Summer 2018

Machine learning techniques implementation in power optimization, data processing, and bio-medical applications

Khalid Khairullah Mezied Al-Jabery

Department: Electrical and Computer Engineering

MACHINE LEARNING TECHNIQUES IMPLEMENTATION IN POWER OPTIMIZATION, DATA PROCESSING, AND BIO-MEDICAL APPLICATIONS

by

KHALID KHAIRULLAH MEZIED AL-JABERY

A DISSERTATION

Presented to the Faculty of the Graduate School of the

MISSOURI UNIVERSITY OF SCIENCE AND TECHNOLOGY

In Partial Fulfillment of the Requirements for the Degree

DOCTOR OF PHILOSOPHY

in

COMPUTER ENGINEERING

2018

Approved by

Dr. Donald C. Wunsch II, Advisor

Dr. Daryl Beetner

Dr. Minsu Choi

Dr. Ian Ferguson

Dr. Abhijit Gosavi

PUBLICATION DISSERTATION OPTION

This dissertation consists of the following four articles which have been published or submitted for publication as follows:

Paper I: Pages 8-46 have been published in the Journal of IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems, Vol. 36, No. 5, pp. 775-788.

Paper II: Pages 47-62 have been published in the 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), 2016.

Paper III: Pages 63-86 have been published in the IEEE Symposium Series on Computational Intelligence (SSCI), 2017.

Paper IV: Pages 87-126 have been submitted to the Journal of Environment International.

ABSTRACT

The rapid progress and development in machine-learning algorithms becomes a key factor in determining the future of humanity. These algorithms and techniques were utilized to solve a wide spectrum of problems extended from data mining and knowledge discovery to unsupervised learning and optimization. This dissertation consists of two study areas. The first area investigates the use of reinforcement learning and adaptive critic design algorithms in the field of power grid control. The second area in this dissertation, consisting of three papers, focuses on developing and applying clustering algorithms on biomedical data. The first paper presents a novel modelling approach for demand side management of electric water heaters using Q-learning and action-dependent heuristic dynamic programming. The implemented approaches provide an efficient load management mechanism that reduces the overall power cost and smooths grid load profile. The second paper implements an ensemble statistical and subspace-clustering model for analyzing the heterogeneous data of the autism spectrum disorder. The paper implements a novel *k*-dimensional algorithm that shows efficiency in handling heterogeneous dataset. The third paper provides a unified learning model for clustering neuroimaging data to identify the potential risk factors for suboptimal brain aging. In the last paper, clustering and clustering validation indices are utilized to identify the groups of compounds that are responsible for plant uptake and contaminant transportation from roots to plants edible parts.

ACKNOWLEDGMENTS

My first gratitude and thankfulness are for my advisor Dr. Donald C. Wunsch II. He was surrounding me with his kind support and guidance during my study. All his advices and recommendations helped to improve the overall quality of my work. I cannot find the proper words to describe his constructive role.

I would also like to thank all my committee members, Drs. Daryl Beetner, Minsu Choi, Ian Ferguson, and Abhijit Gosavi for their valuable suggestions and guidance. I really appreciate that they have devoted their valuable time to improve this work.

Special thanks and gratitude to Drs. Yiyu Shi, Tayo Obafemi-Ajayi, and Gayla Olbricht for their ultimate support and guidance.

My gratitude to my wife Meyyada, my mother Fatimeh for their help, support, and patience on me.

I would like to thank my collaborators and coauthors who were supportive, professional, and I learned a lot from them particularly Dr. Burken, Dr. Xiong, Dr. Wenjian, and Dr. Jhi-young Joo.

My unlimited thanks to my friends who supported and encouraged me.

This project was funded by the Higher Committee for Education Development in Iraq (HCED), Missouri University of Science and Technology, and Mary K. Finley foundation.

TABLE OF CONTENTS

Page

LIST OF ILLUSTRATIONS

LIST OF TABLES

1. INTRODUCTION

1.1. BACKGROUND

The continuous growth in population all around the globe adds new challenges to multiple aspects of life, including the increase in demand on power and better healthcare. These challenges and others require nontraditional solutions that are robust and reliable. Machine-learning algorithms provide the required resources for such solutions due to the revolutionized and adaptive nature of most of these algorithms. Machine learning has become the leading approach in scientific discoveries and innovations that has moved civilization forward in all fields. Machine-learning applications are the reason behind many inventions that are available to us in daily life such as human-machine vocal communication (i.e., speech understanding), machine vision, automatic navigation, route planning, customer segmentation, credit card fraud detection, data mining, internet routing, face recognition, digitization, and medical diagnoses.

Energy management is one of the most challenging problems for power companies. Demand side management is essential in smoothing the grid load profile, which leads to stable and lower cost. Load management techniques have focused on controlling domestic electric water heaters since they have the ability of storing energy. Surveys and official reports show that the load resulting from electric water heaters represents 18% and 23% of the total grid load in the domestic sector in the United States and India, respectively. The peaks in the total grid load profile match those of the electric water heaters' load profile. The peak load management is vital in controlling the total grid demand. Therefore, several approaches have been applied to control domestic electric water heaters. Some studies

shifted the operation of the electric water heaters unconditionally outside load peak periods. Others prioritized their operation after the peak load periods over one or two hours.

The described problem is a multi-objective problem. Solving such problems requires adaptive and robust algorithms. Reinforcement learning (RL) approaches are efficient in solving such problems. Approximate dynamic programming and simulation-based approaches have shown impressive performance in solving high-dimensional problems within a feasible amount of time.

In addition to the energy management in metropolitan cities, healthcare has become a challenging and attractive field for researchers. In recent years, many studies invested machine learning in healthcare industry and biomedical applications such as early disease detection, monitoring devices, and tracking devices. Healthcare and pharmaceutical companies invests billions of dollars annually to improve the quality of their products. The majority of these investments go for funding advance research centers and developing innovative technologies in the field. Due to the advanced tracking and imaging devices, large datasets are generated to record various types of information ranging from clinical test results to routine questionnaires. The biomedical datasets are heterogeneous in general and need advanced preprocessing. The heterogeneity of clinical datasets exists in patient characteristics, illness severity, and treatment responses. Therefore, clustering becomes the most convenient method for understanding and analyzing the heterogeneous clinical data. Clustering is effective in exploring such complex data and identifying important subsets. In general, machine learning aim to discover unknown patterns or relationships that infer new knowledge that can be further used for prevention, prognosis, and treatment.

1.2. REINFORCEMENT LEARNING

Reinforcement learning is a simulation-based dynamic programming approach that can solve complex sequential decision-making problems such as Markov and semi-Markov decision process. Markov decision process (MDP) models are widely used for modeling sequential decision-making problems that appeared in various fields of science and research. However, many real-world problems modeled by MDPs have large state and/or action spaces, leading to the well-known curse of dimensionality and the curse of modeling, which make solving these models infeasible using dynamic programming. Dynamic programming finds the optimal solution for Markov decision problems, but it needs the transition rewards, transition probabilities, and transition times. These values are almost impossible to determine in most cases. Therefore, machine learning methods such as reinforcement learning are used to determine a suboptimal solution for MDPs. The trade-off between finding a suboptimal solution and the optimal solution is necessary when looking into the massive resources and amount of time required to solve high-dimensional problems using standard methods.

1.3. CLUSTERING

Clustering, also known as unsupervised learning, is a field of machine learning that explores and reveals hidden structures in datasets. Recently, clustering has become more important than ever due to the unprecedented increase in data from various disciplines. Therefore, domain experts are no longer able to analyze such massive datasets and the need for automated approaches and machine-learning techniques becomes imminent. Clustering aims to group data samples into distinctive, compact, and homogeneous groups called

clusters. There is no formal definition for a cluster, but it can be described as a group of data samples that share common features with each other more than they do with samples in other clusters. The process of exploring these datasets often leads to groundbreaking discoveries, such as unexpected causes for a specific phenomenon, or a group of samples that requires attention.

Many clustering algorithms are available for data analysts; therefore, selecting the proper one among them is challenging. Choosing a clustering algorithm depends on the nature of the dataset. However, several metrics can be used to evaluate clustering algorithms. There are three types of cluster evaluations: internal criteria, external criteria, and relative criteria. These evaluation techniques are important in determining the quality of the resulting clusters and evaluating the performance of the applied clustering method.

1.4. RESEARCH OBJECTIVES AND CONTRIBUTIONS

In this work, several improvised machine-learning techniques and algorithms were implemented in the areas of smart grid and biomedical data analysis. The first area includes novel approaches in demand side management. In demand side management, a novel modeling approach is implemented to mitigate the peaks in grid load profile by using adaptive control algorithms, such as Q-learning and action dependent heuristic dynamic programming, to control the domestic electric water heaters. The implemented approach includes several novel techniques that were used for the first time in this optimization problem:

- An embedded event-driven simulator was designed to simulate each household hot-water consumption rate.
- The control operation of the domestic electric water heater was modeled as a Markov decision process and solved using reinforcement learning.
- Novel approaches were implemented to estimate water temperature and to identify system states.
- The problem was modeled as a multi-objective optimization problem, with the goals of reducing energy cost and smoothing load grid profile while maintaining desirable temperature for the supplied water.
- The implemented approach has the potential to be utilized in the internet of things.

According to the simulation results, the approximate dynamic programming approaches (Q-learning and action dependent heuristic dynamic programming) outperformed all other scenarios in terms of cost reduction, while maintaining desirable water temperature.

In the area of biomedical data analysis, this dissertation presents three papers. Each paper applied enhanced and improvised clustering techniques on specific biomedical datasets. The first paper in this area (Paper II), implements an ensemble statistical and subspace-clustering model to analyze autism spectrum disorder (ASD) phenotypes. This dataset is challenging due to the complex heterogeneity it conveys from the variability in behavioral phenotypes as well as clinical, physiologic, and pathologic parameters. In this paper, a new *k*-dimensional subspace-clustering algorithm is presented and used to analyze

and cluster the autism dataset. This algorithm is part of the general model that was implemented in this project. The implemented approach is general and can be applied to other biomedical datasets. It incorporates several statistical methods at different levels of the model. The implemented approach successfully sorts out the heterogeneity in the ASD dataset and produces clinically meaningful clusters. This approach is useful in understanding and studying etiology, diagnosis, treatment, and prognosis of ASD.

The second paper (Paper III) in this area applied a robust unified learning framework to cluster subgroups using neuroimaging data for brain volume and white matter. This unified model was used to identify neurological phenotypes that can sort the heterogeneity in cognitive aging and help identify potential risk factors for suboptimal brain aging. The use of machine-learning approaches identified two unique subgroups in healthy older adults with different patterns of white matter integrity and brain volumetric measures. The implemented model identified significant measurements that could potentially serve as biomarkers for delineating clinically meaningful aging subgroups.

The last paper (Paper IV) used machine-learning techniques to study the effect of contaminants and chemical compounds on plant uptake and the causes of pollutants transportation from the environment to vegetation and food. Several approaches were implemented in this paper: neural networks (NN), fuzzy logic, clustering, and statistical methods. The NN model was built to predict transpiration stream concentration factor. Fuzzy logic and clustering were used for predicting TSCF using physicochemical properties of compounds, and examining the interactions between compound properties. Several clustering algorithms have been applied, and they all discovered two major distinct clusters. The clusters resulting from k-means algorithm were the most significant and only

these are presented. Physiochemical property cutoffs, i.e. restrictions, for compounds passing plant roots membrane were shown to be lower than the cutoffs for transmembrane transport in mammalian intestinal systems. Therefore, the human health impacts through consumption of contaminated crops is elucidated and indicated that plant roots are a restrictive barrier to organic pollutants entering our foods. Improved understanding and prediction of plant uptake has significant implications for human health as we continue to shorten our water cycles.

PAPER

I. DEMAND-SIDE MANAGEMENT OF DOMESTIC ELECTRIC WATER HEATERS USING APPROXIMATE DYNAMIC PROGRAMMING

Khalid Al-jabery, Zhezhao Xu, Wenjian Yu, Donald C. Wunsch, II, Jinjun Xiong, and Yiyu Shi

ABSTRACT

In this paper, two techniques based on Q-learning and action dependent heuristic dynamic programming (ADHDP) are demonstrated for the demand-side management of domestic electric water heaters (DEWHs). The problem is modeled as a dynamic programming problem, with the state space defined by the temperature of output water, the instantaneous hot water consumption rate, and the estimated grid load. According to simulation, Q-learning and ADHDP reduce the cost of energy consumed by DEWHs by approximately 26% and 21%, respectively. The simulation results also indicate that these techniques will minimize the energy consumed during load peak periods. As a result, the customers saved about $466 and $367 annually by using Q-learning and ADHDP techniques to control their DEWHs (100 gallons tank size) operation, which is better than the cost reduction that resulted from using the state-of-the-art ($246) control technique under the same simulation parameters. To the best of the authors' knowledge, this is the first work that uses the approximate dynamic programming techniques to solve the DEWH's load management problem.

INDEX TERMS

Approximate dynamic programming (ADP), load management, machine learning, Markov processes, power demand, smart grids, unsupervised learning.

1. INTRODUCTION

The importance of domestic electric water heaters (DEWHs) can be seen from its effect on the overall grid load and energy consumption. For example, in the U.S. the average energy consumed by DEWHs is 18% as shown in Fig. 1. This share did not change during the last decade according to the U.S. Energy Information Administration [1]. The average annual cost of energy consumed by each DEWH is about $500 according to the office of energy and renewable energy [2]. Previous and current researches show that the total energy consumed in a city is highly dependent on the amount of power that the DEWHs consume [3]–[7]. For example, in the city of Quebec, Canada, the peaks in the grid load depends on the water heaters load, as illustrated in Fig. 2. The data plotted in Fig. 2, are from the field studies performed in two different cities [3], [4]. There is a clear relationship between the peaks in the grid load demand and those in the energy consumed for heating water in London and Quebec City. As a result, governmental agencies, policy makers, and of course energy companies, focus on domestic energy consumption with high priority for the water heaters. On April 30, 2015, president Obama signed into law S535, The Energy Efficiency Improvement Act, which established a new product category for large-capacity (75 + gallons) electric resistance “grid-enabled” water heaters for residential demand-response applications. Before that in the 15th of April, the Department of Energy

provided new rule that requires all large capacity (55+ gallons) DEWH would have to be integrated electric heat pump water heaters. The current trend of research according to the peak load management agency is to design efficient grid enabled water heater, which is exactly what we have produced in this paper [8]. The attention to water heaters and load management is not only in the U.S. According to the reports of the Ministry of Power in India, DEWHs consumes nearly 23% of the electricity in the domestic sector [9].

The DEWHs are often selected for demand side management projects because both their load profiles and their average daily load profiles almost follow the same pattern as shown in Fig. 2. Furthermore, DEWH loads are easier to control than other domestic appliances, because of their energy storage ability. Although many researches have been conducted on DEWHs, most of them failed to be widely applied to the DEWH industry for various reasons. There are different demand side management strategies for controlling DEWH loads. In 1998, Nehrir *et al.* [6] introduced a fuzzy logic controller that can shift the DEWH load outside the peak demand period. Some of the suggested approaches significantly affected the temperature of the DEWH's output water, resulting in customer dissatisfaction, plus the complex modeling process it needs [7]. In 2007, Atwa *et al.* [10] used Elman neural network to control the power consumed by water heaters. Some researchers, used detailed analytical methods for modeling the DEWH and provided control strategies based on dividing the load into groups and control them through the thermostats [10]–[17]. In 2011, Moreau [3] described control strategies aimed at distributing and shifting DEWH's operation to within one or two hours outside the peak periods. The new coming technology in water heaters industry is the use of electric heat pump and gas condensing technology for water heaters with tanks that are larger than 55

gallons [18]. However, there are three concerns raised on this new technology, first it needs new installations, second it tends to be used in water heaters with large tanks only. The third concern and the most important one is low temperature operation when the heat pump water heater operates in electric resistance mode, it does not save energy or money compared to a conventional unit [19]. Even if the new heat pump water heaters dominated the market, the approaches demonstrated in this paper still can be used to improve the performance because they are designed to adapt with the user activities, grid load and the temperature of the output water as discussed next.

In this paper, we used action dependent heuristic dynamic programing (ADHDP) [20] and *Q*-learning approaches to solve the DEWH management problem, which is a multi-objectives optimization problem. The objectives are: minimizing the total cost of the energy consumed, reducing the load demand during peak periods, and achieving customer's satisfaction. The *Q*-learning algorithm and the ADHDP approach are both approximate dynamic programming (ADP) techniques [20], [21]. It should be pointed out that this paper is the first study using ADP techniques in the DEWH's load management.

The novelty of this paper lies in the way that the system is modeled. Three factors were used to define and control a DEWH: 1) the temperature of the water delivered to the customers; 2) instantaneous hot water consumption; and 3) estimated grid load demand (i.e., instantaneous energy price). In the *Q*-learning based approach, the three factors are considered linguistic variables. They are categorized as either "high," "medium," or "low." The problem was modeled as a semi-Markov decision process (SMDP) with two possible actions in each state: 1) "ON" and 2) "OFF." Specific fuzzy rules are used to determine the system's current state. Each DEWH is considered to be an artificial agent that was trained

to adapt to the diversity within a user's consumption profile and grid load demand. The agent learns how to adapt, in the *Q*-learning approach, after finding the final *Q*-factors that specified the operating's policy during real time operation. In the ADHDP approach, the adaptation process consists of two phases: 1) critic training and 2) action training. The system learns to estimate the correct cost value during training the critic network, and then uses that cost in training the action network. The system tries to minimize the cost by adapting the weights of the action network [20].

These techniques can be applied to any DEWH, regardless of its capacity, heating elements, or operating environment. Furthermore, according to the simulation results, the approaches are able to reduce the energy consumed by DEWH more than the existing state-of-the-art methods [3]. The experiments show that the *Q*-learning and ADHDP controllers have reduced the cost of the energy consumed by the DEWHs by approximately 26% and 21%, respectively when using large (100 gallon tank) DEWH. As a result, both techniques will save about $466 and $367 per year, respectively, for the customers who use them to control their DEWH's operation. In comparison, only $246 would be saved annually by using the state-of-the-art control strategy [3], as illustrated in Table 3.

2. SYSTEM MODELING AND THE APPROXIMATE DYNAMIC PROGRAMMING

ADP techniques have been used effectively in solving optimization problems that consist of sequences of control actions whose efficiency remains unknown until the end of sequence. For the demand side management problem of DEWH, two ADP techniques: 1) ADHDP and 2) *Q*-learning (which is a special case of ADHDP) [20], were considered. The

system modeling, and the training and controlling processes of both techniques are explained in the following sections.

2.1. SYSTEM MODEL

The DEWH model is defined by three variables: *the output water temperature (T_h), Hot water consumption rate (W_{hc}), and the grid load (G_L). W_{hc}* is generated randomly (See Section 3.3), T_h is calculated using (2) and (3) based on the selected action, G_L depends on the city grid load profile. Therefore, the three variables are not correlated. However, the values of these variables were used differently for the two approaches presented in this paper.

In the Q-learning based approach, the variables were converted in to linguistic values using fuzzy membership functions, as illustrated in Fig. 3. The variables (T_h and W_{hc}) have three possible values (*low 'L', medium 'M' or high 'H'*) while G_L has only two possible values low or high. *(This assumption was based on the time-of-use ToU pricing profile that is used in this paper where there is no medium grid load or in other word medium power cost also it is meaningless practically to describe grid load as medium).*

Therefore, a discrete state system can be defined with 3×3×2=18 different states. The demand side management problem is to decide whether to turn the DEWH "On" or "Off" at each event time, which refers to the time when user consumes any quantity of water from the DEWH's tank. In practice, the numeric value of each variable should be fuzzified and mapped to its corresponding linguistic value. The fuzzy membership functions $f_{z,i}(\cdot), (i = 1,2,3)$ are defined for the three variables, and illustrated in Fig. 3. For variables Th and Whc, values of L, M and H correspond to 0, 1 and 2 respectively. For

variable G_L, the values of L and H correspond to 0 and 1, respectively. The water temperature thresholds were specified based on the fact that legionella bacteria begin to die at temperatures above 120° F [22].

Suppose v_1, v_2 and v_3 are the numeric values of the three variables respectively, and $S(v_1, v_2, v_3)$ is the corresponding state's index number. Then,

$$S(v_1, v_2, v_3) = \sum_{i=1}^{3}\left[3^{i-1} f_{z,i}(v_i)\right] + 1 \quad (1)$$

The system states are encoded as listed in Table 1.

Equation (1) is used to determine the system's current state during the training phase. It is used during the simulation as well. The linguistic values are vital for calculating the immediate reward during training (see Section 3). Actions are selected randomly with equal probability during the training phase in order to provide stochastic value iterations and update the Q-factors accordingly. As discussed in Section 3.2.

Variable T_h's numeric value at time (t+1), denoted by $v_1(t+1)$ can be estimated based on the action decision made at time (t). From the law of energy conservation [23], [24], we derive:

$$v_1(t+1)|_{a(t)=1} = \frac{9P\tau}{5K_jV} \cdot \frac{m_1 C_{hw} T_h(t) + [C_{cw} T_c - C_{hw} T_h(t)] \cdot m_2(t)}{m_1 C_{hw} + (C_{cw} - C_{hw}) \cdot m_2(t)} + 32\,, \quad (2)$$

$$v_1(t+1)|_{a(t)=2} = \frac{9}{5} \cdot \frac{m_1 C_{hw} T_h(t) + [C_{cw} T_c - C_{hw} T_h(t)] \cdot m_2(t)}{m_1 C_{hw} + (C_{cw} - C_{hw}) \cdot m_2(t)} + 32\,, \quad (3)$$

where $v_1(t)$ is the current temperature of the water, $a(t)$ is the current action, 1 means "On" and 2 means "Off". m_1 is the total mass of water in the DEWH tank, $m_2(t)$ is the mass of water consumed at the time (t), and $m_2(t) = v_2(t) \times 3.785$ [25]. C_{hw} and C_{cw} represent the heat capacity of hot water and cold water, respectively [23]. T_c is the temperature of the water

supplied to the DEWH, typically 10~13 ^{o} C [26]. P is the power rating of the heating element (4500, 2800 or 36000 Watts per hour in this study). K_j =2.42 W*h/gal*°F, which is the recovery rate calculation constant [27], [28], and V is the total volume of the DEWH tank. τ is the sampling period (30 minutes in this study). $v_1(t)$ and $v_1(t+1)$ are in unit of °F. Heat dissipation and heat exchange between the DEWH metal surface and air is negligible (*less than 0.25* ℃*/hour*) [19]. These two equations are to estimate the water temperature using energy saving formula. According to the behavior and ranges of the calculated temperature at each time step in compare with field studies [3]-[5], the model was acceptable. However, accurate performance to these models have not presented in this work. Due to the involvement of several random functions in profiles generation and the absence of real data.

An event driven simulator was designed to generate users' profiles. The simulator mimics the data that were collected in [3] and [4]. The distribution fitting toolbox in MATLAB was used in this study to determine the random variable distribution. The designed event driven simulator generates the time of the events (which specifies the current grid load G_L) and the quantity of hot water used in each event W_{hc}. The linguistic value of the grid load factor (G_L) is determined based on the event time since previous studies have shown that there are specific periods during the day when the load demand becomes high [3]-[7], [13]. However, a time-of-use (*ToU*) pricing profile is used to calculate the real cost of the consumed power [29] as illustrated in Section 3.3. The instantaneous output water temperature, $T_{h,}$ is calculated using (2) and (3) when the selected action at time (t) is "Off" and "On" respectively. One of the advantages of the approach is that it avoids the complicated thermodynamic and heat transfer operations,

which occur inside the DEWH's tank, by using (2) and (3) to estimate the value of T_h at time (t+1). Furthermore, the presented work does not require any complex calculations such as that described in previous studies [7] to solve the optimization problem.

In the ADHDP approach, the system is modeled as a continuous state space system. The system's state is also defined by the same variables (T_h, W_{hc}, and G_L) used in the Q-learning approach, but there is no fuzzification/ defuzzification process. The state variables are the inputs of the Critic and the Action neural networks in the ADHDP controller. Their normalized numeric values are used to train the neural networks. This will be explained in more detail in Section 3.1.

2.2. APPROXIMATE DYNAMIC PROGRAMMING

Approximate or Adaptive Dynamic programming (ADP), also known as the reinforcement learning, simulation-based dynamic programming, stochastic programming, and neuro-dynamic programming, refers to a group of algorithms designed to solve the problem of Markov and semi-Markov decision processes given by (4) [30].

$$J^*(i) = \max_{a \in A(i)} [\sum_{j=1}^{|S|} p(i,a,j)[r(i,a,j) + \lambda J^*(j)], \quad (4)$$

where $J^*(i)$ is the i-th element of the vector value function associated with the optimal policy. $A(i)$ is the set of all actions allowed in state i, $p(i, a, j)$ represents the transition probability of going from state i to state j under the influence of action a. $r(i, a, j)$ is an immediate reward earned when action a is selected in state i and the system transfers to state j as a result. S represents the set of states in the Markov chain, and λ is the discounting factor.

2.3. Q-LEARNING ALGORITHM

Watkins published the Q-learning algorithm in 1989. He defined this method as "a form of model free reinforcement learning and it can be viewed as a method of asynchronous dynamic programming" [31]. The Q-learning algorithm associates a scalar value, the Q-factor, with each state action pair. It solves (4) by updating the Q-factors associated with an optimal policy instead of approximating the cost function of a particular policy. Furthermore, it uses policy iteration (PI), as described in (5), to avoid the evaluation of multiple policies. The PI serves as the Q-factor version of the Bellman equation [21], [30].

$$\boldsymbol{Q(i,a) = \sum_{j=1}^{|S|} p(i,a,j)[r(i,a,j) + \lambda \max_{b \in A(j)} Q(j,b)]}, \quad (5)$$

where Q(*i*, *a*) and Q(*j*, *b*) are the Q-factors associated with state-action pairs (*i*, *a*) and (*j*, *b*), respectively.

Equation (5) still requires the transition probabilities. Therefore, the Robbins-Monro algorithm [32] was used to estimate the optimal Q-factors. The optimal Q-factors' estimation was achieved by expressing every Q-factor as an average of a random variable. Equation (6) represents the Q-factor version of the value iteration, which is the Q-learning algorithm for a discounted Markov Decision Process (MDP). The derivation of (6) from (5) can be found in [30].

$$\boldsymbol{Q^{n+1}(i,a) = (1 - \alpha^{n+1})\, Q^{n}(i,a) + \alpha^{n+1}[r(i,a,j) + \lambda \max_{b \in A(j)} Q(j,b)]}, \quad (6)$$

where $\boldsymbol{\alpha^{n+1}}$ represents the adaptive learning rate and attenuating with time. $\boldsymbol{Q^{n+1}}$ is the updated Q-factor, $\boldsymbol{n}$ is the time step.

Algorithm 1: Q-learning

1. ***Set up the training parameter: imax, and initialize Q-factors=0 and t=0.***
2. ***Randomly select initial state and action (S_0,a_0).***
3. ***Repeat until (number of iterations > imax).***
   - ***Apply action a(t) on DEWH model and read the current and the estimated values of vi(t) and vi(t+1), respectively.***
   - ***Determine $S_{(t+1)}$ using (1).***
   - ***Calculate immediate Reward $r(S_t,a_t,S_{t+1})$***
   - ***Update Total Reward $R_t=R_t+ r(S_t,a_t,S_{t+1})$.***
   - ***Update: $t=t+1$; $S_{(t-1)}=S_{(t)}$;$S_{(t)}=S_{(t+1)}$.***
   - [a] ***Update learning rate: $\alpha = \alpha^{t+1}$.***
   - ***Update Q-factors using (6).***
4. ***Construct the final policy from (S, a) pairs with higher Q-factors using the following formula on each state (i):***
   $P(i) = arg^{max}_{b\in A(i)} Q(i,b)$; $P(i)$ is the policy at state (i) (i.e. action that lead to maximum reward on the long run).
5. ***Record $\widehat{P}$(optimum policy for all states) and stop.***

[a] There are multiple ways to update the learning rate (α) [21]. In this work, we update α using: $\alpha^{t+1} = \frac{c_1}{c_2+t}$, where the positive constants c_1 and c_2 fulfills $c_1 < c_2$. (e.g. we set $c1 = 200$ and $c2 = 220$. *More discussion on learning rate selection can be found in [21]).*

In this study, the Q-learning version of the value iteration was used to solve the pre-described SMDP problem, which can be viewed as a discounted reward for the reinforcement learning based on the stochastic value iteration. However, the Q-learning version used in this paper uses a specific cost function to generate the immediate cost/reward for each system state transition, as illustrated in Section 3.1. This cost function eliminates the need for the transition reward matrix TRM, which is usually used in Q-learning algorithms. The same cost function was used to evaluate the system's transition in the ADHDP approach as well (See Section 2.4).

2.4. ACTION DEPENDENT HDP

The family of adaptive critic design (ACD) controllers has been presented by Werbos [33]. HDP, and its action dependent ADHDP forms, have a critic network that estimates the cost-to-go function J^* in (4) which calculates the Bellman equation of dynamic programming [20]. The standard structure of HDP and ADHDP is illustrated in Fig. 4. ADHDP is a generalization of Q-learning for the continuous domain system. In ADHDP, the critic is trained to provide an accurate estimation for the cost-to-go function *J* and to minimize the following error *E*(t):

$$E(t) = J[X(t)] - \gamma J[X(t+1)] - U(t), \tag{7}$$

where *X*(t) is the vector of observations/variables that define the system's current state.

The ADHDP controller adaptation process consists of two phases: Critic and Action networks training. These two processes are implemented continuously in sequence but not in parallel.

The critic network is designed to minimize a back propagated error signal (7), and the gradient of *J* with respect to the weights of the critic *Wc* is given by:

$$\Delta W_c = -\eta_c [E(t)] \times \frac{\partial J}{\partial W_c}, \tag{8}$$

where η_c is a positive learning rate ($0 < \eta_c \leq 1$). The action network is connected as shown in Fig. 4(b) in order to minimize *J* in the next time step and optimize the total cost over the entire domain. In the action network's adaptation phase, the gradient of *J* with respect to *A* (i.e. $\partial J/\partial A$) is back propagated as illustrated in Fig. 4(b) and in (9):

$$\Delta W_A = -\eta_a \times \frac{\partial J}{\partial W_A} \quad , \tag{9}$$

where $\partial J/\partial W_A = (\frac{\partial J}{\partial A}) \times (\frac{\partial A}{\partial W_A})$, is the gradient of the cost-to-go function ***J*** with respect to

the weights of the action network W_A. η_a is the action network learning rate (η_a doesn't have to be equal to η_c in (8.)

In HDP the immediate cost or the utility function U(t) is approximated as well using neural networks, while in ADHDP, U(t) is calculated using a model and the action network is connected directly to the critic [20]. The approaches described in this paper were designed to overcome the limitations presented by previous solutions. These improvements focused on the following:

- Reducing the need for permanent communications and synchronizations between DEWHs and the smart grid infrastructure [7], [12].
- Either shifting or eliminating the peaks with in the grid load [3].
- Reducing power consumption during peak periods and as a result minimizing the cost of the power consumed without sacrificing customer satisfaction.

3. TRAINING AND IMPLEMENTATION

This section contains the discussion on the processes required to transform a normal DEWH into a smart appliance. The discussion clarifies and compares the technical implementation of the presented approaches: the ADHDP and the Q-learning.

3.1. ADHDP IMPLEMENTATION

The ADHDP controller is illustrated in Fig. 5. It consists of the following components:

- The DEWH system module: which was described in (2) and (3) is the same module used during Q-learning and the final simulation (to be discussed further in Section 3.2 and Section 4).
- The critic network used in this work consists of: one input layer with four neurons (for each state variable and the action network output), one hidden layer with 30 neurons each of which has a hyperbolic tangent sigmoid activation function, and one neuron in the output layer with a linear activation function. The output of the critic is $J(t)$, if the inputs were $X(t)$ and $A(t)$ or $J(t+1)$ if the inputs were $X(t+1)$ and $A(t+1)$.
- The action network is almost identical to the critic network, but it only has three neurons in the input layer, and the sigmoid activation function is used in the hidden and also the output layer. The action network generates the control action (On or Off) during normal operation and simulation.

The transition evaluation module (or utility function) was designed to replace the TRM as illustrated in Section 2.3. This utility function (*in some literature called a cost function*) calculates the immediate transition reward/cost for each system's state transition. The same function was used in Q-learning as well. The utility function provides a more efficient evaluation than the TRM. The utility function calculates the immediate transition cost based on the energy consumed during the transition and all the other control variables ($T_h(t)$, $W_{hc}(t)$, $G_L(t)$), as illustrated in Algorithm 2. The pseudo code of the cost function illustrated in Algorithm 2 is of major importance because it highly reduced the complexity of the training process for both ADHDP and Q-learning compared to previous work [34]. The utility function rewarded the agent for each gallon of output water supplied with $T_h >$ 120 °F and penalized failure to do so. It also considers each consumed kWh as a penalty,

but that penalty depends also on whether it was consumed during peak or normal load. If G_L is low, the penalty will be mitigated through dividing the kWh by (a) and vice versa, as illustrated in Algorithm 2. The guidance factors (a and b) provide control over the multi-objective optimization process. (i.e. they encourage the agent to turn the heating element on during low load periods and to reduce the effect of the different scale between energy and output water units). Experimental results showed that: choosing a and b such that ($a \geq 2b, \forall b \geq 1$), provides better performance for the ADHDP controller. However, these factors have no effect at all on the Q-learning performance, as illustrated later in Table 3.

Algorithm 2: Utility function

```
Calculate_cost(power in kwh, state: X(t)= [Th, Whc, GL], t)
1   if  GL is high
        J=energy in kwh*a;    % penalty P1 {peak load}
        if Th< threshold
            J=J+Whc;          % increase cost {penalty}
        else: J=J-b*Whc;      % decrease cost {reward}
2   else
        if time (t) is between 3 and 5:30 am
            J= energy in kwh/a;     % P2<P1 low load
        Else
            J= energy in kwh/b;    % P1>P3>P2
        if Th< threshold
            J=J+Whc*a;        % increase cost  {penalty}
        else: J=J-Whc;        % decrease cost {reward}
3   If Q_learning             % see Section 3.1
        J=-J;
4   Return J;
```

The adaptation of the presented ADHDP controller was implemented in two phases:

- Offline training: In the offline training, the critic network was trained first using the data generated from the Q-learning algorithm. The critic training stops when the back

propagated error signal from (7) becomes less than a pre-specified small value, or when the training lasts for the maximum number of iterations. The critic training of the presented ADHDP is illustrated in Fig. 6. Furthermore, the action network is also trained during the offline phase using the same data used in critic adaptation. The size of the data sets depends on for how many simulation days Q-learning was trained, and each day contained about 150 samples. It was noted during experiences that repeating the offline training after the online training enhances the ADHDP controller's performance. The selection of the guidance factors has major effect on the ADHDP performance too as illustrated in Sections 4 and 5.

- Online training: The online training is executed during the simulation phase. The action network keeps adapting during the simulation to minimize the system's cost to go (i.e., *J).* Simulation here is the same as real time operation, because of the use of the event driven simulator that explained in Section 4. The adaptation of the action network is illustrated in Section 2.4.

The presented ADHDP controller showed better performance in cost reduction as the number of iterations increased. The critic adaptation is illustrated in Fig. 6 and was recorded during training for 100 simulation days with a total data set size of about 15000 samples.

3.2. Q-LEARNING IMPLEMENTATION

The second optimization approach implemented in this work is the Q-learning algorithm. The actions are selected randomly with equal probability at each time step to achieve better exploration of the solution space during the training phase. The algorithm

then receives the control variables' estimated values at (t+1) from the DEWH model and evaluates the performed action based on the utility function illustrated in Algorithm 2, which is the same utility function used in ADHDP. In step 3 from Algorithm 2, the calculated cost is negated. This is because Q-learning selects the optimum policy based on the Q-factor with the maximum value, unlike ADHDP which seeks to minimize the cost. The Q-factor associated with $(\boldsymbol{S_t}, \boldsymbol{a_t})$ was last updated using (6). This algorithm repeats the same procedures and continues until the maximum number of iterations are performed. The Q-factors have been stored in an (18x2) scalar matrix. The optimum policy is then derived, as illustrated in Algorithm 1 and Fig. 7.

Each iteration here represents a one-day simulation that contains about 150 time steps based on the event driven simulator used. The best value for the discount factor $\boldsymbol{\lambda}$ was derived heuristically as 0.9. Note that the same notation used in Fig. 5, to indicate the fact the same discount factor was used in ADHDP as well. The training phase for both of the presented approaches (ADHDP and Q-learning) was conducted using a DEWH's module with the following parameters:

1) The heating element power= 36, 4.5, 4.5 or 2.8 kWh.
2) Tank size=120,100, 70 or 40 gallons respectively.
3) Discount factor λ=0.9.

The state's variables linguistic values were derived as illustrated in Fig. 3, for Q-learning. The same specifications were used with in the simulation for all the other simulated approaches as well.

3.3. DATA PROCESSING

This section includes discussion on the process of generating the control variables' numerical values (T_h, W_{hc}, and G_L) and illustrates the *event driven simulator*.

The *event driven simulator* presented in this work provided comparable results with those obtained from previous field studies [3], [7]. The simulator is designed to mimic human activities in consuming hot water. This simulation was required to provide a reliable assessment for the presented DEWH's control approaches. The simulator generates (*per simulation day)* unique and random profiles for each DEWH used in the simulation as shown in Fig. 8. The simulator assumes 4 occupants in each house, for simplicity we avoided considering the ages, gender, and other social factors for the occupants that may affect their hot water consumption rate. The generated profile shown in Fig. 8, is comparable to the profiles obtained from the field studies in the British Department for Environment, Food and Rural Affairs' report [4].

The simulation consists of two phases: 1) Profile generation which was used in training and in performance evaluation or comparison and 2) comparator simulator, in which an evaluation process implemented among the presented approaches and the state_of_the_art approaches [3]. The profile's generator also provides the time of using the hot water and how much hot water was used. In the evaluation phase, different models for the DEWHs were used, uncontrolled "reference", Scenario0, 1, and 2, Q-learning, and ADHDP. Each group of DEWH have the same number of DEWHs units, the same tank size (120, 100, 70 or 40 gallons) and heating element (36, 4.5 or 2.8 kWh). The simulator assumes that, we are creating six parallel universes or copies from every house dwelling and give each copy different brand (i.e. version) of DEWH. The performance of each group is

evaluated during the simulation based how much energy each group can save with respect to the uncontrolled scenario (All scenarios operated for the same period of time and using the same user profile).

This criteria is used to guarantee a fair comparison among the different approaches. Otherwise, it is difficult to present an accurate comparison among the different control strategies. The user profiles which includes events' time indices and the consumed hot water quantities were generated using special combination of Poisson random variables. The choice of these variables is based on using the distribution fitting tool box from MATLAB on the previous studies' data. The profile generator function was adjusted empirically till it provides user profiles similar to the actual user profiles obtained by previous studies [3]-[5]. Artificial profiles needed due to the limitations of real data profiles. The numerical variables used for calculating the variables are explained as follows.

- Water temperature: In this work, instead of diving deep inside the thermodynamic operations of the DEWH, we utilized the law of energy preservation to provide a reasonable estimation for the temperature of the output water at the hot water faucet, as in (2) and (3). Since calculating the output water's exact instantaneous temperature is almost impossible without using an expensive embedded system to calculate the temperature of the DEWH's output water [23]-[25].
- Hot Water consumption rate: Estimating or predicting any human activity is extremely difficult. This study relied on statistics from field surveys, which have been performed in London, UK and Québec, Canada [3], [4]. However, to generate the required data, an embedded event driven simulator was designed as illustrated in the previous section.

- Grid Load "Energy Cost": The estimated instantaneous grid load can be obtained from the local utility companies, and they are time dependent as illustrated in Fig. 3 (c.) As mentioned earlier the grid load characteristics used in this work are based on data obtained from Quebec and London [3], [4]. However, the numeric values of this factor are the time indices of the operation. The load peak periods occurred approximately between 5:30 am and 10:00 am and between 4:30 pm and 10:00 pm. Furthermore, the final comparison was conducted using a time-of-use profile [29].
- Energy consumed by the DEWH's heating element: The amount of the consumed energy is calculated using the module described in (2) and (3). The values of the immediate energy consumption in kWh were used to calculate the value of the utility function at each system transition as illustrated in Figs. 5 and 7.

The control variables' numerical values are normalized before being used as inputs for the action and the critic networks in the ADHDP controller. The same profiles generated during the Q-learning process were used in the ADHDP approach as well.

4. SIMULATION AND EVALUATION

The simulation process was designed to provide the same operating conditions for all the simulated scenarios as discussed in the previous section. Five different approaches were simulated under the same operating conditions. These operating conditions are as follow:

1) The DEWH specifications as listed in Section 3.2.

2) All DEWHs must supply water in a temperature higher than 120 °F.

3) A soft threshold specified to be 125 °F. This soft threshold was used to prevent the output water's temperature to fall below 120 °F for all the compared approaches [22] (*To maintain customers' satisfaction.)* The DEWH heating element should be turned "On" whenever the water temperature fell below the soft threshold. However, the simulator recorded even the quantities of water outputted to users below thresholds to provide more accurate evaluation to each of the control strategies as illustrated in Tables 3 and 4.

The evaluation comparison is performed among five groups of DEWHs plus the uncontrolled operation as a reference. Each group has the same number of identical DEWHs. The results in this paper were derived from simulating the operation of 100 DEWHs in each group. Any number of DEWHs can be used and from experiences no effect on the comparison. Since the same user profiles is being used for the different groups. In other words: the same group of DEWHs were simulated using 5 different control scenarios and the uncontrolled scenario. The final assessment was presented based on the percentage cost reduction of the consumed energy cost using a *ToU* pricing profile. The *ToU* profile gives three different prices for the kWh during the day: 5.62¢, 10.29¢ and 23.26¢ [29]. These prices were applied to the load profile of the city of Quebec that used in this study and the state-of-the-art work that is compared with. As a result, the pricing profile that was used in our comparison simulator is as follow for the energy unit price:

- 23.26 ¢ for each kWh consumed between 5:30 and 10:30 am (load peak 1.)
- 10.29 ¢ for each kWh consumed between 4 pm and 9 pm (load peak 2.)
- 5.62 ¢ for each kWh consumed other times of the day (low grid load period.)

The different scenarios presented in the state-of-the-art work in the field (i.e., Scenarios 0, 1, and 2 in the list) [3], the uncontrolled scenario (i.e., Scenario 4), and our scenarios (scenarios 3 and 6) are all discussed below.

1) Scenario 0: The demand pick-up at the end of the load shifting period is not controlled.

2) Scenario 1: The pick-up is controlled according to a prioritized random function that was spread over a range of one hour after the peak period ended. In this scenario, the agent turns the heating element off during the peak period.

3) Scenario 2: The pick-up is controlled according to a prioritized random function that is spread over a range of two hours after the peak period ended. The success of the simulator can be verified by looking at the energy consumption curves in Figs 10-13.

4) Scenario 3 (Q-learning): The entire operation of every DEWH in the group is controlled according to the policy selected after the presented Q-learning algorithm (also known as the trained group in the comparison charts), which is used to train the agent.

5) Scenario 4 (Ref. or uncontrolled scenario): The DEWHs that are simulated under this scenario are operating under no artificial control. The heating element is turned "On" whenever the water temperature became less than the specified soft threshold and "Off" if it exceeds140 °F. This scenario (also known as the uncontrolled group in the comparison charts) is used as a reference to calculate the performance of all other scenarios.

6) Scenario 5 (ADHDP): All the DEWHs simulated with in this group are trained, as illustrated in Section 3.1, using the adaptive critic technique ADHDP.

Scenarios 3 and 5 are the techniques implemented in this work. Both scenarios (Q-learning and ADHDP) perform well and even better than the state-of-the-art strategies (Scenarios 0, 1, and 2) in the existing work [3].

In Scenario 0, the agent simply deactivated the heating element during peak periods unless T_h fell below the soft threshold, which is in this work was 125 °F. The controller turned the heating element "On" all the time it was outside the specified peak periods unless their (T_h) exceeded the maximum allowed temperature (140 °F). New peaks appeared when the heating elements for all DEWHs were reactivated simultaneously. In Scenarios 1 and 2, the agent randomly reactivated the water heaters at the end of the load shifting period, giving priority to those that were having the lowest water temperature to be turned "On" first. The time required for the water heaters to reactivate at the end of the shifting load was based on a random function. It was also based on the water's temperature at the end of the load shifting period.

Scenario 3 and 5 represented the control approaches that were presented in this study. In Scenario 3, the operation of the DEWH's heating element was entirely controlled by the suboptimal policy that was achieved during the Q-learning's training phase. The same simulator that generated the control variables during training was used to calculate them during comparison as well.

Furthermore, in all scenarios, the DEWH controller overrode its control scenario on two occasions: when T_h either decreased below or exceeded the pre-specified soft-threshold (125 °F) or maximum (140 °F) thresholds, respectively. The soft threshold was used in the comparator simulator in order to guarantee the same degree of customer's satisfaction for all scenarios (when all scenarios maintained their water's temperature above the hard threshold of 120 °F. It also provided clear performance measurement for the different scenarios, based on the consumed power cost only. The cost of the consumed power was calculated using a *ToU* pricing profile [29] as illustrated in Section 3.3.

5. RESULTS

The event driven simulator, the described system's modelling, the Q-learning process, ADHDP controller back propagation training, and the simulator used for evaluating the performance of all approaches were all designed using MATLAB. Many simulations were conducted in this work using different training parameters. As a result, the best learning schemes for Q-learning in terms of distance between Q-factors and policy stability were obtained using a discount factor of λ=0.9. Table 2 illustrates the optimum policies selected by Q-learning for 30 and 100 iterations for DEWH with tank size 70 gallons.

The Q-learning agent showed the best performance in all experiences. The Q-learning approach implemented here is more advanced and comprehensive than that presented in previous work [34]. The current work uses realistic time events as generated by the event driven simulator. (*In the previous work [34], the controller made a decision every 30 minutes*). The ADHDP approach is also conducted in all experiments and it outperformed the state_of_the_art approaches [3] when setting the appropriate values for the guidance parameters (a *and* b, See Section 3.1). The ADHDP approach is based on the continuous state space version of the problem, not a discrete state space like Q-learning [20]-[21], [30]-[31], [35]. Several experiments were implemented as illustrated in Tables 3 and 4. Table 3 contains results for the experiments that were implemented using the same profiles during training and evaluation, with user profiles illustrated in Fig.8. Table 3 also includes the results for different values for the guidance parameters (a *and* b) and clearly shows their effect on the ADHDP performance. Table 4 contains results for experiments

that uses extreme profiles during the evaluation phase, as illustrated in Fig.9. All the experiments in Table 3 were implemented twice using two different combinations of the guidance factors ($a\ and\ b$). The results are recorded twice for Q-learning and ADHDP, since these are the only approaches that may get affected by the influence of the guidance factors on the cost function. The remaining results were recorded when using ($a = 2 * b$). There were slight fluctuations in the values due to the random profile generation.

In experiment 1, a 100 typical DEWH with tank size of 70 gallons and 4.5 kWh heating element were simulated for 100 iterations (i.e. simulation days.) Experiment 2 repeats experiment 1 but using 30 iterations only. Experiment 3 evaluates the performance of all approaches for smaller tank size DEWH. DEWH of 40 gallons tank was used in this experiment. Experiments 4 and 5 show and compare the performance of all the different scenarios using DEWHs with larger tanks (i.e. 100 and 120 gallons). In experiment 5 a commercial DEWH model with 36 kWh heating element was simulated. The experiments listed in Table 4 (Experiments 6 and 7) were performed using extreme user profiles during the evaluation process, in order to measure the robustness of the presented approaches. The results obtained from experiments 1-5, showed outstanding performance for the Q-learning approach regardless of the guidance factors ($a\ and\ b$). The ADHDP approach outperformed the state_of_the_art techniques when $a = 2 * b; b = 2$. But it performed poorly when setting $a = b = 1$. It was observed during some additional experiments that ADHDP has shown better performance in cost reduction when setting a >>b (e.g. $a = 8 * b; b = 1$).

The comparison simulation was implemented as illustrated in Section 4. The simulation parameters were the same for the different scenarios, and each scenario had the

same number of DEWHs. Furthermore, the ADP approaches were tested using more extreme user profiles than those they were trained with, yet the Q-learning method was able to provide better cost reduction rates than those presented in the state_of_the_art techniques, as illustrated in Table 4, and Fig. 13.

Tables 3 & 4 compares the performance of the ADP approaches with the state-of-the-art control strategies [3] in terms of energy cost reduction and customers satisfaction. The amounts in the first row of each experiment were calculated in the code by accumulating the instantaneous costs of energy consumed in all the 100 dwellings. The cost is based on the ToU profile described earlier [29]. The percentage cost reduction rates (i.e. numeric values in the 2nd row) were calculated using (10).

$$\%\,\boldsymbol{Cost\ Reduction} = \frac{\boldsymbol{Approach's\ Total\ saving}}{\boldsymbol{Ref.\ energy\ cost}} \times \%\mathbf{100} \tag{10}$$

Where approach's total saving = (Cost of energy consumed using the uncontrolled approach 'Ref.' – The cost of energy consumed using the approach). The uncontrolled scenario is used as an index for comparison or to evaluate between all the control strategies. The estimated annual saving (EAS) was calculated from multiplying the per-day saving by 365 as illustrated in (11). The customer's satisfaction evaluation was based on the quantities of the output water with temperature less than '*th*' (i.e. 120 deg. Fahrenheit) and the number of times that happened in the entire 100 dwellings.

$$\boldsymbol{EAS} = \frac{\boldsymbol{Approach's\ Total\ saving}}{\boldsymbol{\#\ DEWHs}} \times \frac{\mathbf{365}}{\boldsymbol{\#\ simulation\ days}} \tag{11}$$

The results illustrated in Figs. 10-13, and in Tables 3 & 4 indicate that, in most cases, using Q-learning to control the 100 DEWHs reduced the cost of the consumed power by 22% which is twice the cost reduction resulted from the state_of_the_art technique

"Scenario-0". The ADHDP approach also outperforms the state_of_the_art techniques, through reducing the total cost by 16% and 12% when trained for 100 and 30 days respectively.

In experiment 3, Q-learning reduced the total cost by 6.6% which is still higher than the reduction from other techniques (e.g. %5.7 for Scenario-0). ADHDP in this experiment reduced the cost by only 5.5% only which is less than Scenario-0, but higher than Scenarios 1 and 2. However, the best performance recorded was in experiment 4 with tank size=100 gallons and heating element of 4.5 kWh. The percentage cost reduction rates were $\approx \mathbf{26}\% \boldsymbol{and} \mathbf{21}\%$ for the Q-learning and ADHDP respectively. The annual savings were approximately $453 and $367 for Q-learning and ADHDP respectively.

In experiments 6 and 7 (Table 4), a higher user profile was used in the evaluation than that used during training. The Q-learning also outperformed all other scenarios by producing about 15% cost reduction. According to the simulation results, the Q-learning controller maintained the temperature of the output water above the pre-specified threshold (120° F), except when the 40 gallons tank size was used, it provided a small amount of water slightly below the threshold, "*th"*. The ADHDP approach outperformed the state_of_the_art scenarios if the training was performed on the same data used during evaluation and a large tank size (70 gallons) was used.

6. CONCLUSIONS

In conclusion, the Q-learning approach can at least save a family of 4 persons between $102, $393 and $453 annually if they are using a DEWH with 40, 70 and 100

gallons respectively. Even for the commercial product the ADP approaches provide excellent annual saving of ($394) for DEWH with larger heating element (36 kWh) and tank size of 120 gallons. Furthermore, the simulation showed that the Q-learning controller maintained the water temperature above 120° F, indicating an opportunity to enhance the system further by designing a flexible threshold. The simulation results shown in Tables 3 & 4 clearly illustrate that Q-learning has the best performance in terms of cost reduction, customer's satisfaction, and even in terms of load peak elimination or shifting as illustrated in Figs. 10-13. Another opportunity for further enhancement would be using real user profiles to control DEWH in real time. ADP may also be used to improve the most recent Heat pump water heaters. The presented techniques don't depend on the technology used in the DEWH, but depend only on the grid load demand (i.e. instantaneous energy cost), the temperature of the output water, and the user profile. The authors therefore believe that various ADP approaches are worth further investigation. Q-learning outperformed ADHDP and previous state-of-the-art methods in these experiments. The authors speculate that ADHDP will still prove useful in scenarios that play to its strengths in continuous state spaces and dynamic environments such as adaptive thresholds. The authors are also aware that some of the references cited in this paper demonstrated better performance with ADHDP than with Q-learning. However, for these experiments, Q-learning outperformed it and all other methods, probably due to the reduced state space and the limited complexity of the implemented state space model. This is encouraging; for Q-learning is a simple and robust, easily deployable ADP approach. The results presented here strongly suggest it should not be difficult to use simple machine learning techniques to achieve substantial cost savings and environmental benefits.

Table 1. System states encoding (L: low=0, M: Medium=1, H: High=2).

S_i	S_1	S_2	S_3	S_4	S_5	S_6	S_7	S_8	S_9	S_{10}	S_{11}	S_{12}	S_{13}	S_{14}	S_{15}	S_{16}	S_{17}	S_{18}
T_h	L	M	H	L	M	H	L	M	H	L	M	H	L	M	H	L	M	H
W_{hc}	L	L	L	M	M	M	H	H	H	L	L	L	M	M	M	H	H	H
G_L	L	L	L	L	L	L	L	L	L	H	H	H	H	H	H	H	H	H

Table 2. The optimum policies selected by q-learning (1= On, 0= Off).

S_t[a]	S_1	S_2	S_3	S_4	S_5	S_6	S_7	S_8	S_9	S_{10}	S_{11}	S_{12}	S_{13}	S_{14}	S_{15}	S_{16}	S_{17}	S_{18}
a(t)[b]	1	1	1	1	0	1	1	1	1	1	0	0	1	0	0	1	0	0
a(t)[c]	1	1	1	1	1	0	1	1	0	1	0	0	1	0	0	1	0	0

[a] States (1-18)
[b] policy for 30 iterations
[c] policy for 100 iterations

Table 3. Simulation results from different experiments using the same user profiles in training and simulation.

Experiment .1: Tank size:70 gallons, 100 simulation days, heating element 4500Wh								
Approaches	**Scenario-0**	**Scenario-1**	**Scenario-2**	**Q-learning**		**ADHDP**		**Ref.**
				$a = b = 1$	$a = 2 * b$	$a = b = 1$	$a = 2 * b$	
Energy Cost in US $ for all DEWHs	43765	45373	45325	37820	37996	54683	42917	48578
Percentage of cost's reduction	9.9%	6.6%	6.7%	22.4%	21.78%	-12.2%	11.65%	-
Estimated Customer's Annual saving	$173	$115	$117	$393	$381	$-213	$204	-
# times users receives water below '*th*'	815	1998	1916	0	0	2042	954	132
Output water below '*th*' in gallons	2190	5035	5006	0	0	9114	2674	331
Experiment .2: Tank size: 70 gallons, 30 simulation days, heating element 4500Wh								
Energy Cost in US $ for all DEWHs	13296	13766	13775	11438	11571	16203	12522	14848
Percentage of cost's reduction	10.45%	7.29%	7.23%	22.25%	22%	-10%	15.67%	-
Estimated Customer's Annual saving	$186	$130	$129	$393	$393	$-179	$279	-
# times users receives water below '*th*'	382	608	570	0	0	20	261	38
Output water below '*th*' in gallons	1031	1521	1453	0	0	56	692	124
Experiment .3: Tank size: 40 gallons, 100 simulation days, heating element 2800Wh								
Energy Cost in US $ for all DEWHs	40400	40922	40930	40014	40016	45661	40485	42843
Percentage of cost's reduction	5.7%	4.48%	4.46%	6.6%	6.6%	-6.6%	5.5%	-
Estimated Customer's Annual saving	$87.97	$69.16	$68.86	$101	$102	$-102	$85	-
# times users receives water below '*th*'	17707	19763	19880	0	4	5388	13421	19528
Output water below '*th*' in gallons	35423	39515	39747	0	4.49	10830	26902	39072
Experiment .4: Tank size: 100 gallons, 30 simulation days, heating element 4500Wh								
Energy Cost in US $ for all DEWHs	12574	13188	13048	10819	10856	16900	11566	14628
Percentage of cost's reduction	14%	9.84%	10.8%	26.4%	25.78%	-14.98%	20.93%	-
Estimated Customer's Annual saving	$246	$173	$190	$466	$453	$-264	$367	-
# times users receives water below '*th*'	33	288	315	3	0	1416	58	0
Output water below '*th*' in gallons	22.2	672.1	693.5	3.17	0	594.5	182	0
Experiment .5: Tank size: 120 gallons, 100 simulation days, heating element 36000Wh (Commercial product)								
Energy Cost in US $ for all DEWHs	48807	46592	46441	36647	36585	48391	44915	47582
Percentage of cost's reduction	-2.54%	2.08%	2.4%	22.98%	22.77%	-1.7%	5.6%	-
Estimated Customer's Annual saving	$-44	$35.7	$41	$394	$388	$-29.45	$96	-
# times users receives water below '*th*'	0	0	10	0	0	0	0	0
Output water below '*th*' in gallons	0	0	5.3	0	0	0	0	0

Table 4. Simulation results from different experiments using user profile in fig. 8 for training and the profile in figure 9 in evaluation.

Experiment .6: Tank size:70 gallons, 100 simulation days, heating element 4500 Wh						
Approaches	**Scenario-0**	**Scenario-1**	**Scenario-2**	**Q-learning**	**ADHDP**	**Ref.**
Energy Cost in US $ for all DEWHs	**56011**	**57796**	**57785**	**53500**	**57114**	**62841**
Percentage of cost's reduction	**10.87%**	**8.03%**	**8.05%**	**14.86%**	**9.12%**	**-**
Estimated Customer's Annual saving	**$249.29**	**$184.14**	**$184.54**	**$340.95**	**$209.04**	**-**
# times users receives water below 'th'	**49709**	**60584**	**60199**	**0**	**57242**	**18645**
Output water below 'th' in gallons	**150140**	**182500**	**181250**	**0**	**172760**	**56596**
Experiment .7: Tank size: 70 gallons, 30 simulation days, heating element 4500 Wh						
Energy Cost in US $ for all DEWHs	**16666**	**17227**	**17226**	**16032**	**16864**	**18873**
Percentage of cost's reduction	**11.52%**	**8.55%**	**8.553%**	**14.9%**	**10.47%**	**-**
Estimated Customer's Annual saving	**$268.52**	**$200.26**	**$200.39**	**$345.66**	**$244.43**	**-**
# times users receives water below 'th'	**12708**	**14694**	**14530**	**0**	**14548**	**3599**
Output water below 'th' in gallons	**38931**	**44903**	**44446**	**0**	**44464**	**11156**

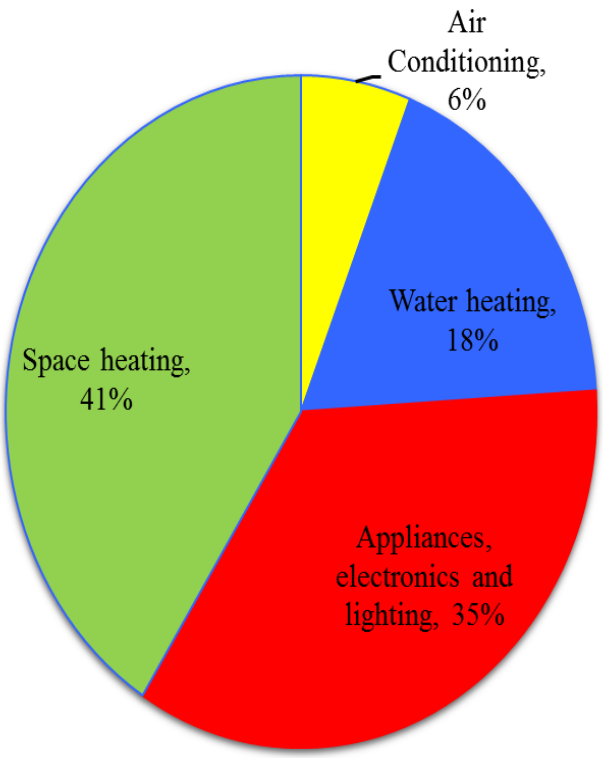


Figure 1. House hold energy use distribution in the US (Aug. 2013) [1].

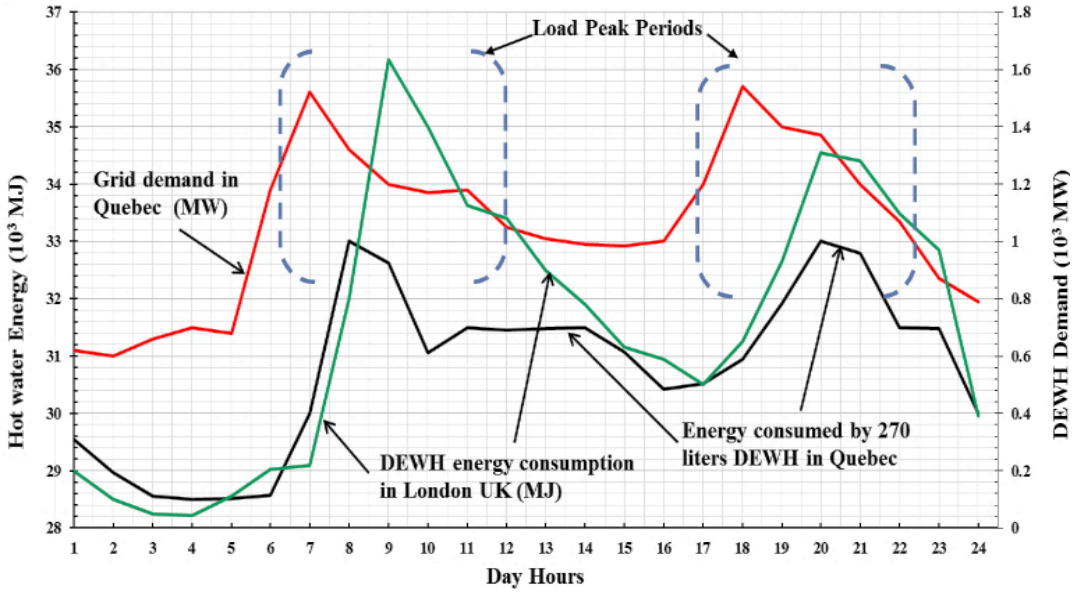


Figure 2. Similarity in load peak periods between grid load demand and energy consumed by DEWH in Quebec CA and London UK [2].

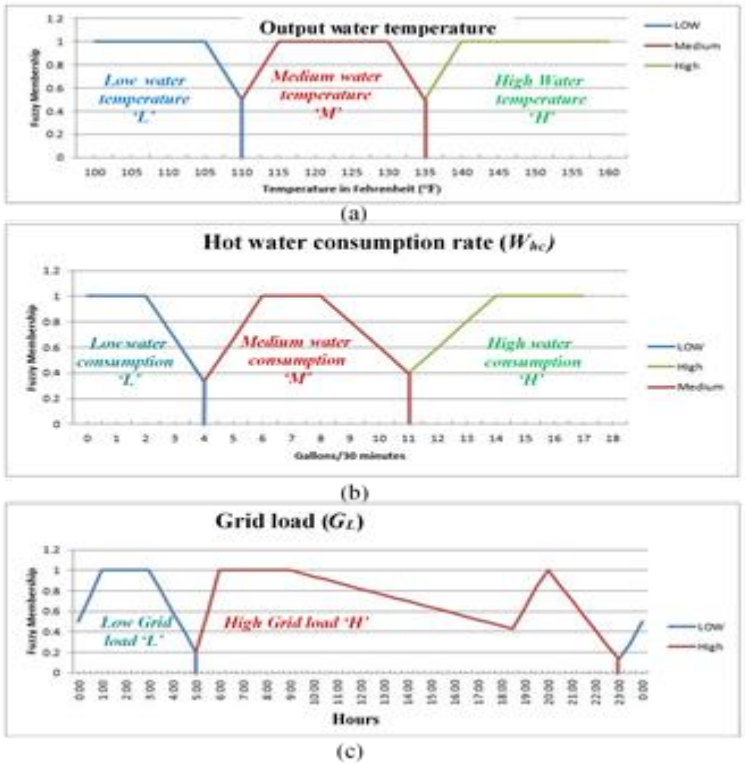


Figure 3. The illustrations of the fuzzy membership functions: (a) $f_{(z,1)}(\cdot)$ for T_h, (b) $f_{(z,2)}(\cdot)$ for W_{hc}, (c) $f_{(z,3)}(\cdot)$ for G_L. Where GL is a function of time in hour.

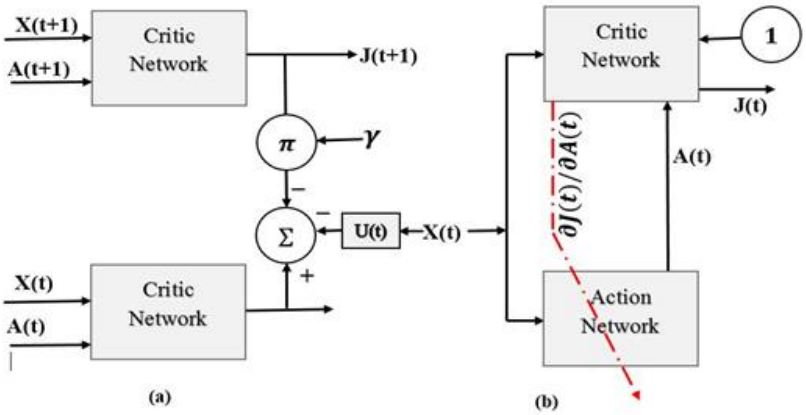


Figure 4. (a) Critic adaptation in ADHDP/HDP. This is the same critic network in two consecutive moments in time. The critic's output J(t+1) is necessary in order to give us the training signal γJ(t + 1)+U(t), which is the target value for J(t). (b) Action adaptation. X is a vector of observables, and A is a control vector. We use the constant ∂J/∂J= 1 as the error signal in order to train the action network to minimize J. This figure is adapted from [20].

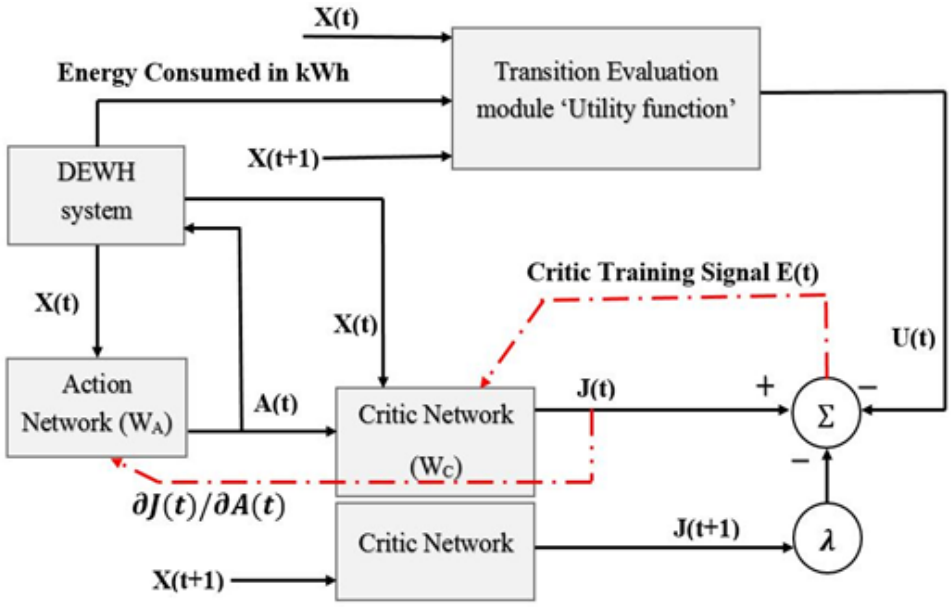


Figure 5. Implemented ADHDP controller's structure. X: state, J: cost, A: action, any variable with (t) means current (t+1) means next. U(t) immediate transition cost "Utility", WC,WA: Critic and Actor networks weight matrices respectively.

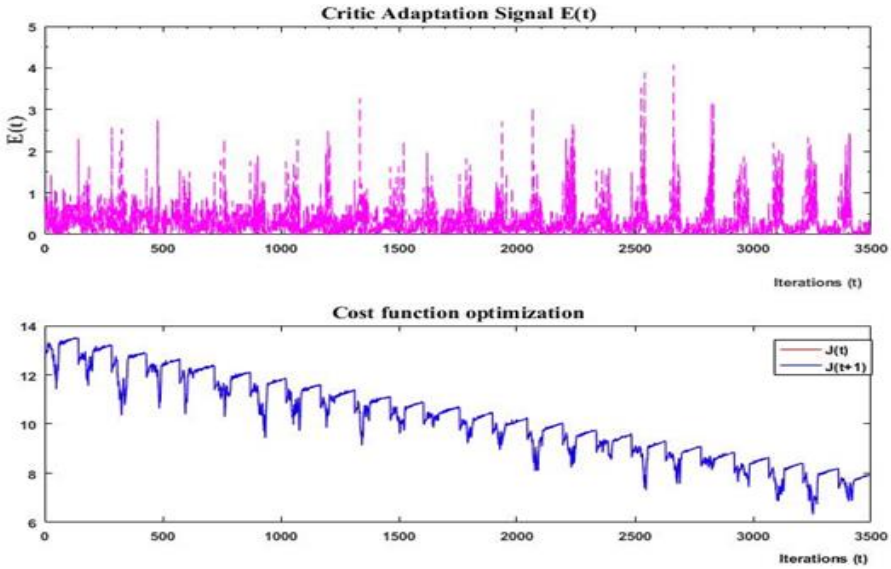


Figure 6. ADHDP Critic Network Adaptation. The top figure shows instantaneous error signals during critic adaptation. The lower figure illustrates the total cost reduction during critic adaptation.

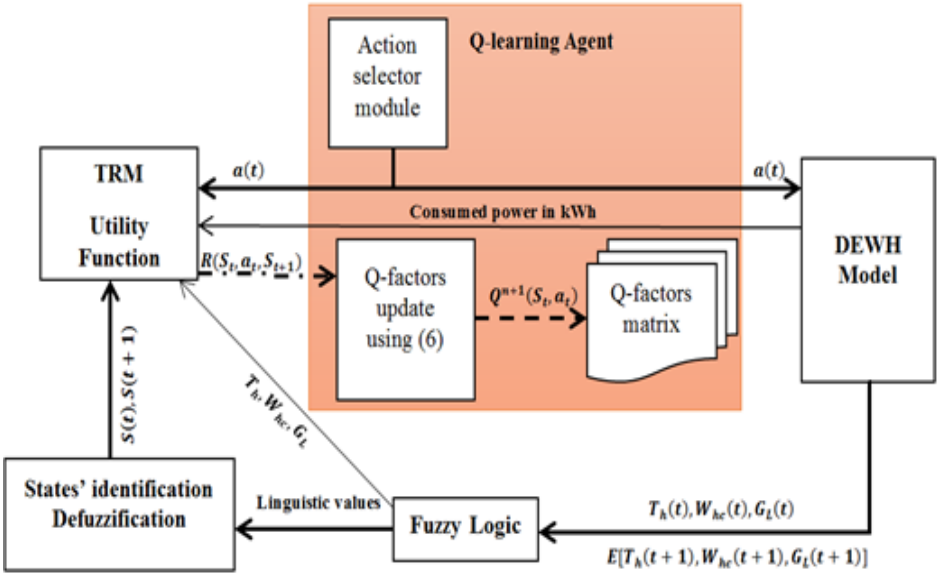


Figure 7. Q-learning's schematic diagram. Qn+1:new value of Q-factor, R(St,at,St+1):immediate reward due to system's transition from current state St to St+1 using the current action at, E[…] estimating the values of the enclosed factors, Q-factors' matrix is an 18 x 2 matrix.

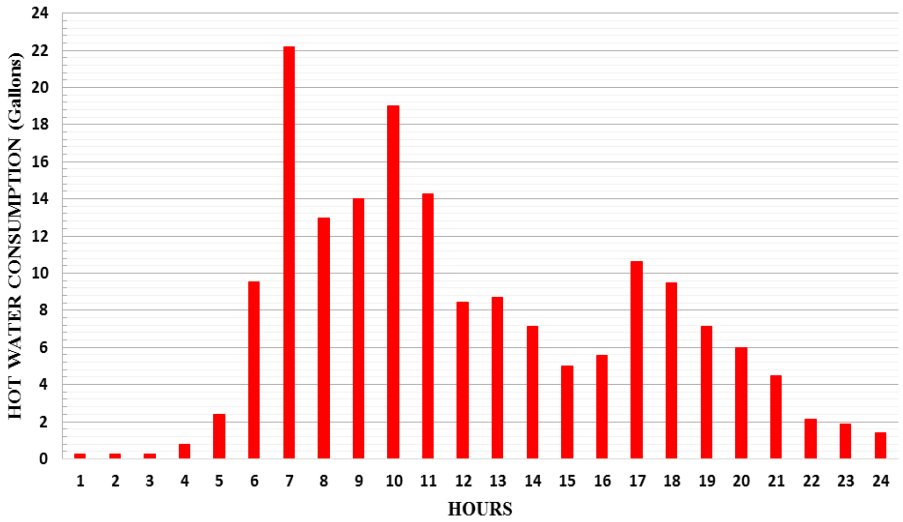


Figure 8. Sample user profile generated from the event driven simulator. Where the profile generated using embedded code as (See Section 3.3).

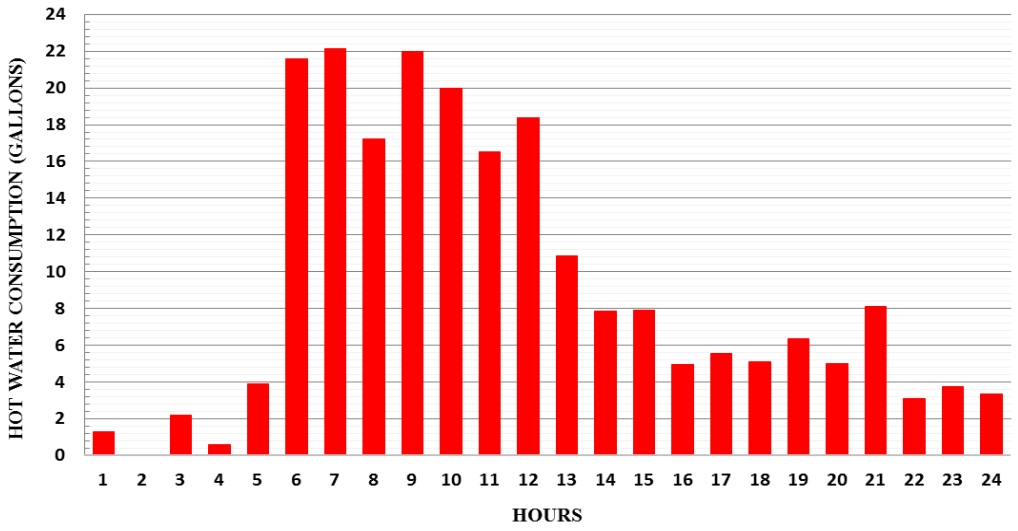


Figure 9. Extreme user profile used in Experiments 6 and 7. In this profile higher rate of hot water consumption is assumed for users.

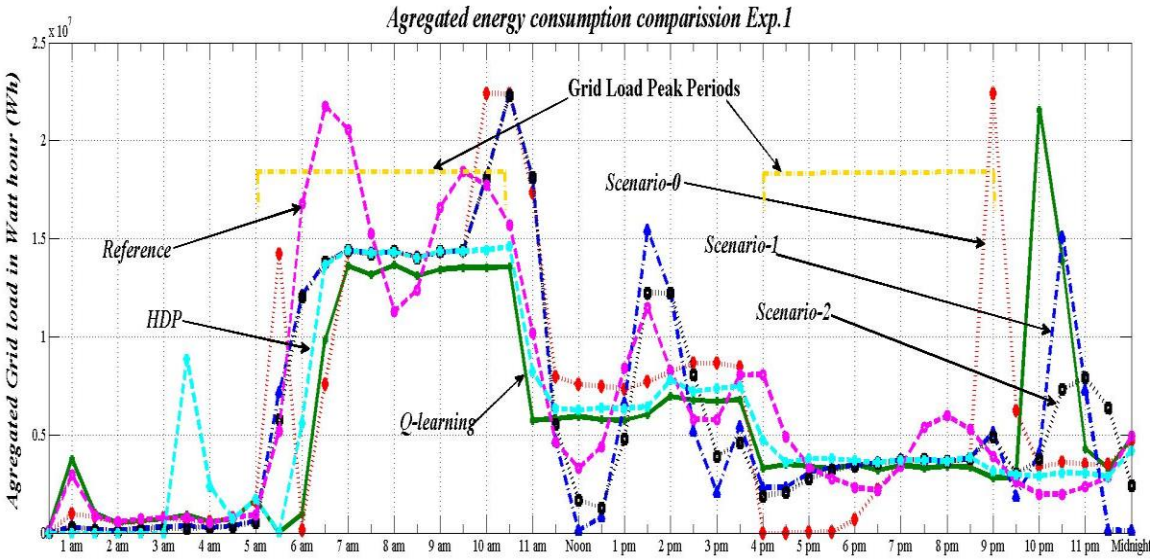


Figure 10. Aggregated energy consumed by all approaches during exp.1. DEWH's Tank size 70 gallons, 100 simulation days.

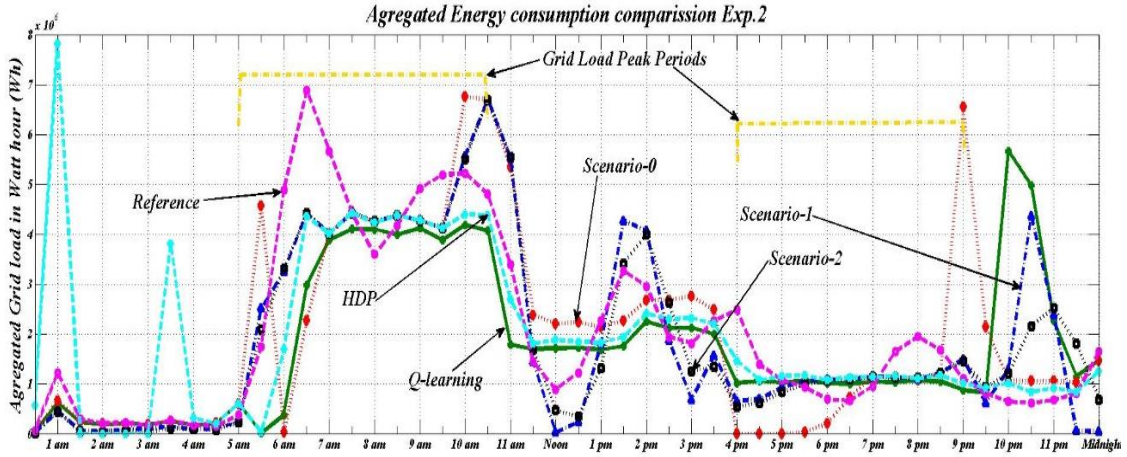


Figure 11. Aggregated energy consumed by all approaches during exp.2. DEWH's Tank size 70 gallons, 30 simulation days.

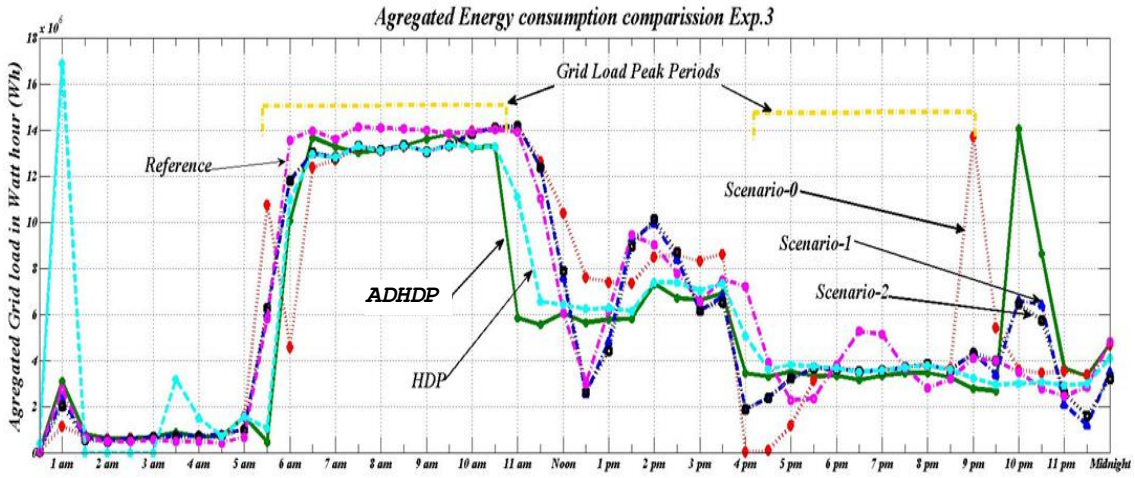


Figure 12. Aggregated energy consumed by all approaches during exp.3. DEWH's Tank size 40 gallons, 100 simulation days.

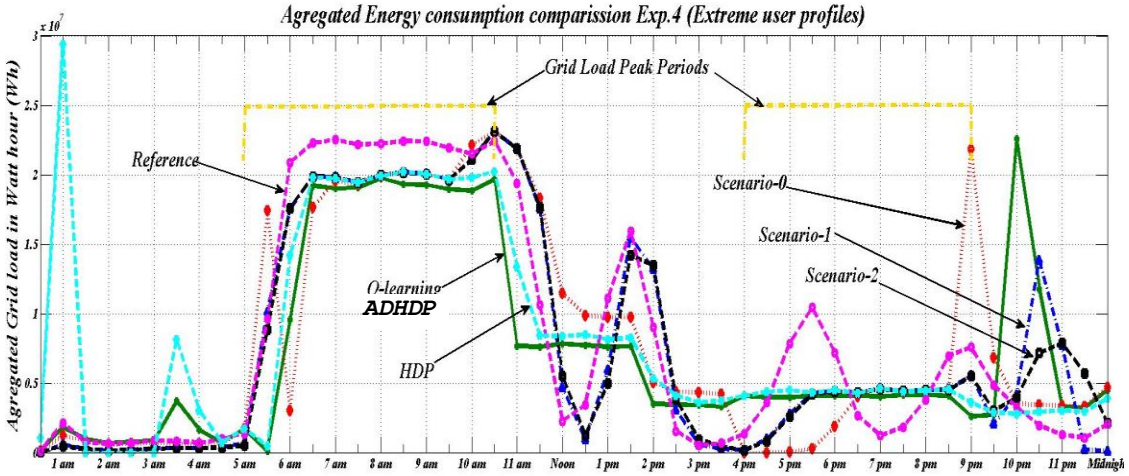


Figure 13. Aggregated energy consumed by all approaches during exp.6. Users' profiles in Figure 9 were used in comparison.

REFERENCES

[1] Eia.gov,. (2015). Residential Energy Consumption Survey (RECS) - Analysis & Projections - U.S. Energy Information Administration (EIA). Retrieved 24 October 2015, from http://www.eia.gov/consumption/residen

[2] Energy.gov,. (2015). Office of Energy Efficiency & Renewable Energy | Department of Energy. Retrieved 25 October 2015, from http://energy.gov/eere/office-energy-efficiency-renewable-energy

[3] Moreau, A. (2011). Control strategy for domestic water heaters during peak periods and its impact on the demand for electricity. *Energy Procedia 12* (1074 – 1082). Chengdu, China: Elsevier.

[4] Department for Environment, Food and Rural Affairs. (2008). Measurement of domestic hot water consumption in dwellings. London; UK: DEFRA.

[5] Lu, S. and Kintner-Meyer, M. (2008). Scoping study for the Demand Response DFT II Project in Morgantown. (PNNL-17474). Richland, Washington; USA: Pacific Northwest National Laboratory.

[6] Nehrir, M. H., LaMeres, B. J., and Gerez, V. (1999, January). A customer-interactive electric water heater demand-side management strategy using fuzzy logic. In *Power Engineering Society 1999 Winter Meeting, IEEE* (Vol. 1, pp. 433-436). IEEE.

[7] Sepulveda, A., Paull, L., Morsi, W. G., Li, H., Diduch, C. P., and Chang, L. (2010, August). A novel demand side management program using water heaters and particle swarm optimization. In *Electric Power and Energy Conference (EPEC), 2010 IEEE* (pp. 1-5). IEEE.

[8] Peakload.org,. (2015). 2014 Annual Report to Members - Peak Load Management Alliance. Retrieved 25 October 2015, from http://www.peakload.org/?page=2014Report.

[9] Powermin.nic.in,. (2016). Annual Reports Year-wise Indian Ministry of Power. Retrieved: 2 February 2016, from http://powermin.nic.in/annual-reports-year-wise.

[10] Atwa, Y. M., El-Saadany, E. F., & Salama, M. M. (2007, October). DSM Approach for Water Heater Control Strategy Utilizing Elman Neural Network. In Electrical Power Conference, 2007. EPC 2007. IEEE Canada (pp. 382-386). IEEE.

[11] Saker, N., Petit, M., Vannier, J. C., & Coullon, J. L. (2011). Demand Side Management of Electrical Water Heaters and Evaluation of the Cold Load Pick-Up characteristics. In 17th Power System Computation Conference (p. 8p).

[12] S.Lefebvre and C.Desbiens. "Residential load modeling for predicting distribution transformer load behavior, feeder load and cold load pickup". International Journal of Electrical Power & Energy Systems, vol.24, pp.285-293, May 2002.

[13] Diduch, C., Shaad, M., Errouissi, R., Kaye, M. E., Meng, J., and Chang, L. (2012, June). Aggregated domestic electric water heater control - Building on smart grid infrastructure. In Power Electronics and Motion Control Conference (IPEMC), 2012 7th International (Vol. 1, pp. 128-135). IEEE.

[14] Ramanathan, B., and Vittal, V. (2008). A framework for evaluation of advanced direct load control with minimum disruption. *Power Systems, IEEE Transactions on*, *23*(4), pp. 1681-1688.

[15] Tiptipakorn, S., and Lee, W. J. (2007, September). A residential consumer-centered load control strategy in real-time electricity pricing environment. In *Power Symposium, 2007. NAPS'07. 39th North American* (pp. 505-510). IEEE.

[16] Lu, N., and Katipamula, S. (2005, June). Control strategies of thermostatically controlled appliances in a competitive electricity market. In *Power Engineering Society General Meeting, 2005. IEEE* (pp. 202-207). IEEE.

[17] Rautenbach, B., and Lane, I. E. (1996). The multi-objective controller: A novel approach to domestic hot water load control. *Power Systems, IEEE Transactions on*, *11*(4), 1832-1837.

[18] Aceee.org,. (2015). Water heaters get an efficiency makeover courtesy of the Department of Energy | ACEEE. Retrieved 30 October 2015, from http://aceee.org/blog/2015/02/water-heaters-get-efficiency-makeover.

[19] Appliance-standards.org,. (2015). The good news, and the not-so-good news, on the new DOE water heater test procedure | ASAP Appliance Standard Awareness Project. Retrieved 30 October 2015, from http://www.appliance-standards.org/blog/good-news-and-not-so-good-news-new-doe-water-heater-test-procedure.

[20] Prokhorov, D. V., & Wunsch, D. C. (1997). Adaptive critic designs. Neural Networks, IEEE Transactions on, 8(5), 997-1007.

[21] Bertsekas, D. P. (1995). *Dynamic programming and optimal control* (Vol. 1, No. 2). Belmont, MA: Athena Scientific.

[22] Who.int,. (2015). Retrieved 13 November 2015, from http://www.who.int/water_sanitation_health/e

[23] Sonntag, R. E., Borgnakke, C., Van Wylen, G. J., and Van Wyk, S. (1998).*Fundamentals of Thermodynamics* (pp. 356-57). New York: Wiley.

[24] Klemes, J., Smith, R., and Kim, J. K. (Eds.). (2008). *Handbook of water and energy management in food processing*. Elsevier.

[25] Cardarelli, F. (2003). *Encyclopaedia of scientific units, weights and measures: Their SI equivalences and origins*. Springer.

[26] Omer, A. M. (2008). Energy, environment and sustainable development.*Renewable and Sustainable Energy Reviews*, *12*(9), pp. 2265-2300.

[27] Mujumdar, A. S. (2006). A review of: "Mathematical Principles of Heat Transfer." *Drying Technology*, 24(2), 245.

[28] Formulas and Facts. (n.d.). Contractorsinstitute.com. Retrieved January 13th, 2014 from http:// www.contractorsinstitute.com/downloads/Solar/Contractors'%20Domestic%20Hot%20Water%20Educational%20PDF's/Hot%20Water%20Formulas%20and%20Facts.pdf.

[29] Borenstein, S. (2005). Time-varying retail electricity prices: Theory and practice. Electricity deregulation: choices and challenges, 111-130.

[30] Watkins, C. J., and Dayan, P. (1992). Q-learning. *Machine Learning*, *8*(3-4), pp. 279-292.

[31] Gosavi, A. (2003). *Simulation-based optimization: Parametric optimization techniques and reinforcement learning* (Vol. 25). Springer.

[32] Robbins, H., & Monro, S. (1951). A stochastic approximation method. The annals of mathematical statistics, 400-407.

[33] Werbos, P. J. (1994). The roots of backpropagation: from ordered derivatives to neural networks and political forecasting (Vol. 1). John Wiley & Sons.

[34] Al-jabery, K., Wunsch, D. C., Xiong, J. and Shi, Y. A novel grid load management technique using electric water heaters and Q-learning. 2014 IEEE International Conference On Smart Grid Communications.

[35] Venayagamoorthy, G., Harley, R., & Wunsch, D. (2002). Comparison of heuristic dynamic programming and dual heuristic programming adaptive critics for neurocontrol of a turbogenerator. IEEE Trans. Neural Netw., 13(3), 764-773.

II. ENSEMBLE STATISTICAL AND SUBSPACE CLUSTERING MODEL FOR ANALYSIS OF AUTISM SPECTRUM DISORDER PHENOTYPES

Khalid Al-jabery, Tayo Obafemi-Ajayi, Gayla R. Olbricht, T. Nicole Takahashi, Stephen Kanne and Donald Wunsch

ABSTRACT

Heterogeneity in Autism Spectrum Disorder (ASD) is complex including variability in behavioral phenotype as well as clinical, physiologic, and pathologic parameters. The fifth edition of the Diagnostic and Statistical Manual of Mental Disorders (DSM-5) now diagnoses ASD using a 2-dimensional model based social communication deficits and fixated interests and repetitive behaviors. Sorting out heterogeneity is crucial for study of etiology, diagnosis, treatment and prognosis. In this paper, we present an ensemble model for analyzing ASD phenotypes using several machine learning techniques and a k-dimensional subspace clustering algorithm. Our ensemble also incorporates statistical methods at several stages of analysis. We apply this model to a sample of 208 probands drawn from the Simon Simplex Collection Missouri Site patients. The results provide useful evidence that is helpful in elucidating the phenotype complexity within ASD. Our model can be extended to other disorders that exhibit a diverse range of heterogeneity.

1. INTRODUCTION

Children with Autism Spectrum Disorder (ASD) make up a heterogeneous population varying widely in the type, number and severity of social deficits, behavioral

problems, communication, language and cognitive difficulties, and etiologic conditions. Among all child-onset psychiatric disorders, autism is perhaps the most serious, intractable and challenging to address because of its vast heterogeneity observed along a spectrum of pathology [1]. The Diagnostic and Statistical Manual of Mental Disorders fifth edition (DSM-5) diagnostic criteria [2] for ASD are based on a severity gradient using a two-dimensional model: social communication deficits (SCD) and fixated interests and repetitive behaviors (FIRB). Heterogeneity in ASD is multidimensional and complex including variability in phenotype as well as clinical, physiologic, and pathologic parameters. Based on numerous studies that have attempted to illuminate the pathogenic mechanisms underlying ASD, it is widely accepted as a behavioral disorder with strong genetic components, given its high heritability index [3]. According to Miles [4], we would not be able to understand the heterogeneity until we have found specific autism genes. Connecting the genetic and etiological data to the behavioral and phenotypic data is critical to making strides in discoveries and effective solution. Hence, a better understanding of phenotypic heterogeneity in autism itself would generate useful information for the study of etiology, diagnosis, treatment and prognosis of the disorder.

Cluster analysis is very useful in this context as it seeks to separate an unlabeled data set into a discrete set of "natural," hidden data structures [5]. Different cluster methodologies [6-10] have been introduced to analyze ASD phenotype features with the objective of obtaining more homogenous meaningful subgroups. A key phase in any clustering framework is feature selection (or extraction). The presence of irrelevant features or of correlations among subsets of features heavily influences the appearance of clusters. Feature selection methods attempt to globally remove irrelevant/redundant

features prior to clustering [11]. By viewing clustering as a multidimensional optimization problem of input features, we automatically remove redundant features by combining statistical correlation analysis with a two-phase subspace clustering framework. We apply our ensemble statistical and clustering model to analyze a population of ASD patients using a set of 27 ASD phenotype features that span the following categories: ASD-specific symptoms, cognitive and adaptive functioning, language & communication skills, and behavioral problems. This approach, as illustrated in Fig. 1, incorporates statistical techniques to validate and interpret the results.

This ensemble statistical and clustering model is an efficient and scalable solution that can be applied to analysis of features from any complex biomedical dataset characterized by high dimensionality with unknown underlying subgroupings. It is applicable to other disorders that exhibit a diverse range of heterogeneity.

2. METHODOLOGY

2.1. DATA

The study sample includes 208 ASD subjects that are part of the Simons Simplex Collection (SSC) [12]. (Simplex families have one child in the family with ASD with unaffected parents and siblings). Studying an SSC sample population guarantees that clinical and phenotype data is comprehensive, rigorous, reliable and consistent. It also makes replication for future studies easier since the entire dataset spans 12 different SSC sites and includes genotype data, which will be useful in translating these results to future genomic analysis. These 208 ASD subjects were recruited and diagnosed with ASD by the

University of Missouri (MU) Thompson Center for Autism and Neurodevelopmental Disorders (the Missouri SSC project site). ASD diagnoses were made using the Autism Diagnostic Interview – Revised (ADI-R) [13] and Autism Diagnostic Observation Schedules (ADOS) [14]. They also completed the SSC protocol, which included clinical, medical, behavioral, and family histories, physical, neurologic and dysmorphology examinations. The experimental procedures involving human subjects described in this paper were conducted under the guidelines and approval of Missouri University of Science and Technology (S&T) Institutional Review Board.

2.2. ENSEMBLE CLUSTERING AND STATISTICAL ANALYSIS MODEL

The ensemble model (Fig. 1) consists of five phases:

Data Processing: The goal of this phase is to convert the features to a normalized numeric representation. Missing values are common in medical data, even in a rigorous dataset such as the SSC project. Several strategies have been developed to address this problem in machine learning. A review of the literature [15] reveals that the efficacy of the proposed methods (mean, missing data imputation, k-nearest neighbors, etc.) depends strongly on the problem domain (e.g., number of cases, number of variables, missingness patterns), and thus there is no clear indication that favors one method over the others. In our data set, the number of missing values is very minor, hence we applied the mean technique: a known simplistic model. The missing values are replaced with their average across the entire sample for that specific feature. Each feature was normalized between 0 and 1, using known standard ranges for the feature, as guided by the ASD domain experts (S.K and T.N.T).

Correlation Analysis: We perform pairwise Pearson's correlations to quantify the level of correlation present among the input numerical features. This is useful to refine the input features by retaining features that exhibit low correlation among each other. The refinement of features based on the results from the correlation analysis should be guided by the application domain experts to pick an appropriate threshold level for the tolerable level of correlation suited for the application data.

Uni-dimensional Clustering: This is the initial phase of our feature-based subspace clustering. The uni-dimensional clustering is performed for each feature x^j in the data set (x denotes the sample and j is the feature index). We cluster the data considering each input feature by itself using the Expectation-Maximization (EM) algorithm [16]. EM algorithm assumes a Gaussian distribution of the dataset and assigns a probability distribution to each feature, which indicates the probability of it belonging to each of the clusters. EM can decide how many clusters to create by cross validation or one may specify a priori how many clusters to generate. We allow EM to determine the optimal number of clusters for each feature in this phase. The number of solutions generated is proportional to m number of input features.

At the end of uni-dimensional clustering, we have m clustering results. To rank the features in terms of their "goodness" in clustering the data, we apply an internal cluster validation index (Davies-Bouldin (DB) [17]) to the set of clustering solutions. We quantify the goodness of each feature based on its ability to singularly cluster the entire data into a set of clusters that mathematically demonstrate a strong degree of compactness within cluster and separation from other clusters. The DB index measures the average value of the similarity between each cluster and its most similar cluster given by:

$$DB = \frac{1}{\mathrm{NC}} \sum_{\mathrm{i}} \max_{\mathrm{j},\mathrm{j}\neq\mathrm{i}} \left\{ \left[\frac{1}{\mathrm{n_i}} \sum_{\mathrm{x}\in\mathrm{C_i}} \mathrm{d}(\mathrm{x}, \mathrm{c_i}) + \frac{1}{\mathrm{n_j}} \sum_{\mathrm{x}\in\mathrm{C_j}} \mathrm{d}(\mathrm{x}, \mathrm{c_j}) \right] / \mathrm{d}(\mathrm{c_i}, \mathrm{c_j}) \right\} \qquad (1)$$

where NC denotes number of clusters; C_i: the i–th cluster; n_i: number of objects in C_i; c_i: center of C_i; $d(x, y)$: distance between x and y.

A lower DB index implies a better cluster configuration. We remove indiscriminant features (i.e. the features that clustered the entire sample as one). Thus, the output of this phase is a ranked set of features with possibly fewer features.

k-dimensional Clustering: This phase clusters the data samples based on the results of the uni-dimensional phase using *k* of *n* total possible number of features. The dimension of the subspace is *k,* and *n* denotes the full space of features. The clusters are constructed by merging data samples that belong to same uni-dimensional clusters. The algorithm iterates through samples that shared the same uni-dimension clusters along *k* dimensions. If number of specified dimensions is *n*, then the clustering process iterates through all possible features: full-space clustering. In this phase, different subspace clustering methods can be applied based on the selected dimensions. The total possible number of subspace clusterings is $\sum_{i=1}^{n} \boldsymbol{P_i^n}$ where $\boldsymbol{P_i^n}$: denotes the permutations of *n* using *i* dimensions.

Clusters Evaluation: The *k*-dimensional phase yields different cluster configurations results. To determine which cluster configuration best fits the data, we evaluate the results of cluster analysis in a quantitative and objective fashion using a three step process: majority cluster validation indices voting, univariate analysis and multivariate analysis. During the uni-dimensional phase, for simplicity, we had used only a single cluster validation index (DB index) to rank the "quality" of the individual features. In this last phase of the model, to determine the optimal clustering solution we employ more than

one validation index. To ensure a robust model, we applied three validation indices (DB index, Silhoutte index, Calinski-Harabasz [18]) to the clustering results to measure the goodness of the clusters. We ranked the solutions based on the majority voting of the indices to determine the top three optimal clustering results. The goal of univariate analysis is to evaluate which input features are significantly discriminant among the resulting clusters.

A feature to determine if there are any significant differences in average values (based on the global F-test p-values) between clusters. A Fisher's exact test was employed to detect significant differences between clusters for nominal variables. The multivariate analysis [19] step involves performing a discriminant canonical analysis to determine which predictors (set of features) best discriminate between the clusters. For each clustering output, the following results are obtained:

Squared canonical correlation gives the proportion of variation explained in the cluster grouping variable (feature) for each canonical variable (discriminating function).

P-value to test for significance of each canonical variable (discriminating function). The number of canonical variables is the number of clusters – 1.

Pooled within-group correlations between each variable and the standardized canonical discriminant functions. The closer to +1 or -1, the more important the variable is in distinguishing the clusters.

Cross-validation error rate which is percent incorrectly classified with leave-one-out cross-validation.

3. EXPERIMENTAL RESULTS

3.1. ASD PHENOTYPE FEATURES AND CORRELATION ANALYSIS

The set of 27 ASD phenotype features selected as input features are listed in Table 1. We combined the CBCL (Child Behavior Check List) 2-5 scores with the CBCL 6-18 scores after normalizing them separately. We intentionally included features that possibly had strong correlations with other features to test the ability of our model to filter out redundant features. Some of the features are actually derived from the same measure such as the Full IQ, Verbal and non-Verbal IQ scores (F1-F3). The results of the correlation analysis on the 27 input features are presented in Table 1. We highlighted features that had a high pairwise correlation value of >0.7. Features in bold indicate a correlation value of ≥0.9.

3.2. FEATURE-BASED 2-PHASE SUBSPACE CLUSTERING RESULTS

Out of 27 features, 18 were selected by the EM uni-clustering phase. The 9 discarded features appear in Table 1. These exhibit a high level of similarity. Thus the resulting features for the k-dimensional phase had a lower level of correlation compared to the initial set. The unidimensional phase succeeded in filtering out most of the highly correlated features though the set of correlated IQ scores are left. However, full IQ score is ranked lowly by the uni-dimensional phase. (At this phase, one could also choose to manually further minimize the level of correlation by eliminating one from each pair of remaining highly correlated features.) Based on the ranking of the remaining 18 features by the DB cluster validation index, we grouped the features into 3 levels. Each level has 6

features, where level1 contains features {4, 1, 13, 8, 2, 11}, level2 = {14, 9, 12, 15, 6, 7} and level3 = {7, 16, 17, 3, 10, 18, 5}. The level1 features had the best validation index scores followed by level2 and level3. In the multi- dimensional subspace clustering phase, we performed 35 different subspace iterations by clustering the samples base on clusters in features of level1 vs. level2 and level3 and varying the number of difference allowed among the features. We also clustered using the full space (i.e. all features) as well as combinations of the levels.

3.3. CLUSTER EVALUATION

The top 3 clustering results (based on the cluster validation ranking of the 35 different clustering configurations, aided by the visualization using Principal Component Analysis) is in Table 2. The ranking was based on the majority voting of the 3 indices (DB, Silhouette & Calinski-Harabasz) though only the DB index value is shown in Table 2. The univariate analysis in Table 1 demonstrates that all the features except F27 exhibited significant differences (p-value<0.05) between clusters for at least one of the top three clustering results. This demonstrates that our model yields solutions that maximize variance of the entire set of features among the clusters though only a subspace of the features are used to cluster.

The multivariate discriminant analysis (MDA) in Table 2 reveals similar ranking of clustering configurations by the cross-validation error rate as those by the validation indices. Many of the features in the feature subspace were identified as being important ($\geq|0.3|$) for distinguishing clusters in the MDA, providing further validation of results. For clustering solution #1 (Table 2), 5/6 features in the subspace were identified as important in MDA;

whereas 7/12 and 4/7 were identified as important in discriminating clusters in MDA for clustering solutions #2 and #3, respectively. A plot of the canonical variable scores for output 1 (Fig. 2) is shown to assess how well the canonical variable is able to separate the clusters. Clustering solutions #2 and #3 have two canonical variables and biplots (Figs. 3 and 4) of the scores of these two variables are provided to assess individually how well each canonical variable does at separating the clusters. One can clearly see canonical variable 1 does a reasonable job of separating the clusters in all three clustering configurations obtained. The canonical plot (Fig. 1) biplots in Figs 2 and 3 visually demonstrate that clustering solution 1 is the best.

By applying a unified model that combines cluster validation indices with univariate analysis and MDA, we are able to confidently assess clustering solution 1 as the optimal clustering configuration for our data sample. Our cluster validation phase is enhanced beyond a cluster validation index value. This is a key contribution of our model.

Finally, we analyze the optimal clustering solution clinically to see if the results are meaningful. The SSC data collection project was completed prior to DSM-5, hence we have no information on severity gradient levels of these patients. We used a more quantitative variable: the ADOS Calculated Severity Scores (ADOS CSS) which is calculated separately from the ADOS social communication and RRBs scores. The ADOS CSS scores provide a continuous measure of overall ASD symptom severity that is less influenced by child characteristics, such as age and language skills, than raw totals [20]. These scores can be used to compare ASD symptom severity across individuals of different developmental levels. As such, they provide a “purer” metric of overall ASD severity. A higher level implies higher severity with 10 as the highest level of severity. As one can

observe from Table 3, cluster 2 (the smaller group) is distinct with a higher level of severity which corresponds to a lower overall IQ score (<70) and a lower Vineland II Composite score. Hence, our clustering model identified a distinct subgroup with higher severity levels compared to the overall sample population.

4. CONCLUSIONS

We applied a unique ensemble method, consisting of five stages of statistical and machine learning approaches, to achieve a subspace clustering of ASD data. The clustering results show promise for sorting out the heterogeneity that is characteristic of these patients. Multiple techniques were also combined for the validation of the identified clusters.

Table 1. Summary of ASD phenotype features; pairwise correlation; EM uni-dimensional clustering & univariate analysis.

Feature	***Description of Features***	***Correlates highly* with***	***EM Uni-dimensional Ranking†***	***Univariate Analysis†† p-values***		
				Output #1	***Output #2***	***Output #3***
F1	Overall Verbal IQ	F2, **F3**, F16, **F17**	2	<.0001	0.05	<.0001
F2	Overall Nonverbal IQ	F1, **F3**, F17	5	<.0001	0.0384	<.0001
F3	Full Scale IQ	**F1, F2**, F16, **F17**	15	<.0001	0.0195	<.0001
F4	Module of ADOS Administered		1	<.0001	0.5314	<.0001
F5	ADI-R B Non Verbal Communication Total		18	<.0001	0.0645	0.0013
F6	ADOS Communication Social Interaction Total	**F8**	11	<.0001	<.0001	<.0001
F7	ADI-R A Total Abnormalities in Reciprocal Social Interaction		12	<.0001	0.0025	0.0467
F8	ADOS Social Affect Total	**F6**	4	<.0001	<.0001	0.0001
F9	ADI-R C Total Restricted Repetitive & Stereotyped Patterns of Behavior		8	0.3475	0.0113	0.8393
F10	ADOS Restricted and Repetitive Behavior (RBB) Total		16	0.0001	0.0016	<.0001
F11	Repetitive Behavior Scale-Revised (RBS-R) Overall Score	F25	6	0.0346	<.0001	0.4622
F12	Aberrant Behavior Checklist (ABC) Total Score		9	<.0001	<.0001	0.2312
F13	Regression		3	0.1881	0.4774	0.006
F14	Vineland II Composite Standard Score	F15, **F16**, F24	discarded			
F15	Vineland II Daily Living Skills Standard Score	F14, F16	discarded			
F16	Vineland II Communication Standard Score	F1, F3, **F14**, F15,F24	discarded			
F17	Peabody Picture Vocabulary Test (PPVT4A) Standard Score	F1, F2, F3	7	<.0001	0.0183	<.0001
F18	Social Responsiveness Scale (SRS) Parent -Awareness Raw Score	F20, F23	discarded			
F19	SRS Parent - Cognition Raw Score	F23	discarded			
F20	SRS Parent - Communication Raw Score	F18, F22, **F23**	10	<.0001	0.0004	0.3525
F21	SRS Parent - Mannerisms Raw Score	F23	discarded			
F22	SRS Parent - Motivation Raw Score	F20, F23	discarded			
F23	SRS Parent Total Raw Score	F18,F19, **F20**,F21,F22	discarded			
F24	Vineland II Socialization Standard Score	F14, F16	discarded			
F25	RBS-R Subscale V Sameness Behavior	F11	13	0.162	<.0001	0.7986
F26	Child Behavior Checklist (CBCL) Internalizing Problems Total		14	0.1906	0.688	0.0005
F27	CBCL Externalizing Problems Total		17	0.5288	0.2127	0.0717

***pairwise correlation value >0.7. Features in bold have pairwise correlation value ≥0.9. †: Ranking determined by Davies-Bouldin (DB) validation index. ††: Analysis done for the top three clustering results.**

Table 2. Top 3 clustering configurations by validation index and multivariate discriminant analysis result.

Clustering Solution⁺	**Feature subspace (*k*-dimensional clustering)**	**DB Index†**	**# of Clusters**	**Multivariate Analysis Features important for distinguishing clusters**	**SCC[1]**	**CER[2]**
1	F6, F7, F9, F12, F17, F20	2.14	2	***Can 1***: **F6**, F8. F3, **F17**, F2, F1, **F7**, **F12**, F4, **F20**, F10	29.9%	8.65%
2	F1, F2, F4, F6, F7, F8, F9, F11, F12, F13, F17, F20	2.84	3	***Can 1***: **F6**, **F8**, F3, F11, **F12**, F25, **F20**, **F7**, F10 ***Can2***: **F13**	35.7% 9.7%	15.4%
3	F3, F5, F7, F10, F25, F26, F27	3.33	3 *	***Can 1***: **F3**, F2, F1, F17, **F10**, F4, F8, F6, **F26**, F13, **F5** ***Can2***: **F10**, F6	25.5% 13.2%	34.6%

Table 3. ASD severity analysis of clustering solution 1.

Cluster (size)	ADOS Calculated Severity Score* Mean (Std)	Overall IQ Score Mean (Std)	Vineland II Composite Mean (Std)
1 (189)	7.41 (1.58)	81.83 (27.55)	74.88 (11.82)
2 (19)	9.0 (1.12)	49.32(11.70)	61.47(8.17)

*ADOS SSC scores available for of patients due to age limit.. All three variables has mean values that were significant in differences between clusters: p<0.001 for all

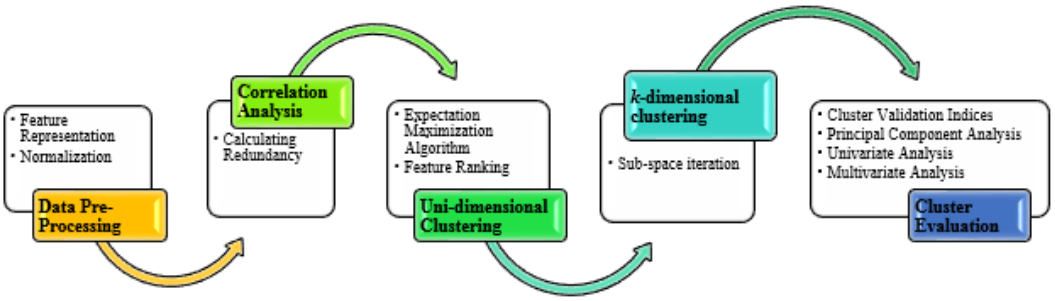


Figure 1. Overview of ensemble statistical and two-phase feature-based subspace clustering model. Note that this is a unique combination of various statistical and machine learning techniques.

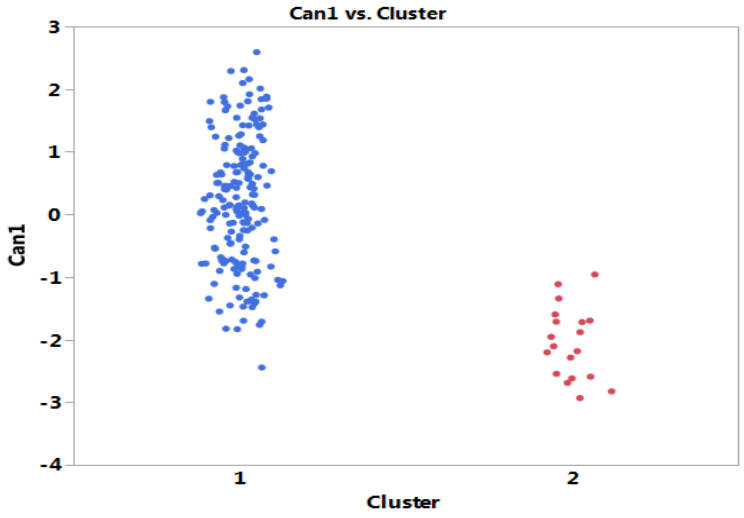


Figure 2. Plot of canonical variable for clustering solution 1. One can clearly see that canonical variable 1 does a reasonable job of separating the two clusters. Cluster 2 has mainly negative scores on Can 1, while cluster 1 has positive or negative scores closer to zero.

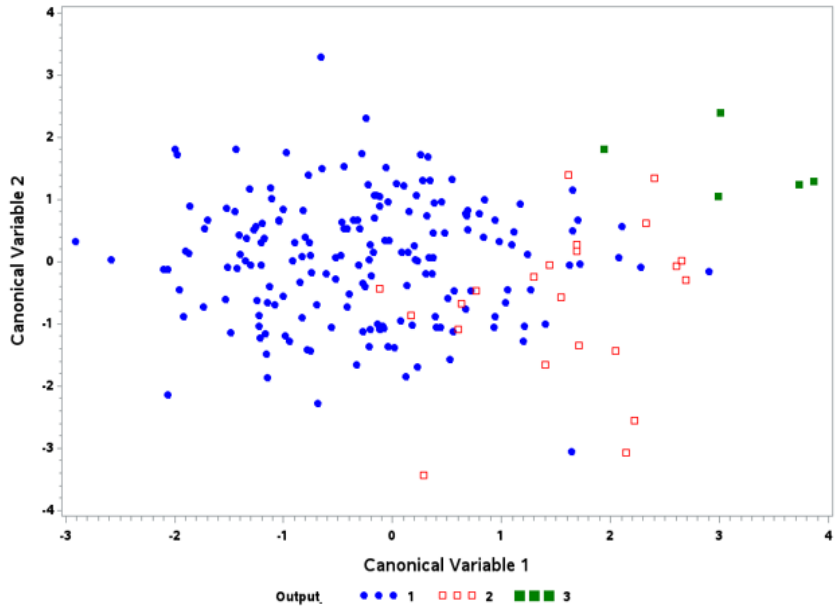


Figure 3. Biplot of the canonical variables scores vs. clusters for clustering solution 2. Canonical variable 1 does a reasonable job of separating the clusters. Canonical variable 2 does seem to be able to distinguish clusters 2 and 3 but neither can be distinguished from cluster 1.

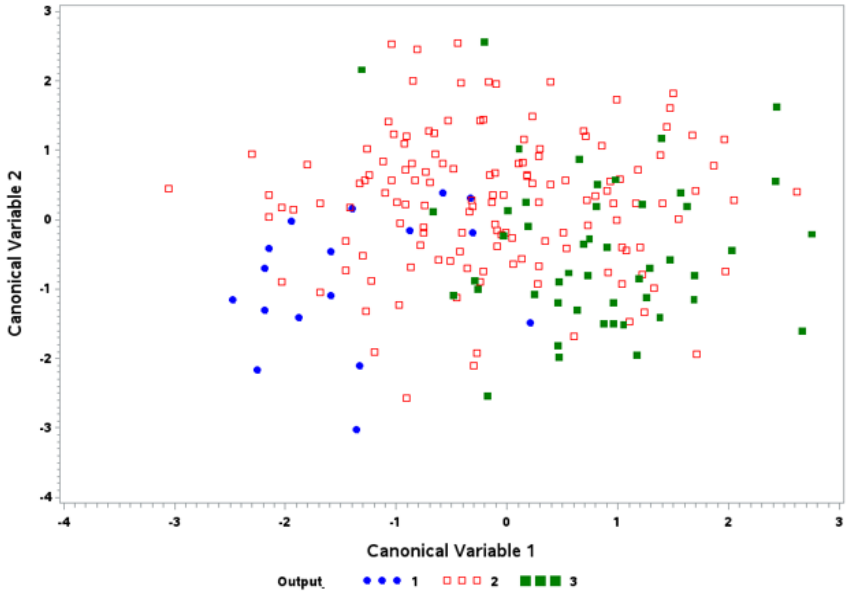


Figure 4. Biplot of the canonical variables scores vs. clusters for clustering solution 3. One can observe more overlap between the clusters in Canonical variable 1 in this plot compared to Fig. 3. We can observe that clustering solution 2 (Fig. 3) is indeed better.

REFERENCES

[1] Georgiades, S., Szatmari, P. and Boyle, M., 2013. Importance of studying heterogeneity in autism. *Neuropsychiatry*, *3*(2), pp.123-125.

[2] American Psychiatric Association, Diagnostic and statistical manual of mental disorders, 5th ed., American Psychiatric Publishing, 2013.

[3] Abrahams, B.S. and Geschwind, D.H., 2008. Advances in autism genetics: on the threshold of a new neurobiology. *Nature reviews genetics*, *9*(5), p.341.

[4] Miles, J.H., 2011. Autism spectrum disorders a genetics review. *Genetics in Medicine*, *13*(4), p.278.

[5] Xu, R. and Wunsch, D., 2008. *Clustering* (Vol. 10). John Wiley & Sons.

[6] Stevens, M.C., Fein, D.A., Dunn, M., Allen, D., Waterhouse, L.H., Feinstein, C. and Rapin, I., 2000. Subgroups of children with autism by cluster analysis: A longitudinal examination. *Journal of the American Academy of Child & Adolescent Psychiatry*, *39*(3), pp.346-352.

[7] Ingram, D.G., Takahashi, T.N. and Miles, J.H., 2008. Defining autism subgroups: a taxometric solution. *Journal of autism and developmental disorders*, *38*(5), pp.950-960.

[8] Georgiades, S., Szatmari, P., Boyle, M., Hanna, S., Duku, E., Zwaigenbaum, L., Bryson, S., Fombonne, E., Volden, J., Mirenda, P. and Smith, I., 2013. Investigating phenotypic heterogeneity in children with autism spectrum disorder: a factor mixture modeling approach. *Journal of Child Psychology and Psychiatry*, *54*(2), pp.206-215.

[9] Veenstra-VanderWeele, J., Christian, S.L. and Cook, Jr, E.H., 2004. Autism as a paradigmatic complex genetic disorder. *Annu. Rev. Genomics Hum. Genet.*, *5*, pp.379-405.

[10] Obafemi-Ajayi, T., Lam, D., Takahashi, T.N., Kanne, S. and Wunsch, D., 2015, August. Sorting the phenotypic heterogeneity of autism spectrum disorders: A hierarchical clustering model. In Computational Intelligence in Bioinformatics and Computational Biology (CIBCB), 2015 IEEE Conference on (pp. 1-7). IEEE.

[11] Alelyani, S., Tang, J. and Liu, H., 2013. Feature Selection for Clustering: A Review. *Data Clustering: Algorithms and Applications*, *29*, pp.110-121.

[12] Fischbach, G.D. and Lord, C., 2010. The Simons Simplex Collection: a resource for identification of autism genetic risk factors. *Neuron*, *68*(2), pp.192-195.

[13] Lord, C., Rutter, M. and Le Couteur, A., 1994. Autism Diagnostic Interview-Revised: a revised version of a diagnostic interview for caregivers of individuals with possible pervasive developmental disorders. *Journal of autism and developmental disorders*, *24*(5), pp.659-685.

[14] Lord, C., Rutter, M., Goode, S., Heemsbergen, J., Jordan, H., Mawhood, L. and Schopler, E., 1989. Austism diagnostic observation schedule: A standardized observation of communicative and social behavior. *Journal of autism and developmental disorders*, *19*(2), pp.185-212.

[15] Jerez, J.M., Molina, I., García-Laencina, P.J., Alba, E., Ribelles, N., Martín, M. and Franco, L., 2010. Missing data imputation using statistical and machine learning methods in a real breast cancer problem. *Artificial intelligence in medicine*, *50*(2), pp.105-115.

[16] Fraley, C. and Raftery, A.E., 2002. Model-based clustering, discriminant analysis, and density estimation. *Journal of the American statistical Association*, *97*(458), pp.611-631.

[17] D. L. Davies and D. W. Bouldin, "A cluster separation measure," Pattern Analysis and Machine Intelligence, IEEE Transactions on, vol. 2, pp. 224-227, 1979.

[18] Liu, Y., Li, Z., Xiong, H., Gao, X. and Wu, J., 2010, December. Understanding of internal clustering validation measures. In *Data Mining (ICDM), 2010 IEEE 10th International Conference on* (pp. 911-916). IEEE.

[19] Johnson, R.A. and Wichern, D.W. Applied Multivariate Statistical Analysis, 6th ed. Pearson, 2008.

[20] Hus, V., Gotham, K. and Lord, C., 2014. Standardizing ADOS domain scores: Separating severity of social affect and restricted and repetitive behaviors. *Journal of autism and developmental disorders*, *44*(10), pp.2400-2412.

III. NEUROIMAGING BIOMARKERS OF COGNITIVE DECLINE IN HEALTHY OLDER ADULTS VIA UNIFIED LEARNING

Tayo Obafemi-Ajayi, Khalid Al-Jabery, Lauren Salminen, David Laidlaw, Ryan Cabeen, Donald Wunsch and Robert Paul

ABSTRACT

Cognitive aging in healthy adults exhibits significant and heterogeneous variability. In this study, we apply a robust unified learning framework to cluster subgroups using neuroimaging data (brain volume and white matter), to identify neurological phenotypes that can sort out the heterogeneity in cognitive aging and help identify potential risk factors for suboptimal brain aging. Using machine learning analytics, results revealed two unique subgroups in healthy older adults with different patterns of white matter integrity and brain volumetric measures. The classification of phenotypical subgroups in healthy older adults may inform the understanding of the complexity of brain changes before the onset of clinical symptoms. The identified neuroimaging features that defined group classification are recognized as important structures that subserve cognitive performance. Further analysis of these potential biomarkers that help predict trajectory of cognitive decline in symptom free individuals could lead to the detection of early stages of neurodegenerative diseases.

1. INTRODUCTION

Machine learning models are very useful in the analytics and exploration of biomedical data. The goal is to discover unknown patterns or relationships that infer new

knowledge which can be further applied for prevention, prognosis and treatment [1], [2], [3]. A growing prolific area of research is the mining of heterogeneous medical data to identify quantifiable and objective criteria, known as biomarkers, for delineation of more homogeneous and meaningful subgroups. In this work, we apply machine learning techniques to address the problem of cognitive aging in otherwise healthy older adults.

Aging is very complex. It is marked by significant diversity across individual organisms, among organs and systems, and within organs cellular elements, particularly the brain [4]. Noninvasive neuroimaging techniques (magnetic resonance imaging (MRI)) and have greatly advanced the inquiry into age-related changes in human brain structure [5]. Numerous studies describe significant inter-individual variability in brain aging [6], [7], which is presumed to reflect discordant risk for age-related degenerative diseases, including Alzheimer's disease and subcortical ischemic vascular disease [8], [9], [10]. The phenotypic manifestations of age-related degenerative conditions reflect underlying structural changes to brain integrity and these anatomical alterations are evident using highresolution neuroimaging [11], [12].

There has been significant progress in understanding the structural changes that occur in the brain as we age [5], [6], [7], [13], [14]. The results of crosssectional studies reveal that gray matter volume declines linearly with age, whereas white matter volume increases during young age, plateaus during middle age, and declines during old age [13], [14]. However, additional work is needed to determine whether specific neuroimaging signatures (biomarkers) reflect unique underlying processes of brain aging among healthy older adults. Mining of neuroimaging data offers great potential to identify these biomarkers and discover subgroups in normal aging that could inform current models of

brain aging prior to the expression of clinical signs of age-related neurodegenerative symptoms.

This study investigates a unified learning model to sort out the heterogeneity associated with aging, identify viable neuroimaging biomarkers of aging and construct a prediction model applicable for further analysis. We examine variability in cognitive aging beyond traditional approaches that are focused on MRI analysis of specific or localized regions of the brain [7], [15] by employing unsupervised learning techniques. Previous studies have attempted to sort out the heterogeneity of aging by applying cluster analysis on data obtained from multiple neuropsychological evaluations [6], [7], [16]. In contrast, the hypothesis of this work is that analysis of neuroimaging data (specifically MRI brain volume and white matter fiber bundle integrity [17], [18]), using robust machine learning techniques will yield more objective results. The methodology presented here integrates statistical methods, outlier detection and removal as well as a rigorous feature selection process that uses structural MRI and diffusion tractography outcomes to identify subgroups of healthy adults that represent distinct aging phenotypes and potential biomarkers. In [19], we demonstrated an initial variation of this framework of an ensemble statistical and clustering model on the cluster analysis of autism spectrum disorders phenotype data. The model presented in this work is more robust and advanced to include a correlation filter algorithm and outlier detection for improved clustering results. The key strengths of this unified learning framework are: i) minimizing presence of irrelevant/redundant features that could bias cluster analysis using correlation filter algorithm; ii) robust iterative outlier detection and removal; iii) validation of results using both internal validation metrics and statistical techniques; iv) robust feature selection; and v) prediction model using supervised

learning algorithm. The integrated model is an efficient and scalable solution that can be applied to the analysis of complex biomedical features characterized by high dimensionality and a relatively small sample set to discover and validate underlying subgroupings.

The remainder of this paper is structured as follows. Section 2 provides a description of the unsupervised learning framework. The experimental setup and results obtained are presented in section 3. Section 4 provides an analysis and discussion of results obtain while we conclude in Section 5.

2. ROBUST UNIFIED LEARNING APPROACH

The unified learning framework, as illustrated in Fig. 1, consists of four phases with some iteration expected between them. These phases address the key challenges of medical data analytics. Data preprocessing (Phase 1) consists of steps to improve quality of data by handling the issue of missing values, data normalization, and redundancy among attributes/features. The dimensionality reduction strategies also deal with another characteristic of medical data: high dimensional data space corresponding to few examples or patients. Phase 2 consists of unsupervised learning (clustering) techniques to infer the underlying homogeneous subgroups present in the data based on the inherent structure. Clustering methods are applied, as this is the first line of defense to mining unlabeled data. This phase also addresses the issue of possible outliers in the data. Feature selection analysis (Phase 3) identifies key quantifiable biomarkers/key phenotypes that discriminate the subgroups. The feature selection phase presented in this work exploits multiple machine

learning based feature selection algorithms. It also includes an extensive statistical analysis to increase confidence in the results obtained and eliminate any erroneous features. The last phase is the prediction model, which involves training a model using the discriminant set of features (phase 3) as well as the "labels" learned from phase 2 to predict subgroup membership of the subjects.

Let **S** = $s_1, s_2, \ldots, s_n$ denote a set of *n* data objects to be clustered to obtain *k* partitions. Each data point s_i is represented by a set of *D* features: $f_1, f_2, \ldots, f_D$. The key phases of the approach are described below. It is particularly suited to the problem domain where the dimension of the feature space, *D*, is almost the same or much higher than the number of samples, *n*, available.

Phase1: Data Preprocessing

The set of *D* features are extracted from raw neuroimaging data consisting of brain volumes and white matter fiber bundle measurements (described in section 3). The data are normalized to transform these original measurements into a comparable format by mapping them as numeric representations ([0, 1]). Missing values among the measurements are replaced with the mean of the set of measurements across all examples [20]. As discussed in [19], missing values are a common phenomena in medical data. In the machine learning community, several strategies have been proposed to address this problem. A review of the literature [20] reveals that the efficacy of the proposed methods (mean, missing data imputation, k-nearest neighbors, etc.) depends strongly on the problem domain and thus there is no clear indication that favors one method over the others. Given that the data analyzed in this work has relatively few missing values, we employ the mean-value replacement technique. Highly correlated features are known to bias clustering outcomes,

therefore to eliminate redundancy among the input features, we employ the correlation filter algorithm described below.

Correlation Filter Algorithm: This involves applying statistical techniques to reduce the feature space f of size D to a minimal set $\{|f| = m, m < D\}$ to maximize the efficiency and effectiveness of the subsequent cluster analysis (phase 2). The algorithm identifies and filters highly correlated features using pairwise Pearson correlation function $r(fi, fj)$ based on a specified threshold value τ . The final minimal set of features employed is $\{f | \forall fi, fj \in f, r(fi, fj) < \tau\}$. The filter algorithm acts as a dimensionality reduction technique by reducing the original feature space.

Phase 2: Ensemble Clustering

The clustering analysis consists of two phases. The initial phase involves outlier detection followed by the actual cluster analysis, which can be applied using a single clustering algorithm or a combination of algorithms. In this work, we investigate algorithm selection and how to leverage learning opportunities to optimize clustering results using cluster validation techniques.

Outlier Detection: Outliers are known to significantly bias clustering results, when the underlying assumption is that every data point has to reside within a cluster. Ott et al. [21] discuss a framework in which outlier detection is integrated with clustering and modeled as an integer programming optimization task that requires prior knowledge of the number of outliers. However, such prior knowledge may not be feasible in certain applications such as the one presented in this work. Similar to [22], an iterative step is employed to identify possible outliers, if any, and exclude them from further cluster analysis. The technique applied further extends the process described in [22] based on

hierarchical agglomerative methods to improve its robustness using a combination of cluster validation metrics to determine the stopping criteria.

Cluster Evaluation/Validation: After excluding outliers, it is important to compare and exploit different algorithms, as they can vary significantly in performance. Clustering is a multidimensional optimization problem. For a single algorithm, multiple results can be obtained by varying different parameters. In this work, we employ K-means, K-medoids (also known as Partitioning About Medoids) and hierarchical clustering methods (ward, complete and average linkage) [23] and vary the number of clusters, *k* from 2 to 10. Validation indices assess the fit between the structure imposed by the clustering algorithm (clustering) and the data based on two main criteria: separation and compactness of clusters to determine the best fit for the data. To determine the optimal clustering configuration, we employ three commonly used internal validation metrics: Silhouette index [24], [25], Davies-Bouldin (DB) index [26] and Calinski-Harabasz (CH) index [27]. To ensure a robust model, the optimal configuration was determined by majority voting of the three cluster validation metrics.

Let k denotes number of clusters; *ki*: the ith cluster; *ni*: number of objects in *ki*; *ci*: center of *ki*; d(x, y): distance between x and y. The SI index is a composite index that measures both the compactness (using the distance between all the points in the same cluster) and separation of clusters (based on the nearest neighbor distance) as defined below.

$$SI = \frac{1}{k}\Sigma_i \left\{ \frac{1}{n_i} \sum_{x \in k_i} \frac{b(x) - a(x)}{(\max[b(x) - a(x)]} \right\} \quad (1)$$

where a(x): the average dissimilarity of x to all other objects in *ki*; b(x): minimum distance between x and all objects in other clusters (*kj* : *j* _= *i*).

The DB index computes the dispersion of a cluster and a dissimilarity measure between pairs of clusters given by:

$$DB = \frac{1}{Nc}\sum_i \max_{j, j\neq i}\left\{\frac{1}{n_i}\sum_{x\in C_i} d(x, c_i) + \frac{1}{n_j}\sum_{x\in C_j} d(x, c_j)\right\} \quad (2)$$

The CH index computes the between-cluster isolation and within-cluster coherence. It is defined as:

$$CH = \frac{\sum_i n_i d^2(c_i, c)/(k-1)}{\sum_i \sum_{x\in k_i} d^2(x, c_i)/(n-k)} \quad (3)$$

For both the SI and CH indices, a maximum value determines the optimal clustering configuration while for the DB index, the optimal is given by the minimum value.

Phase 3: Feature Selection Analysis

After estimating the number of meaningful subgroups in the data, the next step is to determine the most parsimonious set of discriminating features. We hypothesis that these neuroimaging signatures would represent potential biomarkers that inform the differences among the subgroups and identify potential risk factors for subsequent medical analysis. The subset of features obtained from this phase will be useful to build a reliable prediction model (phase 4). An ensemble feature selection approach is employed by combining the results of two feature selection algorithms: evolutionary feature selection method [28] and best-first selection method [29], along with feature ranking based on their information gain score [30]. To optimize the feature selection output, we incorporate the statistical techniques via univariate and multivariate discriminant analysis. The pooled within group correlation score between each feature and the standardized canonical discriminant

functions is employed as a filter to ensure that the final subset of features are indeed statistically significant and not over-fitted to the data. The closer to +1 or -1, the more important the feature is in distinguishing the clusters.

Phase 4: Prediction Model

Having employed a rigorous process to obtain a minimal set of discriminant features, the last phase in the unified learning model is to generate a supervised learning model to predict subgroup membership based on these features. We apply a known machine learning non-linear classifier - support vector machines (SVM) [31]. Given the small sample size, overfitting is a possible issue for any classification model. We apply two different cross-validation strategies and compare the results to minimize over-fitting and increase likelihood of generalization of classification model.

3. EXPERIMENTAL SETUP AND RESULTS

3.1. NEUROIMAGING DATA: BACKGROUND AND ACQUISITION

The neuroimaging data analyzed in the study were obtained from 71 cognitively normal individuals (male and female English-speaking adults older than 50 years) [14]. Individuals were excluded based on the following criteria: 1) Lifetime history of substance use disorder according to DSM IV criteria [32]; 2) Major psychiatric illness including schizophrenia, bipolar disorder, personality disorders, and clinically significant depression); 3) Medical/neurological disorders related to brain abnormalities; 4) Developmental disorders [32]; 5) Contraindications for MRI (metallic implants,

claustrophobia); and 6) Self-reported impairment in hearing or vision. Approvals were obtained from the local institutional review boards of all participating institutions.

Two types of neuroimaging data were analyzed in this study: structural MRI (sMRI) to quantify brain volumes, and quantitative tractography diffusion MRI (qtdMRI) to quantify white matter fiber bundle integrity [18], [17]. Brain volumetric data represents macrostructural measurements of brain integrity, particularly in subcortical regions. The qtdMRI is a novel imaging approach that combines diffusion tensor imaging (DTI) scalar metrics with tractography models to estimate bundle-specific properties such as average fiber bundle length (FBL), total fiber length (TL), and average scalar metrics across the fiber bundle. The primary advantage of qtdMRI compared to regional scalar metrics is the integration of data across a more complete biological system. Brain tractography models derived from DTI data can yield valuable insights about the topography and overall structural integrity of white matter [18].

All neuroimaging measurements were obtained from acquisitions conducted using a head-only Magnetom Allegra 3T MRI scanner at Washington University in St. Louis, MO. High-performance gradients (maximum strength 40 mT/m in a 100-ms rise time; maximum slew rate 400 T/m/s) were used to minimize scan times. Quality assurance was conducted daily to ensure data fidelity. Axial diffusion weighted imaging was acquired using a customized single-shot multislice echo-planar tensor-encoded pulse sequence. Thirty-one noncollinear diffusion-encoded directions were used in the acquisition consisting of 24 main directions. Pulse sequence and acquisition parameters were optimized for tractography, wide directional coverage as well as signal-to-noise ratio efficiency (see reference to parent study for further details). Individual DWI scans were

registered to the I0 image using FMRIBs Software Library (FSL) FMRIBs Linear Image Registration Tool (mutual information metric) to correct for subject motion, and the b-vectors were adjusted to account registration-induced rotation [14]. Brain tissue was extracted automatically using FSL's Brain Extraction Tool. Tensors and fractional anisotropy (FA) values were reconstructed by linear least squares with trilinear interpolation of the diffusion-weighted signal. Wholebrain streamline tractography was deterministic and performed with the principal eigenvector, one random seed per voxel, second-order Runge-Kutta integration, angle threshold of 35, FA threshold of 0.15, and a minimum-length threshold of 10 mm. The primary dependent variable from the diffusion sequence focused on average fiber bundle length (FBL) in each white matter tract of interest. (See parent study for details).

The bundles included in this analysis are the uncinate fasciculus (UNC), inferior fronto-occipital fasciculus (IFOF), superior longitudinal fasciculus (SLF), inferior longitudinal fasciculus (ILF), arcuate fasciculus (ARC), anterior thalamic radiation (ATR), superior thalamic radiation (STR), anterior commissure (AC), cingulum of the cingulate (CGC), cingulum of the hippocampus (CHC), the corticospinal tract (CST), fornix, and subdivisions of the corpus callosum (CC). These were modeled separately for each hemisphere for a total of 16 bundles. From all fibers found by whole brain tractography, a fiber was included in a bundle if 80% or more of its arc length was contained in the associated white matter mask. FBLs were quantified by combining DTI scalars with tractography models to estimate bundle-specific properties including the sum of fiber lengths within a given bundle (sumFBL), and the total weighted length (twl) of both FA

and mean diffusivity (MD). Average FBL was normalized according to total intracranial brain volume.

3.2. EXPERIMENTAL SETUP

Multiple experiments were performed using three different sets of data: sMRI measurements, qtdMRI measurements, and combined-qtdMRI measurements. Combined-qtdMRI measurements are the measurements obtained by summing all left and right hemisphere of the same variable in the qtdMRI data. For example, combined meanFA ARC is obtained by summing both meanFA ARC of the left hemisphere (lh) and meanFA ARC of the right hemisphere (rh) contained in the data. The objective of conducting experiments 1 (sMRI) and 2 (qtdMRI) separately was to determine if these varied feature sets would yield similar subgroups, and which set of features would be more effective in predicting individual subgroup membership. Experiment 3 (combined-qtdMRI) examines the hypothesis that summing left and right hemispheres of the white matter fiber bundle measurements is a natural dimensionality reduction method. Confirmation of this hypothesis would provide further evidence that cognitive aging in otherwise healthy adults is a symmetrical process. The results obtained for the different experiments are presented in the following section. The cluster validation indices were implemented using the using Cluster Validity Analysis Platform (CVAP) [33] while the feature selection and SVM algorithms were implemented using WEKA [34]. All statistical techniques were conducted in IBM SPSS Statistics 23. The software code developed for this unified learning model is available on our GitHub page (https://git.mst.edu/acil-group/MIMHaging).

3.3 EXPERIMENTAL RESULTS

Experiment 1: (Analysis of brain volumes using sMRI features). The correlation filter algorithm reduced the initial set of 51 features to 38 using a threshold value $\tau = 0.9$. The features removed included the right hemisphere measurements of the following regions of interest: lateral ventricle, temporal horn of lateral ventricle, cerebellum white matter, cerebellar cortex, thalamus, caudate. Total volumes of certain regions were also filtered such as total cortical gray matter volume, subcortical gray matter volume, total gray matter volume, supratentorial volume. Subsequent cluster analysis (Table 1) identified two distinct subgroups (Cluster 1, $n = 34$ and Cluster 2, $n = 36$) and one outlier. Results of the ensemble feature selection analysis revealed 19 features that discriminated the two subgroups. Cluster 1 is characterized by significantly higher mean brain volumes in the select regions listed in Table 1 compared to Cluster 2. These include both left and right of the hippocampus, amygdala, ventral diencephalon, pallidum, and general regions such as cortical white matter volumes, cortical gray matter, and brain stem. The differences of the means between the two clusters for these 19 features were all statistically significant, p-value < 0.001, according the *t*-test. Results of the supervised prediction model (Table 2) revealed a mean classification accuracy of 95.7% for the SVM classifier for both 5-fold and 7-fold cross-validation strategies.

Experiment 2: (Analysis of white matter fiber bundle using qtdMRI features). The correlation filter algorithm reduced the initial set of 186 features to 107 features using a threshold value $\tau = 0.8$. All the FBLsum features were filtered out leaving a subset of the mean FBL, mean and twl FA measurements along with the MD measurements. Similar to the volumetric outcomes, cluster analysis (Table 3) identified

two subgroups (Cluster 1, n = 31 and Cluster 2, n = 38) and two outliers. The ensemble feature selection analysis revealed 24 discriminating features with statistically significant differences in mean between both groups (student *t*-test pvalue: < 0.005). Cluster 1 has statistically higher mean FA and twl FA lengths compared to Cluster 2. However, cluster 1 had lower mean diffusivity (MD) value of superior thalamic radiation (STR) compared to Cluster 2. Similar ROIs between the sMRI features and the qtdMRI features include the corpus callosum. These were subsequently used to train and build the SVM classification model. The overall supervised mean classification accuracy for SVM was 91.3% for both models of cross-validation strategies.

Experiment 3: (Hemisphere analyses using combined qtdMRI features). The number of features obtained by summing the right and left hemisphere bundles resulted in a reduced set of 108 features. Applying the correlation filter algorithm, using the same threshold value as in experiment 2, further reduced it to 52 features. The cluster analysis (Table 4) also identified two subgroups (Cluster 1, *n* = 31 and Cluster 2, *n* = 40). The discriminant feature subset obtained from the feature selection analysis identified 19 features, of which 12 features were similar to the FBLs discriminant features. Similarly to experiment 2, Cluster 1 has statistically higher mean FA and twl FA lengths compared to Cluster 2. In terms of MD, cluster 1 had a lower value for the cingulate segment of the cingulum. Comparison of cluster membership of both qtdMRI results (experiments 2 and 3) revealed highly similar group membership. These results suggest that the mechanisms of brain aging impact both hemispheres without preferential involvement of lateralized brain systems. Results of the supervised prediction model also revealed higher classification accuracy of 97.1% for the 5-fold cross-validation runs and 98.6% for the 7-

fold crossvalidation runs. It appears the hemisphere classification model did not generalize well and over-fitted, hence the varying and overly high classification accuracy.

The resulting clusters per experiment were also evaluated in terms of demography information (Table 5). Paired student *t*-tests were completed to examine the significance of the measures between the resulting clusters. Differences in age, gender, and education were statistically significant for all results with the exception of education for the qtdMRI clusters. Fig. 2 presents the visualization of the multidimensional clusters obtained from cluster analysis for the multiple experiments.

4. DISCUSSION

Identification of subgroups in normal aging has potential to inform current models of brain aging prior to the expression of clinical signs of age-related neurodegenerative symptoms. The results from all three sets of experiments identified two distinct groups of normal aging, suggesting unique outcomes and predictors of these outcomes among these individuals. The subgroups with higher brain volumetric measurements (or higher white matter fiber bundle integrity measures) had significant lower mean age, were predominantly female and had significantly higher number of years of education. (The sample analyzed consisted of 64.8% female and 35.2% female.) Decline in cognitive function associated with advanced age is widely recognized and characterized by reductions in psychomotor speed, learning efficiency, and executive functions [35].

From the findings on the sMRI feature analysis, the individual neuroimaging features that defined group classification are recognized as important structures that

subserve cognitive performance. Prior work identifies the sensitivity of these regions to neuropathological mechanisms of aging. Specifically, the pallidum (subcortex) and the cerebellum (which is itself connected to the cortical frontal lobe and also critical for motor function) are critically known cognitive variables while the amygdala and nucleus accumbens represent “key” regions of emotional behavior. The neuroimaging phenotypes (Table 1) that discriminate these subgroups may aid understanding of substrates of age related diseases and help elucidate the complexity of brain changes before the onset of clinical symptoms.

The qtdMRI feature analysis (Table 3) provide additional evidence that both twl and mean FA lengths provide potentially non-redundant information about white matter integrity [18]. The subgroup defined by lower mean values in the features listed in Table 3 is consistent with the literature on cognitive aging and represents the more impaired cognitive group [14]. Deterioration in white matter micro-structure interferes with the synchrony and speed of neural communication, resulting in slowed responses and poor performances on cognitive tasks that require mental manipulation and attentional control. Baker et al. [14] revealed a significant inverse relationship between age and FBL in the ATR and UNC, both of which are important fiber pathways that intersect frontal brain systems known to be comprised with advanced age. In this study, we identified significant differences in the mean and twl FA of the CC, SLF, CGC, IFOF. The hemisphere analyses require more investigation and more in-depth comparisons of the features identified to further inform our knowledge of bilateral mechanisms of brain aging.

The main limitation of this study is the modest sample size of 71 individuals. A larger sample size could help with assessing outliers, further improve prediction accuracy,

and help determine or eliminate other possible causal factors that may not have been evaluated. The classification model could have suffered from over-fitting which may have inflated the accuracy of subgroup discrimination especially for the combined-qtdMRI-based classifier. The learning framework presented here included statistical measures at various levels to increase confidence that real phenomena are being identified, rather than artifacts of the learning approach. Nevertheless, larger data sets can further improve confidence in these results.

5. CONCLUSIONS

Cognitive aging in healthy adults exhibits significant and heterogeneous variability. In this paper, we designed and applied a robust unified learning framework to sort out the heterogeneity associated with aging, identify viable neuroimaging biomarkers, and construct a prediction model applicable for further analysis. Our results revealed unique and populous subgroups in healthy older adults with greater than 90% accuracy. We identified significant measurements that could potentially serve as biomarkers for delineating clinically meaningful aging subgroups.

Table 1. Discriminant features outcomes of cluster analysis using sMRI features.

Features[1]	Information Gain Ranking	Normalized Mean Volume (mm^3) (SD^2) Cluster 1 ($N = 36$)	Cluster 2 ($N = 34$)	Pooled within group correlations[3]
Cortical White Matter Volume (lh)	0.63	0.55 (0.15)	0.35 (0.15)	0.54
Cortical Gray Matter (lh)	0.52	0.71 (0.14)	0.52 (0.20)	0.52
Right Hippocampus	0.50	0.60 (0.21)	0.37 (0.20)	0.49
Left Ventral Diencephalon	0.49	0.55 (0.16)	0.36 (0.14)	0.67
Right Ventral Diencephalon	0.45	0.65 (0.15)	0.35 (0.15)	0.58
Left Hippocampus	0.43	0.69 (0.17)	0.45 (0.14)	0.48
Left Cerebellar Cortex	0.42	0.68 (0.14)	0.46 (0.14)	0.35
Brain Stem	0.41	0.60 (0.15)	0.28 (0.14)	0.61
Left Amygdala	0.39	0.63 (0.15)	0.36 (0.14)	0.50
Left Thalamus	0.38	0.62 (0.18)	0.36 (0.18)	0.52
Left Cerebellum White Matter	0.36	0.60 (0.23)	0.32 (0.20)	0.42
Left Putamen	0.36	0.39 (0.17)	0.23 (0.11)	0.41
Right Amygdala	0.31	0.46 (0.20)	0.25 (0.12)	0.45
Right Nucleus Accumbens	0.29	0.65 (0.15)	0.40 (0.16)	0.41
Right Pallidum	0.29	0.60 (0.16)	0.36 (0.12)	0.43
Central Corpus Callosum	0.24	0.65 (0.16)	0.39 (0.16)	0.35
Left Pallidum	0.23	0.60 (0.14)	0.42 (0.13)	0.43
Mid Anterior Corpus Callosum	0.23	0.49 (0.17)	0.30 (0.11)	0.39
Left Caudate	0.17	0.64 (0.15)	0.36 (0.15)	0.34

[1]All 19 features are statistical significance in difference in means between both clusters, p-value of two-tailed t-test < 0:001;
[2]SD: Standard Deviation.;
[3]Pooled within-groups correlations between discriminating features and standardized canonical discriminant functions extracted from structure matrix of discriminant analysis. The closer to +1 or -1 the more important that variable was in distinguishing the clusters. Features with absolute value < 0:3 is commonly deemed as less important.

Table 2. Classification performance of SVM prediction model.

		sMRI		qtdMRI		Combined-qtdMRI	
		Cluster 1	Cluster 2	Cluster 1	Cluster 2	Cluster 1	Cluster 2
SVM (5-fold CV[1])	Precision	0.919	1.0	0.903	0.921	0.968	0.975
	Recall	1.0	0.917	0.903	0.921	0.968	0.975
	Overall Accuracy	95.7%		91.3%		97.1%	
SVM (7-fold CV[1])	Precision	0.919	1.0	0.903	0.921	1.0	0.976
	Recall	1.0	0.917	0.903	0.921	0.968	1.0
	Overall Accuracy	95.7%		91.3%		98.6%	

[1] CV: Cross Validation

Table 3. Discriminant features outcomes of cluster analysis using qtdMRI features.

Feature[1,2]	Information Gain Ranking	Normalized Mean Volume (mm^3) (SD^3) Cluster 1 ($N = 31$)	Cluster 2 ($N = 38$)	Pooled within group correlations[4]
twl FA posterior CC	0.50	0.65 (0.16)	0.36 (0.15)	0.55
FA SLF (rh)	0.49	0.70 (0.17)	0.40 (0.21)	0.47
FA CGC (rh)	0.47	0.76 (0.15)	0.47 (0.21)	0.44
mean FA anterior CC	0.38	0.65 (0.13)	0.41 (0.14)	0.51
FA IFOF (lh)	0.38	0.66 (0.15)	0.41 (0.15)	0.50
Mean FBL Fornix (lh)	0.36	0.64 (0.21)	0.37 (0.19)	0.39
FA posterior CC	0.32	0.81 (0.09)	0.66 (0.19)	0.29
twl FA anterior CC	0.27	0.59 (0.20)	0.34 (0.19)	0.39
FA ARC (rh)	0.26	0.60 (0.19)	0.37 (0.16)	0.40
twl FA CGC (rh)	0.26	0.42 (0.20)	0.22 (0.12)	0.36
FA CGC (lh)	0.25	0.52 (0.18)	0.28 (0.16)	0.42
MD STR (lh)	0.25	0.36 (0.16)	0.60 (0.19)	-0.41
FA ATR (lh)	0.24	0.61 (0.14)	0.43 (0.19)	0.32
twl FA ARC (lh)	0.22	0.47 (0.15)	0.28 (0.18)	0.31
twl FA SLF (lh)	0.22	0.54 (0.22)	0.27 (0.18)	0.41
FA central CC	0.21	0.78 (0.10)	0.63 (0.12)	0.38
FA ILF (lh)	0.21	0.67 (0.18)	0.48 (0.17)	0.30
twl FA IFOF (lh)	0.21	0.51 (0.23)	0.28 (0.17)	0.34
mean FBL CGC (rh)	0.21	0.58 (0.14)	0.45 (0.18)	0.25
FA ARC (lh)	0.19	0.57 (0.15)	0.38 (0.20)	0.32
twl FA IFOF (rh)	0.15	0.29 (0.18)	0.18 (0.13)	0.21
FA UNC (rh)	0.15	0.59 (0.17)	0.40 (0.18)	0.31

[1]twl: total weighted length, FA: fractional anisotropy, FBL: fiber bundle length; MD: mean diffusivity; lh: left hemisphere, rh: right hemisphere, CC: corpus allosum, SLF: superior longitudinal fasciculus, CGC: cingulate segment of the cingulum, IFOF: inferior fronto-occipital fasciculus, ARC: arcuate fasciculus, STR: superior thalamic radiation ATR:anterior thalamic radiation, ILF: inferior longitudinal fasciculus, UNC: uncinate fasciculus; [2]All features listed are statistical significance in difference in means between both clusters, p-value of two-tailed t-test < 0:005; [3]SD: Standard Deviation; [4] Pooled within-groups correlations between discriminating features and standardized canonical discriminant functions extracted from structure matrix of discriminant analysis. The closer to +1 or -1 the more important that variable was in distinguishing the clusters. Features with absolute value < 0:3 is commonly deemed as less important.

Table 4. Discriminant features outcomes of cluster analysis using combined qtdMRI features.

Feature[1,2]	Information Gain Ranking	Normalized Mean Volume (mm^3) (SD^3) Cluster 1 ($N = 31$)	Cluster 2 ($N = 40$)	Pooled within group correlations[4]
twl FA posterior CC	0.47	0.66 (0.16)	0.36 (0.15)	0.52
MD CGC (rh+lh)	0.44	0.28 (0.16)	0.54 (0.18)	-0.42
FA anterior CC	0.40	0.66 (0.13)	0.41 (0.14)	0.50
FA ILF (rh+lh)	0.38	0.70 (0.18)	0.44 (0.17)	0.35
FA IFOF (rh+lh)	0.37	0.60 (0.19)	0.35 (0.17)	0.40
twl FA IFOF (rh+lh)	0.31	0.45 (0.20)	0.24 (0.16)	0.31
FA CGC (rh+lh)	0.29	0.70 (0.17)	0.40 (0.18)	0.45
twl FA ATR (rh+lh)	0.29	0.36 (0.20)	0.16 (0.12)	0.35
FA ARC (rh)	0.26	0.60 (0.19)	0.37 (0.16)	0.40
twl FA CGC (rh)	0.26	0.42 (0.20)	0.22 (0.12)	0.36
FA posterior CC	0.28	0.81 (0.10)	0.65 (0.21)	0.26
FA ATR (rh+lh)	0.25	0.67 (0.16)	0.43 (0.18)	0.39
twl FA anterior CC	0.25	0.60 (0.21)	0.34 (0.19)	0.37
FA mid posterior	0.25	0.75 (0.11)	0.57 (0.16)	0.36
twl FA CGC (rh+lh)	0.24	0.42 (0.21)	0.23 (0.10)	0.33
twl FA SLF (rh+lh)	0.23	0.61 (0.21)	0.41 (0.22)	0.27
FA ARC (rh+lh)	0.22	0.55 (0.16)	0.32 (0.19)	0.36
twl FA AC (rh+lh)	0.21	0.45 (0.25)	0.25 (0.13)	0.29
mean FBL ARC (rh+lh)	0.21	0.71 (0.14)	0.53 (0.20)	0.29
mean FBL central CC	0.16	0.81 (0.08)	0.71 (0.16)	0.21
mean FA AC (rh+lh)	0.13	0.57 (0.24)	0.39 (0.21)	0.22

[1] twl: total weighted length, FA: fractional anisotropy, FBL: fiber bundle length; MD: mean diffusivity; lh: left hemisphere, rh: right hemisphere, CC: corpus callosum, CGC: cingulate segment of the cingulum, ILF: inferior longitudinal fasciculus, IFOF: inferior fronto-occipital fasciculus, ATR: anterior thalamic radiation, SLF: superior longitudinal fasciculus, ARC: arcuate fasciculus, AC: anterior commissure; [2]All features listed are statistical significance in difference in means between both clusters, p-value of two-tailed t-test < 0:005; 3SD: Standard Deviation.; [4] Pooled within-groups correlations between discriminating features and standardized canonical discriminant functions extracted from structure matrix of discriminant analysis. The closer to +1 or -1 the more important that variable was in distinguishing the clusters. Features with absolute value < 0:3 is commonly deemed as less important.

Table 5. Demographics per cluster.

Demographics	sMRI ($N = 70$) Cluster 1	Cluster 2	qtdMRI ($N = 69$) Cluster 1	Cluster 2	qtdMRI (L+R hemispheres) ($N = 71$) Cluster 1	Cluster 2
Age	58.8 (5.68)	65.5 (8.78)	60.2 (5.76)	64.2 (9.12)	59.6 (5.77)	64.6 (9.05)
Gender (Female/Male)	18/16	28/8	16/15	30/8	15/16	31/9
Education (number of years)	16.1 (2.52)	14.7 (2.4)	15.9 (2.54)	14.8 (2.45)	16.1 (2.46)	14.7 (2.47)

All measures were statistically significant between clusters according to student t-test (p-value¡.05) with the exception of the education for the qtdMRI clusters.

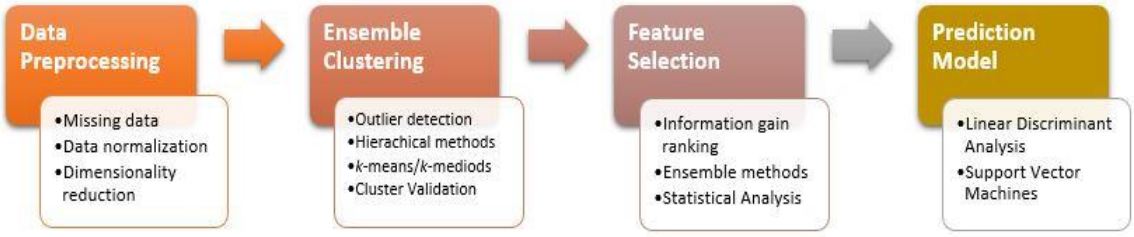
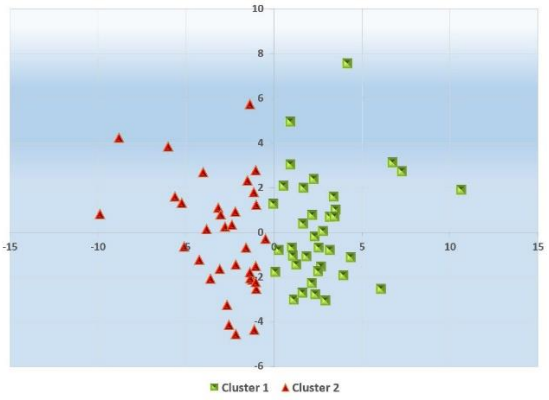
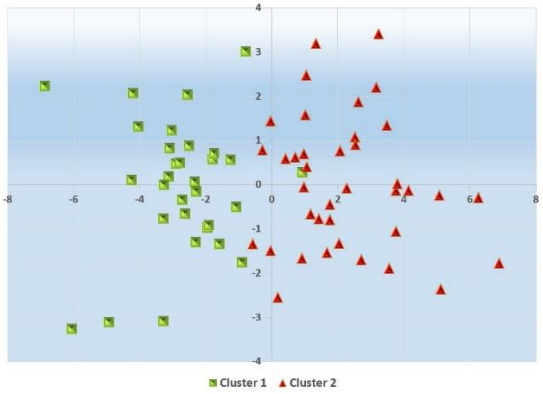


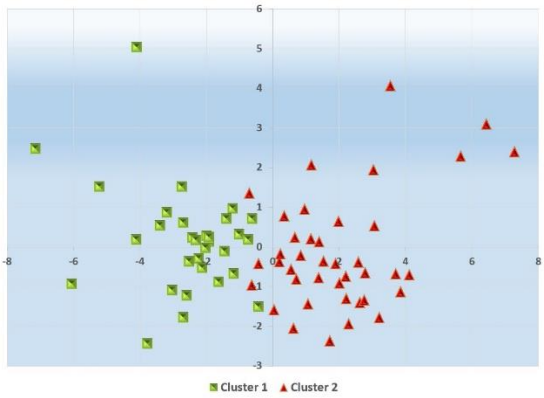
Figure 1. Overview of robust unified learning model for clustering unlabeled data, extracting statistically significant features and informing a prediction model.



(a)



(b)



(c)

Figure 2. Visualization of multidimensional clusters obtained from analysis of neuroimaging data of otherwise healthy adults using Principal Component Analysis (PCA). The first two axes of the PCA account for >95% of the variance among the features. (a) Clusters based on sMRI Features, (b) Clusters based on qtdMRI Features, (c) Clusters based on combined qtdMRI Features.

REFERENCES

[1] Bellazzi, R. and Zupan, B., 2008. Predictive data mining in clinical medicine: current issues and guidelines. International journal of medical informatics, 77(2), pp.81-97.

[2] Xu, R. and Wunsch, D.C., 2010. Clustering algorithms in biomedical research: a review. *IEEE Reviews in Biomedical Engineering*, *3*, pp.120-154.

[3] Fang, R., Pouyanfar, S., Yang, Y., Chen, S.C. and Iyengar, S.S., 2016. Computational health informatics in the big data age: A survey. *ACM Computing Surveys (CSUR)*, *49*(1), p.12.

[4] Raz, N., Rodrigue, K.M. and Haacke, E.M., 2007. Brain aging and its modifiers. *Annals of the New York Academy of Sciences*, *1097*(1), pp.84-93.

[5] Raz, N., 2001. Ageing and the brain. *eLS*..

[6] Foss, M.P., Formigheri, P. and Speciali, J.G., 2009. Heterogeneity of cognitive aging in Brazilian normal elderls. *Dementia & Neuropsychologia*, *3*(4), pp.344-351.

[7] Ardila, A., 2007. Normal aging increases cognitive heterogeneity: Analysis of dispersion in WAIS-III scores across age. *Archives of Clinical Neuropsychology*, *22*(8), pp.1003-1011.

[8] Gold, G., Kövari, E., Herrmann, F.R., Canuto, A., Hof, P.R., Michel, J.P., Bouras, C. and Giannakopoulos, P., 2005. Cognitive consequences of thalamic, basal ganglia, and deep white matter lacunes in brain aging and dementia. *Stroke*, *36*(6), pp.1184-1188.

[9] Gouw, A.A., Seewann, A., Vrenken, H., Van Der Flier, W.M., Rozemuller, J.M., Barkhof, F., Scheltens, P. and Geurts, J.J.G., 2008. Heterogeneity of white matter hyperintensities in Alzheimer's disease: post-mortem quantitative MRI and neuropathology. *Brain*, *131*(12), pp.3286-3298.

[10] Fein, G., Di Sclafani, V., Tanabe, J., Cardenas, V., Weiner, M.W., Jagust, W.J., Reed, B.R., Norman, D., Schuff, N., Kusdra, L. and Greenfield, T., 2000. Hippocampal and cortical atrophy predict dementia in subcortical ischemic vascular disease. *Neurology*, *55*(11), pp.1626-1635.

[11] Mungas, D., Jagust, W.J., Reed, B.R., Kramer, J.H., Weiner, M.W., Schuff, N., Norman, D., Mack, W.J., Willis, L. and Chui, H.C., 2001. MRI predictors of cognition in subcortical ischemic vascular disease and Alzheimer's disease. *Neurology*, *57*(12), pp.2229-2235.

[12] Mok, V.C., Liu, T., Lam, W.W., Wong, A., Hu, X., Guo, L., Chen, X.Y., Tang, W.K., Wong, K.S. and Wong, S., 2008. Neuroimaging predictors of cognitive impairment in confluent white matter lesion: volumetric analyses of 99 brain regions. *Dementia and geriatric cognitive disorders*, *25*(1), pp.67-73.

[13] Cabeza, R., Nyberg, L. and Park, D.C. eds., 2016. *Cognitive neuroscience of aging: Linking cognitive and cerebral aging*. Oxford University Press.

[14] Baker, L.M., Laidlaw, D.H., Conturo, T.E., Hogan, J., Zhao, Y., Luo, X., Correia, S., Cabeen, R., Lane, E.M., Heaps, J.M. and Bolzenius, J., 2014. White matter changes with age utilizing quantitative diffusion MRI. *Neurology*, *83*(3), pp.247-252.

[15] Grieve, S.M., Williams, L.M., Paul, R.H., Clark, C.R. and Gordon, E., 2007. Cognitive aging, executive function, and fractional anisotropy: a diffusion tensor MR imaging study. *American Journal of Neuroradiology*, *28*(2), pp.226-235.

[16] Ylikoski, R., 2000. The relationship of neuropsychological functioning with demographic characteristics, brain imaging findings, and health in elderly individuals.

[17] Cabeen, R.P., Bastin, M.E. and Laidlaw, D.H., 2016. Kernel regression estimation of fiber orientation mixtures in diffusion MRI. Neuroimage, 127, pp.158-172.

[18] Correia, S., Lee, S.Y., Voorn, T., Tate, D.F., Paul, R.H., Zhang, S., Salloway, S.P., Malloy, P.F. and Laidlaw, D.H., 2008. Quantitative tractography metrics of white matter integrity in diffusion-tensor MRI. *Neuroimage*, *42*(2), pp.568-581.

[19] Al-Jabery, K., Obafemi-Ajayi, T., Olbricht, G.R., Takahashi, T.N., Kanne, S. and Wunsch, D., 2016, August. Ensemble statistical and subspace clustering model for analysis of autism spectrum disorder phenotypes. In *Engineering in Medicine and Biology Society (EMBC), 2016 IEEE 38th Annual International Conference of the* (pp. 3329-3333). IEEE.

[20] Fraley, C. and Raftery, A.E., 2002. Model-based clustering, discriminant analysis, and density estimation. *Journal of the American statistical Association*, *97*(458), pp.611-631.

[21] Ott, L., Pang, L., Ramos, F.T. and Chawla, S., 2014. On integrated clustering and outlier detection. In *Advances in neural information processing systems* (pp. 1359-1367).

[22] Obafemi-Ajayi, T., Lam, D., Takahashi, T.N., Kanne, S. and Wunsch, D., 2015, August. Sorting the phenotypic heterogeneity of autism spectrum disorders: A hierarchical clustering model. In *Computational Intelligence in Bioinformatics and Computational Biology (CIBCB), 2015 IEEE Conference on* (pp. 1-7). IEEE.

[23] Xu, R. and Wunsch, D., 2008. *Clustering* (Vol. 10). John Wiley & Sons.

[24] Lam, D., Wei, M. and Wunsch, D., 2015. Clustering data of mixed categorical and numerical type with unsupervised feature learning. *IEEE Access*, *3*, pp.1605-1613.

[25] Rousseeuw, P.J., 1987. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, *20*, pp.53-65.

[26] Halkidi, M., Batistakis, Y. and Vazirgiannis, M., 2001. On clustering validation techniques. *Journal of intelligent information systems*, *17*(2-3), pp.107-145.

[27] Bolshakova, N. and Azuaje, F., 2003. Cluster validation techniques for genome expression data. *Signal processing*, *83*(4), pp.825-833.

[28] Kim, Y., Street, W.N. and Menczer, F., 2000, August. Feature selection in unsupervised learning via evolutionary search. In *Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 365-369). ACM.

[29] Kohavi, R. and John, G.H., 1997. Wrappers for feature subset selection. *Artificial intelligence*, *97*(1-2), pp.273-324.

[30] Guyon, I. and Elisseeff, A., 2003. An introduction to variable and feature selection. *Journal of machine learning research*, *3*(Mar), pp.1157-1182.

[31] Cristianini, N., and Shawe-Taylor, J., "An introduction to support vector machines and other kernel-based learning methods". Cambridge university press, 2000.

[32] A. P. Association et al., Diagnostic and statistical manual of mental disorders (DSM-5). American Psychiatric Pub, 2013.

[33] Wang, K., Wang, B. and Peng, L., 2009. CVAP: validation for cluster analyses. *Data Science Journal*, *8*, pp.88-93.

[34] Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P. and Witten, I.H., 2009. The WEKA data mining software: an update. *ACM SIGKDD explorations newsletter*, *11*(1), pp.10-18.

[35] Paul, R.H., Lawrence, J., Williams, L.M., Richard, C.C., Cooper, N. and Gordon, E., 2005. Preliminary validity of "integneuroTM": A new computerized battery of neurocognitive tests. *International Journal of Neuroscience*, *115*(11), pp.1549-1567.

IV. A DEEPER LOOK AT PLANT UPTAKE OF ENVIRONMENTAL CONTAMINANTS AND ASSOCIATED HUMAN HEALTH RISKS USING INTELLIGENT APPROACHES

Majid Bagheri, Khalid Al-jabery, Donald Wunsch, and Joel G. Burken

ABSTRACT

Emergence of new contaminants in the environment is prevalent due to the increased production, inadvertant spills, improper disposal, wastewater discharges, and reuse. These multiple sources of pollutants have the potential to impact vegetation and the food quality as a prominent part of the human exposome. Uptake of contaminants from the groundwater is one pathway of interest, and efforts have been made to relate root exposure to translocation throughout the plant, termed the transpiration stream concentration factor (TSCF). This work utilized machine learning neural networks (NN), fuzzy logic and clustering for predicting TSCF using physicochemical properties of compounds, and examining the interactions between compound properties. The NN predicted the TSCF with improved accuracy compared to mechanistic models. It also delevered new insight to compound properteis and their importance in transmembrane migration. The sensitivity anlysis indicated that Log K_{ow}, molecular weight, hydrogen bond donor, and rotatable bonds are the most significant properties. The results of fuzzy logic demonstrated that the relationship between molecular weight (MW) and Log K_{ow} with TSCF are both bell-shape and sigmoidal. Several clustering algorithms have been applied, and they all discovered two major distinct clusters. The clusters resulting from k-means algorithm were the clearest and only these are presented. Physiochemical property cutoffs, i.e. restrictions, for compounds passing plant roots membrane were shown to be lower than the cutoffs for

transmembrane transport in mammalian intestinal systems. Therefore, the human health impacts through consumption of contaminated crops is elucidated and indicated that plant roots are a restrictive barrier to organic pollutants entering our foods. Improved understanding and prediction of plant uptake has significant implications for human health as we continue to shorten our water cycles.

KEYWORDS

Emerging contaminants, plant uptake**,** food safety**,** human health, machine learning.

NOMENCLATURE

TSCF	transpiration stream concentration factor
Log K_{ow}	octanol/water partition coefficient
NN	neural network
MW	molecular wright
HBD	hydrogen bond donor
HBA	hydrogen bond acceptor
RB	rotatable bonds
PSA	polar surface area
R	correlation coefficient
MSE	mean squared error
DB	Davies-Bouldin

CH	Calinski-Harabasz
PCA	principal component analysis

1. INTRODUCTION

Our global reliance on anthropogenic organic compounds has risen exponentially and our ability to understand the growing human exposure to these molecules has not kept pace. The number of anthropogenic organic chemicals has expanded tremendously in: numbers, types and functions. The persistence and environmental fate of these molecules is often not well known nor predictable with current tools and knowledge. As a result, we are increasing societal exposure to a wide variety of organic compounds. Additionally, we are living in a closer proximity to each other and to many sources as we continue to urbanize globally.

To add to the human exosome, we are distinctly shortening the water cycle with our increased need for freshwater. The use of reclaimed wastewater had grown to be a notable source of freshwater. While many waters undergo advanced treatment to destroy recalcitrant organics, many other municipal wastewaters and agricultural run-off still contain considerable synthetic organic molecules. Irrigation of crops with reclaimed wastewater and application of agricultural chemicals have in part ameliorated water shortage problems and also enhanced agricultural productivity [1]. Along with the irrigation of reclaimed wastewater and application of agricultural chemicals the exposure of plants to these compounds has increased. In the case of food crops, the concern of food safety has increased. The concerns remain for food safety and contaminant exposure

mainly because some organic molecules can migrate across plants' root membranes [2]. The chemical contaminants in the soil have the potential of being transported to foliage through plant evapotranspiration. Thus, quantification and prediction of transmembrane migration and transport from soil to foliage has direct linkage to potential human health impacts. Russell and Shorrocks [3] introduced transpiration stream concentration factor (TSCF) to show the possibility of transporting a given chemical to foliage. TSCF is the ratio of a chemical concentration in the xylem sap to the concentration of that chemical in the solution used by the roots.

The high cost of experimental studies and inconsistent testing guidelines have resulted in generating few number of experimental TSCF values for a limited number of chemicals and plant species. The variability of the reported TSCF values for a given chemical and plant species is large due to the lack of consistent testing guidelines and difficulty in measuring metabolism and volatilization losses during the experiments [4]. The estimation of TSCF values for new contaminants not only helps to have predictive tools on efficiency of a specific molecule to be translocated by plants but also helps researchers focus efforts on contaminants with likely translocation capacity. Since 1974 several studies have been conducted to introduce a relationship between the physical properties of organic chemicals and their translocation in plants [4-7]. These studies introduced single-parameter relationships relating the TSCF to octanol/water partition coefficient (Log K_{ow}) which is as a term to describe hydrophobicity.

Due to the low precision of single-parameter relationships and the limitation of these relationships for applying to a wide variety of contaminants and plants species, researchers came up with more complicated prediction models for plant uptake and

translocation. The single-parameter relationships were outperformed with the advent of recent models correlating multiple compound properties to TSCF [1, 2]. More complex approaches are compartmental models [8-10], which consider more chemical and environmental properties, and also incorporate the complexity of uptake and translocation processes into their mechanistic relationships. These modeling approaches still have limited accuracy in many cases in spite of improving our understanding of plant uptake and translocation of contaminants. However, in the majority of these modeling efforts, the models are calibrated or verified with specific plan species and chemicals tested in the laboratory portion. Plant selection and experimental design can certainly impact findings and modeling efforts. The current field of contaminant uptake has developed a large data pool [1, 2, 11] that can be used to investigate the comprehensive data sets for uptake of a wide array of compounds, by a range of plants, and in multiple laboratory arrangements, thereby limiting impacts of any one arrangement or data set. The assessment of these large data agglomerations can be challenging given the complexity of the data, and thus needing advanced data assessment methods and tools.

Simulation has been a useful approach to deal with various problems in different fields of science and engineering [12, 13]. NNs and fuzzy logic have been used for monitoring, control, classification, and simulation of engineering and environmental problems in particular [14-17]. One of the most important application of NNs is function approximation. NNs have been used to study complicated ecosystems from activated sludge cultures to land use and land cover systems, in which many factors are acting together [18, 19]. The successful application of fuzzy logic NNs in providing accurate and practical models for these systems has outperformed traditional modeling [20]. Such

complicated systems can be compared with complexity of plant uptake and translocation of contaminants due to the interaction of plant roots with many subsurface factors.

In this work, a NN predicts the plant uptake and translocation of environmental contaminants. The NN uses physicochemical properties of compounds to assess past data collections and predict TSCF. The physicochemical properties of compounds are analyzed using statistical analysis to determine the importance of each property to the TSCF values. Fuzzy logic was used to examine the interactions between important physicochemical properties as predictors of experimentally-determined TSCF values. The results of fuzzy logic accurately project TSCF values and can be used for screening of chemicals that are uptaken or excluded by plants and offer insight on the concerns regarding human health due to plant uptake and translocation of environmental contaminants. Furthermore, in this study, clustering techniques were utilized to determine any distinct groups and hidden data structures in the comprehensive dataset. Another contribution of this work is providing a statistical estimate for the TSCF threshold of high and low plant uptake based on clustering validation indices.

In order to build the NN model, a comprehensive selection of TSCF data was compiled from published literature [5-7, 21-57]. The compound properties including Log K_{ow}, molecular weight (MW), hydrogen bond donor (HBD), hydrogen bond acceptor (HBA), rotatable bonds (RB) and polar surface area (PSA), were obtained from chemical structure databases. Table 1 shows compound properties used in this study for the neural network modeling.

2. BACKGROUND AND IMPLEMENTED APPROACHES

2.1. LIPINSKI'S RULE OF FIVE AND DRUG DEVELOPMENT

The advent of the high throughput screening method in drug discovery enabled researchers to screen a large number of drug-like compounds across in vitro assays. High throughput screening method utilizes robotics, data processing software, liquid handling devices, and sensitive detectors to quickly conduct millions of chemical or pharmacological tests [58]. Lipinski and his colleagues [59] introduced the rule of five, which can be considered as a primary step in high throughput screening and drug discovery. The rule of five, which focuses on psychochemical properties of compounds, says that an orally administrated compound is likely to be absorbed by intestine when molecular weight is less than 500 Da, log K_{ow} is less than 5, the number of hydrogen bond donors is less than 5, and the number of oxygen plus nitrogen atoms is less than 10. Application of a method like high throughput screening in study of plant uptake and translocation of emerging contaminants, and introducing a rule for compounds passing roots membrane improves understanding of the relation between contaminates in the environment and their possible risks for human health.

2.2. NEURAL NETWORK MODEL

A NN is a massively parallel distributed processor consisting of simple processing units that have the ability to learn from experience [60],[61]. The NN implemented in this work consists of three layers of interconnected neurons. A single-output NN with M neurons in the hidden layer is expressed by Eq. (1):

$$y(w,x) = \varphi_{out}(\sum_{i=1}^{M}(W_{i,out} \times x_i) + b_{out}) \quad (1)$$

where φ_{out} is the activation function of the output layer, $W_{i,out}$ is the weight between the *i*th neuron in the hidden layer and the output neuron, b_{out} is the bias of output neuron, and x_i is the output of each neuron in the hidden layer and is calculated by Eq. (2):

$$y_m = \varphi_h(\sum_{i=1}^{M}(W_{i,m} \times x_i) + b_m) \quad (2)$$

where φ_h is the activation function of hidden layer, *M* is the number of input parameters, $W_{i,m}$ is the weight between the *i*th input parameter and the *m*th neuron in the hidden layer, b_m is the bias of *m*th neuron in the hidden layer, and x_i is the *i*th input parameter.

In this study, a NN was used for the first time to predict the uptake and translocation of emerging contaminants in plants. The NN modeling was utilized due to its higher accuracy than current mechanistic models [8]. The input layer of the NN model consisted of six neurons for each of the six inputs (Log K_{ow}, MW, HBD, HBA, RB, and PSA). These molecular descriptors were considered as inputs of the NN because the properties have been cited as important parameters in the modeling of TSCF in recent studies [1, 2]. The network architecture that was implemented in this work is a feedforward NN. The dataset was divided randomly into three parts, 70% for training, 15% for testing, and 15% for validation of the NN model. The code was written in MATLAB R2014a. More information regarding various NNs can be found in [60]. The prediction performance of the NN model for the TSCF was measured using correlation coefficient (R) and mean squared error (MSE), as illustrated in Figs. 1-3.

2.3. STATISTICAL ANALYSIS

In this study, stepwise and forward regression were used to determine the most important predictors in modeling TSCF. Forward regression is the simplest data-driven selection approach in which one predictor is added to the model at a time. Forward regression starts with no predictors in the model and calculates the p-value (a criterion for selection) for all predictors not in the model [62]. It adds the predictors with p-value less than p-critical, and this process is repeated until no new variable can be selected and added to the model. Stepwise regression is a modification of forward regression. In this method, a predictor may be added or removed from the model at a time. Like forward regression, it starts with no predictors in the model and calculates the p-value for all predictors not in the model. After adding a predictor to the model, all variables in the model are examined to find if their importance has been reduced to a predefined limit [63]. The predictors in the model with importance less than a predefined limit are removed from the model. Forward and stepwise regression were performed using SPSS 16.0 software in order to determine the significant predictors.

2.4. FUZZY LOGIC

Fuzzy logic is based on degrees of truth rather than true or false logic that assigns values to an imprecise spectrum of data in order to solve problems. The fuzzy approach is efficient addressing different types of uncertainties associated with environmental problems. Fuzzy logic is distinguished from familiar approaches such as Boolean algebra due to its ability to present results in the form of recommendations[16, 64]. Fuzzy logic is conducted through four main steps: fuzzification, generating fuzzy rules, generating a

fuzzy inference system, and defuzzification. More discussion on fuzzy logic can be found in [65-67].

In this study, an adaptive neuro-fuzzy inference system was used to examine simultaneous impacts of compound properties on the uptake and translocation. Fuzzy logic was utilized to screen the capacity of chemical compounds for uptake and translocation in plants through correlating TSCF values with various compound properties at the same time, using MATLAB R2014a.

2.5. CLUSTERING ALGORITHMS

Clustering algorithms are used to explore and reveal hidden relationships and structures in data [68]. In this work, two types of clustering algorithms have been applied: hierarchical and partitional clustering.

Hierarchical clustering provides divisions for the data on all levels, from singleton clusters to a cluster that contains the entire dataset. Partitional clustering divides data samples into a pre-specified number of clusters (partitions) regardless of their hierarchy (further discussion on these approaches can be found in [68]). To compare between the clustering algorithms used in this study, internal validation indices (see Section 2.6) and visualization-based methods were used. The best clusters according to the validation indices and visual assessment were obtained by k-means [69] (See Section 3.4). In fact, when using majority voting in internal validation indices, both k-means (k=2) and hierarchical clustering were the same. They both isolated a distinct group of samples that do not share any variables with the remaining dataset. However, the second in rank was the k-means (k=3), where the results further divided the large cluster into two distinct and

visually-isolated clusters, as illustrated in Section 3.4. Furthermore, in this study, we provided an estimate for a threshold that classifies the samples with high TSCF from those that have low TSCF and provide an estimated definition for the terms “High” and “Low” TSCF. We have tested the data divisions using multiple TSCF’s thresholds [0.4, 0.5, 0.6, 0.7, 0.8 and 0.9], as discussed in Section 3.4.

2.6. CLUSTER EVALUATION AND VISUALIZATION

In this paper, we have used two approaches for evaluating clusters: internal validation indices and visual representation. The internal validation indices are Silhouette (S), Davies-Bouldin (DB), and Calinski-Harabasz (CH). The Silhouette index [70] evaluates the clustering performance based on the pairwise difference of between and within cluster distances. The optimal cluster number is determined by maximizing the value of this index. The Davies-Bouldin index [71] is calculated as follows. For each cluster C, the similarities between C and all other clusters are computed, and the highest value is assigned to C as its cluster similarity. Then the DB index can be obtained by averaging all the cluster similarities. The smaller the index is, the better the clustering result is. The Calinski-Harabasz index [72] evaluates the cluster validity based on the average between and within the cluster sum of squares. The mathematical formulas and further discussion on the above cluster validation indices can be found in [68] or [73].

The second approach is the visualization-based approach. In order to provide a descriptive visualization, principal component analysis (PCA) visualized the resulting clusters from different algorithms using the first two components, as illustrated in Section 3.4. PCA is the process of data transformation from higher dimensions to lower dimensions

without losing a significant amount of information [60, 74]. Usually the first three variables of the resulting matrix represent more than 90% of the original data.

3. RESULTS AND DISCUSSION

3.1. OPTIMAL NEURAL NETWORK ARCHITECTURE AND PERFORMANCE

In this paper, the best results were achieved after 30 epochs using a NN with 50 neurons in its hidden layer. Table 2 illustrates the different performance evaluations for the various NN architectures. The experiments showed that increasing the interaction between input parameters including Log Kow, MW, HBD, HBA, RB, and PSA did not notably increase the accuracy of the NN prediction. The data was partitioned into 70% for training, 15% for validation and 15% for testing. The validation subset is used to validate the NN model during training to avoid overfitting and to ensure NN generalization. The test subset is used to assess the performance of NN.

The NN learned an acceptable fit between predicted and measured values, as the regression line for measured values of TSCF (Target) and predicted values of TSCF (Output) were in agreement. The Correlation levels between predicted and original values were approximately 0.83, 0.73 and 0.79 for training, validation and testing respectively, as illustrated in Fig. 1. The minimum calculated MSE was 0.0159 and 0.059 for training and testing respectively. These results indicate that the NN has adapted well during training and can be used as a general model for predicting TSCF using the given parameters, whereas traditional models were proven to be inaccurate for some compounds [1].

The results of the model for training, validation, and testing datasets have been plotted versus the frequency of data in Fig. 2b. The determination of a normal distribution of residuals is important, as such distribution is one of the assumptions for regression analysis [75] and validates the lack of bias in NN model predictions.

The NN is able to predict the plant uptake and translocation of environmental contaminants with higher accuracy than traditional modeling approaches [1]. The testing model demonstrated a positive prediction performance in comparison with traditional models, and the MSE value for this model was 0.0372 vs 0.25 to 0.3 for traditional models[1].

The NN learned to predict the TSCF from the compound properties, and the multi-parameters model captured the general changes in TSCF values with an acceptable accuracy and no notable oscillations, as shown in Fig. 3. Earlier modeling approaches reliant on a single parameter have shown inconsistency in predicting TSCF values, notably for hydrophilic compounds [1]. The results of previous studies show that the models did not capture the general changes of output and were unreliable. In contrast, the implemented NN model demonstrated a good generalization ability for learning and predicting plant uptake and translocation efficiency, which originates from the inherent ability of these approaches in function approximation and considering multi-parameter complexity.

The accuracy of the model may be improved if a dataset were expanded to include a more uniform range of records for the input parameters. These results were achieved despite a lack of uniform datasets, in part resulting from inconsistent data reporting and testing guidelines for experimental studies. Two recently published works offer recommendations for experimental studies to follow guidelines in order to achieve a strong

dataset for the modeling of uptake and translocation [1, 76]. Expansion of the current datasets will certainly advance the ability of current and future modeling methods and offer more advanced *in-silico* capabilities to project plant uptake of organic molecules.

3.2. SENSITIVITY ANALYSIS FOR PREDICTORS

Stepwise and forward regression methods were used to analyze the sensitivity of TSCF to the input parameters including Log K_{ow}, MW, HBD, HBA, RB, and PSA. The results of sensitivity analysis were highly correlated for both methods. Both selection methods confirmed that Log K_{ow} is the most important predictor. With a p-value less than 0.05 and t-statistics equal to 6.36, the Log K_{ow} is a compound property with highly significant impacts for the plant uptake and translocation models. The molecular weight and hydrogen bond donor were the second and third significant predictors for the plant uptake and translocation as functions of compound properties. The t-statistics value for MW and HBD was 1.86 and 1.61, respectively, as illustrated in Table 3. The results of this study were in line with the findings of previous studies trying to determine the importance of physicochemical predictors. Using a desirability function indicated that Log K_{ow}, MW, and HBD are the three important predictors in the modeling of plant uptake and translocation of organic contaminants [2]. Based on the results of our regression methods, the Log K_{ow} is eight times more governing than MW as a predictor. The MW is also three times more important than HBD. The results of the stepwise and forward regression methods show that the RB is the fourth most effective predictor. In a recently published paper, Millar et al. [1] used a desirability model for plant uptake of pharmaceutical and personal care products. They found that the number of rotatable bonds is another important

predictor. The findings of this study indicate that RB is a predictor with almost the same importance as HBD in the modeling of plant uptake and translocation of contaminants.

3.3. SIMULTANEOUS IMPACTS OF COMPOUND PROPERTIES

The results of fuzzy logic for examining the interaction between compound properties improved our understanding of plant uptake and translocation problems. Previous studies introduced a bell-shape relationship between Log K_{ow} and TSCF [2]. The most readily translocated compounds were relatively hydrophobic compounds with Log K_{ow} between 1 and 4. The MW and TSCF had a bell-shape relationship, and the MW was less 350 Da for the most trans-locatable compounds. Fig. 5 demonstrates the relationship between both Log K_{ow} and MW with TSCF using an adaptive neuro-fuzzy inference system. The results of our study also confirm that generally there are bell-shape relationships between both Log K_{ow} and MW with TSCF. However, the interaction between Log K_{ow} and MW is important and affects the translocation of compounds. It was observed that for the MW less than 120 Da, the relationship between Log K_{ow} and TSCF is sigmoidal instead of bell-shape. For the MW from 150 to 400 Da, the relationship between Log K_{ow} and TSCF is bell-shape with high possibility of translocation. For the MW higher than 400 Da, the relationship between Log K_{ow} and TSCF is still bell-shape with lower possibility of translocation for compounds. The results of this study indicated that for the Log K_{ow} less than 1, the relation between MW and TSCF is sigmoidal instead of bell-shape. The compounds have a high tendency for translocation in plants when the MW is less than 120 Da and Log K_{ow} less than 1. Previous studies have reported the translocation of hydrophilic compounds in addition to the moderately hydrophilic compounds [37, 77]. The

zones (MW vs. Log K_{ow}) with high possibility of uptake and translocation have been determined using oval shapes in Fig. 4. The first zone includes the Log K_{ow} from 1 to 3, and MW from 150 to 400 Da (hydrophobic compounds). The second zone includes Log K_{ow} values less than 1, and MW less than 120 Da (hydrophilic compounds). The findings of previous studies using the desirability function show a sigmoidal relationship between the hydrogen bond donor and TSCF. The desirability function demonstrated that compounds with HBD equal or less than 5 have a higher possibility for uptake and translocation [1, 2]. Fig. 5 demonstrates the relationship between both Log K_{ow} and HBD with TSCF using an adaptive network fuzzy inference system. In a good agreement with the results of previous studies, the Log K_{ow} has a bell-shape, and HBD has a sigmoidal relationship with TSCF. The zones (HBD vs. Log K_{ow}) with high possibility of uptake and translocation have been determined using oval shapes in Fig. 6. The results show that compounds with 5 or less HDB have a high potential for uptake and translocation. The compounds with 1 or 2 HBD have the highest capacity for uptake and translocation in plants.

The desirability function also demonstrated that there is a sigmoidal relationship between rotatable bonds and TSCF. The desirability function showed that compounds with 7 or less RB are more likely to be translocated in plants. The results of our study using fuzzy logic conform the sigmoidal relationship between RB and TSCF (Fig. 6). It was observed that compounds with 15 or less RB are more likely to be translocated by plants. Moreover, for this range of RB, the compounds with Log K_{ow} less than zero have a higher possibility of translocation than compounds with Log K_{ow} higher than 4. The value of RB for high possibility of translocation from this study is a little different from the findings of

desirability functions. Although our database is the most comprehensive database ever used for modeling of plant uptake and translocation of environmental contaminants, there are limited number of compounds with RB higher than 15. Thus, with increase in the number of published data for all chemical compounds, the accuracy of modeling approaches increases accordingly. The zones (RB vs. Log K_{ow}) with a high possibility of uptake and translocation are shown using oval shapes in Fig. 6.

3.4. RESULTED CLUSTERS AND TSCF THRESHOLD ESTIMATION

As discussed earlier, this paper presented the results from applying multiple machine learning algorithms and statistical tools on the comprehensive dataset. Since the compiled dataset includes a considerable number of various compounds and plants, clustering algorithms were used to examine the chemical compounds and plants that have more distinct behavior than others.

Hierarchical and partitional clustering have been applied. All the applied algorithms have isolated six samples sharing almost no variables with the remaining dataset (see cluster3 in Fig. 7). However, statistical metrics (see Section 2.6) and visual-evaluation select the most descriptive clusters. The selected clustering criteria was using k-means algorithm when k=3, as shown in Fig. 7 and Table 4.

The k-means algorithm clustered the data into three distinct groups with different features. Cluster 3 was distinct from cluster 1 and cluster 2 in terms of chemical compounds and physicochemical properties. The compounds in cluster 3 have Log K_{ow} values higher than 4, and an RB value of more than 21 to 32. The compounds in cluster 3 have shown a low capacity for uptake and translocation with TSCF values mainly less than 0.05. Such

compounds have a small capacity to cross the membrane of the plant roots and mainly accumulate in the root [7].

This paper also presented a statistical estimation for the optimum threshold for dividing the dataset into ‘’High’’ and ‘’Low’’ TSCF. The method divides the samples into two groups using different TSCF values and evaluates the resulting clusters using the internal validation indices discussed in Section 2.6. Majority voting was used to choose the optimum threshold. Both DB and CH selected a value of TSCF=0.6 as the best threshold to classify the given data into two classes, as illustrated in Table. 5. Therefore, we adopted this threshold to classify the given 300 data points into “High” and “Low” uptake and translocation.

The graphical representation for the isolated groups is shown in Fig. 8. The x-y plane in Fig. 8 was formed from the 1^{st} and the 2^{nd} PCA while the altitude was formed from the TSCF values.

The clusters analysis also showed interesting results for tomato, wheat, and corn as three plant species, which are used in many uptake experimental studies. Based on different thresholds of uptake considered in this study (0.4, 0.5, 0.6, 0.7, 0.8, and 0.9), the tomato and wheat mostly were in the group with higher uptake and translocation capacity than the considered limit. The corn almost in all cases was in the group with less uptake and translocation capacity than the considered limit. Therefore, it is recommended to start uptake studies with tomato, wheat, and then corn, if there is not enough data about the uptake capacity of a given compounds.

3.5. PLANT UPTAKE AND HUMAN HEALTH

Many studies have demonstrated that wastewater treatment plants with conventional activated sludge systems do not remove many environmental contaminants such as pharmaceutical and personal care products. Additionally, water shortage in many regions all over the world has encouraged the use of reclaimed wastewater as a source for crops irrigation. The uptake and translocation of pharmaceutical and personal care products can be considered as a potential threat to human health. Lipinski et al., [59] studied the uptake of pharmaceuticals by human intestine. They indicated that an orally administrated compound is likely to be absorbed by human intestine if the compound has log $K_{ow} < 5$, MW < 500 Da, HBD < 5 , and HBD <10. Several other studies investigated uptake of central nervous system drugs by blood brain barrier and developed similar rules [78]. The results of this study show that physicochemical properties for pharmaceuticals that pass plant roots membrane system are lower than Lipinski's cutoffs for the human intestinal system, as presented in [2]. Thus, the pharmaceutical and personal care products that pass plant root membrane are likely to be absorbed by the human intestine.

Environmental contaminants such as pharmaceutical and personal care products could be uptaken by plants and translocated to all parts of them, such as fruit. Humans may be at risk of long term low level exposure to many compounds through consumption of contaminated crops. For those compounds, even if the direct toxicological risk for human health is minimal through consumption of contaminated crops, they can still increase or decrease plant hormones or other endogenous plant compounds that can jeopardize human health [1].

Giving antibiotics in animal food as supplements for promoting growth of food animals is common in husbandry. However, due to incomplete absorption of antibiotics in the animal gut, a part of antibiotics ends up in manure through excreting urine and feces [79]. Using manure as a source of plant nutrients and organic matter to improve soil quality has increased the exposure of plant roots to antibiotics. Many patients do not fully use antibiotics prescribed by their doctors because they get better before finishing them. This has led to more and more antibiotics ending up in the environment due to lack of suitable disposal methods. Amoxicillin, erythromycin, levofloxacin, norfloxacin and tetracycline are examples of reported antibiotics found in the environment. According to their physicochemical properties and considering the findings of this study, they all have a high possibility of uptake and translocation by plants. Along with improper use of antibiotics by humans, chronic exposure to antibiotics through consumption of contaminated fruits and vegetables should be investigated for its possible contribution to enhanced antimicrobial resistance as a global problem threat.

3.6. BROADER IMPACTS AND CONTRIBUTION OF THIS WORK

An improved understanding of emerging and fugitive contaminants in plants provides benefits across a number of disciplines. The transport of organic compounds across biological membranes is an area of interest in mammalian systems. The ability of drugs to cross biological membranes (intestinal membrane, and blood brain barrier) is an important issue for drug development. The absorption efficiency of newly invented drugs is tested on expensive laboratory mice before human use. Thus, replacing mice in drug development with plants is interesting since it has financial benefits, and also saves the

lives of mice. Our study is a base for application of plants in drug development since it indicates that the compounds crossing plant roots membrane can also cross the intestinal membrane and be absorbed by humans. As explained in previous sections, the similarities of mass transfer in biological barriers and plant roots also shed light on the human exposure to contaminants in the environment. The findings of this paper motivate further study of the human health risks associated with production of new chemicals.

The development in biological sciences has led us to believe that organisms may be optimally designed or are evolved to optimize their tasks [80]. The cardiovascular system for example works so optimally that blood circulates through a network of vessels (arteries) throughout the body to provide individual cells with oxygen and nutrients and to dispose metabolic wastes (veins). Similarly, the transport of water and nutrients in plants through xylem and phloem has allowed growth in height and colonization of diverse habitats [81]. Despite many studies on the cardiovascular system [82, 83], it is still challenging. Modeling it improves our understanding of physiology and the interactions among the driving factors. Murray [83] introduced a theory for optimal cardiovascular design that solves for and predicts the sizes of blood vessels. McCulloh et al. [84] indicated that these conduits conform to Murray's law under some assumptions by measuring plant xylem. Due to the possible similarities between xylem and phloem in plants with the cardiovascular system, modeling of water and nutrient transport in plants via xylem and phloem improves our understanding of blood transport in the cardiovascular system. A model can be developed for plants, which can improve understanding of the cardiovascular system by:

1. Utilizing high and rapid screening methods to monitor transport of compounds in xylem and phloem, and generate data for model building.
2. Applying machine learning techniques
3. Considering mechanisms involved in transport of compounds, using generated data.

Vascular plants develop an extensive subsurface root system and an expansive aerial network of leaves with a tremendous surface area [85]. Plants have been used as indicators of their surrounding environments for millennia. That can be traced back to Roman times when willow and poplar stands indicated the presence of a shallow groundwater table and a good location for placement of drinking wells [86]. Prediction of translocatable compounds using machine learning opens the road for using plants as biosensors of subsurface contamination, termed as phytoforensics introduced by Burken et al. [85], to delineate contaminants from past and present. Finding the concentrations of contaminants in trees enabled us to sustainably delineate plumes at numerous sites, allowing more effective remediation strategies at lower cost. This study also impacts phytoremediation as a low cost technique to remove contaminants from soils, sediments, surface water and groundwater using plants [87, 88]. Predicting the capacity of compounds for uptake and translocation by plants gives useful information regarding the applicability and efficiency of phytoremediation in contaminated sites.

The number of emerging and fugitive contaminants in the environment threating human health is increasing. For many of these environmental contaminants there is no data available. Thus, studies were performed to find the possible exposure of human to new emerging contaminants through uptake and translocation by plants. Such experimental

studies require a lot of time and money. The introduced model in this research is a tool to approximate the uptake and translocation capacity of new contaminants by plants before conducting any research. In fact, the model can direct the efforts of researchers to work on contaminants, which are more probable to jeopardize human health.

4. CONCLUSIONS

The introduction of mathematical relationships relating compound properties to the possibility of compounds uptake and translocation in plants has received considerable attention since 1974. The previous studies introduced one dimensional relationships to relate Log K_{ow} to TSCF. However, these relationships are limited to specific compounds and plant species, and are inaccurate in some cases. To the best of our knowledge, this study is the first utilizing machine learning algorithms (i.e. NN, Fuzzy logic and unsupervised learning) to predict TSCF and examine the interaction between compound properties. The NN predicted the TSCF with high accuracy using physicochemical properties of compounds, and also offered insight to the impact and interactions of the different parameters. Log K_{ow}, molecular weight, hydrogen bond donor, and rotatable bonds were proved to be the most important compound properties in governing uptake of organic molecules. The findings of this study indicate that Log K_{ow} is the most significant property, molecular weight is the second most important, and hydrogen bond donor and rotatable bonds are two properties with similar level of importance. The results of fuzzy logic indicated that interaction between compound properties is an important factor to consider. The Log K_{ow} and molecular weight had both bell-shape and sigmoidal

relationships with TSCF, not just a bell-shape one. Clustering algorithms also revealed previously undiscovered structures in the dataset. The findings of this study shows the importance of hydrophilic compounds in addition to the moderately hydrophobic compounds. This study has also provided an estimation based on statistical evaluation metrics for the TSCF threshold. The impacts of plant uptake and translocation of environmental contaminants on human health should be considered seriously, since the cutoffs for compounds passing plant roots membrane are lower than cutoffs for drugs absorbed by the human intestinal system.

ACKNOWLEDGEMENTS

This work was supported by National Science Foundation under Award Number 1606036, the Mary K. Finley Endowment, and the Missouri S&T Intelligent Systems Center. The authors declare no competing financial interests.

Table 1. Characteristics of measured parameters used in the NN modeling process.

Input variable	**Min-Max**	**Output parameter**	**Min-Max**
Log K_{ow}	-2.19–6.75	**TSCF**	0.001-1.16
MW (g/mol)	32–616.4		
HBD	0–6		
HBA	0–16		
RB	0–36		
PSA (A^2)	0–196.2		

Table 2. Results of the MLPANN models with different architectures.

Number of neurons	R	MSE
5	0.75458	0.0460
10	0.73438	0.0486
15	0.78087	0.0412
20	0.78375	0.0405
30	0.73693	0.0481
40	0.78842	0.0377
50	0.79107	0.0372
60	0.78723	0.0379

Table 3. Results of compounds sensitivity analysis.

Selection method		**Log K_{ow}**	**MW**	**HBD**	**RB**
Stepwise	P-value	0.001	0.044	0.107	0.258
	t-statistics	6.368	1.861	1.617	1.133
Forward	P-value	0.001	0.045	0.109	0.258
	t-statistics	6.364	1.855	1.614	1.133

Table 4. Evaluation for clusters resulted from using different clustering algorithms.

Algorithms	**k-means**			**k-medoids**			**Hierarchical Clustering**		
# of clusters K=	**2**	**3**	**4**	**2**	**3**	**4**	**2**	**3**	**4**
Davies-Bouldin	**0.5495**	0.8625	1.1012	1.0383	1.0860	0.8226	**0.5495**	0.9588	0.9007
Silhouette	**0.8712**	0.5299	0.4925	0.4382	0.4309	0.4814	**0.8712**	0.4682	0.4702
Calinski-Harabasz	82.454	**176.490**	145.003	159.145	161.318	161.334	82.454	144.869	129.087

Table 5. Evaluation for clusters resulted from choosing different TSCF threshold using internal validation indices.

TSCF threshold	**0.4**	**0.5**	**0.6**	**0.7**	**0.8**	**0.9**
Silhouette	0.046639	0.021414	0.019997	-0.035096	-0.065708	-0.093345
Davies-Bouldin	4.0954	4.7232	**3.7488**	4.2451	5.2696	6.978
Calinski-Harabasz	12.096	8.7088	**12.357**	7.844	3.5053	1.2195

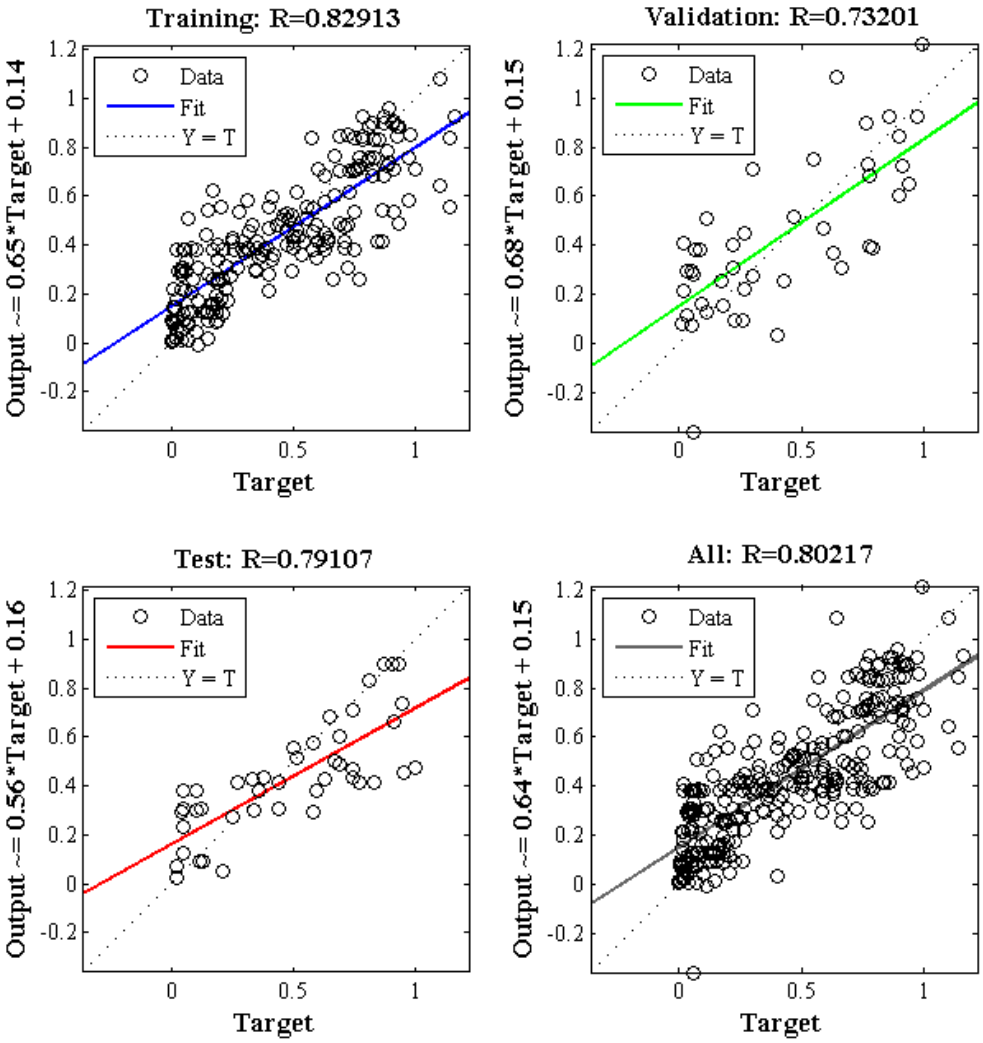


Figure 1. Regression plots of the MLPANN model for training, validation, testing, and all datasets.

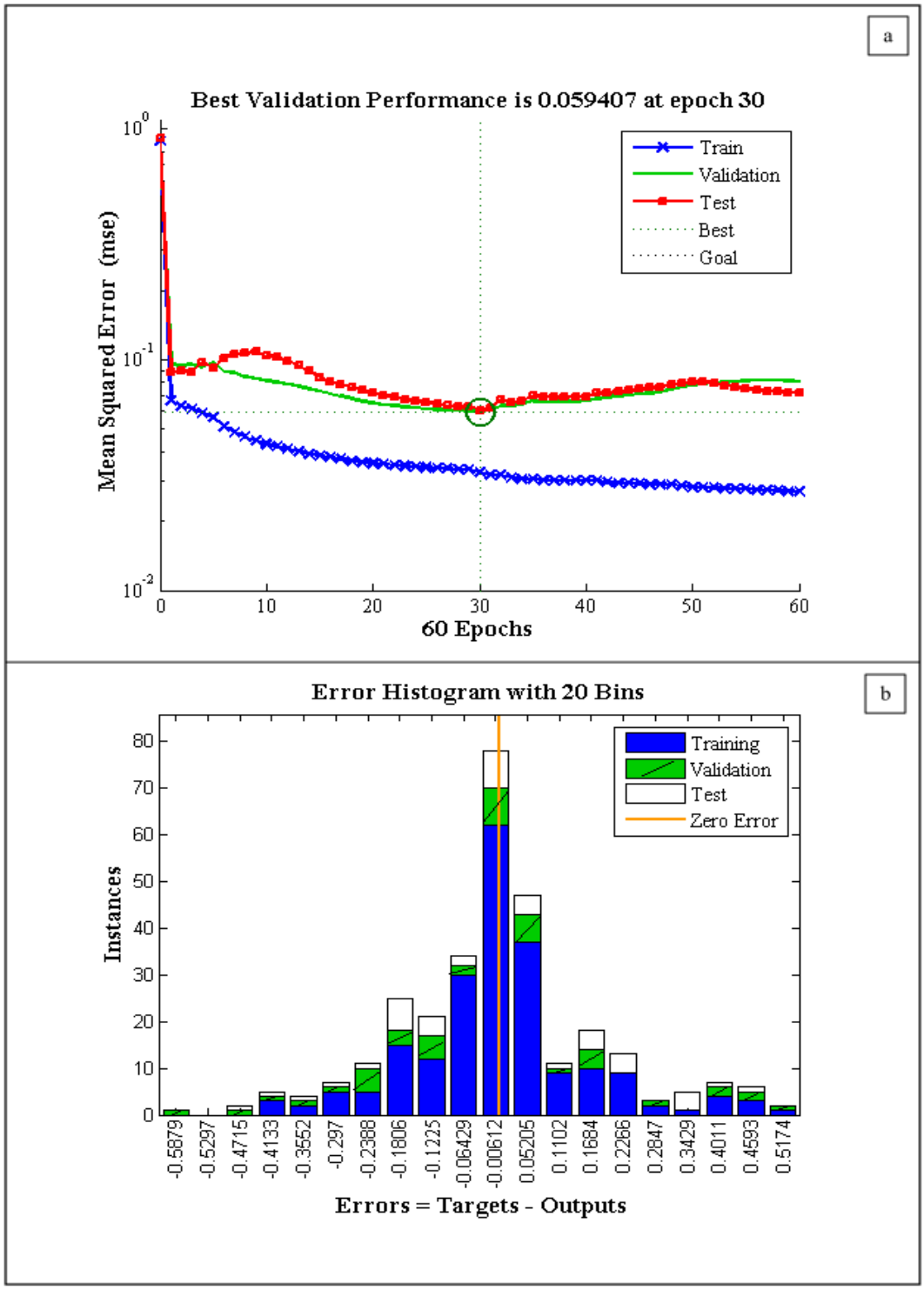


Figure 2. Performance of the training, testing, and validation model (a), and residual of the MLPANN models (b).

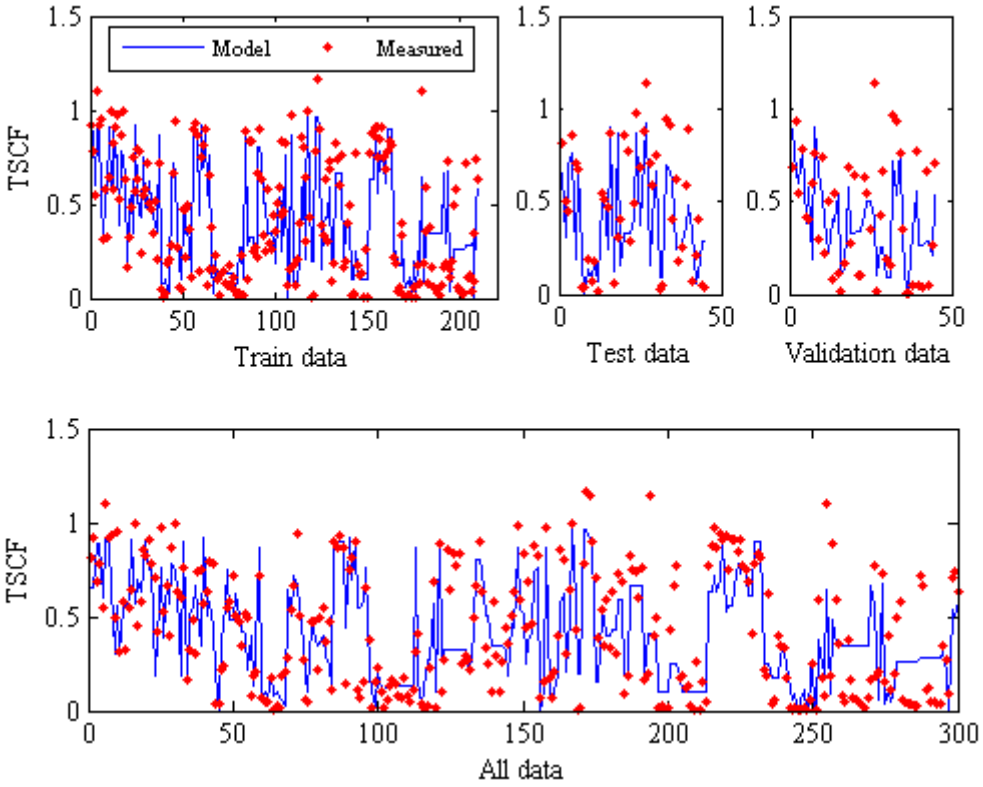


Figure 3. Results of intelligent MLPANN model for predicting TSCF based on training, validation, testing, and all dataset.

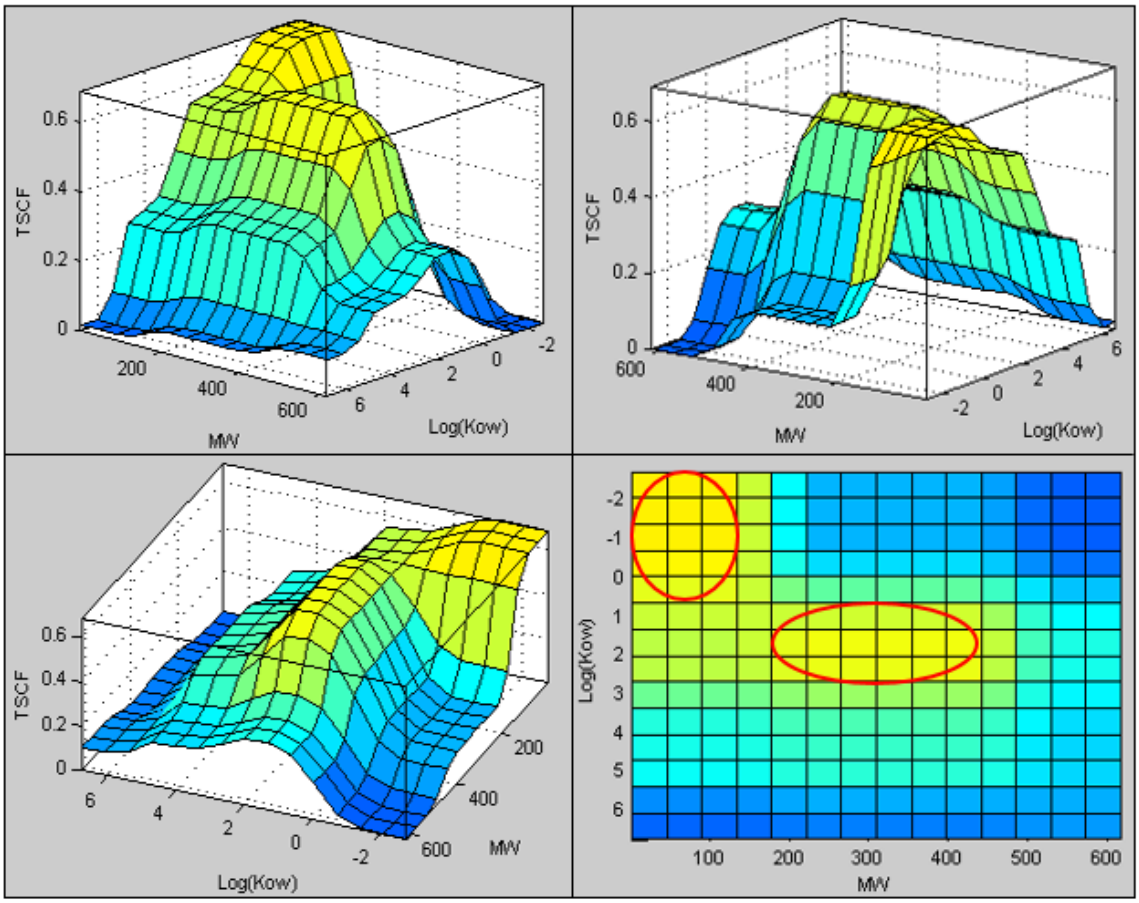


Figure 4. Relationship between Log K_{ow} and MW with TSCF using adaptive network fuzzy inference system.

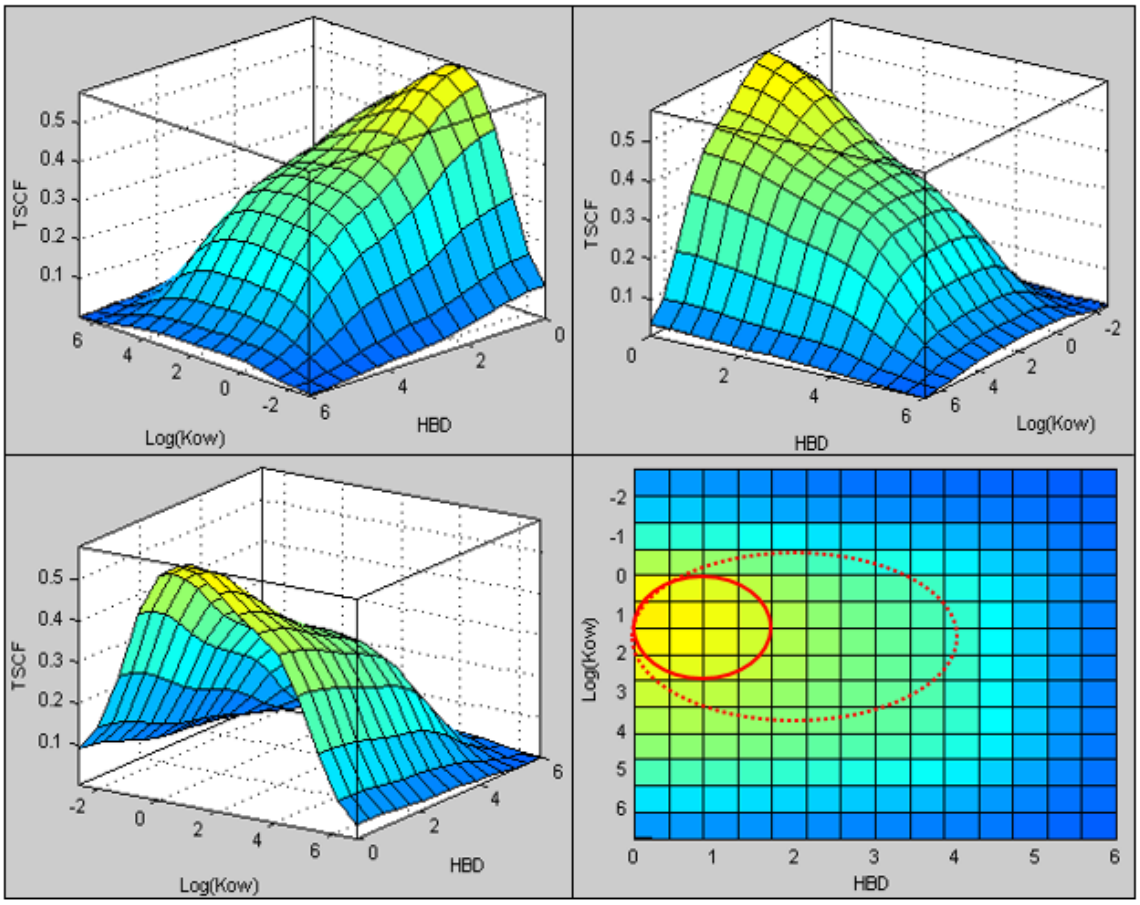


Figure 5. Relationship between Log K_{ow} and HBD with TSCF using adaptive network fuzzy inference system.

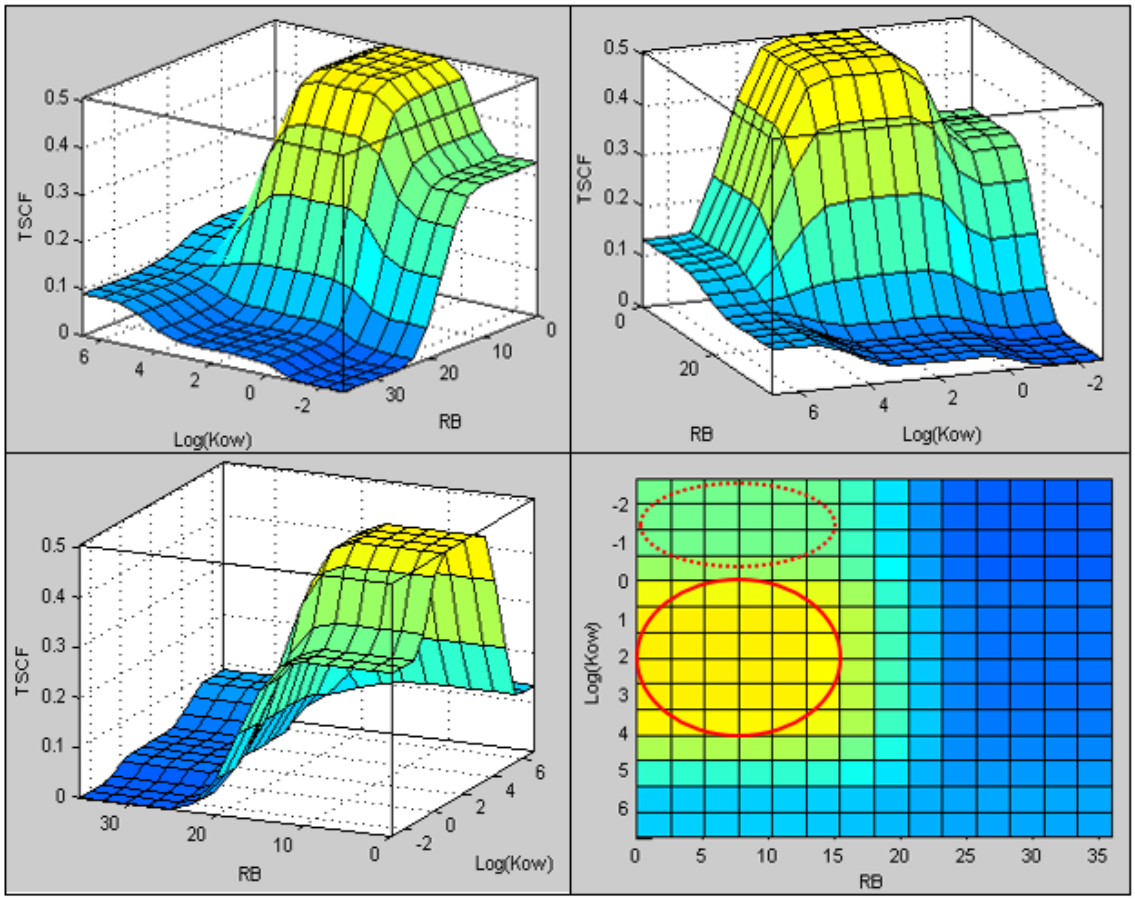


Figure 6. Relationship between Log K_{ow} and RB with TSCF using adaptive network fuzzy inference system.

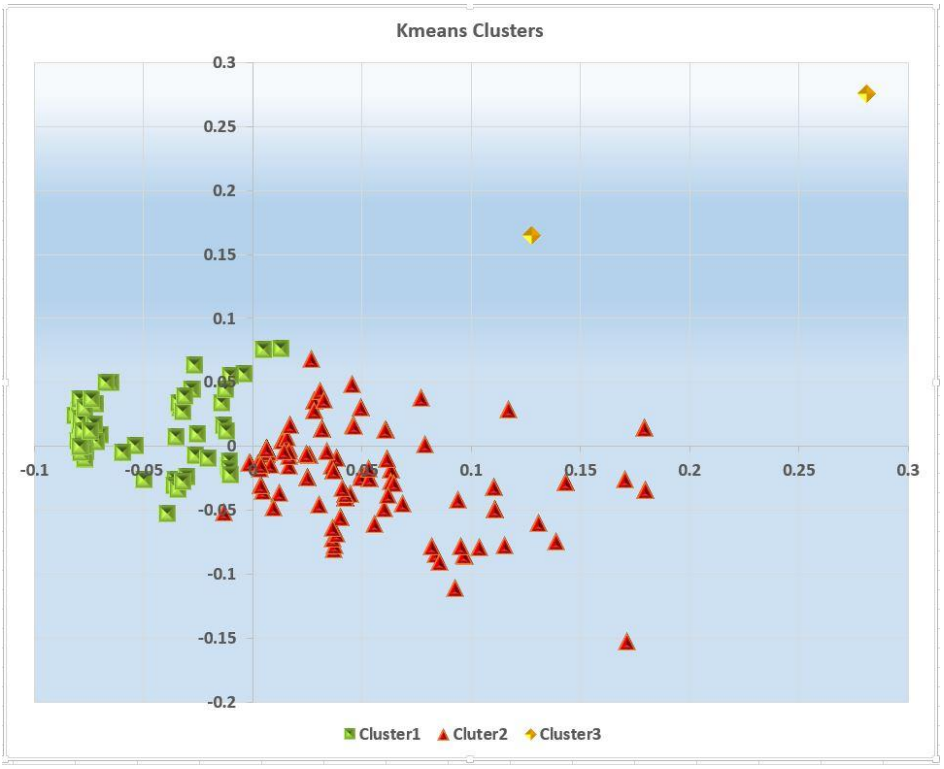


Figure 7. Resulting clusters from using k-means algorithm to generate 3 clusters (k=3), note that clusters 3 is very distant from the other data samples which clustered as one cluster when k=2.

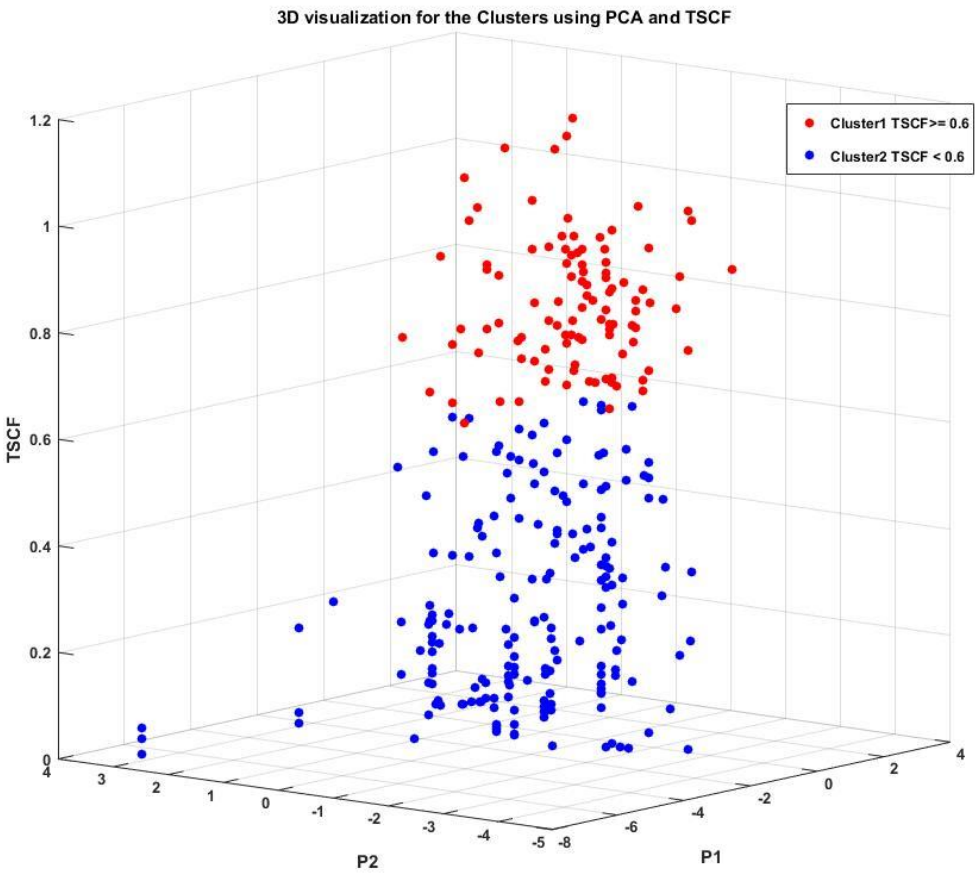


Figure 8. Three dimensional representation of the data used in the clustering.

REFERENCES

[1] Miller, E.L., et al., Root uptake of pharmaceuticals and personal care product ingredients. Environmental science & technology, 2016. 50(2): p. 525-541.

[2] RUSSELL, R.S. and V. Shorrocks, *The relationship between transpiration and the absorption of inorganic ions by intact plants.* Journal of Experimental Botany, 1959. 10(2): p. 301-316.

[3] Dettenmaier, E.M., W.J. Doucette, and B. Bugbee, *Chemical hydrophobicity and uptake by plant roots.* Environmental Science & Technology, 2008. 43(2): p. 324-329.

[4] Briggs, G.G., R.H. Bromilow, and A.A. Evans, *Relationships between lipophilicity and root uptake and translocation of non-ionised chemicals by barley.* Pest Management Science, 1982. 13(5): p. 495-504.

[5] Shone, M., B. Bartlett, and A.V. Wood, *A comparison of the uptake and translocation of some organic herbicides and a systemic fungicide by barley: II. Relationship between uptake by roots and translocation to shoots.* Journal of Experimental Botany, 1974. 25(2): p. 401-409.

[6] Burken, J.G. and J.L. Schnoor, *Predictive relationships for uptake of organic contaminants by hybrid poplar trees.* Environmental Science & Technology, 1998. 32(21): p. 3379-3385.

[7] Manzoni, S., A. Molini, and A. Porporato. *Stochastic modelling of phytoremediation*. in *Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*. 2011. The Royal Society.

[8] Collins, C.D. and E. Finnegan, *Modeling the plant uptake of organic chemicals, including the soil− air− plant pathway.* Environmental science & technology, 2010. 44(3): p. 998-1003.

[9] Undeman, E., G. Czub, and M.S. McLachlan, *Addressing temporal variability when modeling bioaccumulation in plants.* Environmental science & technology, 2009. 43(10): p. 3751-3756.

[10] Felizeter, S., M.S. McLachlan, and P. De Voogt, *Root uptake and translocation of perfluorinated alkyl acids by three hydroponically grown crops.* Journal of agricultural and food chemistry, 2014. 62(15): p. 3334-3342.

[11] Sofizadeh, A., et al., *Predicting the Distribution of Phlebotomus papatasi (Diptera: Psychodidae), the Primary Vector of Zoonotic Cutaneous Leishmaniasis, in Golestan Province of Iran Using Ecological Niche Modeling: Comparison of MaxEnt and GARP Models.* Journal of medical entomology, 2016. 54(2): p. 312-320.

[12] Mollalo, A., et al., *A 24-year exploratory spatial data analysis of Lyme disease incidence rate in Connecticut, USA.* Geospatial Health, 2017. 12(2).

[13] French, J., et al. *Artificial Neural Network forecasting of storm surge water levels at major estuarine ports to supplement national tide-surge models and improve port resilience planning*. in *EGU General Assembly Conference Abstracts*. 2017.

[14] Samli, R., V.Z. Sonmez, and N. Sivri. *Modeling the toxicity of textile industry wastewater using artificial neural networks*. in *Electric Electronics, Computer Science, Biomedical Engineerings' Meeting (EBBT), 2017*. 2017. IEEE.

[15] Bagheri, M., A. Bazvand, and M. Ehteshami, *Application of artificial intelligence for the management of landfill leachate penetration into groundwater, and assessment of its environmental impacts.* Journal of Cleaner Production, 2017. 149: p. 784-796.

[16] López, M.E., et al., *Modelling the removal of volatile pollutants under transient conditions in a two-stage bioreactor using artificial neural networks.* Journal of hazardous materials, 2017. 324: p. 100-109.

[17] Basse, R.M., et al., *Land use changes modelling using advanced methods: cellular automata and artificial neural networks. The spatial and explicit representation of land cover dynamics at the cross-border region scale.* Applied Geography, 2014. 53: p. 160-171.

[18] Han, H.-G., et al., *A soft computing method to predict sludge volume index based on a recurrent self-organizing neural network.* Applied Soft Computing, 2016. 38: p. 477-486.

[19] Gujer, W., et al., *Activated sludge model No. 3.* Water Science and Technology, 1999. 39(1): p. 183-193.

[20] Ciucani, G., et al., *Measurement of xylem translocation of weak electrolytes with the pressure chamber technique.* Pest management science, 2002. 58(5): p. 467-473.

[21] Sicbaldi, F., et al., *Root uptake and xylem translocation of pesticides from different chemical classes.* Pest Management Science, 1997. 50(2): p. 111-119.

[22] Hsu, F.C., R.L. Marxmiller, and A.Y. Yang, *Study of root uptake and xylem translocation of cinmethylin and related compounds in detopped soybean roots using a pressure chamber technique.* Plant Physiology, 1990. 93(4): p. 1573-1578.

[23] Orchard, B.J., et al., *Uptake of trichloroethylene by hybrid poplar trees grown hydroponically in flow-through plant growth chambers.* Environmental Toxicology and Chemistry, 2000. 19(4): p. 895-903.

[24] Aitchison, E.W., et al., *Phytoremediation of 1, 4-dioxane by hybrid poplar trees.* Water Environment Research, 2000. 72(3): p. 313-321.

[25] Dettenmaier, E. and W.J. Doucette, *Mineralization and plant uptake of 14C-labeled nonylphenol, nonylphenol tetraethoxylate, and nonylphenol nonylethoxylate in biosolids/soil systems planted with crested wheatgrass.* Environmental Toxicology and Chemistry, 2007. 26(2): p. 193-200.

[26] Yoon, J.M., et al., *Uptake and leaching of octahydro-1, 3, 5, 7-tetranitro-1, 3, 5, 7-tetrazocine by hybrid poplar trees.* Environmental science & technology, 2002. 36(21): p. 4649-4655.

[27] Doucette, W.J., et al., *Uptake of nonylphenol and nonylphenol ethoxylates by crested wheatgrass.* Environmental toxicology and chemistry, 2005. 24(11): p. 2965-2972.

[28] Geissbühler, H., et al., *THE FATE OF N'-(4-CHLOROPHENOXY)-PHENYL-NN-DIMETHYLUREA (C-1983) IN SOILS AND PLANTS.* Weed Research, 1963. 3(4): p. 277-297.

[29] Doucette, W., et al., *Uptake of sulfolane and diisopropanolamine (DIPA) by cattails (Typha latifolia).* Microchemical Journal, 2005. 81(1): p. 41-49.

[30] Hong, M.S., et al., *Phytoremediation of MTBE from a groundwater plume.* Environmental science & technology, 2001. 35(6): p. 1231-1239.

[31] Crowdy, S. and D.R. Jones, *The translocation of sulphonamides in higher plants: I. Uptake and translocation in broad beans.* Journal of Experimental Botany, 1956. 7(3): p. 335-346.

[32] Hayashi, O., M. Kameshiro, and K. Satoh, *Intrinsic bioavailability of 14C-heptachlor to several plant species.* Journal of Pesticide Science, 2010. 35(2): p. 107-113.

[33] Yifru, D.D. and V.A. Nzengung, *Uptake of N-nitrosodimethylamine (NDMA) from water by phreatophytes in the absence and presence of perchlorate as a co-contaminant.* Environmental science & technology, 2006. 40(23): p. 7374-7380.

[34] Davis, L.C.V., S.; Dana, J.; Selk, K.; Smith, K.; Goplen, B.; Erickson, L. E., , *Movement of Chlorinated Solvents and Other Volatile Organics through Plants Monitored by Fourier Transform Infrared (FT-IR) Spectrometry.* Journal of Hazardous Substance Research, 1998. 1: p. 1-26.

[35] Tanoue, R., et al., *Plant uptake of pharmaceutical chemicals detected in recycled organic manure and reclaimed wastewater.* Journal of agricultural and food chemistry, 2012. 60(41): p. 10203-10211.

[36] Dettenmaier, E., W. Doucette, and B. Bugbee, *Chemical hydrophobicity and uptake by plant roots.* Environmental Science & Technology, 2009. 43(2): p. 324-329.

[37] Su, Y.H. and Y.C. Liang, *Transport via xylem of atrazine, 2, 4-dinitrotoluene, and 1, 2, 3-trichlorobenzene in tomato and wheat seedlings.* Pesticide biochemistry and physiology, 2011. 100(3): p. 284-288.

[38] Fujisawa, T., et al., *Improved uptake models of nonionized pesticides to foliage and seed of crops.* Journal of agricultural and food chemistry, 2002. 50(3): p. 532-537.

[39] Farlane, C.M., T. Pfleeger, and J. Fletcher, *Effect, uptake and disposition of nitrobenzene in several terrestrial plants.* Environmental Toxicology and Chemistry, 1990. 9(4): p. 513-520.

[40] Thompson, P.L., L.A. Ramer, and J.L. Schnoor, *Hexahydro-1, 3, 5-trinitro-1, 3, 5-triazine translocation in poplar trees.* Environmental toxicology and chemistry, 1999. 18(2): p. 279-284.

[41] Crowdy, S. and D. Pramer, *The Occurrence of Translocated Antibiotics in Expressed, Plant Sap.* Annals of Botany, 1955. 19(1): p. 79-86.

[42] Edwards, N., B. Ross-Todd, and E. Garver, *Uptake and metabolism of 14C anthracene by soybean (Glycine max).* Environmental and experimental botany, 1982. 22(3): p. 349-357.

[43] Kim, J., M.C. Drew, and M.Y. Corapcioglu, *Uptake and phytotoxicity of TNT in onion plant.* Journal of Environmental Science and Health, Part A, 2004. 39(3): p. 803-819.

[44] Krstich, M.A. and O.J. Schwarz, *Characterization of xenobiotic uptake utilizing an isolated root uptake test (IRUT) and a whole plant uptake test (WPUT)*, in *Plants for Toxicity Assessment*. 1990, ASTM International.

[45] Sheets, T., *Uptake and distribution of simazine by oat and cotton seedlings.* Weeds, 1961. 9(1): p. 1-13.

[46] Pussemier, L., *Model calculations and measurements of uptake and translocation of carbamates by bean plants.* Chemosphere, 1991. 22(3-4): p. 327-339.

[47] Trapp, S., M. Matthies, and C. McFarlane, *Model for uptake of xenobiotics into plants: validation with bromacil experiments.* Environmental Toxicology and Chemistry, 1994. 13(3): p. 413-422.

[48] Behrendt, H. and R. Brüggemann, *Modelling the fate of organic chemicals in the soil plant environment: model study of root uptake of pesticides.* Chemosphere, 1993. 27(12): p. 2325-2332.

[49] Garvin, N., W.J. Doucette, and J.C. White, *Investigating differences in the root to shoot transfer and xylem sap solubility of organic compounds between zucchini, squash and soybean using a pressure chamber method.* Chemosphere, 2015. 130: p. 98-102.

[50] Paraíba, L.C., et al., *Bioconcentration factor estimates of polycyclic aromatic hydrocarbons in grains of corn plants cultivated in soils treated with sewage sludge.* Science of the total environment, 2010. 408(16): p. 3270-3276.

[51] Qiu, J., et al., *In vivo tracing of organophosphorus pesticides in cabbage (Brassica parachinensis) and aloe (Barbadensis).* Science of the Total Environment, 2016. 550: p. 1134-1140.

[52] Qiu, J., et al., *In vivo tracing of organochloride and organophosphorus pesticides in different organs of hydroponically grown malabar spinach (Basella alba L.).* Journal of hazardous materials, 2016. 316: p. 52-59.

[53] Briggs, G.G., R.L. Rigitano, and R.H. Bromilow, *Physico-chemical factors affecting uptake by roots and translocation to shoots of weak acids in barley.* Pest Management Science, 1987. 19(2): p. 101-112.

[54] San Miguel, A., P. Ravanel, and M. Raveton, *A comparative study on the uptake and translocation of organochlorines by Phragmites australis.* Journal of hazardous materials, 2013. 244: p. 60-69.

[55] Boonsaner, M., S. Borrirukwisitsak, and A. Boonsaner, *Phytoremediation of BTEX contaminated soil by Canna× generalis.* Ecotoxicology and environmental safety, 2011. 74(6): p. 1700-1707.

[56] Su, Y.H., T. Liu, and Y.C. Liang, *Transport via xylem of trichloroethylene in wheat, corn, and tomato seedlings.* Journal of hazardous materials, 2010. 182(1): p. 472-476.

[57] Macarron, R., et al., *Impact of high-throughput screening in biomedical research.* Nature reviews Drug discovery, 2011. 10(3): p. 188-195.

[58] Lipinski, C.A., et al., *Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings.* Advanced drug delivery reviews, 1997. 23(1-3): p. 3-25.

[59] Haykin, S.S., et al., *Neural networks and learning machines*. Vol. 3. 2009: Pearson Upper Saddle River, NJ, USA.

[60] Samarasinghe, S., *Neural networks for applied sciences and engineering: from fundamentals to complex pattern recognition*. 2016: CRC Press.

[61] Guo, Y., et al., *An iterative orthogonal forward regression algorithm.* International Journal of Systems Science, 2015. 46(5): p. 776-789.

[62] Nazarpour, A., G.R. Paydar, and E.J.M. Carranza, *Stepwise regression for recognition of geochemical anomalies: Case study in Takab area, NW Iran.* Journal of Geochemical Exploration, 2016. 168: p. 150-162.

[63] Zadeh, L.A., *Fuzzy Sets.* Journal of Information and Control, 1965. 8: p. 338-353.

[64] Ross, T.J., *Fuzzy logic with engineering applications*. 2009: John Wiley & Sons.

[65] Klir, G. and B. Yuan, *Fuzzy sets and fuzzy logic*. Vol. 4. 1995: Prentice hall New Jersey.

[66] Kosko, B., *Fuzzy engineering*. 1997: Prentice-Hall.

[67] Xu, R. and D. Wunsch, *Clustering*. 2009: John Wiley & Sons.

[68] Forgy, E.W., *Cluster analysis of multivariate data: efficiency versus interpretability of classifications.* biometrics, 1965. 21: p. 768-769.

[69] Rousseeuw, P.J., *Silhouettes: a graphical aid to the interpretation and validation of cluster analysis.* Journal of computational and applied mathematics, 1987. 20: p. 53-65.

[70] Davies, D.L. and D.W. Bouldin, *A cluster separation measure.* IEEE transactions on pattern analysis and machine intelligence, 1979(2): p. 224-227.

[71] Caliński, T. and J. Harabasz, *A dendrite method for cluster analysis.* Communications in Statistics-theory and Methods, 1974. 3(1): p. 1-27.

[72] Liu, Y., et al. *Understanding of internal clustering validation measures*. in *Data Mining (ICDM), 2010 IEEE 10th International Conference on*. 2010. IEEE.

[73] Wold, S., K. Esbensen, and P. Geladi, *Principal component analysis.* Chemometrics and intelligent laboratory systems, 1987. 2(1-3): p. 37-52.

[74] Montgomery, D.C., E.A. Peck, and G.G. Vining, *Introduction to linear regression analysis*. 2015: John Wiley & Sons.

[75] Fantke, P., J.A. Arnot, and W.J. Doucette, *Improving plant bioaccumulation science through consistent reporting of experimental data.* Journal of environmental management, 2016. 181: p. 374-384.

[76] Trapp, S., *Fruit tree model for uptake of organic compounds from soil and air.* SAR and QSAR in Environmental Research, 2007. 18(3-4): p. 367-387.

[77] van de Waterbeemd, H., et al., *Estimation of blood-brain barrier crossing of drugs using molecular size and shape, and H-bonding descriptors.* Journal of drug targeting, 1998. 6(2): p. 151-165.

[78] Kumar, K., et al., *Antibiotic uptake by plants from soil fertilized with animal manure.* Journal of environmental quality, 2005. 34(6): p. 2082-2085.

[79] Mark, R., *Architecture and evolution.* American scientist, 1996. 84(4): p. 383-389.

[80] Raven, J.A., *The evolution of vascular land plants in relation to supracellular transport processes.* Advances in botanical research, 1977. 5: p. 153-219.

[81] Luczak, H. and F. Raschke, *A model of the structure and the behaviour of human heart rate control (author's transl).* Biological cybernetics, 1975. 18(1): p. 1.

[82] Murray, C.D., *The physiological principle of minimum work I. The vascular system and the cost of blood volume.* Proceedings of the National Academy of Sciences, 1926. 12(3): p. 207-214.

[83] McCulloh, K.A., J.S. Sperry, and F.R. Adler, *Water transport in plants obeys Murray's law.* Nature, 2003. 421(6926): p. 939-942.

[84] Burken, J.G., D.A. Vroblesky, and J.C. Balouet, *Phytoforensics, dendrochemistry, and phytoscreening: new green tools for delineating contaminants from past and present*. 2011, ACS Publications.

[85] Balouet, C., M. Chalot, and G. O'Sullivan, *Phytoforensics: Sampling and Analytical Methods*, in *Environmental Forensics*. 2014. p. 200-229.

[86] Burken, J.G. and J.L. Schnoor, *Phytoremediation: plant uptake of atrazine and role of root exudates.* Journal of Environmental Engineering, 1996. 122(11): p. 958-963.

[87] Ma, X., et al., Phytoremediation of MTBE with hybrid poplar trees. International Journal of Phytoremediation, 2004. 6(2): p. 157-167.

SECTION

2. SUMMARY, CONCLUSIONS, AND RECOMMENDATIONS

2.1. SUMMARY OF RESEARCH WORK

This dissertation aims to study the effect of using improvised and novel machine-learning techniques in two areas of applications: smart grid (demand side management) and biomedical data analysis.

In demand side management, a novel approach was implemented to solve the multi-objective problem of minimizing the total cost of the power consumed, reducing the power demand during the peak period, and achieving customer satisfaction through controlling domestic electric water heaters. The novelty of the implemented approach lies in modeling the problem as Markov decision process, which opens the door for other researchers to adopt similar modelling in their studies, see for example the works of Ruelens, Matoki, and Zang in 2015 and 2017. The simulation results showed outstanding performance of using approximate dynamic programming in solving the described problem.

The second part of this dissertation is dedicated to the use of clustering in biomedical data analysis. Due to the heterogeneity and complexity of clinical and biomedical data, clustering becomes one of the most effective techniques in exploring these datasets. In this dissertation, three studies were implemented, in which machine-learning and statistical techniques were combined to develop ensemble, robust and unified models for analyzing and sorting out the heterogeneity in the studied datasets. The model includes the use of a new *k*-dimensional subspace-clustering algorithm that was used for exploring the autism dataset. The results paved the way for further analysis and facilitated studying

etiology, diagnosis, treatment, and prognosis. The models is general and can be extended to other disorders that exhibit a diverse range of heterogeneity. Three different datasets were studied in this part: autism spectrum disorder, neuroimaging, and environmental contaminants datasets. The results showed the efficiency of the implemented approaches in revealing unique and meaningful clusters. These findings ranged from sorting out heterogeneity in cognitive and phenotypes of autism and neuroimaging datasets to specifying the compounds that are responsible for plants uptake and contaminants transportation from the environment to food, vegetation, and mammalians' intestines.

2.2. CONCLUSIONS

This section summarizes the conclusions from the studies implemented in this dissertation. In demand side management, this study demonstrated the following conclusions:

- The Q-learning approach can at least save a family of four persons between $102, $393, and $453 annually if they are using a DEWH with 40, 70, and 100 gallons, respectively. Even for the commercial product the ADP approaches provide excellent annual saving of ($394) for DEWH with larger heating element (36 kWh) and tank size of 120 gallons. The simulation showed that the Q-learning controller maintained the water temperature above 120 ◦F.

- The implemented ADP approaches do not depend on the technology used in manufacturing the domestic electric water heaters, but depend only on the grid load demand (i.e., instantaneous energy cost), the temperature of the output water, and the user profile.
- The experiments results show that the use of machine-learning techniques can achieve substantial cost saving and environmental benefits.
- There are commercial potentials for adopting the implemented approaches and using it in real life.

In the area of biomedical-data analysis, this dissertation demonstrates the following conclusions:

- Clustering and statistical analyses for biomedical datasets is essential and could reveal important facts and distill vital information that may lead to better treatment and/or early detection for specific neurological and mental diseases.
- The implemented *k*-dimensional clustering approach showed efficiency in exploring complicated and heterogeneous datasets.
- The implemented models are general, robust, and can be used to analyze various types of complicated and mixed datasets.
- The experiments results show that sorting out the heterogeneity in autism and neuroimaging datasets may reveal significant measurements that could potentially serve as biomarkers for delineating clinically meaningful groups.

2.3. RECOMMENDATIONS

Based on the objectives and scope of work of this study, the following aspects are recommended for future research:

1. Further investigations are required to determine the efficiency of controlling electric water heaters using reinforcement learning techniques by building the model in hardware and using it in real life.
2. Integrating the implemented demand-side management approaches into smart home control systems and internet of things, which will potentially lead to better demand management and more saving for customers.
3. Applying more experiments using more reinforcement learning approaches (e.g. average reward and heuristic approaches) in solving unit commitment and day-ahead scheduling problem to get better evaluation and find more optimum approach.
4. The *k*-dimensional subspace clustering is a new approach that needs to be investigated and evaluated efficiently by applying it to more datasets.
5. The implemented unified learning approach has been used on 208 samples only, applying it on larger datasets is recommended to provide better insights for analyzing the biomarkers in aging datasets.

REFERENCES

A. P. Association et al., Diagnostic and statistical manual of mental disorders (DSM-5). American Psychiatric Pub, 2013.

Abrahams, B.S. and Geschwind, D.H., 2008. Advances in autism genetics: on the threshold of a new neurobiology. *Nature reviews genetics*, *9*(5), p.341.

Aceee.org,. (2015). Water heaters get an efficiency makeover courtesy of the Department of Energy | ACEEE. Retrieved 30 October 2015, from http://aceee.org/blog/2015/02/water-heaters-get-efficiency-makeover.

Aghaei, J. and Alizadeh, M.I., 2013. Critical peak pricing with load control demand response program in unit commitment problem. *IET Generation, Transmission & Distribution*, *7*(7), pp.681-690.

Aitchison, E.W., et al., *Phytoremediation of 1, 4-dioxane by hybrid poplar trees.* Water Environment Research, 2000. 72(3): p. 313-321.

Alelyani, S., Tang, J. and Liu, H., 2013. Feature Selection for Clustering: A Review. *Data Clustering: Algorithms and Applications*, *29*, pp.110-121.

Al-Jabery, K., Obafemi-Ajayi, T., Olbricht, G.R., Takahashi, T.N., Kanne, S. and Wunsch, D., 2016, August. Ensemble statistical and subspace clustering model for analysis of autism spectrum disorder phenotypes. In *Engineering in Medicine and Biology Society (EMBC), 2016 IEEE 38th Annual International Conference of the* (pp. 3329-3333). IEEE.

Al-Jabery, K., Wunsch, D.C., Xiong, J. and Shi, Y., 2014, November. A novel grid load management technique using electric water heaters and Q-learning. In *Smart Grid Communications (SmartGridComm), 2014 IEEE International Conference on* (pp. 776-781). IEEE.

Al-Jabery, K., Xu, Z., Yu, W., Wunsch, D.C., Xiong, J. and Shi, Y., 2017. Demand-side management of domestic electric water heaters using approximate dynamic programming. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, *36*(5), pp.775-788.

American Psychiatric Association, Diagnostic and statistical manual of mental disorders, 5th ed., American Psychiatric Publishing, 2013.

Appliance-standards.org,. (2015). The good news, and the not-so-good news, on the new DOE water heater test procedure | ASAP Appliance Standard Awareness Project. Retrieved 30 October 2015, from http://www.appliance-standards.org/blog/good-news-and-not-so-good-news-new-doe-water-heater-test-procedure.

Ardila, A., 2007. Normal aging increases cognitive heterogeneity: Analysis of dispersion in WAIS-III scores across age. *Archives of Clinical Neuropsychology*, *22*(8), pp.1003-1011.

Arora, V. and Chanana, S., 2016, March. Solution to unit commitment problem using Lagrangian relaxation and Mendel's GA method. In *Emerging Trends in Electrical Electronics & Sustainable Energy Systems (ICETEESES), International Conference on* (pp. 126-129). IEEE.

Atwa, Y. M., El-Saadany, E. F., & Salama, M. M. (2007, October). DSM Approach for Water Heater Control Strategy Utilizing Elman Neural Network. In Electrical Power Conference, 2007. EPC 2007. IEEE Canada (pp. 382-386). IEEE.

Bagheri, M., A. Bazvand, and M. Ehteshami, *Application of artificial intelligence for the management of landfill leachate penetration into groundwater, and assessment of its environmental impacts.* Journal of Cleaner Production, 2017. 149: p. 784-796.

Baker, L.M., Laidlaw, D.H., Conturo, T.E., Hogan, J., Zhao, Y., Luo, X., Correia, S., Cabeen, R., Lane, E.M., Heaps, J.M. and Bolzenius, J., 2014. White matter changes with age utilizing quantitative diffusion MRI. *Neurology*, *83*(3), pp.247-252.

Balouet, C., M. Chalot, and G. O'Sullivan, *Phytoforensics: Sampling and Analytical Methods*, in *Environmental Forensics*. 2014. p. 200-229.

Bard, J.F., 1988. Short-term scheduling of thermal-electric generators using Lagrangian relaxation. *Operations Research*, *36*(5), pp.756-766.

Basse, R.M., et al., *Land use changes modelling using advanced methods: cellular automata and artificial neural networks. The spatial and explicit representation of land cover dynamics at the cross-border region scale.* Applied Geography, 2014. 53: p. 160-171.

Bathe, K.J. and Wilson, E.L., 1976. Numerical methods in finite element analysis (Vol. 197). Englewood Cliffs, NJ: Prentice-Hall.

Behrendt, H. and R. Brüggemann, *Modelling the fate of organic chemicals in the soil plant environment: model study of root uptake of pesticides.* Chemosphere, 1993. 27(12): p. 2325-2332.

Bellazzi, R. and Zupan, B., 2008. Predictive data mining in clinical medicine: current issues and guidelines. International journal of medical informatics, 77(2), pp.81-97.

Bellman, R., 1954. Dynamic programming and a new formalism in the calculus of variations. *Proceedings of the national academy of sciences*, *40*(4), pp.231-235.

Bertsekas, D. P. (1995). *Dynamic programming and optimal control* (Vol. 1, No. 2). Belmont, MA: Athena Scientific.

Bertsekas, D., Lauer, G., Sandell, N. and Posbergh, T., 1983. Optimal short-term scheduling of large-scale power systems. *IEEE Transactions on Automatic Control*, *28*(1), pp.1-11.

Bolshakova, N. and Azuaje, F., 2003. Cluster validation techniques for genome expression data. *Signal processing*, *83*(4), pp.825-833.

Boonsaner, M., S. Borrirukwisitsak, and A. Boonsaner, *Phytoremediation of BTEX contaminated soil by Canna× generalis.* Ecotoxicology and environmental safety, 2011. 74(6): p. 1700-1707.

Borenstein, S. (2005). Time-varying retail electricity prices: Theory and practice. Electricity deregulation: choices and challenges, 111-130.

Bragin, M.A., Luh, P.B., Yan, J.H. and Stern, G.A., 2015, July. Novel exploitation of convex hull invariance for solving unit commitment by using surrogate Lagrangian relaxation and branch-and-cut. In *Power & Energy Society General Meeting, 2015 IEEE* (pp. 1-5). IEEE.

Briggs, G.G., R.H. Bromilow, and A.A. Evans, *Relationships between lipophilicity and root uptake and translocation of non‐ionised chemicals by barley.* Pest Management Science, 1982. 13(5): p. 495-504.

Briggs, G.G., R.L. Rigitano, and R.H. Bromilow, *Physico‐chemical factors affecting uptake by roots and translocation to shoots of weak acids in barley.* Pest Management Science, 1987. 19(2): p. 101-112.

Burken, J.G. and J.L. Schnoor, *Phytoremediation: plant uptake of atrazine and role of root exudates.* Journal of Environmental Engineering, 1996. 122(11): p. 958-963.

Burken, J.G. and J.L. Schnoor, *Predictive relationships for uptake of organic contaminants by hybrid poplar trees.* Environmental Science & Technology, 1998. 32(21): p. 3379-3385.

Burken, J.G., D.A. Vroblesky, and J.C. Balouet, *Phytoforensics, dendrochemistry, and phytoscreening: new green tools for delineating contaminants from past and present*. 2011, ACS Publications.

Cabeen, R.P., Bastin, M.E. and Laidlaw, D.H., 2016. Kernel regression estimation of fiber orientation mixtures in diffusion MRI. *Neuroimage*, *127*, pp.158-172.

Cabeza, R., Nyberg, L. and Park, D.C. eds., 2016. *Cognitive neuroscience of aging: Linking cognitive and cerebral aging*. Oxford University Press.

Caliński, T. and J. Harabasz, *A dendrite method for cluster analysis.* Communications in Statistics-theory and Methods, 1974. 3(1): p. 1-27.

Cardarelli, F. (2003). *Encyclopaedia of scientific units, weights and measures: Their SI equivalences and origins*. Springer.

Chen, C., Duan, S., Cai, T. and Liu, B., 2011. Online 24-h solar power forecasting based on weather type classification using artificial neural network. Solar Energy, 85(11), pp.2856-2870.

Ciucani, G., et al., *Measurement of xylem translocation of weak electrolytes with the pressure chamber technique.* Pest management science, 2002. 58(5): p. 467-473.

Cohen, A.I. and Sherkat, V.R., 1987. Optimization-based methods for operations scheduling. *Proceedings of the IEEE*, *75*(12), pp.1574-1591.

Collins, C.D. and E. Finnegan, *Modeling the plant uptake of organic chemicals, including the soil− air− plant pathway.* Environmental science & technology, 2010. 44(3): p. 998-1003.

Correia, S., Lee, S.Y., Voorn, T., Tate, D.F., Paul, R.H., Zhang, S., Salloway, S.P., Malloy, P.F. and Laidlaw, D.H., 2008. Quantitative tractography metrics of white matter integrity in diffusion-tensor MRI. *Neuroimage*, *42*(2), pp.568-581.

Cristianini, N., and Shawe-Taylor, J., "An introduction to support vector machines and other kernel-based learning methods". Cambridge university press, 2000.

Crowdy, S. and D. Pramer, *The Occurrence of Translocated Antibiotics in Expressed, Plant Sap.* Annals of Botany, 1955. 19(1): p. 79-86.

Crowdy, S. and D.R. Jones, *The translocation of sulphonamides in higher plants: I. Uptake and translocation in broad beans.* Journal of Experimental Botany, 1956. 7(3): p. 335-346.

Dalal, G. and Mannor, S., 2015, June. Reinforcement learning for the unit commitment problem. In *PowerTech, 2015 IEEE Eindhoven* (pp. 1-6). IEEE..

Davies, D.L. and D.W. Bouldin, *A cluster separation measure.* IEEE transactions on pattern analysis and machine intelligence, 1979(2): p. 224-227.

Davis, L.C.V., S.; Dana, J.; Selk, K.; Smith, K.; Goplen, B.; Erickson, L. E., , *Movement of Chlorinated Solvents and Other Volatile Organics through Plants Monitored by Fourier Transform Infrared (FT-IR) Spectrometry.* Journal of Hazardous Substance Research, 1998. 1: p. 1-26.

Department for Environment, Food and Rural Affairs. (2008). Measurement of domestic hot water consumption in dwellings. London; UK: DEFRA.

Dettenmaier, E. and W.J. Doucette, *Mineralization and plant uptake of 14C-labeled nonylphenol, nonylphenol tetraethoxylate, and nonylphenol nonylethoxylate in biosolids/soil systems planted with crested wheatgrass.* Environmental Toxicology and Chemistry, 2007. 26(2): p. 193-200.

Dettenmaier, E., W. Doucette, and B. Bugbee, *Chemical hydrophobicity and uptake by plant roots.* Environmental Science & Technology, 2009. 43(2): p. 324-329.

Diduch, C., Shaad, M., Errouissi, R., Kaye, M. E., Meng, J., and Chang, L. (2012, June). Aggregated domestic electric water heater control - Building on smart grid infrastructure. In Power Electronics and Motion Control Conference (IPEMC), 2012 7th International (Vol. 1, pp. 128-135). IEEE.

Doucette, W., et al., *Uptake of sulfolane and diisopropanolamine (DIPA) by cattails (Typha latifolia).* Microchemical Journal, 2005. 81(1): p. 41-49.

Doucette, W.J., et al., *Uptake of nonylphenol and nonylphenol ethoxylates by crested wheatgrass.* Environmental toxicology and chemistry, 2005. 24(11): p. 2965-2972.

Dudek, G., 2013. Genetic algorithm with binary representation of generating unit start-up and shut-down times for the unit commitment problem. Expert Systems with Applications, 40(15), pp.6080-6086.

Edwards, N., B. Ross-Todd, and E. Garver, *Uptake and metabolism of 14C anthracene by soybean (Glycine max).* Environmental and experimental botany, 1982. 22(3): p. 349-357.

Eia.gov,. (2015). Residential Energy Consumption Survey (RECS) - Analysis & Projections - U.S. Energy Information Administration (EIA). Retrieved 24 October 2015, from http://www.eia.gov/consumption/residen

Energy.gov,. (2015). Office of Energy Efficiency & Renewable Energy | Department of Energy. Retrieved 25 October 2015, from http://energy.gov/eere/office-energy-efficiency-renewable-energy

Ercan, I., Öztopal, A., Yerli, B., Kemal Kaymak, M., & Şahin, A. D. (2012). Short–mid-term solar power prediction by using artificial neural networks. Solar Energy, 725-733.

Fang, R., Pouyanfar, S., Yang, Y., Chen, S.C. and Iyengar, S.S., 2016. Computational health informatics in the big data age: A survey. *ACM Computing Surveys (CSUR)*, *49*(1), p.12.

Fantke, P., J.A. Arnot, and W.J. Doucette, *Improving plant bioaccumulation science through consistent reporting of experimental data.* Journal of environmental management, 2016. 181: p. 374-384.

Farlane, C.M., T. Pfleeger, and J. Fletcher, *Effect, uptake and disposition of nitrobenzene in several terrestrial plants.* Environmental Toxicology and Chemistry, 1990. 9(4): p. 513-520.

Fein, G., Di Sclafani, V., Tanabe, J., Cardenas, V., Weiner, M.W., Jagust, W.J., Reed, B.R., Norman, D., Schuff, N., Kusdra, L. and Greenfield, T., 2000. Hippocampal and cortical atrophy predict dementia in subcortical ischemic vascular disease. *Neurology*, *55*(11), pp.1626-1635.

Felizeter, S., M.S. McLachlan, and P. De Voogt, *Root uptake and translocation of perfluorinated alkyl acids by three hydroponically grown crops.* Journal of agricultural and food chemistry, 2014. 62(15): p. 3334-3342.

Fernandez-Jimenez, L. A., Muñoz-Jimenez, A., Falces, A., Mendoza-Villena, M., Garcia-Garrido, E., Lara-Santillan, P. M., Zorzano-Santamaria, P. J. (2012). Short-term power forecasting system for photovoltaic plants. Renewable Energy, 311-317.

Fischbach, G.D. and Lord, C., 2010. The Simons Simplex Collection: a resource for identification of autism genetic risk factors. *Neuron*, *68*(2), pp.192-195.

Forgy, E.W., *Cluster analysis of multivariate data: efficiency versus interpretability of classifications.* biometrics, 1965. 21: p. 768-769.

Formulas and Facts. (n.d.). Contractorsinstitute.com. Retrieved January 13th, 2014 from http:// www.contractorsinstitute.com/downloads/Solar/Contractors'%20Domestic%20Hot%20Water%20Educational%20PDF's/Hot%20Water%20Formulas%20and%20Facts.pdf.

Foss, M.P., Formigheri, P. and Speciali, J.G., 2009. Heterogeneity of cognitive aging in Brazilian normal elderls. *Dementia & Neuropsychologia*, *3*(4), pp.344-351.

Fraley, C. and Raftery, A.E., 2002. Model-based clustering, discriminant analysis, and density estimation. *Journal of the American statistical Association*, *97*(458), pp.611-631.

French, J., et al. *Artificial Neural Network forecasting of storm surge water levels at major estuarine ports to supplement national tide-surge models and improve port resilience planning*. in *EGU General Assembly Conference Abstracts*. 2017.

Fujisawa, T., et al., *Improved uptake models of nonionized pesticides to foliage and seed of crops.* Journal of agricultural and food chemistry, 2002. 50(3): p. 532-537.

Garvin, N., W.J. Doucette, and J.C. White, *Investigating differences in the root to shoot transfer and xylem sap solubility of organic compounds between zucchini, squash and soybean using a pressure chamber method.* Chemosphere, 2015. 130: p. 98-102.

Geissbühler, H., et al., *THE FATE OF N'-(4-CHLOROPHENOXY)-PHENYL-NN-DIMETHYLUREA (C-1983) IN SOILS AND PLANTS.* Weed Research, 1963. 3(4): p. 277-297.

Georgiades, S., Szatmari, P. and Boyle, M., 2013. Importance of studying heterogeneity in autism. *Neuropsychiatry*, *3*(2), pp.123-125.

Georgiades, S., Szatmari, P., Boyle, M., Hanna, S., Duku, E., Zwaigenbaum, L., Bryson, S., Fombonne, E., Volden, J., Mirenda, P. and Smith, I., 2013. Investigating phenotypic heterogeneity in children with autism spectrum disorder: a factor mixture modeling approach. *Journal of Child Psychology and Psychiatry*, *54*(2), pp.206-215.

Gold, G., Kövari, E., Herrmann, F.R., Canuto, A., Hof, P.R., Michel, J.P., Bouras, C. and Giannakopoulos, P., 2005. Cognitive consequences of thalamic, basal ganglia, and deep white matter lacunes in brain aging and dementia. *Stroke*, *36*(6), pp.1184-1188.

Gosavi, A., 2015. Simulation-based optimization: an overview. In Simulation-Based Optimization (pp. 29-35). Springer, Boston, MA.

Gouw, A.A., Seewann, A., Vrenken, H., Van Der Flier, W.M., Rozemuller, J.M., Barkhof, F., Scheltens, P. and Geurts, J.J.G., 2008. Heterogeneity of white matter hyperintensities in Alzheimer's disease: post-mortem quantitative MRI and neuropathology. *Brain*, *131*(12), pp.3286-3298.

Grainger, J.J.S., Grainger, W.D.J.J. and Stevenson, W.D., 1994. Power system analysis.disruption. Power Systems, IEEE Transactions on, 23(4), pp. 1681-1688.

Grieve, S.M., Williams, L.M., Paul, R.H., Clark, C.R. and Gordon, E., 2007. Cognitive aging, executive function, and fractional anisotropy: a diffusion tensor MR imaging study. *American Journal of Neuroradiology*, *28*(2), pp.226-235.

Guan, X., Luh, P.B., Yan, H. and Amalfi, J.A., 1992. An optimization-based method for unit commitment. *International Journal of Electrical Power & Energy Systems*, *14*(1), pp.9-17.

Gujer, W., et al., *Activated sludge model No. 3.* Water Science and Technology, 1999. 39(1): p. 183-193.

Guo, Y., et al., *An iterative orthogonal forward regression algorithm.* International Journal of Systems Science, 2015. 46(5): p. 776-789.

Guyon, I. and Elisseeff, A., 2003. An introduction to variable and feature selection. *Journal of machine learning research*, *3*(Mar), pp.1157-1182.

Halkidi, M., Batistakis, Y. and Vazirgiannis, M., 2001. On clustering validation techniques. *Journal of intelligent information systems*, *17*(2-3), pp.107-145.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P. and Witten, I.H., 2009. The WEKA data mining software: an update. *ACM SIGKDD explorations newsletter*, *11*(1), pp.10-18.

Han, H.-G., et al., *A soft computing method to predict sludge volume index based on a recurrent self-organizing neural network.* Applied Soft Computing, 2016. 38: p. 477-486.

Hargreaves, J.J. and Hobbs, B.F., 2012. Commitment and dispatch with uncertain wind generation by dynamic programming. IEEE Transactions on sustainable energy, 3(4), pp.724-734.

Hayashi, O., M. Kameshiro, and K. Satoh, *Intrinsic bioavailability of 14C-heptachlor to several plant species.* Journal of Pesticide Science, 2010. 35(2): p. 107-113.

Haykin, S.S., et al., *Neural networks and learning machines*. Vol. 3. 2009: Pearson Upper Saddle River, NJ, USA:.

Henze, G. P., & Dodier, R. H. (2003). Adaptive Optimal Control of a Grid-Independent Photovoltaic System. Transaction of the ASME, 34-42.

Hong, M.S., et al., *Phytoremediation of MTBE from a groundwater plume.* Environmental science & technology, 2001. 35(6): p. 1231-1239.

Hsu, F.C., R.L. Marxmiller, and A.Y. Yang, *Study of root uptake and xylem translocation of cinmethylin and related compounds in detopped soybean roots using a pressure chamber technique.* Plant Physiology, 1990. 93(4): p. 1573-1578.

Hus, V., Gotham, K. and Lord, C., 2014. Standardizing ADOS domain scores: Separating severity of social affect and restricted and repetitive behaviors. *Journal of autism and developmental disorders*, *44*(10), pp.2400-2412.

Ingram, D.G., Takahashi, T.N. and Miles, J.H., 2008. Defining autism subgroups: a taxometric solution. *Journal of autism and developmental disorders*, *38*(5), pp.950-960.

Jasmin, E.A. and TP, I.A., 2009, December. Reinforcement Learning solution for Unit Commitment Problem through pursuit method. In *Advances in Computing, Control, & Telecommunication Technologies, 2009. ACT'09. International Conference on* (pp. 324-327). IEEE.

Jasmin, E.A., Ahamed, T.I. and Remani, T., 2016, December. A function approximation approach to Reinforcement Learning for solving unit commitment problem with Photo voltaic sources. In *Power Electronics, Drives and Energy Systems (PEDES), 2016 IEEE International Conference on* (pp. 1-6). IEEE.

Jerez, J.M., Molina, I., García-Laencina, P.J., Alba, E., Ribelles, N., Martín, M. and Franco, L., 2010. Missing data imputation using statistical and machine learning methods in a real breast cancer problem. *Artificial intelligence in medicine*, *50*(2), pp.105-115.

Johnson, R.A. and Wichern, D.W. Applied Multivariate Statistical Analysis, 6th ed. Pearson, 2008.

Ke, X., Wu, D., Lu, N. and Kintner-Meyer, M., 2015, July. A modified priority list-based MILP method for solving large-scale unit commitment problems. In Power & Energy Society General Meeting, 2015 IEEE (pp. 1-5). IEEE.

Kim, J., M.C. Drew, and M.Y. Corapcioglu, *Uptake and phytotoxicity of TNT in onion plant.* Journal of Environmental Science and Health, Part A, 2004. 39(3): p. 803-819.

Kim, Y., Street, W.N. and Menczer, F., 2000, August. Feature selection in unsupervised learning via evolutionary search. In *Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 365-369). ACM.

Klemes, J., Smith, R., and Kim, J. K. (Eds.). (2008). *Handbook of water and energy management in food processing*. Elsevier.

Klir, G. and B. Yuan, *Fuzzy sets and fuzzy logic*. Vol. 4. 1995: Prentice hall New Jersey.

Kohavi, R. and John, G.H., 1997. Wrappers for feature subset selection. *Artificial intelligence*, *97*(1-2), pp.273-324.

Kosko, B., *Fuzzy engineering*. 1997: Prentice-Hall.

Krstich, M.A. and O.J. Schwarz, *Characterization of xenobiotic uptake utilizing an isolated root uptake test (IRUT) and a whole plant uptake test (WPUT)*, in *Plants for Toxicity Assessment*. 1990, ASTM International.

Kumar, K., et al., *Antibiotic uptake by plants from soil fertilized with animal manure.* Journal of environmental quality, 2005. 34(6): p. 2082-2085.

Lam, D., Wei, M. and Wunsch, D., 2015. Clustering data of mixed categorical and numerical type with unsupervised feature learning. *IEEE Access*, *3*, pp.1605-1613.

Lefebvre, S. and Desbiens, C., 2002. Residential load modeling for predicting distribution transformer load behavior, feeder load and cold load pickup. *International journal of electrical power & energy systems*, *24*(4), pp.285-293.

Limmer, M.A. and J.G. Burken, *Plant translocation of organic compounds: molecular and physicochemical predictors.* Environmental Science & Technology Letters, 2014. 1(2): p. 156-161.

Lipinski, C.A., et al., *Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings.* Advanced drug delivery reviews, 1997. 23(1-3): p. 3-25.

Liu, Y., et al. *Understanding of internal clustering validation measures*. in *Data Mining (ICDM), 2010 IEEE 10th International Conference on*. 2010. IEEE.

Liu, Y., Li, Z., Xiong, H., Gao, X. and Wu, J., 2010, December. Understanding of internal clustering validation measures. In *Data Mining (ICDM), 2010 IEEE 10th International Conference on* (pp. 911-916). IEEE.

López, M.E., et al., *Modelling the removal of volatile pollutants under transient conditions in a two-stage bioreactor using artificial neural networks.* Journal of hazardous materials, 2017. 324: p. 100-109.

Lord, C., Rutter, M. and Le Couteur, A., 1994. Autism Diagnostic Interview-Revised: a revised version of a diagnostic interview for caregivers of individuals with possible pervasive developmental disorders. *Journal of autism and developmental disorders*, *24*(5), pp.659-685.

Lord, C., Rutter, M., Goode, S., Heemsbergen, J., Jordan, H., Mawhood, L. and Schopler, E., 1989. Austism diagnostic observation schedule: A standardized observation of communicative and social behavior. *Journal of autism and developmental disorders*, *19*(2), pp.185-212.

Lu, N., and Katipamula, S. (2005, June). Control strategies of thermostatically controlled appliances in a competitive electricity market. In *Power Engineering Society General Meeting, 2005. IEEE* (pp. 202-207). IEEE.

Lu, S. and Kintner-Meyer, M. (2008). Scoping study for the Demand Response DFT II Project in Morgantown. (PNNL-17474). Richland, Washington; USA: Pacific Northwest National Laboratory.

Luczak, H. and F. Raschke, *A model of the structure and the behaviour of human heart rate control (author's transl).* Biological cybernetics, 1975. 18(1): p. 1.

Ma, X., et al., *Phytoremediation of MTBE with hybrid poplar trees.* International Journal of Phytoremediation, 2004. 6(2): p. 157-167.

Macarron, R., et al., *Impact of high-throughput screening in biomedical research.* Nature reviews Drug discovery, 2011. 10(3): p. 188-195.

Manzoni, S., A. Molini, and A. Porporato. *Stochastic modelling of phytoremediation*. in *Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*. 2011. The Royal Society.

Mark, R., *Architecture and evolution.* American scientist, 1996. 84(4): p. 383-389.

McCulloh, K.A., J.S. Sperry, and F.R. Adler, *Water transport in plants obeys Murray's law*. Nature, 2003. 421(6926): p. 939-942.

Mellit, A., Pavan, A. M., & Lughi, V. (2014). Short-term forecasting of power production in a large-scale photovoltaic plant. Solar Energy, 401-413.

MIDC, N., 2010. National Renewable Energy Laboratory Measurement and Instrumentation Data Center (NREL MIDC) Solar Position and Intensity (SOLPOS) Calculator. Online at http://www. nrel. gov/midc/solpos/solpos. html.

Miles, J.H., 2011. Autism spectrum disorders—a genetics review. *Genetics in Medicine*, *13*(4), p.278.

Miller, E.L., et al., *Root uptake of pharmaceuticals and personal care product ingredients.* Environmental science & technology, 2016. 50(2): p. 525-541.

Mok, V.C., Liu, T., Lam, W.W., Wong, A., Hu, X., Guo, L., Chen, X.Y., Tang, W.K., Wong, K.S. and Wong, S., 2008. Neuroimaging predictors of cognitive impairment in confluent white matter lesion: volumetric analyses of 99 brain regions. *Dementia and geriatric cognitive disorders*, *25*(1), pp.67-73.

Mollalo, A., et al., *A 24-year exploratory spatial data analysis of Lyme disease incidence rate in Connecticut, USA.* Geospatial Health, 2017. 12(2).

Montgomery, D.C., E.A. Peck, and G.G. Vining, *Introduction to linear regression analysis*. 2015: John Wiley & Sons.

Moreau, A. (2011). Control strategy for domestic water heaters during peak periods and its impact on the demand for electricity. *Energy Procedia 12* (1074 – 1082). Chengdu, China: Elsevier.

Mujumdar, A. S. (2006). A review of: "Mathematical Principles of Heat Transfer." *Drying Technology*, 24(2), 245.

Mungas, D., Jagust, W.J., Reed, B.R., Kramer, J.H., Weiner, M.W., Schuff, N., Norman, D., Mack, W.J., Willis, L. and Chui, H.C., 2001. MRI predictors of cognition in subcortical ischemic vascular disease and Alzheimer's disease. *Neurology*, *57*(12), pp.2229-2235.

Murray, C.D., *The physiological principle of minimum work I. The vascular system and the cost of blood volume.* Proceedings of the National Academy of Sciences, 1926. 12(3): p. 207-214.

Nazarpour, A., G.R. Paydar, and E.J.M. Carranza, *Stepwise regression for recognition of geochemical anomalies: Case study in Takab area, NW Iran.* Journal of Geochemical Exploration, 2016. 168: p. 150-162.

Nehrir, M. H., LaMeres, B. J., and Gerez, V. (1999, January). A customer-interactive electric water heater demand-side management strategy using fuzzy logic. In *Power Engineering Society 1999 Winter Meeting, IEEE* (Vol. 1, pp. 433-436). IEEE.

Nikovski, D. and Zhang, W., 2010, September. Factored markov decision process models for stochastic unit commitment. In Innovative Technologies for an Efficient and Reliable Electricity Supply (CITRES), 2010 IEEE Conference on (pp. 28-35). IEEE.

O'Leary, D., & Kubby, J. (2017). Feature Selection and ANN Solar Power Prediction. Journal of Renewable Energy.

Obafemi-Ajayi, T., Lam, D., Takahashi, T.N., Kanne, S. and Wunsch, D., 2015, August. Sorting the phenotypic heterogeneity of autism spectrum disorders: A hierarchical clustering model. In *Computational Intelligence in Bioinformatics and Computational Biology (CIBCB), 2015 IEEE Conference on* (pp. 1-7). IEEE.

Omer, A. M. (2008). Energy, environment and sustainable development.*Renewable and Sustainable Energy Reviews*, *12*(9), pp. 2265-2300.

Orchard, B.J., et al., *Uptake of trichloroethylene by hybrid poplar trees grown hydroponically in flow-through plant growth chambers.* Environmental Toxicology and Chemistry, 2000. 19(4): p. 895-903.

Ordoudis, C., Pinson, P., Morales González, J. M., & Zugno, M. (2016). An Updated Version of the IEEE RTS 24-Bus System for Electricity Market and Power System Operation Studies. Technical University of Denmark (DTU).

Ott, L., Pang, L., Ramos, F.T. and Chawla, S., 2014. On integrated clustering and outlier detection. In *Advances in neural information processing systems* (pp. 1359-1367).

Padhy, N.P., 2004. Unit commitment-a bibliographical survey. IEEE Transactions on power systems, 19(2), pp.1196-1205.

Pal, A., Dasgupta, K., Banerjee, S. and Chanda, C.K., 2016, March. An analysis of Economic Load Dispatch with Ramp-rate limit constraints using BSA. In *Electrical, Electronics and Computer Science (SCEECS), 2016 IEEE Students' Conference on* (pp. 1-6). IEEE..

Paraíba, L.C., et al., *Bioconcentration factor estimates of polycyclic aromatic hydrocarbons in grains of corn plants cultivated in soils treated with sewage sludge.* Science of the total environment, 2010. 408(16): p. 3270-3276.

Paul, R.H., Lawrence, J., Williams, L.M., Richard, C.C., Cooper, N. and Gordon, E., 2005. Preliminary validity of "integneuroTM": A new computerized battery of neurocognitive tests. *International Journal of Neuroscience*, *115*(11), pp.1549-1567.

Peakload.org,. (2015). 2014 Annual Report to Members - Peak Load Management Alliance. Retrieved 25 October 2015, from http://www.peakload.org/?page=2014Report.

Powermin.nic.in,. (2016). Annual Reports Year-wise Indian Ministry of Power. Retrieved: 2 February 2016, from http://powermin.nic.in/annual-reports-year-wise.

Prokhorov, D. V., & Wunsch, D. C. (1997). Adaptive critic designs. Neural Networks, IEEE Transactions on, 8(5), 997-1007.

Pussemier, L., *Model calculations and measurements of uptake and translocation of carbamates by bean plants.* Chemosphere, 1991. 22(3-4): p. 327-339.

Qiu, J., et al., *In vivo tracing of organochloride and organophosphorus pesticides in different organs of hydroponically grown malabar spinach (Basella alba L.).* Journal of hazardous materials, 2016. 316: p. 52-59.

Qiu, J., et al., *In vivo tracing of organophosphorus pesticides in cabbage (Brassica parachinensis) and aloe (Barbadensis).* Science of the Total Environment, 2016. 550: p. 1134-1140.

Ramanathan, B., and Vittal, V. (2008). A framework for evaluation of advanced direct load control with minimum disruption. *Power Systems, IEEE Transactions on*, *23*(4), pp. 1681-1688.

Rautenbach, B., and Lane, I. E. (1996). The multi-objective controller: A novel approach to domestic hot water load control. *Power Systems, IEEE Transactions on*, *11*(4), 1832-1837.

Raven, J.A., *The evolution of vascular land plants in relation to supracellular transport processes.* Advances in botanical research, 1977. 5: p. 153-219.

Raz, N., 2001. Ageing and the brain. *eLS*.

Raz, N., Rodrigue, K.M. and Haacke, E.M., 2007. Brain aging and its modifiers. *Annals of the New York Academy of Sciences*, *1097*(1), pp.84-93.

Robbins, H., & Monro, S. (1951). A stochastic approximation method. The annals of mathematical statistics, 400-407.

Ross, T.J., *Fuzzy logic with engineering applications*. 2009: John Wiley & Sons.

Rousseeuw, P.J., 1987. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, *20*, pp.53-65.

Ruelens, F., Claessens, B., Quaiyum, S., De Schutter, B., Babuska, R. and Belmans, R., 2016. Reinforcement learning applied to an electric water heater: from theory to practice. IEEE Transactions on Smart Grid.

Ruelens, F., Claessens, B.J., Vandael, S., De Schutter, B., Babuška, R. and Belmans, R., 2017. Residential demand response of thermostatically controlled loads using batch reinforcement learning. IEEE Transactions on Smart Grid.

Ruelens, F., Claessens, B.J., Vandael, S., Iacovella, S., Vingerhoets, P. and Belmans, R., 2014, August. Demand response of a heterogeneous cluster of electric water heaters using batch reinforcement learning. In Power Systems Computation Conference (PSCC), 2014 (pp. 1-7). IEEE.

Russell, R.S. and V. Shorrocks, *The relationship between transpiration and the absorption of inorganic ions by intact plants.* Journal of Experimental Botany, 1959. 10(2): p. 301-316.

Saker, N., Petit, M., Vannier, J. C., & Coullon, J. L. (2011). Demand Side Management of Electrical Water Heaters and Evaluation of the Cold Load Pick-Up characteristics. In 17th Power System Computation Conference (p. 8p).

Samarasinghe, S., *Neural networks for applied sciences and engineering: from fundamentals to complex pattern recognition*. 2016: CRC Press.

Samli, R., V.Z. Sonmez, and N. Sivri. *Modeling the toxicity of textile industry wastewater using artificial neural networks*. in *Electric Electronics, Computer Science, Biomedical Engineerings' Meeting (EBBT), 2017*. 2017. IEEE.

San Miguel, A., P. Ravanel, and M. Raveton, *A comparative study on the uptake and translocation of organochlorines by Phragmites australis.* Journal of hazardous materials, 2013. 244: p. 60-69.

Saravanan, B., Sikri, S., Swarup, K.S. and Kothari, D.P., 2013, February. Unit commitment using DP—An exhaustive working of both classical and stochastic approach. In Power, Energy and Control (ICPEC), 2013 International Conference on (pp. 382-385). IEEE.

Seki, T., Yamashita, N. and Kawamoto, K., 2010. New local search methods for improving the Lagrangian-relaxation-based unit commitment solution. IEEE Transactions on Power Systems, 25(1), pp.272-283.

Senjyu, T., Shimabukuro, K., Uezato, K. and Funabashi, T., 2003. A fast technique for unit commitment problem by extended priority list. IEEE Transactions on Power Systems, 18(2), pp.882-888.

Sepulveda, A., Paull, L., Morsi, W. G., Li, H., Diduch, C. P., and Chang, L. (2010, August). A novel demand side management program using water heaters and particle swarm optimization. In *Electric Power and Energy Conference (EPEC), 2010 IEEE* (pp. 1-5). IEEE.

Sheets, T., *Uptake and distribution of simazine by oat and cotton seedlings.* Weeds, 1961. 9(1): p. 1-13.

Shone, M., B. Bartlett, and A.V. Wood, *A comparison of the uptake and translocation of some organic herbicides and a systemic fungicide by barley: II. Relationship between uptake by roots and translocation to shoots.* Journal of Experimental Botany, 1974. 25(2): p. 401-409.

Shukla, A. and Singh, S.N., 2016. Multi-objective unit commitment with renewable energy using hybrid approach. IET Renewable Power Generation, 10(3), pp.327-338.

Sicbaldi, F., et al., *Root uptake and xylem translocation of pesticides from different chemical classes.* Pest Management Science, 1997. 50(2): p. 111-119.

Sofizadeh, A., et al., *Predicting the Distribution of Phlebotomus papatasi (Diptera: Psychodidae), the Primary Vector of Zoonotic Cutaneous Leishmaniasis, in Golestan Province of Iran Using Ecological Niche Modeling: Comparison of MaxEnt and GARP Models.* Journal of medical entomology, 2016. 54(2): p. 312-320.

Sonntag, R. E., Borgnakke, C., Van Wylen, G. J., and Van Wyk, S. (1998).*Fundamentals of Thermodynamics* (pp. 356-57). New York: Wiley.

Stevens, M.C., Fein, D.A., Dunn, M., Allen, D., Waterhouse, L.H., Feinstein, C. and Rapin, I., 2000. Subgroups of children with autism by cluster analysis: A longitudinal examination. *Journal of the American Academy of Child & Adolescent Psychiatry*, *39*(3), pp.346-352.

Strelec, M. and Berka, J., 2013, October. Microgrid energy management based on approximate dynamic programming. In Innovative Smart Grid Technologies Europe (ISGT EUROPE), 2013 4th IEEE/PES (pp. 1-5). IEEE.

Su, Y.H. and Y.C. Liang, *Transport via xylem of atrazine, 2, 4-dinitrotoluene, and 1, 2, 3-trichlorobenzene in tomato and wheat seedlings.* Pesticide biochemistry and physiology, 2011. 100(3): p. 284-288.

Su, Y.H., T. Liu, and Y.C. Liang, *Transport via xylem of trichloroethylene in wheat, corn, and tomato seedlings.* Journal of hazardous materials, 2010. 182(1): p. 472-476.

Tanoue, R., et al., *Plant uptake of pharmaceutical chemicals detected in recycled organic manure and reclaimed wastewater.* Journal of agricultural and food chemistry, 2012. 60(41): p. 10203-10211.

Thompson, P.L., L.A. Ramer, and J.L. Schnoor, *Hexahydro-1, 3, 5-trinitro-1, 3, 5-triazine translocation in poplar trees.* Environmental toxicology and chemistry, 1999. 18(2): p. 279-284.

Tiptipakorn, S., and Lee, W. J. (2007, September). A residential consumer-centered load control strategy in real-time electricity pricing environment. In *Power Symposium, 2007. NAPS'07. 39th North American* (pp. 505-510). IEEE.

Tong, S.K., Shahidehpour, S.M. and Ouyang, Z., 1991. A heuristic short-term unit commitment. IEEE Transactions on Power Systems, 6(3), pp.1210-1216.

Trapp, S., *Fruit tree model for uptake of organic compounds from soil and air.* SAR and QSAR in Environmental Research, 2007. 18(3-4): p. 367-387.

Trapp, S., M. Matthies, and C. McFarlane, *Model for uptake of xenobiotics into plants: validation with bromacil experiments.* Environmental Toxicology and Chemistry, 1994. 13(3): p. 413-422.

Trivedi, A., Srinivasan, D., Biswas, S. and Reindl, T., 2015. Hybridizing genetic algorithm with differential evolution for solving the unit commitment scheduling problem. Swarm and Evolutionary Computation, 23, pp.50-64.

Tseng, C. L. (1998) "On Power System Generation Unit Commitment Problems" Ph D Thesis Department of Industry Engineering and Operations Research University of California at Berkeley.

Tseng, C. L., Oren, S. S., Cheng, C. S., Li, C. A., Svoboda, A. J., & Johnson, R. B. (1998). A transmission-constrained unit commitment method. In System Sciences, 1998., Proceedings of the Thirty-First Hawaii International Conference on (Vol. 3, pp. 71-80). IEEE.

Undeman, E., G. Czub, and M.S. McLachlan, *Addressing temporal variability when modeling bioaccumulation in plants.* Environmental science & technology, 2009. 43(10): p. 3751-3756.

Van de Waterbeemd, H., et al., *Estimation of blood-brain barrier crossing of drugs using molecular size and shape, and H-bonding descriptors.* Journal of drug targeting, 1998. 6(2): p. 151-165.

Veenstra-VanderWeele, J., Christian, S.L. and Cook, Jr, E.H., 2004. Autism as a paradigmatic complex genetic disorder. *Annu. Rev. Genomics Hum. Genet.*, *5*, pp.379-405.

Venayagamoorthy, G., Harley, R., & Wunsch, D. (2002). Comparison of heuristic dynamic programming and dual heuristic programming adaptive critics for neurocontrol of a turbogenerator. IEEE Trans. Neural Netw., 13(3), 764-773.

Wang, J., Shahidehpour, M., & Li, Z. (2008). Security-constrained unit commitment with volatile wind power generation. IEEE Transactions on Power Systems, 23(3), 1319-1327.

Wang, K., Wang, B. and Peng, L., 2009. CVAP: validation for cluster analyses. *Data Science Journal*, *8*, pp.88-93.

Watkins, C. J., and Dayan, P. (1992). Q-learning. *Machine Learning*, *8*(3-4), pp. 279-292.

Wen, Z., O'Neill, D. and Maei, H., 2015. Optimal demand response using device-based reinforcement learning. IEEE Transactions on Smart Grid, 6(5), pp.2312-2324.

Werbos, P. J. (1994). The roots of backpropagation: from ordered derivatives to neural networks and political forecasting (Vol. 1). John Wiley & Sons.

Who.int,. (2015). Retrieved 13 November 2015, from http://www.who.int/water_sanitation_health/e

Wold, S., K. Esbensen, and P. Geladi, *Principal component analysis.* Chemometrics and intelligent laboratory systems, 1987. 2(1-3): p. 37-52.

Wong, P., Albrecht, P., Billinton, R., Chen, Q., Fong, C., Haddad, S., LI, W., Mukerji, R., Patton, D., Schneider, A. and Shahidehpour, M., 1999. A report prepared by the Reliability Test System Task Force of the Application of Probability Methods Subcommittee.

Wood, A. J., & Wollenberg, B. F. (2012). Power generation, operation, and control. John Wiley & Sons.

Xu, R. and Wunsch, D., 2008. *Clustering* (Vol. 10). John Wiley & Sons.

Xu, R. and Wunsch, D.C., 2010. Clustering algorithms in biomedical research: a review. *IEEE Reviews in Biomedical Engineering*, *3*, pp.120-154.

Yesilbudak, M., Çolak, M. and Bayindir, R., 2016, November. A review of data mining and solar power prediction. In Renewable Energy Research and Applications (ICRERA), 2016 IEEE International Conference on (pp. 1117-1121). IEEE.

Yifru, D.D. and V.A. Nzengung, *Uptake of N-nitrosodimethylamine (NDMA) from water by phreatophytes in the absence and presence of perchlorate as a co-contaminant.* Environmental science & technology, 2006. 40(23): p. 7374-7380.

Ylikoski, R., 2000. The relationship of neuropsychological functioning with demographic characteristics, brain imaging findings, and health in elderly individuals.

Yoon, J.M., et al., *Uptake and leaching of octahydro-1, 3, 5, 7-tetranitro-1, 3, 5, 7-tetrazocine by hybrid poplar trees.* Environmental science & technology, 2002. 36(21): p. 4649-4655.

Yu, Y., Luh, P.B., Litvinov, E., Zheng, T., Zhao, J. and Zhao, F., 2015. Grid integration of distributed wind generation: Hybrid Markovian and interval unit commitment. IEEE Transactions on Smart Grid, 6(6), pp.3061-3072.

Zadeh, L.A., *Fuzzy Sets.* Journal of Information and Control, 1965. 8: p. 338-353.

Zeynal, H., Hui, L.X., Jiazhen, Y., Eidiani, M. and Azzopardi, B., 2014, December. Improving Lagrangian Relaxation Unit Commitment with Cuckoo Search Algorithm. In Power and Energy (PECon), 2014 IEEE International Conference on (pp. 77-82). IEEE.

Zhang, X., Bao, T., Yu, T., Yang, B. and Han, C., 2017. Deep transfer Q-learning with virtual leader-follower for supply-demand Stackelberg game of smart grid. Energy, 133, pp.348-365.

VITA

Khalid Al-Jabery was born in Basrah, Iraq in 1983. He received his bachelor and master's degree in Computer Engineering, both with distinction, from University of Basrah - Iraq in 2005 and 2009 respectively. He started his graduate studies with research focused on computer networks and embedded systems in October 2005.

He worked as an IT engineer for the Iraqi-ministry of Oil / South Oil company from 2006 to 2012.

In 2011, he won a scholarship to complete his PhD at Missouri University of Science and Technology and started his Ph.D. program in Computer Engineering. He worked as a research and teaching assistant in the Applied Computational Intelligence Laboratory and Electrical and Computer Engineering department at Missouri S&T in August 2012. His research focused on smart grid, computational intelligence, data analysis, and clustering. He published several papers in IEEE {International Conference on Smart Grid Communications (SmartGridComm), Transactions on Computer-Aided Design of Integrated Circuits and Systems, Annual International Conference of Engineering in Medicine and Biology Society (EMBC), Symposium Series on Computational Intelligence (SSCI), and the American Society of Civil Engineers}. He received his PhD in Computer Engineering from Missouri University of Science and Technology in July 2018.