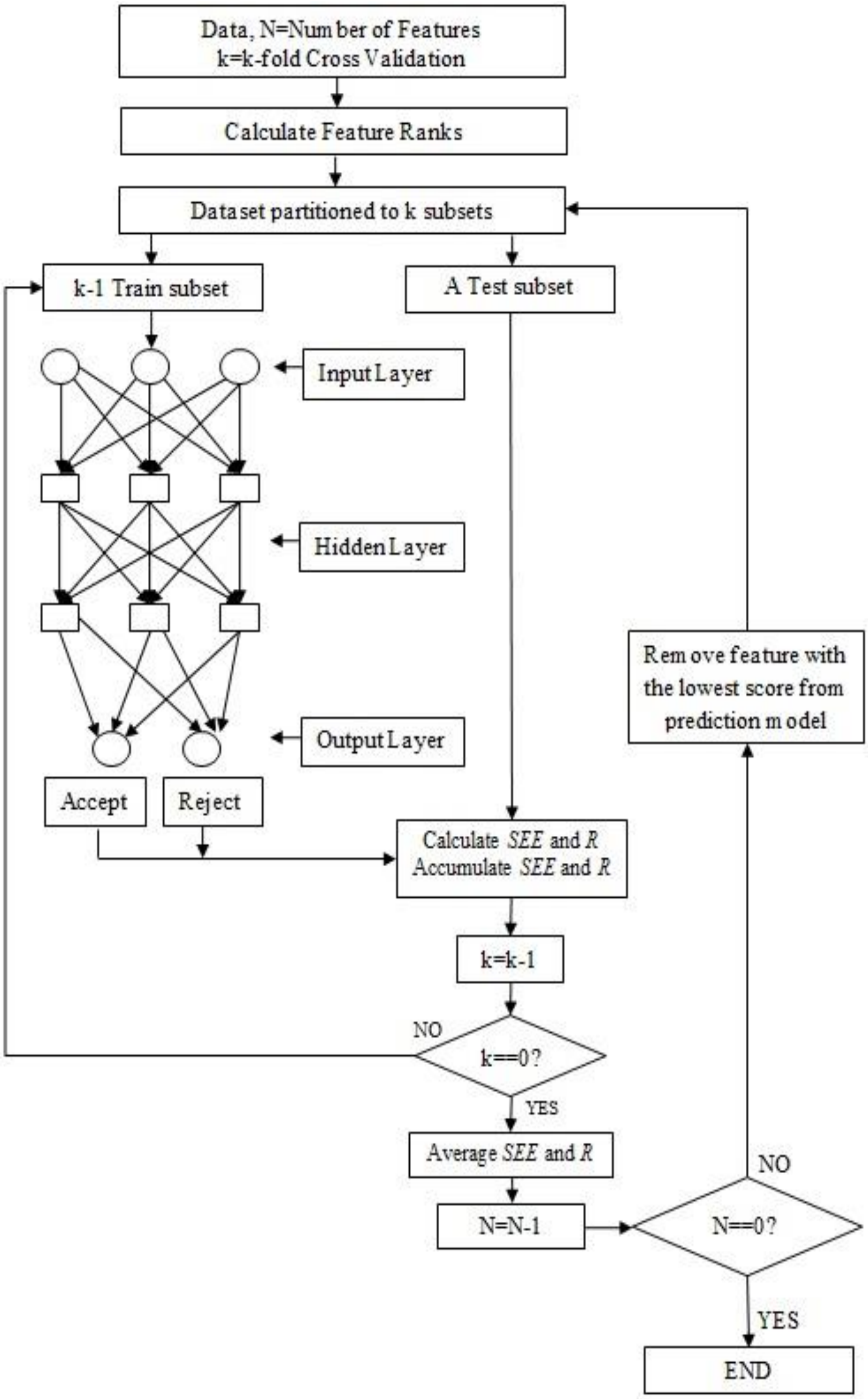


Figure 3.3. Flow chart of MLP-based model

### 3.5. RBF- based Prediction Models

The flow chart of the RBF-based UBP prediction model is shown in Figure 3.3. The first step includes the assignment of the scores to each feature and ranking of the features based on the assigned scores. Then, a neuron is added to the network at each step until the desired error has been reached in the training stage. After the weight for output layer is computed, the error at output of network is calculated. If the error is acceptable, the performance of the network is measured for the test set. Otherwise, the process goes back to the section of adding a new neuron. Once the error is acceptable, the *SEE* and *R* values are calculated. Finally, the feature having the minimum score has been removed. At the end of the flow, if the number of features equals zero, the process is terminated, otherwise, the flow returns to the part of dataset partitioning.

The parameters affecting the performance of RBF based models are the number of neurons, radius and lambda. Table 3.23 lists the ranges for values of the utilized parameters for RBF based prediction models.

Table 3.23. Values of the utilized parameters for the RBF models

| **Parameter** | **Value** |
|---|---|
| Number of neurons | [3, 100] |
| Radius | [0.01 , 400] |
| Lambda | [0.001, 10] |

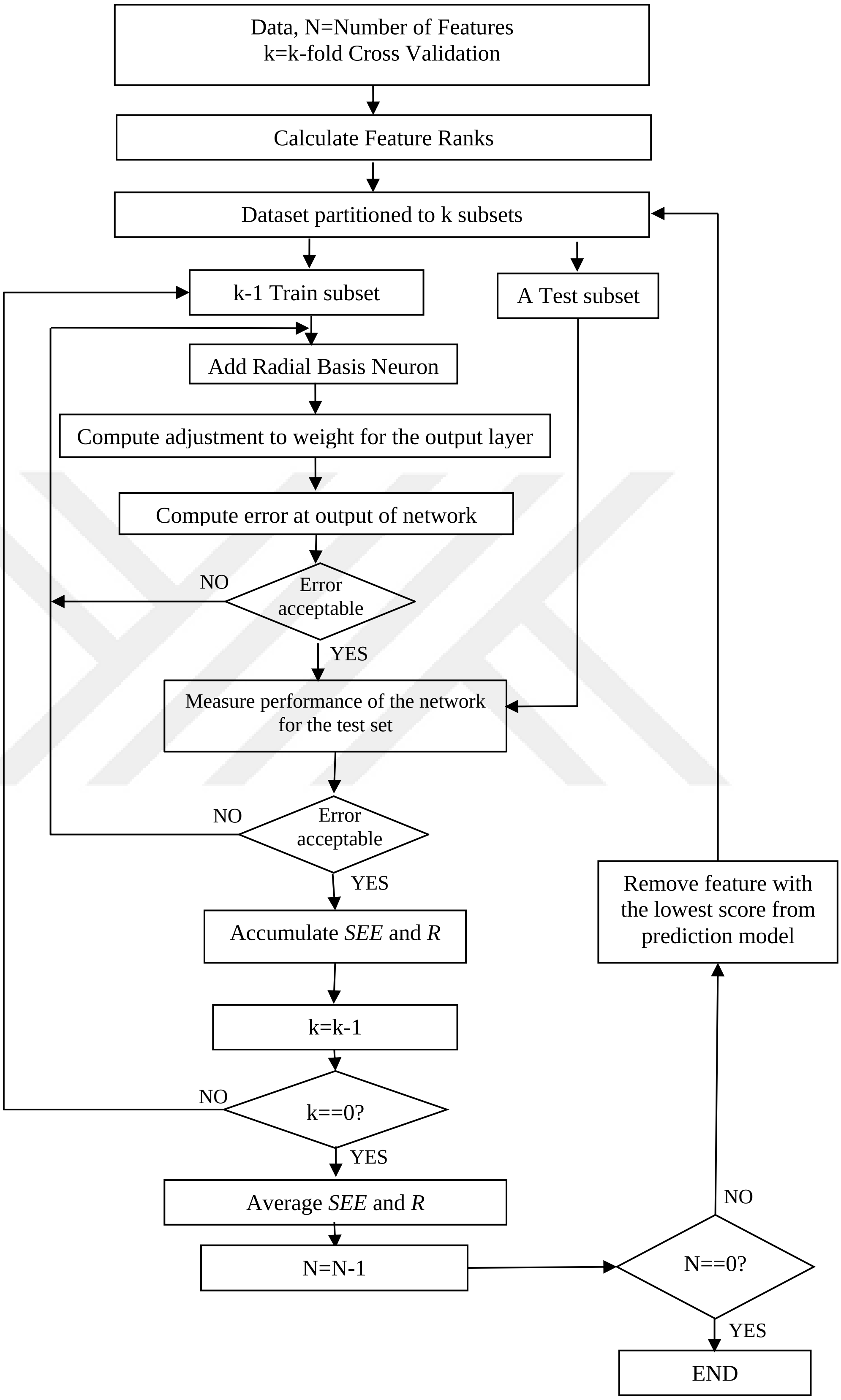


Figure 3.4. Flow chart of RBF-based model

### 3.6. SDT- based Prediction Models

There are three important parameters that affect the performance of SDT-based prediction models. One of the parameters is known as the minimum rows in a node and the others are known as the minimum size node to split and maximum tree levels. The list of intervals for values of the utilized parameters for SDT-based prediction models is shown in Table 3.24.

Table 3.24. Values of the utilized parameters for the SDT models

| **Parameter** | **Value** |
|---|---|
| Minimum rows in a node | [2, 10] |
| Minimum size node to split | [6, 8] |
| Maximum tree levels | [7, 10] |

# 4. RESULTS AND DISCUSSION

As described in Section 1, two datasets have been used in this thesis. The first data set comprises the features Protocol, Gender, Age, Weight, Height, $VO_2$max, HRmax, Time and HRLT. The second data set includes Protocol, Gender, Age, BMI, $VO_2$max, HRmax, Time and HRLT. Both datasets have the two target variables $UBP_{10}$ and $UBP_{60}$. The prediction models have been developed on these two datasets with the integration of Relief-F and mRMR feature selection algorithms and CFS feature subset selection algorithm in GRNN, RBF, MLP, SVM, SDT and TB.

## 4.1. Results and Discussion for Prediction Models with Relief-F

The *SEE*'s related to the prediction of both $UBP_{10}$ and $UBP_{60}$ for UBP-set1 are listed in ascending order Table 4.1 and Table 4.2, respectively.

Table 4.1. *SEE*'s for Relief-F-based $UBP_{10}$ prediction models (UBP-set1)

| **Models** | **SEE** | | | | | |
|---|---|---|---|---|---|---|
| | **GRNN** | **SVM** | **MLP** | **TB** | **RBF** | **SDT** |
| Model 5 | 24.73 | 27.80 | 29.26 | 29.34 | 30.92 | 38.83 |
| Model 6 | 25.23 | 28.73 | 29.57 | 32.85 | 33.03 | 39.30 |
| Model 1 | 26.76 | 28.81 | 29.91 | 33.20 | 34.47 | 39.83 |
| Model 2 | 26.91 | 29.22 | 30.25 | 33.26 | 37.25 | 40.19 |
| Model 3 | 27.80 | 29.42 | 30.68 | 33.87 | 38.63 | 40.46 |
| Model 4 | 28.42 | 29.58 | 31.13 | 33.88 | 38.71 | 40.51 |
| Model 7 | 28.47 | 30.67 | 31.38 | 34.09 | 40.13 | 40.52 |
| Model 8 | 43.19 | 43.50 | 44.87 | 45.69 | 50.84 | 51.68 |
| Model 9 | 57.80 | 59.36 | 60.94 | 61.16 | 63.05 | 63.45 |

Table 4.2. *SEE*’s for Relief-F-based $UBP_{60}$ prediction models (UBP-set1)

| Models | SEE | | | | | |
|---|---|---|---|---|---|---|
| | GRNN | SVM | TB | MLP | RBF | SDT |
| Model 14 | 14.94 | 18.73 | 19.51 | 20.42 | 22.15 | 24.59 |
| Model 13 | 15.66 | 18.75 | 19.59 | 20.57 | 22.48 | 25.11 |
| Model 15 | 17.14 | 19.47 | 19.79 | 20.66 | 23.08 | 26.28 |
| Model 12 | 17.38 | 19.64 | 20.49 | 20.98 | 23.38 | 29.43 |
| Model 10 | 17.62 | 20.19 | 20.69 | 22.49 | 23.94 | 29.75 |
| Model 11 | 18.34 | 20.43 | 20.79 | 22.71 | 24.85 | 30.06 |
| Model 16 | 19.93 | 20.69 | 21.47 | 23.90 | 25.24 | 30.59 |
| Model 17 | 36.96 | 40.85 | 41.93 | 42.41 | 42.65 | 43.78 |
| Model 18 | 40.88 | 42.68 | 43.49 | 45.17 | 45.32 | 48.42 |

The *R*'s related to the prediction models of both $UBP_{10}$ and $UBP_{60}$ for UBP-set1 are listed in descending order Table 4.3 and Table 4.4, respectively.

Table 4.3. *R*’s for Relief-F-based $UBP_{10}$ prediction models (UBP-set1)

| Models | R | | | | | |
|---|---|---|---|---|---|---|
| | GRNN | SVM | MLP | TB | RBF | SDT |
| Model 5 | 0.937 | 0.919 | 0.910 | 0.909 | 0.899 | 0.835 |
| Model 6 | 0.934 | 0.914 | 0.908 | 0.885 | 0.884 | 0.831 |
| Model 1 | 0.925 | 0.913 | 0.906 | 0.883 | 0.873 | 0.826 |
| Model 2 | 0.925 | 0.910 | 0.904 | 0.882 | 0.849 | 0.822 |
| Model 3 | 0.919 | 0.909 | 0.901 | 0.878 | 0.837 | 0.820 |
| Model 4 | 0.915 | 0.908 | 0.897 | 0.877 | 0.836 | 0.819 |
| Model 7 | 0.915 | 0.901 | 0.896 | 0.876 | 0.823 | 0.818 |
| Model 8 | 0.791 | 0.787 | 0.772 | 0.762 | 0.694 | 0.681 |
| Model 9 | 0.574 | 0.541 | 0.505 | 0.499 | 0.450 | 0.438 |

Table 4.4. *R*'s for Relief-F-based $UBP_{60}$ prediction models (UBP-set1)

| **Models** | **R** | | | | | |
|---|---|---|---|---|---|---|
| | **GRNN** | **SVM** | **TB** | **MLP** | **RBF** | **SDT** |
| Model 14 | 0.960 | 0.937 | 0.932 | 0.925 | 0.911 | 0.889 |
| Model 13 | 0.956 | 0.937 | 0.931 | 0.924 | 0.908 | 0.884 |
| Model 15 | 0.948 | 0.932 | 0.929 | 0.923 | 0.903 | 0.872 |
| Model 12 | 0.946 | 0.930 | 0.924 | 0.920 | 0.900 | 0.836 |
| Model 10 | 0.945 | 0.926 | 0.923 | 0.908 | 0.895 | 0.832 |
| Model 11 | 0.940 | 0.925 | 0.922 | 0.906 | 0.886 | 0.828 |
| Model 16 | 0.929 | 0.923 | 0.917 | 0.895 | 0.882 | 0.822 |
| Model 17 | 0.725 | 0.649 | 0.624 | 0.613 | 0.607 | 0.578 |
| Model 18 | 0.648 | 0.606 | 0.586 | 0.540 | 0.536 | 0.431 |

The *MAPE*'s related to the prediction models of both $UBP_{10}$ and $UBP_{60}$ for UBP-set1 are listed in ascending order Table 4.5 and Table 4.6, respectively.

Table 4.5. *MAPE*'s for Relief-F-based $UBP_{10}$ prediction models (UBP-set1)

| **Models** | **MAPE (%)** | | | | | |
|---|---|---|---|---|---|---|
| | **GRNN** | **SVM** | **MLP** | **TB** | **RBF** | **SDT** |
| Model 5 | 8.63 | 9.79 | 10.87 | 11.05 | 11.05 | 13.80 |
| Model 6 | 8.98 | 9.89 | 10.90 | 11.17 | 11.84 | 13.89 |
| Model 1 | 9.51 | 10.65 | 11.06 | 11.44 | 12.42 | 14.72 |
| Model 2 | 9.60 | 10.73 | 11.35 | 11.68 | 12.71 | 14.78 |
| Model 3 | 9.75 | 10.81 | 11.37 | 11.94 | 13.51 | 14.84 |
| Model 4 | 9.92 | 10.90 | 11.38 | 11.96 | 14.41 | 14.85 |
| Model 7 | 9.97 | 11.05 | 11.81 | 11.99 | 14.62 | 14.86 |
| Model 8 | 11.54 | 16.06 | 16.28 | 16.77 | 16.91 | 17.35 |
| Model 9 | 15.50 | 23.64 | 23.93 | 24.78 | 25.67 | 26.49 |

Table 4.6. *MAPE*'s for Relief-F-based $UBP_{60}$ prediction models (UBP-set1)

| **Models** | **MAPE (%)** | | | | | |
|---|---|---|---|---|---|---|
| | **GRNN** | **SVM** | **MLP** | **TB** | **RBF** | **SDT** |
| Model 14 | 6.83 | 7.84 | 8.68 | 8.96 | 9.60 | 11.66 |
| Model 13 | 7.26 | 7.91 | 9.13 | 9.64 | 9.73 | 11.74 |
| Model 15 | 7.58 | 9.01 | 9.15 | 10.51 | 10.52 | 12.29 |
| Model 12 | 7.71 | 9.33 | 9.43 | 10.60 | 10.61 | 13.38 |
| Model 10 | 7.85 | 9.61 | 9.98 | 10.85 | 11.05 | 13.53 |
| Model 11 | 8.16 | 9.66 | 10.02 | 11.06 | 11.40 | 13.61 |
| Model 16 | 8.88 | 9.95 | 10.48 | 11.58 | 11.67 | 13.66 |
| Model 17 | 16.61 | 19.71 | 19.92 | 20.07 | 21.34 | 22.42 |
| Model 18 | 20.07 | 20.57 | 20.96 | 21.53 | 21.65 | 22.61 |

The following points can be stated for UBP prediction models developed using UBP-set1:

- With respect to the ranking of the Relief-F algorithm, the features exercise time and weight exhibit the highest scores for $UBP_{10}$ and $UBP_{60}$ predictions, respectively. On the other hand; the feature, having the minimum Relief-F score for both $UBP_{10}$ and $UBP_{60}$ prediction, is protocol.
- Using less number of features than the full set of features yield lower *SEE*'s and higher *R*'s for $UBP_{10}$ and $UBP_{60}$ predictions. In more details; Model 5 and Model 14, which include less number features, lead up to 13.16% and 20.98% reductions in *SEE*'s for prediction of $UBP_{10}$ and $UBP_{60}$, in the order given.
- The prediction model including just the feature exercise time yields the highest *SEE*'s for $UBP_{10}$ prediction. On the other hand; the model containing the feature weight only yields the highest *SEE*'s for $UBP_{60}$ prediction. In particular, SDT based models give the highest *SEE*'s (63.45 W, 48.42 W) and the lowest *R*'s (0.438, 0.431) for prediction of $UBP_{10}$ and $UBP_{60}$, respectively.
- Observation of the *SEE*'s of Model 7 including the features time, weight, gender and Model 8 including the features time and weight suggests that the feature gender has significant effect for $UBP_{10}$ prediction. In more detail, the

inclusion of gender in the aforementioned model leads in average 36.30% reduction in *SEE* for predicting $UBP_{10}$. A similar case is observed when the *SEE*'s of the Model 16 including weight, age, gender and the Model 17 including weight and age are compared. In particular, the addition of gender in the related model causes in average 75.28% decrement in *SEE* for prediction of $UBP_{60}$.

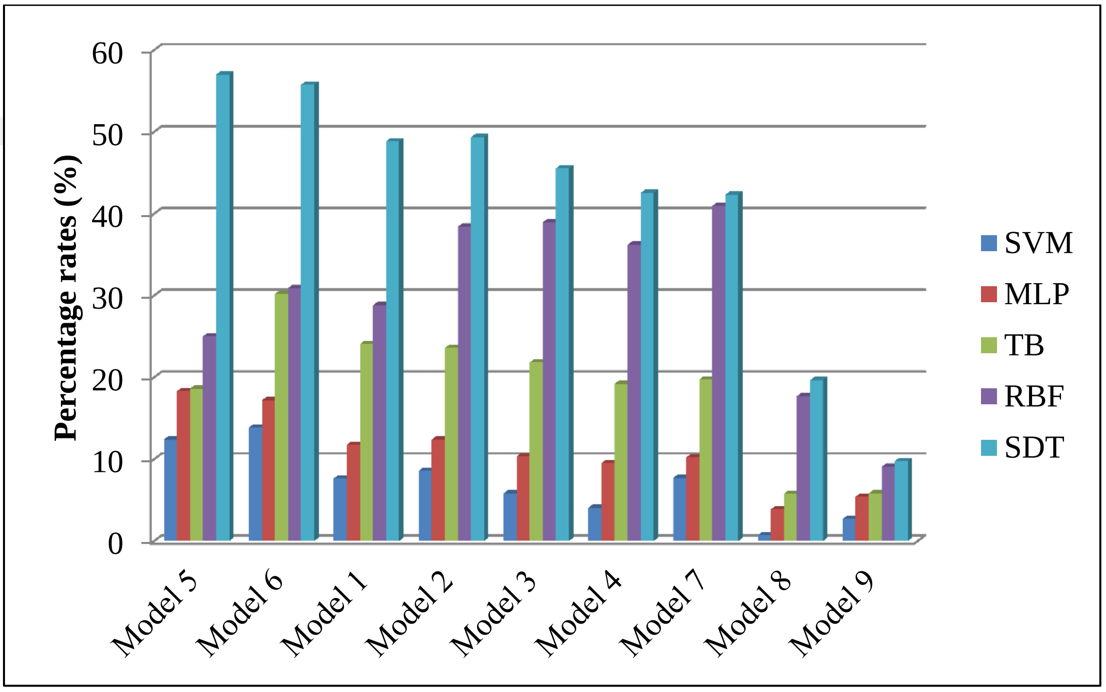


Figure 4.1. Percentage *decrease* rates in *SEE* values of $UBP_{10}$ for GRNN-based models compared to *SEE* values obtained by SVM, MLP, TB, RBF and SDT-based models using Relief-F on UBP-set1

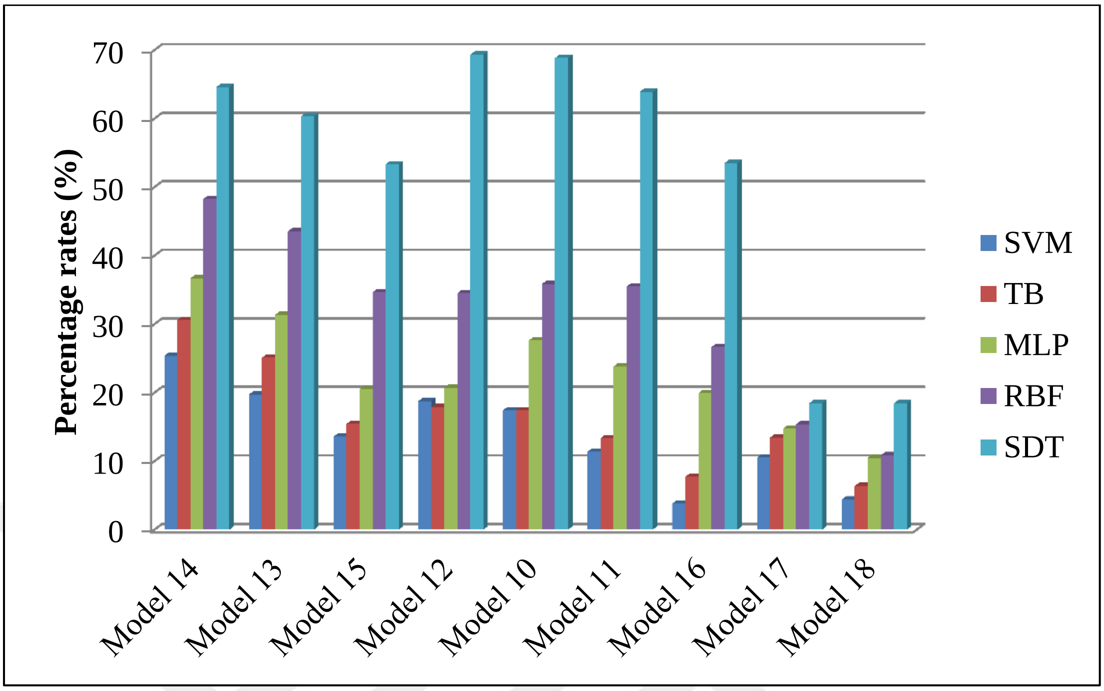


Figure 4.2. Percentage decrease *rates* in *SEE* values of $UBP_{60}$ for GRNN-based models compared to *SEE* values obtained by SVM, TB, MLP, RBF and SDT-based models using Relief-F on UBP-set1

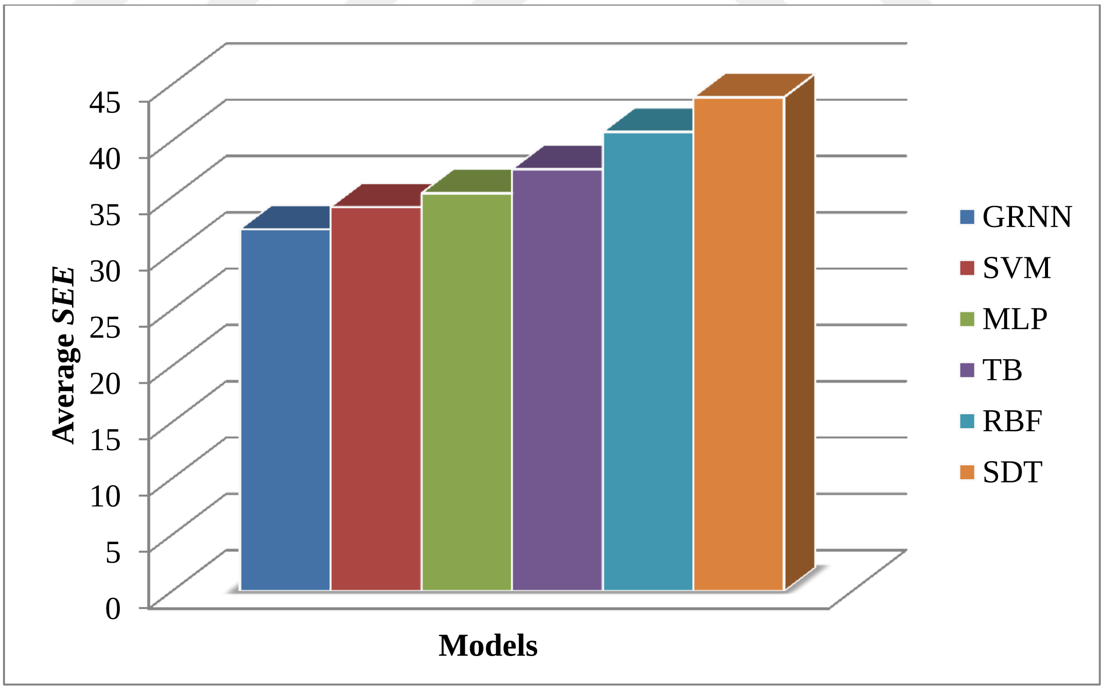


Figure 4.3. Average *SEE*'s of *prediction* models for $UBP_{10}$ using Relief-F on UBP-set1

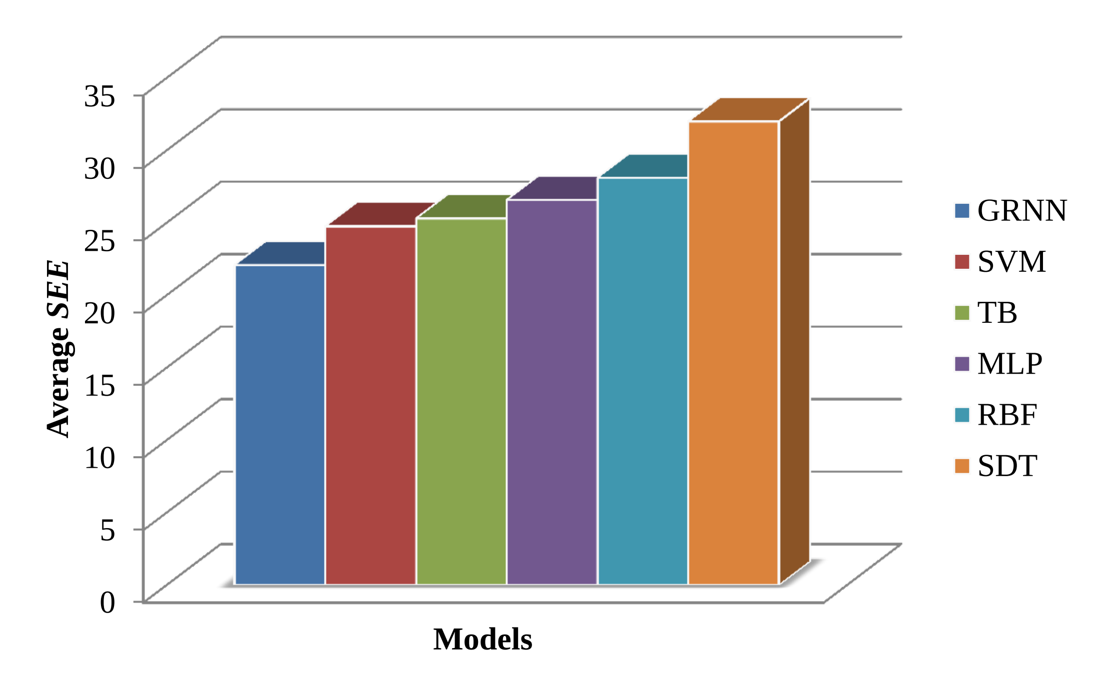


Figure 4.4. Average *SEE*'s of *prediction* models for $UBP_{60}$ using Relief-F on UBP-set1

- As compared to the *SEE*'s obtained by SVM, MLP, TB, RBF and SDT, the maximum percentage decrement rates in *SEE*'s obtained by GRNN are in average 7.06%, 11.02%, 18.77%, 29.57% and 41.20% for prediction of $UBP_{10}$, respectively.
- Similarly, as compared to the *SEE*'s obtained by SVM, TB, MLP, RBF and SDT, the maximum percentage decrement rates in *SEE*'s obtained by GRNN are in average 13.89%, 16.38%, 22.88%, 31.69% and 52.30% for prediction of $UBP_{60}$, in the order given.
- For $UBP_{10}$ and $UBP_{60}$, prediction models including the features weight, height, gender and age give lower *SEE*'s than other models. This in turn means that the correlation of UBP with these four features is quite high.

The *SEE*'s related to the prediction models of both $UBP_{10}$ and $UBP_{60}$ for UBP-set2 are listed in ascending order Table 4.7 and Table 4.8, respectively.

Table 4.7. *SEE*'s for Relief-F-based $UBP_{10}$ prediction models (UBP-set2)

| Models | SEE | | | | | |
|---|---|---|---|---|---|---|
| | **GRNN** | **SVM** | **MLP** | **TB** | **RBF** | **SDT** |
| Model 24 | 32.00 | 32.37 | 33.64 | 35.53 | 37.54 | 39.81 |
| Model 22 | 32.01 | 32.84 | 34.13 | 35.84 | 38.67 | 40.45 |
| Model 23 | 32.67 | 33.11 | 34.34 | 37.52 | 39.24 | 40.91 |
| Model 19 | 32.83 | 33.91 | 34.43 | 37.55 | 39.28 | 42.76 |
| Model 20 | 32.96 | 34.05 | 35.77 | 38.07 | 39.35 | 42.89 |
| Model 21 | 33.35 | 34.14 | 37.93 | 39.14 | 40.29 | 43.29 |
| Model 25 | 36.15 | 36.16 | 39.71 | 39.85 | 40.32 | 43.77 |
| Model 26 | 40.02 | 40.11 | 40.17 | 40.20 | 40.39 | 45.09 |

Table 4.8. *SEE*'s for Relief-F-based $UBP_{60}$ prediction models (UBP-set2)

| Models | SEE | | | | | |
|---|---|---|---|---|---|---|
| | **GRNN** | **SVM** | **TB** | **MLP** | **RBF** | **SDT** |
| Model 32 | 18.45 | 21.40 | 21.92 | 23.44 | 24.71 | 26.51 |
| Model 31 | 19.90 | 21.72 | 22.52 | 23.95 | 24.86 | 26.92 |
| Model 30 | 20.23 | 21.73 | 22.77 | 24.54 | 25.19 | 27.53 |
| Model 27 | 20.36 | 21.88 | 23.34 | 24.60 | 25.94 | 28.32 |
| Model 28 | 21.61 | 21.89 | 24.02 | 25.61 | 26.86 | 29.68 |
| Model 29 | 21.70 | 21.99 | 24.29 | 26.84 | 27.77 | 29.85 |
| Model 33 | 24.69 | 24.87 | 25.16 | 27.16 | 28.54 | 30.02 |
| Model 34 | 28.86 | 29.24 | 29.25 | 29.26 | 29.46 | 31.14 |

The *R*'s related to the prediction of both $UBP_{10}$ and $UBP_{60}$ for UBP-set2 are listed in descending order Table 4.9. and Table 4.10, respectively.

Table 4.9. *R*'s for Relief-F-based $UBP_{10}$ prediction models (UBP-set2)

| Models | R | | | | | |
|---|---|---|---|---|---|---|
| | **GRNN** | **SVM** | **MLP** | **TB** | **RBF** | **SDT** |
| Model 24 | 0.891 | 0.889 | 0.879 | 0.864 | 0.847 | 0.826 |
| Model 22 | 0.891 | 0.885 | 0.875 | 0.861 | 0.837 | 0.820 |
| Model 23 | 0.886 | 0.883 | 0.874 | 0.847 | 0.831 | 0.815 |
| Model 19 | 0.885 | 0.877 | 0.873 | 0.847 | 0.831 | 0.796 |
| Model 20 | 0.884 | 0.876 | 0.862 | 0.842 | 0.830 | 0.794 |
| Model 21 | 0.881 | 0.875 | 0.843 | 0.832 | 0.821 | 0.790 |
| Model 25 | 0.859 | 0.859 | 0.827 | 0.825 | 0.821 | 0.785 |
| Model 26 | 0.824 | 0.823 | 0.822 | 0.822 | 0.820 | 0.769 |

Table 4.10. *R*'s for Relief-F-based $UBP_{60}$ prediction models (UBP-set2)

| Models | R | | | | | |
|---|---|---|---|---|---|---|
| | **GRNN** | **SVM** | **TB** | **MLP** | **RBF** | **SDT** |
| Model 32 | 0.939 | 0.917 | 0.913 | 0.900 | 0.888 | 0.869 |
| Model 31 | 0.929 | 0.914 | 0.908 | 0.895 | 0.886 | 0.865 |
| Model 30 | 0.926 | 0.914 | 0.905 | 0.889 | 0.883 | 0.858 |
| Model 27 | 0.925 | 0.913 | 0.900 | 0.889 | 0.875 | 0.849 |
| Model 28 | 0.915 | 0.913 | 0.894 | 0.879 | 0.866 | 0.833 |
| Model 29 | 0.914 | 0.912 | 0.892 | 0.866 | 0.856 | 0.831 |
| Model 33 | 0.888 | 0.886 | 0.883 | 0.862 | 0.847 | 0.829 |
| Model 34 | 0.843 | 0.838 | 0.838 | 0.838 | 0.836 | 0.814 |

The *MAPE*'s related to the prediction of both $UBP_{10}$ and $UBP_{60}$ for UBP-set2 are listed in ascending order Table 4.11. and Table 4.12, respectively.

Table 4.11. *MAPE*'s for Relief-F-based $UBP_{10}$ prediction models (UBP-set2)

| **Models** | **MAPE (%)** | | | | | |
|---|---|---|---|---|---|---|
| | **GRNN** | **SVM** | **MLP** | **TB** | **RBF** | **SDT** |
| Model 24 | 10.96 | 11.56 | 12.16 | 12.72 | 13.94 | 14.40 |
| Model 22 | 11.19 | 12.28 | 12.29 | 13.35 | 13.95 | 14.58 |
| Model 23 | 11.42 | 12.31 | 12.40 | 13.81 | 14.45 | 15.39 |
| Model 19 | 11.42 | 12.36 | 12.48 | 13.82 | 14.45 | 15.44 |
| Model 20 | 12.09 | 12.43 | 12.62 | 13.87 | 14.47 | 15.55 |
| Model 21 | 12.14 | 12.44 | 12.87 | 14.67 | 14.67 | 15.55 |
| Model 25 | 13.82 | 14.26 | 14.91 | 14.92 | 15.20 | 15.90 |
| Model 26 | 15.46 | 15.50 | 15.53 | 15.74 | 15.75 | 16.51 |

Table 4.12. *MAPE*'s for Relief-F-based $UBP_{60}$ prediction models (UBP-set2)

| **Models** | **MAPE (%)** | | | | | |
|---|---|---|---|---|---|---|
| | **GRNN** | **SVM** | **MLP** | **TB** | **RBF** | **SDT** |
| Model 32 | 8.01 | 9.59 | 9.86 | 11.24 | 11.26 | 12.51 |
| Model 31 | 9.36 | 9.98 | 10.29 | 11.52 | 12.10 | 12.66 |
| Model 30 | 9.39 | 10.49 | 11.09 | 11.63 | 12.21 | 12.82 |
| Model 27 | 9.46 | 10.52 | 11.17 | 11.92 | 12.25 | 13.04 |
| Model 28 | 9.73 | 10.61 | 11.20 | 12.07 | 12.64 | 13.74 |
| Model 29 | 10.41 | 10.63 | 11.74 | 12.37 | 14.00 | 13.94 |
| Model 33 | 12.13 | 12.20 | 12.26 | 12.55 | 15.33 | 13.94 |
| Model 34 | 14.28 | 14.37 | 14.43 | 14.62 | 15.34 | 15.50 |

In a similar fashion, the following statements are valid for UBP prediction models developed using UBP-set2:

- With respect to the ranking of the Relief-F algorithm, the feature gender exhibits the highest scores for $UBP_{10}$ and $UBP_{60}$ predictions, respectively. On the other hand; the feature, having the minimum Relief-F score for both $UBP_{10}$ and $UBP_{60}$ prediction, is protocol.
- Using less number of features than the full set of features yield lower *SEE*'s and higher *R*'s for $UBP_{10}$ and $UBP_{60}$ predictions. In more details; Model 24 and Model 32, which include less number features, lead up to 7.41% and 10.35% reductions in *SEE*'s for prediction of $UBP_{10}$ and $UBP_{60}$, in the order given.
- The prediction model consisting of only the feature gender (i.e. Model 26 and Model 34) yields the highest *SEE*'s and lowest *R*'s for both prediction of $UBP_{10}$ and $UBP_{60}$. In particular, SDT based models yield the highest *SEE*'s (45.09 W, 31.14 W) and the lowest *R*'s (0.769, 0.814) for prediction of $UBP_{10}$ and $UBP_{60}$, respectively.
- As compared to the *SEE*'s obtained by SVM, MLP, TB, RBF and SDT, the maximum percentage decrement rates in *SEE*'s obtained by GRNN are in average 1.79%, 6.78%, 11.97%, 16.32% and 24.99% for prediction of $UBP_{10}$, in the order given.
- Similarly, as compared to the *SEE*'s obtained by SVM, TB, MLP, RBF and SDT, the maximum percentage decrement rates in *SEE*'s obtained by GRNN are in average 5.59%, 10.69%, 17.89%, 22.59% and 32.32% for prediction of $UBP_{60}$, respectively.

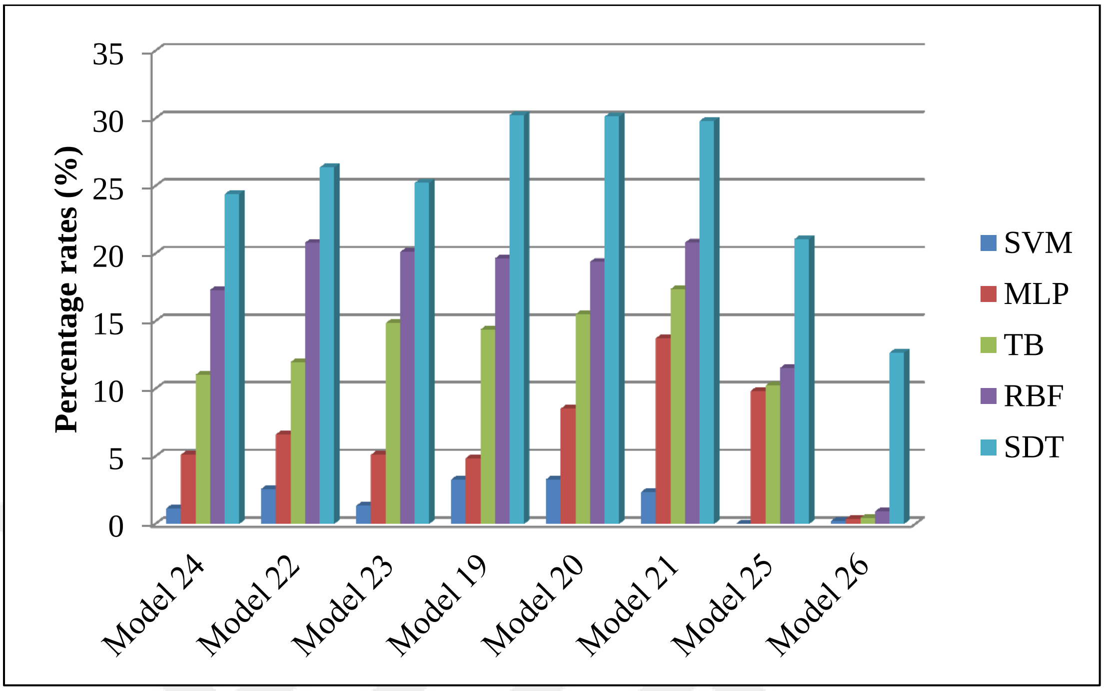


Figure 4.5. Percentage decrease rates in *SEE* values of $UBP_{10}$ for GRNN-based models compared to *SEE* values obtained by SVM, MLP, TB RBF and SDT-based models using Relief-F on UBP-set2

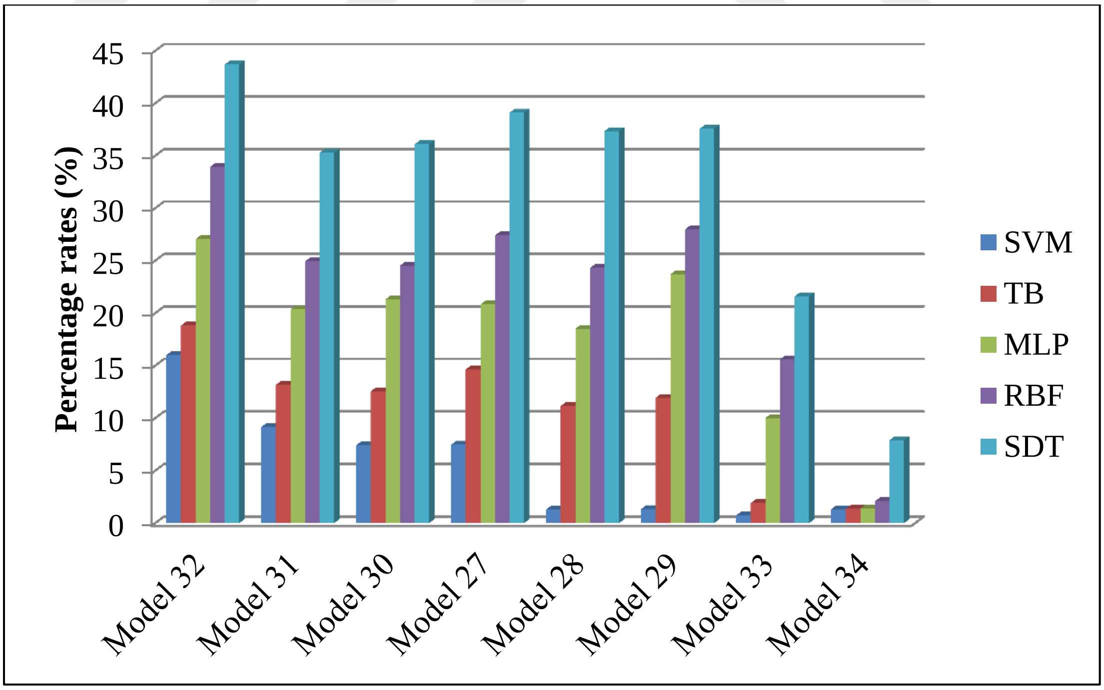


Figure 4.6. Percentage decrease rates in *SEE* values of $UBP_{60}$ for GRNN-based models compared to *SEE* values obtained by SVM, TB, MLP RBF and SDT-based models using Relief-F on UBP-set2

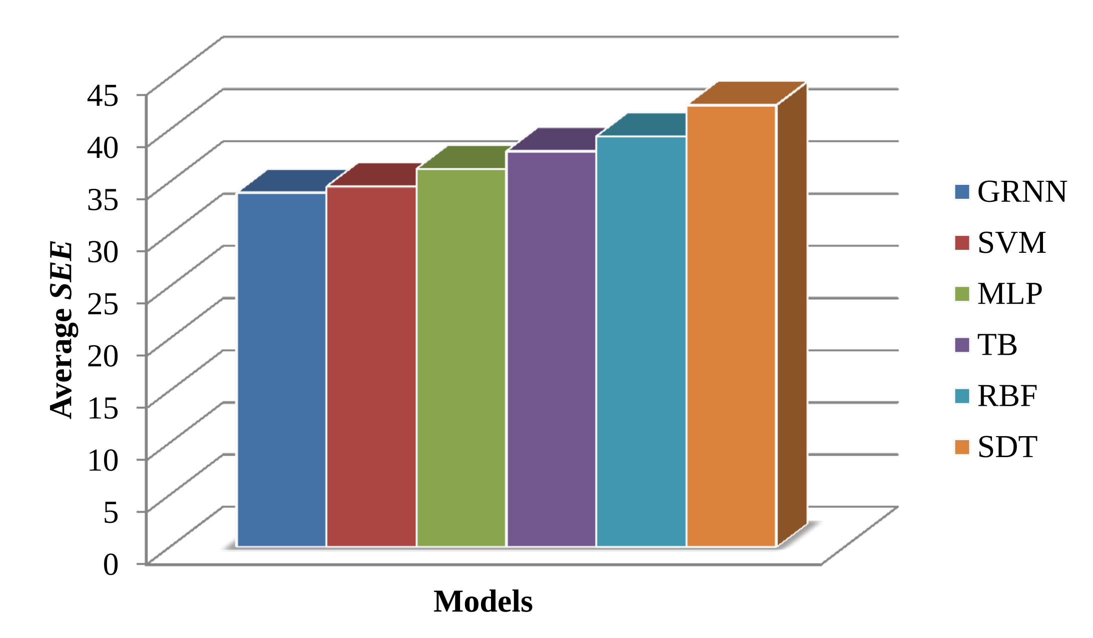


Figure 4.7. Average *SEE*'s of prediction models for $UBP_{10}$ using Relief-F on UBP-set2

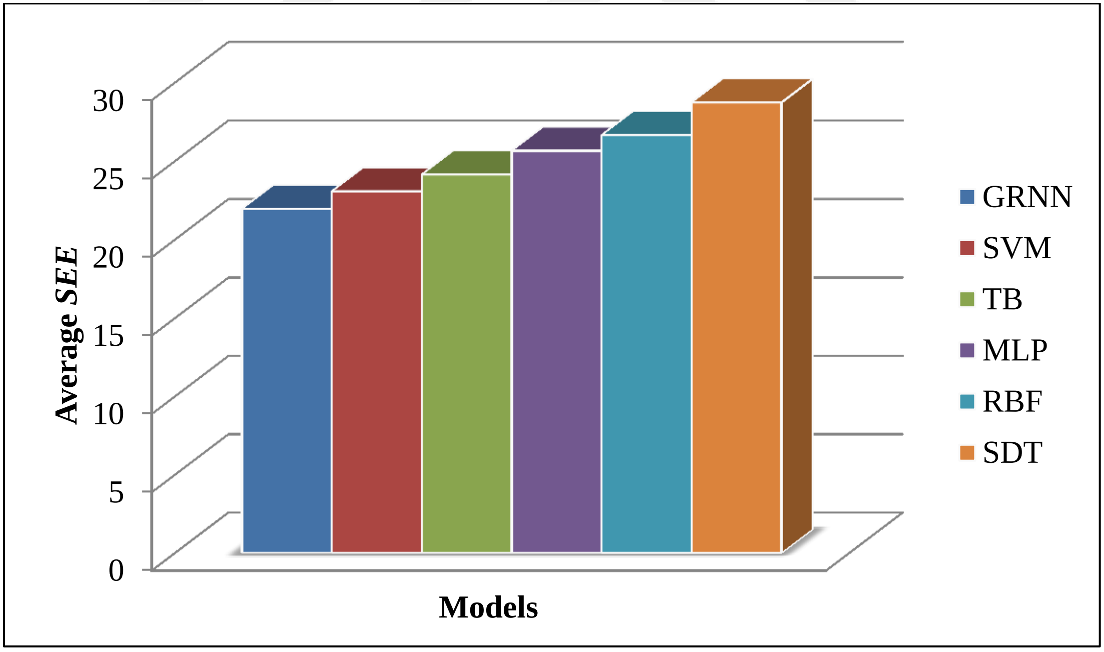


Figure 4.8. Average *SEE*'s of prediction models for $UBP_{60}$ using Relief-F on UBP-set2

- As compared to the *SEE*'s obtained by SDT, the maximum percentage decrement rates in *SEE*'s obtained by GRNN, SVM, MLP, TB and RBF are in average 24.99%, 22.75%, 17.09%, 11.64% and 7.55% for prediction of $UBP_{10}$, respectively.
- Also, as compared to the *SEE*'s obtained by SDT, the maximum percentage decrement rates in *SEE*'s obtained by GRNN, SVM, TB, MLP and RBF are in average 32.32%, 25.31%, 19.37%, 12.11% and 7.86% for prediction of $UBP_{60}$, in the order given.
- When all results are taken into consideration, Model 24 including the features gender, time, BMI and Model 25 including the features gender and time suggests that the feature BMI provides a significant improvement for prediction of $UBP_{10}$. In more detail, the inclusion of BMI in the aforementioned model leads in average 11.89% reduction in *SEE* for predicting $UBP_{10}$. However, when the *SEE*'s of the Model 32 including gender, BMI, age and the Model 33 including gender and BMI are compared, the addition of gender in the related model causes in average 17.60% decrement for prediction of $UBP_{60}$.
- In general, prediction models including the features gender, time and BMI lead to much lower *SEE's* than other models for $UBP_{10}$. Similarly, for $UBP_{60}$ prediction models including the features gender, BMI and age lead to much lower *SEE's* than other models, suggesting that UBP is highly correlated to these three features.

### 4.2. Results and Discussion for Prediction Models with mRMR

The *SEE*'s related to the prediction of both $UBP_{10}$ and $UBP_{60}$ for UBP-set1 are listed in ascending order Table 4.13 and Table 4.14, respectively.

Table 4.13. *SEE*'s for mRMR-based $UBP_{10}$ prediction models (UBP-set1)

| **Models** | **SEE** | | | | | |
|---|---|---|---|---|---|---|
| | **GRNN** | **SVM** | **MLP** | **TB** | **RBF** | **SDT** |
| Model 37 | 25.23 | 27.12 | 28.34 | 31.15 | 32.78 | 37.42 |
| Model 39 | 26.15 | 27.50 | 28.59 | 31.40 | 32.85 | 38.75 |
| Model 36 | 26.67 | 27.89 | 28.74 | 32.08 | 33.79 | 38.88 |
| Model 38 | 26.74 | 28.18 | 29.48 | 32.51 | 34.24 | 39.47 |
| Model 35 | 26.76 | 28.81 | 29.91 | 33.20 | 34.47 | 39.83 |
| Model 41 | 30.34 | 31.29 | 32.39 | 33.69 | 35.67 | 40.65 |
| Model 42 | 31.32 | 32.24 | 32.48 | 33.87 | 38.63 | 40.89 |
| Model 40 | 31.39 | 32.27 | 33.49 | 34.24 | 40.32 | 41.19 |
| Model 43 | 40.02 | 40.11 | 40.17 | 40.20 | 40.39 | 45.09 |

Table 4.14. *SEE*'s for mRMR-based $UBP_{60}$ prediction models (UBP-set1)

| **Models** | **SEE** | | | | | |
|---|---|---|---|---|---|---|
| | **GRNN** | **SVM** | **TB** | **MLP** | **RBF** | **SDT** |
| Model 46 | 17.10 | 18.01 | 19.52 | 20.24 | 21.05 | 24.59 |
| Model 44 | 17.62 | 20.19 | 20.69 | 22.49 | 23.94 | 29.75 |
| Model 45 | 17.86 | 20.21 | 20.84 | 22.84 | 24.18 | 29.82 |
| Model 49 | 17.92 | 20.35 | 20.93 | 22.93 | 24.85 | 29.91 |
| Model 48 | 18.21 | 20.39 | 21.02 | 23.42 | 25.10 | 30.10 |
| Model 47 | 18.93 | 20.54 | 21.26 | 23.85 | 25.17 | 30.48 |
| Model 50 | 19.93 | 20.69 | 21.47 | 23.90 | 25.24 | 30.59 |
| Model 51 | 24.62 | 24.66 | 25.05 | 25.25 | 28.20 | 30.75 |
| Model 52 | 28.86 | 29.24 | 29.25 | 29.26 | 29.46 | 31.14 |

The *R*'s related to the prediction of both $UBP_{10}$ and $UBP_{60}$ for UBP-set1 are listed in descending order Table 4.15 and Table 4.16, respectively.

Table 4.15. *R*'s for mRMR-based $UBP_{10}$ prediction models (UBP-set1)

| **Models** | **R** | | | | | |
|---|---|---|---|---|---|---|
| | **GRNN** | **SVM** | **MLP** | **TB** | **RBF** | **SDT** |
| Model 37 | 0.934 | 0.923 | 0.919 | 0.897 | 0.884 | 0.842 |
| Model 39 | 0.929 | 0.921 | 0.917 | 0.896 | 0.881 | 0.839 |
| Model 36 | 0.926 | 0.919 | 0.911 | 0.888 | 0.878 | 0.834 |
| Model 38 | 0.925 | 0.916 | 0.908 | 0.885 | 0.876 | 0.828 |
| Model 35 | 0.925 | 0.913 | 0.906 | 0.883 | 0.873 | 0.826 |
| Model 41 | 0.903 | 0.896 | 0.889 | 0.866 | 0.827 | 0.780 |
| Model 42 | 0.896 | 0.890 | 0.888 | 0.837 | 0.825 | 0.778 |
| Model 40 | 0.895 | 0.889 | 0.823 | 0.824 | 0.824 | 0.770 |
| Model 43 | 0.824 | 0.823 | 0.822 | 0.822 | 0.820 | 0.769 |

Table 4.16. *R*'s for mRMR-based $UBP_{60}$ prediction models (UBP-set1)

| **Models** | **R** | | | | | |
|---|---|---|---|---|---|---|
| | **GRNN** | **SVM** | **TB** | **MLP** | **RBF** | **SDT** |
| Model 46 | 0.959 | 0.942 | 0.932 | 0.926 | 0.920 | 0.889 |
| Model 44 | 0.945 | 0.926 | 0.923 | 0.908 | 0.895 | 0.832 |
| Model 45 | 0.952 | 0.937 | 0.930 | 0.924 | 0.902 | 0.875 |
| Model 49 | 0.951 | 0.932 | 0.929 | 0.916 | 0.886 | 0.873 |
| Model 48 | 0.943 | 0.931 | 0.928 | 0.915 | 0.884 | 0.860 |
| Model 47 | 0.936 | 0.926 | 0.924 | 0.908 | 0.883 | 0.833 |
| Model 50 | 0.929 | 0.923 | 0.917 | 0.895 | 0.882 | 0.822 |
| Model 51 | 0.888 | 0.887 | 0.884 | 0.882 | 0.859 | 0.820 |
| Model 52 | 0.843 | 0.838 | 0.838 | 0.838 | 0.836 | 0.814 |

The *MAPE*s related to the prediction of both $UBP_{10}$ and $UBP_{60}$ for UBP-set1 are listed in ascending order Table 4.17 and Table 4.18, respectively.

Table 4.17. *MAPE*'s for mRMR-based $UBP_{10}$ prediction models (UBP-set1)

| **Models** | **MAPE (%)** | | | | | |
|---|---|---|---|---|---|---|
| | **GRNN** | **SVM** | **MLP** | **TB** | **RBF** | **SDT** |
| Model 37 | 9.02 | 9.88 | 9.94 | 10.85 | 11.14 | 13.50 |
| Model 39 | 9.03 | 9.98 | 10.39 | 11.01 | 11.16 | 13.92 |
| Model 36 | 9.30 | 10.34 | 10.75 | 11.23 | 11.28 | 14.36 |
| Model 38 | 9.41 | 10.59 | 10.89 | 11.23 | 11.29 | 14.67 |
| Model 35 | 9.51 | 10.65 | 11.06 | 11.44 | 12.42 | 14.72 |
| Model 41 | 10.88 | 10.96 | 11.35 | 11.51 | 12.98 | 14.81 |
| Model 42 | 10.93 | 11.05 | 11.44 | 11.99 | 13.76 | 14.86 |
| Model 40 | 11.04 | 11.60 | 12.31 | 14.62 | 15.13 | 15.84 |
| Model 43 | 15.46 | 15.50 | 15.53 | 15.74 | 15.75 | 16.51 |

Table 4.18. *MAPE*'s for mRMR-based $UBP_{60}$ prediction models (UBP-set1)

| **Models** | **MAPE (%)** | | | | | |
|---|---|---|---|---|---|---|
| | **GRNN** | **SVM** | **TB** | **MLP** | **RBF** | **SDT** |
| Model 46 | 6.84 | 7.84 | 9.21 | 9.60 | 9.96 | 11.66 |
| Model 44 | 7.85 | 9.61 | 9.98 | 10.85 | 11.05 | 13.53 |
| Model 45 | 7.88 | 9.68 | 10.01 | 10.96 | 11.47 | 13.55 |
| Model 49 | 7.98 | 8.75 | 10.20 | 11.01 | 11.55 | 13.59 |
| Model 48 | 8.38 | 8.82 | 10.31 | 11.39 | 11.65 | 13.62 |
| Model 47 | 8.60 | 9.84 | 10.42 | 11.51 | 11.66 | 13.65 |
| Model 50 | 8.88 | 9.95 | 10.48 | 11.58 | 11.67 | 13.66 |
| Model 51 | 11.47 | 11.66 | 12.14 | 12.57 | 14.10 | 14.46 |
| Model 52 | 14.28 | 14.37 | 14.43 | 14.62 | 15.34 | 15.50 |

The following statements are valid for UBP prediction models developed using UBP-set1:

- With respect to the ranking of the mRMR algorithm, the features HRLT and time exhibit the highest scores for $UBP_{10}$ and $UBP_{60}$ predictions, respectively. On the other hand; the feature, having the minimum Relief-F score for both $UBP_{10}$ and $UBP_{60}$ prediction, is gender.
- The prediction models including gender always yield the highest *SEE*'s for $UBP_{10}$ and $UBP_{60}$.
- As compared to the *SEE*'s obtained by SVM, MLP, TB, RBF and SDT, the percentage decrement rates in SEE's obtained by GRNN are in average 4.37%, 7.66%, 15.35%, 23.26%, 38.57% for prediction of $UBP_{10}$ on UBP-set1, respectively.
- Additionally, compared to the *SEE*'s obtained by SVM, TB, MLP, RBF Network and SDT, the percentage decrement rates in SEE's obtained by GRNN are in average 8.04%, 11.51%,20.03%, 27.45%, 51.01% for prediction of $UBP_{60}$ on UBP-set1, in the order given.
- With regard to $UBP_{10}$ prediction models for GRNN, SVM, MLP, TB, RBF and SDT, adding the feature time to Model 40 yields an average decrement rate of 15.43% for UBP-set1. As for $UBP_{60}$ prediction models for GRNN, SVM, TB, MLP, RBF and SDT, adding the feature $VO_2$max to Model 50 yields an average decrement rate of 3.92% on UBP-set1.

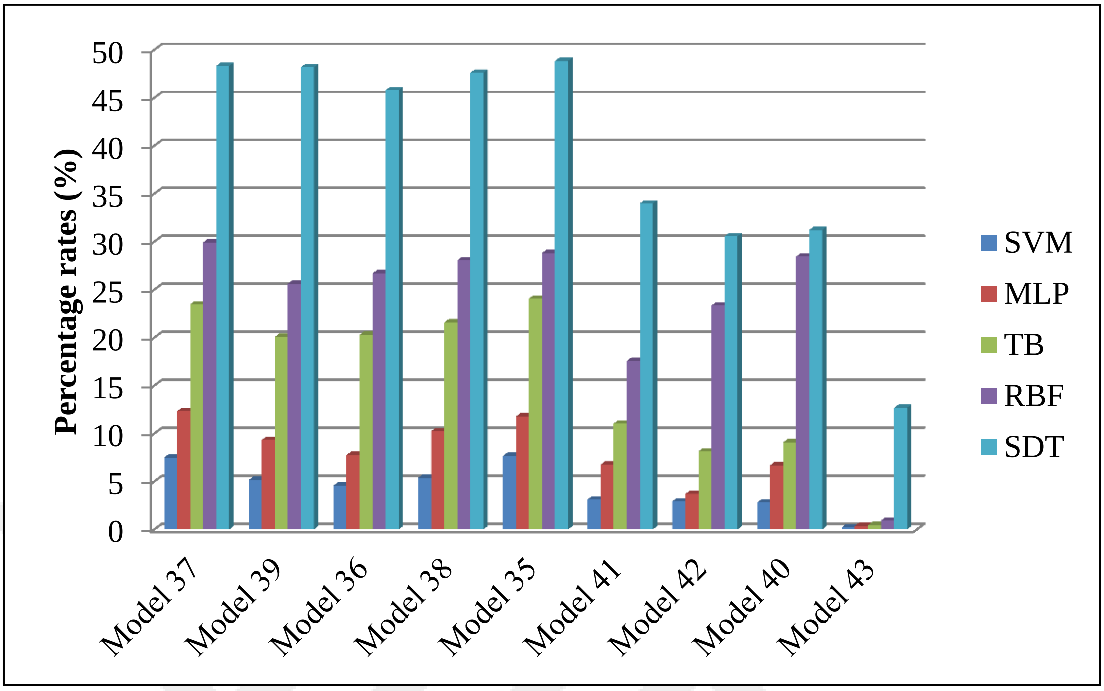


Figure 4.9. Percentage decrease rates in *SEE* values of $UBP_{10}$ for GRNN-based models compared to *SEE* values obtained by SVM, MLP, TB, RBF and SDT-based models using mRMR on UBP-set1

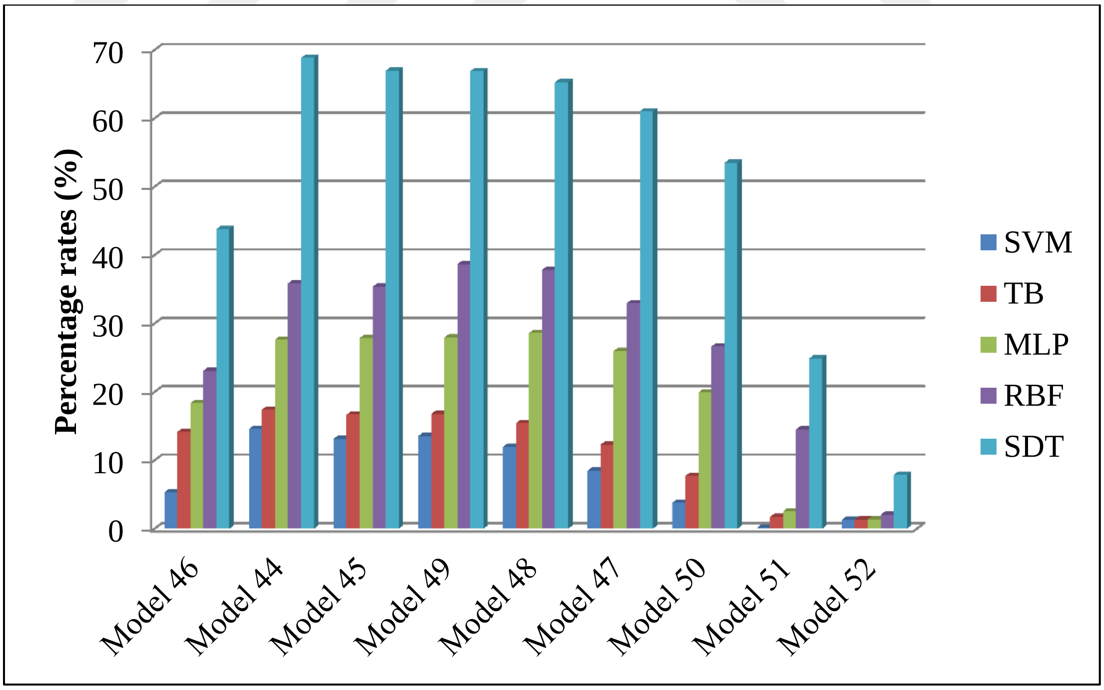


Figure 4.10. Percentage decrease rates in *SEE* values of $UBP_{60}$ for GRNN-based models compared to *SEE* values obtained by SVM, TB, MLP, RBF and SDT-based models using mRMR on UBP-set1

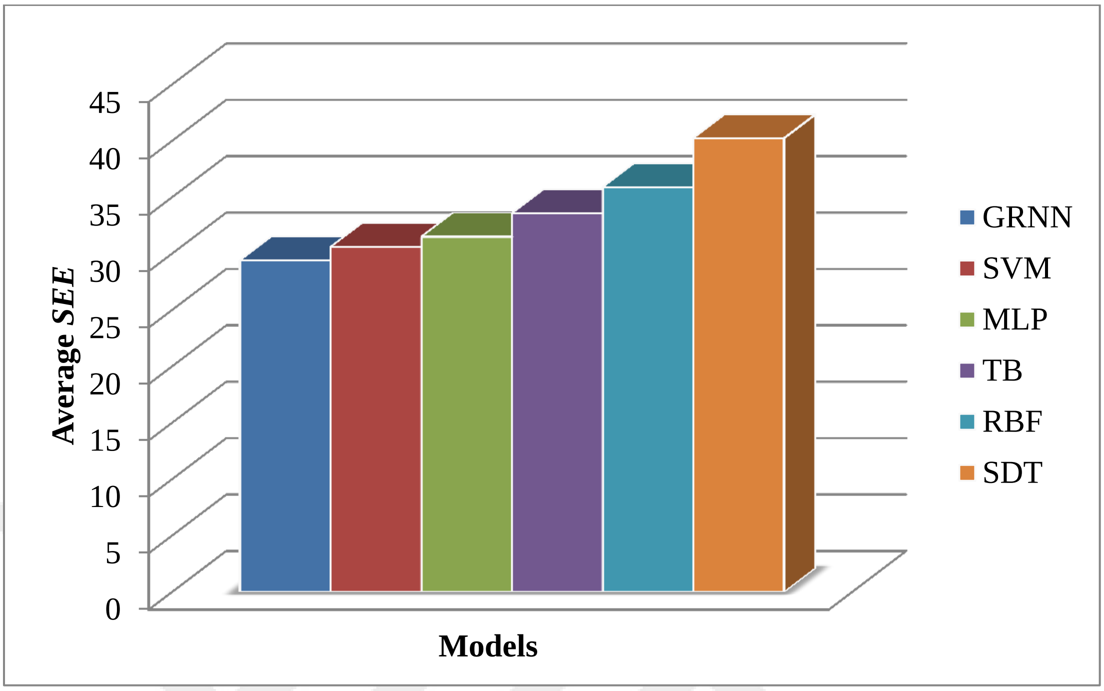


Figure 4.11. Average *SEE*'s of prediction models for $UBP_{10}$ using mRMR on UBP-set1

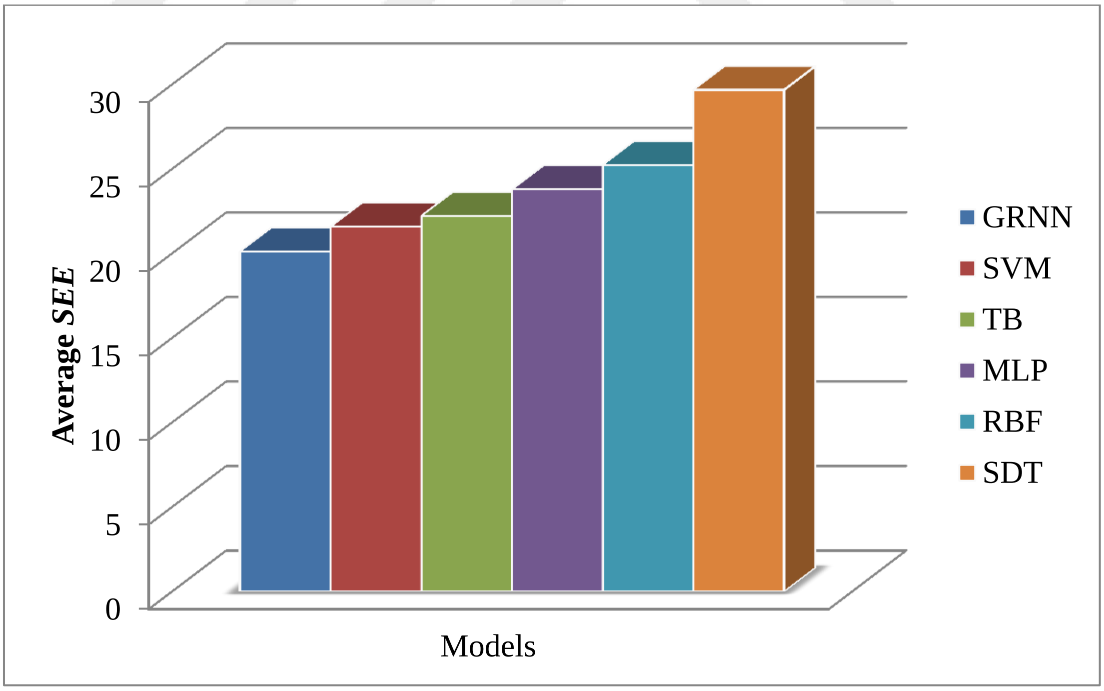


Figure 4.12. Average *SEE*'s of prediction models for $UBP_{60}$ using mRMR on UBP-set1

- The results show that $UBP_{10}$ and $UBP_{60}$ can be predicted with lower *SEE*'s and higher *R*'s by using less number of features instead of the full set of features. In particular, for UBP-set1, Model 37 and Model 46 yield up to 6.58% and 6.22% lower *SEE*'s than the ones comprising the full set of features.

In a similar fashion, the following statements are valid for UBP prediction models developed using UBP-set2:

- With respect to the ranking of the mRMR algorithm, the features protocol and time exhibit the highest scores for $UBP_{10}$ and $UBP_{60}$ predictions, respectively. On the other hand; the feature, having the minimum Relief-F score for both $UBP_{10}$ and $UBP_{60}$ prediction, is gender.
- The results show that $UBP_{10}$ and $UBP_{60}$ can be predicted with lower *SEE*'s and higher *R*'s by using less number of features instead of the full set of features. In particular, Model 56 and Model 66 yield up to 6.10% and 10.35% lower *SEE*'s than the ones including the full set of features for UBP-set2.
- As compared to the *SEE*'s obtained by SVM, MLP, TB, RBF and SDT, the percentage decrement rates in *SEE*'s obtained by GRNN are in average 1.84%, 7.97%,12.55%, 15.82% and 25.20% for prediction of $UBP_{10}$ on UBP-set2, respectively. Also, as compared to the *SEE*'s obtained by SVM, TB, MLP, RBF and SDT, the percentage decrement rates in *SEE*'s obtained by GRNN are in average 6.93%, 11.56%, 16.45%, 23.59% and 33.31% for prediction of $UBP_{60}$ on UBP-set2, in the order given.
- With regard to $UBP_{10}$ prediction models for GRNN, SVM, MLP, TB, RBF and SDT, adding the feature time to Model 57 yields an average decrement rate of 8.50% on UBP-set2. As for $UBP_{60}$ prediction models for GRNN, SVM, TB, MLP, RBF and SDT, adding the feature BMI to Model 67 yields an average decrement rate of 16.80% on UBP-set2.

The *SEE*'s related to the prediction of both $UBP_{10}$ and $UBP_{60}$ for UBP-set2 are listed in ascending order Table 4.19 and Table 4.20, respectively.

Table 4.19. *SEE*'s for mRMR-based $UBP_{10}$ prediction models (UBP-set2)

| **Models** | **SEE** | | | | | |
|---|---|---|---|---|---|---|
| | **GRNN** | **SVM** | **MLP** | **TB** | **RBF** | **SDT** |
| Model 56 | 31.88 | 32.84 | 33.56 | 35.84 | 39.14 | 40.30 |
| Model 53 | 32.83 | 33.91 | 34.43 | 37.55 | 39.28 | 42.80 |
| Model 54 | 32.96 | 34.05 | 35.77 | 38.07 | 39.35 | 42.90 |
| Model 55 | 33.35 | 34.14 | 37.93 | 39.14 | 40.29 | 43.30 |
| Model 57 | 34.71 | 35.08 | 38.14 | 39.85 | 40.32 | 43.31 |
| Model 59 | 35.10 | 35.49 | 38.87 | 40.04 | 40.35 | 43.37 |
| Model 58 | 36.07 | 36.20 | 39.82 | 40.18 | 40.37 | 44.73 |
| Model 60 | 40.02 | 40.11 | 40.17 | 40.20 | 40.39 | 45.09 |

Table 4.20. *SEE*'s for mRMR-based $UBP_{60}$ prediction models (UBP-set2)

| **Models** | **SEE** | | | | | |
|---|---|---|---|---|---|---|
| | **GRNN** | **SVM** | **MLP** | **TB** | **RBF** | **SDT** |
| Model 66 | 18.45 | 21.40 | 21.92 | 23.44 | 24.71 | 26.50 |
| Model 64 | 19.35 | 21.50 | 22.47 | 23.74 | 24.75 | 26.65 |
| Model 65 | 19.90 | 21.72 | 22.52 | 23.95 | 24.86 | 26.90 |
| Model 61 | 20.36 | 21.88 | 23.34 | 24.60 | 25.94 | 28.30 |
| Model 62 | 20.84 | 21.94 | 23.78 | 24.79 | 27.09 | 28.94 |
| Model 63 | 21.35 | 22.41 | 24.02 | 25.14 | 27.31 | 29.68 |
| Model 67 | 24.62 | 24.66 | 25.05 | 25.25 | 28.20 | 30.75 |
| Model 68 | 28.86 | 29.24 | 29.25 | 29.26 | 29.46 | 31.14 |

The R's related to the prediction of both $UBP_{10}$ and $UBP_{60}$ for UBP-set2 are listed in descending order Table 4.21 and Table 4.22, respectively.

Table 4.21. *R*'s for mRMR-based $UBP_{10}$ prediction models (UBP-set2)

| **Models** | **R** | | | | | |
|---|---|---|---|---|---|---|
| | **GRNN** | **SVM** | **MLP** | **TB** | **RBF** | **SDT** |
| Model 56 | 0.892 | 0.885 | 0.879 | 0.861 | 0.832 | 0.820 |
| Model 53 | 0.885 | 0.877 | 0.873 | 0.847 | 0.831 | 0.796 |
| Model 54 | 0.884 | 0.876 | 0.862 | 0.842 | 0.830 | 0.794 |
| Model 55 | 0.881 | 0.875 | 0.843 | 0.832 | 0.821 | 0.790 |
| Model 57 | 0.871 | 0.868 | 0.841 | 0.826 | 0.821 | 0.788 |
| Model 59 | 0.868 | 0.864 | 0.840 | 0.825 | 0.820 | 0.777 |
| Model 58 | 0.860 | 0.859 | 0.827 | 0.823 | 0.820 | 0.773 |
| Model 60 | 0.824 | 0.823 | 0.822 | 0.822 | 0.820 | 0.769 |

Table 4.22. *R*'s for mRMR-based $UBP_{60}$ prediction models (UBP-set2)

| **Models** | **R** | | | | | |
|---|---|---|---|---|---|---|
| | **GRNN** | **SVM** | **MLP** | **TB** | **RBF** | **SDT** |
| Model 66 | 0.939 | 0.917 | 0.913 | 0.900 | 0.888 | 0.869 |
| Model 64 | 0.933 | 0.916 | 0.908 | 0.897 | 0.887 | 0.868 |
| Model 65 | 0.929 | 0.914 | 0.908 | 0.895 | 0.886 | 0.865 |
| Model 61 | 0.925 | 0.913 | 0.900 | 0.889 | 0.875 | 0.849 |
| Model 62 | 0.922 | 0.912 | 0.896 | 0.887 | 0.863 | 0.847 |
| Model 63 | 0.921 | 0.909 | 0.894 | 0.885 | 0.861 | 0.833 |
| Model 67 | 0.888 | 0.887 | 0.884 | 0.882 | 0.859 | 0.820 |
| Model 68 | 0.843 | 0.838 | 0.838 | 0.838 | 0.836 | 0.814 |

The *MAPE*s related to the prediction of both $UBP_{10}$ and $UBP_{60}$ for UBP-set2 are listed in ascending order Table 4.23 and Table 4.24, respectively.

Table 4.23. *MAPE*'s for mRMR-based $UBP_{10}$ prediction models (UBP-set2)

| **Models** | **MAPE (%)** | | | | | |
|---|---|---|---|---|---|---|
| | **GRNN** | **SVM** | **MLP** | **TB** | **RBF** | **SDT** |
| Model 56 | 10.96 | 11.58 | 12.16 | 12.72 | 14.38 | 14.39 |
| Model 53 | 11.42 | 12.36 | 12.48 | 13.82 | 14.45 | 15.44 |
| Model 54 | 12.09 | 12.43 | 12.62 | 13.87 | 14.47 | 15.55 |
| Model 55 | 12.14 | 12.44 | 12.87 | 14.67 | 14.67 | 15.55 |
| Model 57 | 12.16 | 12.45 | 13.42 | 14.68 | 14.92 | 15.55 |
| Model 59 | 12.22 | 13.25 | 13.81 | 14.92 | 14.96 | 15.55 |
| Model 58 | 12.45 | 13.97 | 15.20 | 15.46 | 15.53 | 15.55 |
| Model 60 | 15.46 | 15.50 | 15.53 | 15.74 | 15.75 | 16.51 |

Table 4.24. *MAPE*'s for mRMR-based $UBP_{60}$ prediction models (UBP-set2)

| **Models** | **MAPE (%)** | | | | | |
|---|---|---|---|---|---|---|
| | **GRNN** | **SVM** | **TB** | **MLP** | **RBF** | **SDT** |
| Model 66 | 8.54 | 9.59 | 9.86 | 11.24 | 11.26 | 11.26 |
| Model 64 | 8.83 | 9.98 | 10.12 | 11.63 | 11.67 | 12.40 |
| Model 65 | 8.86 | 10.29 | 10.36 | 11.65 | 12.64 | 12.92 |
| Model 61 | 8.92 | 10.39 | 11.12 | 11.68 | 12.81 | 13.29 |
| Model 62 | 9.99 | 10.52 | 11.16 | 11.86 | 13.26 | 13.89 |
| Model 63 | 10.04 | 10.53 | 11.37 | 12.07 | 13.89 | 14.36 |
| Model 67 | 11.47 | 11.66 | 12.14 | 12.57 | 14.10 | 14.46 |
| Model 68 | 14.28 | 14.37 | 14.43 | 14.62 | 15.34 | 15.50 |

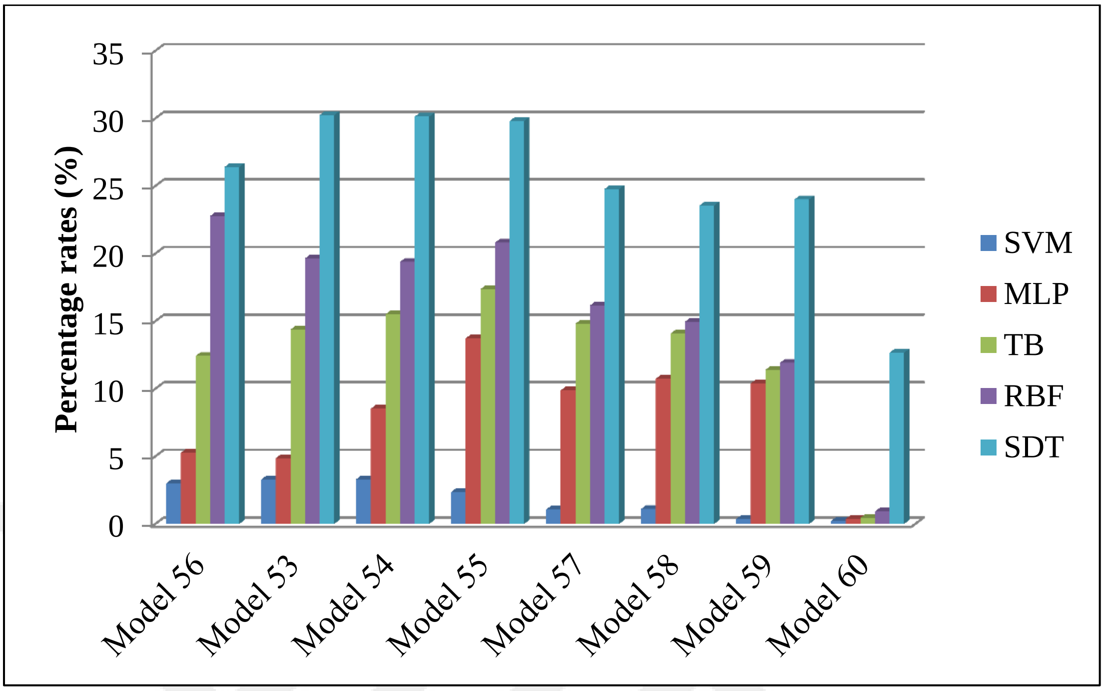


Figure 4.13. Percentage decrease rates in *SEE* values of $UBP_{10}$ for GRNN-based models compared to *SEE* values obtained by SVM, MLP, TB, RBF and SDT-based models using mRMR on UBP-set2

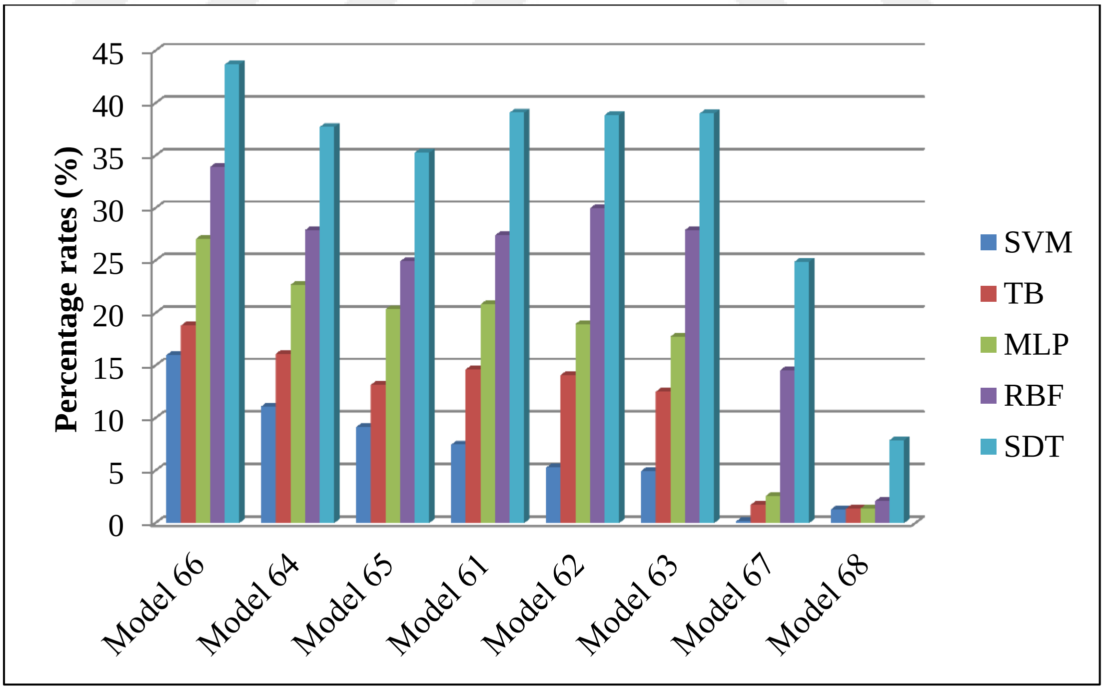


Figure 4.14. Percentage decrease rates in *SEE* values of $UBP_{60}$ for GRNN-based models compared to *SEE* values obtained by SVM, TB, MLP, RBF and SDT-based models using mRMR on UBP-set2

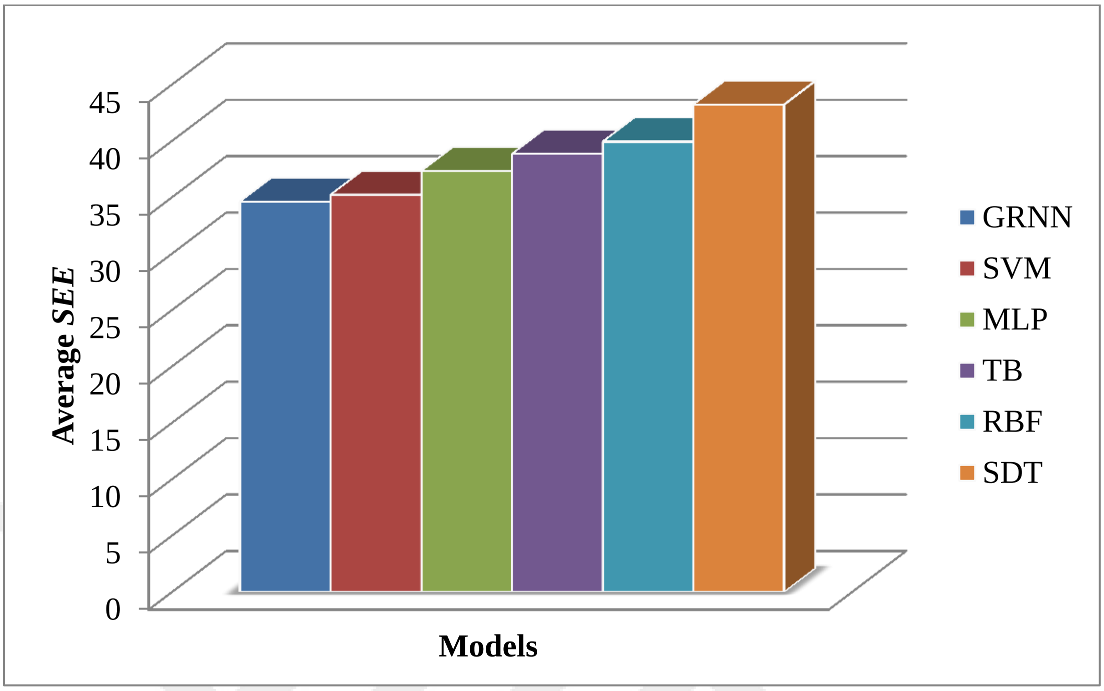


Figure 4.15. Average *SEE*'s of prediction models for $UBP_{10}$ using mRMR on UBP-set2

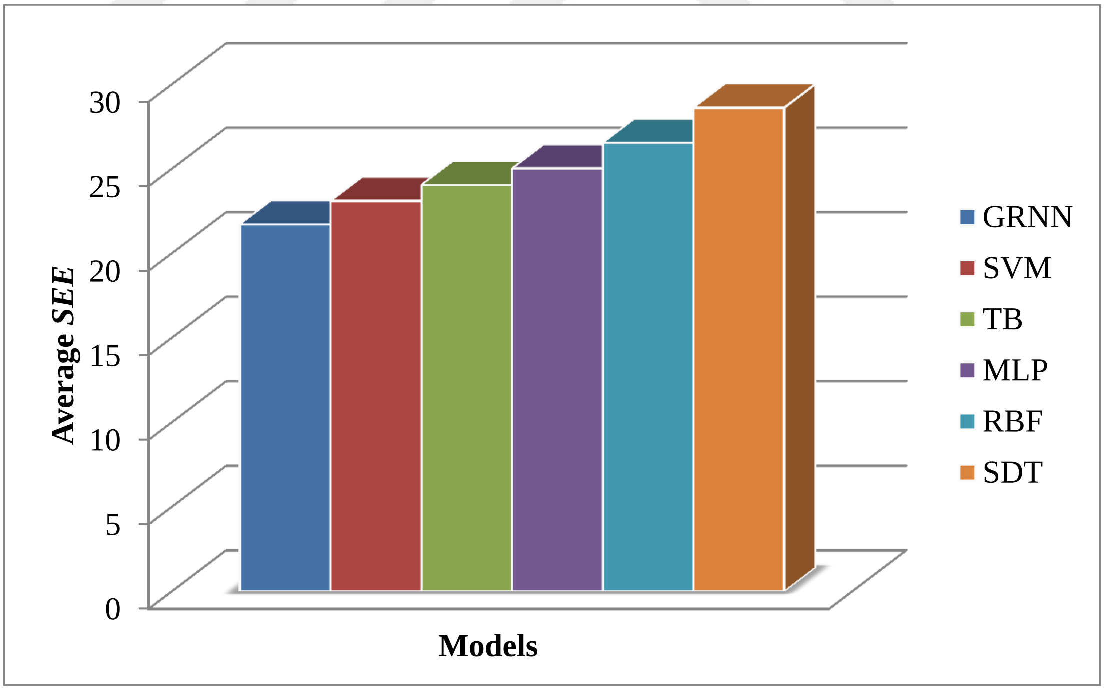


Figure 4.16. Average *SEE*'s of prediction models for $UBP_{60}$ using mRMR on UBP-set2

### 4.3. Results and Discussion for Prediction Models with CFS

By using the CFS feature selection algorithm, nine prediction models have been created for both data sets. In contrast to Relief-F and mRMR algorithms, the CFS algorithm which is combined with diffirent search methods such as BestFirst, GeneticSearch, GreedyStepwise and RandomSearch gives a set of selected variables to develop four different models for UBP-set1 and five different models for UBP-set2.

The BestFirst search method has produced Model 69 that includes gender, weight, height, $VO_2$max and time for $UBP_{10}$ on UBP-set1. The GreedyStepwise has created Model 70 comprising of gender, weight, height, HRmax, HRLT and time for $UBP_{10}$ on UBP-set1. Also, the BestFirst search method has produced Model 71 consisting of protocol, gender, weight, height, $VO_2$max and time for $UBP_{60}$ on UBP-set1. The GreedyStepwise has developed Model 72 including gender, weight, height, HRmax, HRLT and time for $UBP_{60}$ on UBP-set1.

The *SEE*'s related to the prediction of both $UBP_{10}$ and $UBP_{60}$ for UBP-set1 are listed in ascending order Table 4.25 and Table 4.26, respectively.

.Table 4.25. *SEE*’s for CFS-based $UBP_{10}$ prediction models (UBP-set1)

| **Models** | **SEE** | | | | | |
|---|---|---|---|---|---|---|
| | **GRNN** | **SVM** | **MLP** | **TB** | **RBF** | **SDT** |
| Model 69 | 26.15 | 27.50 | 28.59 | 31.40 | 32.85 | 38.75 |
| Model 70 | 29.60 | 31.57 | 34.72 | 35.44 | 38.82 | 46.88 |

Table 4.26. *SEE*’s for CFS-based $UBP_{60}$ prediction models (UBP-set1)

| **Models** | **SEE** | | | | | |
|---|---|---|---|---|---|---|
| | **GRNN** | **SVM** | **TB** | **MLP** | **RBF** | **SDT** |
| Model 71 | 18.16 | 20.96 | 21.92 | 22.17 | 29.97 | 30.75 |
| Model 72 | 22.83 | 23.96 | 25.82 | 30.68 | 33.78 | 37.01 |

The *R*'s related to the prediction of both $UBP_{10}$ and $UBP_{60}$ for UBP-set1 are listed in descending order Table 4.27 and Table 4.28, respectively.

Table 4.27. *R*'s for CFS-based $UBP_{10}$ prediction models (UBP-set1)

| Models | R | | | | | |
|---|---|---|---|---|---|---|
| | **GRNN** | **SVM** | **MLP** | **TB** | **RBF** | **SDT** |
| Model 69 | 0.929 | 0.921 | 0.917 | 0.896 | 0.881 | 0.839 |
| Model 70 | 0.908 | 0.894 | 0.871 | 0.865 | 0.835 | 0.748 |

Table 4.28. *R*'s for CFS-based $UBP_{60}$ prediction models (UBP-set1)

| Models | R | | | | | |
|---|---|---|---|---|---|---|
| | **GRNN** | **SVM** | **TB** | **MLP** | **RBF** | **SDT** |
| Model 71 | 0.941 | 0.921 | 0.913 | 0.911 | 0.829 | 0.819 |
| Model 72 | 0.905 | 0.895 | 0.877 | 0.821 | 0.777 | 0.724 |

The *MAPE*'s related to the prediction of both $UBP_{10}$ and $UBP_{60}$ for UBP-set1 are listed in ascending order Table 4.29 and Table 4.30, respectively.

Table 4.29. *MAPE*'s for CFS-based $UBP_{10}$ prediction models (UBP-set1)

| Models | MAPE (%) | | | | | |
|---|---|---|---|---|---|---|
| | **GRNN** | **SVM** | **MLP** | **TB** | **RBF** | **SDT** |
| Model 69 | 9.03 | 9.98 | 10.39 | 11.01 | 11.16 | 13.92 |
| Model 70 | 11.03 | 11.29 | 11.75 | 11.90 | 12.45 | 13.96 |

Table 4.30. *MAPE*'s for CFS-based $UBP_{60}$ prediction models (UBP-set1)

| Models | MAPE (%) | | | | | |
|---|---|---|---|---|---|---|
| | **GRNN** | **SVM** | **TB** | **MLP** | **RBF** | **SDT** |
| Model 71 | 9.51 | 9.83 | 10.06 | 10.41 | 12.38 | 14.29 |
| Model 72 | 10.10 | 10.53 | 10.68 | 12.23 | 13.88 | 14.15 |

The following statements are valid for UBP prediction models developed using UBP-set1.

- The *SEE*'s of $UBP_{60}$ prediction models are in average 26.50% lower than the *SEE*'s of $UBP_{10}$ prediction models.
- The features protocol and age were eliminated by using either the BestFirst or GreedyStepwise search method for $UBP_{10}$.
- Samely, the feature age was eliminated by using both the BestFirst and GreedyStepwise search method for $UBP_{60}$.
- Model 69 including the features gender, weight, height, $VO_2$max and time yields lower *SEE's* than Model 70 for prediction of $UBP_{10}$. In particular, the *SEE* of Model 69 is in average 17.16% lower than the *SEE* of Model 70 for $UBP_{10}$.
- Model 71 comprising of the features protocol, gender, weight, height, $VO_2$max and time leads to lower *SEE's* than Model 72 for prediction of $UBP_{10}$. In particular, the *SEE* of Model 72 is in average 20.95% lower than the *SEE* of Model 71 for $UBP_{60}$.
- As compared to the *SEE*'s obtained by SVM, MLP, TB, RBF and SDT, the percentage decrement rates in SEE's obtained by GRNN are in average 5.91%, 13.31%,19.90%, 28.39% and 53.28% for prediction of $UBP_{10}$ on UBP-set1, respectively. Also, as compared to the *SEE*'s obtained by SVM, TB, MLP, RBF and SDT, the percentage decrement rates in *SEE*'s obtained by GRNN are in average 10.18%, 16.90%, 28.23%, 56.50% and 65.72% for prediction of $UBP_{60}$ on UBP-set1, in the order given.

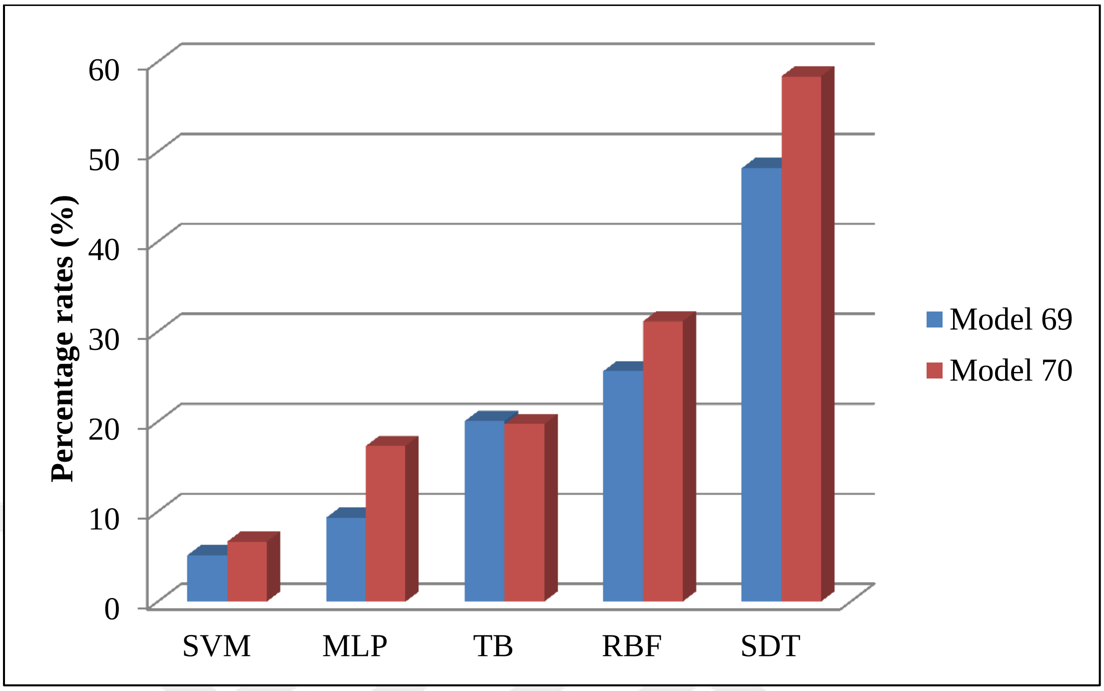


Figure 4.17. Percentage decrease rates in *SEE* values of $UBP_{10}$ GRNN-based models compared to *SEE* values obtained by for SVM, MLP, TB, RBF and SDT-based models using CFS on UBP-set1

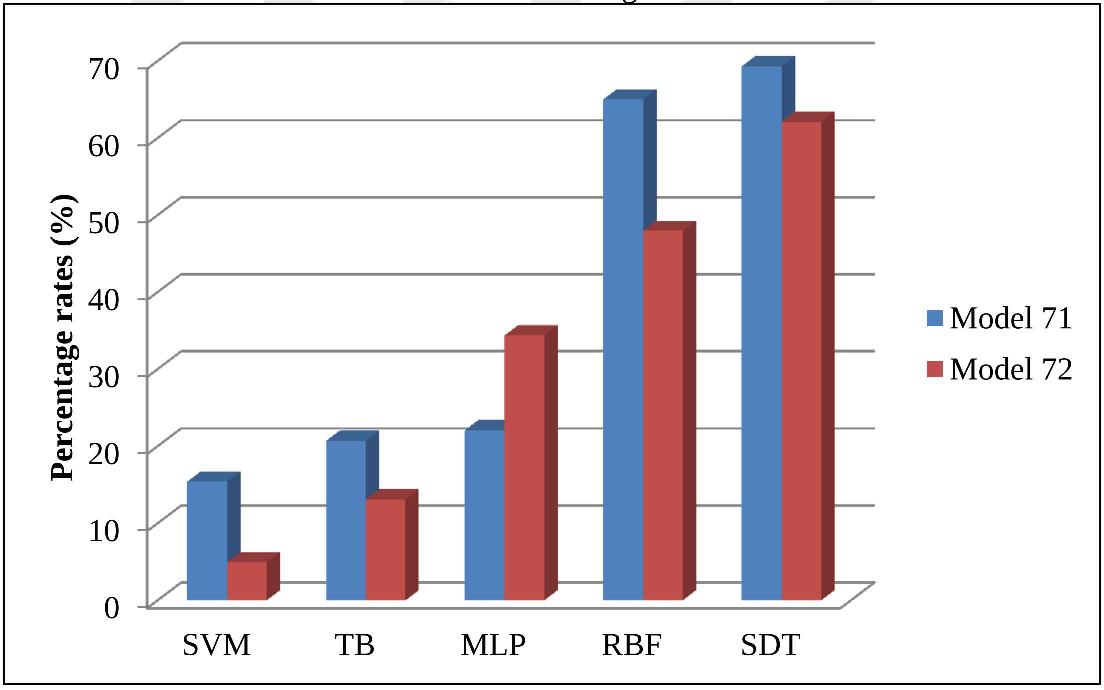


Figure 4.18. Percentage decrease rates in *SEE* values of $UBP_{60}$ for GRNN-based models compared to *SEE* values obtained by SVM, TB, MLP, RBF and SDT-based models using CFS on UBP-set1

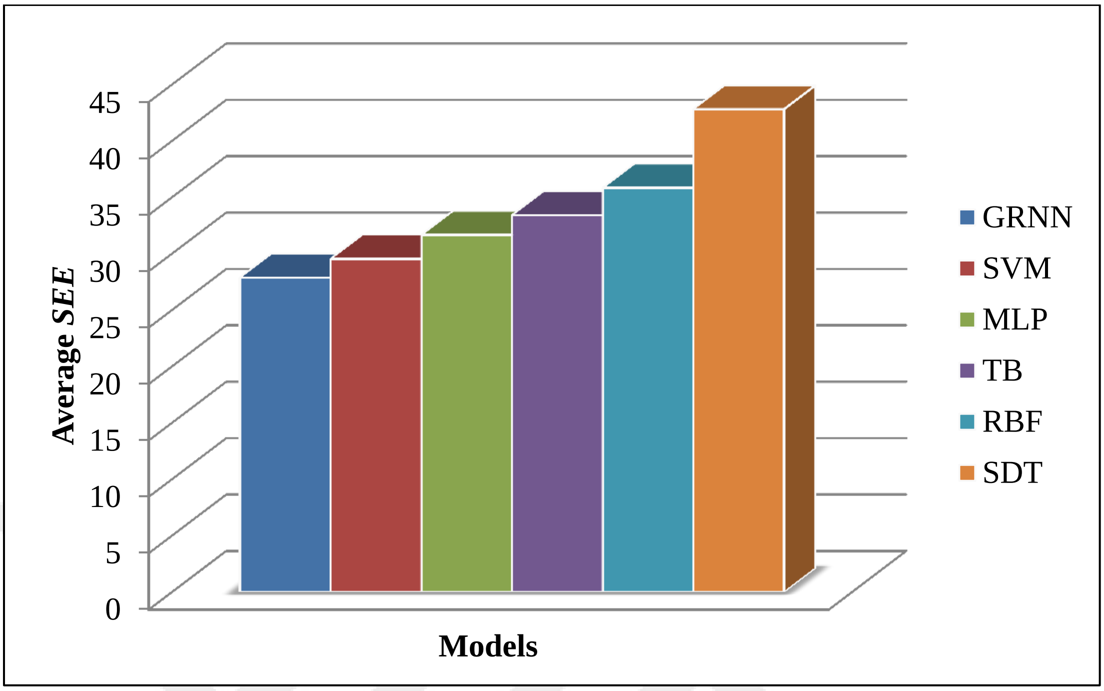


Figure 4.19. Average *SEE*'s of prediction models for $UBP_{10}$ using CFS on UBP-set1

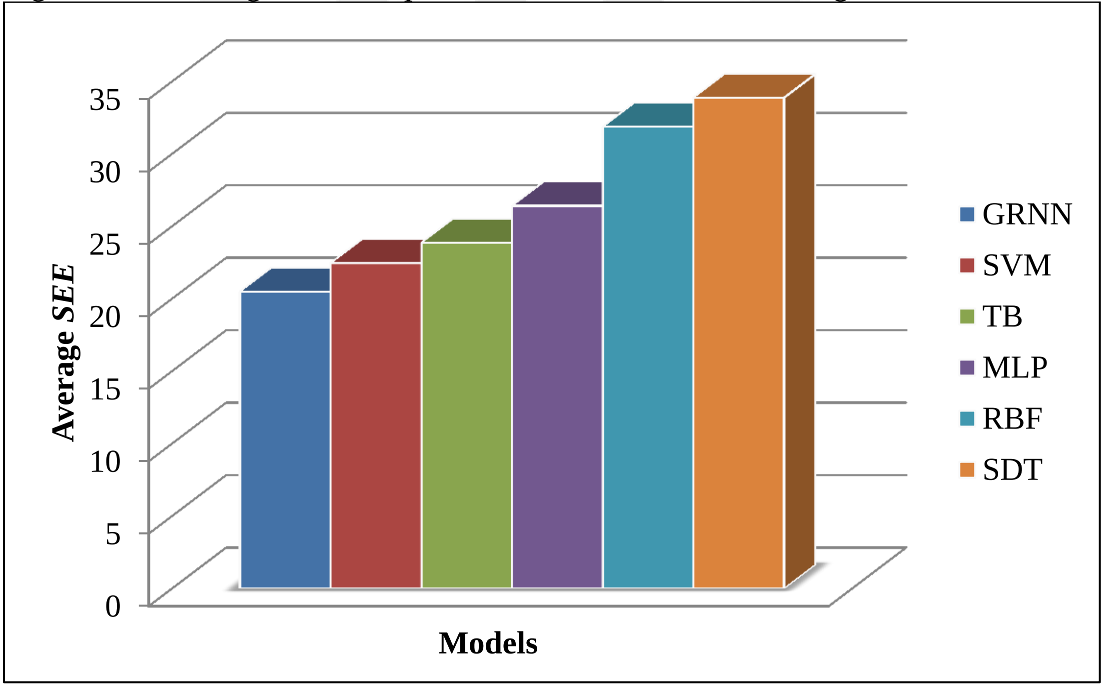


Figure 4.20. Average *SEE*'s of prediction models for $UBP_{60}$ using CFS on UBP-set1

- Upon observation of Figure 4.19 and Figure 4.20, one can claim that the average *SEE* of GRNN based models is lower than the ones obtained by other regression methods for both $UBP_{10}$ and $UBP_{60}$.

In a similar fashion, the following statements are valid for UBP prediction models developed using UBP-set2:

The GeneticSearch method has created Model 73 consisting of the features gender, BMI, $VO_2$max, HRLT and time for prediction of $UBP_{10}$ on UBP-set2. Additionally, the BestFirst search method has produced Model 74 including the features gender, BMI, HRLT and time for prediction of $UBP_{10}$ on UBP-set2.

The GeneticSearch method has developed Model 75 including the features protocol, gender, $VO_2$max, HRmax, HRLT and time for prediction of $UBP_{60}$ on UBP-set2. Also, the RandomSearch method has created Model 76 comprising of the features protocol, gender, $VO_2$max, HRLT and time for prediction of $UBP_{60}$ on UBP-set2. The BestFirst search method has developed Model 77 consisting of the features gender, HRLT, HRmax and time for prediction of $UBP_{60}$ on UBP-set2.

The *SEE*'s related to the prediction of both $UBP_{10}$ and $UBP_{60}$ for UBP-set1 are listed in ascending order Table 4.31 and Table 4.32, respectively.

Table 4.31. *SEE*'s for CFS-based $UBP_{10}$ prediction models (UBP-set2)

| **Models** | **SEE** | | | | | |
|---|---|---|---|---|---|---|
| | **GRNN** | **SVM** | **MLP** | **TB** | **RBF** | **SDT** |
| Model 73 | 31.88 | 32.84 | 33.56 | 35.84 | 39.14 | 40.30 |
| Model 74 | 32.67 | 33.11 | 34.34 | 37.52 | 39.24 | 40.91 |

Table 4.32. *SEE*'s for CFS-based $UBP_{60}$ prediction models (UBP-set2)

| **Models** | **SEE** | | | | | |
|---|---|---|---|---|---|---|
| | **GRNN** | **SVM** | **TB** | **MLP** | **RBF** | **SDT** |
| Model 75 | 27.07 | 27.59 | 28.46 | 28.57 | 29.52 | 30.83 |
| Model 76 | 27.22 | 27.71 | 28.57 | 28.75 | 30.28 | 34.01 |
| Model 77 | 28.09 | 29.26 | 29.35 | 30.32 | 32.38 | 35.17 |

The *R*'s related to the prediction of both $UBP_{10}$ and $UBP_{60}$ for UBP-set1 are listed in descending order Table 4.33 and Table 4.34, respectively.

Table 4.33. *R*'s for CFS-based $UBP_{10}$ prediction models (UBP-set2)

| **Models** | **R** | | | | | |
|---|---|---|---|---|---|---|
| | **GRNN** | **SVM** | **MLP** | **TB** | **RBF** | **SDT** |
| Model 73 | 0.892 | 0.885 | 0.879 | 0.861 | 0.832 | 0.820 |
| Model 74 | 0.886 | 0.883 | 0.874 | 0.847 | 0.831 | 0.815 |

Table 4.34. *R*'s for CFS-based $UBP_{60}$ prediction models (UBP-set2)

| **Models** | **R** | | | | | |
|---|---|---|---|---|---|---|
| | **GRNN** | **SVM** | **TB** | **MLP** | **RBF** | **SDT** |
| Model 75 | 0.863 | 0.858 | 0.848 | 0.846 | 0.835 | 0.818 |
| Model 76 | 0.862 | 0.856 | 0.846 | 0.844 | 0.826 | 0.773 |
| Model 77 | 0.837 | 0.852 | 0.825 | 0.838 | 0.797 | 0.755 |

The *MAPE*'s related to the prediction of both $UBP_{10}$ and $UBP_{60}$ for UBP-set1 are listed in ascending order Table 4.35 and Table 4.36, respectively.

Table 4.35. *MAPE*'s for CFS-based $UBP_{10}$ prediction models (UBP-set2)

| **Models** | **MAPE (%)** | | | | | |
|---|---|---|---|---|---|---|
| | **GRNN** | **SVM** | **MLP** | **TB** | **RBF** | **SDT** |
| Model 73 | 10.96 | 11.58 | 12.16 | 12.72 | 14.38 | 14.39 |
| Model 74 | 11.42 | 12.31 | 12.40 | 13.81 | 14.45 | 15.39 |

Table 4.36. *MAPE*'s for CFS-based $UBP_{60}$ prediction models (UBP-set2)

| **Models** | **MAPE (%)** | | | | | |
|---|---|---|---|---|---|---|
| | **GRNN** | **SVM** | **TB** | **MLP** | **RBF** | **SDT** |
| Model 75 | 13.64 | 13.98 | 14.11 | 14.58 | 15.71 | 15.29 |
| Model 76 | 14.24 | 14.40 | 14.42 | 15.49 | 16.98 | 15.77 |
| Model 77 | 14.43 | 14.80 | 15.12 | 15.61 | 16.98 | 16.89 |

- The *SEE's* of $UBP_{60}$ prediction models are in average 21.36% lower than the *SEE's* of $UBP_{10}$ prediction models.
- The features protocol, age and HRmax were eliminated by using either the GeneticSearch or BestFirst search method for $UBP_{10}$.
- The features age and BMI were eliminated regardless of which search methods are utilized for $UBP_{60}$.
- The prediction model comprising of the features protocol, gender, $VO_2$max, HRmax, HRLT and time (Model 73) yields the lowest *SEE*'s and highest *R's* for prediction of $UBP_{10}$.
- Model 75 comprising of the features protocol, gender, $VO_2$max, HRmax, HRLT and time leads to the lowest *SEE's* than the ones for prediction of $UBP_{60}$. In particular, the *SEE*s'of Model 75 are in average 2.62% and 7.29% lower than the ones for $UBP_{60,}$ resperctively.
- As compared to the *SEE*'s obtained by SVM, MLP, TB, RBF and SDT, the percentage decrement rates in SEE's obtained by GRNN are in average 2.18%, 5.19%,13.63%, 21.44% and 25.82% for prediction of $UBP_{10}$ on UBP-set2, in the order given. Also, as compared to the *SEE*'s obtained by SVM, TB, MLP, RBF and SDT, the percentage decrement rates in *SEE*'s obtained by GRNN are in average 2.63%, 4.86%, 6.37%, 11.85% and 21.35% for prediction of $UBP_{60}$ on UBP-set2, respectively.

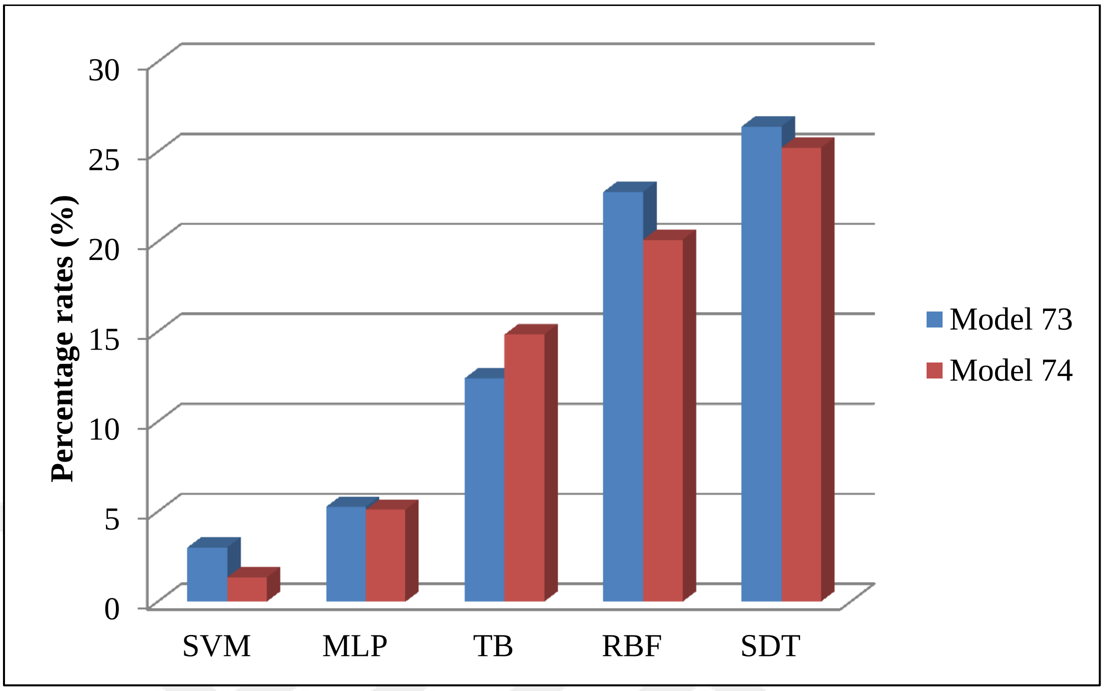


Figure 4.21. Percentage decrease rates in *SEE* values of $UBP_{10}$ for GRNN-based models compared to *SEE* values obtained by SVM, MLP, TB, RBF and SDT-based models using CFS on UBP-set2

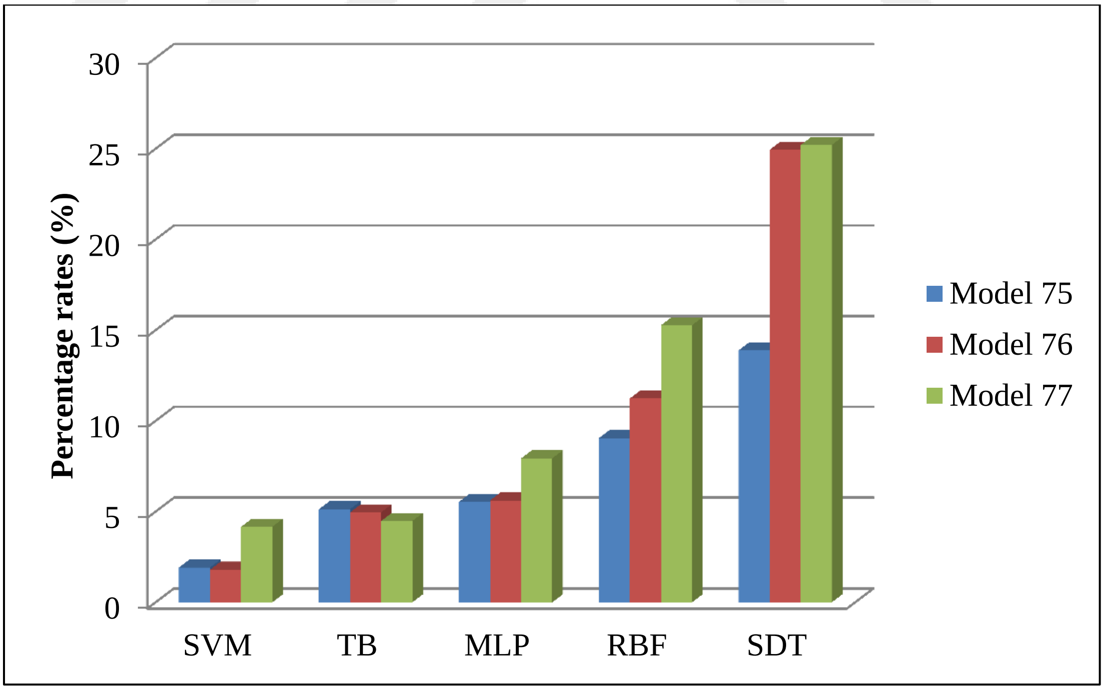


Figure 4.22. Percentage decrease rates in *SEE* values of $UBP_{60}$ for GRNN-based models compared to *SEE* values obtained by SVM, TB, MLP, RBF and SDT-based models using CFS on UBP-set2

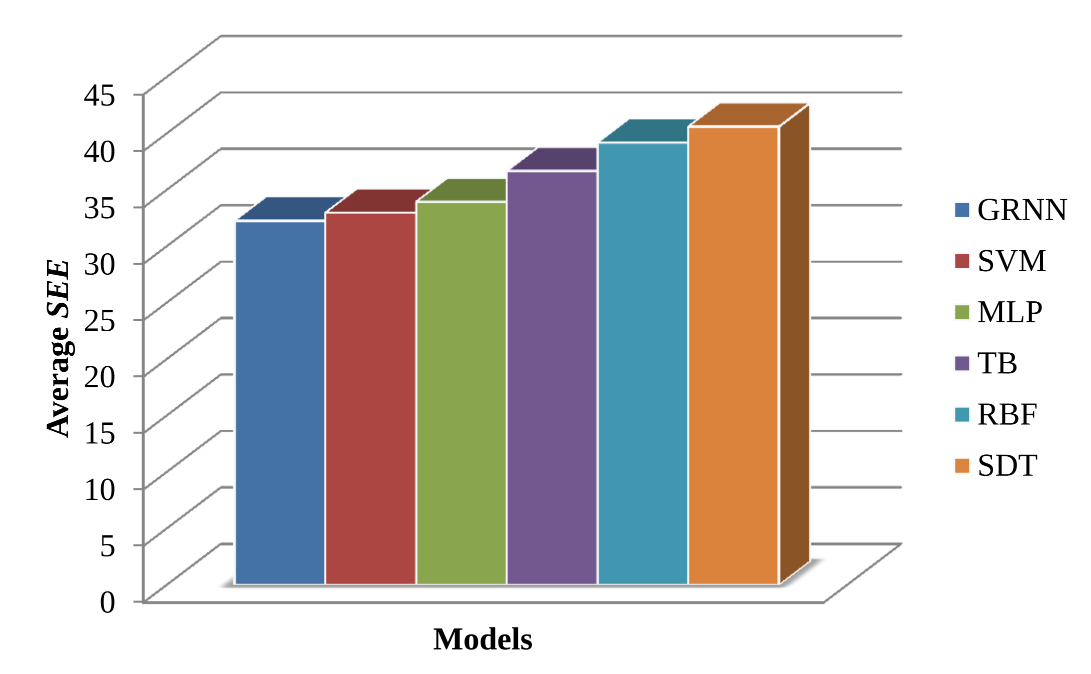


Figure 4.23. Average *SEE*'s of prediction models for $UBP_{10}$ using CFS on UBP-set2

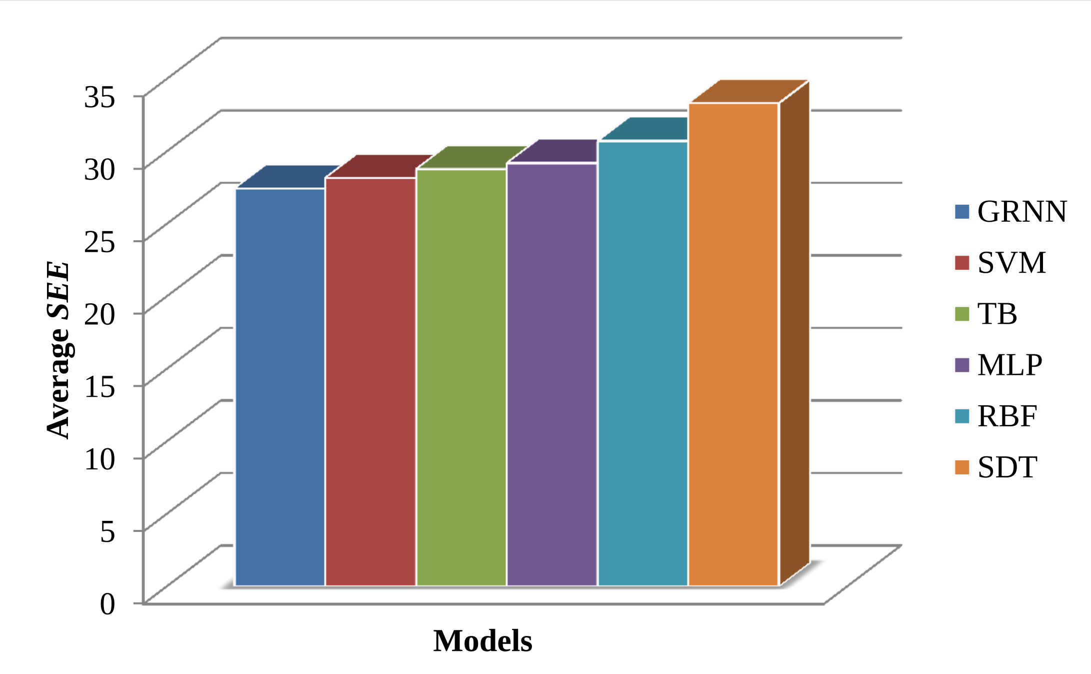


Figure 4.24. Average *SEE*'s of prediction models for $UBP_{60}$ using CFS on UBP-set2

### 4.4. General Discussion

- For both $UBP_{10}$ and $UBP_{60}$ predictions, GRNN shows superior performance and SDT exhibits the poorest prediction performance. In particular, the order of the regression methods in terms of their prediction performance, from the highest to the lowest, is GRNN, SVM, MLP, TB, RBF and SDT on UBP-set1. Similarly, the order of the regression methods, from the highest to the lowest, is GRNN, SVM, TB, MLP, RBF and SDT on UBP-set2.
- It can be deduced from the *SEE*'s of all prediction models that $UBP_{10}$ and $UBP_{60}$ can be predicted with lower *SEE*'s, *MAPE*'s and higher *R*'s using smaller number of features instead of the full set of features.
- In general, inclusion of height and weight in place of BMI in the prediction models give lower error rates for the prediction of $UBP_{10}$ and $UBP_{60}$ regardless of which regression methods have been used.
- According to the ranking given by Relief-F and mRMR selection algorithms, the feature "gender" generally has the highest score, whereas the feature "protocol" generally has the lowest score.
- The prediction models in average having the lowest *SEE*'s and *MAPE*'s, also, the highest *R*'s were produced by the mRMR feature selection algorithms. This statement is true for both $UBP_{10}$ and $UBP_{60}$ predictions on UBP-set1 and UBP-set2.
- The order of the feature selection algorithms in terms of their contribution to the performance of the prediction models, from the highest to the lowest, is mRMR, Relief-F and CFS.

# 5. CONCLUSION

The application of machine learning methods in sports science has long been an active research area. Although machine learning methods have been applied on different problems in various sports branches, some of the problems require further investigation. One such sports field is cross-country skiing. More specifically, predicting the UBP of cross-country skiers and identifying the discriminative features affecting UBP are the two important problems in this field.

In this thesis, various $UBP_{10}$ and $UBP_{60}$ prediction models have been developed by using Relief-F, mRMR and CFS feature selection algorithms, which have been integrated into GRNN, SVM, MLP, TB, RBF and SDT. The effects of the features on UBP have also been investigated. In more details, utilization of Relief-F and mRMR feature selection algorithms produced the ranking of the features. Then, based on these ranking scores, several models have been developed by removing the feature with the lowest score at a time. In contrast to Relief-F and mRMR algorithms, the CFS algorithm gave a set of selected features to develop a single UBP prediction model for each search algortihm. Using 10-fold cross validation, *R* and *SEE* have been used to assess the performance of all prediction models.

Several conclusions can be deduced with respect to the obtained results. In the first place, using less amount of features than the full set of features is feasible for prediction of $UBP_{10}$ and $UBP_{60}$. Secondly, the model comprising of the features time, weight, gender, height and age in general yields the minimum *SEE* (24.73 W) and maximum *R* (0.937) for $UBP_{10}$ prediction. On the other hand, the model including the features weight, age, gender, height and $VO_2$max in general gives the minimum *SEE* (14.94 W) and maximum *R* (0.960) for $UBP_{60}$ prediction. These both statements are valid independent of the regression method utilized. Using these two models instead of the full set of predictor variables yields up to 13.16% and 20.98% decrement rates in *SEE* for prediction of $UBP_{10}$ and $UBP_{60,}$ respectively. Thirdly, error rates related to the prediction of $UBP_{60}$ are always lower than that of $UBP_{10}$ because of the high standard deviation of $UBP_{10}$. In particular, when all results are investigated, the *SEE* of $UBP_{60}$ prediction is in average 27.74% lower than the *SEE* of $UBP_{10}$ prediction.

Fourthly, GRNN outperforms the other regression methods whereas SDT has the lowest ranking in terms of prediction performance among the utilized machine learning methods. Lastly, the prediction models that are comprised of the features height, gender, weight and BMI show the highest correlation with $UBP_{10}$ and $UBP_{60}$ as these models have the best prediction performance.

Although this is a detailed study, further research can be carried out in various aspects. Several other machine learning methods and feature selection algorithms can be combined to further analyze the effects of features on UBP. Upon collaboration with scientists from the sports field, new data sets can be created. The effect of questionnaire variables on UBP prediction also deserves further investigation. Development of hybrid models for prediction of UBP is definitely another challenging area to make research on.

## REFERENCES

ACID, S., DE CAMPOS, L. M., and FERNANDEZ, M., 2011. Minimum redundancy maximum relevancy versus score-based methods for learning Markov boundaries. In Intelligent Systems Design and Applications (ISDA), 11th International Conference, IEEE, 619-623.

AKAY, M. F., INAN, C., BRADSHAW, D. I., and GEORGE, J. D., 2009. Support vector regression and multilayer feed forward neural networks for non-exercise prediction of VO 2 max. Expert Systems with Applications, 36(6):10112-10119.

AKAY, M.F., ABUT, F., DANESHVAR, S., and HEIL, D., 2015a. Prediction of upper body power of cross-country skiers using support vector machines. Int. Arabian Journal for Science and Engineering, 40(4), 1045–1055.

AKAY, M.F., ABUT, F., OZCILOGLU, M., and HEIL, D., 2015b. Identifying the discriminative predictors of upper body power of cross-country skiers using support vector machines combined with feature selection. Int. J. Neural Computing and Applications, doi: 10.1007/s00521-015-1986-9.

AKGOL, D., AKAY, M.F., TURHAN, I., and HEIL, D., 2014. Prediction of upper body power of cross-country skiers using support vector machines combined with feature selection. 2nd International Symposium on Engineering, Artificial Intelligence & Applications, North Cyprus, Nov 5-7, 7-9.

AKGOL, D., AKAY, M.F., and HEIL, D., 2015a. Development of new models for predicting upper body power of cross-country skiers using machine learning methods. 1st International Symposium on Sport Science, Engineering and Technology, Istanbul, Turkey, May 10-13, 34-40.

AKGOL, D., and AKAY, M.F., 2015b. Predicting upper body power of cross-country skiers using machine learning methods combined with feature selection. 23th Signal Processing and Communications Applications Conference, Malatya, Turkey, May 16-19, 148-151.

ALSOBROOK, N. G., and HEIL, D. P., 2009a. Upper body power as a determinant of classical cross-country ski performance. European Journal of Applied Physiology, 105:633-641.

ALSOBROOK, N.G., HEIL, D.P., 2009b. Anaerobic and aerobic upper body power as determinants of classical cross-country ski performance. Eur. J. Appl. Physiol. 105, 633–641.

ALSOBROOK, N. G., 2005. The role of upper body power in classical cross-country skiing performance (Doctoral dissertation, Montana State University-Bozeman, College of Education, Health & Human Development).

APPEL, R., FUCHS, T., DOLL'AR, P. , and PERONA, P., 2013. Quickly Boosting Decision Trees Pruning Underachieving Features Early. JMLR Workshop and Conference Proceedings, 28:594-602.

ARMSTRONG, J. S., 2011. Illusions in regression analysis. Available at SSRN 1969740.

CELIKOGLU, H. B., 2006. Application of radial basis function and generalized regression neural networks in non-linear utility function specification for travel mode choice modelling. Mathematical and Computer Modelling, 44(7):640-658.

DANESHVAR, S., and AKAY, M.F., 2013. Prediction of Upper Body Power of Cross-Country Skiers Using Artificial Neural Networks. First International Symposium on Engineering, Artificial Intelligence & Applications (ISEAIA2013), North Cyprus, 6-8 November, pp. 30.

DANESHVAR, S., ABUT, F., YILDIZ, I., and AKAY, M.F., 2015. Development of new upper body power prediction models for cross-country skiers by using different machine learning methods. Signal Processing and Communications Applications Conference, Malatya, Turkey, May 16-19, 260-263.

DING, C., and PENG, H., 2005. Minimum redundancy feature selection from microarray gene expression data. Journal of bioinformatics and computational biology, 3(02):185-205.

DOWNING, J. J., and WILCOX, A. R., 2003. Effects of an eight-week training program on upper-body power in women cross-country skiers. J. Strength Cond. Res., 17( 4):726–733.

FRIEDMAN, J. H., 2002. Stochastic gradient boosting. Computational Statistics & Data Analysis, 38:367-378.

FRIEDRICHS, F., and IGEL, C., 2005. Evolutionary tuning of multiple SVM parameters. Neurocomputing, 64:107-117.

GHOSH-DASTIDAR, S., ADELI, H., and DADHMER, N., 2008. Principal Component Analysis Enhanced Cosine Radial Basis Function Neural Network for Robust Epilepsy and Seizure Detection. Biomedical Engineering, 55(2):512 – 518.

GUNN S. R., 1998. Support Vector Machines for Classification and Regression. ISIS Technical Report ISIS-1-98, Image Speech & Intelligent Systems Research Group, University of Southampton, 14.

GUO, X. C., LIANG, Y. C., WU, C. G., and WANG, C. Y., 2006. PSO-based hyper-parameters selection for LS-SVM classifiers. In Neural Information Processing, Springer Berlin Heidelberg, 1138-1147.

HALL, M. A., and SMITH, L. A., 1997. Feature Subset Selection: a correlation based filter approach. International Conference on Neural Information Processing and Intelligent Information Systems, Springer, Berlin, 855-858.

HAN, J.W., KAMBER, M., and PEI, J., 2006. Data mining: concepts and techniques, 2rd edition. Morgan Kaufmann Publishers, San Francisco, Canada, USA.

HASTIE, T., TIBSHIRANI, R., FRIEDMAN, J. H., 2009. 10. Boosting and Additive Trees. The Elements of Statictical Learning. Second Edition, Springer, New York, 337–384. ISBN 0-387-84857-6.

HAYKIN, S., 1994. NN: A comprehensive foundation.

HEIL, D.P., MURPHY, O.F., MATTINGLY, A.R., and HIGGINSON, B.K., 2001. Prediction of uphill time-trial bicycling performance in humans with a scaling-derived protocol. European Journal of Applied Physiology, 85: 374-382.

HEIL, D.P., ENGEN, J., and HIOGGINSON, B.K., 2004. Influence of ski pole grip on peak upper body power output in cross-country skiers. European Journal of Applied Physiology, 91: 481-487.

HEIL, D.P., WILLIS, S.J., 2012. Determinants of both classic and skate cross country ski performance in competitive junior and collegiate skiers. Science and Skiing V. In: E. Muller, S. Lindinger and S. Stoggl (Eds.), 513-522.

HEIL, D. P., and CAMENISCH, K., 2014. Static Flexibility as a Correlate of Skate Roller Skiing Economy Within Collegiate Cross-Country Skiers. *International Journal of Sports Science*, *4*(5):188-197.

HINDMAN, STEVE, 2005. Cross-Country Skiing: Building Skills for Fun and Fitness. Seattle: The Mountaineers Books. p. 264. ISBN 978-0-89886-862-3.

HSU, C. W., CHANG, C. C., and LIN, C. J., 2003. A practical guide to support vector classification"(http://www. csie. ntu. edu. tw/~ cjlin/papers/guide/guide. pdf.

HUANG, Y., 2009. Advances in artificial neural networks–methodological development and application. Algorithms, 2(3):973-1007.

HU, Q., CHE, X., ZHANG, L., and YU, D., 2010. Feature evaluation and selection based on neighborhood soft margin. Neurocomputing, 73(10): 2114-2124.

ISOGLU, O., AKAY, M.F., DANESHVAR, S., and HEIL, D., 2014. Predicting the upper body of cross-country skiers using support vector machines. 7th Engineering and Technology Symposium, Ankara, Turkey, May 15-16, 21-24.

IVANOVICH, E. V., and HAMID, K., 2014. An Alternative Approach for Assessing Sediment Impact on Aquatic Ecosystems Using Single Decision Tree (SDT). Journal of Water Sustainability, 4(3):181-204.

KANTARDZIC, M., 2003. Data Mining: Concepts, models, methods, and algorithms. John Wiley & Sons, USA.

KOMI PV, 1987. Force measurement during cross-country skiing. Int JSport Biomech, 3:370–381.

LI, C. P., ZHI, X. Y., MA, J., CUI, Z., ZHU, Z. L., ZHANG, C., and HU, L. P., 2012. Performance comparison between Logistic regression, decision trees, and multilayer perceptron in predicting peripheral neuropathy in type 2 diabetes mellitus. Chinese medical journal, 125(5):851-857.

LINDINGER, S.J., HOLMBERG, H.C., MULLER, E., and RAPP, W., 2009a. Changes in upper body muscle activity with increasing double poling velocities in elite cross-country skiing. European Journal of Applied Physiology, 106:353-363.

LINDINGER, S.J., STOGGL, T., MULLER, E., and HOLMBERG, H.C., 2009b. Control of speed during the double poling technique performed by elite cross-country skiers. Medical Science Sports Exercise, 41:210-220.

LINDINGER, S.J., HOLMBERG, H.C., 2011. How do elite cross-country skiers adapt to different double poling frequencies at low to high speeds? European Journal of Applied Physiology, 111:1103-1119.

MARSLAND, F., LYONS, K., ANSON, J., WADDINGTON, G., MACINTOSH, C., and CHAPMAN, D., 2012. Identification of cross-country skiing movement patterns using micro-sensors. Sensors (Basel, Switzerland), 12(4):5047–5066.

NADARAYA, E.A., 1964. On estimating regression, Theory Probab. Appl. 10, 186–190.

NASSIF, A. B., CAPRETZ, L. F., HO, D., and AZZEH, M., 2012. A treeboost model for software effort estimation based on use case points. In Machine Learning and Applications (ICMLA), 11th International Conference on IEEE, 2:314-319.

OZCILOGLU, M., AKAY, M.F., ABUT, F., and HEIL, D., 2015a. Feature Selection For Prediction Of Upper Body Power Of Cross-Country Skiers Using Different Machine Learning Methods. 1st International Symposium on Sport Science, Engineering and Technology, Istanbul, Turkey, Nov 4-6, 1-5.

OZCILOGLU, M., AKAY, M.F., and HEIL, D., 2015b. New Prediction Models for Upper Body Power of Cross-Country Skiers Using Machine Learning Methods with Minimum-Redundancy Maximum-Relevance Feature Selection Algorithm. Third International Symposium on Engineering, Artificial Intelligence & Applications, North Cyprus, Nov 4-6, 16-17.

OZCILOGLU, M., and AKAY, M.F., 2016a. Tree Based Methods Combined with Feature Selection for Predicting Upper Body Power of Cross Country Skiers. International Conference on Natural Science and Engineering, Kilis, March 19-20, 213.

OZCILOGLU, M., AKAY, M.F., and HEIL, D., 2016b. On the Application of Artificial Intelligence and Feature Selection in Sports Science Education and Research: A Case Study. 5th Cyprus International Conference on Educational Research, North Cyprus, April 1-2.

PARK, J., SANDBERG, I.W., 1991. Universal approximation using radial basis function networks. Neural Computation, 3(2): 246–257.

PENG, H., LONG, F., and DING, C., 2005. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 27(8):1226-1238.

QUINLAN, J. R., 1986. Induction of decision trees. Machine learning, 1(1):81-106.

ROBNIK-SIKONJO, M., and KONONENKO, I., 2003. Theoretical and empirical analysis of ReliefF and RReliefF. Machine learning, 53(1-2): 23-69.

RUSKO, H., 2003. Physiology of cross-country skiing. In Handbook of sports medicine and science: Cross-Country Skiing. Blackwell Publishing Company, 1-31.

SALTIN, B., 1996. The physiology of competitive c.c. skiing across a four decade perspective; with a note on training induced adaptations and role of training at medium altitude. In E. Muller (Ed.), London: Spon Press. Science and skiing, 435–469.

SANLI, T.E., ABUT, F., AKAY, M.F., and HEIL, D., 2014. Investigating the Effects of Maximum Oxygen Uptake and Maximum Heart Rate Variables on the Prediction of Upper Body Power of Cross-Country Skiers Using Support Vector Machines. Second International Symposium on Engineering, Artificial Intelligence & Applications (ISEAIA2014), North Cyprus, 5-7 Nov, 8-9.

SCHIOLER, H., HATMANN, U., 1992. Mapping neural network derived from the Parzen window estimator. Neural Netw. 5 (6):903–909.

SPECHT, D.F., 1991. A general regression neural network. IEEE Trans. Neural Network, 2 (6):568–576.

STREET, GM, 1989. Kinetic *analysis* of the V1 skate technique during roller skiing. 472. Medicine & Science in Sports and Exercise, 21(2):S79.

STREET, G.M., 1992. Technological advances in cross-country ski equipment. Medicine and Science in Sports and Exercise, 24(9):1048–1054.

SUN, Y., LOU, X., and BAO, B., 2011. A novel relief feature selection algorithm based on mean-variance model. Journal of Information & Computational Science, 8(16):3921-3929.

VAPNIK, V.N., 1995. The Nature of Statistical Learning Theory. New York: Springer.

VAPNIK, V.N., GOLOWICH, S., and SMOLA, A., 1997. Support vector method for function, approximation, regression estimation and signal processing. In M. Mozer, M. Jordan, & T. Petsche (Eds.).Advances in Neural Information Processing Systems, Cambridge, MA: MIT Press., 9:281–287.

VAPNIK, V.N., 2000. The Nature of Statistical Learning Theory. New York: Springer, 2nd edition.

WATSON, G.S.,1964. Smooth regression analysis. Sankhya, Ser. A, 26(4):359–372.

WITTEN, I. H., FRANK, E., and HALL, M. A., 2011. Data mining: practical machine learning tools and techniques. Morgan Kaufmann, Amsterdam.

WRIGHT, G. B., FLYER, N., FORNBERG, B., 2013. Radial Basis Function-generated Finite Differences: A Mesh-free Method for Computational Geosciences, October 17, 1-30.

## CURRICULUM VITAE

Mustafa Mikail ÖZÇİLOĞLU was born in Kilis-Turkey, in 1985. He completed the primary education and went to Hacı Mehmet Koçarslan Anatolian high school in Kilis. He graduated from the department of Computer Engineering, Mersin University, in 2006. He received MSc degree in the department of Computer Engineering from TOBB Economy and Technology University.

He has been working as a computer engineer in IT department of Kilis University since 2010.