ÇUKUROVA UNIVERSITY
INSTITUTE OF NATURAL AND APPLIED SCIENCES

PhD THESIS

Kasım ZOR

RESEARCH AND APPLICATION OF REAL-TIME SHORT-TERM ELECTRICAL ENERGY CONSUMPTION FORECASTING USING ARTIFICIAL INTELLIGENCE BASED TECHNIQUES

DEPARTMENT OF ELECTRICAL AND ELECTRONICS ENGINEERING

ADANA, 2019

ÇUKUROVA UNIVERSITY
INSTITUTE OF NATURAL AND APPLIED SCIENCES

RESEARCH AND APPLICATION OF REAL-TIME SHORT-TERM ELECTRICAL ENERGY CONSUMPTION FORECASTING USING ARTIFICIAL INTELLIGENCE BASED TECHNIQUES

Kasım ZOR

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 29/08/2019.

.....	.....	.....
Assoc. Prof. Dr. Ahmet TEKE SUPERVISOR	Asst. Prof. Dr. Adnan TAN MEMBER	Asst. Prof. Dr. Necdet Sinan ÖZBEK MEMBER
.....	.....	
Asst. Prof. Dr. Ercan AVŞAR MEMBER	Asst. Prof. Dr. İpek ABASIKELEŞ TURGUT MEMBER	

This PhD Thesis is written at the Department of Electrical and Electronics Engineering 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 Faculty Development Programme Coordination Unit of Çukurova University.

ABSTRACT

PhD THESIS

RESEARCH AND APPLICATION OF REAL-TIME SHORT-TERM ELECTRICAL ENERGY CONSUMPTION FORECASTING USING ARTIFICIAL INTELLIGENCE BASED TECHNIQUES

Kasım ZOR

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

Supervisor: Assoc. Prof. Dr. Ahmet TEKE
Co-supervisor: Asst. Prof. Dr. Hatice Başak YILDIRIM

Year: 2019, Pages: 149

Jury: Assoc. Prof. Dr. Ahmet TEKE
Asst. Prof. Dr. Adnan TAN
Asst. Prof. Dr. İpek ABASIKELEŞ TURGUT
Asst. Prof. Dr. Ercan AVŞAR
Asst. Prof. Dr. Necdet Sinan ÖZBEK

Short-term electrical energy consumption forecasting is crucial for efficient, reliable, and economic operations of hospitals due to serving 24/7, and they require round-the-clock energy. The main objectives of this thesis are to gather real-time electrical energy consumption and meteorological data of a large hospital complex by utilising an energy logger connected to a humidity-temperature transducer on-site and MERRA-2 data to form a very-short term raw data set in RStudio environment wherein R programming language is used; to develop a novel methodology that has the capability of identifying missing and erroneous values in the raw data set, replacing these values with NA values, imputing the NA values with different methods, and completing the conversion process by creating a cleansed short-term data set; to forecast short-term electrical energy consumption of the hospital by using artificial intelligence based techniques; and to present benchmark analyses of the obtained results by evaluating coefficient of determination, coefficient of variation, mean absolute error, root mean squared error, mean absolute percentage error, and discussing the future perspectives respectively.

Key Words: Short-Term, Electrical Energy Consumption Forecasting, Artificial Intelligence Techniques, Imputation Methods, Hospital.

ÖZ

DOKTORA TEZİ

GERÇEK ZAMANLI KISA DÖNEM ELEKTRİK ENERJİSİ TÜKETİM TAHMİNİNİN YAPAY ZEKA TABANLI YÖNTEMLER KULLANILARAK ARAŞTIRILMASI VE UYGULANMASI

Kasım ZOR

ÇUKUROVA ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ
ELEKTRİK-ELEKTRONİK MÜHENDİSLİĞİ ANABİLİM DALI

Danışman: Doç. Dr. Ahmet TEKE
İkinci Danışman: Dr. Öğr. Üyesi Hatice Başak YILDIRIM
Yıl: 2019, Sayfa: 149
Jüri: Doç. Dr. Ahmet TEKE
Dr. Öğr. Üyesi Adnan TAN
Dr. Öğr. Üyesi İpek ABASIKELEŞ TURGUT
Dr. Öğr. Üyesi Ercan AVŞAR
Dr. Öğr. Üyesi Necdet Sinan ÖZBEK

7 gün 24 saat sürekli çalışmaları nedeniyle enerji ihtiyacı devamlılık arz eden hastanelerin verimli, güvenilir ve ekonomik işleyişi için kısa dönem elektrik enerjisi tüketim tahmini hayati önem taşımaktadır. Bu tezin ana hedefleri sırasıyla sahadaki nem-sıcaklık sensörüne bağlı bir enerji kayıt cihazından ve MERRA-2 verilerinden faydalanarak, R programlama dilinin kullanıldığı RStudio ortamında çok kısa dönem bir ham veri seti oluşturmak için büyük bir hastane kompleksinin gerçek zamanlı elektrik enerjisi tüketim ve meteorolojik verilerini toplamak; oluşturulmuş ham veri setindeki kayıp ve hatalı değerleri saptama, bu değerleri NA ile değiştirme, NA değerlerinin yerine farklı metotlar ile veri atama ve dönüştürme işlemini tamamlayarak temizlenmiş bir kısa dönem veri seti yaratma yeteneklerine sahip özgün bir metodoloji geliştirmek; yapay zeka tabanlı yöntemler kullanarak hastanenin kısa dönem elektrik enerjisi tüketimini tahmin etmek; belirleme katsayısı, varyasyon katsayısı, ortalama mutlak hata, kök ortalama karesel hata ile ortalama mutlak yüzdesel hatayı değerlendirerek ve gelecek öngörülerini tartışarak elde edilen sonuçların karşılaştırmalı analizlerini sunmaktır.

Anahtar Kelimeler: Kısa Dönem, Elektrik Enerjisi Tüketim Tahmini, Yapay Zeka Yöntemleri, Veri Atama Metotları, Hastane.

EXTENDED SUMMARY

More recently, energy forecasting applications not only on the grid side of electric power systems, but also on the customer side for load and demand prediction purposes have become crucial after the deregulation of electricity market and advancements in the smart grid technologies. In this manner, short-term electrical energy consumption forecasting (STEF), which covers hour, day, and week ahead predictions, is essential to energy management and planning of all buildings from households and residences in the small-scale to huge building complexes in the large-scale.

Hospitals may be described as highly sophisticated organisations from the point of view of functional, technological, economic, managerial, and procedural aspects. The reliability of uninterrupted energy flow has utmost importance for hospitals owing to their continuous duty for 24/7 operation without any excuses. STEF is an essential tool that is not merely required for the integration of smart grids to current electric power systems, but enhances hospital's quality of energy management and planning as well by monitoring energy consumption, finding base and peak demands, reducing losses, minimising risks, securing reliability for uninterrupted operation, playing an active role in making viable decisions in regard to maintenance planning and future investments including both renewable and non-renewable energy technologies such as photovoltaic (PV), landfill, tri-generation fuelled by natural gas.

In this thesis, a study of research and application of real-time STEF using artificial intelligence (AI) based techniques has been realised for a large hospital complex in Adana, Turkey. Initially, in order to form a very-short term raw data set with a sampling period of 10 minutes in RStudio environment wherein R

programming language is used, an energy logger connected to a humidity-temperature transducer has been utilised for gathering real-time electrical energy consumption and meteorological data on-site, and MERRA-2 (Modern-Era Retrospective analysis for Research and Applications, Version 2) data of Global Modelling and Assimilation Office of the US National Aeronautics and Space Administration.

Very short-term raw data set has 3 input variable categories and 1 target variable. The input variable categories are electrical, meteorological, and calendar variables. Target variable is the actual electrical energy consumption. Electrical variables can be stated as the historical electrical energy consumption variables belonging to previous 10 minute, previous 1 day (the same time in the previous day), and previous 1 week (the same time and day in the previous week). Meteorological variables have been obtained from two different sources. The first source is on-site humidity-temperature transducer which provides the indoor relative humidity and temperature of the ambient where the transducer is located with the energy logger. The second source is MERRA-2 data consisting of outdoor temperature, relative humidity, pressure, wind direction, wind speed, rainfall, and short-wave irradiation. Calendar variables include month of year, week of year, day of month, hour of day, sample number of hour, and day type. The very short-term raw data set is represented as a $52,416 \times 19$ matrix in RStudio environment.

Afterwards, the created very-short term raw data set is converted to a cleansed short-term data set by applying a novel methodology named as forecast time horizon converter (FTHC) which is developed to identify missing and erroneous values in the very short-term data set, replace these values with NA values, impute the NA values with different methods, and complete the conversion process.

Missing and erroneous values in the very short-term raw data set is occurred due to power outages in the hospital. For this case, the actual variables where data are missing are not the cause of the incomplete data. Instead, the cause of the missing data is due to some other external influence which is power outage for the case of hospital, and hence missing data mechanism for this case is missing at random (MAR).

The number of variables possessing NA values for the very short-term data set is 6 and they are actual, previous 10 minute, previous 1 day, and previous 1 week electrical energy consumption, indoor relative humidity, and indoor temperature acquired from the energy logger and the humidity-temperature transducer. The number of rows having NAs varies between 379 and 398, while the proportion of NAs for each variable changes from 0.723% to 0.759%. In total, 2,333 NAs in a $52,416 \times 19$ matrix correspond to a proportion of 0.234%.

To the best of one's knowledge, first and foremost there is no certain threshold to ignore the imputation, but the most of the literature suggests applying complete case analysis to a data set with a missing data mechanism of MAR if the proportion of missing data is below 5%. However, imputation looks very tempting, because performing complete case analysis (CCA) by using listwise deletion to a data set with a missing data mechanism of MAR sometimes wastes a whole row for just one NA value in the row and frequently introduces bias which causes a loss in efficiency. Therefore, Kalman filters (KalmanARIMA and KalmanStructTS), interpolation (linear, Stineman, and spline), weighted moving average (simple, linearly weighted, and exponentially weighted), kNN imputation, and persistence (last observation carried forward and next observation carried backward) methods are employed to compare the performances of different imputation methods and reveal the impacts of each method on a data set especially prepared for energy

forecasting. After the completion of the conversion process, the cleansed short-term data set is represented as $8,736 \times 18$ matrix in RStudio environment and names of all variables remain the same except previous 10 minute owing to renaming as previous 1 hour. In addition to those, sample number of hour variable is not used in the short-term data set.

Next, STEF of the hospital is implemented by using multiple linear regression (MLR) as a statistical technique for benchmarking purposes and AI based techniques containing support vector machines (SVM), gene expression programming (GEP), gradient boosted decision trees (GBDT), and artificial neural networks (ANN) consisting of multilayer perceptron neural networks (MLPNN), radial basis function neural networks (RBFNN), generalised regression neural networks (GRNN), and group method of data handling polynomial neural networks (GMDHNN) under identical constraints such as random sampling for training and validation.

Furthermore, an algorithm containing sensitivity analysis is utilised for all statistical and artificial intelligence techniques to calculate relative importance of input variables in which the values of each variable are randomised and the effect on the quality of the model is computed out of hundred. Depending on the results of this analysis, within all input variables; electrical variable previous 1 hour, meteorological variable short-wave irradiation, and calendar variable hour of day have been identified as prerequisite for one hour ahead STEF. In order to reach satisfactory results for one hour ahead STEF problem; the necessity of having at least electrical variables containing previous 1 hour, 1 day, and 1 week; meteorological variables including short-wave irradiation, indoor and outdoor temperatures and rainfall; and calendar variables consisting of hour of day, week of year, and month of year has been emphasised.

As a consequence, benchmark analyses of the achieved results; in which GBDT surpass other statistical and AI techniques, and imputed data sets with KalmanARIMA, linear interpolation, and next observation carried backward are ranked among the top three; are presented meticulously by evaluating coefficient of determination (R^2), coefficient of variation (CV), mean absolute error (MAE), root mean squared error (RMSE), mean absolute percentage error (MAPE), and the future perspectives are discussed after conclusions respectively.

GENİŞLETİLMİŞ ÖZET

Elektrik piyasasının serbestleşmesi ve akıllı şebeke teknolojilerindeki gelişmeler sonrasında elektrik güç sistemlerinin sadece şebeke tarafında değil, yük ve talep tahmini amaçlarıyla müşteri tarafında da enerji tahmin uygulamaları yakın zamanda önem kazanmıştır. Bu sebeple, saatlik, günlük ve haftalık ileriye dönük öngörüleri kapsayan kısa dönem elektrik enerjisi tüketim tahmini (KDET) küçük ölçekteki konut ve meskenlerden geniş ölçekteki devasa bina komplekslerine kadar tüm binaların enerji yönetimi ve planlaması için gereklilik arz etmektedir.

Hastaneler işlevsel, teknolojik, ekonomik, yönetimsel ve prosedürel bakış açılarından son derece karmaşık organizasyonlar olarak tanımlanabilirler. Mazeret olmaksızın 7 gün 24 saat sürekli çalışmaları nedeniyle kesintisiz enerji akışının güvenilirliği hastaneler için azami önem taşımaktadır. KDET yalnızca akıllı şebekelerin mevcut elektrik güç sistemlerine entegrasyonu için gerekli değil; enerji tüketimini gözlemleyerek, dip ve tepe taleplerini bularak, kayıpları azaltarak, riskleri en aza indirgeyerek, kesintisiz çalışma için güvenilirliği tesis ederek, fotovoltaik (FV), katı atık ve doğal gaz yakıtlı trijenerasyon gibi hem yenilenebilir hem de yenilenemeyen enerji teknolojilerini içeren gelecek yatırımlar ve bakım planlaması ile ilgili uygulanabilir kararların alınmasında aktif rol oynarak hastanenin enerji yönetimi ve planlamasının kalitesinin arttırılmasında da gerekli bir araçtır.

Bu tezde, Adana, Türkiye'deki büyük bir hastane kompleksi için gerçek zamanlı kısa dönem elektrik enerjisi tüketim tahmininin yapay zeka tabanlı yöntemler kullanılarak araştırılması ve uygulanması çalışması gerçekleştirilmiştir. İlk olarak, R programlama dilinin kullanıldığı RStudio ortamında 10 dakika örnekleme zamanlı çok kısa dönem bir ham veri setini oluşturmak amacıyla sahadaki nem-sıcaklık sensörüne bağlı bir enerji kayıt cihazından ve ABD Ulusal

Havacılık ve Uzay Dairesi Global Modelleme ve Asimilasyon Ofisi'nin MERRA-2 (Araştırma ve Uygulamalar İçin Modern Çağ Retrospektif Analizi, Versiyon 2) verilerinden faydalanılmıştır.

Çok kısa dönem ham veri seti 3 kategoride giriş değişkenine ve 1 hedef değişkenine sahiptir. Giriş değişkeni kategorileri elektriksel, meteorolojik ve takvim değişkenlerinden oluşmaktadır. Hedef değişkeni ise güncel elektrik enerjisi tüketimidir. Elektriksel değişkenler geçmiş 10 dakika, geçmiş 1 gün (bir önceki günün aynı zamanı) ve geçmiş 1 haftayı (bir önceki haftanın aynı gün ve zamanı) bünyesinde barındıran geçmiş elektrik enerjisi tüketim değişkenleri olarak belirtilebilirler. Meteorolojik değişkenler iki farklı kaynaktan elde edilmişlerdir. İlk kaynak enerji kayıt cihazının bulunduğu ortamın bina içi bağıl nem ve sıcaklığının ölçümünü sağlayan sahadaki nem-sıcaklık sensörüdür. İkinci kaynak ise bina dışı sıcaklık, bağıl nem, basınç, rüzgar yönü, rüzgar hızı, yağış miktarı ve kısa dalga ışınımını içeren MERRA-2 verileridir. Takvim değişkenleri yılın ayı, yılın haftası, ayın günü, günün saati, saatin örnekleme numarası ve gün tipini içermektedir. Çok kısa dönem ham veri seti RStudio ortamında 52.416×19'luk bir matrisle gösterilmektedir.

Sonrasında, oluşturulan çok kısa dönem ham veri seti, tahmin zaman ufku dönüştürücü (TZUD) olarak adlandırılan ve geliştirilme amacı bu veri setindeki kayıp ve hatalı değerleri saptamak, saptanan değerleri NA ile değiştirmek, NA değerlerinin yerine farklı metotlar ile veri atamak ve kısa dönem veri setine dönüştürme işlemini tamamlamak olan özgün bir metodoloji uygulanarak temizlenmiş bir kısa dönem veri setine dönüştürülmektedir.

Çok kısa dönem ham veri setindeki kayıp ve hatalı değerler hastanedeki enerji kesintileri nedeniyle meydana gelmektedir. Bu durumda, eksik verinin sebebi kayıp verinin bulunduğu mevcut değişkenler değildir. Buna karşılık, eksik veri

başka bir dış etki nedeniyle meydana gelmiş olup, bu etki hastane örneği için enerji kesintisidir ve bu yüzden bu örnek için kayıp veri mekanizması rastlantısal kayıptır.

Çok kısa dönem veri setindeki NA değerlerine sahip değişkenlerin sayısı 6 olup, enerji kayıt cihazı ve nem-sıcaklık sensöründen elde edilen güncel, geçmiş 10 dakika, geçmiş 1 gün ve geçmiş 1 hafta elektrik enerjisi tüketimi, bina içi bağıl nem ve sıcaklık değişkenleridir. Her değişken için NA'lara sahip satırların sayısı 379 ve 398 arasında çeşitlenmekteyken, NA'ların oranları ise %0,723 ile %0,759 arasında değişmektedir. Toplamda, 52.416×19'luk matristeki 2.333 NA değeri %0,234'lük bir orana karşılık gelmektedir.

Bilindiği kadarıyla, veri atamanın göz ardı edilmesi için her şeyden önce bir eşik bulunmamakta, fakat literatürün büyük kısmı kayıp veri mekanizması rastlantısal kayıp olan bir veri setinin eğer kayıp veri oranı %5'in altında ise tam vaka analizi (TVA) uygulamayı önermektedir. Bununla birlikte, kayıp veri mekanizması rastlantısal kayıp olan bir veri setine listesel silme yöntemiyle tam vaka analizi uygulamanın bazen satırdaki tek bir NA değeri için tüm satırı boşa harcaması ve verim kaybına neden olacak şekilde sıklıkla sapma eklemesi yüzünden veri atama çok cazip görünmektedir. Bu nedenle, özellikle enerji tahmini için hazırlanmış bir veri seti üzerinde farklı veri atama metotlarının performanslarını karşılaştırmak ve her metodun etkilerini ortaya çıkarmak için Kalman filtreleri (KalmanARIMA ve KalmanStructTS), enterpolasyon (doğrusal, Stineman ve spline), ağırlıklı hareketli ortalama (basit, doğrusal ağırlıklandırılmış ve üssel ağırlıklandırılmış), k en yakın komşu ve sürerlik (önceki gözlemi ileriye taşıma ve sonraki gözlemi geriye taşıma) yöntemleri kullanılmaktadır. Dönüştürme işleminin tamamlanmasından sonra, temizlenmiş kısa dönem veri seti RStudio ortamında 8.736×18'lik bir matrisle gösterilmektedir ve adını geçmiş 1 saat olarak değiştiren geçmiş 10 dakika dışında tüm değişkenlerin adı aynı kalmaktadır. Ek olarak, saatin

örnekleme numarası değişkeni kısa dönem veri setinde kullanılmamaktadır.

Bundan sonra, hastanenin KDET, karşılaştırma amacıyla bir istatistiki yöntem olan çoklu doğrusal regresyon (ÇDR) ve yapay zeka tabanlı yöntemlerden destek vektör makineleri (DVM), gen ifadeli programlama (GİP), gradyan artırmalı karar ağaçları (GAKA) ve yapay sinir ağlarına (YSA) dayanan çok katmanlı algılayıcı sinir ağları (ÇKASA), radyal temelli fonksiyon sinir ağları (RTFSA), genelleştirilmiş regresyon sinir ağları (GRSA) ve veri işleme grup yöntemi polinom sinir ağları (VİGYPSA) eğitim ve doğrulama için rastgele örnekleme gibi özdeş kısıtlar altında kullanılarak uygulanmaktadır.

Ayrıca, tüm istatistiki ve yapay zeka yöntemlerinin giriş değişkenlerinin bağıl öneminin hesaplanmasında, her değişkenin değerlerinin rastgeleleştirildiği ve modelin kalitesine olan etkisinin yüz üzerinden hesaplandığı hassasiyet analizi içeren bir algoritma kullanılmaktadır. Bu analiz sonucuna bağlı olarak, tüm giriş değişkenleri arasında geçmiş 1 saat elektriksel değişkeni, kısa dalga ışınımı meteorolojik değişkeni ve günün saati takvim değişkeninin bir saat sonraki KDET için ön koşul oldukları tespit edilmiştir. Ayrıca, bir saat sonraki KDET probleminde kabul edilebilir sonuçlara ulaşmak için ise en azından geçmiş 1 saat, 1 gün ve 1 hafta elektrik değişkenleri; kısa dalga ışınımı, bina içi sıcaklığı, bina dışı sıcaklığı ve yağış miktarı meteorolojik değişkenleri ile günün saati, yılın haftası ve yılın ayı takvim değişkenlerinin kullanılmasının gerekliliği vurgulanmıştır.

Sonuç olarak, belirleme katsayısı (R^2), varyasyon katsayısı (VK), ortalama mutlak hata (OMH), kök ortalama karesel hata (KOKH) ve ortalama mutlak yüzdesel hatanın (OMYH) değerlendirilmesi ile GAKA'nın diğer istatistiki ve yapay zeka yöntemlerini geride bıraktığı, KalmanARIMA, doğrusal enterpolasyon ve sonraki gözlemi geriye taşıma ile veri ataması yapılan veri setlerinin ilk 3 sırayı aldığı sonuçların karşılaştırmalı analizleri titizlikle sunulmakta ve gelecek

öngörüleri sonuçların ardından sırasıyla tartışılmaktadır.

To my lonely and beautiful country

ACKNOWLEDGEMENTS

First and foremost, I would like to express my deepest gratitude to my supervisor Assoc. Prof. Dr. Ahmet TEKE for his excellent tutelage and complete support. I also want to present my sincere thanks to my co-supervisor Asst. Prof. Dr. Hatice Başak YILDIRIM for her continuous help and guidance.

In addition, I would like to thank the members of thesis progress committee Asst. Prof. Dr. Adnan TAN and Asst. Prof. Dr. İpek ABASIKELEŞ TURGUT for making valuable comments and suggestions, and the members of jury Asst. Prof. Dr. Ercan AVŞAR and Asst. Prof. Dr. Necdet Sinan ÖZBEK for analysing and evaluating my work.

I would like to appreciate Dr. Jethro Browell from University of Strathclyde for hosting me, widening my horizon by sharing his extensive knowledge, introducing the R programming language, and giving invaluable recommendations about my work. By the way, I would like to acknowledge the National Agency of Turkey and the EU for Erasmus+ Staff Mobility Grant for Training in the University of Strathclyde.

In the meantime, this thesis was supported by the Scientific Research Project Unit of Çukurova University [grant number FBA-2017-8252 and FBA-2017-9344]. Herein, I would also like to express my thanks to the personnel of local electric distribution company Toroslar EDAŞ, Çukurova University's Directorate of Construction and Technical Works, especially to Mr. Emirullah KAYA and Mr. Selçuk YAPICI from the Section of Maintenance and Operation, and also to Hospital Administration for their collaboration.

I am deeply indebted to Dr. Oğuzhan TİMUR for his moral and material support during this work. I would also like to thank my room-mate at work Res.

Asst. Özgür ÇELİK for his comradeship and patience. Special thanks go to Caner YILDIRIM, Emrah Mehmet GÜLLÜOĞLU, Mustafa AKAR, Burhan EVGİN, and Çağatay CEBECİ for their supports not only during this thesis, but also during my whole life.

Finally, I would like to present my everlasting thanks to my family, especially to my wife for her unconditional love and endless support, and also to my precious daughter for stealing my heart and being my joie de vivre.

CONTENTS **PAGE**

LIST OF TABLES **PAGE**

LIST OF FIGURES **PAGE**

LIST OF SYMBOLS

a : Either Coefficient or Weight Vector

b : Weights of Hidden Layer

b : Bias

b_0 : Bias of Hidden Layer Node

b_{0j} : Bias of the jth Node of Output Layer

C : Cost Controlling Empirical Risk Degree of Model

c_i : Centre of the ith Hidden Node

$d_{i,j}$: Gower Distance among ith and jth Observation

e : Error Term

$\mathrm{e}^{D(x_i)}$: Output of Each Neuron i

$F_m(x)$: Summation of m Decision Trees

F_t : State-Transition Parameter at t

$f(x)$: Flat Function

$f(x,y)$: Joint Probability Density Function

$f(\Gamma)$: Fitting Error of Γ

$f_1(x)$: First-Order Interpolating Polynomial

$g(\cdot)$: Transfer Function

H : Number of Hidden Nodes

H_t : Measurement Parameter at t

$h_m(x)$: Decision Trees of Constant Size

I : First Successive Point

I : Total Number of Inputs p_i

i : Index

i : Neuron

J : Second Successive Point

j : Index

K : Third Successive Point

K : Number of Columns

$K(x_i, x_j)$: Nonlinear Kernel Function

M : Identity Matrix

m : Number of Outputs

m_{ij} : Identity Matrix

n : Number of Rows

n : Number of Nodes in the Hidden Layer

n : Number of Measured Samples

n : Number of Neurons in a Layer

n : Number of Observations

$o_{i,j}$: The jth Output of the ith Sample

$P(x)$: Cubic Polynomials

p : Number of Neurons in the Input Layer

p : Dimension of x

p_i : Inputs

r_k : Range of the kth Variable

$r_{m,i,t}$: Model Fitting Current Residuals at Iteration m

S_D : Denominator

S_N : Numerator

$S(x)$: Sequence of Cubic Polynomials

s_j : Slope of the Line Segment

t : An Instant in Time

v : Neuron for Each Input Variable

v : Learning Rate

X : Input Vector

x : Input Vector

x : Vector for Input Random Variable

x : Set of Input Variables

x : Column Vector

x_i : An Electrical Energy Consumption Value

x_i : Value of Independent Variables

x_i : Training Sample

x_i : Input Instance

$x_{i,j}$: The jth Input of the ith Sample

$x_{i,k}$: The Value of kth Variable of the ith Observation

x_j : Rectangular Coordinates

x_j : Input Instance

$x_{\max}$: Maximum Value of x

$x_{\min}$: Minimum Value of x

x_{norm} : Normalised Column Vector Converted From x

x_t : State Vector at t

x_0 : Prediction Sample

Y : Representation of a Data Set

$Y_{i,t+k}$: Predicted Response

Y_{obs} : Observations of Y

Y_{miss} : Missing Values of Y

Y_t : Actual Term

$\hat{Y}_{t+1}$: One Step Ahead Term

y : Consumed Energy Amount

y : Scalar Indicating Output Random Variable

y : Output

y_i : Actual or Measured Output

y_{ij} : Representation of a Data Set

y_j : Rectangular Coordinates

$y_{\max}$: Maximum Value of y

$y_{\min}$: Minimum Value of y

y_t : Reciprocating Measurement Vector at t

$\acute{y}_j$: Slope of the Curve at jth Interior Point

$\acute{y}_m$: Slope of the Curve at mth End Point

$\hat{y}$: Predicted Output

$\overline{y}$: Mean of y_i

α_j : Lagrangian Function Multiplier

α_j^* : Lagrangian Function Multiplier

β_i : Regression parameters with respect to x_i

δ : Weights of Output Layer

δ_d : Tolerance Deviation

$\delta_{i,j,k}$: Contribution of kth Variable

δ_{0i} : Bias of Output Layer Node

ε_t : Random State Noise Term at t

Γ : Quality of Formula

γ : Gamma Controlling Gaussian Function Width

ω : Weight

ω_i : Weight

ω_{ij} : Weight of ith Node of the Hidden Layer

ω_k : Weight

ω_t : Measurement Error Term at t

ϕ : Unknown Parameters

σ : Spread Value

σ : Spread Factor

σ^2 : Variance

τ : Sequence of Real Numbers

ε : Error Term

ε : Epsilon Controlling ε-Insensitive Zone's Width

φ_i : Radial Basis Function with c_i being its Centre

φ_x : Mapping Function

ξ_i : Penalty for Observation out of ε Margin

ξ_i^* : Penalty for Observation out of ε Margin

LIST OF ABBREVIATIONS

AI : Artificial Intelligence

AMR : Automatic Meter Reading

ANFIS : Adaptive Neuro-Fuzzy Inference System

ANN : Artificial Neural Networks

AR : Auto-Regressive

ARIMA : Auto-Regressive Integrated Moving Average

ARMA : Auto-Regressive Moving Average

ARMAX : Auto-Regressive Moving Average with Exogeneous Inputs

BP : Back-Propagation

BPNN : Back-Propagation Neural Networks

BSc : Bachelor of Science

BSOD : Backward Second Order Difference

CCA : Complete Case Analysis

CCHP : Combined Cooling, Heat, and Power

CHP : Combined Heat and Power

CRBM : Conditional Restricted Boltzmann Machine

CV : Coefficient of Variation

DL : Deep Learning

DOI : Digital Object Identifier

EF : Electrical Energy Consumption Forecasting

ELM : Extreme Learning Machines

EMRA : Energy Market Regulatory Authority

FIS : Fuzzy Inference System

FL : Fuzzy Logic

FSOD : Forward Second Order Difference

FTHC : Forecast Time Horizon Converter

FTV : Fair Temperature Value

GA : Genetic Algorithms

GBDT : Gradient Boosted Decision Trees

GBM : Generalised Boosted Regression Models

GEP : Gene Expression Programming

GMAO : Global Modelling and Assimilation Office

GMDHNN : Group Method of Data Handling Polynomial Neural Networks

GRNN : Generalised Regression Neural Network

GRU : Gated Recurrent Unit

HMM : Hidden Markov Models

HVAC : Heating, Ventilation, and Air-Conditioning

IEEE : Institute of Electrical and Electronics Engineers

kNN : k-Nearest Neighbour

LEED : Leadership in Energy and Environmental Design

LI : Linear Interpolation

LOCF : Last Observation Carried Forward

LSSVM : Least Squares Support Vector Machines

LSTM : Long Short-Term Memory

LTEF : Long-Term Electrical Energy Consumption Forecasting

LWMA : Linearly Weighted Moving Average

MAE : Mean Absolute Error

MAPE : Mean Absolute Percentage Error

MAR : Missing at Random

MARSplines : Multivariate Adaptive Regression Splines
MCAR : Missing Completely at Random
MERRA-2 : Modern-Era Retrospective Analysis for Research and Applications, Version 2
MICE : Multivariate Imputation by Chained Equations
MLPNN : Multilayer Perceptron Neural Networks
MLR : Multiple Linear Regression
MOGA : Multi Objective Genetic Algorithm
MSc : Master of Science
MTEF : Medium-Term Electrical Energy Consumption Forecasting
MTU : Motoren und Turbinen Union GmbH
NA : Not Available
NAR : Nonlinear Auto-Regressive
NARX : Nonlinear Auto-Regressive with Exogeneous Inputs
NASA : National Aeronautics and Space Administration
NMAR : Not Missing at Random
NOCB : Next Observation Carried Backward
PARIMA : Profiles Auto-Regressive Integrated Moving Average
PCA : Principal Component Analysis
PhD : Doctor of Philosophy
PLM : Possibilistic Linear Model
PSO : Particle Swarm Optimisation
PV : Photovoltaic
RBFNN : Radial Basis Function Neural Networks
RMSE : Root Mean Squared Error

RNN : Recurrent Neural Networks

RSNNS : R Stuttgart Neural Network Simulator

RT : Regression Trees

R^2 : Coefficient of Determination

SARIMA : Seasonal Auto-Regressive Integrated Moving Average

SCG : Scaled Conjugate Gradient

SMA : Simple Moving Average

SpI : Spline Interpolation

StI : Stineman Interpolation

STEF : Short-Term Electrical Energy Consumption Forecasting

STLF : Short-Term Load Forecasting

SVM : Support Vector Machines

SVR : Support Vector Regression

S&AI : Statistical and Artificial Intelligence

S/AI : Statistical or Artificial Intelligence

TVB : Tao's Vanilla Benchmark

UCTEA : Union of Chambers of Turkish Engineers and Architects

UK : The United Kingdom of Great Britain and the Northern Ireland

URL : Uniform Resource Locator

US : The United States of America

USB : Universal Serial Bus

VDM : Variability Decomposition Method

VIM : Visualisation and Imputation of Missing Values

VSTEF : Very Short-Term Electrical Energy Consumption Forecasting

XGBoost : Extreme Gradient Boosting Trees

1. INTRODUCTION

In this chapter, background and motivation, objectives, methodology, contributions, and organisations of the thesis are emphasised respectively.

1.1. Background and Motivation

Recently, energy forecasting applications not only on the grid side of electric power systems, but also on the customer side for load and demand prediction purposes have come into prominence after the privatisation and deregulation of electricity market and technological advancements in the smart grid.

According to time horizon, there are four categories of electrical energy consumption forecasting (EF): (1) long-term electrical energy consumption forecasting (LTEF), among 3-year and 50-year energy consumption is predicted, (2) if the forecast ranges from 2 weeks to 3-year, then it is considered as medium-term electrical energy consumption forecasting (MTEF), (3) short-term electrical energy consumption forecasting (STEF) refers to hour, day, or week ahead predictions, and (4) very short-term electrical energy consumption forecasting (VSTEF) which includes few minutes to an hour ahead forecasting of energy consumption (Hong and Fan, 2016).

LTEF is utilised for the long-term power system planning in accordance with the future energy demand and energy policy of a country. MTEF is being used for the efficient operation and maintenance of the large facilities in the power system. MTEF is especially interesting for companies and large state facilities operating in a deregulated environment, as it provides them valuable insights on the market need of energy, maintenance schedule, fuel supplies, and occasional needs for further investments related to renewable energy technologies (Kaboli et al.,

2017). VSTEF and STEF play a crucial role in optimum unit commitment, control of spinning reserve, evaluation of sales or purchase contracts between several companies (Raza and Khosravi, 2015). EF applications and classification are illustrated in the following figure.

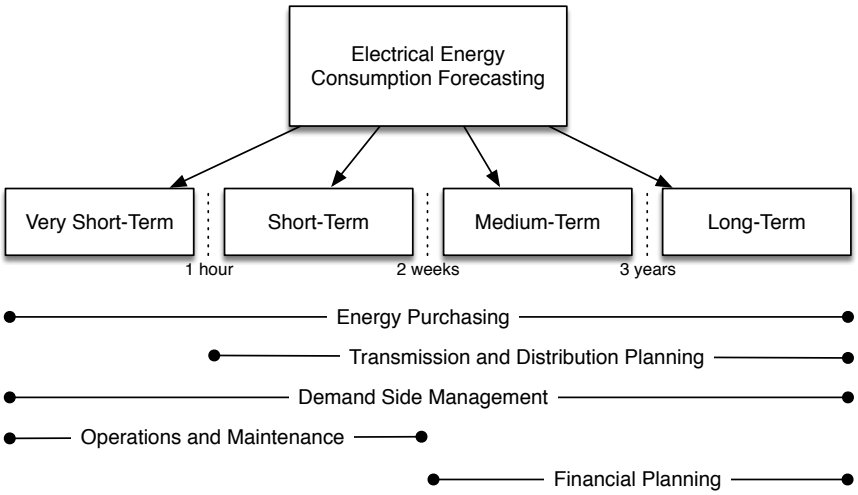

Figure 1.1. EF applications and classification (Zor et al., 2017b)

The motivation of the thesis are expressed as follows:

- After the deregulation of electricity, STEF is a necessity of energy management and planning of all buildings from households and residences in the small-scale to huge building complexes in the large-scale.
- In Turkey, hospitals are seeking for renewable energy investment opportunities such as installation of a PV system to supply electricity as a reliable alternative to the grid and to sell excess electrical energy to the grid with an advantageous tariff price containing incentives which starts from 13.3 US cents/kWh up to 20.0 US cents/kWh in case

of using domestically manufactured equipment (EMRA, 2019).

- The hospitals are also planning to install tri-generation plants (so called CCHP plant which stands for combined cooling, heat, and power plant) fuelled by natural gas due to a directive (Teksan et al., 2017) by Republic of Turkey Ministry of Health which states that hospitals equipped with more than 200 beds and spanning over 20,000 m^2 of closed area shall install either co-generation (so called CHP which stands for combined heat and power) or tri-generation plant in order to supply their own energy demands for improving energy efficiency in health facilities (Teke and Timur, 2014; Teke et al., 2015).
- Before installing the mentioned plants for generating electricity to hospitals operating in a deregulated environment; STEF, which is implemented by statistical and AI (S&AI) techniques, is an essential tool that is not only required for the integration of smart grids to present electric power systems, but also improves hospital's energy management and planning quality by monitoring energy consumption, finding base and peak demands, reducing losses, minimising risks, ensuring reliability for continuous operation, taking an active role in making viable decisions related to maintenance planning and future investments including both renewable and non-renewable energy technologies such as PV, landfill, tri-generation fuelled by natural gas.

1.2. Objectives

The main objectives of the thesis are described as follows:

- To acquire real-time electrical energy consumption and meteorological data of a large hospital complex by utilising an energy logger

connected to a humidity-temperature transducer on-site and MERRA-2 data to form a very-short term raw data set with a sampling period of 10 minutes in RStudio environment,

- To develop a novel methodology named as forecast time horizon converter (FTHC) which has the capability of identifying missing and erroneous values in the raw data set, replacing these values with NA values, imputing the NA values with different imputation methods, and completing the conversion process by creating a cleansed short-term data set with a sampling period of 1 hour,
- To forecast short-term electrical energy consumption of the hospital by using multiple linear regression (MLR) as a statistical technique for benchmarking purposes and AI based techniques containing support vector machines (SVM), gene expression programming (GEP), gradient boosted decision trees (GBDT), and artificial neural networks (ANN) consisting of multilayer perceptron neural networks (MLPNN), radial basis function neural networks (RBFNN), generalised regression neural networks (GRNN), and group method of data handling polynomial neural networks (GMDHNN) under identical constraints, and to calculate relative importance of input variables,
- To present a benchmark analysis of the obtained results by evaluating coefficient of determination (R^2), coefficient of variation (CV), mean absolute error (MAE), root mean squared error (RMSE), and mean absolute percentage error (MAPE).

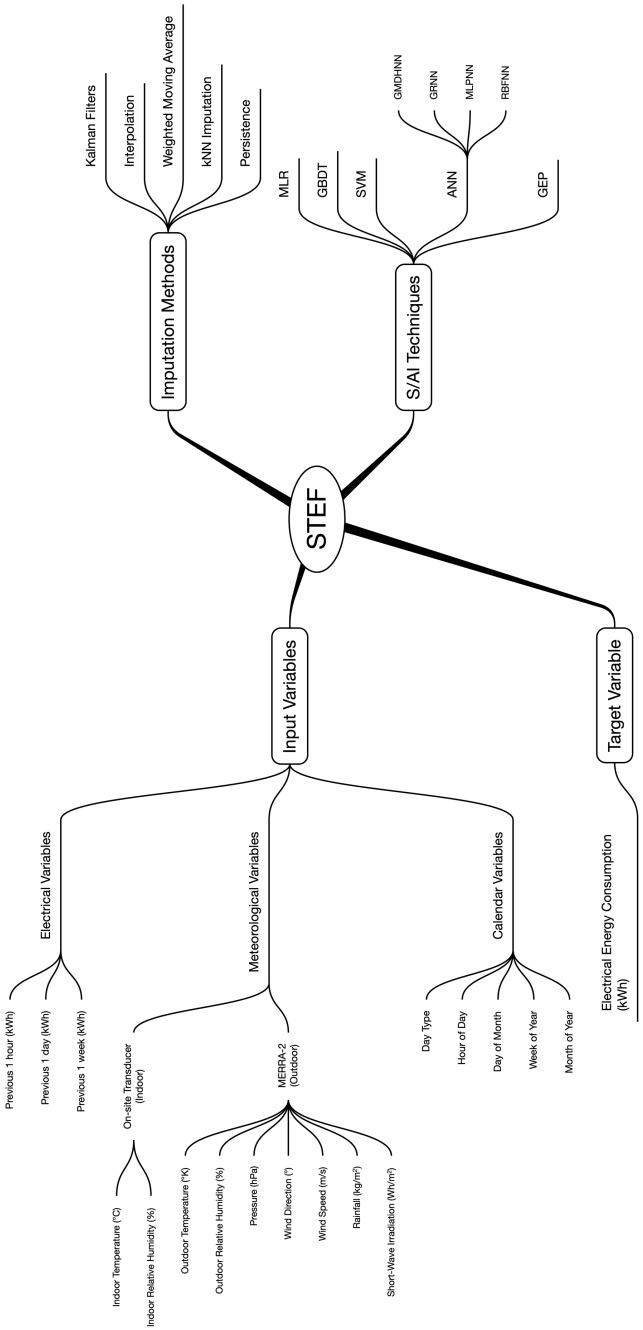

Figure 1.2. Steps of the applied methodology

1.3. Methodology

Steps of the applied methodology for the thesis are summarised in the previous figure and given as follows:

- Primarily, in order to form a very-short term raw data set with a sampling period of 10 minutes in RStudio environment wherein R programming language is used, an energy logger connected to a humidity-temperature transducer is utilised for gathering real-time electrical energy consumption and meteorological data on-site, and MERRA-2 data.
- Very short-term raw data set has 3 input variable categories and 1 target variable. The input variable categories are electrical, meteorological, and calendar variables. Target variable is the actual electrical energy consumption. Electrical variables can be stated as the historical electrical energy consumption variables belonging to previous 10 minute, previous 1 day (the same time in the previous day), and previous 1 week (the same time and day in the previous week). Meteorological variables are obtained from two different sources. The first source is on-site humidity-temperature transducer which provides the indoor relative humidity and temperature of the ambient where the transducer is located with the energy logger. The second source is MERRA-2 data consisting of outdoor temperature, relative humidity, pressure, wind direction, wind speed, rainfall, and short-wave irradiation. Calendar variables include month of year, week of year, day of month, hour of day, sample number of hour, and day type. The very short-term raw data set is represented as a $52,416 \times 19$

matrix in RStudio environment.

- Afterwards, the created very-short term raw data set is converted to a cleansed short-term data set by applying a novel methodology named as FTHC which is developed to identify missing and erroneous values in the very short-term data set, replace these values with NA values, impute the NA values with different methods, and complete the conversion process.
- Missing and erroneous values in the very short-term raw data set is occurred due to power outages in the hospital. For this case, the actual variables where data are missing are not the cause of the incomplete data. Instead, the cause of the missing data is due to some other external influence which is power outage for the case of hospital, and hence missing data mechanism for this case is missing at random (MAR).
- The number of variables possessing NA values for the very short-term data set is 6 and they are actual, previous 10 minute, previous 1 day, and previous 1 week electrical energy consumption, indoor relative humidity, and indoor temperature acquired from the energy logger and the humidity-temperature transducer. The number of rows having NAs varies between 379 and 398, while the proportion of NAs for each variable changes from 0.723% to 0.759%. In total, 2,333 NAs in a $52,416 \times 19$ matrix correspond to a proportion of 0.234%.
- To the best of one's knowledge, first and foremost there is no certain threshold to ignore the imputation, but the most of the literature suggests applying complete case analysis (CCA) to a data set with a missing data mechanism of MAR if the proportion of missing data is

below 5%. However, imputation looks very tempting, because performing CCA by using listwise deletion to a data set with a missing data mechanism of MAR sometimes wastes a whole row for just one NA value in the row and frequently introduces bias which causes a loss in efficiency. Therefore, Kalman filters, interpolation, weighted moving average, kNN imputation, and persistence methods are employed to compare the performances of different imputation methods and reveal the impacts of each method on a data set especially prepared for energy forecasting. After the completion of the conversion process, the cleansed short-term data set is represented as $8,736 \times 18$ matrix in RStudio environment and previous 10 minute is renamed as previous 1 hour because of the conversion. In addition to those, sample number of hour variable is not used in the short-term data set.

- Next, STEF of the hospital is implemented by using MLR as a statistical technique for benchmarking purposes and AI based techniques containing SVM, GEP, GBDT, and ANN consisting of MLPNN, RBFNN, GRNN, and GMDHNN under identical constraints such as random sampling for test and validation. Moreover, an algorithm employing sensitivity analysis is performed to all S&AI techniques for computation of relative importance of input variables where the values of each variable are randomised and the effect on the quality of the model is measured out of hundred.
- Consequently, a benchmark analysis of the achieved results, in which GBDT surpass other S&AI techniques, while imputed data sets with KalmanARIMA, linear interpolation, and next observation carried

backward are ranked among the top three, is presented meticulously by R^2, CV, MAE, RMSE, and MAPE.

1.4. Contributions

The main contributions of the thesis are detailed as follows:

- It is the first time in Turkey, an application of real-time electrical energy forecasting study with comprehensive meteorological observations and high reliability is conducted for a large-scale nonindustrial building complex in order to create a novel data set with sampling periods in accordance with very short-term and short-term horizon by utilising FTHC. The created data set may be used for undergraduate and graduate level courses in relation with management, planning, economics, and analytics of electrical energy in electrical and electronics engineering. For future studies, researchers may benefit from the data set for electrical energy consumption, electric load and demand studies from 10 minutes to 1 hour ahead forecasting by the FTHC. In addition to those, studies in the STEF literature is limited especially for real-time applications, and this thesis is thought to fill the gap in the STEF literature.
- R programming language and thereby RStudio are valuable assets in terms of S&AI applications, data manipulation, wrangling, and visualisation. Despite this, it is considered that electrical energy studies in the related literature do not give R the well-deserved credit for its capabilities. One of the key contribution of the thesis is using R from beginning to end for a sophisticated problem in power engineering by presenting R from the electrical and electronics

engineering's point of view.

- Missing data are almost never processed in the energy forecasting literature apart from a few studies. One of the most significant contribution of the thesis is to investigate the individual effects of a variety of imputation methods on an energy forecasting data set by not just performing CCA and deleting the rows having NAs. To the best of our knowledge, this thesis is the first study in the literature that implements Kalman filtering, interpolation, weighted moving average, persistence, and kNN imputation approaches to missing data for a real-time STEF problem. It is suggested that KalmanARIMA and linear interpolation may be applied to energy forecasting data sets for STEF problem in future studies.
- The novel FTHC methodology debuting in the thesis is mainly developed for energy forecasting applications due to the steady nature of electric power systems where fluctuations are evaluated as unwanted circumstances. Nonetheless, it is considered that deviation and tolerance structure of the FTHC may be employed for applications where sudden changes in the magnitude of parameters are undesirable.
- An algorithm including sensitivity analysis is used to calculate the relative importance of input variables. Comprehensive analyses conducted by using the algorithm indicated that previous 1 hour, short-wave irradiation, and hour of day variables should be underlined as prerequisites among all input variables for one hour ahead STEF. By the way, electrical variables containing previous 1 hour, 1 day, and 1 week; meteorological variables including short-wave irradiation, indoor and outdoor temperatures and rainfall; and calendar variables

consisting of hour of day, week of year, and month of year should be at least provided for one hour ahead STEF.

- STEF is performed for a 1-year period using S&AI techniques under identical constraints. Carrying out analyses with the same criteria unveils the genuine performance of the S&AI techniques for a better comparison in terms of R^2, CV, MAE, RMSE, and MAPE. It can be inferred from the acquired detailed results that GBDT delivered an outstanding performance by any measure in comparison with other S&AI techniques for STEF. Therefore, it is recommended that GBDT should employ for STEF more frequently.
- Obtained results from benchmark analyses fulfilled in the scope of this work are invaluable in reference to administrations of both university and hospital owing to the fact that they may utilise this work for hospital's interest not only in the forthcoming integration of the smart grid, but also in improving energy management and planning quality by reducing losses, minimising risks, securing reliability for uninterrupted operation, making viable decisions particularly for maintenance scheduling and future investments related to PV, landfill, or natural gas fuelled tri-generation.
- The last contribution of the thesis is to find out the correlativity of seasonality and electrical energy consumption of heating, ventilation, and air-conditioning (HVAC) systems. The experimental results revealed that seasonality has a dominant influence on STEF performance of the hospitals in which HVAC systems constitute the major part of electrical energy consumption. On the other hand, it is seen that day type variable, which determines whether a day is either

weekend and public holiday or working day, is a preeminent classifier in STEF of hospitals thanks to the operation of polyclinics.

Much of the work in this thesis is an extension on the work presented in the following publications:

- Teke, A., Zor, K., and Timur, O., 2015. A simple methodology for capacity sizing of cogeneration and trigeneration plants in hospitals: A case study for a university hospital. Journal of Renewable and Sustainable Energy, 7(053102):1–15. DOI: 10.1063/1.4930064
- Zor, K., Timur, O., and Teke, A., 2017. A state-of-the-art review of artificial intelligence techniques for short-term electric load forecasting. Proceedings of the 6th International Youth Conference on Energy (IYCE2017), Budapest, 1–7. DOI: 10.1109/IYCE.2017.8003734
- Zor, K., Timur, O., Çelik, Ö., Yıldırım, H. B., and Teke, A., 2017. Interpretation of error calculation methods in the context of energy forecasting. Proceedings of 12th Conference on Sustainable Development of Energy, Water and Environment Systems (SDEWES2017), Dubrovnik, 0722:1–9.
- Zor, K., Çelik, Ö., Timur, O., Yıldırım, H. B., and Teke, A., 2018. Simple approaches to missing data for energy forecasting applications. Proceedings of the 16th International Conference on Clean Energy (ICCE-2018), Gazimağusa, FORC-03:1–4.
- Zor, K., Timur, O., Çelik, Ö., Yıldırım, H. B., and Teke, A., 2018. Very Short-Term Electrical Energy Consumption Forecasting of a Household for the Integration of Smart Grids. Official Conference Proceedings of

the European Conference on Sustainability, Energy & the Environment 2018 (ECSEE2018), Brighton, 1–14.

- Timur, O., Zor, K., Çelik, Ö., Teke, A., and İbrikçi, T., In Press. Application of Statistical and Artificial Intelligence Techniques for Medium-Term Electrical Energy Forecasting: A Case Study for a Regional Hospital. Journal of Sustainable Development of Energy, Water and Environment Systems, 1–17. DOI: Unassigned

1.5. Organisation of the Thesis

Organisation of the thesis is summarised and demonstrated as follows:

- **Chapter 1:** An introduction to the thesis with elucidate designs is presented by stating background and motivation, objectives, methodology, original contributions, and organisation of the thesis.
- **Chapter 2:** A brief summary of related literature including PhD theses, journal and conference publications are given meticulously.
- **Chapter 3:** In this chapter, data acquisition stage for the realisation of STEF, a brief introduction to R programming language, and data wrangling and visualisation with RStudio are rigorously explained.
- **Chapter 4:** The novel FTHC is expressed in details by the help of an explanatory flowchart and graphics with high visuality.
- **Chapter 5:** Individual effects of a variety of imputation methods on a real-time energy forecasting data set is investigated throughout this chapter.
- **Chapter 6:** MLR as a statistical technique and AI based techniques containing SVM, GEP, GBDT, and ANN consisting of MLPNN, RBFNN, GRNN, and GMDHNN are thoroughly processed.

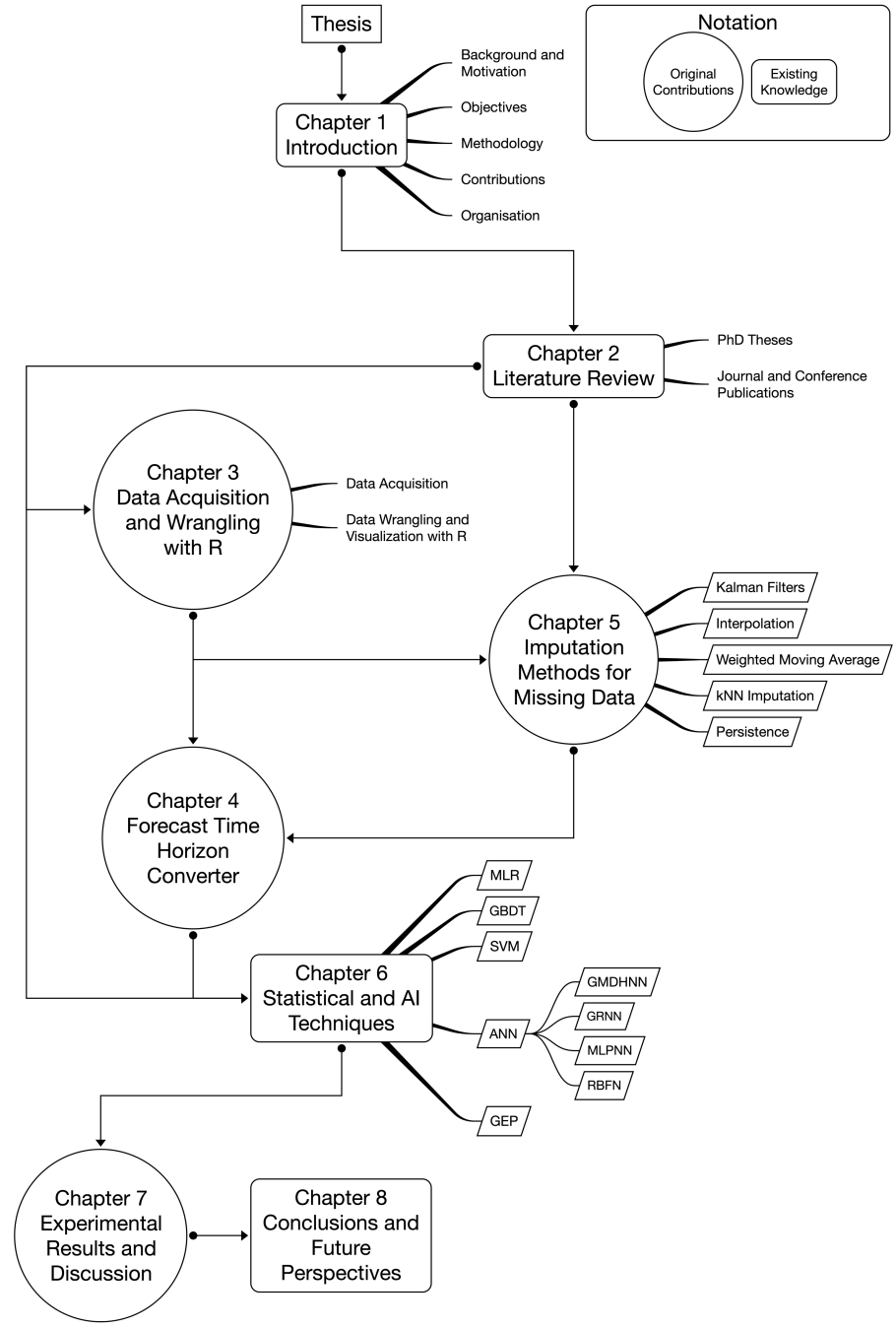

Figure 1.3. Organisation of the thesis

- **Chapter 7:** The detailed results of benchmark analyses are discussed during this chapter.
- **Chapter 8:** Conclusions and future perspectives belonging to the thesis are presented respectively.

In the following chapter, literature review of the thesis is given according to journal and conference publications, and PhD theses.

2. LITERATURE REVIEW

In this chapter, a brief summary of related literature including PhD theses, journal and conference publications, which are contained by ProQuest Dissertations and Theses Database, Institute of Electrical and Electronics Engineers (IEEE) Xplore and ScienceDirect Digital Libraries, and other databases and libraries, are presented meticulously.

2.1. Related PhD Theses in the Literature

The related theses in the literature start with Al-Madfai's thesis work named as "Weather Corrected Electricity Demand Forecasting" which was published in 2002. According to Al-Madfai, the complexity of short-term load forecasting (STLF) problem is based on the multiple seasonal components, the change in consumer behaviour during holiday seasons, and other social and religious events which affect electrical energy consumption. The purpose of his research was to create models for electric demand that can be utilised to further the understanding of the dynamics of electrical energy consumption in South Wales, to generate weather corrected forecasts, and to obtain short-term load forecasts. In his work, a novel explanatory collective measure of temperature, entitled "Fair Temperature Value" (FTV) was introduced, and two time series modelling approaches, namely profiles auto-regressive integrated moving average (PARIMA) and variability decomposition method (VDM) were employed to model the daily, weekly, monthly, and quarterly demand series. VDM model with FTV was used for weather correction because of the fact that it was the most successful model in the analyses (Al-Madfai, 2002).

Topallı proposed a method which employs recurrent neural networks (RNN) in order to perform one day ahead forecasting for Turkey's total electric load. In her

work, she developed a hybrid learning scheme, which integrates off-line learning with real-time forecasting, to utilise historical data for adjusting the weights and neuron relations with respect to the changing conditions. She offered a solution for all load types such as working days, weekends, and public holidays, and she also suggested a technique based on principal component analysis (PCA) for the selection of input variables. As a consequence, she compared her model with an auto-regressive moving average (ARMA) model for the same data, and the proposed method provided lower errors (1.60% as MAPE), especially for loads belonging to public holidays (Topallı, 2003).

Owing to the introduction of deregulated market structure, Fay presented a strategy to reduce costs from electric demand forecast error employing models developed particularly for the Irish system under three categories including data segmentation, modelling technique, and the approach to minimise the impact of errors exist in weather inputs. In his work, a variety of segmentation strategies were analysed and the one suitable for Irish data was determined. Linear and non-linear techniques are compared to find the optimal model (2.89% as MAPE), and a novel method is offered for minimising the impact of weather forecast errors (Fay, 2004).

Yang suggested applying forward second order difference (FSOD) and backward second order difference (BSOD) in Shanghai Power Grid and German data to detect bad sectors between the adjacent continuous segments by regressing in a quadratic form for power system STLF. After detecting the bad data and replacing them with reasonable data, regression trees (RT), support vector regression (SVR), and a combination method of RT and SVM were proposed (2.63% as MAPE). According to Yang, proposed methods should not only be limited to STLF, but also be used for other horizons of load forecasting (Yang, 2006).

Hassnain preferred to use particle swarm optimisation (PSO) algorithm in

order to train ANN due to the drawbacks of back-propagation (BP) algorithm for STLF. In addition, Hassnain employed modularised approach for capturing the trend to overcome the effects of seasonality. From the point of view of Hassnain, the developed approach gave better trained models and resulted in fairly accurate forecasts (Hassnain, 2009).

According to Hong, with the advancements in the smart grid technologies, load forecasting comes into prominence by the reason of its broad applications in the planning of demand side management, electric vehicles, distributed energy resources, and so on. In his work, he proposed an integrated forecasting framework with the focus on STLF engine which can be easily connected to several forecasts. Hong implemented MLR, possibilistic linear model (PLM), and ANN based load forecasters, and compared them as a benchmark (2.96% as MAPE) for STLF (Hong, 2010).

Zhao applied a variety of SVM methods for the prediction of building energy consumption. An approach based on recursive deterministic perceptron neural networks was proposed for the enhancement of a diagnostic method which may detect faults of a particular equipment. Also, a feature selection method is suggested to reduce the input dimension of SVR. Lastly, a new parallel approach was developed to optimise the training of SVM for multi-core and multiprocessor systems, and compared the approach with Libsvm and Pisvm. According to the benchmark tests conducted in Zhao's thesis, the new approach overwhelmed the others (Zhao, 2011).

Matijas applied meta-learning to electric load forecasting on two levels which are meta-level and task level. For forecasting multivariate time series, the proposed approach employed an ensemble of seven algorithms for classification at the meta-level, and seven different algorithms at the task level of load forecasting (Matijas,

2013).

Guan presented a method of multilevel wavelet neural networks with data prefiltering. The main idea was to utilise a spike filtering technique in order to perceive load spikes and smooth them without changing the ordinary load. After filtering, wavelet decomposition was employed and separate neural networks were implemented (Guan, 2013).

Ben Taieb managed to narrow the gap at the intersection of time series forecasting and machine learning by addressing the problem of multi-step-ahead time series forecasting from the perspective of machine learning. In his work, he conducted a detailed study by decomposing the multi-step mean squared forecast errors into the bias and variance components and analysing their behaviour over the forecast horizon for different lengths of time series. He also developed multi-stage forecasting strategies that seek to combine the best features of the recursive and direct strategies. As a last contribution, he created and analysed multi-horizon forecasting strategies that utilise information of other horizons while learning the model for each horizon (Ben Taieb, 2014).

Grant designed a real-time energy monitoring system for a large government building to reduce the peak demand. Data set was obtained from the monitoring system, and implementation of ANN model (3.9% as MAPE) was compared to simple moving average (SMA), linear regression, and multivariate adaptive regression splines (MARSplines) models (Grant, 2014).

Tuunanen suggested a methodology, which consists of a spatial analysis, clustering, end-use modelling, and scenarios and simulation methods, to forecast electrical energy consumption in the distribution networks by using automatic meter reading (AMR) data. Tuunanen's thesis work also contained a case study, and for the case study, the effects of future energy technologies on the distribution network

were analysed by means of electrical energy and power (Tuunanen, 2015).

For STLF problem, Li presented two models based on ANN and modified statistical learning techniques (2.56% as MAPE) to predict future electric demands depending upon historical hourly load and hourly weather data (Li, 2015).

In Liu's thesis work, both point and probabilistic load forecasting were embroidered. In the thesis work, Tao's vanilla benchmark (TVB) model was conducted firstly, and then an STLF model with recency effect was developed to produce sister models and forecasts. Liu also reduced the computation duration of the forecasting process using high performance computing techniques (Liu, 2016).

Torres proposed three main contributions to the load forecasting literature such as improving the accuracy of forecast (3.73% as MAPE) and the adaptiveness of model, and automatising the execution of applied load forecasting strategies by implementing machine learning, computational intelligence, evolvable networks, expert systems, and regression approaches (Torres, 2017).

Lastly, Jain designed a Mamdani fuzzy inference system (FIS) for STLF with three input membership functions which are load, temperature, and humidity differences using trapezoidal memberships, and one output membership function that is the correction factor using triangular membership. Moreover, Jain combined clustering and regression to form a framework of clustering based regression methodology for STLF (Jain, 2018).

2.2. Related Journal and Conference Publications in the Literature

In this section, related journal and conference publications in the literature are mentioned. It should be noted that the literature is narrowed according to the following criteria:

- Firstly, "short-term" is determined as the time horizon of the literature

survey,

- Among all other applications, “building” is identified,
- Then, the combination of the terms “electric”, “electrical”, “electricity”, and “energy”; “power”, “demand”, “use”, and “consumption”; “forecast”, “forecasting”, “prediction”, “predicting”, “estimation”, and “estimating” are searched,
- Afterwards, studies appertaining to “household” and “residential” buildings are excluded,
- Finally, temporal granularity is limited as an hour or less than an hour.

After sorting irrelevant studies out rigorously, the number of the related journal and conference publications are counted as 60 and 18 respectively. Quantitative information of studies with respect to the name of publishers and publication type is visualised in Figure 2.1. Similarly, quantitative information of studies in reference to the name of journal and on-line access of proceedings is illustrated in Figure 2.2.

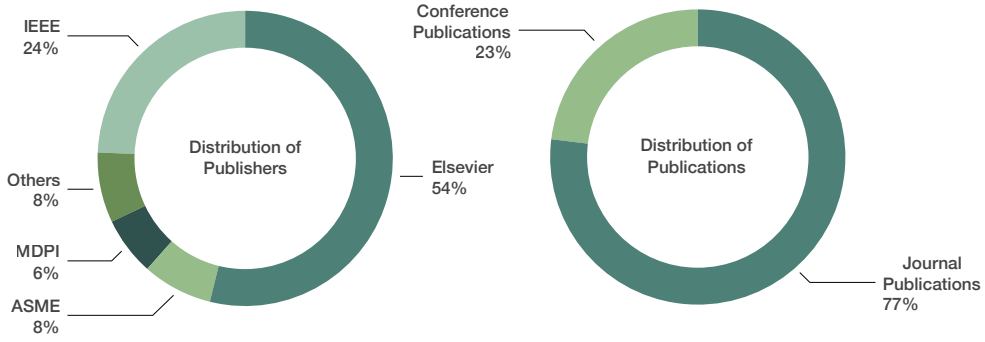

Figure 2.1. Distributions with respect to publisher name and publication type

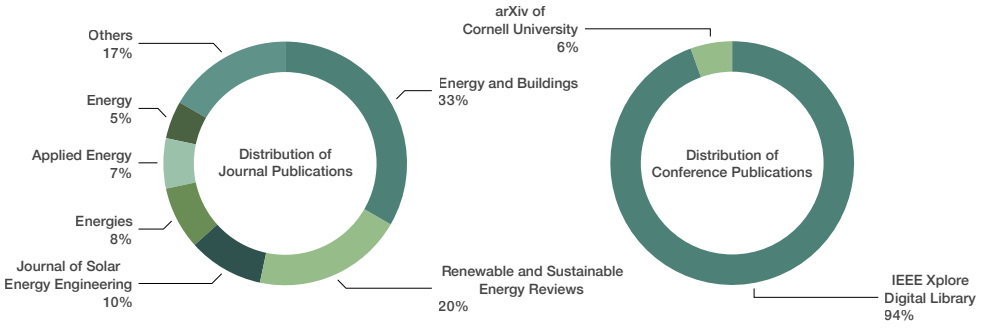

Figure 2.2. Distributions in reference to journal name and access of proceedings

As a beginning, Seem and Braun presented an adaptive algorithm containing a deterministic and a stochastic model with three auto-regressive (AR) parameters to forecast from one hour to 24-hour ahead electrical demand for a grocery store and a restaurant (Seem and Braun, 1991). Kreider et al. used RNN for one hour ahead building energy prediction of an engineering centre without historical energy consumption (Kreider et al., 1995). Dhar et al. employed Fourier series functional forms for modelling of both weather independent and dependent hourly energy use in commercial buildings (Dhar et al., 1998). Similarly, they performed temperature based Fourier series to predict hourly heating and cooling energy use in commercial buildings in which only meteorological variable is the outdoor temperature (Dhar et al., 1999a). Also, they suggested a generalised Fourier series approach for the modelling of hourly energy use in commercial buildings by separating the days of the year (Dhar et al., 1999b).

Kalogirou and Bojic utilised RNN with dampened feedback for the prediction of hourly energy consumption of a passive solar building (Kalogirou and Bojic, 2000). Krarti reviewed AI-based methods for building energy systems such as STLF and weather forecasting and systems modelling by ANN, controls of thermal energy storages by ANN and genetic algorithms (GA), and fault detection and

diagnostic by a fuzzy logic (FL) model based approach (Krarti, 2003). Dodier and Henze applied ANN to the building energy use data of the Energy Prediction Shootout II Contest by excluding irrelevant inputs according to the Wald test (Dodier and Henze, 2004).

Yalcintas and Akkurt managed to predict electrical energy consumption of a 42 storey building in downtown Honolulu, Hawaii by implementing ANN in consideration of climate and a variety of building functions (Yalcintas and Akkurt, 2005). Yang et al. proposed two adaptive ANN models with different training algorithms named as accumulative and sliding window training for the hourly energy prediction of synthetic data belonging to The Laval office building located in Montreal (Yang et al., 2005). Gonzalez and Zamarreno employed a feedback ANN model trained by means of a hybrid algorithm for the prediction of hourly energy consumption in buildings (Gonzalez and Zamarreno, 2005). Karatasou et al. predicted building hourly energy consumption acquired from two data sets consisting of the energy use data of the Energy Prediction Shootout I Contest and an office building in Athens by performing ANN (Karatasou et al., 2006). Neto and Fiorelli compared a physical model and ANN model in forecasting hourly energy consumption of the administration building of the University of Sao Paulo (Neto and Fiorelli, 2008). Pedersen et al. developed a load prediction method for heat and electricity demand in buildings for the planning of mixed energy distribution systems (Pedersen et al., 2008).

Zhao and Magoules brought parallel implementation of SVM in the literature to accelerate model training process in the prediction of multiple buildings energy consumption (Zhao and Magoules, 2010). They also introduced an implementation for multi-core and multiprocessor systems named as MRPsim which outperforms Libsvm and Pisvm by means of training speed in building

energy consumption predictions (Zhao and Magoules, 2011). Cherkassy et al. suggested a computational technique that combines regression and clustering methods for the prediction of electric power consumption of commercial buildings (Cherkassy et al., 2011). Li et al. compared ANN and hybrid neuro-fuzzy system for forecasting hourly energy consumption obtained from the Energy Prediction Shootout I and a library located in Zhejiang University, Hangzhou (Li et al., 2011). Mathieu et al. described new ways of visualising electric load data, introduced a time-of-week indicator variable for electric load regression models, avoided the use of change-point models but tried to capture a nonlinear relationship between electrical and meteorological variables, defined new parameters to characterise electric load profiles and demand response behaviour, and applied the modelling methods to evaluate demand response effectiveness (Mathieu et al., 2011).

Zhao and Magoules reviewed on the prediction of building energy consumption by highlighting the prediction methods including engineering methods, statistical methods, and AI methods (Zhao and Magoules, 2012a). In addition, they proposed a feature selection method for building energy consumption prediction that can ensure the prediction accuracy and reduces the SVR computational time for data analysing (Zhao and Magoules, 2012b). Durijic and Novakovic identified important variables of energy utilisation in low energy office building in Trondheim, Norway by implementing multivariate analysis (Durijic and Novakovic, 2012).

Twanabasu and Bremdal addressed demand side flexibility in a smart grid oriented building of Ostfold University in Halden by employing ARIMA, ANN, and SVM (Twanabasu and Bremdal, 2013). Yoo and Hur developed a load forecast model switching scheme which provides improved robustness to changes in building hourly electrical demand by utilising auto-regressive moving average with exogeneous inputs (ARMAX) and Kalman filtering (Yoo and Hur, 2013). Foucquier

et al. reviewed the state-of-the-art studies in building modelling and energy prediction by depicting physical models for building thermal behaviour modelling, statistical methods for machine learning, and hybrid models with their advantages and limitations (Foucquier et al., 2013). Roldan-Blay et al. enhanced an upgrade, which employs a time temperature curve forecast model, for ANN in forecasting hourly electrical consumption (Roldan-Blay et al., 2013).

Mocanu et al. presented a comparison of machine learning methods including conditional restricted Boltzmann machine (CRBM), ANN, and hidden Markov models (HMM) for estimating total and lighting energy consumption of a Dutch office building (Mocanu et al., 2014). Mai et al. applied RBFNN to Shenzen's 39 storey commercial office building's hourly electric load forecasting model utilising meteorological and historical load data (Mai et al., 2014). Gulin et al. implemented ANN to STLF of a building belonging to the Faculty of Electrical Engineering and Computing of the University of Zagreb for microgrid power flow optimisation (Gulin et al., 2014). Ahmad et al. reviewed on the applications of ANN and SVM for building EF by emphasising the potential of hybrid method merging GMDHNN and least squares SVM (LSSVM) (Ahmad et al., 2014). Jetcheva et al. employed neural network model ensembles for building-level hourly electricity load forecasts in comparison with seasonal auto-regressive integrated moving average (SARIMA) (Jetcheva et al., 2014). Lazos et al. presented a comprehensive review on the optimisation of energy management in commercial buildings with meteorological forecasting inputs (Lazos et al., 2014).

Vinagre et al. preferred an ANN based methodology for forecasting short-term electrical energy consumption by utilising external facility data obtained from supervisory control and data acquisition system along with meteorological variables (Vinagre et al., 2015). Heylman et al. acquired hourly electric power data

from over 200 buildings on the University of Virginia's campus to apply linear model and SARIMA in order to forecast energy trends and especially peak power (Heylman et al., 2015). Paudel et al. used pseudo dynamic approach with SVM in energy demand prediction of Ecole des Mines de Nantes office building in France (Paudel et al., 2015). Liu et al. implemented a time series forecasting method for building hourly energy consumption employing SVR (Liu et al., 2015). Raza and Khosravi conducted a review on AI based load demand forecasting techniques for smart grid and buildings by explaining all stages of STLF in details (Raza and Khosravi, 2015). Touretzky and Patil used ARMAX models and physics based modelling approaches for forecasting power demand of a commercial building (Touretzky and Patil, 2015). Li et al. performed an improved PSO algorithm to modify the weights and thresholds of ANN along with PCA for electricity consumption prediction of two data sets containing Energy Prediction Shootout Contest I and a campus building situated in East China (Li et al., 2015). Tardioli et al. summarised data-driven approaches for building energy consumption prediction at urban level by classifying prediction methods as white-box, gray-box, and black-box approaches (Tardioli et al., 2015). Platon et al. conducted hourly electricity consumption prediction of an institutional Canadian facility located in Calgary, Alberta by applying case-based reasoning and ANN as AI techniques and PCA for feature selection of inputs (Platon et al., 2015). Chitsaz et al. proposed self-recurrent wavelet neural network with Levenberg-Marquardt learning algorithm for hourly electric load forecasting of a building within a microgrid (Chitsaz et al., 2015). Massana et al. implemented MLR, MLPNN, and SVR to forecast short-term loads of office building of University of Girona, and analysed to find the most relevant variable in forecast (Massana et al., 2015).

Manjhi and Dhar performed a hybrid algorithm containing PSO in the

training stage of ANN and gravitational search algorithm in the optimisation stage of ANN to short-term energy consumption forecasting (Manjhi and Dhar, 2016). Ke et al. employed three different method consisting of direct fitting method, similar day approach, and MLR to forecast short-term loads of Centennial Campus of North Carolina State University (Ke et al., 2016). Khosravani et al. compared nonlinear auto-regressive with exogeneous inputs (NARX) ANN model and ANN based models generated by multi objective genetic algorithm (MOGA) to predict short-term energy consumption of Solar Energy Research Centre (CIESOL) located in the south-east of Spain (Khosravani et al., 2016). Ruiz et al. presented a case study application of nonlinear auto-regressive (NAR) neural networks and NARX neural networks for energy consumption prediction of public buildings in University of Granada (Ruiz et al., 2016). Chae et al. performed ANN model with Bayesian regularisation algorithm to forecast sub-hourly electricity usage in a commercial office building complex (Chae et al., 2016). Sarduy et al. employed linear and nonlinear models to forecast the peak load of a campus of the University of Sao Paulo in order to choose the best one for generalisation (Sarduy et al., 2016). Massana et al. implemented SVR for STLF of non-residential buildings by highlighting artificial occupancy attributes (Massana et al., 2016). Zhang et al. utilised the energy data belonging to a university campus in Singapore to forecast half-hourly electrical energy consumption by using weighted SVR with differential evolution algorithm (Zhang et al., 2016).

Xypolytou et al. focused on the accurate energy consumption prediction of office buildings and conducted a case study by applying ANN (Xypolytou et al., 2017). Yildiz et al. presented a review and analysis of regression and machine learning models for electricity load forecasting of Kensington Campus and Tyree Energy Technologies Building at the University of New South Wales, and according

to their study ANN with Bayesian regulation BP overwhelmed other models (Yildiz et al., 2017). Wang and Srinivasan contrasted single and ensemble prediction models for AI based building energy consumption prediction within a review study (Wang and Srinivasan, 2017). Deb et al. conducted an in-depth review on times series forecasting techniques for building energy consumption by expressing advantages and disadvantages of each technique (Deb et al., 2017). Daut reviewed on the building electrical energy consumption forecasting analysis using conventional and AI methods rigorously (Daut et al., 2017). Li et al. proposed an extreme deep learning (DL) approach, which is a combination of stacked autoencoders and extreme learning machines (ELM), and compared with back-propagation neural networks (BPNN), SVR, generalised RBFNN, and MLR for energy consumption forecasting of a retail store in Freemont, CA (Li et al., 2017). Pombeiro et al. compared MLR, fuzzy modelling, and ANN to assess low-complexity models for the prediction of electricity consumption in an institutional building (Pombeiro et al., 2017). Pino-Mejias et al. tried to develop and compare linear regression models and ANN in order to predict the electrical energy consumption and other demands of office buildings in Chile (Pino-Mejias et al., 2017). Molina-Solana et al. carried out a review to indicate how data science has been implemented to solve the most difficult problems in the field of energy management, particularly for the building-scale (Molina-Solana et al., 2017). Liu et al. proposed sliding window empirical mode decomposition, a new feature selection algorithm, and a hybrid forecast engine for building energy consumption prediction (Liu et al., 2017). Chen et al. used SVR model for STLF to calculate the baseline of the demand response for office buildings (Chen et al., 2017). Chen and Tan created a hybrid model combining wavelet decomposition and SVR for the application of hourly electric demand forecasting to buildings including a hotel and a mall (Chen

and Tan, 2017). Ahmad et al. compared ANN and random forest models for HVAC short-term electrical energy consumption prediction (Ahmad et al., 2017).

Naug and Biswas implemented long short-term memory (LSTM) networks as a data driven method to predict energy demand of commercial buildings and they also compared the performance of LSTM networks with SVR and AdaBoost regression on a data set belonging to Alumni Hall building at Vanderbilt University (Naug and Biswas, 2018). Chandramitasari et al. preferred a combination of LSTM neural networks and ANN to forecast half an hourly electricity consumption of a manufacturing company in Japan (Chandramitasari et al., 2018). Nichiforov et al. applied LSTM layers to RNN for forecasting hourly electric loads of two buildings from university campuses in Chicago and Zurich (Nichiforov et al., 2018). Wang et al. employed ensemble bagging trees to predict hourly energy consumption of Leadership in Energy and Environmental Design (LEED) Gold certificated Rinker Hall building in the University of Florida (Wang et al., 2018a). Wei et al. presented a review of data-driven approches for prediction and classification of building energy consumption by mentioning practical applications of the approaches (Wei et al., 2018). Similarly, Amasyali and El-Gohary reviewed data-driven building energy consumption prediction studies by particularly focusing on the scopes of prediction, data properties and preprocessing methods, machine learning algorithms, and performance measures (Amasyali and El-Gohary, 2018). Sala-Cardoso et al. developed a hybrid methodology using RNN and adative neuro-fuzzy inference system (ANFIS) for STLF of HVAC thermal power demand and tested the metholodogy in a smart building which is also a research ecosystem of the Polytechnic University of Catalonia (Sala-Cardoso et al., 2018). Li et al. enhanced a modified deep belief network based hybrid model to predict energy consumption of buildings and contrasted the enhanced model with BPNN, generalised RBFNN,

ELM, and SVR by testing the model on a data set of a retail store in Freemont, CA (Li et al., 2018). Seyedzadeh et al. reviewed four machine learning techniques containing ANN, SVM, Gaussian process and mixture models, and clustering algorithms for estimation of building energy consumption and performance (Seyedzadeh et al., 2018). Wang et al. utilised random forest approach for hourly energy consumption prediction of two institutional buildings, namely Rinker Hall Building and Fine Arts Building C in the University of Florida, and compared with RT and SVR approaches (Wang et al., 2018b).

Haque et al. conducted a study to determine the impact of HVAC set point adjustments on building-level by performing SVR based hourly electric load forecasting in a commercial building in Chicago area (Haque et al., 2019). With the same data set, Jing et al. applied Levenberg-Marquardt, scaled conjugate gradient (SCG) BP, and Bayesian regularisation training algorithms to ANN-based building-scale hourly electric load forecasting from HVAC point of view (Jing et al, 2019). Fan et al. employed DL based feature engineering methods along with MLR, ANN, SVR, and extreme gradient boosting trees (XGBoost) for an improved energy prediction for an educational building in Hong Kong (Fan et al., 2019). Shan et al. proposed an ensemble prediction model integrating the gated recurrent unit (GRU) model and the proposed logarithmic electricity consumption gravity model to forecast hourly electricity consumption of a five-star hotel building in Shangai and an office building in Hangzhou by comparing 9 benchmarks mainly containing generalised linear regression, SVM, nearest neighbour, decision trees (DT), MLPNN, ARIMA, and LSTM (Shan et al., 2019). Liu et al. introduced a deep reinforcement learning algorithm named as deep deterministic policy gradient for hourly energy consumption prediction of a HVAC system belonging to a 9 storey office building in Henan, China (Liu et al., In Press).

In addition to the related journal and conference publications, studies in energy forecasting literature for hospitals or healthcare facilities are very limited and specifically expressed as follows:

Chen et al. proposed STEF of air-conditioners of a hospital by using ANN for three scenarios (Chen et al., 2005). Morinigo-Sotelo et al. presented STEF by using MLPNN with a sigmoid activation function for a hospital in Castile and Leon region of Spain (Morinigo-Sotelo et al., 2011). Bagnasco et al. performed a day ahead STEF by using ANN for both a large university hospital located in Rome (Bagnasco et al., 2014) and the Cellini medical clinic of Turin (Bagnasco et al., 2015). Guillen-Garcia et al. presented a methodology for VSTEF in a hospital in Castile and Leon region of Spain considering harmonics, inter-harmonics, and power quality disturbances (Guillen-Garcia et al., 2017). Damrongsak et al. analysed the factor impacts on the energy usage of 14 hospitals in Thailand by executing MLR for MTEF (Damrongsak et al, 2018). Gordillo-Orquera et al. performed MTEF in a hospital and primary care centre in Fuenlabrada, Madrid for a 1-year horizon by using multivariate analysis (Gordillo-Orquera et al., 2018).

In the next chapter, data acquisition, wrangling, and visualisation with R are explained.

3. DATA ACQUISITION AND WRANGLING WITH R

In this chapter, data acquisition stage for the realisation of STEF, a brief introduction to R programming language, and data wrangling with RStudio are rigorously explained.

3.1. Data Acquisition

A general information about the hospital, which is the data acquisition terminal and case study of the thesis, has to be mentioned before explaining the data acquisition stage of STEF.

3.1.1. General Information about Hospital

Hospitals may be described as highly sophisticated organisations from the point of view of functional, technological, economic, managerial, and procedural aspects. The reliability of continuous energy flow has utmost importance for hospitals owing to their uninterrupted duty for 24/7 operation without any excuses.

With its full name, Çukurova University Balcalı Health Application and Research Hospital, is a large hospital complex and a pioneer health institution situated in Campus Balcalı of Çukurova University in Sarıçam district of Adana, Turkey.

Since 1987, the hospital has been serving unceasingly to a region in the Southern Turkey that covers the area containing Adana, Mersin, Hatay, Osmaniye, Kahramanmaraş, Gaziantep, and Kilis. Therefore, it has nonstop demands to supply electricity for one emergency service, forty-two polyclinics, twelve intensive care units, twenty-three operating rooms, forty-three clinical services, five laboratories, one radiology unit, nuclear medicine, one blood centre, one burn unit, one

sterilisation unit, and one pharmacy, laundries, kitchens, and a morgue (Timur et al., In Press).

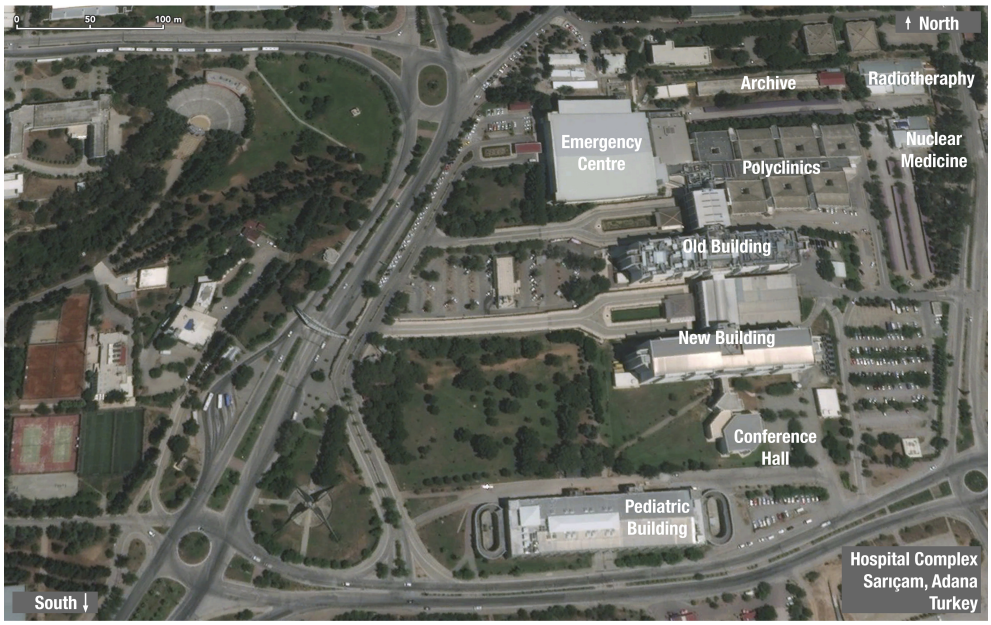

Figure 3.1. Aerial view of the hospital complex

The hospital has 1,200 beds, serves more than 3,500 patients per day with over 4,000 academic and administrative staff, and has an installed transformer capacity around 18 MVA (Zor, 2015; Timur, 2018). Aerial view of the hospital is illustrated in Figure 3.1.

3.1.2. Data Acquisition Stage

Data acquisition terminal for the hospital is the medium-voltage switchgear building where electricity meter of the hospital is located. The building is situated in the same campus with the hospital.

Geographical information related to the building is as follows:

- Latitude : 37.05891623°N,
- Longitude: 35.36113376°E,
- Altitude : 144.774 meters.

Energy logger connected to the humidity-temperature transducer properly logged electrical energy consumption, indoor temperature, and indoor humidity data between 13:52:57 on 22 September 2017 to 11:10:00 on 11 December 2018. In order to have historical electrical energy consumption data such as a certain time in a certain day of the previous week, starting period of data is specified as "2017-10-02 00:00:00" in POSIXct format in RStudio (to include 25 September 2017's data as the previous week's same day). Thus, data period of the thesis is also determined as between 2 October 2017 and 1 October 2018 as 1-year (52,416 rows).

Energy logger conducts logging by using the connections of current and voltage transformer in the terminal box of the electricity meter. Energy logger settings are adjusted to the multiplying factors of current and voltage transformers properly.

Equipment list contains the equipment necessary for the data acquisition is given in Table 3.1. In addition, data acquisition stage is visualised in Figure 3.2 by a connection schematic that combines hardware and software components.

Table 3.1. Equipment List

Equipment Name	Features
Energy Logger	Portable, Three-phase with USB Memory with Auxiliary Cable with High Sensitive Current Probes with Kensington Lock Mechanism
Humidity-Temperature Transducer	4-20 mA Dual Output

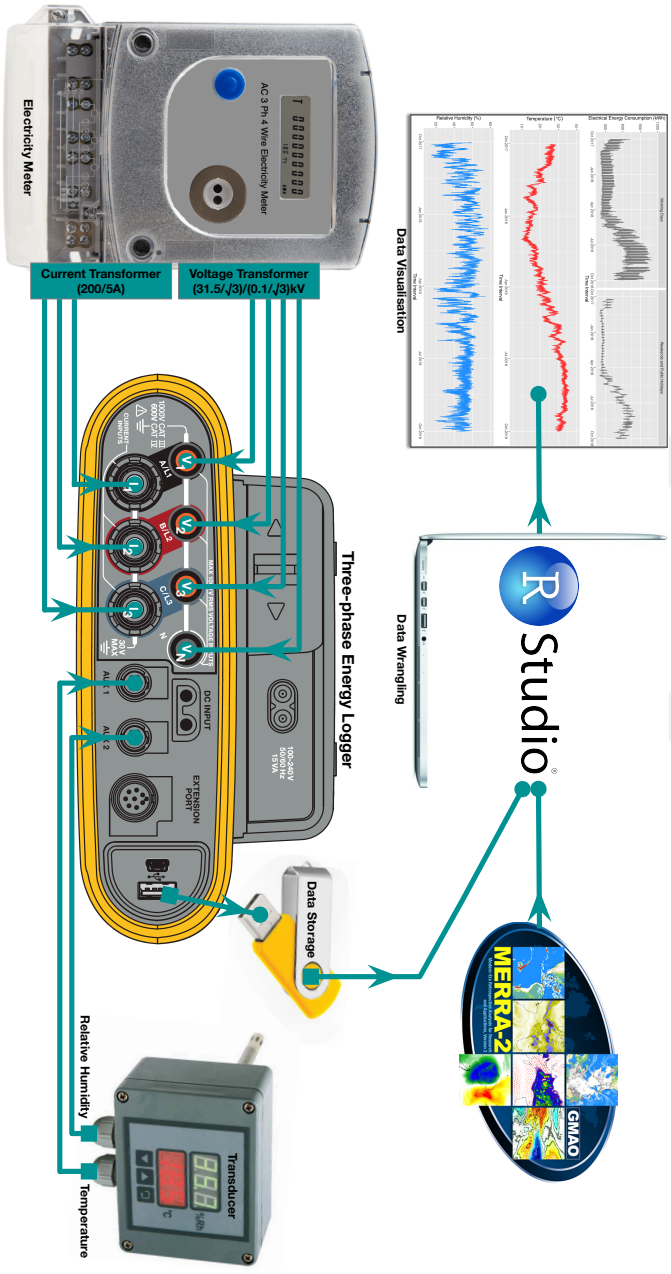

Figure 3.2. Demonstration of data acquisition stage

By using the equipment in Table 3.1; voltage, current, total harmonic current and voltage distortion, frequency, active power, reactive power, apparent power, power factor, active energy, nonactive energy, apparent energy, and auxiliary values such as relative humidity and temperature can be obtained as minimum, maximum, average, and total series.

Meanwhile, MERRA-2 data set is a database available worldwide of meteorological variables hosted by NASA and generated by the Goddard Space Flight Center. All results are produced by a numerical weather forecast model. The spatial resolution is 0.625° in latitude and 0.5° in longitude (approximately 50 km) (Soda-Pro, 2019).

The MERRA-2 reanalysis shares time series of temperature (K) and relative humidity (%) at 2 meters above ground, pressure (hPa) at ground level, wind speed (m/s) and direction (°) at 10 meters above ground, rainfall (kg/m^2), snowfall (kg/m^2), snow depth (m), and short-wave irradiation (Wh/m^2) for global horizontal irradiance from 1 minute up to 1 month (Gelaro et al., 2017).

MERRA-2's nearest grid point for the hospital has a site altitude of 81 meters, but the data acquisition terminal building is 144.774 m. In order to modulate the temperature with the actual altitude of the selected point, the following equation is given as

$$^{\circ}C = K - 272.15 - 0.65 \times \frac{(Altitude_{MERRA-2} - Altitude_{Actual})}{100}$$

where K stands for temperature in Kelvin taken from MERRA-2, $Altitude_{MERRA-2}$ is the site altitude of MERRA-2 nearest grid point, $Altitude_{Actual}$ represents the actual altitude (it is the data acquisition terminal building for the case of the thesis), and $^{\circ}C$ corresponds to the temperature in Celcius degree (Zor et al., 2018b).

For calendar variables, month of year (1-12), week of year (1-53), day of

month (1-31), day type (0 for working days and 1 for weekends and public holidays), hour of day (0-23), sample number of hour (0-5) variables are employed by using lubridate (Grolemund and Wickham, 2011) package in RStudio environment for very short-term data set.

After converting very short-term data set to short-term data set, there is no need to use the sample number of hour variable, hence it is removed during the conversion process.

Furthermore, public holidays in Turkey between 2 October 2017 and 1 October 2018 are listed as follows:

- Republic Day: 29 October 2017 (Half-holiday for Saturday, Sunday),
- New Year's Day: 1 January 2018 (*Monday),
- National Sovereignty and Children's Day: 23 April 2018 (*Monday),
- Labour and Solidarity Day: 1 May 2018 (*Tuesday),
- Commemoration of Atatürk, Youth, and Sports Day: 19 May 2018 (Saturday),
- Religious Holiday: 15–17 June 2018 (*Half-Holiday for Thursday, *Friday, Saturday, and Sunday),
- Religious Holiday: 21–24 August 2018 (*Half-Holiday for Monday, *Tuesday, *Wednesday, *Thursday, and *Friday),
- Victory Day: 30 August 2018 (*Thursday).

For listed public holidays above, it should be noted that if the name of the day is marked with an asterisk (*), then the day type variable corresponding to that day is evaluated as weekends and public holidays. For half-holidays, the starting period of holiday is determined as 13:00 of each day having half-holiday. For example on June 14, 2017; the rows between 00:00 and 13:00 are assigned as 0, but from 13:00

to 24:00 are marked as 1 under the column of day type.

3.2. Data Wrangling and Visualisation with R

R is a programming language and environment for statistical computing and graphics. R provides a variety of statistical, AI, and graphical techniques. R is designed to write programs as functions. For computationally-intensive tasks, C, C++, and Fortran code can be linked and called at run time. Advanced users can write C code to manipulate R objects directly. R can be simply extended via packages. There are about eight packages supplied with the R distribution and many more are available through the CRAN family of Internet sites covering a very wide range of modern statistics (R Core Team, 2018).

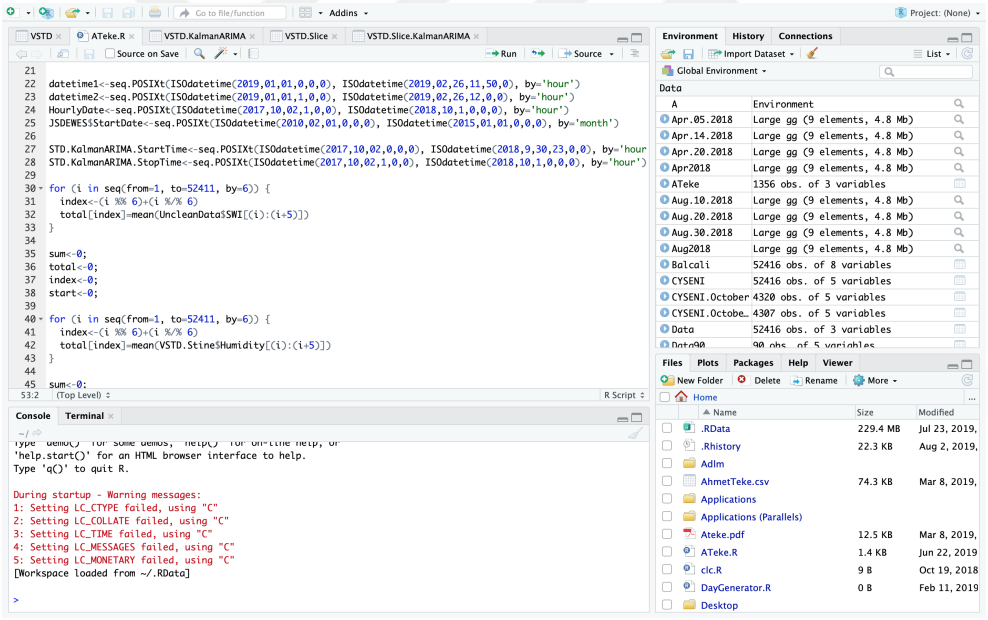

Figure 3.3. A screen shot of RStudio environment

RStudio is an open source integrated development environment for R programming language. In the context of the thesis, packages including tidyverse

(Wickham, 2017), dplyr (Wickham et al., 2018), tidyr (Wickham and Henry, 2018), ggplot2 (Wickham, 2016), scales (Wickham, 2018), gridExtra (Auguie, 2017), imputeTS (Moritz and Bartz-Beielstein, 2017), VIM (Kowarik and Templ, 2016), naniar (Tierney et al., 2018), mice (Van Buuren and Groothuis-Oudshoorn, 2011), and lubridate (Grolemund and Wickham, 2011) are utilised for data wrangling and visualisation purposes.

For the thesis, a version 1.1.456 of RStudio is executed on a MacBook Pro (Retina-Late 2013) whose processor is 2.4 GHz Intel Core i5, memory is 8 GB 1600 MHz DDR3, graphics is Intel Iris 1536 MB, and operating system is macOS Mojave 10.14.6.

In the following chapter, FTHC is described by elucidate graphs and a detailed flowchart.

4. FORECAST TIME HORIZON CONVERTER

For a safe data logging, sampling periods should be determined by regarding the worst case scenarios in the future.

For instance, selecting the demand interval as 1 hour may initially seem advantageous for 1 hour ahead forecasting because of the fact that this kind of selection ordinarily lightens the work load for data wrangling and stores much more data belonging to a longer period indeed, but the selection brings major drawbacks with it.

At first, it is not possible to divide a data set with a sampling period of 1 hour into 1 minute, 5, 10, 15, and 30 minutes. In addition to this, vice-versa is possible. Combining very-short intervals to short, medium, or long intervals is an accomplishable task.

Another drawback reveals in case of a power outage or equipment failure. Assume a power outage takes half an hour long. In case of a sampling period selection of 1 hour, it is impossible to determine the exact correspondence of 30 minutes within 1 hour. If the sampling period is chosen as 5 minutes, then it will be easier to examine the precise interval of loss. One simple solution to forenamed drawbacks is to pick the least possible alternative for sampling period as demand interval which is provided by the logger device according to its measuring sensitivity.

On the other hand, conversion of data sets from one time horizon to another is an arduous challenge. To the best of our knowledge, there is not a customised converter as a toolkit or a software program in the literature especially for energy forecasting applications. Therefore, a novel FTHC is developed for both satisfying the need and filling the gap.

Throughout this chapter, the FTHC is expressed in details by giving examples

from the hospital as part of this thesis.

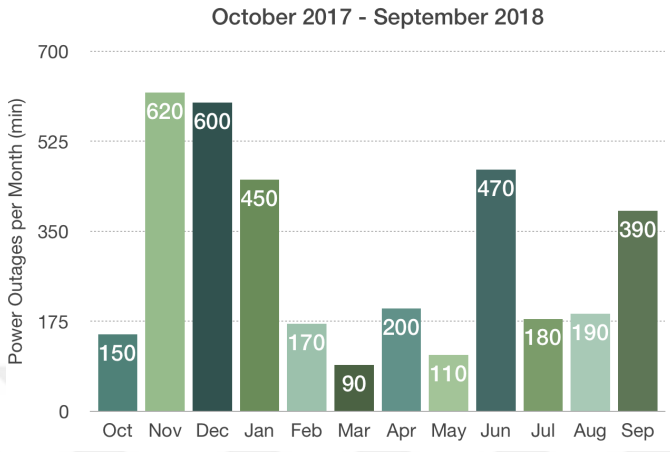

Figure 4.1. Monthly power outages at the hospital

In the context of this thesis, the target is 1 hour ahead STEF of the hospital, but the data acquisition sampling period is selected as 10 minutes because the aforementioned benefits. To do so, a converter methodology is required.

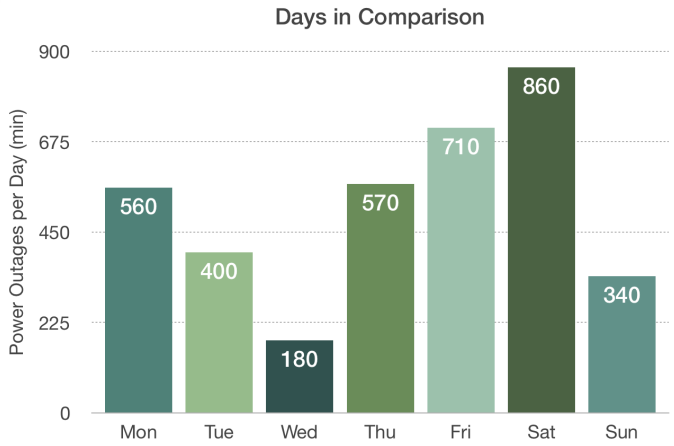

Figure 4.2. Daily power outages at the hospital

Only converting very short-term raw data to short-term data is not adequate

for a methodology owing to the needs of any electrical energy consuming facility. As an example from the hospital case, the hospital had power outages for 3,620 minutes between October 2017 and September 2018 as indicated in Figure 4.1 and Figure 4.2 with respect to months and days respectively.

Due to these events, missing and erroneous (partially missing) data occurred in electrical energy consumption, indoor temperature, and indoor relative humidity data as illustrated in Figure 4.3. Thus, the converter has to detect missing and erroneous data, and perform either an imputation or deletion process.

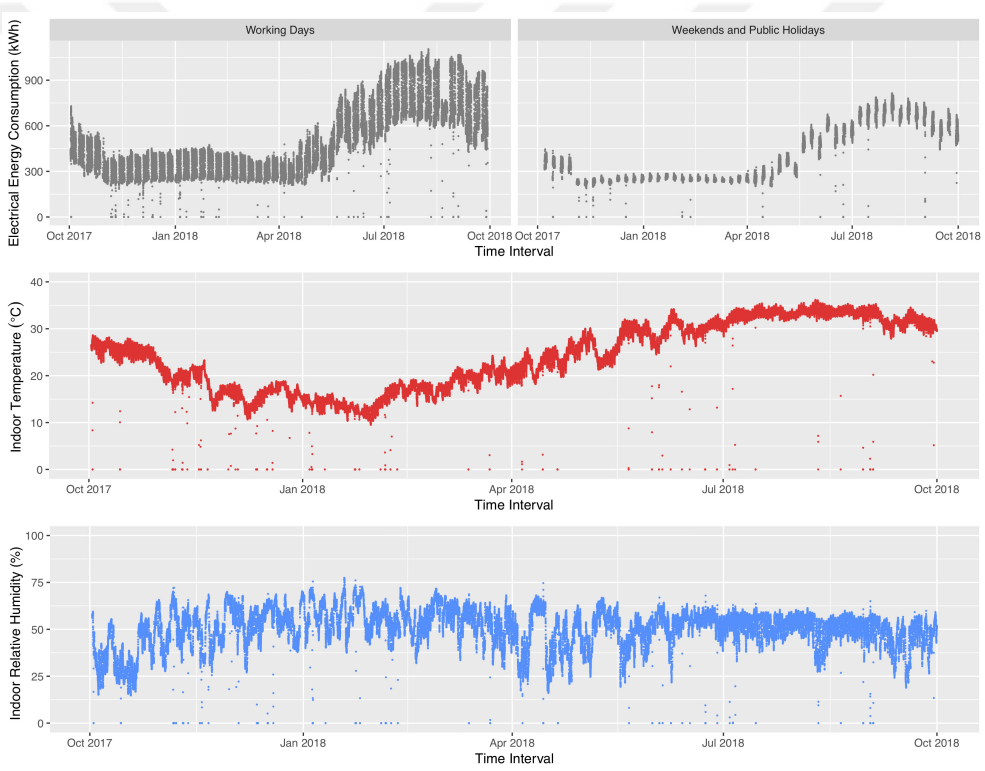

Figure 4.3. Very short-term raw data of the hospital

Novel FTHC methodology was divided into three phases named as phase 1, 2, and 3 respectively as visualised in Figure 4.4.

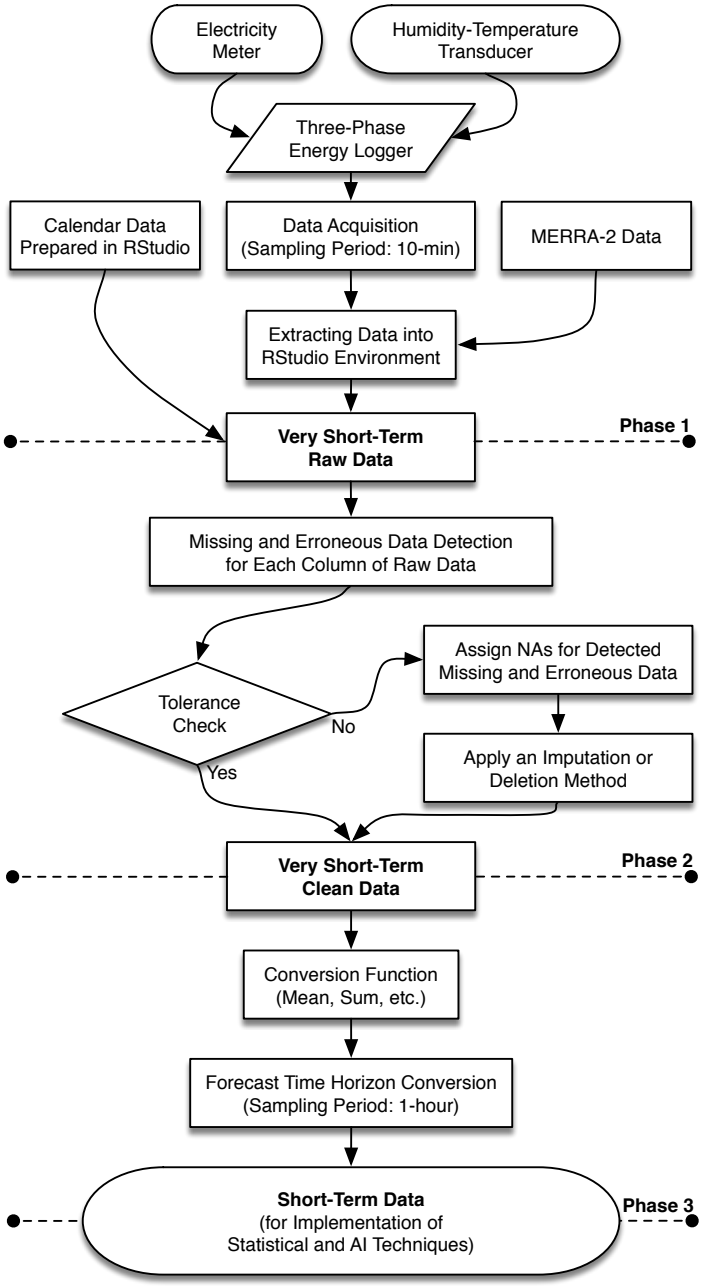

Figure 4.4. Flowchart of forecast time horizon converter

In phase 1, an energy logger measuring from the electricity meter of the hospital and connected to a humidity-temperature transducer has been utilised for gathering real-time electrical energy consumption and meteorological data on-site, and MERRA-2 data in order to form a very-short term raw data set with a sampling period of 10 minutes. Electrical and meteorological variables are imported into RStudio environment, then calendar data are prepared by using lubridate (Grolemund and Wickham, 2011) package in RStudio, and finally very-short-term raw data set is constituted.

In phase 2, very short-term raw data set has been formed. This data set has 3 input variable categories and 1 target variable. The input variable categories are electrical, meteorological, and calendar variables. Target variable is the actual electrical energy consumption. Electrical variables can be stated as the historical electrical energy consumption variables belonging to previous 10 minute, previous 1 day (the same time in the previous day), and previous 1 week (the same time and day in the previous week). Meteorological variables have been obtained from two different sources. The first source is on-site humidity-temperature transducer which provides the indoor relative humidity and temperature of the ambient where the transducer is located with the energy logger. The second source is MERRA-2 data consisting of outdoor temperature, relative humidity, pressure, wind direction, wind speed, rainfall, and short-wave irradiation. Calendar variables include month of year, week of year, day of month, hour of day, sample number of hour, and day type. The very short-term raw data set is represented as a $52,416 \times 19$ matrix in RStudio environment.

Tolerance check mechanism is the core of FTHC methodology. It checks every column of the data set cell by cell and detects outliers according to a user defined deviation parameter in order to create an allowed region for both

unidirectional and bidirectional tolerance check as demonstrated in Figure 4.5.

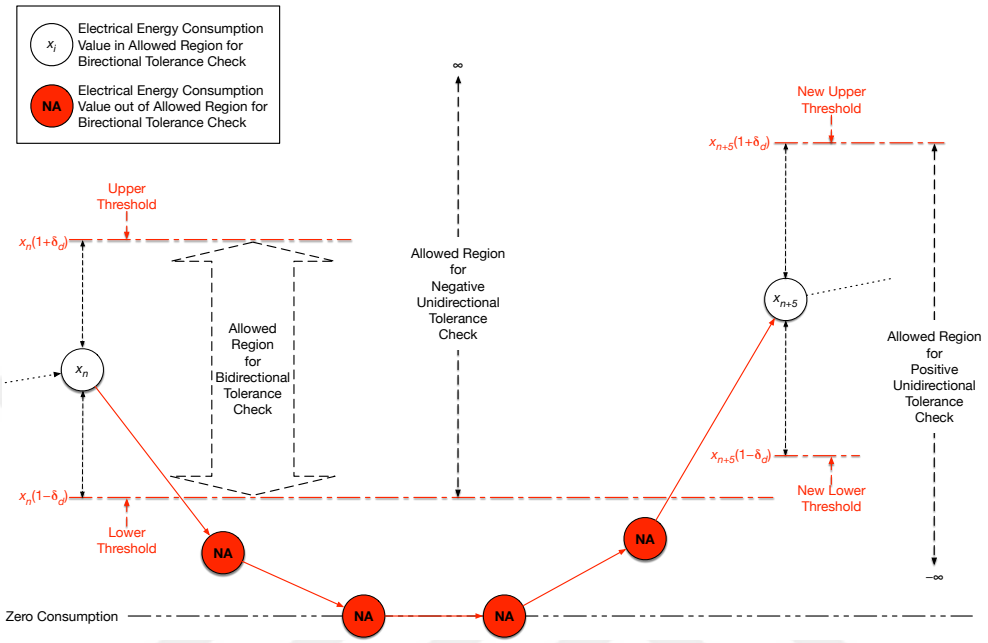

Figure 4.5. Tolerance check mechanism of FTHC

Before expressing the mechanism, the deviation parameter which is a percentage value and symbolised as δ_d must be defined. Then, according to the type of variable, operational scenario of the mechanism must be chosen either bidirectional, negative unidirectional, or positive unidirectional tolerance check. Assume that a cell in a data column is indicated as x_i. The mechanism firstly creates an allowed region for bidirectional tolerance check by using the equation

$$x_i(1-\delta_d) < x_{i+1} < x_i(1+\delta_d)$$

similarly an allowed region for negative unidirectional tolerance check by the equation

$$x_{i+1} > x_i(1-\delta_d)$$

and an allowed region for positive unidirectional tolerance check by the following equation

$$x_{i+1} < x_i(1+\delta_d)$$

where x_{i+1} stands for the next cell after x_i.

Note that if the next cell x_{i+1} does not satisfy the equation corresponding to the selected operational scenario of the mechanism, then it is out of the allowed region and evaluated as an outlier, hence the value within the cell will be replaced by NA for the application of imputation or deletion. Additionally, bidirectional tolerance check has lower threshold $x_i(1-\delta_d)$ and upper threshold $x_i(1+\delta_d)$, negative unidirectional tolerance check possesses the lower threshold with no upper threshold, and positive unidirectional tolerance check owns the upper threshold without a lower threshold. Deviation parameter should be carefully determined for different types of variables. Since, a deviation for electrical energy consumption data should not be the same with a deviation for temperature data. Selecting the optimal deviation is considered diversely.

In this thesis, a negative unidirectional tolerance check mechanism with deviations of 15% for electrical energy consumption and 20% for indoor temperature is performed. An example of implementation is illustrated in Figure 4.6 for the power outages at the hospital on 12, 17, and 18 November 2017. The dots surrounded by red circles in the figure are out of allowed region according to the lower threshold $x_{i+1} > x_i(1-0.15)$ and hence the values of the dots are replaced with NA. For data wrangling and data visualisation, dplyr (Wickham et al., 2018), tidyr (Wickham and Henry, 2018), and ggplot2 (Wickham, 2016) packages under tidyverse (Wickham, 2017) package were used along with scales (Wickham, 2018) and gridExtra (Auguie, 2017) packages.

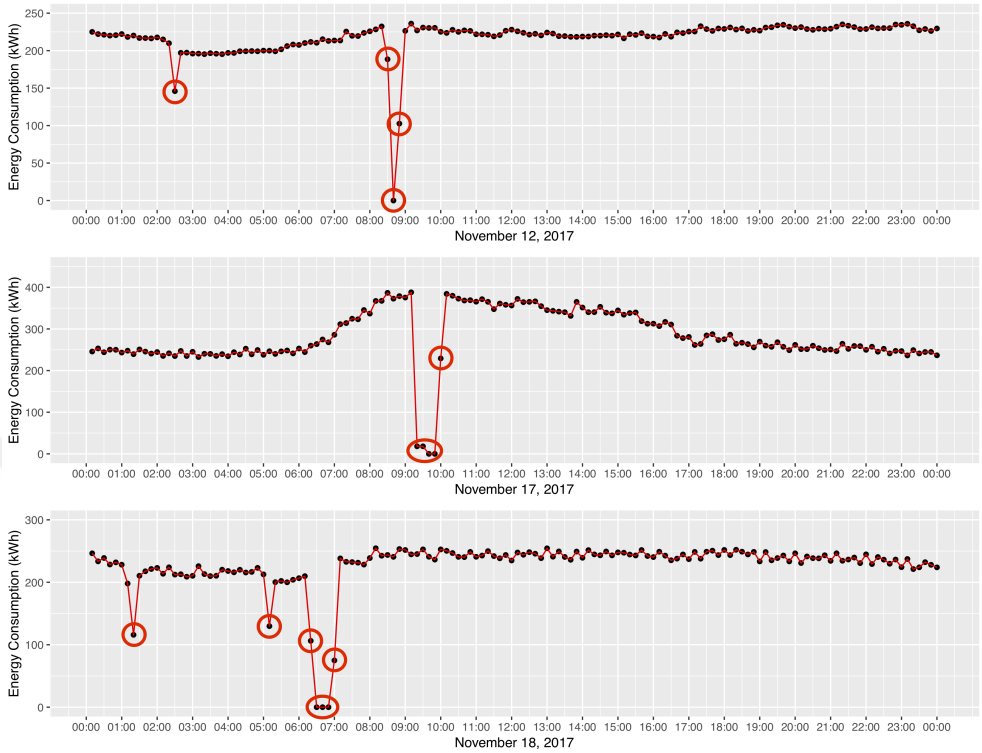

Figure 4.6. An example of tolerance check mechanism

The number of variables possessing NA values for the very short-term data set is 6 and they are actual, previous 10 minute, previous 1 day, and previous 1 week electrical energy consumption, indoor relative humidity, and indoor temperature acquired from the energy logger and the humidity-temperature transducer. The number of rows having NAs varies between 379 and 398, while the proportion of NAs for each variable changes from 0.723% to 0.759%. In total, 2,333 NAs in a $52,416 \times 19$ matrix correspond to a proportion of 0.234%. For further details regarding NAs, Figure 4.7 can be glanced over. For the application of imputation or deletion, and visualisation of NAs; imputeTS (Moritz and Bartz-Beielstein, 2017), VIM (Kowarik and Templ, 2016), naniar (Tierney et al., 2018), and mice (Van

Buuren and Groothuis-Oudshoorn, 2011) packages are employed in RStudio.

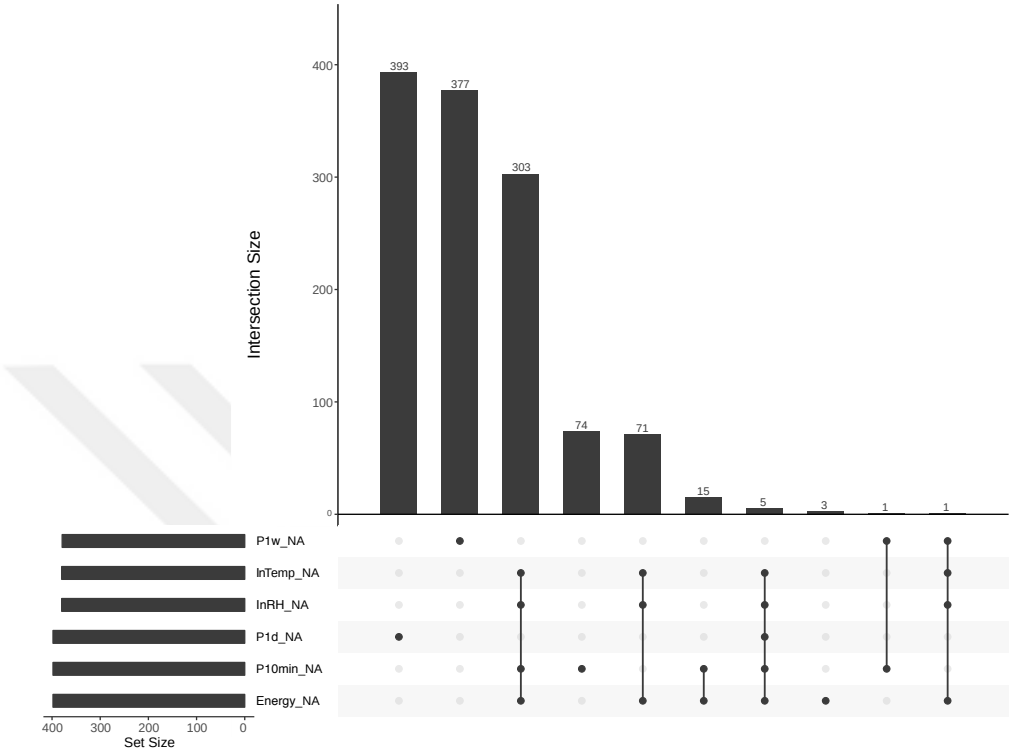

Figure 4.7. Details of variables having missing data

In Figure 4.8, linear interpolation is performed for NA values in electrical energy consumption, indoor temperature, and indoor relative humidity very-short term data before the conversion.

In phase 3, conversion process of treated very short-term raw data to short-term data was expressed by defining a conversion function (sum function for electrical variables, rainfall, and short-wave irradiation; and mean function for other meteorological variables) for each input parameter.

After phase 3, the cleansed short-term data set is represented as $8,736 \times 18$ matrix in RStudio environment and is ready for implementation of statistical and AI

techniques not only for energy related applications, but also for a variety of applications which may have suitable structure with the FTHC methodology.

Points to be paid attention to the end of the conversion, names of all variables remain the same except previous 10 minute owing to renaming as previous 1 hour. In addition to those, sample number of hour variable is not used in the short-term data set.

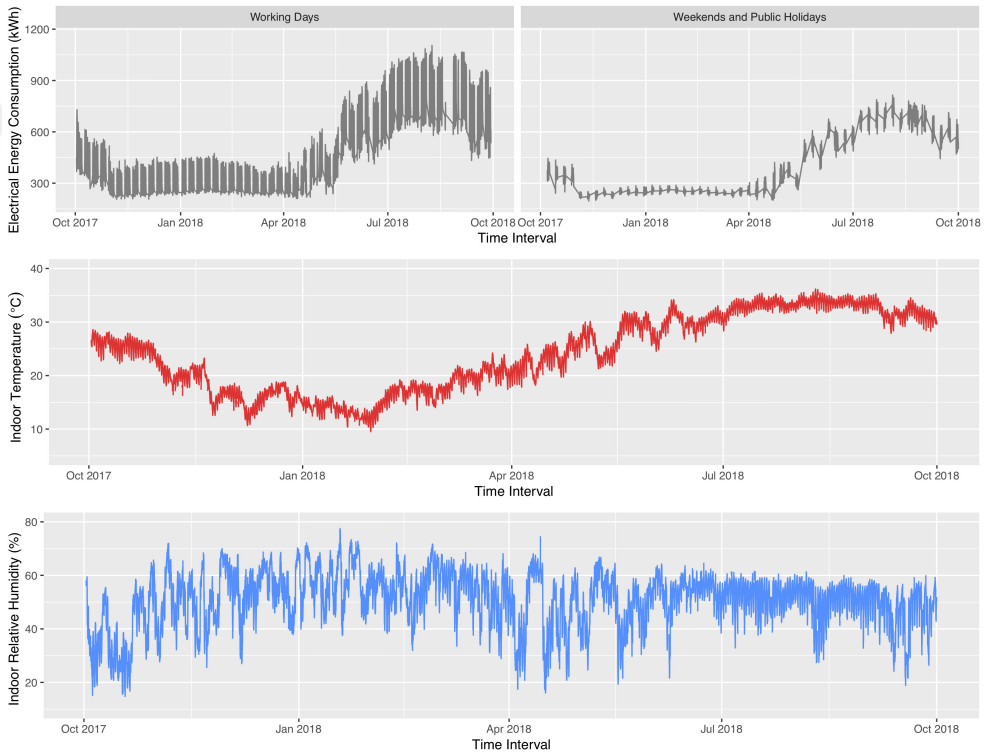

Figure 4.8. Linearly interpolated part of very-short term clean data set

In the following figures, graphs belonging to historical electrical variables and several meteorological variables of short-term clean data set are indicated.

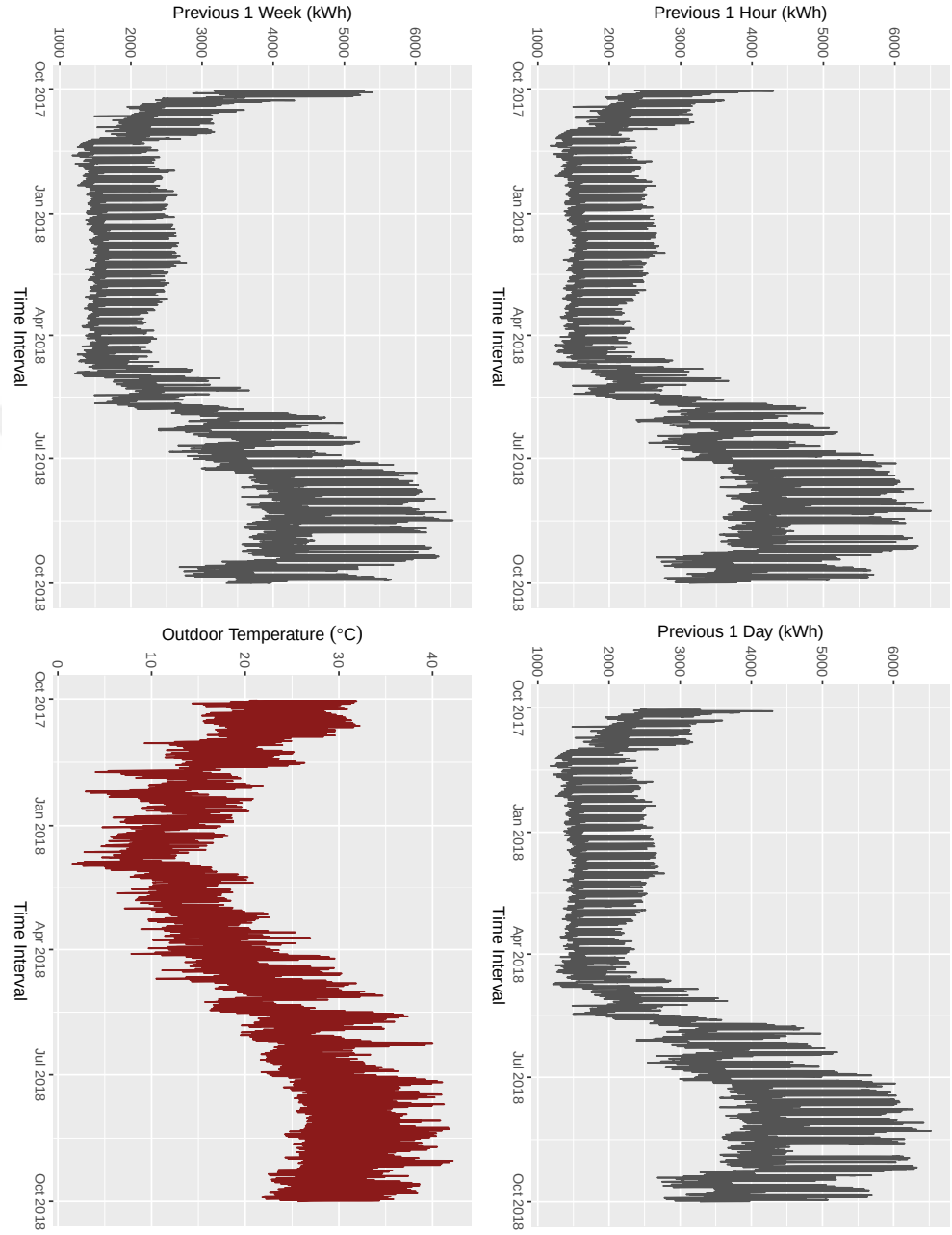

Figure 4.9. Historical electrical variables and outdoor temperature graphs

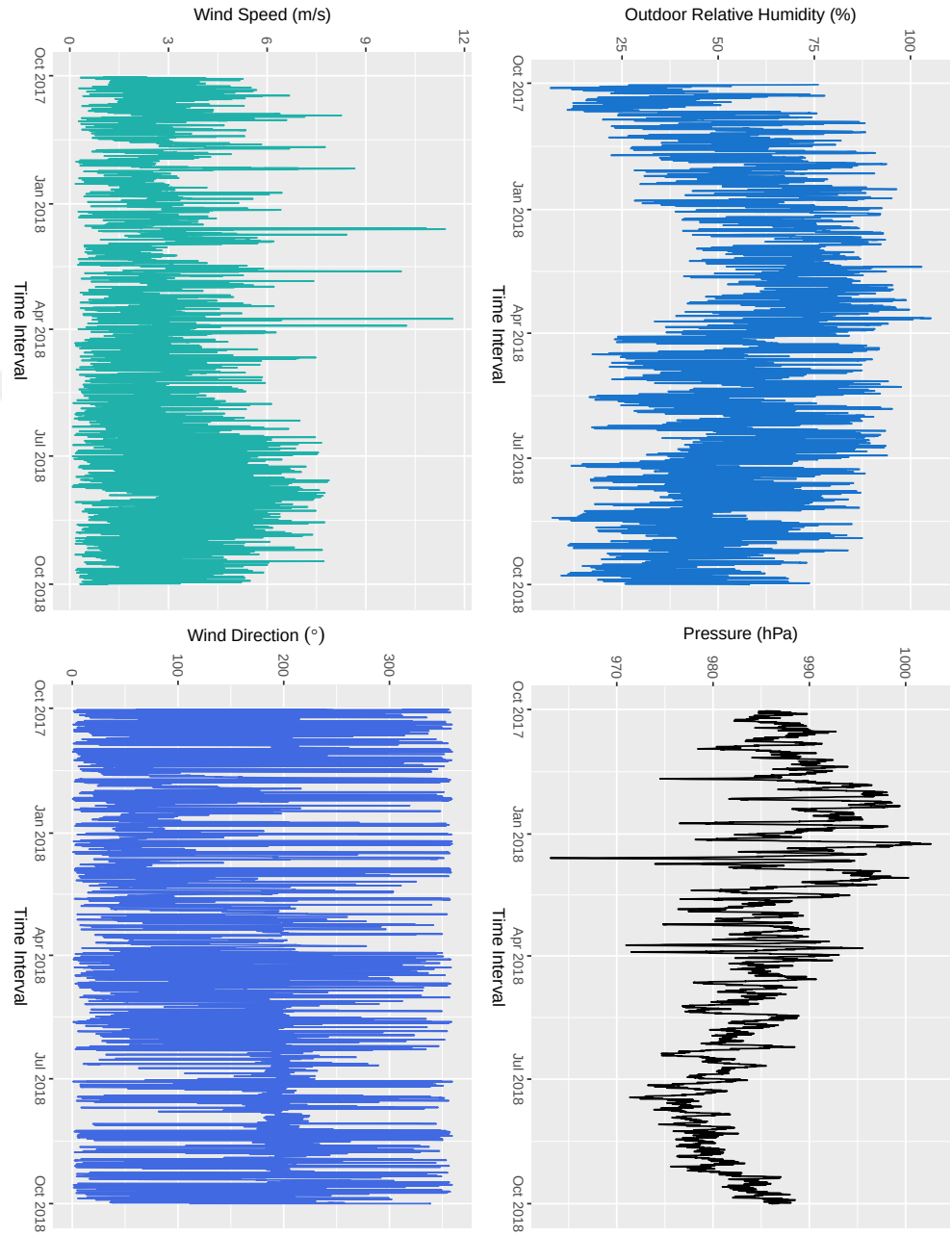

Figure 4.10. Outdoor relative humidity, pressure, wind speed and direction graphs

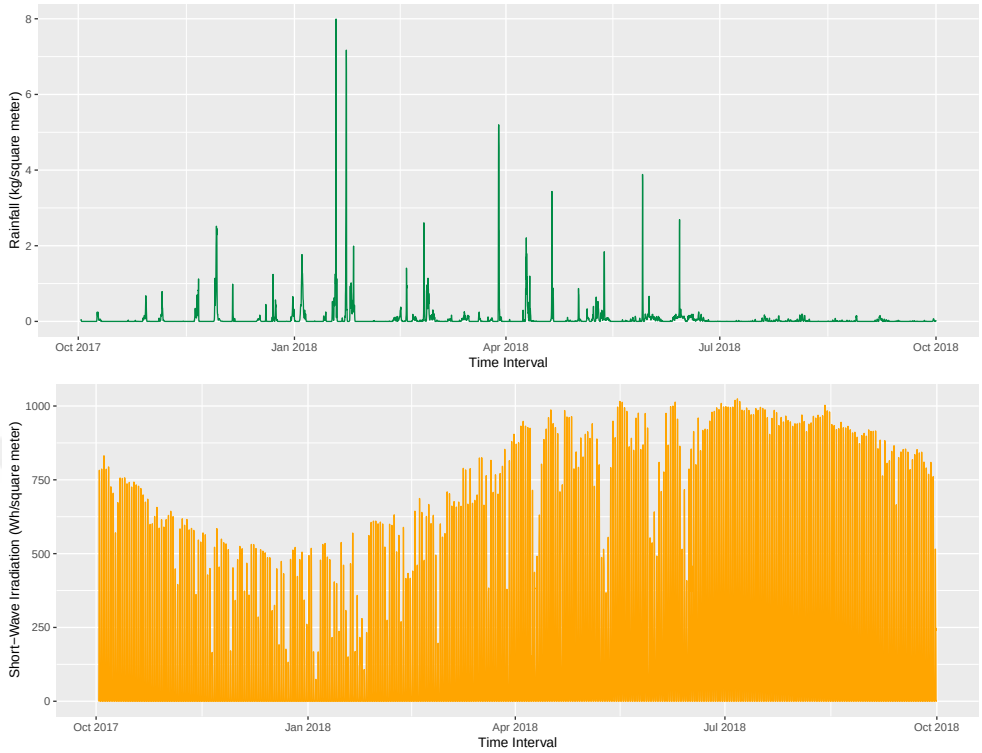

Figure 4.11. Rainfall and short-wave irradiation graphs

In the next chapter, fundamentals of imputation methods for missing data and their individual effects on a real-time energy forecasting data set are expressed.

5. IMPUTATION METHODS FOR MISSING DATA

In this chapter, the individual effects of a variety of imputation methods on a real-time energy forecasting data set is investigated.

5.1. Missing Data

Data sets are inseparable parts of energy forecasting studies and tackling the presence of missing values in the data sets improves the quality of data wrangling while enhancing the accuracy. Missing data are ubiquitous that reveal not only in energy forecasting applications, but also appear in many real-world situations. Frequently, missing data can show up owing to a power outage or a malfunctioned sensor during data acquisition stage of energy forecasting (Zor et al., 2018a).

5.1.1. Missing Data Mechanisms

In the literature, there are three missing data mechanisms named as missing completely at random (MCAR), missing at random (MAR), and not missing at random (NMAR).

First of all, data are MCAR when the probability of a case having a missing value for a variable does not depend on either the known values or the missing data. Assume that $Y = y_{ij}$ is a $(n \times K)$ data set having each row $y_i = (y_{i1}, \ldots, y_{ik})$ set of y_{ij} values of feature Y_j for instance i. Consider that Y_{obs} states the observations of Y and Y_{miss} that represents the missing values. Suppose that M expresses the identity matrix $M = m_{ij}$ for missing data, where $m_{ij} = 1$ if y_{ij} is missing and $m_{ij} = 0$ if y_{ij} is not missing. MCAR can be defined as

$$f(M|Y, \phi) = f(M|\phi) \quad \forall Y, \phi$$

where ϕ corresponds to unknown parameters.

Secondly, data are MAR when the probability of a case having a missing value for a variable may depend on the known values but not on the value of the missing data itself. Hence, it is less delimiting than MCAR and can be expressed as

$$f(M|Y,\phi) = f(M|Y_{obs},\phi) \quad \forall Y_{miss},\phi.$$

Lastly, the pattern of missing data is non-random and depends on the missing variable. In this situation, the missing variable in the NMAR case cannot be predicted only from the available variables in the database. NMAR can be described as

$$f(M|Y,\phi) = f(M|Y_{miss},\phi) \quad \forall Y_{miss},\phi.$$

Missing data in the context of this thesis is occurred due to power outage in the hospital. For this case, the actual variables where data are missing are not the cause of the incomplete data. Instead, the cause of the missing data is due to some other external influence (power outage) which obviously states that missing mechanism for this case is MAR (Little and Rubin, 2002; Schmitt et al., 2015; Poulos and Valle, 2018).

5.1.2. Proportion of Missing Data

For the missing data mechanism MAR, there is not a certain threshold for the proportion of missing data in a data set to apply either complete case analysis (CCA) so called listwise deletion, or single or multiple imputation methods (Dong and Peng, 2013).

In the literature, Schafer proposed that 5% missing data in a data set is a lower threshold below which multiple imputation benefits negligibly (Schafer, 1999). Munguia and Armando emphasised that if the missing data mechanism is either MAR or NMAR, CCA may introduce bias which causes a loss in efficiency that cannot be

negligible (Munguia and Armando, 2014). In contrast to Schafer, Alice stated that 5% missing data is the maximum upper threshold for large data sets (Alice, 2018). More recently, Hughes et al. mentioned that multiple imputation is a valid approach for all MAR mechanisms when compared to CCA, and Madley-Dowd et al. pointed out that multiple imputation with auxiliary information enhanced estimation efficiency at any proportion of missing data in comparison with CCA (Hughes et al., 2019; Madley-Dowd et al., 2019).

Due to the aforementioned arguments, a variety of imputation methods are implemented instead of CCA in this thesis.

5.2. Imputation Methods

In this section, imputation methods applied to missing data in the scope of this thesis are expressed in depth. In order to do so, imputeTS (Moritz and Bartz-Beielstein, 2017) and VIM (Kowarik and Templ, 2016) packages of RStudio are employed for imputation of missing data. Furthermore, tidyverse (Wickham, 2017) and scales (Wickham, 2018) packages of RStudio are used to manipulate imputed data sets for visualisation of figures illustrated throughout the chapter.

5.2.1. Kalman Filters

In this subsection, two approaches for the imputation of missing data namely, KalmanARIMA and KalmanStructTS are described. Briefly, KalmanARIMA utilises Kalman smoothing on the state-space representation of an ARIMA model, while KalmanStructTS adopts the same smoothing method on structural time series models for the imputation (Moritz and Bartz-Beielstein, 2017).

Mathematically, Kalman filters implemented in two phases that are

fundamentally based on the state-space models given in the following equations as

$$x_t = F_t x_{t-1} + \varepsilon_t \tag{5.1}$$

$$y_t = H_t x_t + \omega_t \tag{5.2}$$

where x_t is the state vector of a given system at an instant in time t, y_t is the reciprocating measurement vector at t, F_t is the state-transition parameter of the system, ε_t is the random state noise term, H_t is the measurement parameter, and ω_t is the measurement error term.

In the first phase, the state and the corresponding variance of the system is estimated by using eq. (5.1). In the second phase, the estimated phase is updated

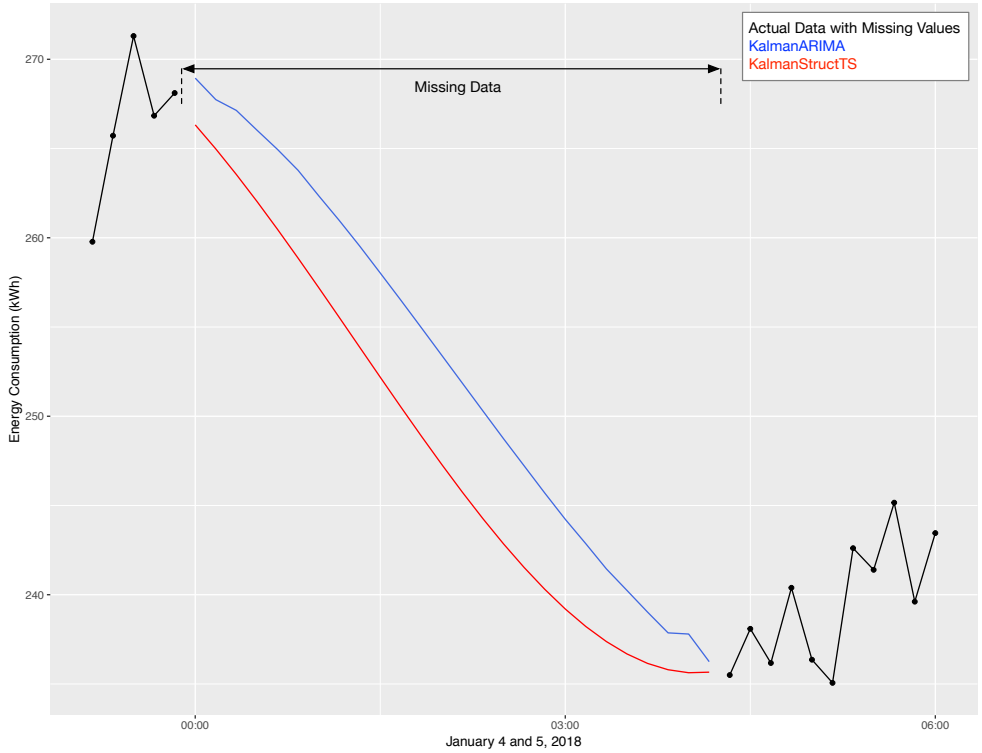

Figure 5.1. Missing data imputation with Kalman filters

by performing both eq. (5.1) and eq. (5.2). In this thesis, KalmanARIMA and KalmanStructTS are employed as transition models. KalmanARIMA utilises forecast (Hyndman et al., 2018) package's auto.arima function that carries out a search in order to find the possible model and the auto.arima function returns the best ARIMA model (Hyndman and Khandakar, 2008). On the other hand, a linear state-space model, which is named as local level model with Gaussian errors where ε_t and ω_t follow Gaussian distribution at the same time by StructTS function of stats (R Core Team, 2018) package, is fitted to apply KalmanStructTS (Demirhan and Renwick, 2018).

An illustration of missing data imputation with Kalman filters is presented in Figure 5.1 to impute the longest missing data sequence in the data set.

5.2.2. Interpolation

Interpolation methods employed in the context of this thesis can be classified as linear interpolation (LI), spline interpolation (SpI), and Stineman interpolation (StI).

The simplest form of interpolation is called as LI which is used to connect two data points with a straight line (Kassam et al., 2014). Using similar triangles,

$$\frac{f_1(x)-f(x_0)}{x-x_0}=\frac{f(x_1)-f(x_0)}{x_1-x_0}(x-x_0)$$

which can be arranged to yield

$$f_1(x)=f(x_0)+\frac{f(x_1)-f(x_0)}{x_1-x_0}(x-x_0)$$

which is the formula of LI. The notation $f_1(x)$ indicates that this is a first-order interpolating polynomial. In general, a smaller interval between the data points results in a better approximation. This is due to the fact that, as the interval

decreases, a continuous function will be better approximated by a straight line (Chapra and Canale, 2010). For LI, approx function of stats (R Core Team, 2018) package is performed within na_interpolation function of imputeTS package.

SpI utilises a nonlinear spline function which can be expressed as

$$S(x) = \begin{cases} P_0(x), & x \in (-\infty, \tau_1); \\ P_j(x), & x \in (\tau_j, \tau_{j+1}), \quad j = 1, \ldots, r-1; \\ P_0(x), & x \in (\tau_r, \infty) \end{cases}$$

where S: $\mathbb{R} \to \mathbb{R}$, $\{P_0(x), P_2(x), \ldots, P_r(x)\}$ is a sequence of cubic polynomials, and $\tau_1 < \tau_2 < \ldots < \tau_r$ is a sequence of real numbers called knots of spline space (Villiers, 2012). For SpI, spline function of stats package is performed within na_interpolation function of imputeTS package.

The last interpolation method employed in the scope of this thesis is StI that utilised stinterp function of stinepack (Johannesson et al., 2018) package within na_interpolation function of imputeTS package. The function returns the values of an interpolating function that runs through a set of points in the xy-plane according to the algorithm of Stineman (Stineman, 1980). According to Stineman, polynomial interpolation such as SpI frequently gives undesirable results near an abrupt change of slope. To avoid those kinds of results, Stineman offered the following interpolation procedure,

1. If values of the ordinates of the specified points change monotonically, and the slopes of the line segments joining the points change monotonically, then the interpolating curve and its slope will change monotonically.
2. If the slopes of the line segments joining the specified points change

monotonically, then the slopes of the interpolating curve will change monotonically.

3. Suppose that the conditions in (1) or (2) are satisfied by a set of points, but a small change in the ordinate or slope at one of the points will result conditions (1) or (2) being not longer satisfied. Then making this small change in the ordinate or slope at a point will cause no more than a small change in the interpolating curve (Stineman, 1980).

More mathematically, consider that x_j and y_j are rectangular coordinates on a curve's jth point, while $\acute{y}_j$ is the slope of the curve at jth point for $j = 1,\ldots,n$ and $x_j < x_{j+1}$ for $j = 1,\ldots,n-1$. Afterwards, the interpolated value y can be computed by following the steps given as

1. For a given x fulfilling $x_j \leq x \leq x_{j+1}$, compute the slope of the line segment joining the points j and $j+1$ by $s_j = (y_{j+1} - y_j)/(x_{j+1} - x_j)$.
2. Compute the ordinate corresponding to x by $y_0 = y_j + s_j(x - x_j)$.
3. Compute the vertical distance from the point $(x - y_0)$ to a line through $(x_j - y_j)$ with the slope $\acute{y}_j$ by $\Delta y_j = y_j + \acute{y}_j(x - x_j) - y_0$ for the points j and $j+1$.
4. Compute the interpolated value $y = y_0(\Delta y_j \Delta y_{j+1})/(\Delta y_j + \Delta y_{j+1})$ if $\Delta y_j \Delta y_{j+1} > 0$, and also compute for the other scenario as $y = y_0[\Delta y_j \Delta y_{j+1}(2x - x_j - x_{j+1})]/[(\Delta y_j - \Delta y_{j+1})(x_{j+1} - x_j)]$ else if $\Delta y_j \Delta y_{j+1} < 0$.

In order to apply the above algorithm, $\acute{y}_j$ must be foreknown. If $\acute{y}_j$ for interior points and $\acute{y}_m$ for the end point m are not known in the first place, these values should be computed as follows. Consider that I, J, and K are any three successive points fulfilling either $(\acute{IJ}) < \acute{y}_j < (\acute{JK})$ or $(\acute{IJ}) > \acute{y}_j > (\acute{JK})$, where $\acute{(\cdot)}$ represents the slope

of inner curve segment. The slope $\acute{y}_j$ is computed for interior points as shown in the below equation

$$\acute{y}_j = \frac{(y_j - y_i)[(x_k - x_j)^2 + (y_k - y_j)^2] + (y_k - y_j)[(x_j - x_i)^2 + (y_j - y_i)^2]}{(x_j - x_i)[(x_k - x_j)^2 + (y_k - y_j)^2] + (x_k - x_j)[(x_j - x_i)^2 + (y_j - y_i)^2]}$$

and for the end point m, the slope $\acute{y}_m$ is computed as shown in the below

$$\acute{y}_m = 2s - \acute{y}_j$$

where s represents the slope of line segments joining points J and end points (Demirhan and Renwick, 2018).

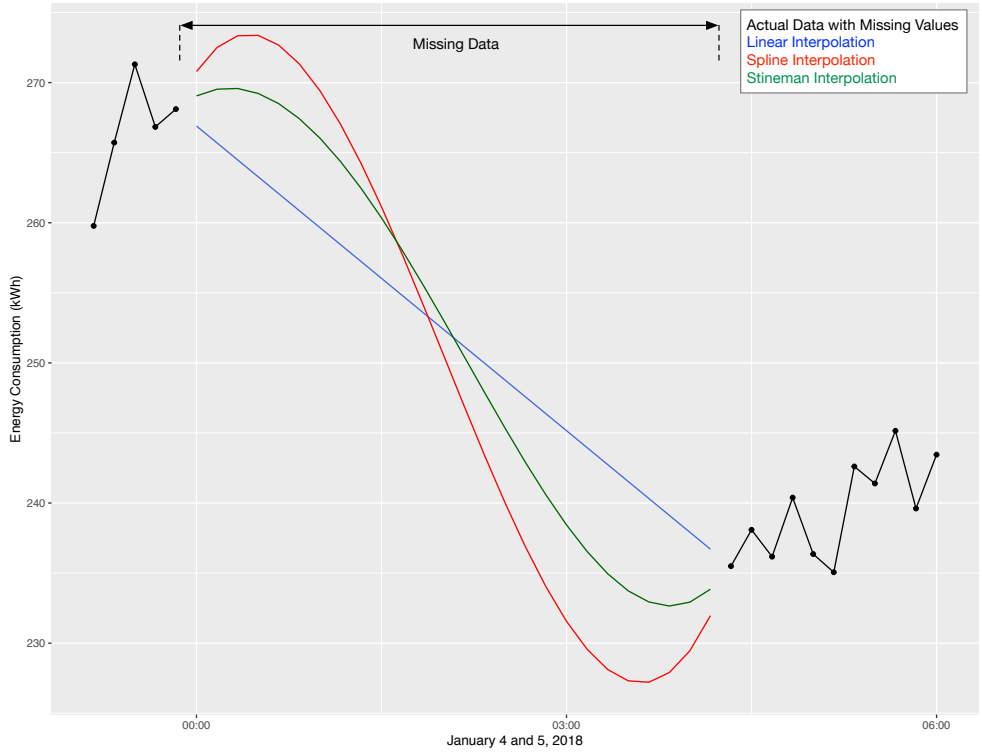

Figure 5.2. Missing data imputation with interpolation methods

A demonstration of missing data imputation with interpolation methods is illustrated in Figure 5.2 to impute the longest missing data sequence in the data set.

5.2.3. Weighted Moving Average

Weighted moving average (WMA) approaches utilised in this thesis can be classified as simple moving average (SMA), linearly weighted moving average (LWMA), and exponentially weighted moving average (EWMA). In general, all WMA approaches employ a semi-adaptive window size to satisfy the fact that all of missing values in the data set are imputed. For all WMA approaches, na_ma function of imputeTS package is used. In order to change the weighting scenario, the weighting option in the function can be declared as "simple" for SMA, "linear" for LWMA, and "exponential" for EWMA respectively.

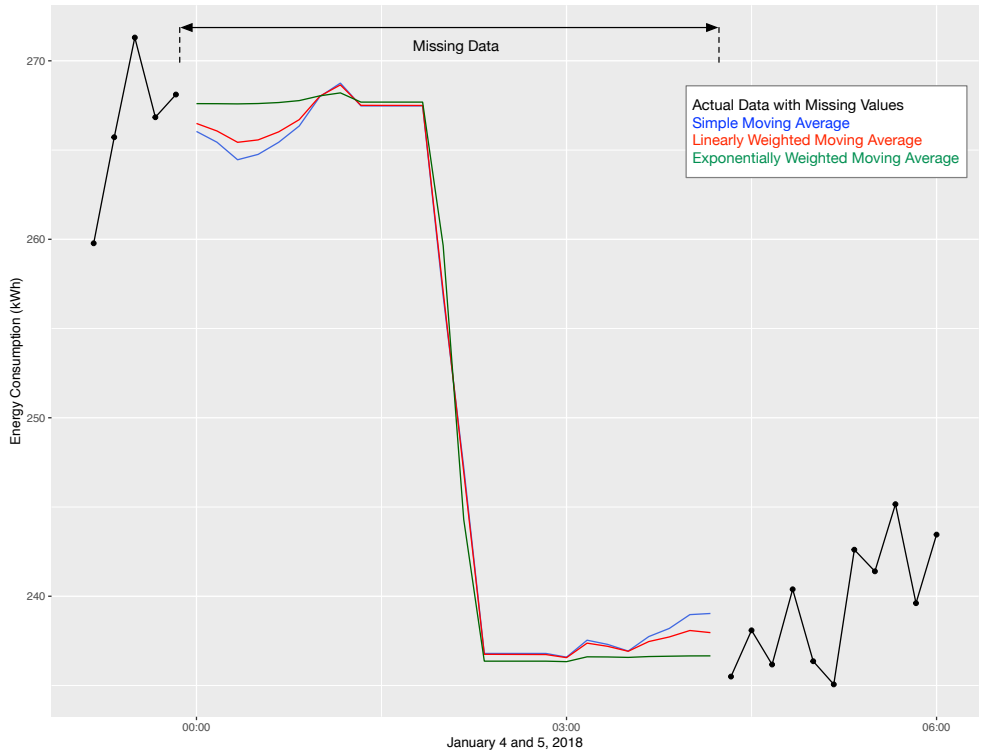

Figure 5.3. Missing data imputation with weighted moving average methods

Missing data imputation with WMA approaches is indicated in Figure 5.3 to

impute the longest missing data sequence in the data set.

Theoretically, the average in WMA obtained from the same number of observations on either side of a central value. In SMA, all observations in the window are evenly weighted to compute the average. On the other hand, in LWMA, weights reduce by arithmetical sequence such as 1/2, 1/3, 1/4, and so on. Similarly, EWMA implements weighting factors that reduce exponentially such as 1/2, 1/4, 1/8, and so forth (Moritz and Bartz-Beielstein, 2017).

Mathematically, a general expression satisfying all WMA approaches can be derived for one-step ahead WMA forecast such that

$$\hat{Y}_{t+1} = \sum_{i=-k}^{k} \omega_i Y_{t+1+i}$$

where $\omega_{-k}, \omega_{-k+1}, \ldots, \omega_k$ symbolise the weights by paying attention to the fact that $\{Y_t, t = 1, \ldots, T\}$ is the time series of interest (Demirhan and Renwick, 2018).

5.2.4. kNN Imputation

In kNN imputation, an aggregation of k values of the nearest neighbours is employed to impute the individual missing value. Gower distance is the baseline for distance calculation of the nearest neighbours (Gower, 1971). The distance among two observations is the weighted average of the contributions of each variable, while the weight corresponds to the importance of the variable. Thus, the distance among the ith and jth observation can be stated as

$$d_{i,j} = \frac{\sum_{k=1}^{p} \omega_k \delta_{i,j,k}}{\sum_{k=1}^{p} \omega_k}$$

where ω_k is the weight and $\delta_{i,j,k}$ is the contribution of the kth variable. $\delta_{i,j,k}$ can be calculated for continuous variables as shown below

$$\delta_{i,j,k} = \frac{| x_{i,k} - x_{j,k} |}{r_k}$$

where $x_{i,k}$ is the value of kth variable of the ith observation and r_k is the range of the kth variable (Kowarik and Templ, 2016).

In order to implement kNN imputation, kNN function of VIM (Kowarik and Templ, 2016) package is performed. For $k = 2$ and $k = 144$ values, missing data imputation with kNN is demonstrated in Figure 5.4 to impute the longest missing data sequence in the data set.

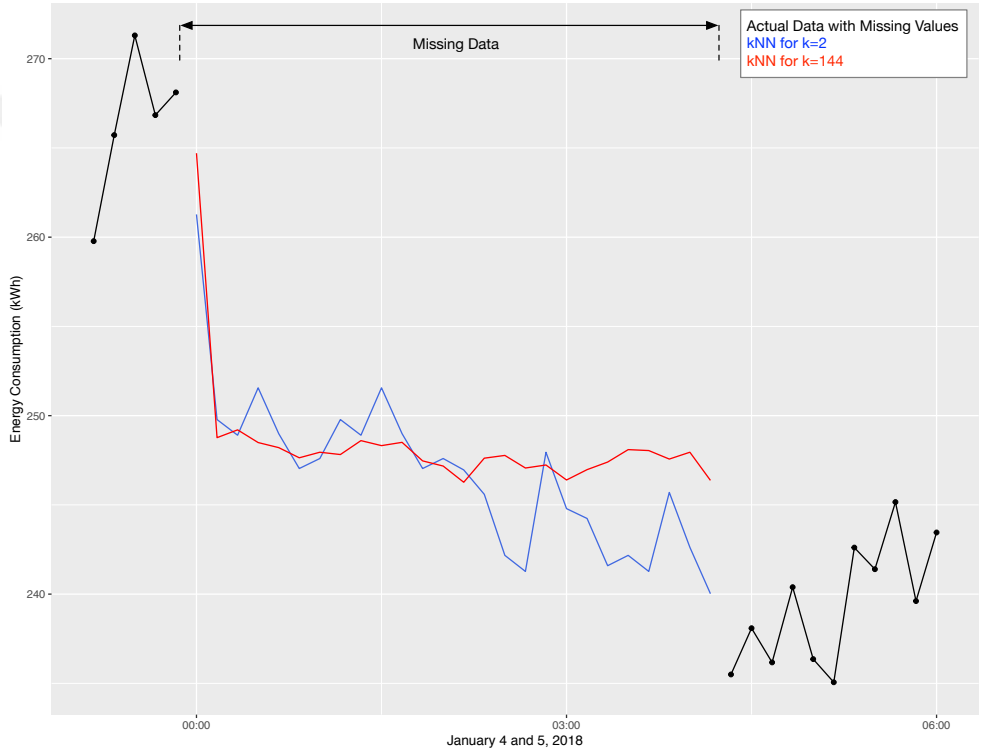

Figure 5.4. Missing data imputation with kNN

5.2.5. Persistence

Persistence approaches used in the context of this thesis can be divided into two categories, namely last observation carried forward (LOCF) and next observation carried backward (NOCB).

For both LOCF and NOCB imputation methods, na_locf function of imputeTS package is employed in RStudio. LOCF imputes the missing observations in the forward direction by replacing NAs with the last observed value, while NOCB fills the missing observations in the backward direction by replacing NAs with the next observed value as indicated in Figure 5.5 in order to impute the longest missing data sequence in the data set.

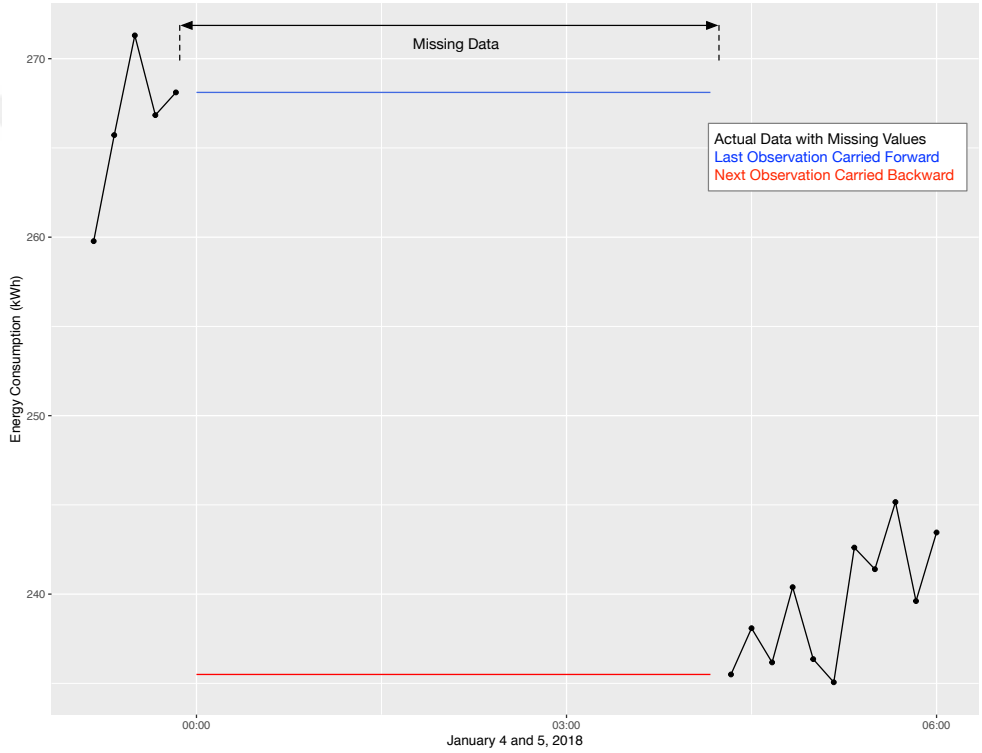

Figure 5.5. Missing data imputation with persistence methods

In the following chapter, fundamentals of S&AI techniques utilised in the thesis are investigated in details.

6. STATISTICAL AND AI TECHNIQUES

In this chapter, MLR as a statistical technique and AI based techniques containing SVM, GEP, GBDT, and ANN consisting of MLPNN, RBFNN, GRNN, and GMDHNN are thoroughly investigated.

6.1. Statistical Technique

Statistical techniques are not in the scope of this thesis, but one of the most common of them, MLR is expressed and employed for benchmarking purposes.

6.1.1. Multiple Linear Regression

In the field of electrical energy consumption modelling, the goal of MLR as a statistical technique is to formalise the relationship among various explanatory variables such as weather and calendar information, and a dependent variable which is the amount of electrical energy demand as a linear function in order to predict the consumed energy amount as closely as possible (Hong et al., 2010). Model using MLR is expressed as

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n + e$$

where y is the consumed energy amount, x_i is the value of independent variables, β_i is regression parameters with respect to x_i, and e represents error (Amral et al., 2007). In MLR, the error term corresponds to a set of random variables independent and identically distributed with a Gaussian distribution having zero mean (Hong et al., 2010).

The main reasons of the selection of MLR in this thesis as the only statistical technique are its simplicity for understanding, its ease to use, and its faster operation in comparison with other techniques. For MLR, stats package (R Core Team, 2018)

is used in RStudio environment. Benefits and drawbacks of MLR are described in the below table.

Table 6.1. Benefits and drawbacks of MLR (Timur et al., In Press)

Benefits	Drawbacks
1. Easy to calculate 2. Fast 3. Has good interpretability 4. Requires less memory	1. Its nature of assuming a linear relationship between dependent and independent variables 2. Oversimplifies real-world problems 3. Can cause severe multicollinearity 4. Prone to outliers

6.2. Artificial Intelligence Techniques

In this section, AI techniques including SVM, GEP, GBDT, and ANN containing MLPNN, RBFNN, GRNN, and GMDHNN are expressed in details.

6.2.1. Artificial Neural Networks

The interpretation of the neuron doctrine was initially credited by McCulloch and Pitts in 1943, Frank Rosenblatt actualised mathematical analysis, digital computer simulation, and experiments with special purpose parallel analog systems that neural networks with variable weight connections can be trained for the classification of spatial patterns into prespecified categories (Nagy, 1991). Nearly five decades after Rosenblatts approach, a variety of ANN methodologies are very trendy as AI techniques, especially for energy forecasting applications (Zor et al., 2017b).

In this thesis, ANN methodologies are investigated by dividing into 4 categories named as MLPNN, RBFNN, GRNN, and GMDHNN.

6.2.1.1. Multilayer Perceptron Neural Networks

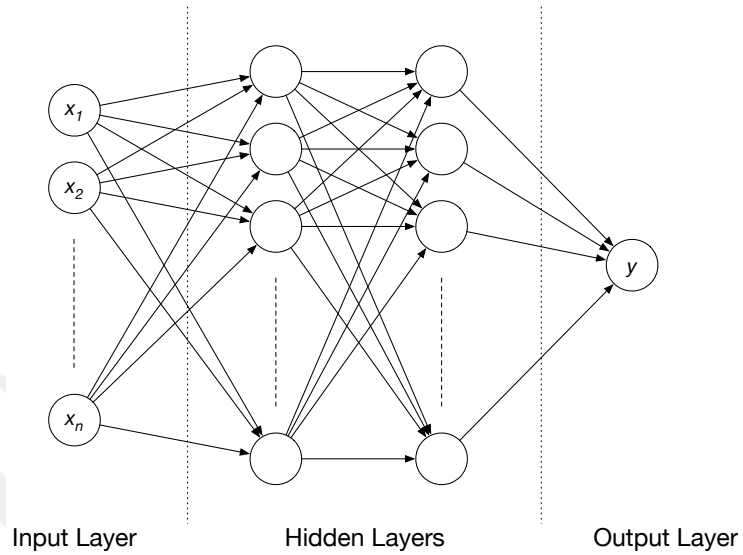

Figure 6.1. A basic feed-forward MLPNN topology

The simplest and smallest unit of ANN is an artificial neuron which has the capability of managing complex behaviours between the operative neurons and weight parameters (Yıldırım et al., 2018). In general, the fundamental topology of ANN is represented by a feed-forward MLPNN with one hidden layer which is constituted as demonstrated in Figure 6.1 by three types of neuron layers, namely input layer, hidden layer, and output layer

$$\hat{y} = b_0 + \sum_{h=1}^{H} b_h g(\delta_{0i} + \sum_{i=1}^{I} \delta_{hi} p_i)$$

where I represents the number of inputs p_i and H is the number of hidden nodes in the network. The weights $\omega = (b, \delta)$, where $b = [b_1, \ldots, b_H]$ and $\delta = [\delta_{11}, \ldots, \delta_{H1}]$, are for the hidden and output layer sequentially. b_0 and δ_{0i} are the biases of each node, and the transfer function $g(\cdot)$ may be nonlinear and is usually either the sigmoid logistic or the hyperbolic tangent function (Barrow and Kourentzes, 2016; Timur et

al., In Press).

The fundamental reasons behind the selection of MLPNN in this thesis as AI techniques are its fully approximation for any complex nonlinear relationship, its high-speed search in determining the ideal number of hidden layers and the optimal number of neurons within the layers, and capable of learning and adapting to unknown or uncertain systems. For MLPNN, RSNSS package (Bergmeier and Benitez, 2012) is employed in RStudio environment. Benefits and drawbacks of MLPNN are described in the following table.

Table 6.2. Benefits and drawbacks of MLPNN (Timur et al., In Press)

Benefits	Drawbacks
1. Needs less formal statistical training 2. Implicitly detects complex relationships between dependent and independent variables 3. Detects all possible interactions between predictor variables 4. Has access to multiple training algorithms	1. Its nature of being a black-box 2. Has greater computational cost 3. Tends to overfitting 4. The empirical nature of the model development

6.2.1.2. Radial Basis Function Neural Networks

Recently, RBFNN have been considered as a prospering alternative to MLPNN owing to their broad spectrum of applications and quicker learning ability. In comparison with traditional sigmoidal MLPNN, RBFNN have minimal interaction between RBF units, because each RBF unit is typically influenced by smaller portions of input patterns (Yu et al., 2011). In RBFNN, merely one hidden layer is in existence and neurons of the hidden layer contain radial basis activation functions as shown in Figure 6.2. Hence, output of an RBF network is tantamount to

the weighted summation of the responses of the hidden neurons can be explained as

$$y_j = \sum_{i=1}^{n} \omega_{ij} \varphi_i(||x - c_i||) + b_{0j}, \quad (j = 1, 2, \ldots, n)$$

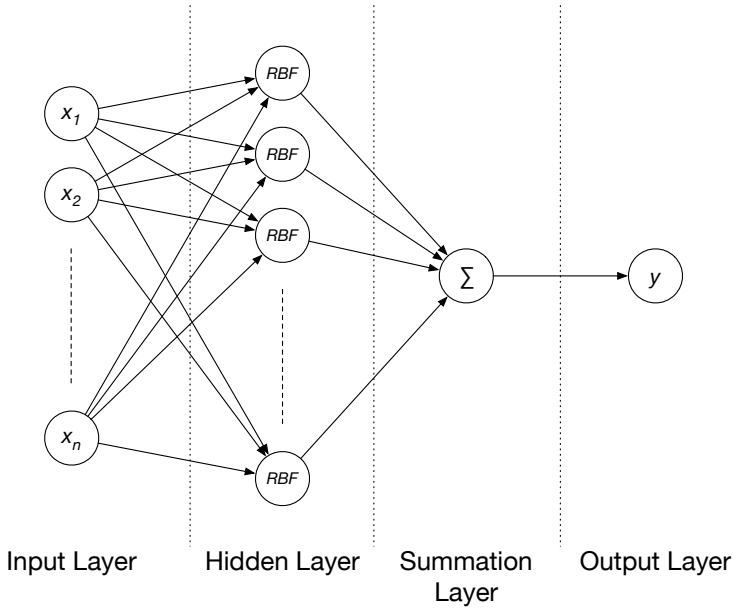

Figure 6.2. RBFNN topology

where the number of nodes in the hidden layer is n, input vector is x, the centre of the ith hidden node is c_i, the weight of ith node of the hidden layer is ω_{ij}; the radial basis function with c_i being its centre is φ_i, and the bias of the jth node of output layer is b_{0j}. The mapping from the input layer to the hidden layer is nonlinear, while it is linear from the hidden layer to the output layer (Timur et al., In Press). Herein, the radial basis function is a Gaussian function and input-to-centre distance is designated by utilising simple Euclidean distance as indicated below

$$\varphi_i(x) = \exp\left(-\frac{||x - c_i||^2}{\sigma^2}\right)$$

where $\sigma < 0$ is a predefined spread value of the function. It should be noted that three key parameters are decided in RBFNN which are spread, centres, and inter-network weights (Hossain et al., 2017).

The basic reasons of the selection of RBFNN in this thesis as AI techniques are its fast on-line learning ability, its strong tolerance to noisy input data, and its easy implementation. For RBFNN, RSNSS package (Bergmeier and Benitez, 2012) is executed in RStudio environment. Benefits and drawbacks of RBFNN are described in the following table.

Table 6.3. Benefits and drawbacks of RBFNN (Mai et al., 2014)

Benefits	Drawbacks
1. Fast on-line learning ability	1. Its black-box structure
2. Strong tolerance to noisy input data	2. Cannot be operated without training
3. Good generalisation	data owing to supervised learning
4. Easy implementation	3. No guarantee of success

6.2.1.3. Generalised Regression Neural Networks

GRNN are highly parallel RBFNN based on a nonlinear regression analysis named as kernel regression (Liang et al., 2019).

GRNN are constituted of four layers, namely input layer, pattern layer, summation layer, and output layer as visualised in Figure 6.3. Each layer has a certain function to perform nonlinear regression. The input layer has p number of neurons resulting in p dimension input features of the samples.

Every one of the neurons in the pattern layer is the centre of a cluster, that computes the exponential form of Euclidean distance between prediction sample x_0 and training sample x_i, and the output of each neuron i is $e^{D(x_i)}$. In the pattern layer, the number of neurons is the same with the number of training samples.

The neurons in the summation layer take the outputs of the neurons in the pattern layer. Summation layer consists of two units named as numerator and denominator. In order to reach the output layer, the units in the summation layer are divided to acquire the final output.

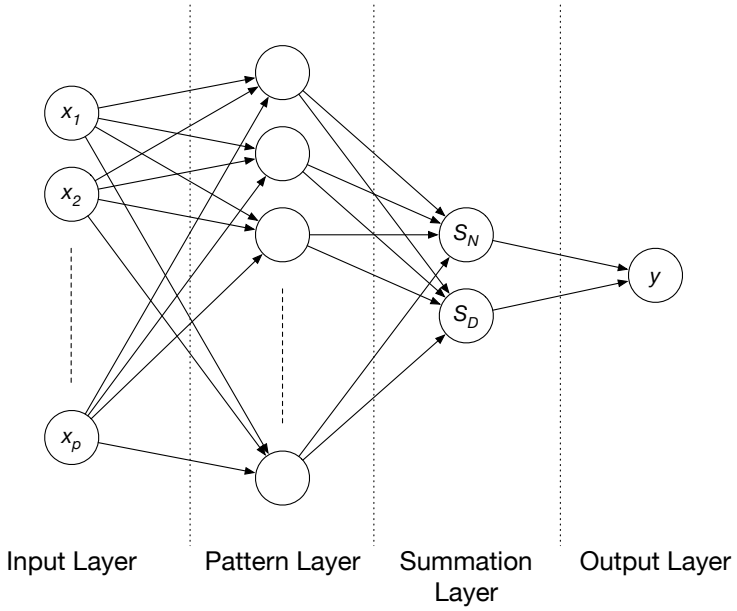

Figure 6.3. Visualisation of GRNN

The joint probability density function having random variables x and y is represented as $f(x,y)$. x is a vector for input random variable, while y is a scalar that indicates output random variable. If $f(x,y)$ is known, then condition mean of y on a given x_0 can be computed by the following equation

$$\hat{y}(x_0) = E(y|x_0) = \frac{\int_{-\infty}^{\infty} y f(x_0,y)dy}{\int_{-\infty}^{\infty} f(x_0,y)dy}$$

In spite of the fact that $f(x,y)$ is rarely known in most of the systems, hence estimation of $f(x,y)$ is a necessity for output parameter calculation. GRNN utilise measured x and y values together for the estimation of $f(x,y)$. Assume that a

measured sample data point $\{x_i, y_i | i = 1, 2, \ldots, N\}$ is given. Then the estimated joint probability density function can be stated as

$$\hat{f}(x,y) = \frac{1}{n(2\pi)^{(p+1)/2}\sigma^{(p+1)}} \sum_{i=1}^{n} e^{D(x_i)} \cdot e^{D(y_i)}$$

$$D(x_i) = -\frac{(x-x_i)^2}{2\sigma^2}$$

$$D(y_i) = -\frac{(y-y_i)^2}{2\sigma^2}$$

where n represents the number of measured samples, p corresponds to the dimension of x, and σ stands for the spread factor of the Gauss function for smoothing.

For a given x_0, the condition mean of y can be updated as the division of numerator (S_N) by denominator (S_D) for the summation layer

$$\hat{y}(x_0) = \frac{S_N}{S_D} = \frac{\sum_{i=1}^{n} y_i e^{D(x_i)}}{\sum_{i=1}^{n} e^{D(x_i)}}$$

The estimated condition mean can be considered as the weighted average belonging to all observed values of y_i in which each observed value is exponentially weighted with respect to its Euclidean distance from x_0 (Xie et al., 2019).

Table 6.4. Benefits and drawbacks of GRNN (Onwubolu, 2015; Stepashko et al., 2017)

Benefits	Drawbacks
1. Presents adaptive network topologies which can be customised to the given problem 2. Finds locally good weights owing to the reliability of the fitting technique 3. Can be trained rapidly by sparse connectivity	1. Tends to produce quite complex polynomials for simple systems 2. Do not guarantee building up the true structure 3. Biased estimates of coefficients due to the least squares method

In this thesis, GRNN are chosen as an AI technique due to its ability to perform excellent approximation, to have fast learning speed, and to converge to the

optimal regression surface (Li et al., 2013). For GRNN, grnn package (Chasset, 2013) is employed in RStudio environment. Benefits and drawbacks of SVM are described in the previous table.

6.2.1.4. Group Method of Data Handling Polynomial Neural Networks

GMDHNN principally operate as self-organising networks where neuron connections, number of selected neurons, layers, and neurons in hidden layers are not constant and are self-acting along with training in order to reach an optimal model for maximum accuracy without overfitting (De Giorgi et al., 2016). To do so, GMDHNN uses least squares regression to find the best mathematical relation among input and output variables by a reference function which can be expressed as

$$y = a_0 + \sum_{i=1}^{n} a_i x_i + \sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij} x_i x_j + \sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{n} a_{ijk} x_i x_j x_k + \dots$$

where y corresponds to the output, $X = (x_1, x_2, \dots, x_n)$ represents the input vector, and a symbolises either the coefficient or weight vector (Xiao et al., 2018).

Ordinarily, the previous equation is utilised in the quadratic form of two variables such that

$$y = a_0 + a_1 x_i + a_2 x_j + a_3 x_i x_j + a_4 x_i^2 + a_5 x_j^2$$

In GMDHNN, input layer contain neurons for each input variables indicated by v. Each neuron in the first layer acquires its inputs from two of the neurons in the input layer. The neurons in the second and the third layers obtain their inputs from two of the neurons in the previous layer and this process continues to output layer. The output layer takes two of its inputs from the previous layer and generates the final result that shows the most suitable mathematical expression in satisfying the relationship between input and output variables.

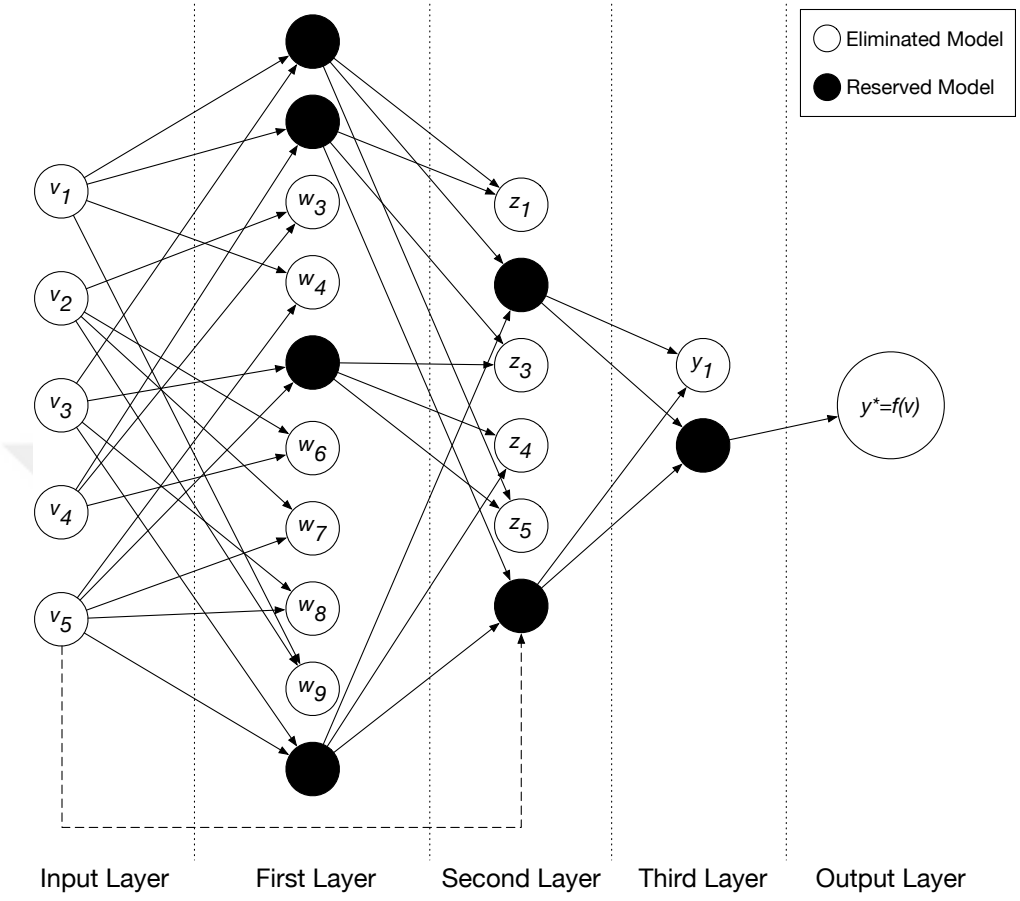

Figure 6.4. Modelling process of GMDHNN

If n is the number of neurons in a layer in GMDHNN, then the number of candidate neurons in the next layer will be calculated as $n \times (n-1)/2$ for two variable polynomials. Additionally, it should be noted that one neuron also may skip layers directly from the input variables to one of the next layers in GMDHNN as demonstrated with dashed lines from v_5 to z_6 in Figure 6.4 as an example.

The best GMDHNN model acquired from the data set imputed with EWMA

in the scope of this thesis can be computed and found as

$$\begin{aligned}
N(4) =& 2,729.142 - 10.703 \times InTemp - 5.419 \times InTemp^2 + 1,270.456 \times \\
& P1h - 1.425 \times P1h^2 - 0.211 \times InTemp \times P1h \\
N(7) =& 2,728.972 + 18.535 \times SWI - 5.167 \times SWI^2 + 1,255.867 \times P1h - \\
& 4.762 \times P1h^2 + 8.236 \times SWI \times P1h \\
N(3) =& -1.196 - 0.315 \times N(4) - 0.011 \times N(4)^2 + 1.316 \times N(7) - 0.011 \times \\
& N(7)^2 + 0.023 \times N(4) \times N(7) \\
N(1) =& 15.503 + 8.136 \times WSpeed - 0.397 \times WSpeed^2 + 0.987 \times N(3) + 2 \\
& \times 10^{-6} \times N(3)^2 - 0.004 \times WSpeed \times N(3) \\
N(11) =& 1.694 \times \mathrm{e}^{13} \times DayType - 1.694 \times \mathrm{e}^{13} \times DayType^2 + 1,242.292 \times \\
& P1d - 56.781 \times P1d^2 - 85.375 \times DayType \times P1d \\
N(10) =& 0.520 + 0.003 \times N(11) - 6 \times 10^{-6} \times N(11)^2 + 0.997 \times N(7) - 5 \times \\
& 10^{-6} \times N(7)^2 + 11 \times 10^{-6} \times N(11) \times N(7) \\
N(14) =& -5.934 - 2.091 \times HoD + 4.125 \times HoD^2 + N(7) + 4.870 \times \mathrm{e}^{-8} \times \\
& N(7) + 52 \times 10^{-6} \times HoD \times N(7) \\
N(9) =& 12.536 + 0.812 \times N(10) - 0.317 \times N(10)^2 + 0.182 \times N(14) - 0.316 \\
& \times N(14)^2 + 0.633 \times N(10) \times N(14) \\
y =& 0.655 - 0.082 \times N(1) - 0.011 \times N(1)^2 + 1.081 \times N(9) - 0.011 \times \\
& N(9)^2 + 0.022 \times N(1) \times N(9)
\end{aligned}$$

where for a quadratic reference function used by the network which is stated as

$$y = p_1 + p_2 + x_1 + p_3 x_1^2 + p_4 x_2 + p_5 x_2^2 + p_6 x_1 x_2$$

In this thesis, GMDHNN are chosen as an AI technique due to its ability to

create mathematical model for analyses, its structural and parametrical configurability (Ahmad et al., 2014). For GMDHNN, GMDH package (Dag and Yozgatligil, 2012) is utilised in RStudio environment. Benefits and drawbacks of SVM are described in the following table.

Table 6.5. Benefits and drawbacks of GMDHNN (Onwubolu, 2015; Stepashko et al., 2017)

Benefits	Drawbacks
1. Presents adaptive network topologies which can be customised to the given problem 2. Finds locally good weights owing to the reliability of the fitting technique 3. Can be trained rapidly by sparse connectivity	1. Tends to produce quite complex polynomials for simple systems 2. Do not guarantee building up the true structure 3. Biased estimates of coefficients due to the least squares method

6.2.2. Support Vector Machines

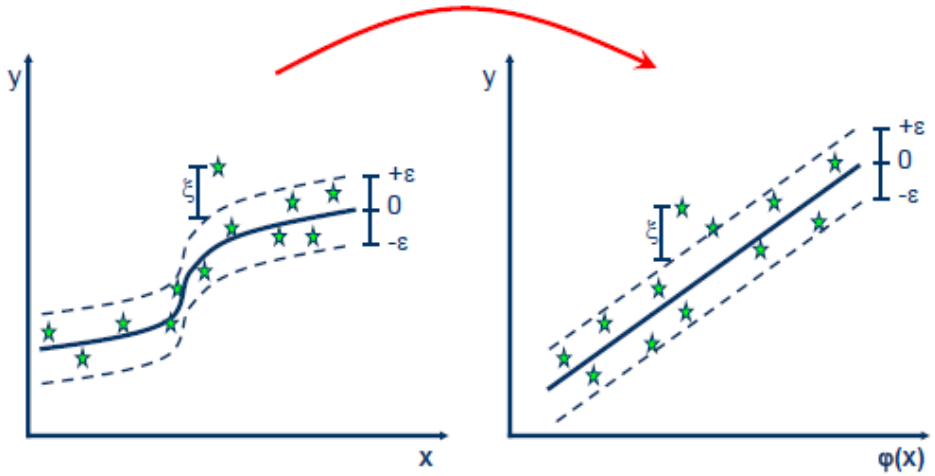

Figure 6.5. Nonlinear to linear mapping (Sayad, 2019)

SVM is an AI technique for binary classification problems. With an extension to SVM, the technique can also be applied to regression problems (i.e. SVR) for function estimation (Avşar, 2017). SVR is utilised to constitute a quite flat function $f(x)$, which is a linear regression function, that has the capability to get the nearest vector representing the real output with a tolerance ε indicating error term. Energy forecasting has nonlinear solutions just as most problems encountered on the Earth, hence input data are mapped into a higher-dimensional space by using SVR in order to find out probable linearities for training data, and linear regression technique can be applied to the consequent space

$$f(x) = \omega \cdot \varphi(x) + b$$

where $\varphi(x)$ is a function used for mapping from nonlinear space to linear space as shown in Figure 6.5, and b corresponds to the bias (Zendehboudi et al., 2018). In order to guarantee the flatness of $f(x)$, a function having a minimum norm value of $||\omega||^2$ should be obtained for each residual possessing a value smaller than ε (Timur et al., In Press). In practice, a cost can be defined for residuals that are not smaller than or equal to ε, because such function may not be obtained. For this optimisation problem, formulation of nonlinear ε-insensitive SVR (ε-SVR) is as follows

$$\min_{\omega,b} \frac{1}{2}\omega^2 + C\sum_{i=1}^{\ell}(\xi_i + \xi_i^*)$$

$$s.t. \begin{cases} y_i - f(x) \leq \varepsilon + \xi_i \\ f(x) - y_i \leq \varepsilon + \xi_i^* \\ \xi_i, \xi_i^* \geq 0 \end{cases}$$

where the penalty imposed on observations that lie outside the ε margin is controlled by C and shown by ξ_i and ξ_i^*. Dual optimisation problem of ε-SVR can be acquired

by introducing a Lagrangian function with multipliers α_j and α_j^*. Each instance must conform to Karush-Kuhn-Tucker (KKT) conditions as well. Lagrangian multipliers for all instances throughout the margin are zero. Instances having multipliers, that are not equal to zero, are support vectors. In that case, the function $f(x)$ is stated as

$$f(x_i) = \sum_{j=1}^{\ell} (a_j - a_j^*) K(x_i, x_j) + b$$

where $K(x_i, xj)$ is a nonlinear kernel function (Vrablecova et al., 2018). Linear, sigmoid, polynomial, and Gaussian RBF are universally utilised kernels. Due to its simpleness and computational efficiency over the years, RBF kernel has been qualified as one of the best kernels (Yaslan and Bican, 2017). RBF kernel function is expressed as

$$K(x_i, x_j) = \exp\left(-\frac{||x_i - x_j||^2}{2\sigma^2}\right) = \exp\left(-\gamma \cdot ||x_i - x_j||^2\right)$$

where x_i and x_j are input instances, σ^2 is variance, and $||x_i - x_j||^2$ can be described as the squared Euclidean distance among two instances (Vrablecova et al., 2018).

Moreover, cost (C) controls the SVR model's empirical risk degree, gamma (γ) controls the Gaussian function width, and epsilon (ε) controls the ε-insensitive zone's width sequentially. For the performance of SVR models, C, γ, and ε parameters should be well-determined in order to have a more accurate ε-SVR model (Zhang et al., 2017; Timur et al., In Press).

In this thesis, SVM are chosen as an AI technique due to its ability to be performed with less parameters, its kernel trick which simplifies nonlinear relationships into linear ones by mapping, and its capability in improving generalisation performance. For SVM, e1071 package (Meyer et al.,2019) is executed in RStudio environment. Benefits and drawbacks of SVM are described in the following table.

Table 6.6. Benefits and drawbacks of SVM (Timur et al., In Press)

Benefits	Drawbacks
1. Avoids overfitting by regularisation parameter 2. Kernel trick 3. Can be defined by an optimisation problem having no local minima and there are efficient methods for solution of it	1. The first and biggest limitation depends on the choice of kernel 2. The second limitation in speed and size for both in training and testing stages 3. Significantly slow in the testing stage

6.2.3. Gene Expression Programming

GEP is an enhanced methodology primarily based on GA and genetic programming (GP) (Ferreira, 2001). GEP contains five basic components, namely function set, terminal set, fitness function, control parameters, and termination condition. Although parse tree demonstration is used in traditional GP, GEP employs a fixed length of character strings for illustrating solutions to the problems, which are then visualised as parse trees (Hosseini and Gandomi, 2012). The illustration of trees in GEP is named as expression tree (ET) and shown in Figure 6.6.

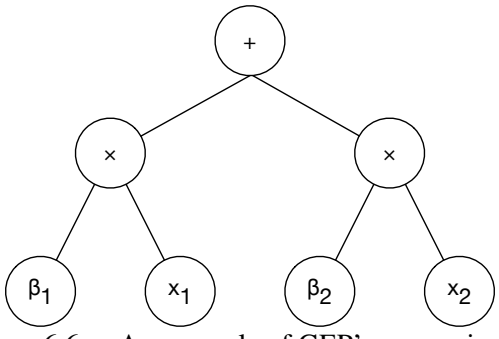

Figure 6.6. An example of GEP's expression tree

The ET indicated in Figure 6.6 corresponds to the below equation

$$y = \beta_1 x_1 + \beta_2 x_2$$

Among GEP applications, symbolic regression is a broadly utilised method to obtain a mathematical formula for a desired output from input variables of a given data set. Each sample of the data set contains input variables and outputs which can be stated as

$$\{x_{i,1}, x_{i,2}, \ldots, x_{i,n}, o_{i,1}, \ldots, o_{i,m}\}$$

where n represents the number of input variables and m corresponds to the number of outputs, $x_{i,j}$ and $o_{i,j}$ are the jth input and output of the ith sample. RMSE is frequently used for the accuracy of fitting. The symbolic regression needs to find the optimal Γ^* which minimises the RMSE for the given data set

$$\Gamma^* = \arg_{\Gamma} \min f(\Gamma)$$

where Γ is the quality of the formula, $f(\Gamma)$ gives the fitting error of Γ (Zhong et al., 2017).

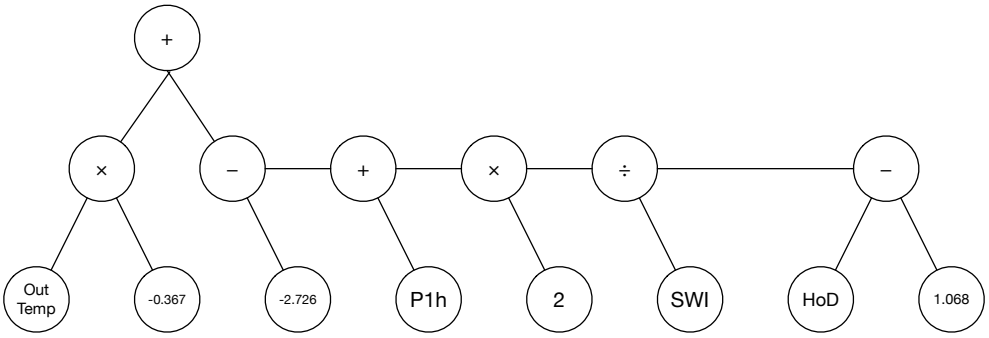

Figure 6.7. The best GEP model for the data set imputed with KalmanARIMA

The best GEP model obtained from the data set imputed with KalmanARIMA in the context of this thesis is demonstrated in Figure 6.7 and yields

the following equation

$$y = (OutTemp \times -0.367) + ((P1h + (2 \times (SWI/(HoD - 1.068)))) - 2.726)$$

where y is forecasted electrical energy consumption, $OutTemp$ and SWI represents the outdoor temperature and short-wave irradiation values taken from MERRA-2, $P1h$ corresponds to the electrical energy consumption value for the previous one hour, and HoD is the value of calendar variable standing for hour of day.

In this thesis, GEP is chosen as an AI technique due to its capability of being universal, its simpleness to understand, and its capability in containing advantages of GA and GP. For GEP, gepR package (Liu, 2018) is operated in RStudio environment. Benefits and drawbacks of GEP are described in the following table.

Table 6.7. Benefits and drawbacks of GEP (Li et al., 2005; Gan et al., 2007)

Benefits	Drawbacks
1. Extremely versatile 2. Easy to understand with 3. Faster than old GAs its linear and ramified structure 4. Has no invalid individuals 5. Overcomes the shortcomings of GA and GP	1. Does not ensure that the levels of functional complexity in the phenotype are also directly reflected in the genotype 2. The best individual is maintained, but some of better individuals may be lost 3. Needs much additional computation owing to mutations, crossovers, and rotations before reaching an optimal solution 4. Indicates premature convergence

6.2.4. Gradient Boosted Decision Trees

Boosting is a series approach for aggregating weighted outputs of several simple models recurrently to obtain an enhanced accuracy of prediction by minimising loss functions (Touzani et al., 2018).

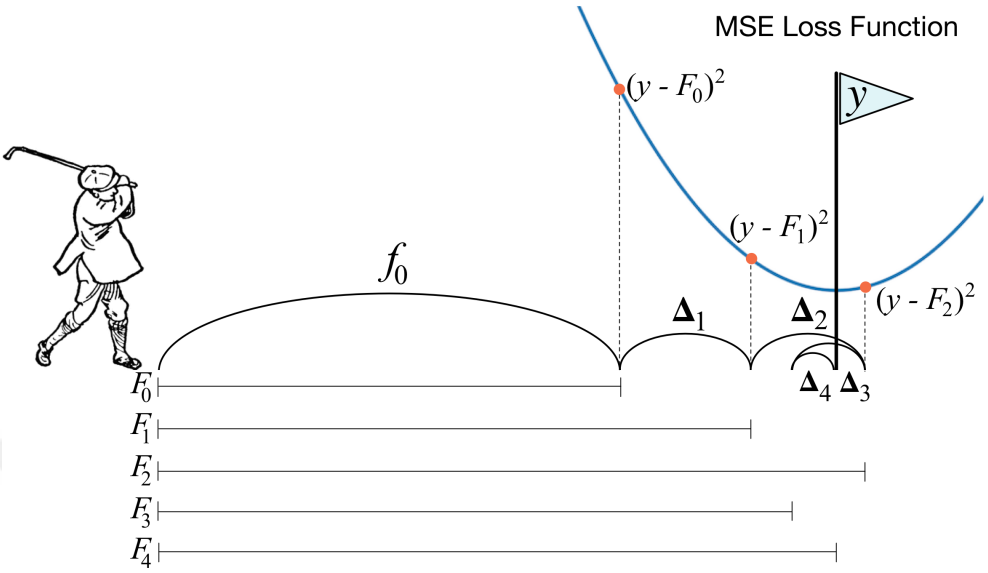

Figure 6.8. An illustration of GBDT (Parr and Howard, 2019)

Gradient boosting employs additive models which are trained in a forward stage-wise manner of the form

$$F_m(x) = F_{m-1}(x) + h_m(x)$$

where $h_m(x)$ are decision trees of constant size, $F_m(x)$ is the summation of m decision trees, and x is the set of input variables. In order to predict the response $Y_{i,t+k}$ from the training set for the best h_m

$$F_m(x_{i,t}) = F_{m-1}(x_{i,t}) + h_m(x_{i,t}) = Y_{i,t+k}$$

which yields to

$$h_m(x_{i,t}) = Y_{i,t+k} - F_{m-1}(x_{i,t})$$

where h_m also corresponds to $r_{m,i,t}$ which stands for the model fitting the current residuals at iteration m. Note that current residuals are the negative gradients of the

squared error loss function

$$r_{m,i,t} = -\frac{\partial \frac{1}{2}(Y_{i,t+k} - F_{m-1}(x_{i,t}))^2}{\partial F_{m-1}(x_{i,t})}$$

which also indicates that h_m equals to the negative gradient of the squared error loss function.

In GBDT, a learning rate v so called shrinkage factor, is defined to scale the contribution of each decision tree for a regularisation strategy which is utilised to avoid overfitting, may be stated as

$$F_m(x) = F_{m-1}(x) + vh_m(x) \qquad v \in [0;1]$$

where small values of v is recommended for better test error (Persson et al., 2017).

In this thesis, GBDT are chosen as an AI technique due to its capability of superior accuracy of prediction, being computationally fast and efficient, and being flexible for different loss functions. For GBDT, gbm package (Greenwell et al., 2019) is used in RStudio environment. Benefits and drawbacks of GBDT are described in the following table.

Table 6.8. Benefits and drawbacks of GBDT

Benefits	Drawbacks
1. Superior accuracy	1. Complex models can not be visualised
2. Computationally fast and efficient	2. Computationally expensive
3. Flexible for different loss functions and hyper-parameter tuning options	3. Requires a large grid search during tuning
4. Imputation is not necessary	4. Less interpretable

In the next chapter, experimental results and discussion of the thesis are presented.

7. EXPERIMENTAL RESULTS AND DISCUSSION

In this chapter, the detailed results of benchmark analyses are discussed from both imputation methods and S&AI techniques' point of view after expressing normalisation and evaluation criteria.

7.1. Normalisation and Evaluation Criteria

Normalisation process is essentially performed to eliminate the units of different data types in the data set, to maintain data integrity for decreasing execution time and occupying less memory, and to compare performances of heterogeneous data in a similar manner (Timur et al., In Press). In order to have a data distribution between 0 and 1 for each column vector representing different data type, the following formula can be applied for $y_{\min} = 0$ and $y_{\max} = 1$

$$x_{norm} = (y_{\max} - y_{\min}) \times (\frac{x - x_{\min}}{x_{\max} - x_{\min}}) + y_{\min}$$

where x is a column vector, $x_{\min}$ and $x_{\max}$ correspond to minimum and maximum values of the x, $y_{\min}$ and $y_{\max}$ are boundaries for distribution, and x_{norm} represents a normalised column vector converted from x respectively (He and Zheng, 2018).

Before evaluation, normalised data have to be de-normalised to calculate performance metrics which can only be compared between models whose errors are measured in the similar units such as MAE and RMSE. De-normalisation formula is the same with normalisation formula, but $y_{\max}$ and $y_{\min}$ represent the minimum and maximum known values of the previously normalised column vector x.

In order to evaluate the performances of different S&AI techniques, R^2, CV, MAE, RMSE, and MAPE are employed in this thesis. Firstly, R^2 corresponds to the coefficient of determination which is the proportion of the variance in the dependent variable that can be predicted from the independent variables (Çelik et al., 2016).

Secondly, the CV of the RMSE evaluates the relative closeness of the predictions to the actual values and can be calculated as dividing the RMSE by the mean. Thirdly, MAE measures the accuracy of continuous variables by the mean of errors without taking their direction into account. Mathematically, the MAE is the average over the verification sample of the absolute values of the differences between predicted and the actual observation. The MAE is a linear score which means that all the individual differences are weighted equally in the average.

Furthermore, RMSE is a quadratic scoring rule for the square root of the variance which also represents the average of the root forecasting error squares (Akay and Abasıkeleş, 2010). However, there is no precise criterion for an optimum value of the RMSE, hence it is based on the scales of the measured variables and the size of the sample. The RMSE can be only compared among models whose errors are measured in the same units (Salkind, 2010). On the other hand, MAPE performance metric does not depend on the magnitude of the unit of measurement. Similar to the CV, MAE, and RMSE, if the MAPE is small, then the model is accurate. The MAPE is the most widely used error measure in energy forecasting (Zor et al., 2017a). Formulae of the R^2, CV, MAE, RMSE, and MAPE are as follows

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

$$CV_{RMSE}(\%) = \frac{100}{\bar{y}}\sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}}$$

$$MAPE(\%) = \frac{100}{n}\sum_{i=1}^{n}\frac{|y_i - \hat{y}_i|}{|y_i|}$$

where y_i is actual or measured output, $\hat{y}$ is predicted output, $\bar{y}$ is mean of y_i, and n indicates the number of observations (Timur et al., In Press).

7.2. Benchmark Analyses of Experimental Results

In this section, benchmark analyses of experimental results are presented by evaluating according to R^2, CV, MAE, RMSE, and MAPE.

7.2.1. Application of Identical Constraints to S&AI Techniques

For all S&AI techniques, the following constraints are applied identically as follows:

- The same number of input categories and variables are implemented to each technique in order to produce results for the same target variable,
- For input and output variables, identical resolution is performed as temporal granularity,
- For model testing and validation, random sampling method is applied to all S&AI techniques in such a manner that 20% of data set is used to constitute training data and 80% of the data set is employed to form validation data randomly,
- For calculation of the relative importance of input variables, the same algorithm executing sensitivity analysis is implemented to all S&AI techniques in which the values of each variable are randomised and the effect on the quality of the model is measured out of 100 as percentage.

7.2.2. Experimental Results of Imputation Methods

Under this subsection, experimental results of the data set imputed with different methods for missing data are presented according to the best average

MAPE performances. The data set imputed with NOCB persistence method appeared in the first place according to the mean MAPEs of S&AI techniques as shown in Table 7.1. NOCB was followed by the data set imputed with LWMA method which were ranked as the second as illustrated in Table 7.2.

Table 7.1. Results of the data set imputed with NOCB

Imputation Method	S/AI Model	R^2 (%)	CV (%)	MAE (kWh)	RMSE (kWh)	MAPE (%)
NOCB	GBDT	99.982	0.620	11.154	16.867	0.427
	SVM	99.975	0.727	12.994	19.798	0.491
	GMDHNN	99.959	0.934	16.783	25.426	0.625
	GEP	99.951	1.026	17.935	27.926	0.674
	GRNN	99.972	0.777	17.748	26.566	0.678
	MLR	99.955	0.986	17.882	26.826	0.683
	$MLPNN_1$	99.957	0.959	18.019	26.091	0.696
	RBFNN	99.941	1.121	20.873	30.517	0.826
	$MLPNN_2$	99.932	1.204	23.457	32.759	0.927
	Average	99.958	0.928	17.427	25.844	0.670

Table 7.2. Results of the data set imputed with LWMA

Imputation Method	S/AI Model	R^2 (%)	CV (%)	MAE (kWh)	RMSE (kWh)	MAPE (%)
LWMA	GBDT	99.982	0.613	11.275	16.681	0.431
	SVM	99.976	0.710	12.799	19.325	0.484
	GMDHNN	99.961	0.917	16.589	24.958	0.623
	$MLPNN_1$	99.960	0.919	17.443	25.029	0.670
	GRNN	99.972	0.779	15.587	21.218	0.670
	MLR	99.956	0.973	17.740	26.486	0.677
	GEP	99.954	0.993	18.517	27.021	0.707
	RBFNN	99.934	1.188	21.727	32.346	0.868
	$MLPNN_2$	99.931	1.215	23.922	33.068	0.947
	Average	99.958	0.923	17.289	25.126	0.675

The data set imputed with KalmanARIMA method took the third place as indicated in Table 7.3, while the data set imputed with SpI were ranked fourth as demonstrated in Table 7.4.

Table 7.3. Results of the data set imputed with KalmanARIMA

Imputation Method	S/AI Model	R^2 (%)	CV (%)	MAE (kWh)	RMSE (kWh)	MAPE (%)
KalmanARIMA	GBDT	99.984	0.593	10.952	16.137	0.423
	SVM	99.976	0.714	12.812	19.441	0.484
	GMDHNN	99.960	0.921	16.674	25.067	0.620
	GEP	99.955	0.980	17.159	26.660	0.641
	GRNN	99.973	0.760	15.186	20.680	0.655
	MLR	99.956	0.974	17.765	26.513	0.679
	$MLPNN_1$	99.956	0.976	18.343	26.558	0.701
	RBFNN	99.917	1.337	22.462	36.381	0.914
	$MLPNN_2$	99.907	1.409	27.363	38.343	1.075
	Average	99.954	0.962	17.635	26.198	0.688

Table 7.4. Results of the data set imputed with SpI

Imputation Method	S/AI Model	R^2 (%)	CV (%)	MAE (kWh)	RMSE (kWh)	MAPE (%)
SpI	GBDT	99.981	0.632	11.329	17.212	0.437
	SVM	99.975	0.735	13.019	19.991	0.493
	GMDHNN	99.959	0.939	16.842	25.553	0.627
	GRNN	99.972	0.772	15.431	21.019	0.665
	MLR	99.954	0.991	17.944	26.965	0.686
	GEP	99.951	1.020	18.424	27.752	0.695
	$MLPNN_1$	99.955	0.978	18.278	26.612	0.705
	RBFNN	99.908	1.405	22.201	38.237	0.906
	$MLPNN_2$	99.922	1.292	25.728	35.171	1.031
	Average	99.953	0.974	17.688	26.501	0.694

Moreover, the data sets imputed with LOCF persistence method and KalmanStructTS were ranked the fifth and sixth as visualised in Table 7.5 and Table 7.6 respectively.

Table 7.5. Results of the data set imputed with LOCF

Imputation Method	S/AI Model	R^2 (%)	CV (%)	MAE (kWh)	RMSE (kWh)	MAPE (%)
LOCF	GBDT	99.983	0.607	11.125	16.511	0.429
	SVM	99.975	0.726	13.040	19.762	0.492
	GMDHNN	99.961	0.919	16.728	25.011	0.622
	GRNN	99.972	0.777	15.549	21.158	0.669
	$MLPNN_1$	99.960	0.924	17.525	25.139	0.676
	MLR	99.955	0.976	17.748	26.566	0.678
	GEP	99.953	1.004	18.158	27.342	0.688
	RBFNN	99.866	1.691	22.746	46.016	0.935
	$MLPNN_2$	99.902	1.450	27.882	39.468	1.088
	Average	99.947	1.008	17.833	27.441	0.697

Table 7.6. Results of the data set imputed with KalmanStructTS

Imputation Method	S/AI Model	R^2 (%)	CV (%)	MAE (kWh)	RMSE (kWh)	MAPE (%)
KalmanStructTS	GBDT	99.982	0.621	11.405	16.892	0.435
	SVM	99.976	0.712	12.815	19.377	0.485
	GMDHNN	99.960	0.921	16.671	25.080	0.620
	GRNN	99.973	0.761	15.209	20.710	0.656
	GEP	99.955	0.978	17.670	26.614	0.670
	MLR	99.956	0.973	17.759	26.496	0.679
	$MLPNN_1$	99.953	1.007	19.167	27.408	0.746
	RBFNN	99.937	1.163	20.439	31.656	0.822
	$MLPNN_2$	99.902	1.449	29.792	39.435	1.218
	Average	99.955	0.954	17.881	25.963	0.703

Furthermore, the data sets imputed with SMA and EWMA took the seventh and eighth place as given in Table 7.7 and Table 7.8 sequentially.

Table 7.7. Results of the data set imputed with SMA

Imputation Method	S/AI Model	R^2 (%)	CV (%)	MAE (kWh)	RMSE (kWh)	MAPE (%)
SMA	GBDT	99.982	0.623	11.456	16.955	0.437
	SVM	99.977	0.703	12.782	19.140	0.485
	GMDHNN	99.961	0.916	16.549	24.973	0.622
	GRNN	99.972	0.770	15.397	20.955	0.663
	GEP	99.953	1.003	17.881	27.292	0.677
	MLR	99.956	0.973	17.730	26.483	0.677
	$MLPNN_1$	99.956	0.965	18.448	26.261	0.718
	RBFNN	99.925	1.267	23.207	34.486	0.945
	$MLPNN_2$	99.898	1.480	28.458	40.296	1.109
	Average	99.953	0.967	17.990	26.316	0.704

Table 7.8. Results of the data set imputed with EWMA

Imputation Method	S/AI Model	R^2 (%)	CV (%)	MAE (kWh)	RMSE (kWh)	MAPE (%)
EWMA	GBDT	99.982	0.618	11.451	16.823	0.437
	SVM	99.976	0.711	12.655	19.365	0.477
	GMDHNN	99.960	0.920	16.554	25.050	0.619
	MLR	99.956	0.974	17.759	26.523	0.678
	GRNN	99.970	0.805	16.074	21.906	0.688
	GEP	99.952	1.012	18.486	27.558	0.704
	$MLPNN_1$	99.945	1.089	21.404	29.639	0.854
	RBFNN	99.876	1.629	22.942	44.342	0.941
	$MLPNN_2$	99.916	1.341	27.032	36.498	1.087
	Average	99.948	1.011	18.262	27.523	0.720

Additionally, StI and kNN for $k = 2$ imputed data sets were ranked as the ninth and tenth.

Table 7.9. Results of the data set imputed with StI

Imputation Method	S/AI Model	R^2 (%)	CV (%)	MAE (kWh)	RMSE (kWh)	MAPE (%)
StI	GBDT	99.982	0.615	11.292	16.738	0.433
	SVM	99.977	0.704	12.586	19.173	0.476
	GMDHNN	99.960	0.921	16.694	25.071	0.628
	MLR	99.956	0.976	17.819	26.560	0.681
	GRNN	99.970	0.804	16.064	21.897	0.687
	GEP	99.953	1.001	18.243	27.242	0.696
	$MLPNN_1$	99.958	0.949	18.456	25.840	0.731
	RBFNN	99.802	2.060	22.707	56.066	0.922
	$MLPNN_2$	99.872	1.657	34.204	45.104	1.397
	Average	99.937	1.076	18.674	29.299	0.739

Table 7.10. Results of the data set imputed with kNN for $k = 2$

Imputation Method	S/AI Model	R^2 (%)	CV (%)	MAE (kWh)	RMSE (kWh)	MAPE (%)
kNN for $k = 2$	GBDT	99.966	0.847	12.237	23.055	0.473
	SVM	99.959	0.933	13.221	25.405	0.507
	GRNN	99.973	0.758	15.142	20.637	0.654
	GMDHNN	99.937	1.157	18.435	31.487	0.695
	GEP	99.931	1.217	18.819	33.142	0.726
	MLR	99.930	1.225	19.064	33.358	0.738
	$MLPNN_1$	99.927	1.246	20.190	33.925	0.780
	RBFNN	99.879	1.612	24.440	43.879	1.008
	$MLPNN_2$	99.892	1.517	27.972	41.303	1.130
	Average	99.933	1.168	18.835	31.839	0.746

Lastly, the data sets imputed with LI and kNN for $k = 144$ took the eleventh and the last place.

Table 7.11. Results of the data set imputed with LI

Imputation Method	S/AI Model	R^2 (%)	CV (%)	MAE (kWh)	RMSE (kWh)	MAPE (%)
LI	GBDT	99.983	0.603	11.109	16.404	0.427
	SVM	99.976	0.709	12.621	19.309	0.476
	GMDHNN	99.960	0.919	16.665	25.024	0.627
	GRNN	99.972	0.767	15.343	20.885	0.661
	MLR	99.956	0.974	17.784	26.518	0.679
	GEP	99.953	1.000	18.084	27.223	0.683
	RBFNN	99.932	1.203	20.824	32.750	0.826
	$MLPNN_1$	99.947	1.069	21.095	29.102	0.848
	$MLPNN_2$	99.754	2.296	46.484	62.505	1.833
	Average	99.937	1.060	20.001	28.858	0.784

Table 7.12. Results of the data set imputed with kNN for $k = 144$

Imputation Method	S/AI Model	R^2 (%)	CV (%)	MAE (kWh)	RMSE (kWh)	MAPE (%)
kNN for $k = 144$	GBDT	99.965	0.864	12.751	23.530	0.504
	SVM	99.967	0.846	13.671	23.041	0.546
	GMDHNN	99.943	1.106	17.960	30.125	0.686
	MLR	99.936	1.169	18.877	31.828	0.739
	GEP	99.934	1.189	19.388	32.385	0.758
	$MLPNN_1$	99.934	1.190	20.624	32.409	0.821
	GRNN	99.949	1.040	20.547	28.320	0.846
	RBFNN	99.918	1.326	23.350	36.113	0.956
	$MLPNN_2$	99.887	1.556	29.790	42.364	1.213
	Average	99.937	1.143	19.962	31.124	0.785

7.2.2.1. Summary of Results of Imputation Methods

A variety of imputation methods applied in combination with S&AI techniques to identify individual impact of each imputation method on forecasting performance by means of R^2, CV, MAE, RMSE, and MAPE. Summary table that states the rankings based on average MAPE values of each imputation method for all S&AI techniques is given in Table 7.13.

Table 7.13. Summary of results of imputation methods

Imputation Method	R^2 (%)	CV (%)	MAE (kWh)	RMSE (kWh)	MAPE (%)
NOCB	99.958	0.928	17.427	25.844	0.670
LWMA	99.958	0.923	17.289	25.126	0.675
KalmanARIMA	99.954	0.962	17.635	26.198	0.688
SpI	99.953	0.974	17.688	26.501	0.694
LOCF	99.947	1.008	17.833	27.441	0.697
KalmanStructTS	99.955	0.954	17.881	25.963	0.703
SMA	99.953	0.967	17.990	26.316	0.704
EWMA	99.948	1.011	18.262	27.523	0.720
StI	99.937	1.076	18.674	29.299	0.739
kNN for $k = 2$	99.933	1.168	18.835	31.839	0.746
LI	99.937	1.060	20.001	28.858	0.784
kNN for $k = 144$	99.937	1.143	19.962	31.124	0.785
Average	99.947	1.014	18.290	27.669	0.717

In regard to the Table 7.13; NOCB, LWMA, and KalmanARIMA imputation methods took the first three place with respect to the experimental results for the evaluation of their general performance on all S&AI techniques.

7.2.3. Experimental Results of S&AI Techniques

Experimental results of the S&AI techniques imputed with different imputation methods for STEF are presented according to the best average MAPE

performances.

GBDT models came in the first among all S&AI techniques, the experimental results of the models are illustrated in Table 7.14, and parameters employed for each GBDT model in the conducted analyses for all imputation methods are stated as follows:

- Minimum and maximum number of trees in a series: 10 and 400,
- Depth of individual trees: 5,
- Minimum size node to split: 10,
- Proportion of rows for each tree: 50%,
- Quantile cut-off for Huber's loss function: 90%,
- Influence trimming factor: 1%,
- Number of minimum spikes for smoothing: 5,
- Pruning of series is carried out according to minimum error.

Table 7.14. Results of GBDT models

S/AI Model	Imputation Method	R^2 (%)	CV (%)	MAE (kWh)	RMSE (kWh)	MAPE (%)
GBDT	KalmanARIMA	99.984	0.593	10.952	16.137	0.423
	LI	99.983	0.603	11.109	16.404	0.427
	NOCB	99.982	0.620	11.154	16.867	0.427
	LOCF	99.983	0.607	11.125	16.511	0.429
	LWMA	99.982	0.613	11.275	16.681	0.431
	StI	99.982	0.615	11.292	16.738	0.433
	KalmanStructTS	99.982	0.621	11.405	16.892	0.435
	EWMA	99.982	0.618	11.451	16.823	0.437
	SMA	99.982	0.623	11.456	16.955	0.437
	SpI	99.981	0.632	11.329	17.212	0.437
	kNN for $k = 2$	99.966	0.847	12.237	23.055	0.473
	kNN for $k = 144$	99.965	0.864	12.751	23.530	0.504
	Average	99.979	0.655	11.461	17.817	0.441

SVM models took the second position between all S&AI techniques, the experimental results of the models are demonstrated in Table 7.15, and parameters used in the experimental analyses are given as

- Regression type: ε-SVR,
- Kernel function: Gaussian radial basis function,
- Grid and pattern search for parameter optimisation,
- 5-fold cross validation for parameter optimisation,
- Parameter optimisation in reference to minimising total error,
- Search range for model parameters:
  - C from 10^{-1} to 5000,
  - γ between 10^{-3} and 50,
  - ε among 10^{-4} and 100.

Table 7.15. Results of SVM models

S/AI Model	Imputation Method	R^2 (%)	CV (%)	MAE (kWh)	RMSE (kWh)	MAPE (%)
SVM	StI	99.977	0.704	12.586	19.173	0.476
	LI	99.976	0.709	12.621	19.309	0.476
	EWMA	99.976	0.711	12.655	19.365	0.477
	LWMA	99.976	0.710	12.799	19.325	0.484
	KalmanARIMA	99.976	0.714	12.812	19.441	0.484
	SMA	99.977	0.703	12.782	19.140	0.485
	KalmanStructTS	99.976	0.712	12.815	19.377	0.485
	NOCB	99.975	0.727	12.994	19.798	0.491
	LOCF	99.975	0.726	13.040	19.762	0.492
	SpI	99.975	0.735	13.019	19.991	0.493
	kNN for $k = 2$	99.959	0.933	13.221	25.405	0.507
	kNN for $k = 144$	99.967	0.846	13.671	23.041	0.546
	Average	99.974	0.744	12.918	20.260	0.491

GMDHNN models ranked the third in all S&AI techniques, the experimental results of the models are demonstrated in Table 7.16, and parameters used in the experimental analyses are expressed as

- Reference function: Quadratic function with two variables
- Number of maximum network layers: 20,
- Maximum polynomial order: 16,
- Convergence tolerance: 10^{-4},
- Number of neurons per layer: A fixed number of 20 neurons,
- Allowed network configurations: Previous layer and original input variables,
- Overfitting protection control: Hold-out sample 20%.

Table 7.16. Results of GMDHNN models

S/AI Model	Imputation Method	R^2 (%)	CV (%)	MAE (kWh)	RMSE (kWh)	MAPE (%)
GMDHNN	EWMA	99.960	0.920	16.554	25.050	0.619
	KalmanStructTS	99.960	0.921	16.671	25.080	0.620
	KalmanARIMA	99.960	0.921	16.674	25.067	0.620
	SMA	99.961	0.916	16.549	24.973	0.622
	LOCF	99.961	0.919	16.728	25.011	0.622
	LWMA	99.961	0.917	16.589	24.958	0.623
	NOCB	99.959	0.934	16.783	25.426	0.625
	LI	99.960	0.919	16.665	25.024	0.627
	SpI	99.959	0.939	16.842	25.553	0.627
	StI	99.960	0.921	16.694	25.071	0.628
	kNN for $k = 144$	99.943	1.106	17.960	30.125	0.686
	kNN for $k = 2$	99.937	1.157	18.435	31.487	0.695
	Average	99.957	0.957	16.929	26.069	0.634

GRNN models took the fourth place among all S&AI techniques and the

experimental results of the models are demonstrated in Table 7.17. Herein, it should be noted that GRNN had better R^2, CV, MAE, and RMSE averages in comparison with GMDHNN during the analyses, but this ranking was constituted based on MAPE averages. Parameters used in the experimental analyses are expressed as

- Kernel function: Gaussian radial basis function,
- σ configuration: σ for each input variable,
- Search range for model parameter σ: from 10^{-4} to 10,
- Step size for searching: 20,
- In order to find the optimal σ values, conjugate gradient algorithm is executed.

Table 7.17. Results of GRNN models

S/AI Model	Imputation Method	R^2 (%)	CV (%)	MAE (kWh)	RMSE (kWh)	MAPE (%)
GRNN	kNN for $k = 2$	99.973	0.758	15.142	20.637	0.654
	KalmanARIMA	99.973	0.760	15.186	20.680	0.655
	KalmanStructTS	99.973	0.761	15.209	20.710	0.656
	LI	99.972	0.767	15.343	20.885	0.661
	SMA	99.972	0.770	15.397	20.955	0.663
	SpI	99.972	0.772	15.431	21.019	0.665
	LOCF	99.972	0.777	15.549	21.158	0.669
	LWMA	99.972	0.779	15.587	21.218	0.670
	NOCB	99.972	0.777	17.748	26.566	0.678
	StI	99.970	0.804	16.064	21.897	0.687
	EWMA	99.970	0.805	16.074	21.906	0.688
	kNN for $k = 144$	99.949	1.040	20.547	28.320	0.846
	Average	99.970	0.797	16.106	22.166	0.683

MLR models ranked the fifth between all S&AI techniques, the experimental results of the models are demonstrated in Table 7.18, and parameters used in the experimental analyses are denoted as

- Confidence interval: 95%,
- Intercept term is included in MLR models as error term e,
- Statistically significant variables for all MLR models according to p-values of hypothesis test results:
  - Electrical variable: Previous 1 hour,
  - Meteorological variables: Indoor temperature, indoor humidity, outdoor temperature, and short-wave irradiation.

Table 7.18. Results of MLR models

S/AI Model	Imputation Method	R^2 (%)	CV (%)	MAE (kWh)	RMSE (kWh)	MAPE (%)
MLR	SMA	99.956	0.973	17.730	26.483	0.677
	LWMA	99.956	0.973	17.740	26.486	0.677
	LOCF	99.955	0.976	17.748	26.566	0.678
	EWMA	99.956	0.974	17.759	26.523	0.678
	KalmanStructTS	99.956	0.973	17.759	26.496	0.679
	KalmanARIMA	99.956	0.974	17.765	26.513	0.679
	LI	99.956	0.974	17.784	26.518	0.679
	StI	99.956	0.976	17.819	26.560	0.681
	NOCB	99.955	0.986	17.882	26.826	0.683
	SpI	99.954	0.991	17.944	26.965	0.686
	kNN for $k = 2$	99.930	1.225	19.064	33.358	0.738
	kNN for $k = 144$	99.936	1.169	18.877	31.828	0.739
	Average	99.952	1.014	17.989	27.593	0.689

GEP models took the sixth place in all S&AI techniques, the experimental results of the models are demonstrated in Table 7.19, and parameters used in the experimental analyses are explained as

- Model building parameters,
  - Population size: 50,

    - Number of the maximum tries for initial population: 10,000,
    - Genes per chromosome: 4,
    - Gene head length: 8,
    - Number of maximum generations: 2,000,
    - Number of generations without improvement: 1,000,
    - Stop for the best chromosome's fitness score: 1.0
- Fitness properties,
    - Fitness function: Mean squared error,
    - Hit tolerance: 1%,
    - Selection range: 100,
- Expression simplification,
    - Algebraic Simplification: Allowed,
    - Parameter values managing to simplify after training,
        - Number of generations for simplification: 500,
        - Number of generations without improvement: 200,
- Allowed functions,
    - Addition (+),
    - Subtraction (–),
    - Multiplication ($\times$),
    - Division (/),
    - Square root ($\sqrt{\ }$),
- Evolution parameters,
    - Mutation rate: 4.4%,
    - Inversion rate: 10%,

    - Insertion sequence transposition rate: 10%,
    - Root insertion sequence transposition rate: 10%,
    - Gene transposition rate: 10%,
    - One-point rate: 30%,
    - Two-point rate: 30%,
    - Gene rate: 10%,
- Link function used for all genes: Addition (+),
- Features of random constants,
    - Random real constants per gene: 10,
    - Minimum and maximum constant values: -10 and 10,
    - Mutation rate: 1%,
    - Nonlinear regression did never enhanced the models.

Table 7.19. Results of GEP models

S/AI Model	Imputation Method	R^2 (%)	CV (%)	MAE (kWh)	RMSE (kWh)	MAPE (%)
GEP	KalmanARIMA	99.955	0.980	17.159	26.660	0.641
	KalmanStructTS	99.955	0.978	17.670	26.614	0.670
	NOCB	99.951	1.026	17.935	27.926	0.674
	SMA	99.953	1.003	17.881	27.292	0.677
	LI	99.953	1.000	18.084	27.223	0.683
	LOCF	99.953	1.004	18.158	27.342	0.688
	SpI	99.951	1.020	18.424	27.752	0.695
	StI	99.953	1.001	18.243	27.242	0.696
	EWMA	99.952	1.012	18.486	27.558	0.704
	LWMA	99.954	0.993	18.517	27.021	0.707
	kNN for $k = 2$	99.931	1.217	18.819	33.142	0.726
	kNN for $k = 144$	99.934	1.189	19.388	32.385	0.758
	Average	99.949	1.035	18.230	28.180	0.693

$MLPNN_1$ models ranked the seventh among all S&AI techniques, the experimental results of the models are showed in Table 7.20, and parameters used in the experimental analyses are explained as

- Number of hidden layers: 1,
- Search for the optimal number of neurons in the hidden layer,
  - Minimum and maximum neuron number: 2 and 20,
  - Neuron step size: 1,
  - Maximum steps without change: 8,
  - Evaluation of each model's quality: Hold-out sample 20%,
- Test data are used to detect overfitting,
- Overfitting parameters,
  - Percentage of training rows for hold-out: 20%,
  - Number of maximum steps without change: 10,
- Activation functions,
  - Hidden layer activation function: Logistic sigmoid,
  - Output layer activation function: Linear,
- Training method: Scaled conjugate gradient,
- Conjugate gradient parameters,
  - Number of convergence trials: 4,
  - Number of maximum iterations: 10^4,
  - Number of iterations without improvement: 100,
  - Convergence tolerance: 10^{-5},
  - Minimum improvement in residuals: 10^{-6},
  - Minimum gradient: 10^{-6}.

Table 7.20. Results of $MLPNN_1$ models

S/AI Model	Imputation Method	R^2 (%)	CV (%)	MAE (kWh)	RMSE (kWh)	MAPE (%)
$MLPNN_1$	LWMA	99.960	0.919	17.443	25.029	0.670
	LOCF	99.960	0.924	17.525	25.139	0.676
	NOCB	99.957	0.959	18.019	26.091	0.696
	KalmanARIMA	99.956	0.976	18.343	26.558	0.701
	SpI	99.955	0.978	18.278	26.612	0.705
	SMA	99.956	0.965	18.448	26.261	0.718
	StI	99.958	0.949	18.456	25.840	0.731
	KalmanStructTS	99.953	1.007	19.167	27.408	0.746
	kNN for $k = 2$	99.927	1.246	20.190	33.925	0.780
	kNN for $k = 144$	99.934	1.190	20.624	32.409	0.821
	LI	99.947	1.069	21.095	29.102	0.848
	EWMA	99.945	1.089	21.404	29.639	0.854
	Average	99.951	1.022	19.083	27.834	0.745

RBFNN models took the eighth place between all S&AI techniques, the experimental results of the models are showed in Table 7.21, and parameters used in the experimental analyses are given as

- Network parameters,
  - Number of maximum neurons: 100,
  - Minimum network tolerance for mean squared error: 10^{-6},
  - Minimum neuron tolerance for mean squared error: 10^{-5},
  - Minimum and maximum spread: 0.01 and 400,
  - Minimum and maximum lambda: 0.001 and 10,
- Tuning parameters for neurons,
  - Size of population: 200,
  - Number of maximum generations: 20,

- Number of consecutive generations without improvement: 5,
- Number of maximum iterations: 50,
- Boosting tolerance: 10^{-4}

Table 7.21. Results of RBFNN models

S/AI Model	Imputation Method	R^2 (%)	CV (%)	MAE (kWh)	RMSE (kWh)	MAPE (%)
RBFNN	KalmanStructTS	99.937	1.163	20.439	31.656	0.822
	LI	99.932	1.203	20.824	32.750	0.826
	NOCB	99.941	1.121	20.873	30.517	0.826
	LWMA	99.934	1.188	21.727	32.346	0.868
	SpI	99.908	1.405	22.201	38.237	0.906
	KalmanARIMA	99.917	1.337	22.462	36.381	0.914
	StI	99.802	2.060	22.707	56.066	0.922
	LOCF	99.866	1.691	22.746	46.016	0.935
	EWMA	99.876	1.629	22.942	44.342	0.941
	SMA	99.925	1.267	23.207	34.486	0.945
	kNN for $k = 144$	99.918	1.326	23.350	36.113	0.956
	kNN for $k = 2$	99.879	1.612	24.440	43.879	1.008
	Average	99.903	1.417	22.326	38.566	0.906

$MLPNN_2$ models ranked the last in all of the S&AI techniques, the experimental results of the models are depicted in Table 7.22, and parameters used in the experimental analyses are highlighted as

- Number of hidden layers: 2,
- Search for the optimal number of neurons in the first hidden layer,
  - Minimum and maximum neuron number: 2 and 20,
  - Neuron step size: 1,
  - Maximum steps without change: 8,
  - Evaluation of each model's quality: Hold-out sample 20%,

Table 7.22. Results of $MLPNN_2$ models

S/AI Model	Imputation Method	R^2 (%)	CV (%)	MAE (kWh)	RMSE (kWh)	MAPE (%)
$MLPNN_2$	NOCB	99.932	1.204	23.457	32.759	0.927
	LWMA	99.931	1.215	23.922	33.068	0.947
	SpI	99.922	1.292	25.728	35.171	1.031
	KalmanARIMA	99.907	1.409	27.363	38.343	1.075
	EWMA	99.916	1.341	27.032	36.498	1.087
	LOCF	99.902	1.450	27.882	39.468	1.088
	SMA	99.898	1.480	28.458	40.296	1.109
	kNN for $k = 2$	99.892	1.517	27.972	41.303	1.130
	kNN for $k = 144$	99.887	1.556	29.790	42.364	1.213
	KalmanStructTS	99.902	1.449	29.792	39.435	1.218
	StI	99.872	1.657	34.204	45.104	1.397
	LI	99.754	2.296	46.484	62.505	1.833
	Average	99.893	1.489	29.348	40.526	1.171

- Number of neurons in the second hidden layer: 4,
- Test data are used to detect overfitting,
- Overfitting parameters,
  - Percentage of training rows for hold-out: 20%,
  - Number of maximum steps without change: 10,
- Activation functions,
  - Hidden layer activation function: Logistic sigmoid,
  - Output layer activation function: Linear,
- Training method: Scaled conjugate gradient,
- Conjugate gradient parameters,
  - Number of convergence trials: 4,
  - Number of maximum iterations: 10^4,

- Number of iterations without improvement: 100,
- Convergence tolerance: 10^{-5},
- Minimum improvement in residuals: 10^{-6},
- Minimum gradient: 10^{-6}.

7.2.3.1. Summary of Results of S&AI Techniques

Several S&AI techniques implemented along with different imputation methods to forecast one hour ahead electrical energy consumption of the hospital and performance of the techniques evaluated with respect to R^2, CV, MAE, RMSE, and MAPE. Summary table that indicates the rankings based on average MAPE values of each S/AI technique in combination with variously imputed data sets is demonstrated in Table 7.23.

Table 7.23. Summary of results of S&AI techniques

S/AI Model	R^2 (%)	CV (%)	MAE (kWh)	RMSE (kWh)	MAPE (%)
GBDT	99.979	0.655	11.461	17.817	0.441
SVM	99.974	0.744	12.918	20.260	0.491
GMDHNN	99.957	0.957	16.929	26.069	0.634
GRNN	99.970	0.797	16.106	22.166	0.683
MLR	99.952	1.014	17.989	27.593	0.689
GEP	99.949	1.035	18.230	28.180	0.693
$MLPNN_1$	99.951	1.022	19.083	27.834	0.745
RBFNN	99.903	1.417	22.326	38.566	0.906
$MLPNN_2$	99.893	1.489	29.348	40.526	1.171
Average	99.947	1.014	18.290	27.669	0.717

According to the Table 7.23, GBDT and SVM techniques overwhelmed the others in terms of the experimental results for the evaluation of their overall performance on diversely imputed data sets.

7.2.4. Experimental Results of Importance of Variables

Along with the conducted experimental analyses, relative importance of each input variable was computed meticulously in reference to an algorithm which uses sensitivity analysis in which the values of each variable are randomised and the effect on the quality of the model is measured out of 100 as percentage.

In Table 7.24, ranking of variable importance for all S&AI techniques is given by means of elaborating the best model having minimum MAPE for each S/AI technique. In details, GBDT-KalmanARIMA, SVM-StI, GMDHNN-EWMA, GRNN-kNN (for $k = 2$), MLR-SMA, GEP-KalmanARIMA, $MLPNN_1$-LWMA, RBFNN-KalmanStructTS, and $MLPNN_1$-NOCB combinations are emphasised in Table 7.24. Abbreviations of input variables are given in the footnotes.

Rankings in Table 7.24 are sequenced between 1 to 17 for each row. If a cell in the table is depicted with a hypen (–), then column variable corresponding to that cell is not used in the analysis of the cell's row implying an S/AI technique. For instance, pressure variable is not employed for MLR implementation. The first three ranked variables in each row are marked in bold.

Table 7.25 states percentage values of each variable's relative importance for each S/AI technique with respect to the algorithm utilised for calculating the relative importance of input variables. The distribution of algorithm is from 0% to 100%. If the value of the cell is between 1% and 100%, it is written in numbers. Else, it is represented by a smaller or less than sign (<). Similarly, unused cells are illustrated by a hypen (–) in Table 7.25 too. Detailed explanation is also provided in the footnote.

In the following pages, Table 7.24 and Table 7.25 are presented consecutively.

Table 7.24. Ranking of variable importance

Category	Electrical			Meteorological									Calendar				
Variable	P1h	P1d	P1w	InTemp	InRH	OutTemp	OutRH	Pressure	WSpeed	WDirection	Rainfall	SWI	HoD	DoM	DayType	WoY	MoY
GBDT	**1**	4	7	6	9	5	10	13	11	12	16	**2**	**3**	14	8	15	17
SVM	**1**	9	8	4	13	7	11	14	16	15	17	6	5	10	12	**2**	**3**
GMDHNN	**1**	**2**	–	6	–	–	–	–	7	–	–	**3**	5	–	4	–	–
GRNN	**1**	5	**3**	16	6	14	11	8	10	13	**2**	7	4	9	12	17	15
MLR	**1**	9	10	6	7	5	13	–	11	12	16	4	14	8	15	**2**	**3**
GEP	**1**	–	–	–	–	4	–	–	–	–	–	**2**	**3**	–	–	–	–
$MLPNN_1$	**1**	7	6	5	13	11	12	16	15	14	17	4	10	8	9	**2**	**3**
RBFNN	**1**	5	11	8	7	12	14	17	10	13	**2**	6	9	15	16	**3**	4
$MLPNN_2$	**1**	5	7	**2**	13	**3**	8	16	12	11	17	4	14	15	9	10	6

*Abbreviations for variables are expressed as follows:

P1h: Previous 1 hour, P1d: Previous 1 day, P1w: Previous 1 week, InTemp: Indoor Temperature, InRH: Indoor Relative Humidity, WSpeed: Wind Speed, WDirection: Wind Direction, SWI: Short-Wave Irradiation, HoD: Hour of Day, DoM: Day of Month, WoY: Week of Year, MoY: Month of Year

Table 7.25. Percentage of relative importance

Category	Electrical			Meteorological									Calendar				
Variable	P1h	P1d	P1w	InTemp	InRH	OutTemp	OutRH	Pressure	WSpeed	WDirection	Rainfall	SWI	HoD	DoM	DayType	WoY	MoY
GBDT	100	<	<	<	<	<	<	<	<	<	<	<	<	<	<	<	<
SVM	100	<	<	<	<	<	<	<	<	<	<	<	<	<	<	<	<
GMDHNN	100	86.5	–	<	–	–	–	–	<	–	–	<	<	–	<	–	–
GRNN	100	<	<	<	<	<	<	<	<	<	7.3	<	<	<	<	<	<
MLR	100	<	<	<	<	<	<	–	<	<	<	<	<	<	<	<	<
GEP	100	–	–	–	–	<	–	–	–	–	–	5.2	4.8	–	–	–	–
$MLPNN_1$	100	<	<	<	<	<	<	<	<	<	<	<	<	<	<	<	<
RBFNN	100	<	<	<	<	<	<	<	<	<	<	<	<	<	<	<	<
$MLPNN_2$	100	<	<	<	<	<	<	<	<	<	<	<	<	<	<	<	<

Table notes are given as follows:

1. Above table contains percentage (%) values which indicate the relative importance of each input variable as a predictor.

2. "<" represents for values below 1% and "–" stands for variables that are not used.

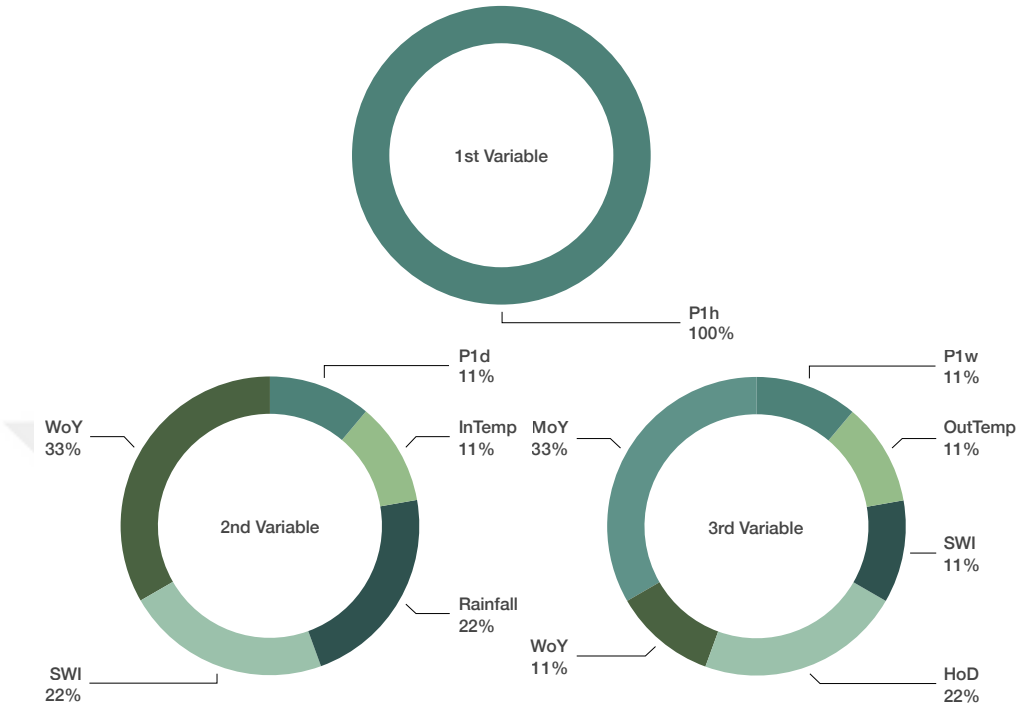

Figure 7.1. Importance of variables: 1st, 2nd, and 3rd

Distribution of the first three most important variables is visualised in Figure 7.1. Previous 1 hour is the most important variable and took the first place for all S&AI techniques according to the relative importance analysis. Week of year, short-wave irradiation, rainfall, previous 1 day, and indoor temperature variables came in second in the conducted analysis. Month of year, hour of day, week of year, previous 1 week, outdoor temperature, and short-wave irradiation ranked the third during the analyses.

In Figure 7.2, distribution of all used variables for the first three positions is shown on the left of the figure which are also mentioned in Figure 7.1 separately, distribution of unused variables in any of analyses is illustrated on the right of the

figure and they are month of year, week of year, day type, day of month, wind direction, wind speed, pressure, outdoor relative humidity, outdoor temperature, indoor relative humidity, rainfall, indoor temperature, previous 1 week, and previous 1 day.

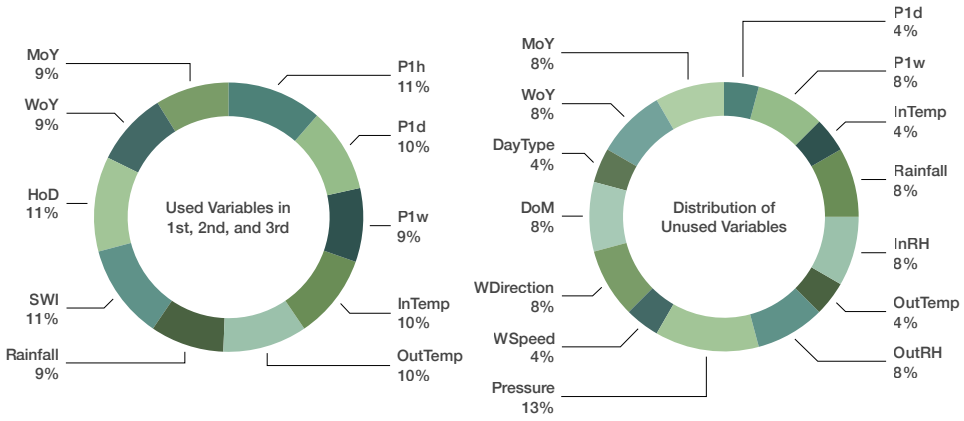

Figure 7.2. Importance of used and unused variables

7.3. Discussion

Based on the results obtained from the comprehensive benchmark analyses conducted rigorously, the significant findings are discussed and deduced as follows:

- The first three best MAPE performances of the combination of both imputation methods and S&AI techniques were found as GBDT-KalmanARIMA (0.423%), GBDT-LI (0.427%), and GBDT-NOCB (0.427%). Therefore, it is recommended that GBDT can be employed as an AI technique for STEF problem, and KalmanARIMA, LI, and NOCB may be the first choices for the imputation of energy forecasting data sets having missing, erroneous, and anomalous values.

- In terms of each S/AI technique separate performance evaluation, S&AI techniques having MAPE values below the average (0.717%) as indicated in Table 7.23 were detected as GBDT (0.441%), SVM (0.491%), GMDHNN (0.634%), GRNN (0.683%), MLR (0.689%), and GEP (0.693%) sequentially. Thus, implementation of GBDT and SVM techniques may yield satisfactory results for STEF problem, while GMDHNN, GRNN, MLR, and GEP techniques can produce reasonable outcomes. Herein, it should be noted that GBDT resulted in an overwhelming performance by any statistical or computational measure in comparison with other S&AI techniques for STEF.
- When FTHC methodology with its tolerance check mechanism is assessed with respect to R^2 values acquired from the analyses, it is thought that the proposed methodology can be performed to any application in which sudden changes in the magnitude of parameters are considered as unwanted circumstances.
- Moreover, within all input variables; electrical variable previous 1 hour, meteorological variable short-wave irradiation, and calendar variable hour of day should be emphasised as prerequisite for one hour ahead STEF. In order to reach satisfactory results for one hour ahead STEF; electrical variables containing previous 1 hour, 1 day, and 1 week; meteorological variables including short-wave irradiation, indoor and outdoor temperatures and rainfall; and calendar variables consisting of hour of day, week of year, and month of year should be at least employed.
- Lastly, a correlativity between seasonality and electrical energy consumption of HVAC systems at the hospital was founded out. For

this reason, it is inferred from the experimental results that seasonality has a dominant influence on STEF performance of the hospitals where HVAC systems constitute the major part of electrical energy consumption. In addition to those, it is understood that day type variable, which determines whether a day is either weekend and public holiday or working day, is a preeminent classifier in STEF of hospitals owing to the operation of polyclinics.

In the following chapter, conclusions and future perspectives belonging to the thesis are presented respectively.

8. CONCLUSIONS AND FUTURE PERSPECTIVES

8.1. Conclusions

In this thesis, a study of research and application of real-time STEF using artificial intelligence (AI) based techniques has been conducted for a large hospital complex in Adana, Turkey. To do so, real-time electrical energy consumption and meteorological data of a large hospital complex are acquired at first by utilising an energy logger connected to a humidity-temperature transducer on-site and MERRA-2 data to form a very-short term raw data set with a sampling period of 10 minutes in RStudio environment. After then, a novel methodology is developed and named as forecast time horizon converter which has the capability of identifying missing and erroneous values in the raw data set, replacing these values with NA values, imputing the NA values with different imputation methods, and completing the conversion process by creating a cleansed short-term data set with a sampling period of 1 hour. Next, short-term electrical energy consumption of the hospital is predicted by using MLR as a statistical technique for benchmarking purposes and AI based techniques containing SVM, GEP, GBDT, and ANN consisting of MLPNN, RBFNN, GRNN, and GMDHNN under identical constraints. Moreover, an algorithm involving sensitivity analysis, in which the values of each variable are randomised and the effect on the quality of the model is computed out of hundred, is implemented to all S&AI techniques for calculation of relative importance belonging to each input variable. Eventually, benchmark analyses of the obtained results are presented by evaluating R^2, CV, MAE, RMSE, and MAPE.

Consequently, detailed results of benchmark analyses indicated that KalmanARIMA (with GBDT resulting in MAPE of 0.423%), LI (via GBDT calculated as MAPE of 0.427%), and NOCB (by performing GBDT computed as

MAPE of 0.427%) may be applied to energy forecasting data sets for STEF problem in future studies. Electrical variables containing previous 1 hour, 1 day, and 1 week; meteorological variables including short-wave irradiation, indoor and outdoor temperatures and rainfall; and calendar variables consisting of hour of day, week of year, and month of year should be at least utilised in order to reach satisfactory results. Herein, previous 1 hour, short-wave irradiation, and hour of day variables should be underlined as prerequisites within all input variables for one hour ahead STEF. In addition, it can be inferred from the results that GBDT (with KalmanARIMA resulting in MAPE of 0.423%) delivered an outstanding performance by any measure in comparison with other S&AI techniques for STEF. Hence, it is suggested to employ GBDT for STEF more frequently in the energy forecasting literature.

8.2. Future Perspectives

Future perspectives of the thesis are described as follows:

- The created data set may be used for undergraduate and graduate level courses in relation with management, planning, economics, and analytics of electrical energy in electrical and electronics engineering.
- For future studies, researchers may benefit from the data set for electrical energy consumption, electric load and demand studies from 10 minutes to 1 hour ahead forecasting by the FTHC.
- In addition to current data set, hospital related new variables may be incorporated into the data set such as occupancy, number of hourly inpatients and staff, number of hourly registered patients at emergency services, polyclinics, or entire hospital, etc.
- In the thesis, R programming language is executed in RStudio

environment from beginning to end. Conducted analyses in the thesis may be verified in another environment utilising Python programming language as a future work.

- In the scope of the thesis, scaled conjugate gradient training algorithm is employed for MLPNN. Additionally, other back-propagation training algorithms may be performed to produce a benchmark study as one more future work. Moreover, logistic sigmoid activation function is used for MLPNN. Other activation functions, for instance, rectified linear unit may be performed to MLPNN as another activation function for a comparative study.
- In the context of this thesis, shallow ANN methodologies are generally used for forecasting purposes. Deep network methodologies including gated recurrent unit networks, convolutional neural networks, and long short-term memory networks may be considered to be implemented in the future.
- In this thesis, best results were obtained by using GBDT technique. An advanced version of GBDT, XGBoost technique may be managed to adapt STEF problem for future studies.
- Throughout this thesis, point forecasting approaches are applied. Probabilistic forecasting approaches may be adapted for future works.

REFERENCES

Ahmad, A. S., Hassan, M. Y., Abdullah, M. P., Rahman, H. A., Hussin, F., Abdullah, H., and Saidur, R., 2014. A review on applications of ANN and SVM for building electrical energy consumption forecasting. Renewable and Sustainable Energy Reviews, 33:102–109. DOI: 10.1016/j.rser.2014.01.069

Ahmad, M. W., Mourshed, M., and Rezgui, Y., 2017. Trees vs Neurons: Comparison between random forest and ANN for high-resolution prediction of building energy consumption. Energy and Buildings, 147:77–89. DOI: 10.1016/j.enbuild.2017.04.038

Akay, M. F., and Abasıkeleş, İ., 2010. Predicting the performance measures of an optical distributed shared memory multiprocessor by using support vector regression. Expert Systems with Applications, 37:6293–6301. DOI: 10.1016/j.eswa.2010.02.092

Alice, M., 2018. Imputing Missing Data with R; MICE package. Data Management in R. Online Source. URL: https://datascienceplus.com/imputing-missing-data-with-r-mice-package/

Al-Madfai, H., 2002. Weather corrected electricity demand forecasting. PhD Thesis, University of Glamorgan School of Technology, Trefforest, 1–304.

Amasyali, K., and El-Gohary, N. M., 2018. A review of data-driven building energy consumption prediction studies. Renewable and Sustainable Energy Reviews, 81:1192–1205. DOI: 10.1016/j.rser.2017.04.095

Amral, N., Ozveren, C. S., and King, D., 2007. Short term load forecasting using multiple linear regression. Proceedings of the 42nd

International Universities Power Engineering Conference, Brighton, 1192–1198. DOI: 10.1109/UPEC.2007.4469121

Auguie, B., 2017. gridExtra: Miscellaneous Functions for ”Grid” Graphics. R package version 2.3. URL: https://CRAN.R-project.org/package=gridExtra

Avşar, E., 2017. Dimensionality reduction for predicting CO conversion in water gas shift reaction over Pt-based catalysts using support vector regression models. International Journal of Hydrogen Energy, 42:23326–23333. DOI: 10.1016/j.ijhydene.2016.12.091

Bagnasco, A., Saviozzi, M., Silvestro, F., Vinci, A., Grillo, S., and Zennaro, E., 2014. Artificial neural network application to load forecasting in a large hospital facility. Proceedings of International Conference on Probabilistic Methods Applied to Power Systems (PMAPS), 16, Durham. DOI: 10.1109/PMAPS.2014.6960579

_____, Fresi, F., Saviozzi, M., Silvestro, F., and Vinci, A., 2015. Electrical consumption forecasting in hospital facilities: An application case. Energy and Buildings, 103:261–270. DOI: 10.1016/j.enbuild.2015.05.056

Ben Taieb, S., 2014. Machine learning strategies for multi-step-ahead time series forecasting. PhD Thesis, Department of Computer Science, Free University of Brussels, Brussels, 1–190.

Bergmeier, C., and Benitez, J. M., 2012. Neural Networks in R Using the Stuttgart Neural Network Simulator: RSNNS. Journal of Statistical Software, 46(7):1–26. DOI: 10.18637/jss.v046.i07

Chae, Y. T., Horesh, R., Hwang, Y., and Lee, Y. M., 2016. Artificial neural network model for forecasting sub-hourly electricity usage in

commercial buildings. Energy and Buildings, 111:184–194. DOI: 10.1016/j.enbuild.2015.11.045

Chandramitasari, W., Kurniawan, B., and Fujimura, S., 2018. Building deep neural network model for short term electricity consumption forecasting. Proceedings of 2018 International Symposium on Advanced Intelligent Informatics (SAIN), Yogyakarta, 43–48. DOI: 10.1109/SAIN.2018.8673340

Chapra, S. C., and Canale, R. P., 2010. Numerical Methods for Engineers, 6th Ed. McGraw-Hill.

Chasset, P.-O., 2013. GRNN: General regression neural network for the statistical software R. Independent scientist. Nancy, France. URL: https://CRAN.R-project.org/package=grnn

Chen, C. R., Shih, S. C., and Hu, S. C., 2005. Short-term electricity forecasting of air-conditioners of hospital using artificial neural networks. Proceedings of IEEE/PES Transmission Distribution Conference Exposition: Asia and Pacific, Dalian, 1–5. DOI: 10.1109/TDC.2005.1547136

Chen, Y., and Tan, H., 2017. Short-term prediction of electric demand in building sector via hybrid support vector regression. Applied Energy, 204:1363–1374. DOI: 10.1016/j.apenergy.2017.03.070

Chen, Y., Xu, P., Chu, Y., Li, W., Wu, Y., Ni, L., Bao, Y., and Wang, K., 2017. Short-term electrical load forecasting using the support vector regression (SVR) model to calculate the demand response baseline for office buildings. Applied Energy, 195:659–670. DOI: 10.1016/j.apenergy.2017.03.034

Cherkassy, V., Chowdhury, S. R., Landenberger, V., Tewari, S., and Bursch,

P., 2011. Prediction of electric power consumption for commercial buildings. Proceedings of International Joint Conference on Neural Networks, San Jose, CA, 666–672. DOI: 10.1109/IJCNN.2011.60332856

Chitsaz, H., Shaker, H., Zareipour, H., Wood, D., and Amjady, N., 2015. Short-term electricity load forecasting of buildings in microgrids. Energy and Buildings, 99:50–60. DOI: 10.1016/j.enbuild.2015.04.011

Çelik, Ö., Teke, A., Yıldırım, H. B., 2016. The optimized artificial neural network model with Levenberg-Marquardt algorithm for global solar radiation estimation in eastern Mediterranean region of Turkey. Journal of Cleaner Production, 116:1–12. DOI: 10.1016/j.jclepro.2015.12.082

Dag, O., and Yozgatligil, C., 2012. GMDH: An R Package for Short Term Forecasting via GMDH-Type Neural Network Algorithms. The R Journal, 8(1):379–386. DOI: 10.32614/RJ-2016-028

Damrongsak, D., Wongsapai, W., and Thinate, N., 2018. Factor impacts and target setting of energy consumption in Thailands hospital building. Chemical Engineering Transactions, 70:1585–1590. DOI: 10.3303/CET1870265

Daut, M. A. M., Hassan, M. Y., Abdullah, H., Rahman, H. A., Abdullah, M. P., and Hussin, F., 2017. Building electrical energy consumption forecasting analysis using conventional and artificial intelligence methods: A review. Renewable and Sustainable Energy Reviews, 70:1108–1118. DOI: 10.1016/j.rser.2016.12.015

De Giorgi, M. G., Malvoni, M., and Congedo, P. M., 2016. Comparison of

strategies for multi-step ahead photovoltaic power forecasting models based on hybrid group method of data handling networks and least square support vector machine. Energy, 107:360–373. DOI: 10.1016/j.energy.2016.04.020

Deb, C., Zhang, F., Yang, J., Lee, S. E., and Shah, K. W., 2017. A review on time series forecasting techniques for building energy consumption. Renewable and Sustainable Energy Reviews, 74:902–924. DOI: 10.1016/j.rser.2017.02.085

Demirhan, H., and Renwick, Z., 2018. Missing value imputation for short to mid-term horizontal solar irradiance data. Applied Energy, 225:998–1012. DOI: 10.1016/j.apenergy.2018.05.054

Dhar, A., Reddy, T. A., and Claridge, D. E., 1998. Modeling hourly energy use in commercial buildings with Fourier series functional forms. Journal of Solar Energy Engineering, 120(3):217–223. DOI: 10.1115/1.2888072

———, Reddy, T. A., and Claridge, D. E., 1999a. A Fourier series model to predict hourly heating and cooling energy use in commercial buildings with outdoor temperature as the only weather variable. Journal of Solar Energy Engineering, 121(1):47–53. DOI: 10.1115/1.2888142

———, Reddy, T. A., and Claridge, D. E., 1999b. Generalization of the Fourier series approach to model hourly energy use in commercial buildings. Journal of Solar Energy Engineering, 121(1):54–62. DOI: 10.1115/1.2888143

Dodier, R. H., and G. P. Henze, 2004. Journal of Solar Energy Engineering, 126(1):592–600. DOI: 10.1115/1.1637640

Dong, Y., and Peng, C. Y. J., 2013. Principled missing data methods for researchers. SpringerPlus, 2(1):1–17. DOI: 10.1186/2193-1801-2-222

Durijic, N., and Novakovic, V., 2012. Identifying important variables of energy use in low energy office building by using multivariate analysis. Energy and Buildings, 45:91–98. DOI: 10.1016/j.enbuild.2011.10.031

EMRA, 2019. Energy Market Regulatory Authority. Law regarding to the use of renewable energy sources for electrical energy production Version 3. URL: http://eskiweb.epdk.org.tr/TR/Dokuman/6777

Fan, C, Sun, Y., Zhao, Y., Song, M., and Wang, J., 2019. Deep learning-based feature engineering methods for improved building energy prediction. Applied Energy, 240:35–45. DOI: 10.1016/j.apenergy.2019.02.052

Fay, D., 2004. A strategy for short-term load forecasting in Ireland. PhD Thesis, Dublin City University, School of Electronic Engineering, Dublin, 1–276.

Ferreira, C., 2001. Gene Expression Programming: A New Adaptive Algorithm for Solving Problems. Complex Systems, 13(2):87–129. URL: https://arxiv.org/pdf/cs/0102027

Foucquier, A., Robert, S., Suard, F., Stephan, L., and Jay, A., 2013. State of the art building modelling and energy performances prediction: A review. Renewable and Sustainable Energy Reviews, 23:272–288. DOI: 10.1016/j.rser.2013.03.004

Gan, Z., Yang, Z., Li, G., and Jiang, M., 2007. Automatic Modeling of Complex Functions with Clonal Selection-Based Gene Expression

Programming. Proceedings of the Third International Conference on Natural Computation (ICNC 2007), Haikou. DOI: 10.1109/ICNC.2007.278

Gelaro, R., McCarty, W., Suarez, M. J., Todling, R., Molod, A., Takacs, L., Randles, C. A., Darmenov, A., Bosilovich, M. G., Reichle, R., Wargan, K., Coy, L., Cullather, R., Draper, C., Akella, S., Buchard, V., Conaty, A., Da Silva, A. M., Gu, W., Kim, G. K., Koster, R., Lucchesi, R., Merkova, D., Nielsen, J. E., Partyka, G., Pawson, S., Putman, W., Rienecker, M., Schubert, S. D., Sienkiewicz, M., and Zhao, B., 2017. The Modern-Era Retrospective Analysis for Research and Applications, Version 2 (MERRA-2). Journal of Climate, 30:5419–5454. DOI: 10.1175/JCLI-D-16-0758.1

Gonzalez, P. A., and Zamarreno, J. M., 2005. Prediction of hourly energy consumption in buildings based on a feedback artificial neural network. Energy and Buildings, 37:595–601. DOI: 10.1016/j.enbuild.2004.09.006

Gordillo-Orquera, R., Lopez-Ramos, L., Munoz-Romero, S., Iglesias-Casarrubios, P., Arcos-Aviles, D., Marques, A., and Rojo-Alvarez, J., 2018. Analyzing and forecasting electrical load consumption in healthcare buildings. Energies, 11(493):1–18. DOI: 10.3390/en11030493

Gower, J. C., 1971. A General Coefficient of Similarity and Some of Its Properties. Biometrics, 27(4):857–871. DOI: 10.2307/2528823

Grant, J. L., 2014. Short-term peak demand forecasting using an artificial neural network with controlled peak demand through intelligent electrical loading. PhD Thesis, University of Miami, Coral Gables,

1–139.

Greenwell, B, Boehmke, B., Cunningham, J., GBM Developers, 2019. gbm: Generalised Boosted Regression Models. R package version 2.1.5. URL: https://CRAN.R-project.org/package=gbm

Grolemund, G., and Wickham, H., 2011. Dates and Times Made Easy with lubridate. Journal of Statistical Software, 40(3):1–25. DOI: 10.18637/jss.v040.i03

Guan, C., 2013. Wavelet neural network based very short-term load forecasting and prediction interval estimation. PhD Thesis, University of Connecticut, Storrs, 1–92.

Guillen-Garcia, E., Zorita-Lamadrid, A., Duque-Perez, O., Morales-Velazquez, L., Osornio-Rios, R., and Romero-Troncoso, R., 2017. Power consumption analysis of electrical installations at healthcare facility. Energies, 10(64):1–14. DOI: 10.3390/en10010064

Gulin, M., Vasak, M., Banjac, G., and Tomisa, T., 2014. Load forecast of a university building for application in microgrid power flow optimization. Proceedings of 2014 IEEE International Energy Conference (ENERGYCON), Cavtat, 1223–1227. DOI: 10.1109/ENERGYCON.2014.6850579

Haque, A., Pipattanasomporn, M., Rahman, S., Kothandaraman, S. R., and Malekpour, A., 2019. An SVR-based Building-level Load Forecasting Method Considering Impact of HVAC Set Points. Proceedings of 2019 IEEE Power & Energy Society Innovative Smart Grid Technologies Conference (ISGT), Washington, DC, 1–5. DOI: 10.1109/ISGT.2019.8791649

Hassnain, S. R. U., 2009. Application of artificial neural networks to short-term load forecasting. PhD Thesis, Department of Electrical and Electronics Engineering, University of Engineering and Technology Peshawar, Peshawar, 1–353.

He, Y., and Zheng, Y., 2018. Short-term power load probability density forecasting based on yeo-johnson transformation quantile regression and gaussian kernel function. Energy, 154:143–156. DOI: 10.1016/j.energy.2018.04.072

Heylman, C., Kim, Y. G., and Wang, J., 2015. Forecasting energy trends and peak usage at the University of Virginia. Proceedings of 2015 Systems and Information Engineering Design Symposium, Charlotsville, VA, 362–368. DOI: 10.1109/SIEDS.2015.7117006

Hong, T., 2010. Short-term electric load forecasting. PhD Thesis, Department of Operations Research and Electrical Engineering, North Carolina State University, Raleigh, 1–157.

_____, and Fan, S., 2016. Probabilistic electric load forecasting: A tutorial review. International Journal of Forecasting, 32:914–938. DOI: 10.1016/j.ijforecast.2015.11.011

_____, Gui, M., Baran, M. E., and Willis, H. L., 2010. Modeling and forecasting hourly electric load by multiple linear regression with interactions. Proceedings of IEEE PES General Meeting, Providence, 1–8. DOI: 10.1109/PES.2010.5589959

Hossain, M. S., Ong, Z. C., Ismail, Z., and Khoo, S. Y., 2017. A comparative study of vibrational response based impact force localization and quantification using radial basis function network and multilayer perceptron. Expert Systems with Applications,

85:87–98. DOI: 10.1016/j.eswa.2017.05.027

Hosseini, S. S. S., and Gandomi, A. H., 2012. Short-term load forecasting of power systems by gene expression programming. Neural Computing and Applications, 21(2):377–389. DOI: 10.1007/s00521-010-0444-y

Hughes, R. A., Heron, J., Sterne, J. A. C., and Tilling, K., 2019. Accounting for missing data in statistical analyses: multiple imputation is not always the answer. International Journal of Epidemiology, dyz032:1–11. DOI: 10.1093/ije/dyz032

Hyndman, R. J., Athanasopoulos, G., Bergmeir, C., Caceres, G., Chhay, L., O'Hara-Wild, M., Petropoulos, F., Razbash, S., Wang, E., and Yasmeen, F., 2018. forecast: Forecasting functions for time series and linear models. R package version 8.4. URL: https://cran.r-project.org/package=forecast

_____, and Khandakar, Y., 2008. Automatic time series forecasting: the forecast package for R. Journal of Statistical Software, 26(3):1–22. DOI: 10.18637/jss.v027.i03

Jain, B. K., 2018. Short term load forecasting for smart power systems. PhD Thesis, International Institute of Information Technology, Hyderabad, 1–173.

Jetcheva, J. G., Majidpour, M., and Chen, W. P., 2014. Neural network model ensembles for building-level electricity load forecasts. Energy and Buildings, 84:214–223. DOI: 10.1016/j.enbuild.2014.08.004

Jing, Z., Cai, M., Pipattanasomporn, M., Rahman, S., Kothandaraman, S. R., and Malekpour, A., 2019. Commercial building load forecasts with artificial neural network. Proceedings of 2019 IEEE Power &

Energy Society Innovative Smart Grid Technologies Conference (ISGT), Washington, DC, 1–5. DOI: 10.1109/ISGT.2019.8791654

Johannesson, T., Bjornsson, H., and Grothendieck, G., 2018. stinepack: Stineman, a Consistently Well Behaved Method of Interpolation. R package version 1.4. URL: https://cran.r-project.org/package=stinepack

Kaboli, S. H. A., Fallahpour, A., Selvaraj, J., and Rahim, N., 2017. Long-term electrical energy consumption formulating and forecasting via optimized gene expression programming. Energy, 126:144–164. DOI: 10.1016/j.energy.2017.03.009

Kalogirou, S. A., and Bojic, M., 2000. Artificial neural networks for the prediction of the energy consumption of a passive solar building. Energy, 25(5):479–491. DOI: 10.1016/S0360-5442(99)00086-9

Karatasou, S., Santamouris, M., and Geros, V., 2006. Modeling and predicting building's energy use with artificial neural networks: Methods and results. Energy and Buildings, 38:949–958. DOI: 10.1016/j.enbuild.2005.11.005

Kassam, A. A., Lee, B. D., and Paredis, C. J. J., 2014. Statistical methods for interpolating meteorological data for use in building simulation. Building Simulation, 7(5):455–465. DOI: 10.1007/s12273-014-0174-7

Ke, X., Jiang, A., and Lu, N., 2016. Load profile analysis and short-term building load forecast for a university campus. Proceedings of 2016 IEEE Power and Energy Society General Meeting (PESGM), Boston, MA, 1–5. DOI: 10.1109/PESGM.2016.7742034

Khosravani, H. R., Castilla, M. D. M., Berenguel, M., Ruano, A. E., and

Ferreira, P. M., 2016. A comparison of energy consumption prediction models based on neural networks of a bioclimatic building. Energies, 9(57):1–24. DOI: 10.3390/en9010057

Kowarik, A., and Templ, M., 2016. Imputation with the R package VIM. Journal of Statistical Software, 74(7):1–16. DOI: 10.18637/jss.v074.i07

Krarti, M., 2003. An overview of artificial intelligence-based methods for building energy systems. Journal of Solar Energy Engineering, 125(3):331–342. DOI: 10.1115/1.1592186

Kreider, J. F., Claridge, D. E., Curtiss, P., Dodier, R., Haberl, J. S., and Krarti, M., 1995. Building energy use prediction and system identification using recurrent neural networks. Journal of Solar Energy Engineering, 117(3):161–166. DOI: 10.1115/1.2847757

Lazos, D., Sproul, A. B., and Kay, M., 2014. Optimisation of energy management in commercial buildings with weather forecasting inputs: A review. Renewable and Sustainable Energy Reviews, 39:587–603. DOI: 10.1016/j.rser.2014.07.053

Li, C., Ding, Z., Yi, J., Lv, Y., and Zhang, G., 2018. Deep belief network based hybrid model for building energy consumption prediction. Energies, 11(242):1–26. DOI: 10.3390/en11010242

---------, Ding, Z., Zhao, D., Yi, J., and Zhang, G., 2017. Building energy consumption prediction: An extreme deep learning approach. Energies, 10(1525):1–20. DOI: 10.3390/en10101525

Li, H. Z., Guo, S., Li, C. J., and Sun, L. Q., 2013. A hybrid annual power load forecasting model based on generalized regression neural network with fruit fly optimization algorithm. Knowledge-Based Systems,

37:378–387. DOI: 10.1016/j.knosys.2012.08.015

Li, K., Hu, C., Liu, G., and Xue, W., 2015. Building's electricity consumption prediction using optimized artificial neural networks and principal component analysis. Energy and Buildings, 108:106–113. DOI: 10.1016/j.enbuild.2015.09.002

_____, Su, H., and Chu, J., 2011. Forecasting building energy consumption using neural networks and hybrid neuro-fuzzy system: A comparative study. Energy and Buildings, 43:2893–2899. DOI: 10.1016/j.enbuild.2011.07.010

Li, M., 2015. Real-time power flow analysis & short-term electricity load forecasting in smart grid. PhD Thesis, Department of Applied Mathematics and Statistics, Stony Brook University, New York, 1–88.

Li, X., Zhou, C, Xiao, W., and Nelson, P. C., 2005. Prefix Gene Expression Programming. Proceedings of Genetic and Evolutionary Computation Conference (GECCO), Washington. URL: https://pdfs.semanticscholar.org/dc20/94393f7044f12dc5c33ec3...

Liang, Y., Niu, D., and Hong, W. C., 2019. Short term load forecasting based on feature extraction and improved general regression neural network model. Energy, 166:653–663. DOI: 10.1016/j.energy.2018.10.119

Little, R. J. A., and Rubin, D. B., 2002. Statistical Analysis with Missing Data, 2nd Ed. John Wiley & Sons.

Liu, B., 2016. Short-term load forecasting with recency effect: models and applications. PhD Thesis, Department of Infrastructure and Environmental Systems, University of North Carolina, Charlotte,

1–135.

Liu, D., Chen, Q., and Mori, K., 2015. Time series forecasting method of building energy consumption using support vector regression. Proceedings of 2015 IEEE International Conference on Information and Automation, Lijiang, 1628–1632. DOI: 10.1109/ICInfA.2015.7279546

Liu, Y., 2018. gepR: Composite linear-nonlinear regression using GEP. URL: https://github.com/profyliu/gepR

Liu, Y., Wang, W., and Ghadimi, N., 2017. Electricity load forecasting by an improved forecast engine for building level consumers. Energy, 139:18–30. DOI : 10.1016/j.energy.2017.07.150

Liu, T., Xu, C., Guo, Y., and Chen, H., In Press. A novel deep reinforcement learning based methodology for short-term HVAC system energy consumption prediction. International Journal of Refrigeration. DOI : 10.1016/j.ijrefrig.2019.07.018

Madley-Dowd, P., Hughes, R., Tilling, K., and Heron, J., 2019. The proportion of missing data should not be used to guide decisions on multiple imputation. Journal of Clinical Epidemiology, 110:63–73. DOI: 10.1016/j.jclinepi.2019.02.016

Mai, W., Chung, C. Y., Wu, T., and Huang, H., 2014. Electric load forecasting for large office building based on radial basis function neural network. Proceedings of 2014 IEEE PES General Meeting — Conference & Exposition, National Harbor, MD, 1–5. DOI: 10.1109/PESGM.2014.6939378

Manjhi, Y., and Dhar, J., 2016. Forecasting energy consumption using particle swarm optimization and gravitational search algorithm.

Proceedings of 2016 International Conference on Advanced Communication Control and Computing Technologies (ICACCCT), Ramanathapuram, 417–420. DOI: 10.1109/ICACCCT.2016.7831673

Massana, J., Pous, C., Burgas, L., Melendez, J., and Colomer, J., 2015. Short-term load forecasting in a non-residential building contrasting models and attributes. Energy and Buildings, 92:322–330. DOI: 10.1016/j.enbuild.2015.02.007

_____, Pous, C., Burgas, L., Melendez, J., and Colomer, J., 2016. Short-term load forecasting for non-residential buildings contrasting artificial occupancy attributes. Energy and Buildings, 130:519–531. DOI: 10.1016/j.enbuild.2016.08.081

Mathieu, J. L., Price, P. N., Kiliccote, S., and Pierce, M. A., 2011. Quantifying changes in building electricity use, with application to demand response. IEEE Transactions on Smart Grid, 2(3):507–518. DOI: 10.1109/TSG.2011.2145010

Matijas, M., 2013. Electric load forecasting using multivariate meta-learning. PhD Thesis, Faculty of Electrical Engineering and Computing, University of Zagreb, Zagreb, 1–143.

Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., and Leisch, F., 2019. e1071: Misc Functions of the Department of Statistics, Probability Group (Formerly: E1071), TU Wien. R package version 1.7-2. URL: https://CRAN.R-project.org/package=e1071

Mocanu, E., Nguyen, P. H., Gibescu, M., and Kling, W. L., 2014. Comparison of machine learning methods for estimating energy consumption in buildings. Proceedings of 2014 International

Conference on Probabilistic Methods Applied to Power Systems (PMAPS), Durham, 1–6. DOI: 10.1109/PMAPS.2014.6960635

Molina-Solana, M., Ros, M., Ruiz, M. D., Gomez-Romero, J., and Martin-Bautista, M. J., 2017. Data science for building energy management: A review. Renewable and Sustainable Energy Reviews, 70:598–609. DOI: 10.1016/j.rser.2016.11.132

Morinigo-Sotelo, D., Duque-Perez, O., Garcia-Escudero, L. A., Fernandez-Temprano, M., Fraile-Llorente, P., Riesco-Sanz, M. V., and Zorita-Lamadrid, A.L., 2011. Short-term hourly load forecasting of a hospital using an artificial neural network. Proceedings of International Conference on Renewable Energies and Power Quality, Las Palmas de Gran Canaria, 441–446. DOI: 10.24084/repqj09.355

Moritz, S., and Bartz-Beielstein, T., 2017. imputeTS: Time Series Missing Value Imputation in R. The R Journal, 9(1):207–218. DOI: 10.32614/RJ-2017-009

Munguia, J. A. T., 2014. Comparison of Imputation Methods for Handling Missing Categorical Data with Univariate Pattern. Journal of Quantitative Methods for Economics and Business Administration, 17:101–120. URL: http://www.upo.es/RevMetCuant/art.php?id=91

Nagy, G., 1991. Neural Networks - Then and Now. IEEE Transactions on Neural Networks, 2(2):316–318. DOI: 10.1109/72.80343

Naug, A., and Biswas, G., 2018. A Data Driven Method for Prediction of Energy Demand in Commercial Buildings. Proceedings of 2018 IEEE 14th International Conference on Automation Science and Engineering (CASE), Munich, 335–340. DOI:

10.1109/COASE.2018.8560520

Neto, A. H., and Fiorelli, F. A. S., 2008. Comparison between detailed model simulation and artificial neural network for forecasting building energy consumption. Energy and Buildings, 40:2169–2176. DOI: 10.1016/j.enbuild.2008.06.013

Nichiforov, C., Stamatescu, G., and Stamatescu, I., 2018. Deep learning techniques for load forecasting in large commercial buildings. Proceedings of the 22nd International Conference on System Theory, Control and Computing (ICSTCC), Sinaia, 492–497. DOI: 10.1109/ICSTCC.2018.8540768

Onwubolu, G., 2015. GMDH-Methodology and Implementation in C. Imperial College Press, Singapore, 1–283. ISBN: 978-1-84816-610-3

Parr, T., and Howard, J., 2019. How to explain gradient boosting. Illustration of GBDT by golf example. URL: https://explained.ai/gradient-boosting/images/golf-MSE.png

Paudel, S., Nguyen, P. H., Kling, W. L., Elmitri, M., Lacarriere, B., and Le Corre, O., 2015. Support vector machine in prediction of building energy demand using pseudo dynamic approach. Proceedings of the 28th International Conference on Efficiency, Cost, Optimization, Simulation and Environmental Impact of Energy Systems (ECOS 2015), Pau, France. URL: https://arxiv.org/abs/1507.05019

Pedersen, L., Stang, J., and Ulseth, R., 2008. Load prediction method for heat and electricity demand in buildings for the purpose of planning for mixed energy distribution systems. Energy and Buildings, 40:1124–1134. DOI: 10.1016/j.enbuild.2007.10.014

Persson, C., Bacher, P., Shiga, T., and Madsen, H., 2017. Multi-site solar power forecasting using gradient boosted regression trees. Solar Energy, 150:423–436. DOI: 10.1016/j.solener.2017.04.066

Pino-Mejias, R., Perez-Fargallo, A., Rubio-Bellido, C., and Pulido-Arcas, J.A., 2017. Comparison of linear regression and artificial neural networks models to predict heating and cooling energy demand, energy consumption and CO_2 emissions. Energy, 118:24–36. DOI: 10.1016/j.energy.2016.12.022

Platon, R., Dehkordi, V. R., and Martel, J., 2015. Hourly prediction of a building's electricity consumption using case-based reasoning, artificial neural networks and principal component analysis. Energy and Buildings, 92:10–18. DOI: 10.1016/j.enbuild.2015.01.047

Pombeiro, H., Santos, R., Carreira, P., Silva, C., and Sousa, J., 2017. Comparative assessment of low-complexity models to predict electricity consumption in an institutional building: linear regression vs. fuzzy modeling vs. neural networks. Energy and Buildings, 146:141–151. DOI: 10.1016/j.enbuild.2017.04.032

Poulos, J., and Valle, R., 2018. Missing data imputation for supervised learning. Applied Artificial Intelligence, 32(2):186–196. DOI: 10.1080/08839514.2018.1448143

R Core Team, 2018. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. Vienna, Austria. URL: https://www.R-project.org/

Raza, M. Q., and Khosravi, A., 2015. A review on artificial intelligence based load demand forecasting techniques for smart grid and buildings. Renewable and Sustainable Energy Reviews,

50:1352–1372. DOI: 10.1016/j.rser.2015.04.065

Roldan-Blay, C., Escriva-Escriva, G., Alvarez-Bel, C., Roldan-Porta, C., and Rodriguez-Garcia, J., 2013. Upgrade of an artificial neural network prediction method for electrical consumption forecasting using an hourly temperature curve model. Energy and Buildings, 60:38–46. DOI: 10.1016/j.enbuild.2012.12.009

Ruiz, L. G. B., Cuellar, M. P., Calvo-Flores, M. D., and Jimenez, M. D. C. P., 2016. An application of non-linear autoregressive neural networks to predict energy consumption in public buildings. Energies, 9(684):1–21. DOI: 10.3390/en9090684

Salkind, N. J., 2010. Encyclopedia of Research Design. SAGE, Thousand Oaks, California, 1287–1288.

Sala-Cardoso, E., Delgado-Prieto, M., Kampouropoulos, K., and Romeral, L., 2018. Activity-aware HVAC power demand forecasting. Energy and Buildings, 170:15–24. DOI: 10.1016/j.enbuild.2018.03.087

Sarduy, J. R. G., Santo, K. G. D., and Saidel, M. A., 2016. Linear and non-linear methods for prediction of peak load at university of Sao Paulo. Measurement, 78:187–201. DOI: 10.1016/j.measurement.2015.09.053

Sayad, S., 2019. Support Vector Machine - Regression (SVR). Figure of Non-linear SVR. URL: https://www.saedsayad.com/images/$SVR_5.png$

Schafer, J. L., 1999. Multiple imputation: a primer. Statistical Methods in Medical Research, 8:3–15. DOI: 10.1177/096228029900800102

Schmitt, P., Mandel, J., and Guedj, M., 2015. A Comparison of Six Methods for Missing Data Imputation. Journal of Biometrics & Biostatistics,

6(1):1–6. DOI: 10.472/2155-6180.1000224

Seem, J. E., and Braun, J. E., 1991. Adaptive methods for real-time forecasting of building electrical demand. ASHRAE Transactions, 1991 Winter Meeting, New York, 97(1):710–721.

Seyedzadeh, S., Rahimian, F. P., Glesk, I., and Roper, M., 2018. Machine learning for estimation of building energy consumption and performance: a review. Visualization in Engineering, 6(5):1–20. DOI: 10.1186/s40327-018-0064-7

Shan, S., Cao, B., and Wu, Z., 2019. Forecasting the short-term electricity consumption of building using a novel ensemble model. IEEE Access, 7:88093–88106. DOI: 10.1109/ACCESS.2019.2925740

Soda-Pro, 2019. Correction of the temperature with the altitude. URL: http://www.soda-pro.com/web-services/meteo-data/merra/info

Stepashko, V., Bulgakova, O., and Zosimov, V., 2017. Construction and Research of the Generalized Iterative GMDH Algorithm with Active Neurons. Advances in Intelligent Systems and Computing, 689:492–510. DOI: 10.1007/978-3-319-70581-$1_3$5

Stineman, R. W., 1980. A Consistently Well-Behaved Method of Interpolation. Creative Computing, 6(7):54–57.

Tardioli, G., Kerrigan, R., Oates, M., O'Donnell, J., and Finn, D., 2015. Data driven approaches for prediction of building energy consumption at urban level. Energy Procedia, 78:3378–3383. DOI: 10.1016/j.egypro.2015.11.754

Teke, A., and Timur, O., 2014. Assessing the energy efficiency improvement potentials of HVAC systems considering economic and environmental aspects at the hospitals. Renewable and

Sustainable Energy Reviews, 33:224–235. DOI: 10.1016/j.rser.2014.02.002

_____, Zor, K., and Timur, O., 2015. A simple methodology for capacity sizing of cogeneration and trigeneration plants in hospitals: A case study for a university hospital. Journal of Renewable and Sustainable Energy, 7(053102):1–15. DOI: 10.1063/1.4930064

Teksan, A. E., Kocar, G., Eryasar, A., and Aytav, E., 2017. Investigation of benefits of co-generation/tri-generation applications in hospitals through a case study. Proceedings of 5th National Electrical Installation Congress and Exhibition, İzmir. URL: http://www.emo.org.tr/ekler/661b2c3bc72f4b5_ek.pdf

Tierney, N., Cook, D., McBain, M., and Fay, C., 2018. naniar: Data Structures, Summaries, and Visualisations for Missing Data. R package version 0.4.1. URL: https://CRAN.R-project.org/package=naniar

Timur, O., 2018. Design and Implementation of a Wireless Sensor Network for Energy Monitoring, Analysis and Management in Smart Buildings. PhD Thesis, Department of Electrical and Electronics Engineering, Çukurova University, Institute of Natural and Applied Sciences, Adana, 1–215.

_____, Zor, K., Çelik, Ö., Teke, A., and İbrikçi, T., In Press. Application of Statistical and Artificial Intelligence Techniques for Medium-Term Electrical Energy Forecasting: A Case Study for a Regional Hospital. Journal of Sustainable Development of Energy, Water and Environment Systems, 1–17. DOI: Unassigned

Topallı, A. K., 2003. Hybrid learning algorithm for intelligent electric

short-term load forecasting. PhD Thesis, Department of Electrical and Electronics Engineering, Middle East Technical University, Graduate School of Natural and Applied Sciences, Ankara, 1–104.

Torres, F. J. G., 2017. Adaptive load consumption modelling on the user side: Contributions to load forecasting modelling based on supervised mixture of experts and genetic programming. PhD Thesis, Universitat Politecnica de Catalunya, Barcelona, 1–221.

Touretzky, C. R., and Patil, R., 2015. Building-level power demand forecasting framework using building specific inputs: Development and applications. Applied Energy, 147:466–477. DOI: 10.1016/j.apenergy.2015.03.025

Touzani, S., Granderson, J., and Fernandes, S., 2018. Gradient boosting machine for modeling the energy consumption of commercial buildings. Energy and Buildings, 158:1533–1543. DOI: 10.1016/j.enbuild.2017.11.039

Tuunanen, J., 2015. Modelling of changes in electricity end-use and their impacts on electricity distribution. PhD Thesis, Lappeenranta University, Lappeenranta, 1–193.

Twanabasu, S. R., and Bremdal, B. A., 2013. Load forecasting in a Smart Grid oriented building. Proceedings of the 22nd International Conference and Exhibition on Electricity Distribution (CIRED 2013), Stockholm, 1–4. DOI: 10.1049/cp.2013.0997

Xiao, J., Li, Y., Xie, L., Liu, D., and Huang, J., 2018. A hybrid model based on selective ensemble for energy consumption forecasting in China. Energy, 159:534–546. DOI: 10.1016/j.energy.2018.06.161

Xie, Y., Li, C., Lv, Y., and Yu, C., 2019. Predicting lightning outages of

transmission lines using generalized regression neural network. Applied Soft Computing Journal, 78:438–446. DOI: 10.1016/j.asoc.2018.09.042

Xypolytou, E., Meisel, M., and Sauter, T., 2017. Short-term electricity consumption forecast with artificial neural networks - a case study of buildings. Proceedings of 2017 IEEE Manchester PowerTech, Manchester, 1–6. DOI: 10.1109/PTC.2017.7980874

Van Buuren, S., and Groothuis-Oudshoorn, K., 2011. mice: Multivariate Imputation by Chained Equations in R. Journal of Statistical Software, 45(3):1–67. DOI: 10.18637/jss.v045.i03

Villiers, J., 2012. Mathematics of approximation. Atlantis Press Paris.

Vinagre, E., Gomes, L., and Vale, Z., 2015. Electrical energy consumption forecast using external facility data. Proceedings of 2015 IEEE Symposium Series on Computational Intelligence, Cape Town, 659–664. DOI: 10.1109/SSCI.2015.101

Vrablecova, P., Ezzeddine, A. B., Rozinajova, V., Sarik, S., and Sangaiah, A. K., 2018. Smart grid load forecasting using online support vector regression. Computers & Electrical Engineering, 65:102–117. DOI: 10.1016/j.compeleceng.2017.07.006

Wang, Z., and Srinivasan, R. S., 2017. A review of artificial intelligence based building energy use prediction: Contrasting the capabilities of single and ensemble prediction models. Renewable and Sustainable Energy Reviews, 75:796–808. DOI: 10.1016/j.rser.2016.10.079

_____, Wang, Y., and Srinivasan, R. S., 2018a. A novel ensemble learning approach to support building energy use prediction. Energy and Buildings, 159:109–122. DOI: 10.1016/j.enbuild.2017.10.085

---------, Wang, Y., Zeng, R., Srinivasan, R. S., and Ahrentzen, S., 2018b. Random forest based hourly building energy prediction. Energy and Buildings, 171:11–25. DOI: 10.1016/j.enbuild.2018.04.008

Wei, Y., Zhang, X., Shi, Y., Xia, L., Pan, S., Wu, J., Han, M., and Zhao, X., 2018. A review of data-driven approaches for prediction and classification of building energy consumption. Renewable and Sustainable Energy Reviews, 82:1027–1047. DOI: 10.1016/j.rser.2017.09.108

Wickham, H., 2016. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 1–267. ISBN: 978-3-319-24277-4. URL: https://ggplot2.tidyverse.org

---------, 2017. tidyverse: Easily Install and Load the ”Tidyverse”. R package version 1.2.1. URL: https://CRAN.R-project.org/package=tidyverse

---------, 2018. scales: Scale Functions for Visualisation. R package version 1.0.0. URL: https://CRAN.R-project.org/package=scales

---------, François, R., Henry, L., and Müller, K., 2018. dplyr: A Grammar of Data Manipulation. R package version 0.7.7. URL: https://CRAN.R-project.org/package=dplyr

---------, and Henry, L., 2018. tidyr: Easily Tidy Data with ”spread()” and ”gather()” Functions. R package version 0.8.2. URL: https://CRAN.R-project.org/package=tidyr

Yalcintas, M., and Akkurt, S., 2005. Artificial neural networks applications in building energy predictions and a case study for tropical climates. International Journal of Energy Research, 29:891–901. DOI: 10.1002/er.1105

Yang, J., 2006. Power system short-term load forecasting. PhD Thesis, Department of Electrical and Computer Engineering, Darmstadt Technical University, Darmstadt, 1–133.

Yang, J., Rivard, H., and Zmeureanu, R., 2005. On-line building energy prediction using adaptive artificial neural networks. Energy and Buildings, 37:1250–1259. DOI: 10.1016/j.enbuild.2005.02.005

Yaslan, Y., and Bican, B., 2017. Empirical mode decomposition based denoising method with support vector regression for time series prediction. Measurement, 103:52–61. DOI: 10.1016/j.measurement.2017.02.007

Yıldırım, H. B., Çelik, Ö., Teke, A., and Barutçu, B., 2018. Estimating daily global solar radiation with graphical user interface in eastern Mediterranean region of Turkey. Renewable and Sustainable Energy Reviews, 82:1528–1537. DOI: 10.1016/j.rser.2017.06.030

Yildiz, B., Bilbao, J. I., and Sproul, A. B., 2017. A review and analysis of regression and machine learning models on commercial building electricity load forecasting. Renewable and Sustainable Energy Reviews, 73:1104–1122. DOI: 10.1016/j.rser.2017.02.023

Yoo, J., and Hur, K., 2013. Load forecast model switching scheme for improved robustness to changes in building energy consumption patterns. Energies, 6:1329–1343. DOI: 10.3390/en6031329

Yu, H., Xie, T., Paszczynski, S., and Wilamowski, B. M., 2011. Advantages of radial basis function networks for dynamic system design. IEEE Transactions on Industrial Electronics, 58:5438–5450. DOI: 10.1109/TIE.2011.2164773

Zendehboudi, A., Baseer, M., and Saidur, R., 2018. Application of support

vector machine models for forecasting solar and wind energy resources: A review. Journal of Cleaner Production, 199:272–285. DOI: 10.1016/j.jclepro.2018.07.164

Zhang, F., Deb, C., Lee, S. E., Yang, J., and Shah, K. W., 2016. Time series forecasting for building energy consumption using weighted support vector regression with differential evolution optimization technique. Energy and Buildings, 126:94–103. DOI: 10.1016/j.enbuild.2016.05.028

Zhao, H. X., 2011. Artificial intelligence models for large-scale buildings energy consumption analysis. PhD Thesis, Ecole Centrale Paris, Paris, 1–125.

_____, and Magoules, F., 2010. Parallel support vector machines applied to the prediction of multiple buildings energy consumption. Journal of Algorithms & Computational Technology, 4(2):231–249. DOI: 10.1260/1748-3018.4.2.231

_____, and Magoules, F., 2011. New parallel support vector regression for predicting building energy consumption. Proceedings of 2011 IEEE Symposium on Computational Intelligence in Multicriteria Decision-Making (MDCM), Paris, 14–21. DOI: 10.1109/SMDCM.2011.5949289

_____, and Magoules, F., 2012a. A review on the prediction of building energy consumption. Renewable and Sustainable Energy Reviews, 16:3586–3592. DOI: 10.1016/j.rser.2012.02.049

_____, and Magoules, F., 2012b. Feature selection for predicting building energy consumption based on statistical learning method. Journal of Algorithms & Computational Technology, 6(1):59–77. DOI:

10.1260/1748-3018.6.1.59

Zhong, J., Feng, L., and Ong, Y. -S., 2017. Gene Expression Programming: A Survey. IEEE Computational Intelligence Magazine, 12(3):54–72. DOI: 10.1109/MCI.2017.2708618

Zor, K., 2015. Developing a Software Program to Determine the Optimal Capacity Rating of Cogeneration and Trigeneration Plants Driven by Gas Engines for Unlicensed Generation of Electricity. MSc Thesis, Department of Electrical and Electronics Engineering, Çukurova University, Institute of Natural and Applied Sciences, Adana, 1–87.

_____, Timur, O., Çelik, Ö., Yıldırım, H. B., and Teke, A., 2017a. Interpretation of error calculation methods in the context of energy forecasting. Proceedings of 12th Conference on Sustainable Development of Energy, Water and Environment Systems (SDEWES), Dubrovnik, 0722:1–9.

_____, Timur, O., and Teke, A., 2017b. A state-of-the-art review of artificial intelligence techniques for short-term electric load forecasting. Proceedings of the 6th International Youth Conference on Energy (IYCE2017), Budapest, 1–7. DOI: 10.1109/IYCE.2017.8003734

_____, Çelik, Ö., Timur, O., Yıldırım, H. B., and Teke, A., 2018a. Simple approaches to missing data for energy forecasting applications. Proceedings of the 16th International Conference on Clean Energy (ICCE-2018), Gazimağusa, FORC-03:1–4.

_____, Timur, O., Çelik, Ö., Yıldırım, H. B., and Teke, A., 2018b. Very Short-Term Electrical Energy Consumption Forecasting of a

Household for the Integration of Smart Grids. Official Conference Proceedings of the European Conference on Sustainability, Energy & the Environment 2018 (ECSEE2018), Brighton, 1–14.

BIOGRAPHY

Kasım Zor was born in Adana in 1987. He graduated from Seyhan ÇEAŞ Anatolian High School in 2004. In 2007, he completed Erasmus Exchange Programme in Linköping Institute of Technology in Sweden, and received his BSc and MSc degrees in Electrical and Electronics Engineering from Çukurova University in 2010 and 2015 respectively.

After participating in Konya Air Defence School Command in 2010, he worked for Military Engineering Branch of Turkish Land Forces 4th Mechanised Infantry Brigade as a third and second lieutenant, Energy Division of Sanko Textile as an electrical maintenance engineer, MTU Onsite Energy of Rolls-Royce Power Systems as a service and commissioning engineer, and Electrical and Electronics Engineering Department of Çukurova University as a research assistant. He has been working as a research assistant at Electrical and Electronic Engineering Department of Adana Alparslan Türkeş Science and Technology University since February 2013.

His research interests include energy analytics, energy forecasting, distributed generation, energy economics, and energy efficiency. He is a member of UCTEA Chamber of Electrical Engineers. He is married and father of a daughter.