DATA-DRIVEN PREDICTIVE ANALYTICS FOR WATER INFRASTRUCTURE CONDITION ASSESSMENT AND MANAGEMENT

by

Fang Shi

B. Eng., Fuzhou University, 2016

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF APPLIED SCIENCE

in

THE COLLEGE OF GRADUATE STUDIES
(Electrical Engineering)

THE UNIVERSITY OF BRITISH COLUMBIA
(Okanagan)

September 2018

© Fang Shi, 2018
The following individuals certify that they have read, and recommend to the College of Graduate Studies for acceptance, the thesis entitled:

DATA-DRIVEN PREDICTIVE ANALYTICS FOR WATER INFRASTRUCTURE CONDITION ASSESSMENT AND MANAGEMENT

submitted by Fang Shi in partial fulfillment of the requirements of the degree of Master of Applied Science.

Dr. Zheng Liu, School of Engineering

Supervisor

Dr. Eric Li, Faculty of Management

Co-Supervisor

Dr. Sumi Siddiqua, School of Engineering

Supervisory Committee Member

Dr. Liwei Wang, School of Engineering

Supervisory Committee Member

Dr. Yang Cao, School of Engineering

University Examiner
Abstract

In North American, aging water infrastructure continually challenges communities’ safety. An efficient infrastructure management system is thus required to manage deteriorating pipeline systems. However, there are several challenges in the water infrastructure management. For instance, water systems are often buried underground, making regular inspection and asset condition evaluation difficult. Furthermore, pipe working environment is always complex and not well-understood. In this thesis, data-driven approaches were used to address these challenges.

The first study examines the relationship between soil properties and cast iron water main deterioration condition. The effects of soil properties on pipe failure were visualized and analyzed. Additionally, a stacking ensemble based method was proposed to overcome the drawbacks of the individual models by optimally combining the predictive results from multiple learners. Using soil property data, a single-model and an ensemble-model were developed to predict the pipe condition. The prediction results demonstrated that the proposed ensemble method outperforms the existing single models.

Next, to investigate the structural response of pipelines to varying soil movements, a machine learning based framework was proposed to predict pipe deformations. The critical predictors contributing to the pipe deformations were first identified by random forest-recursive feature elimination algorithm. Super learning based methods were then employed to predict pipe deformations considering the selected feature subsets. Based on the prediction performance, the scalability and superiority of super learning was validated.

Finally, a novel risk analysis approach was investigated by developing both condition rating and consequence models. For condition evaluation, multiple re-
gression analyses were conducted for pipe structural and operational condition rating, respectively. A geographical information system based method was then applied to determine the multi-variant weighting system criteria for economic, operational, environmental, and social impacts. Finally, a risk matrix was used to integrate the results of condition grades and failure consequence scores, allowing the high-risk areas to be identified.

With the data-driven methodology employed in this research, a cloud-based infrastructure management system which can accommodate the data-driven analysis, was designed to support the decision making for pipe system inspection, maintenance, and rehabilitation.
Lay Summary

In 2010, the American Water Works Association estimated that, by 2050, the total cost of pipeline system management will exceed 1.7 trillion dollar. Thus, it is important to have an integrated water infrastructure management system to optimize the rehabilitation process. In this thesis, machine learning based methods were studied for pipe performance prediction and the analytical solutions for pipe failure consequence and risk level evaluation were obtained. With the proposed approaches, a cloud-based framework was designed to aid in a reliable decision making in infrastructure management of buried pipeline networks.
Preface

This thesis is based on the research work completed in the School of Engineering at the University of British Columbia, Okanagan, under the supervision and guidance of Dr. Zheng Liu and Dr. Eric Li. All published works are included in this thesis.

Chapter 3 is based on the following published papers and used with permission of IOS Press and IEEE:


Chapter 4 is based on the following published paper and used with permission of SPIE:


I am the principle contributor for the work. Prof. Zheng Liu and Prof. Eric Li provided me with some advice on research methodology and experiment design. Dr. Yafei Hu, Dr. Huan Liu, Mr. Xiang Peng, and Mr. Yihao Liu helped me proofread the manuscripts.
# Table of Contents

Abstract ................................................................. iii

Lay Summary ............................................................ v

Preface ................................................................. vi

Table of Contents .................................................... vii

List of Tables ......................................................... x

List of Figures ......................................................... xii

Glossary ............................................................... xiv

Acknowledgments ..................................................... xvi

1 Introduction ........................................................ 1
  1.1 Background and Motivation .................................. 1
  1.2 Literature Review ............................................ 2
    1.2.1 Pipe Deterioration Prediction ........................ 2
    1.2.2 Pipe Deformation Prediction ........................ 3
    1.2.3 Risk Analysis for Pipe Management ............... 5
  1.3 Thesis Outline and Contributions ........................ 6

2 Supervised Machine Learning for Predictive Analytics .... 8
  2.1 General Framework of Supervised Learning ............. 8
2.2 Feature Selection Using Wrapper Methods

2.3 Supervised Learning Models for Regression Problems
  2.3.1 Multiple Linear Regression
  2.3.2 Random Forest
  2.3.3 Super Learning (Stacking Ensemble Model)

2.4 Summary

3 Super Learning based Approaches for Pipe Deterioration Prediction
  3.1 The Overall Methodology for Pipe Performance Prediction
  3.2 Pipeline Data Analysis
    3.2.1 Factors Related to Pipe Performance
    3.2.2 Data Collection
    3.2.3 Soil Properties Data Analysis based on Pipe Deterioration Mechanisms
  3.3 Implementation of Machine Learning Prediction Algorithms
    3.3.1 Data Preparation
    3.3.2 Individual Regression Models
    3.3.3 Super Learning for Pipe Deterioration Prediction
  3.4 Experimental Results and Discussion
  3.5 Summary

4 Prediction of Pipe Deformation Using Super Learner and Recursive Feature Elimination Algorithm
  4.1 Data-Driven Approaches for Pipe Deformation Prediction
    4.1.1 Sensor Data Preprocessing and Summarization
    4.1.2 Data Integration with Climate Factors and Data Analytics
    4.1.3 Feature Selection Using Recursive Feature Elimination
    4.1.4 Pipe Deformation Prediction Using Super Learning
  4.2 Implementation of Pipe Deformation Prediction
    4.2.1 Feature Selection with Random Forest-Recursive Feature Elimination
    4.2.2 Super Learning Framework Implementation
  4.3 Experimental Results
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.3.1</td>
<td>Performance Evaluation Metrics</td>
<td>46</td>
</tr>
<tr>
<td>4.3.2</td>
<td>Results and Discussion</td>
<td>46</td>
</tr>
<tr>
<td>4.4</td>
<td>Summary</td>
<td>48</td>
</tr>
<tr>
<td>5</td>
<td>Risk Assessment Model for Pipe Asset Management</td>
<td>51</td>
</tr>
<tr>
<td>5.1</td>
<td>Risk-based Approaches for Pipe Asset Management</td>
<td>51</td>
</tr>
<tr>
<td>5.1.1</td>
<td>Pipe Condition Modeling</td>
<td>53</td>
</tr>
<tr>
<td>5.1.2</td>
<td>Pipe Failure Consequences</td>
<td>57</td>
</tr>
<tr>
<td>5.1.3</td>
<td>Risk Matrix</td>
<td>58</td>
</tr>
<tr>
<td>5.2</td>
<td>Implementation of Risk Analysis Using GIS Techniques</td>
<td>60</td>
</tr>
<tr>
<td>5.2.1</td>
<td>Research Site: The Village of Lumby</td>
<td>60</td>
</tr>
<tr>
<td>5.2.2</td>
<td>Data Collection and Description</td>
<td>62</td>
</tr>
<tr>
<td>5.2.3</td>
<td>Data Summarization for Pipe Condition Analysis</td>
<td>62</td>
</tr>
<tr>
<td>5.2.4</td>
<td>GIS Techniques for Consequences of Failure Analysis</td>
<td>63</td>
</tr>
<tr>
<td>5.3</td>
<td>Experimental Results</td>
<td>65</td>
</tr>
<tr>
<td>5.4</td>
<td>Summary</td>
<td>68</td>
</tr>
<tr>
<td>6</td>
<td>Decision Support for Asset Management: A Cloud-based Framework</td>
<td>70</td>
</tr>
<tr>
<td>6.1</td>
<td>System Architecture of the Cloud-based Platform</td>
<td>70</td>
</tr>
<tr>
<td>6.2</td>
<td>Pipe Condition Prediction Models Deployment in Cloud</td>
<td>72</td>
</tr>
<tr>
<td>6.3</td>
<td>Decision Analytics Implementation in Cloud</td>
<td>73</td>
</tr>
<tr>
<td>6.4</td>
<td>Summary</td>
<td>74</td>
</tr>
<tr>
<td>7</td>
<td>Conclusions</td>
<td>77</td>
</tr>
<tr>
<td></td>
<td>Bibliography</td>
<td>80</td>
</tr>
</tbody>
</table>
# List of Tables

<table>
<thead>
<tr>
<th>Table</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Table 2.1</td>
<td>Summary of machine learning algorithms used in this thesis</td>
<td>10</td>
</tr>
<tr>
<td>Table 2.2</td>
<td>Random forest - recursive feature elimination algorithm</td>
<td>12</td>
</tr>
<tr>
<td>Table 2.3</td>
<td>Super learning algorithm</td>
<td>13</td>
</tr>
<tr>
<td>Table 3.1</td>
<td>Summary of continuous variables</td>
<td>24</td>
</tr>
<tr>
<td>Table 3.2</td>
<td>Pipe condition rating scale based on RWT (Adapted from [1])</td>
<td>25</td>
</tr>
<tr>
<td>Table 3.3</td>
<td>Stacking ensemble models training steps (Adapted from [2])</td>
<td>30</td>
</tr>
<tr>
<td>Table 3.4</td>
<td>Model assessment with $R^2$ and $RMSE$ (Best results are indicated in bold)</td>
<td>31</td>
</tr>
<tr>
<td>Table 4.1</td>
<td>Description of the sensor data</td>
<td>35</td>
</tr>
<tr>
<td>Table 4.2</td>
<td>Summary of the climate data</td>
<td>40</td>
</tr>
<tr>
<td>Table 4.3</td>
<td>Feature importance ranking results (Variables in bold are features selected by RF-RFE)</td>
<td>43</td>
</tr>
<tr>
<td>Table 4.4</td>
<td>Super learning performance comparison with different feature subsets using $RMSE$, $MAE$, and $R^2$ (Results highlighted in boldface are from super learning combined with RF-RFE)</td>
<td>47</td>
</tr>
<tr>
<td>Table 4.5</td>
<td>Prediction performance comparison of base learners and super learners with best performance feature subset using $RMSE$, $MAE$, and $R^2$ (Significant results are highlighted in boldface)</td>
<td>48</td>
</tr>
<tr>
<td>Table 5.1</td>
<td>Major factors and sub-factors influencing pipe structural condition [3] (Factors in bold were used in the regression model)</td>
<td>54</td>
</tr>
<tr>
<td>Table 5.2</td>
<td>Performance values and predetermined weights for the major CoF factors and sub-factors</td>
<td>59</td>
</tr>
</tbody>
</table>
Table 5.3 The final CoF score distribution ........................................ 60
Table 5.4 Risk matrix [4] .............................................................. 60
Table 5.5 Summary of modeled variables ...................................... 62
# List of Figures

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Figure 3.1</td>
<td>Flowchart of stacking ensemble based methodology.</td>
<td>16</td>
</tr>
<tr>
<td>Figure 3.2</td>
<td>Major factors and sub-factors influencing pipeline performance (Factors in bold were used as input variables in modeling).</td>
<td>18</td>
</tr>
<tr>
<td>Figure 3.3</td>
<td>Box plot of soil property data.</td>
<td>21</td>
</tr>
<tr>
<td>Figure 3.4</td>
<td>Histogram of soil type by sample location (M: Silts, C: Clays, S: Sands, G: Gravels).</td>
<td>22</td>
</tr>
<tr>
<td>Figure 3.5</td>
<td>Machine learning experiment flowchart.</td>
<td>23</td>
</tr>
<tr>
<td>Figure 3.6</td>
<td>Chart of correlation matrix for soil property data and RWT. On the bottom of the diagonal: the bivariate scatter plots with a fitted line. On the top of the diagonal: the value of the correlation plus the significance level as stars.</td>
<td>26</td>
</tr>
<tr>
<td>Figure 3.7</td>
<td>Individual model prediction performance comparison.</td>
<td>29</td>
</tr>
<tr>
<td>Figure 4.1</td>
<td>Data-driven approaches for pipe structural behavior prediction.</td>
<td>34</td>
</tr>
<tr>
<td>Figure 4.2</td>
<td>Daily time series plot of strains and climate data (The blue smooth lines describe the curve trend).</td>
<td>37</td>
</tr>
<tr>
<td>Figure 4.3</td>
<td>Daily time series plot of soil properties.</td>
<td>39</td>
</tr>
<tr>
<td>Figure 4.4</td>
<td>Chart of correlation matrix for soil properties and climate factors. On the bottom of the diagonal: the value of the correlation coefficient. On the top of the diagonal: correlation visualization using circles of different size and colors. Color bar on the right: correlation value scale and corresponding color.</td>
<td>42</td>
</tr>
<tr>
<td>Figure 4.5</td>
<td>RF-RFE results for feature selection (Blue solid points indicate the optimum number of variables used for the prediction).</td>
<td>44</td>
</tr>
<tr>
<td>Figure</td>
<td>Description</td>
<td>Page</td>
</tr>
<tr>
<td>--------</td>
<td>--------------------------------------------------------------------------------------------------------</td>
<td>------</td>
</tr>
<tr>
<td>4.6</td>
<td>Prediction results of six different super learning models.</td>
<td>49</td>
</tr>
<tr>
<td>5.1</td>
<td>Flowchart for pipe risk analysis modeling.</td>
<td>52</td>
</tr>
<tr>
<td>5.2</td>
<td>Factors affecting pipe operational condition [5].</td>
<td>56</td>
</tr>
<tr>
<td>5.3</td>
<td>Chart of correlation matrix for feature data (RT: Road Type, ST: Service Type). Scatter plots with</td>
<td>63</td>
</tr>
<tr>
<td></td>
<td>a fitted line. On the top of the diagonal: the value of the correlation plus the significance level</td>
<td></td>
</tr>
<tr>
<td></td>
<td>as stars.</td>
<td></td>
</tr>
<tr>
<td>5.4</td>
<td>Social and environmental impact mapping visualization.</td>
<td>64</td>
</tr>
<tr>
<td>5.5</td>
<td>Number of pipes with different condition grades.</td>
<td>66</td>
</tr>
<tr>
<td>5.6</td>
<td>Number of pipes with different scores in terms of their CoF impacts.</td>
<td>67</td>
</tr>
<tr>
<td>5.7</td>
<td>Number of pipes in different CoF score categories.</td>
<td>67</td>
</tr>
<tr>
<td>5.8</td>
<td>Number of pipes in different risk.</td>
<td>68</td>
</tr>
<tr>
<td>6.1</td>
<td>Cloud-based data management framework for decision analytics.</td>
<td>71</td>
</tr>
<tr>
<td>6.2</td>
<td>ER modeling with UML diagram.</td>
<td>72</td>
</tr>
<tr>
<td>6.3</td>
<td>The predictive experiment based on saved training model using web service.</td>
<td>73</td>
</tr>
<tr>
<td>6.4</td>
<td>Dashboard design for decision support on asset management.</td>
<td>75</td>
</tr>
<tr>
<td></td>
<td>Histogram in the left: frequency of different condition, failure consequence, and risk level pipes.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Map in the upper center: location of different risk level pipe. Donut in the bottom center:</td>
<td></td>
</tr>
<tr>
<td></td>
<td>proportion of different condition, failure consequence, and risk level pipes. Table in the right:</td>
<td></td>
</tr>
<tr>
<td></td>
<td>list of 'high' and 'very high' risk pipes.</td>
<td></td>
</tr>
<tr>
<td>6.5</td>
<td>Dashboard design for pipe structural health monitoring with real-time series plot.</td>
<td>76</td>
</tr>
<tr>
<td></td>
<td>Left: sensor network data including pipe strain, soil temperature, soil pressure, and soil water</td>
<td></td>
</tr>
<tr>
<td></td>
<td>contents. Right: climate condition data including precipitation and rainfall deficit.</td>
<td></td>
</tr>
</tbody>
</table>
## Glossary

**AHP**  Analytical Hierarchy Process  
**ALS**  Average Longitudinal Strain  
**ANN**  Artificial Neural Networks  
**API**  Application Programming Interface  
**ASCE**  American Society of Civil Engineers  
**BBN**  Bayesian Belief Network  
**CCTV**  Closed-Circuit Television  
**CIT**  Conditional Inference Tree  
**CoF**  Consequences of Failure  
**CS**  Circumferential Strain  
**DLS**  Differential Longitudinal Strain  
**ER**  Entity-Relationship  
**FD**  Frost Depth  
**FEA**  Finite Element Analysis  
**GIS**  Geographical Information System  
**GLM**  Generalized Linear Model
**GRNN**  General Regression Neural Network

**KNN**  K-nearest Neighbors

**MAE**  Mean Absolute Error

**MLP**  Multilayer Perceptron

**MLR**  Multiple Linear Regression

**PC**  Pressure Cell

**PPT**  Precipitation

**RD**  Rainfall Deficit

**RF**  Random Forest

**RFE**  Recursive Feature Elimination

**RF-RFE**  Random Forest-Recursive Feature Elimination

**RMSE**  Root-Mean-Square Error

**RT**  Regression Tree

**RWT**  Remaining Wall Thickness

**SG**  Strain Gauge

**SVM**  Support Vector Machine

**T**  Temperature

**TC**  Thermocouple

**UML**  Unified Modeling Language

**WCR**  Water Content Reflectometer

**WSA**  Winkler Spring Approach
Acknowledgments

I deeply acknowledge my supervisor Prof. Zheng Liu and co-supervisor Prof. Eric Li whose advice, encouragement, and support made this work possible. Not only did they grant me a great freedom in my study, but they also delivered me lots of academic knowledge. It is my honor to study and work under their guidance.

Thanks also to committee members Prof. Liwei Wang and Prof. Sumi Siddiqua for their crucial role in completing my thesis.

I would also like to thank my dear colleagues Mr. Yihao Liu, Mr. Shuo Liu, Mr. Xiang Peng and Dr. Huang Liu for sharing their academic experiences and constructive viewpoints generously with me during our discussions. Also, I need to thank Dr. Yafei Hu from City of Regina, Mr. Max Wang from Stanford University, and Mr. Mamiya, a Ph.D. student at McGill University. It is their brilliant insights that helped me overcome some obstacles in my study.

Finally, sincere thanks to my family and friends for their understanding, encouragement and support in all these years.
Chapter 1

Introduction

1.1 Background and Motivation

The water distribution system is a complex network consisting of a large number of underground pipelines that play an important role in maintaining public health and environment condition. As one of the most important infrastructures in cities, underground water mains require an efficient infrastructure management system to manage complex water distribution networks.

According to the American Society of Civil Engineers (ASCE) infrastructure 2013 report, in the United States, 240,000 water mains breaks occur every year [6]. Since pipeline failure can result in high maintenance costs and pose serious threats to human health and the environment, it is important to develop the technology for water mains condition assessment to avoid any catastrophic failures and associated risks and costs. However, accessibility of buried pipelines remains a practical challenge for directly predicting water mains condition. For instance, the technologies frequently used for this purpose, such as Closed-Circuit Television (CCTV), are not always applicable to inspecting the condition of water mains [7]. Hence, in this study, the data-driven indirect approaches were explored by developing predictive models to assess the pipe performance and analyzing the consequence of pipeline failure to calculate the risk index of each pipe for decision making.

In this thesis, the researches focus on the key components in the decision support system for integrated asset management, i.e., pipe deterioration condition pre-
diction (Chapter 3), pipe deformation behavior modeling (Chapter 4), and risk-based decision analytics (Chapter 5). In the following sections, a detailed description of existing methods and a brief overview of major challenges associated with them are provided.

1.2 Literature Review

1.2.1 Pipe Deterioration Prediction

For water mains assessment, statistical modeling remains the most cost-effective analytical approach. The multivariate exponential model generalized by Kleiner et al. [8], which considered time-dependent variables, allows for simple evaluation and variation prediction of pipe breaks. This model was developed in 2011 and was validated by comparing its performance with that of a transition state-life regression model [7]. More recently, a statistical model, Bayesian Belief Network (BBN), was proposed, whereby soil properties served as predictors of the remaining service life of water mains [9]. The BBN model determined soil corrosion index by sorting the soil parameters into major and minor groups. Subsequently, by combining this approach with a mathematical model, the safety index was calculated using pit depth, corrosion initiation time, and pipe wall thickness as inputs to the BBN. The statistical approaches mentioned above could aid in pipeline management by predicting the remaining service life or the break rate of water mains. However, the complexity of water networks has also led to the growing reliance on data mining techniques as a means of predicting pipe failures [10].

In extant studies, data mining techniques and Artificial Neural Networks (ANN) methods have been employed in pipe condition prediction [11-14]. To enhance pipe failure prediction, a Multilayer Perceptron (MLP) model was developed and investigated by Achim et al. in 2007 [12]. This model achieved 19% improvement in terms of accuracy when compared with a shifted time exponential model. This result has led to the development of a computational model integrating Analytical Hierarchy Process (AHP) and ANN [13]. After utilizing AHP to determine the weights of factors and sub-factors, the supervised ANN and back propagation algorithm were used to predict pipe performance via the deterioration rate. In 2009,
Fahmy and Moselhi applied multiple regression, MLP, and General Regression Neural Network (GRNN) to forecast the remaining useful life of cast iron water mains [14]. Considering two rehabilitation strategies: cement mortar lining and cathodic protection, Asnaashari and colleagues successfully predicted water mains failure rates using an ANN-based model in 2016 [15].

However, the ANN model may not always be the most optimal choice. Recently, when GRNN and feed forward neural networks were used to estimate the pipe failure rates, their performance was inferior to that of a model combining fuzzy clustering and least squares support vector machine [16]. These findings indicate that individual models are not universally applicable in all cases. Hence, a stacking ensemble prediction method was proposed in Chapter 3 to achieve a stronger overall prediction in comparison with the existing models. This method relies on multiple models, whereby a meta-level algorithm is trained to learn the most optimal way to combine different advanced models for each specific case.

1.2.2 Pipe Deformation Prediction

In North America, there are many water distribution networks locating in expansive soils, which has a significant impact on the buried pipe performance [17]. For instance, due to the soil moisture-suction variation, the expansive soil can undergo the extreme shrink-swell movement, which can cause pipe deformation and result in pipe breakage finally [18]. Especially in the case of small diameter pipes (< 200 mm) [19], pipe failure rates are directly affected by the shrink/swell behavior of soil [17]. Hence, it is critical to perform researches to understand the complex pipe-soil interactions [20] and analyze the pipe strain for pipe health monitoring [21].

Based on existing studies, the soil-pipe interaction was modeled using two methods primarily, the Winkler Spring Approach (WSA) and the Finite Element Analysis (FEA) [19]. The WSA was proposed by Winkler in 1876 [22], simulating the pipe-soil interaction by nonlinear soil springs in the axial, lateral, and vertical directions [23]. Although WSA has been advanced to reflect the different physical aspects of the soil-structure interaction, it neglects the impact of rigid soil movements and interaction through soil from location to location [24]. Accordingly, Trickey and Moore implemented FEA to observe the three-dimensional response of
buried pipes in 2017 [24]. In 2014, a Gurson-Tvergaard-Needleman model-based FEA methodology was developed to predict the pipeline tensile strain capacity, which was validated against 4 full-scale tests [25]. However, through a parametric study based on the methodology, the developed equation for the strain prediction is very complex and not well understood. Additionally, the data required to model these physical mechanisms are always rarely available and prohibitively costly to acquire [26, 27]. Although a relatively simplified modeling approach, a three-dimensional finite difference continuum model, was used to analyze pipe behavior in the expansive soil [28], the model was only validated by the experiment results from the laboratory, which was still full of uncertainties when applied practically. Therefore, further studies are needed to better understand the field performance of buried pipes influenced by surrounding unsaturated soil conditions [19].

In addition to soil properties, the local climate is identified as one of crucial factors influencing pipe failure rates. Gould et al. found that seasonality exists in the pipe failure rate database and indicated that pipe failures occur due to a complex interaction of different factors including pipe attributes, soil properties and weather conditions [29]. Furthermore, Rajeev et al. also identified the relationship between climate change and expansive soil volume variation, which results in the majority of the pipe damage in shrink–swell soil [18]. To investigate the effects of climate factors on pipe failure, a number of studies have been performed to predict pipe failure using climate-related data [30–33]. However, these methods are dependent on limit historical pipe breakage records which is always rarely available [34]. And the impacts of climate factors were not characterized clearly as their contribution to the results of pipe failure is vague.

For these purposes, a data-driven approach was investigated in Chapter 4 to characterize pipe structural behavior using soil property data collected from a sensor network along with daily climate data. Machine learning techniques with capability of exploring the hidden relationships in datasets were applied in the proposed framework for pipe structural health assessment, which have been widely used in corrosion monitoring [35], damage detection [36], soil corrosivity analysis [37], and pipe deterioration evaluation [11]. With the objective of building a robust and reliable model for pipe deformation prediction, the machine learning based framework integrates feature selection methods and super learning algorithms for the
optimal feature subset selection and prediction results combination. Since cost of sensors limits the number of deployed sensors [38], the proposed methodology can also advance the sensing systems by filtering the redundant features and hence, reducing the use of some sensor nodes.

1.2.3 Risk Analysis for Pipe Management

Reliable risk analysis of a pipe network facilitates proactive asset management, which in turn helps achieve an optimal balance between system cost and performance [39]. Therefore, it is necessary to develop a risk assessment model that can be employed in managing pipe inspection, rehabilitation, and replacement. Generally, pipe risk analysis consists of pipe condition predication and pipe failure consequence analysis.

When conducting pipe performance assessment, many utility service providers rely on pipe survey techniques, including CCTV inspection, which is one of the most commonly used approaches to identify defects in pipes [40]. However, CCTV reports require manual intervention and human expertise to identify the defects and determine their condition [41]. Hence, many machine learning based techniques have been developed to classify the defects in pipes automatically [42, 43]. Moreover, several protocols, including the Pipeline Assessment and Certification Program rating systems and Water Research Center protocols, are presently used to assist in sewer pipe condition rating based on CCTV reports. However, as CCTV technology can only provide the view of pipes internal surface, no structural data on the pipe wall integrity can be obtained [44].

To predict the pipe condition more accurately and economically, several researchers have adopted the statistics based approaches, which generally follow a similar progression of steps: collection and identification of assets and available data, data analysis, modeling and model evaluation. Chughtai and Tayed [5] used multiple regression techniques to evaluate structural and operational condition of sewer pipes while considering various physical, environmental, and operational factors that have 82 – 86% accuracy. Regression models were subsequently utilized to generate deterioration curves for sewer pipes. In a recent study, Bakry et al. [3] employed regression analysis technique to predict pipe condition using vari-
ables, such as pipe material and rehabilitation type, along with data based on CCTV inspection reports. The complexity of wastewater networks has also prompted utilization of data mining techniques in pipe failure prediction. For instance, in 2005, Kulandaivel [45] developed an ANN model for sewer pipe condition prediction based on previously acquired field data. Based on machine learning techniques, Support Vector Machine (SVM) models were also applied to predict sewer condition with 91% accuracy [46].

Although pipe condition prediction has been the subject of extensive research, failure consequences, as the main factor of a typical risk analysis program, remain insufficiently explored [47]. Baah et al. [4] used a risk matrix and a weighted sum multi-criteria decision method to assess the consequences and risks associated with pipe failure. However, the authors only considered the social and economic impacts of pipe failure without analyzing the operational impact of the pipeline system failure. Anbari et al.’s [47] study suffered from similar deficiencies, as failure consequences were classified based solely on their impact (direct or indirect). More recently, Hu et al. [48] proposed a consequence model based on the impact of failure, which was classified into four categories, namely economic, operational, environmental, and social. However, only the multi-variant weightings for economic category are described in the article. To address issues noted above, a novel risk was proposed analysis approach by developing both condition rating (structural and operational condition) and consequence (economic, operational, social, and environmental impact) models in Chapter 5.

1.3 Thesis Outline and Contributions

This thesis is organized into seven chapters. Chapter 1 gives the background information and motivation of this research. Furthermore, this chapter gives a comprehensive literature review of existing methods, and describes challenges for the rest of this thesis.

Chapter 2 provides detailed technical background for the supervised machine learning techniques including feature selection methods, regression analysis and ensemble algorithms.

In Chapter 3, a stacking ensemble based approach is proposed for cast iron pipe
deterioration prediction. The factors related to pipe performance were categorized and summarized first. After analysis of the relationship between soil properties and pipe break history, an ensemble-model was built to predict the pipe condition considering the soil property data. Additionally, single models were also constructed as a comparison. Finally, the superiority of the proposed ensemble method was verified through its lowest value in the root-mean-square error relative to the individual models. The techniques presented in this work can aid in a reliable decision making in infrastructure management of buried pipeline networks.

In Chapter 4, a data-driven approach is proposed for pipe deformation prediction, where the relationship between pipe deformations and environmental factors including soil properties and weather conditions was analyzed and characterized. Using field monitoring data, random forest - recursive feature elimination algorithm was first investigated for the optimal feature subset selection. Super learning was then applied to predict the pipe structural behavior. The experimental results were finally presented to validate the accuracy and the scalability of proposed methods.

Chapter 5 presents a risk analysis study for water distribution network system management. To determine a variety of risk levels associated with pipe failures, the pipe condition rating was first calculated using data collected from asset management database. Using Geographical Information System (GIS) techniques, the consequence of pipe failure was analyzed from economic, operational, environmental, and social aspects. Finally, the risk level of each pipe was obtained by combining the results of pipe condition rating and consequence of failure analysis with a risk matrix. The proposed methods can support operational decisions pertaining to the maintenance and future development of the pipe infrastructure.

Based on the findings from the researches described in Chapters 3, 4, and 5, a cloud-based framework was proposed in Chapter 6 to provide solutions to implement decision support system for infrastructure monitoring and management. The system design including database modeling, pipe condition prediction model deployment, and dashboard development was also provided.

Chapter 7 draws conclusions for this thesis, where future works are also suggested.
Chapter 2

Supervised Machine Learning for Predictive Analytics

In this chapter, the supervised machine learning techniques on regression problems for the predictive analytics is reviewed. Section 2.1 gives an overview of supervised learning application framework. Subsequently, two key steps in machine learning application, feature selection and modeling, are presented and the fundamental concepts of these algorithms are provided, which was described in Section 2.2 and Section 2.3 respectively. This chapter serves as a technical background for audience better understanding the studies presented in Chapters 3 and 4.

2.1 General Framework of Supervised Learning

Machine learning referring to a vast set of tools for understanding data can be classified as supervised and unsupervised [49]. Supervised learning involves modeling for learning a function mapping one or more inputs to an output based on data with input-output pairs for prediction purpose [50]. There are a number of applications for supervised learning, the most significant of which is predictive modeling, which widely applied in science, finance and industry [51]. In our research, we focus regression predictive problems on non-destructive evaluation and structural health monitoring using supervised learning.

Generally, the process of supervised learning project usually involves a number
of steps as following:

- Define the problem. First step of a machine learning application is to identify the project scope and the problems going to be tackled. This step determines the data that will be collected and used in the modeling and the performance metrics. In our study, the problem includes pipe condition assessment such as pipe deterioration prediction and pipe deformation evaluation.

- Collect data. The collected data used for model training and testing is a set of input variables gathered with corresponding response outputs, either from human experts or from measurements. The quality of the data set plays a key role in machine learning.

- Select the best feature subset. For supervised learning, features selection is always necessary when the number of features is huge or when the relationship between the features and the response variable is vague. The accuracy of the trained model highly relies on how the input object is represented. In this chapter, feature selection methods will be further discussed in Section 2.2.

- Determine the machine learning algorithm. A great number of supervised learning methods have been introduced in the past twenty years [52]. A comparison study is necessary to determine the best fitting models. There are several factors always considered in algorithm selection, i.e., accuracy, scalability, and interpretability. Each learning algorithm will be trained on the gathered training data set and usually certain control parameters should be adjusted via cross-validation. Table 2.1 lists all supervised learning algorithms used in this study along with a brief description, three of which are explained in detail in Section 2.3.

- Evaluate the trained model performance. After model training, the performance of these supervised learning methods should be evaluated on a test set that is separate from the training set using some performance criteria. Regarding to regression problems for predictive analysis, r-squared, root mean square error, and absolute mean error are commonly used for prediction accuracy evaluation.
Table 2.1. Summary of machine learning algorithms used in this thesis

<table>
<thead>
<tr>
<th>Supervised Learning Algorithms</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Artificial Neural Network (ANN)</td>
<td>ANN is a rough model based on the principle of biological neural networks, which comprises of an extensive number of artificial neurons arranged in different layers (input layer, hidden layer and output layer) [53].</td>
</tr>
<tr>
<td>Conditional Inference Tree (CIT)</td>
<td>CIT is one possible decision tree algorithm for recursive binary splitting that uses a significance test procedure for variable selection too avoids bias [54].</td>
</tr>
<tr>
<td>K-nearest Neighbors (KNN)</td>
<td>KNN is a non-parametric model consisting of the k closest training examples in the feature space [55]. The output is the average value of the object k nearest neighbors.</td>
</tr>
<tr>
<td>Multilayer Perceptron (MLP)</td>
<td>MLP is a class of feedforward ANN using backpropagation for training each node is a neuron that uses a nonlinear activation function [56].</td>
</tr>
<tr>
<td>Multiple Linear Regression (MLR)</td>
<td>MLR is a parametric learning algorithm used to explain the relationship between one continuous respond variable and two or more explanatory variables by fitting a linear equation [49].</td>
</tr>
<tr>
<td>Random Forest (RF)</td>
<td>RF is a bagging ensemble learning method by constructing a multitude of decision trees at training time and outputting the class that is the mean prediction of the individual trees [57,58].</td>
</tr>
<tr>
<td>Regression Tree (RT)</td>
<td>RT is a decision tree algorithm for regression problems [11], employing information measures for variable selection.</td>
</tr>
<tr>
<td>Super Learning/ Stacking Ensemble</td>
<td>Super learning is a stacking ensemble learning algorithm that utilizes the strength of individual methods by optimally combining the prediction results from a group of candidate algorithms through a meta learning process [59].</td>
</tr>
<tr>
<td>Support Vector Machine (SVM)</td>
<td>SVM is a universal tool for solving multidimensional function estimation problems that perform non-linear functions by usage of kernel trick [60].</td>
</tr>
</tbody>
</table>

2.2 Feature Selection Using Wrapper Methods

Feature selection refers to the process of identifying and removing as many irrelevant and redundant variables as possible [61]. It is often an important step in applications of machine learning methods as collected data sets often have many variables that are irrelevant or redundant [62]. The advantages of variable selection mainly include improving the prediction performance of the predictors, reducing the algorithm training time, and providing a better understanding of the underlying process that generated the data [63].

Generally, the feature selection methods could be classified into three cate-
gories mainly: filter, embedded, and wrapper methods. Filter methods, generally used as a preprocessing step, rely on the general characteristics of the training data which is independent of machine learning algorithms [64]. Embedded methods are usually implemented by a specific machine learning algorithm that has its own built-in variable selection models. In a wrapper method, learning machine of interest is used as a black box returning a feature ranking based on the variable predictive power [63]. In this section, a detailed description of a wrapper method, Recursive Feature Elimination (RFE), is provided, which was used in the pipe deformation prediction study described in Chapter 4.

RFE is a greedy optimization algorithm based on feature ranking techniques [65]. In this study, RFE method was combined with RF variable importance analysis to select the best feature subset. Table 2.2 provides the algorithm details of Random Forest-Recursive Feature Elimination (RF-RFE). Beginning with using all predictors as inputs, RFE iteratively constructs models by excluding the worst performing feature, which are ranked according to its importance to the model. At each iteration, the variable rankings are recomputed by RF, which algorithm details are provided in Section 2.3. The final ranking of the predictors are based on the order of their elimination [66].

2.3 Supervised Learning Models for Regression Problems

2.3.1 Multiple Linear Regression

Multiple linear regression is a very straightforward approach for predicting a quantitative response $Y$ on the basis of a group of predictor variables $(X_1, X_2, ..., X_i)$, which assumes there is approximately a linear relationship between variables and $Y$ [49]. Mathematically, MLR model is formulated as Eq. 2.1.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_i X_i + \epsilon$$  \hspace{1cm} (2.1)

where $\epsilon$ is the notation of the model deviation and the subscript $i$ refers to the $i_{th}$ predictor. $\beta_0, \beta_1, \beta_2, ..., \beta_i$ are the coefficients of the linear regression model. The most common approach to determine the regression coefficients is minimizing the
random forest - recursive feature elimination algorithm

**Algorithm 1: RF-RFE**

1. Train RF on the data $D = \{x_i, y_i\}_{i=1}^n$ using all predictors
2. Evaluate RF performance
3. Rank the importance of variables
4. **for Each subset size** $S_i, i = 1...n$ **do**
   5. Exclude the least important feature
   6. Train RF with $S_i - 1$ predictors
   7. Evaluate RF performance
8. **end**
9. Calculate the performance profile over the $S_i$
10. Find the best performance feature subset

least-square criterion. MLR is still a useful and widely used statistical learning method, implementation of which is presented in Chapters 3, 4, and 5.

2.3.2 Random Forest

Random forest is an extremely popular and powerful machine learning algorithm, which is a type of ensemble algorithm called bootstrap aggregation or bagging. To construct more powerful prediction models, RF combines multiple decision tree models built from different subsamples of the training dataset by averaging the resulting tree outputs. The advantages of the RF include reduction in overfitting and less variance. In RF, a number of decision trees on bootstrapped training samples are built first, in which a random number of $m$ predictors are chosen as split candidates from the all $p$ features (Typically, $m \approx \sqrt{p}$). This prevents the built bagged trees being highly correlated, which reduces the variance over a single tree in this setting [49]. In this thesis, RF algorithm was studied and explored in the researches presented in Chapters 3 and 4.
2.3.3 Super Learning (Stacking Ensemble Model)

Different from RF that integrates multiple same type models from different sub-samples, super learning is a stacking ensemble approach which combines multiple different type of models through a meta-learning process. Table 2.3 describes the super learning algorithm. Prior to the construction of super learner, a variety of individual machine learning algorithms were evaluated by k-fold validation.

**Table 2.3. Super learning algorithm**

```
Algorithm 2: Super learning

1. Split data $D = \{x_i, y_i\}_{i=1}^n$ into $K$ blocks for validation: $V_i, i = 1,2\ldots k$
2. for Each candidate learner $H_i, i = 1\ldots n$ do
3. for Each validation block $V_i, i = 1,2\ldots k$ do
4. Train candidate learner $H_i$ on $D$ excluding the validation block $V_i$
5. Predict the target value $y$ in the validation block $V_i$ based on $H_i$
6. end
7. end
8. Construct level-one data $D_1 = \{x'_i, y_i\}_{i=1}^n$, where $x'_i = \{V_1(y)\ldots V_k(y)\}$
9. Train the meta learner $M$ on $D_1$ to determined the weight of each candidate learner
10. Train each candidate learner $H_i, i = 1\ldots n$ on full dataset $D$
11. Combine predictions from $H_i, i = 1\ldots n$ with $M$
12. Return super learner $S$ comprising of $H_i, i = 1\ldots n$ and $M$ models
```

After the selection of the candidate algorithms, the first step of super learning development is to evaluate the prediction performance of the candidate learners by k-fold cross-validation. For each candidate learner, the predicted results for each validation sets were constructed as variables for meta-learner training, where the best weighted combination of the base learners were determined by cross-validated risk, i.e., the average error of cross validation. The trained meta learner is then used to combine the base learners trained with the full data set. The super learner is the combination of these base learners with determined weights [67]. Due to the superior prediction ability of super learning, this modeling method was employed in both the pipe deterioration rate evaluation and pipe deformation prediction given
in Chapters 3 and 4.

2.4 Summary

This chapter presents the supervised machine learning theory that provides a theoretical framework for the researches presented in Chapters 3 and 4. In this chapter, feature selection methods were also discussed and the random forest - recursive feature elimination algorithm was described in detail. Additionally, three regression models that play an important role in our supervised learning applications were introduced.
Chapter 3

Super Learning based Approaches for Pipe Deterioration Prediction

In this chapter, a super learning based method, also known as stacking ensemble learning, is proposed for the pipe deterioration prediction, which is structured as follows. The overall structure of the stacking ensemble learning method is presented in Section 3.1 where a brief overview of the pertinent background information is also given. In Section 3.2, the main factors affecting pipe performance are discussed, along with the relationship between soil properties and pipe failure. The classic machine learning and super learning techniques presently used for pipe performance prediction are described in detail in Section 3.3. In Section 3.4, the evaluation metrics and the experimental results are discussed, as a means of demonstrating the effectiveness of the super learner when applied in pipe performance prediction.

3.1 The Overall Methodology for Pipe Performance Prediction

Condition assessment provides the critical information needed to assess the physical condition and functionality of a water mains system. It also allows the remain-
Figure 3.1. Flowchart of stacking ensemble based methodology.
ing service life and asset value to be accurately estimated. A number of individual models have been applied to predict water mains performance [11, 14, 15, 68–70]. However, model-based determination outcomes depend on the type and quality of the data set to be analyzed. As individual models are not always optimal in all cases, model ensemble is a preferred option, as they allow the relevant aspects of all incorporated models to be used in each individual scenario [71]. Hence, to adapt to different data sets, instead of specifying an algorithm that provides the best fit to the input data, in this work, a powerful stacking ensemble learning framework was employed. This approach allows the model to learn the optimal combination of multiple individual candidate learners to increase the scalability of the model to various data generating distributions. The detailed explanation of stacking ensemble is given in Section 3.3.

Figure 3.1 illustrates the structure of computational framework adopted in this study. Based on a comprehensive literature review, a number of factors related to water mains performance were identified. Moreover, a list of algorithms aligned with those employed in pertinent studies were specified as base learner candidates, such as MLR, RF, ANN, and SVM. By analyzing the relationship between soil properties and deterioration of metallic pipes, these regression models were developed and their ability to predict pipe condition using collected soil data was investigated. Based on the comparison of their prediction performance, which was conducted in the test phase, a set of base learners which would be considered in the super learning were selected and the cross-validated predicted values from these learners were collected for the meta-learner training. Additionally, the prediction results of these individual models were compared with those yielded by the stacking ensemble, thus allowing the superiority of super learner to be experimentally verified.

### 3.2 Pipeline Data Analysis

#### 3.2.1 Factors Related to Pipe Performance

Figure 3.2 presents the factors, which could potentially influence the deterioration of water mains. These factors are commonly classified into physical, environment-
tal, and operational categories.

Figure 3.2. Major factors and sub-factors influencing pipeline performance
(Factors in bold were used as input variables in modeling).

Physical Factors

Physical factors such as pipe age, wall thickness, and pit depth are widely used to determine pipe deteriorate rate [11] or the remaining service life [9], which were represented as assessment outcomes. Moreover, pipe length is included to determine failure rate as well. The fact that pipe age, length, diameter etc. are generally well recorded in most municipalities solves the accessibility problem of data [15].
Environmental Factors

Since corrosion is one of the critical factors contributing to pipe failure and aggressive soil environments results in corrosion, several studies have made efforts to calculate the soil corrosivity based on soil properties [37]. Soil resistivity is often cited as the key contributor to pipe deterioration [11]. The groundwater level, traffic type, location, and service type are also used as input variables, if available, to increase the accuracy of predictive models [13]. However, reliable data for estimating the values of these factors cannot be easily acquired in most cases [12]. Most authors also highlight the importance of temperature in pipe system performance [7]. For example, Rajani and colleagues have explored the relation between pipe break and temperature covariates to assess the condition of water mains [72].

Operational Factors

Operational aspects play an important role in the analysis of pipe infrastructure failure rates, since the break category and cathodic protection were ranked as the first and the fourth most essential independent variables in the water mains failure prediction [15]. Moreover, the historical break and maintenance records are necessary for data analysis [10].

3.2.2 Data Collection

The data used in this study were collected from the City of Toronto between January 1998 and October 1999, and pertain to 108 pipe and 98 soil samples [73]. Therefore, the data set mainly includes the information related to pipe physical features and soil properties. In addition, the majority of the data samples were taken from the locations at which water mains breaks occurred, while the remaining samples were taken from access pits dug during cement mortar lining operations [73].

3.2.3 Soil Properties Data Analysis based on Pipe Deterioration Mechanisms

Pipeline corrosion is one of the key factors resulting in the water mains’ failure. Corrosion is classified into internal corrosion (caused by water flowing through the pipe) and external corrosion (which is caused by the soil surrounding the pipe) [74].
The most direct outcome of corrosion are the corrosion pits in pipe structure. Thus, by measuring the corrosion pit properties, the relationship between soil characteristics and water mains structural condition can be explored, as was done in this study.

Through direct interaction with pipes, soil environment causes pipe corrosion, mainly by electrochemical reactions, including galvanic corrosion cells, electrolytic corrosion cells, bacterial corrosion and acid attacks [73]. Bacterial corrosion is also common, as soil environment with high sulfide content promotes the growth of the sulfite reducing bacteria which cause bacterial corrosion, as shown in Fig. 3.3a. However, in the analysis conducted as a part of this investigation, these effects are identified as outliers.

In the case of electrochemical corrosion, soil resistivity is the main factor in determining the pipe deterioration rate, and is thus an important parameter in pipeline construction [75]. As shown by Fig. 3.3b, the soil samples collected from water mains break locations have much lower resistivity, indicating that pipe corrosion increases as the resistivity declines. On the other hand, as the main factor determining the corrosion current density [76], the moisture content of the soil samples collected at pipe break locations has higher value, as shown in Fig. 3.3c. However, the correlation between moisture content and corrosion has not yet been established and proved with sufficient degree of scientific certainty.

Figure 3.3d pertains to the effects of acid attacks, indicating that this corrosion type may not be applicable in this case, since the pH measurements were either neutral or alkaline in nature, and thus have negligible effect on soil corrosivity.

In the present study, soil type was classified as silts, clays, sands, or gravels, based on the Unified Soil Classification System [74]. Figure 3.4 indicates that most of the data collected from the break locations pertains to clay soil environment, while data collected from access pits samples was indicative of sand soil. Thus, it can be seen that clay soil is more corrosive than sand soil. According to the information presented in the box plots in Fig. 3.3c and Fig. 3.3f, the low resistivity and high moisture in clay soil environment may be the reason for its high corrosivity.
Figure 3.3. Box plot of soil property data.

(a) Soil sulphide by sample location.

(b) Soil resistivity by sample location.

(c) Soil moisture by sample location.

(d) pH value by sample location.

(e) Resistivity by soil type.

(f) Moisture by soil type.
3.3 Implementation of Machine Learning Prediction Algorithms

Figure 3.5 provides a detailed description of the machine learning procedure performed in this research. As indicated by the flowchart, after data preparation, the collected data was randomly divided into two sets, whereby 80% of the data set was used to train the models and the remaining 20% was reserved for testing. Using soil resistivity, pH value, sulfide, moisture, soil type, soil corrosivity and pipe wall thickness as input variables, along with Remaining Wall Thickness (RWT) as the output variable, five machine learning models were built using the same training data set. Since k-fold cross-validation with moderate k value (10 ∼ 20) reduces the variance while increasing the bias [77], 15-fold cross-validation was performed to evaluate the training performance of each algorithm more accurately. Subsequently, the trained models re-evaluated by comparing the predicted RWT values and the original response vectors in the testing data set. After finding the best-fitting model, the RWT can be estimated to prioritize the water mains system maintenance and replacement, when new input predictors were collected.
Figure 3.5. Machine learning experiment flowchart.

Input Variables:
1. Resistivity
2. PH
3. Sulfide
4. Moisture
5. Soil type
6. Soil Corrosivity
7. Pipe wall thickness

Target: Condition (RWT)
3.3.1 Data Preparation

Variables and Data Summarization

Continuous variables considered in this prediction modeling are presented in Tab. 3.1, along with pertinent statistical information, such as mean, median, and quartiles. As shown in Tab. 3.1, resistivity has a much greater value compared to the other variables, which should be regularized. Thus, in this machine learning experiment, the log of resistivity was used instead of the original resistivity value.

Table 3.1. Summary of continuous variables

<table>
<thead>
<tr>
<th>Variables</th>
<th>Min.</th>
<th>1st Qu.</th>
<th>Median</th>
<th>Mean</th>
<th>3rd Qu.</th>
<th>Max.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Resistivity ($\Omega$ cm)</td>
<td>424</td>
<td>1500</td>
<td>2288</td>
<td>4439</td>
<td>5831</td>
<td>22600</td>
</tr>
<tr>
<td>pH value</td>
<td>7.70</td>
<td>8.38</td>
<td>8.70</td>
<td>8.68</td>
<td>9.00</td>
<td>9.70</td>
</tr>
<tr>
<td>Sulfide</td>
<td>0.00</td>
<td>0.45</td>
<td>0.92</td>
<td>3.77</td>
<td>1.43</td>
<td>167.98</td>
</tr>
<tr>
<td>Moisture</td>
<td>0.60</td>
<td>8.25</td>
<td>14.55</td>
<td>14.51</td>
<td>19.30</td>
<td>34.80</td>
</tr>
<tr>
<td>Thickness</td>
<td>9.50</td>
<td>10.80</td>
<td>11.85</td>
<td>12.30</td>
<td>13.47</td>
<td>17.90</td>
</tr>
</tbody>
</table>

Pipe Condition Rating and Prediction Target

To evaluate pipe condition, RWT was used as the predictive target in this experiment, which was expressed as given below:

\[
RWT = \frac{t - \max(d_1, d_2, ..., d_5)}{t}
\]

where \(t\) refers to the pipe original wall thickness and \(d\) represents the value of pit depth measured and recorded for a given pipe.

As given in Tab. 3.2, the RWT can be converted into different condition grades based on the condition assessment scale [1]. Furthermore, corresponding maintenance and replacement action decision can be supported to facilitate meeting the asset management needs of the municipalities.

Figure 3.6 depicts the pairwise relationships, with scatter plots and trend lines for the variables in consideration, whereby RWT is the response variable. The upper right of the matrix are the corresponding correlation coefficient between variables where the larger values were highlighted with larger text size and more stars.
Table 3.2. Pipe condition rating scale based on RWT (Adapted from [1])

<table>
<thead>
<tr>
<th>Numeric scale</th>
<th>Linguistic scale</th>
<th>Criteria</th>
<th>Action</th>
</tr>
</thead>
<tbody>
<tr>
<td>8-10</td>
<td>Excellent</td>
<td>No signs of corrosion</td>
<td>No action required</td>
</tr>
<tr>
<td>6-8</td>
<td>Good</td>
<td>RWT &gt;90%</td>
<td>Reassess in 10 years. Schedule for cathodic protection in 5-10 years</td>
</tr>
<tr>
<td>4-6</td>
<td>Moderate</td>
<td>RWT &gt;= 75%</td>
<td>Reassess in 3-5 years. Schedule for lining and rehabilitation in 5-10 years</td>
</tr>
<tr>
<td>3-4</td>
<td>Poor</td>
<td>RWT = 50-75%</td>
<td>Schedule for rehabilitation or replacement in 3-5 years</td>
</tr>
<tr>
<td>&lt;3</td>
<td>Critical</td>
<td>RWT &lt;50%</td>
<td>Immediate repair or replacement required</td>
</tr>
</tbody>
</table>

From Fig. 3.6 it can be seen that resistivity, moisture, and soil type are correlated. Since strong correlation is associated with values exceeding 0.8, none of the three variables are closely correlated. In addition, a negative correlation between soil resistivity and moisture is identified. Similarly, the linear correlation between the soil properties and RWT is weak, as indicated by the plot.

3.3.2 Individual Regression Models

Multiple Linear Regression

In prediction modeling, linear regression models are commonly used, as they are simple to apply. For instance, five multiple linear regression models were developed to predict annual break rates of five types of water mains [68]. Based on the best data subsets selected through best subsets regression, an optimal multiple regression model was built. Through subsequent sensitivity analysis, the model’s effectiveness in predicting the annual break rate of a given water mains system was demonstrated. However, the assessment of water mains condition was inadequate.

Thus, in the present study, based on the analysis of correlations between multiple independent variables and the dependent variable, MLR model was first built to obtain the best-fitting line for the relationship between variables.

Artificial Neural Network

Artificial neural networks have been widely employed in pipe performance prediction with the capability of simulating nonlinear and complex behavior of water networks. Recently, Zangenehmadar and Moselhi used ANN models to pre-
Figure 3.6. Chart of correlation matrix for soil property data and RWT. On the bottom of the diagonal: the bivariate scatter plots with a fitted line. On the top of the diagonal: the value of the correlation plus the significance level as stars.

dict the remaining life of pipelines [78], where $R^2$ value was over 98%. The authors used the remaining useful life as the target value, which was calculated as

$$\text{Target value} = 200 - \text{pipe age}.$$ 

This simplified approach is clearly inadequate to thoroughly describe the pipe condition. Since the linear correlation between the soil properties and RWT is weak, in the present study, ANN with one hidden layer and two neurons in the structure is also developed.
Random Forest

The use of RF has been reported in previous study. Considering five soil properties, namely resistivity, pH value, redox potential, sulfides and soil type, RF was applied to to explore the relationship between soil properties and deterioration of metallic pipes [11]. The authors reported $R^2 = 0.739$ and normalized mean square error of 0.23288. When using RF, as the number of predictors increases, so does the number of trees necessary for good performance. Considering the size of data set used in this study, 300 trees and $m_{try} = 9$ after tuning were used for the RF.

Support Vector Machine

The support vector machine algorithm derived from a nonlinear generalization of generalized portrait algorithm in 1993 and entered the standard methods toolbox of machine learning around 1998 [79]. SVM is scalable, as it can generalize well even when applied to relatively small training data sets [80]. The SVM is applied to regression problems by the introduction of an alternative loss function [81]. In the present study, SVM with a radial basis function kernel was used to predict the pipe performance. In the training process, $\lambda$ and $c$ are parameters of particular importance. While $\lambda$ parameter defines the influence weights of a single training examples, $c$ parameter weights the errors associated with the training samples against the simplicity of the decision surface. After parameter tuning, the final parameters used for the model were $\lambda = 0.02889218$ and $c = 1$.

3.3.3 Super Learning for Pipe Deterioration Prediction

Super learning, namely stacking ensemble, is a generalized loss-based learning framework. It was first introduced by Wolpert [82], and was subsequently formalized by Breiman [83], and theoretically validated by Van der Laan and colleagues [84]. Due to its superior performance in comparison with single algorithms, super learning has found many applications, including estimation of propensity scores and dose-response functions [59]. However, the report on using stacking ensemble for civil infrastructure asset management is not available. Hence, aiming to reduce drawbacks of single models and to support operational decisions pertaining to the maintenance, a stacking ensemble based method was proposed in this
study, focusing on cast iron water mains performance prediction.

**Base Learner Selection**

In stacking ensemble, a group of base learners needs to be specified first. Ideally, the base learners should be diverse and uncorrelated [59]. As described in subsection 3.3.2, therefore, a diverse set of individual learners (MLR, ANN, RF, SVM) were built and tested. Based on their prediction performance when applied to the test data, depicted in Fig. 3.7a and Fig. 3.7b, three best performance algorithms: MLR, RF, and SVM were selected as the base-level models.

**Meta-learning Algorithm**

In the stacking ensemble learning, meta-learning algorithms are implemented to minimize the cross-validation risk related to some loss function and to find the optimal combination of the predictions yielded by different models [85]. In the present study, the meta-learner was built using the outputs from the base learners, known as level-one data, which was generated using k-fold cross validation. The cross-validation was also applied in the meta-learning process, which is actually a machine learning process, where the meta-learner fits the level-one data.

The Generalized Linear Model (GLM) is a broad class of models including linear regression, analysis of variance, Poisson regression, log-linear models, etc. In this study, liner regression was employed for the GLM as the meta-learner to fit the level-one data. Thus, the weights were assigned to the individual models through the process of training the linear model using level-one data.

**Stacking**

Stacking regression is an ensemble technique for forming linear combinations of multiple models to improve prediction accuracy [83]. As described in Tab. 3.3, the first step in the development of a stacking ensemble comprises of training base learners, denoted as $h_1$, $h_2$, and $h_3$, by performing k-fold cross-validation. With cross-validated predicted values and the original outcome, level-one data, $D_1$, can be constructed for the meta-learning. In the next step, as the meta-level model, GLM was trained to fit the level-one data to obtain $\Phi$. The final ensemble fit
Figure 3.7. Individual model prediction performance comparison.
Table 3.3. Stacking ensemble models training steps (Adapted from [2])

<table>
<thead>
<tr>
<th>Algorithm: Stacking Ensemble Modeling</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Input</strong>: training data (D = {x_i, y_i}_{i=1}^n)</td>
</tr>
<tr>
<td><strong>Output</strong>: stacking ensemble models (S)</td>
</tr>
</tbody>
</table>

1. **Step 1**: develop base-level models \(H\) on \(D\)
   - \(h_1 = mlr()\)
   - \(h_2 = rf()\)
   - \(h_3 = svm()\)
2. Perform k-fold cross-validation on base-level models
3. \(D_1 = \{x'_i, y_i\}_{i=1}^n\), where \(x'_i = \{h_1(x_i), h_2(x_i), h_3(x_i)\}\)
4. **Step 2**: construct new data: level-one data \(D_1\)
5. **Step 3**: train meta-level model \(\Phi\) on \(D_1\)
6. Perform k-fold cross-validation on meta-level model
7. \(\Phi = glm()\)
8. Return \(S\) comprising of \(H\) and \(\Phi\) models

comprises the base learner fits to the full training set and the meta-learner fit to the level-one data.

### 3.4 Experimental Results and Discussion

In the experiment, R package called ”caretEnsemble” was used to build the stacking ensemble model and the data sourced from [73] were used to validate the proposed stacking ensemble based method.

To evaluate the performance of different modeling methods, a statistical measure called Root-Mean-Square Error (\(RMSE\)) and the Coefficient of Determination \(R^2\) were used, which are expressed as follows:

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

\[
R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}
\]  
(3)

where \(y_i\) and \(\hat{y}_i\) are the original dependent variables and predicted values, respectively. Moreover, \(R^2\) is a relative measure of fit and \(RMSE\) is an absolute measure.
The lower RMSE is, the better the model fits the input data. $R^2$ in the $[0 - 1]$ range, whereby a higher value (close to 1) indicates a better model.

To obtain an unbiased estimate of the accuracy of a learned model, in this experiment, an independent test set was used to test the hypothesis. In addition, to accurately estimate and compare the average accuracy of machine learning algorithms, 15-fold cross-validation was employed for algorithm selection, which makes efficient use of the data available for testing.

Table 3.4 provides the RMSE and $R^2$ values in the training and testing phases for all individual and ensemble models. The results clearly demonstrate that, based on the RMSE and $R^2$ values, the proposed stacking ensemble method achieved the best prediction in comparison with other models.

Table 3.4. Model assessment with $R^2$ and RMSE (Best results are indicated in bold)

<table>
<thead>
<tr>
<th>Model</th>
<th>Training</th>
<th></th>
<th>Testing</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$R^2$</td>
<td>RMSE</td>
<td>$R^2$</td>
<td>RMSE</td>
</tr>
<tr>
<td>MLR</td>
<td>0.6323</td>
<td>0.1507</td>
<td>0.5526</td>
<td>0.1470</td>
</tr>
<tr>
<td>ANN</td>
<td>0.6456</td>
<td>0.1395</td>
<td>0.5280</td>
<td>0.1507</td>
</tr>
<tr>
<td>RF</td>
<td>0.6206</td>
<td>0.1393</td>
<td>0.6398</td>
<td>0.1391</td>
</tr>
<tr>
<td>SVM</td>
<td>0.7161</td>
<td>0.1538</td>
<td>0.6522</td>
<td>0.1395</td>
</tr>
<tr>
<td>Stacking</td>
<td><strong>0.7514</strong></td>
<td><strong>0.1403</strong></td>
<td><strong>0.7163</strong></td>
<td><strong>0.1302</strong></td>
</tr>
</tbody>
</table>

Using the same dataset, Doyle [73] also performed an analysis with MLR considering the pH and the logarithm of soil resistivity as the variables. However, the analysis could only explain the relationship between two soil properties and the maximum external pitting rate of pipes ambiguously. For pipe performance assessment, according to the extensive literature review, ANN outperforms other models, such as MLR. Nevertheless, when it was applied to the data set in this study, as the data set is small, its performance was poor, with 0.1507 being the highest obtained RMSE value. Hence, its scalability needs further verification.

In contrast, even though the collected data is small, the proposed ensemble modeling method can still achieve a good prediction performance, with the highest $R^2$ value of 0.7163 and the lowest RMSE value of 0.1302 in the test results. Compared with RF and SVM model, the prediction accuracy also has a certain level of
Furthermore, if additional data are collected, the performance of ensemble modeling can be further improved and validated by adding more features. The automated updating of the ensemble model also remains a topic for our future study. Future work will include a more comprehensive algorithm comparison with nonparametric statistical tests as well [86].

3.5 Summary

In this chapter, stacking ensemble was proposed to overcome the disadvantages of the existing machine learning algorithms including MLR, ANN, RF, and SVM. With the objective of more accurate pipe condition forecasting to support water mains asset management program, variables affecting pipe performance were first identified and the impacts of different predictors were discussed in relation to pipe failure mechanisms. The comparison results of the experiments demonstrate the superiority of the proposed methods.
Chapter 4

Prediction of Pipe Deformation Using Super Learner and Recursive Feature Elimination Algorithm

This chapter presents a super learning based approach to characterize the soil-pipe interaction and predict pipe deformation. The methodology of the proposed super learning framework is presented in Section 4.1, where the details of data description and preprocessing along with the theoretical background of recursive feature elimination methods and super learning algorithm are provided. The implementation of super learning based methods are described in Section 4.2. The performance evaluation metrics and the experimental results are presented and discussed in the Section 4.3.

4.1 Data-Driven Approaches for Pipe Deformation Prediction

Figure 5.1 describes the procedure of the pipe structural behavior prediction modeling using super learning techniques. Firstly, preprocessing of sensor data, which was collected from a monitoring system, was performed to obtain clean and in-
formative indicators. Integrating with climate data, the fused data was filtered and analyzed using Pearson correlation. The output of the filter method was then applied to the RF-RFE algorithm. Not only did this module make the model simple and easily interpreted, it helped find the best variable subset to achieve the most precise prediction modeling, as well. Using the selected variables, both base learners and super learning were constructed to predict the pipe structural behavior. For the super learning modeling, a group of machine learning algorithms were trained and combined optimally to generate a super learner. The presented framework can be used to estimate the pipe performance based on soil properties and weather conditions.

4.1.1 Sensor Data Preprocessing and Summarization

In the City of Regina, the volume change of the native clay deposit is always a problem for the underground water distribution pipes. Hence, a field instrumentation program was conducted to monitor the performance of a polyvinyl chloride pipe with 150mm diameter and its surrounding soil condition in south Regina, Saskatchewan, where a high pipe breakage rate was reported. The monitoring system consists of a group of high quality sensors, measuring pipe wall strain, in situ soil temperature, soil pressure, and soil water contents. The data used in the exploratory research was collected from this system in the period between May
5, 2005 to April 21, 2008 [87]. This data set contains 973 records with soil data measured by 23 different sensors as well as pipe deformation information obtained from 8 strain gauges. Table 4.1 provide a brief description of the instruments with its measured attributes.

Table 4.1. Description of the sensor data

<table>
<thead>
<tr>
<th>Sensor type</th>
<th>No.</th>
<th>Measured property</th>
</tr>
</thead>
<tbody>
<tr>
<td>Thermocouple (TC)</td>
<td>13</td>
<td>Soil temperature</td>
</tr>
<tr>
<td>Water Content Reflectometer</td>
<td>6</td>
<td>Soil water content</td>
</tr>
<tr>
<td>Pressure Cell (PC)</td>
<td>4</td>
<td>Soil pressure</td>
</tr>
<tr>
<td>Strain Gauge (SG)</td>
<td>8</td>
<td>Pipe deformation</td>
</tr>
</tbody>
</table>

To observe the changes in soil properties vertically and horizontally, multiple sensors were installed to measure the soil temperature and water contents. This resulted in a large number of features, and more importantly, a correlation between these attributes collected by the same type of sensors. Therefore, prior to super learning, preprocessing data exported from sensor nodes is necessary to filter the redundant variables and to extract features. In this subsection, the description and preprocessing for the indicators of pipe deformations and soil properties collected from sensor network are provided.

**Pipe Deformation**

1) **Circumferential Strain (CS):** In the sensor monitoring network, there were four sensors installed to measure the pipe CS in the top, bottom, and two spring-lines respectively. In this study, the maximum value of the four obtained CS was identified as the indicator of the pipe hoop deformation, shown as Eq. (4.1). Generally, the strains are mainly a consequence of the seasonal pipe temperature variations [88], which is demonstrated in Fig. 4.2, where a seasonal variation can be observed in the measured CS plots by the daily time.

\[
CS = \text{Max}(cs_1, cs_2, cs_3, cs_4) \quad (4.1)
\]
where $cs_1, cs_2, cs_3, cs_4$ indicates pipe CS measured in the top, bottom, and two springlines.

2) **Average Longitudinal Strain (ALS):** In the same spot of the pipe, there were another four strain gauges installed to measure the pipe longitudinal strain. As given in Eq. (4.2), ALS is the mean of the longitudinal strain values measured at four points (top, bottom, and two springlines) of the pipe, which represents the tensile deformation in the longitudinal direction [87]. In Fig. 4.2, the time variation of the ALS is described, where a seasonal pattern can be found as well. However, different from CS, the variation magnitude of ALS decreases and the measurements tend to increase over time.

$$ALS = \text{Avg}(ls_1, ls_2, ls_3, ls_4)$$  \hspace{1cm} (4.2)

where $ls_1, ls_2, ls_3, ls_4$ is the pipe longitudinal strain measured in the top, bottom, and two springlines.

3) **Differential Longitudinal Strain (DLS):** As the indicator of pipe bending deformation, the DLS is also derived from four measured pipe longitudinal strains, which calculation was demonstrated in the Eqs. (4.3) and (4.4). Based on the observation of Fig. 4.2, the DLS curve tends to be smooth after April 2006, during which the value is relatively constant.

$$dls_i = ls_i - ALS (i = 1, 2, 3, 4)$$  \hspace{1cm} (4.3)

$$\text{DLS} = \text{Max}(dls_1, dls_2, dls_3, dls_4)$$  \hspace{1cm} (4.4)

where $ls_i, (i = 1, 2, 3, 4)$ is the pipe longitudinal strain measured in the top, bottom, and two springlines.

**Soil Properties**

1) **Soil Temperature:** To measure the temperature variance in different depths of the soil, there are 13 thermocouples (TCs) installed in this program. Figure 4.3a presents the measured temperature varying vertically, where a periodic variation was observed. It also indicates that the closer to the surface, the larger fluctuation the temperature has. Since Fig. 4.3a demonstrates a strong relationship between
### Figure 4.2
Daily time series plot of strains and climate data (The blue smooth lines describe the curve trend).

the temperature variables, which can cause the multicollinearity problem. Hence, Pearson correlation coefficient, as given in Eq. (4.5), was applied to identify highly correlated soil temperature data in the preprocessing module, which measures a linear dependence between two variables. When two variables have a high correlation, the attribute with larger mean absolute correlation with all other attributes is removed. Setting cutoff value as 0.75, TC2 (represented the temperature near the pipe) and TC10 (represented the temperature near the ground) were remained.

\[
\rho_{X,Y} = \frac{\text{cov}(X,Y)}{\sigma_X \sigma_Y} \tag{4.5}
\]

where \(\text{cov}\) is the covariance; \(\sigma_X\) is the standard deviation of \(X\); \(\sigma_Y\) is the standard...
deviation of $Y$. The value of $\rho_{X,Y}$ is always between $+1$ and 1. The closer its absolute value is to 1, the stronger the relationship between the two variables and a correlation of 0 indicates an absence of the linear relationship.

2) Soil Water Content: Water content is a critical property for the expansive soil, as the high variation in water content will cause soil swell-shrink, which could result in soil deformation. In this field instrumentation program, there are 4 water content reflectometers (WCRs) used to collect volumetric water content data in different depths in the trench backfill, shown as Fig. 4.3b and another 2 water content values measured in the surrounding clays, shown as Fig. 4.3c. As depicted in Fig. 4.3b with an increase in depth, the magnitudes of the variations generally decreased. Furthermore, variation patterns in volumetric water content at 1.07m depth and 0.45m depth with time were similar while those at 1.96m depth and 1.07m depth were different. The same preprocessing method as TC sensor nodes was employed in WCRs as well and water content variables whose Pearson correlation coefficient is greater than 0.75 were removed, i.e., WCR7.4.

3) Soil Pressure: Soil pressure depends mainly on the nature of the soil, its natural saturation degree and the variation in soil moisture [19]. Figure 4.3d shows the variation of earth pressures on four pressure cells (PCs), where two distinct peaks can be observed annually in the vertical soil pressure, while the horizontal soil pressure only has one peak in the summer. Since the earth pressures measured by different sensors have similar variations with little difference in the magnitude, the average value of soil pressure was used to replace two highly related explanatory variables, shown as Eqs. (4.6) and (4.7).

$$PC.H = \text{Avg}(PC4.2, PC4.3)$$  \hspace{1cm} (4.6)

$$PC.V = \text{Avg}(PC4.1, PC4.4)$$  \hspace{1cm} (4.7)

where $PC.H$ refers to the average horizontal soil pressure and $PC.V$ is the notation of the average vertical soil pressure. PC4.2 and PC4.3 are two near horizontal soil pressure measured in the same level. Similarly, PC4.1 and PC4.4 are the measurement of the vertical soil pressure.
(a) Soil temperature

(b) Water content in the backfill.

(c) Water content in the clays.

(d) Vertical and horizontal soil pressure.

Figure 4.3. Daily time series plot of soil properties.
4.1.2 Data Integration with Climate Factors and Data Analytics

Climate conditions including rain deficit, temperature, and freezing index etc. were identified as one of the primary factors affecting the variation of soil moisture-suction [18] and pipe failure [29, 31, 33]. Additionally, climate data are always well-recorded and available for analytics. Therefore, climate data in City of Regina was also obtained from an environment Canada meteorological station in this study and integrated with sensor data based on timestamps index to complement the feature information for pipe deformation prediction. The detailed description about the climate variables was provided below along with their statistical summery, given in Tab. 4.2.

<table>
<thead>
<tr>
<th>Table 4.2. Summary of the climate data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Variable</td>
</tr>
<tr>
<td>-------------</td>
</tr>
<tr>
<td>T (°C)</td>
</tr>
<tr>
<td>PPT (mm)</td>
</tr>
<tr>
<td>RD (mm)</td>
</tr>
<tr>
<td>FD (m)</td>
</tr>
</tbody>
</table>

1) Temperature (T): There are many studies having reported the influence of temperature on pipe break rate [89], which also indicated that the cold ground temperature could result in an increase of pipe circular breaks. Based on the observation of Fig. 4.2, it was found that the temperature and CS has similar periodic variance while temperature arises its peak earlier. Figure 4.3a presents the measured temperature varying vertically along with air temperature, which demonstrates that there is a similar variation between the temperature variables with different peak values.

2) Precipitation (PPT): The PPT describes the daily precipitation amount including rainfall and snowfall. Liu et al. reported that soil moisture changes due to the PPT, evaporation, and transpiration [90]. Recently, it is indicated that the effect of precipitation received by the earth surface on pipe breaks can be quantified in the form of a rain deficit [33]. As depicted in Fig. 4.2 the peak of PPT appears in both summer and winter due to the sufficient rainfall and the heavy snow.
3) **Rainfall Deficit (RD):** The RD is the difference between the received precipitation and the water lost due to evapotranspiration, which is closely related to soil moisture content level [33]. It was reported that break rate increases in dry summer due to the shear stress exerted on the water mains [26]. As depicted in Fig. 4.2, a simultaneous variation between RD and CS can be observed.

4) **Frost Depth (FD):** The FD describes the frost penetration depth that highly correlated with the soil freezing index. Cohen and Fielding underlined the relationship between frost depth and pipeline performance [91] while Kleiner and Rajani reported that high breakage rates primarily results from frost loading during the winter time[26]. Hu and Vu found that the deeper depth of the frost penetration, the higher the vertical soil pressure is [87]. Accordingly, FD was also considered for pipe deformation prediction. As presented in Fig. 4.2, a seasonal variation can be observed, which changes in the opposite direction with other variables.

According to the observation from Fig. 4.4 for the soil properties, it can be seen that the temperature near the surface (TC10) is closely related to T and the measured temperature around pipe (TC2) has a strong correlation with RD and PC.V, as a strong correlation is associated with values exceeding 0.8. And water contents near the pipe (WCR3.6 and WCR5.2) are weakly related to any other variables while water contents measured near the ground WCR9.6, WCR7.3, WCR9.4 are positively related with T and negatively with FD. Figure 4.4 also indicates that vertical soil pressure is linearly correlated with the temperatures and RD. Since there are strong linear correlation between the soil features and climate variables, Pearson correlation coefficient filter was applied to remove highly correlated features to avoid biases in the feature selection procedure, in which TC2 and TC10 were removed.
**Figure 4.4.** Chart of correlation matrix for soil properties and climate factors. On the bottom of the diagonal: the value of the correlation coefficient. On the top of the diagonal: correlation visualization using circles of different size and colors. Color bar on the right: correlation value scale and corresponding color.

### 4.1.3 Feature Selection Using Recursive Feature Elimination

Feature selection, generally, can be classified into three groups: filters, wrappers and embedded, where filter approach is faster than the wrapper and embedded methods as it is independent of the induction algorithm [64]. Widely used as a pre-processing step, in this study, filter methods were implemented based on Pearson correlation coefficient to remove highly correlated feature features in sensor datasets.

After removing the redundant features, an automatic feature selection method, RF-RFE, was used to find the variable subset with the best performance. As an wrapper feature selection algorithm, RFE has been applied in sensor data analy-
Table 4.3. Feature importance ranking results (Variables in bold are features selected by RF-RFE)

<table>
<thead>
<tr>
<th>Outcome</th>
<th>Covariates ordered by feature ranking results</th>
</tr>
</thead>
<tbody>
<tr>
<td>CS</td>
<td>RD, WCR3.6, PC.H, PC.V, WCR7.3, Month, WCR5.2, T, WCR9.4, TDR9.6, FD, PPT</td>
</tr>
<tr>
<td>ALS</td>
<td>WCR3.6, WCR7.3, Month, PC.V, WCR5.2, RD, PC.H, WCR9.4, WCR9.6, T, FD, PPT</td>
</tr>
<tr>
<td>DLS</td>
<td>WCR3.6, Month, WCR7.3, WCR5.2, PC.H, RD, WCR9.4, WCR9.6, PC.V, T, FD, PPT</td>
</tr>
</tbody>
</table>

sis [92], structural damage detection [93], and spectral analysis [66]. The algorithm details have been given in Section 2.2.

4.1.4 Pipe Deformation Prediction Using Super Learning

Considering scalability, stability, and prediction accuracy, an ensemble based method, super learning, hence, was combined with RF-RFE algorithm to predict pipe deformations. Super learning is a technique with ability of finding the optimum way to combine multiple, typically diverse, base learning algorithms to generate a powerful prediction function [94]. Considering a group of prediction models, super learning has capability of taking advantage of the individual strengths of the best performing models [95]. The algorithm has been described in detail in Section 2.3 in Chapter[2].

4.2 Implementation of Pipe Deformation Prediction

4.2.1 Feature Selection with Random Forest-Recursive Feature Elimination

Although a number of redundant features have been removed during data analysis, the importance of each variables was unknown. Therefore, RF-RFE was used to further identify confounding variables within the selected sensor data and, hence, to determine the optimum subset of variables. Constructing three datasets (CS, ALS, and DLS), along with soil property dataset, the key factors contributing to different type of deformations were ranked respectively by RF-RFE.

In this study, the caret package (version 6.0.77) in R (version 3.4.2) were used to implement RF-RFE with the full study data.
Figure 4.5. RF-RFE results for feature selection (Blue solid points indicate the optimum number of variables used for the prediction).
4.2.2 Super Learning Framework Implementation

Candidate Learner Selection

After variable selection, highly correlated soil attributes were removed and the optimum number of variables for modeling was determined by automatic feature selection methods, which also identified the most important factors contributing different type of pipe deformation. Utilizing the filtered data, multiple diverse regression algorithms available in Caret packaged (version 6.0.77) were investigated to determine the candidate learners for super learning.

It is indicated that the prediction algorithm with minimum modeling error using selected variables is always not the one used to select features, as a different modeling strategy probably ignores data particularities found in the feature selection process, leading to a better generalization [66]. In addition, LeDell reported that ideally the base learners should be diverse and uncorrelated [94]. Hence, for base learner selection, 6 machine learning algorithms including MLR, KNN, RT, CIT, MLP, and SVM, were selected as the base learners considering diversity and prediction accuracy.

Super Learner Implementation Details

To implement pipe behavior prediction modeling using super learning, caretEnsemble package (version 2.0.0) was applied in this study. Each dataset was divided into two parts, where 75% of the data was used for model training and the rest for test purpose. 10-fold cross-validation was applied to evaluate each candidate learner in the training process, and then the weights of these fitted model were determined by their prediction performance across the validation sets.

To examine the prediction ability of super learning and the advantages of RF-RFE, 6 super learning models were built to predict three different type of pipe deformations (CS, ALS, DLS) considering different feature subsets in this study. Additionally, base learners were also trained with the best variable subsets, which were used to validated the super learner ability to combine a variety of individual algorithms optimally.
4.3 Experimental Results

4.3.1 Performance Evaluation Metrics

To evaluate the super learning predication performance, three criteria were used in this experiment: RMSE, $R^2$, and Mean Absolute Error (MAE). In statistics, RMSE and MAE are two metrics frequently used to evaluate the prediction modeling accuracy. RMSE is used to measure the differences between predicted values and the observed values while MAE measures the average magnitude of the errors without considering their direction. Since RMSE gives a relatively high weight to large errors, it is more useful when large errors are particularly undesirable. The equations of RMSE and MAE are given in Eqs. (4.8) and (4.10), respectively. Equation (4.9) is the formula of $R^2$, which describes the fitting degree of models in a $[0, 1]$ range, where the larger the $R^2$ value is, the better the model fits the data.

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

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

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

where $n$ is the number of observations in the data set, and $y_i$ and $\hat{y}_i$ represent the observed outcomes and predicted target values, respectively.

4.3.2 Results and Discussion

Figures 4.5a, 4.5b, and 4.5c present the RF-RFE feature selection results with the optimal number of variables for prediction modeling. As shown in Figs. 4.5a, 4.5b, and 4.5c, the ideal feature subsets size for CS, ALS, and DLS prediction is 9, which indicates that the prediction performance can be improved by finding the best feature subset. Table 4.3 provides the feature importance ranking details by RF-RFE based on three difference targets. It can be observed that ALS and DLS shared the same optimal feature subset where WCR3.6 is the most critical factors.
to predict ALS and DLS while RD contributes most to CS prediction. Additionally, among climate factors, PPT, FD, and T are relatively less important than RD and excluded when predicting pipe ALS and DLS. For soil properties, both water contents and soil pressure play an important role on pipe structural behavior evaluation. Ultimately, based on the feature selection results using both filter method and embedded method, the sensor network for soil property measurement can be simplified from 23 sensor nodes to 9.

Table 4.4 reports the RMSE, MAE, and R² values for both training sets and test sets for all super learners trained with difference predictor subsets. According to the observation of Tab. 4.4, the advantages of RF-RFE are demonstrated, which helps super learning achieve higher accuracy with fewer features.

**Table 4.4.** Super learning performance comparison with different feature subsets using RMSE, MAE, and R² (Results highlighted in boldface are from super learning combined with RF-RFE)

| Target | # Features | Training | | | | | Testing | | | |
|--------|------------|----------|----------|----------|----------|----------|----------|----------|----------|----------|----------|
|        |            | RMSE     | MAE      | R²       | RMSE     | MAE      | R²       | RMSE     | MAE      | R²       | RMSE     | MAE      | R²       |
| CS     | 12         | 0.00249  | 0.00169  | 0.99312  | 0.00233  | 0.00151  | 0.99416  |          |          |          |          |          |          |
|        | 9          | 0.00245  | 0.00165  | 0.99330  | 0.00233  | 0.00154  | 0.99420  |          |          |          |          |          |          |
| ALS    | 12         | 0.00090  | 0.00056  | 0.98818  | 0.00093  | 0.00056  | 0.98765  |          |          |          |          |          |          |
|        | 9          | 0.00066  | 0.00045  | 0.99380  | 0.00058  | 0.00042  | 0.99515  |          |          |          |          |          |          |
| DLS    | 12         | 0.00031  | 0.00020  | 0.96583  | 0.00027  | 0.00018  | 0.97248  |          |          |          |          |          |          |
|        | 9          | 0.00026  | 0.00017  | 0.97603  | 0.00021  | 0.00013  | 0.98350  |          |          |          |          |          |          |

In Tab. 4.5, the evaluation results of all base learners and super learners in the training and testing phases were given. Compared with the base learners prediction performance, super learning achieved the best prediction based on the RMSE, MAE, and R² values. From the experimental results given in the Tab. 4.5, the following conclusions were drawn: super learning algorithm is powerful and adaptive to different data sets, which has the capability of predicting the CS, ALS, and DLS accurately; benefiting from RF-RFE algorithm, the super learner prediction performance was enhanced by only using 9 variables; finally, the relationship between the soil properties and pipe deformation was successfully modeled using the proposed super learning based approaches. Figures 4.6a, 4.6b, and 4.6c show the
Table 4.5. Prediction performance comparison of base learners and super learners with best performance feature subset using RMSE, MAE, and $R^2$ (Significant results are highlighted in boldface)

<table>
<thead>
<tr>
<th>Target</th>
<th>Model</th>
<th>Training</th>
<th></th>
<th>Testing</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>RMSE</td>
<td>MAE</td>
<td>$R^2$</td>
<td>RMSE</td>
</tr>
<tr>
<td>CS</td>
<td>MLR</td>
<td>0.00829</td>
<td>0.00651</td>
<td>0.92526</td>
<td>0.00816</td>
</tr>
<tr>
<td></td>
<td>RT</td>
<td>0.01147</td>
<td>0.00905</td>
<td>0.85081</td>
<td>0.01218</td>
</tr>
<tr>
<td></td>
<td>CIT</td>
<td>0.00387</td>
<td>0.00236</td>
<td>0.98323</td>
<td>0.00398</td>
</tr>
<tr>
<td></td>
<td>MLP</td>
<td>0.00900</td>
<td>0.00727</td>
<td>0.92101</td>
<td>0.00946</td>
</tr>
<tr>
<td></td>
<td>SVM</td>
<td>0.00435</td>
<td>0.00278</td>
<td>0.98047</td>
<td>0.00495</td>
</tr>
<tr>
<td></td>
<td>KNN</td>
<td>0.00261</td>
<td>0.00171</td>
<td>0.99241</td>
<td>0.00249</td>
</tr>
<tr>
<td></td>
<td>SL</td>
<td><strong>0.00245</strong></td>
<td><strong>0.00165</strong></td>
<td><strong>0.99330</strong></td>
<td><strong>0.00233</strong></td>
</tr>
<tr>
<td>ALS</td>
<td>MLR</td>
<td>0.00474</td>
<td>0.00369</td>
<td>0.68601</td>
<td>0.00509</td>
</tr>
<tr>
<td></td>
<td>RT</td>
<td>0.00594</td>
<td>0.00449</td>
<td>0.50025</td>
<td>0.00649</td>
</tr>
<tr>
<td></td>
<td>CIT</td>
<td>0.00170</td>
<td>0.00092</td>
<td>0.98710</td>
<td>0.00159</td>
</tr>
<tr>
<td></td>
<td>MLP</td>
<td>0.00608</td>
<td>0.00491</td>
<td>0.54961</td>
<td>0.00672</td>
</tr>
<tr>
<td></td>
<td>SVM</td>
<td>0.00098</td>
<td>0.00067</td>
<td>0.98047</td>
<td>0.00077</td>
</tr>
<tr>
<td></td>
<td>KNN</td>
<td>0.00087</td>
<td>0.00052</td>
<td>0.98739</td>
<td>0.00067</td>
</tr>
<tr>
<td></td>
<td>SL</td>
<td><strong>0.00066</strong></td>
<td><strong>0.00045</strong></td>
<td><strong>0.99380</strong></td>
<td><strong>0.00058</strong></td>
</tr>
<tr>
<td>DLS</td>
<td>MLR</td>
<td>0.00117</td>
<td>0.00092</td>
<td>0.53319</td>
<td>0.00113</td>
</tr>
<tr>
<td></td>
<td>RT</td>
<td>0.00109</td>
<td>0.00082</td>
<td>0.59173</td>
<td>0.00117</td>
</tr>
<tr>
<td></td>
<td>CIT</td>
<td>0.00048</td>
<td>0.00030</td>
<td>0.92080</td>
<td>0.00054</td>
</tr>
<tr>
<td></td>
<td>MLP</td>
<td>0.00179</td>
<td>0.00144</td>
<td>0.02262</td>
<td>0.00167</td>
</tr>
<tr>
<td></td>
<td>SVM</td>
<td>0.00032</td>
<td>0.00018</td>
<td>0.96291</td>
<td><strong>0.00019</strong></td>
</tr>
<tr>
<td></td>
<td>KNN</td>
<td>0.00027</td>
<td>0.00017</td>
<td>0.97161</td>
<td>0.00026</td>
</tr>
<tr>
<td></td>
<td>SL</td>
<td><strong>0.00026</strong></td>
<td><strong>0.00017</strong></td>
<td><strong>0.97603</strong></td>
<td><strong>0.00021</strong></td>
</tr>
</tbody>
</table>

prediction results obtained by best performance super learning models.

4.4 Summary

This chapter presents a data-driven approach for pipe deformations prediction utilizing machine learning techniques. Data used in this study, which includes pipe
Figure 4.6. Prediction results of six different super learning models.
wall strains, in situ soil water content, soil pressure, and temperature, were collected from a monitor system. In this chapter, a library of parametric and non-parametric machine learning algorithms were considered, and more importantly, a novel ensemble learning framework, i.e. super learning, was applied to predict the water mains structural behavior in different soil environments. To investigate the adaptability of super learning to different predictive models, this research employed super learning based methods to three different datasets. The predictive performance was evaluated by R-squared, root-mean-square error and mean absolute error. Based on the prediction performance evaluation, the superiority of super learning was validated and demonstrated by predicting three types of pipe deformations accurately. Additionally, a comprehensive understand of the water mains working environments becomes possible.
Chapter 5

Risk Assessment Model for Pipe Asset Management

This chapter presents multi-layer of decision-making tools for resource-constrained communities to manage their pipe system efficiently. In the present work, a novel risk-based approach was proposed by integrating condition rating algorithms and failure consequence models. Section 5.1 provides the overall methodology for the whole risk analysis study, where condition evaluation was studied and the failure consequence modeling method was described. In Section 5.2, the implementation details of the proposed risk analysis methods were provided. Finally, the experimental results including each pipe condition grading, failure consequence score, and risk level of each pipe were given and discussed in Section 5.3.

5.1 Risk-based Approaches for Pipe Asset Management

As shown in Fig. 5.1, the developed risk assessment approach involves estimating the current status of pipelines as well as determining the Consequences of Failure (CoF). The data utilized for the condition grading and the CoF assessment was obtained from a small-sized community, the Village of Lumby, located in North Okanagan, BC, Canada. The steps of the adopted risk-based methodology are described in detail below:

- Literature review: The review of pertinent works was comprehensive and
Figure 5.1. Flowchart for pipe risk analysis modeling.
was guided by the challenge that was going to be addressed. Thus, the focus was on the current risk-based approaches aimed at effectively managing underground pipeline assets. Methods to be utilized when determining the two key components of the risk analysis, i.e., pipe condition and CoF, were extensively researched and fully mastered. More importantly, factors related to pipe performance were identified and were aligned with those employed in pertinent studies.

- Data collection and preparation: The pipe information was collected in different formats, including asset management database, GIS, CCTV reports, etc. Data sourced from the database and CCTV reports is essential for accurate pipe condition prediction, as asset database contains asset identification and its physical factors, while CCTV provides pipe condition information. For CoF calculation, GIS data is also used when the failure impacts are closely related to physical pipe locations.

- Pipe condition prediction: Multiple regression technique was applied to develop pipe structural and operational condition assessment models. In the data preparation stage, data visualization was used to extract potentially useful information from the collected data, and the data was pre-processed through normalization prior to regression modeling.

- Consequence of failure: GIS technique was adopted when developing a failure consequence model based on a weighted sum multi-criteria matrix [4]. Pipe failure impacts on the economy, water distribution system, society, and the environment were visualized and analyzed.

- Risk identification: Based on the pipe condition and the CoF ratings, a risk matrix was generated, as combining the results assisted in identifying different risk level areas in the system.

5.1.1 Pipe Condition Modeling

Condition assessment is one of the core components of an asset management program. It provides critical information about the current state of asset [44]. According to US Environmental Protection Agency, condition assessment commences
Table 5.1. Major factors and sub-factors influencing pipe structural condition [3] (Factors in bold were used in the regression model)

<table>
<thead>
<tr>
<th>Major factor</th>
<th>Sub-factor</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Physical factors</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age (years)</td>
<td>Installation year</td>
<td></td>
</tr>
<tr>
<td>Pipe depth (m)</td>
<td>Average depth</td>
<td></td>
</tr>
<tr>
<td>Pipe material</td>
<td>Asbestos cement (AC), polyvinyl chloride (PVC), concrete (CONC) etc.</td>
<td></td>
</tr>
<tr>
<td>Pipe diameter (mm)</td>
<td>Nominal diameter of pipes</td>
<td></td>
</tr>
<tr>
<td>Pipe length (m)</td>
<td>Length of pipe between two manholes</td>
<td></td>
</tr>
<tr>
<td>Pipe gradient</td>
<td>Slope of pipe segments between two manholes</td>
<td></td>
</tr>
<tr>
<td><strong>Environmental factors</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Soil type</td>
<td>Aggressive, moderate</td>
<td></td>
</tr>
<tr>
<td>Ground water level</td>
<td>High, moderate, low</td>
<td></td>
</tr>
<tr>
<td>Service type</td>
<td>Industrial, commercial, residential</td>
<td></td>
</tr>
<tr>
<td>Road type</td>
<td>Local, primary, secondary</td>
<td></td>
</tr>
<tr>
<td><strong>Operational factors</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rehabilitation method</td>
<td>CIPP, slip lining, spot repairs</td>
<td></td>
</tr>
<tr>
<td>Rehabilitation age</td>
<td>Date of pipe rehabilitation (years)</td>
<td></td>
</tr>
<tr>
<td>Break history</td>
<td>Break rate</td>
<td></td>
</tr>
</tbody>
</table>

with the collection of data and information through the direct inspection, observation, and investigation, as well as in-direct monitoring and reporting. This is followed by a detailed analysis of the gathered data and information to make determination of the structural and operational performance status of capital infrastructure assets [96]. Unfortunately, for a small-sized community, the actual condition data collected from CCTV inspections is rarely sufficient. Consequently, in line with extant studies conducted in this field, in this work, multiple regression modeling was utilized to perform pipe performance prediction considering historical data. The condition index ranged from 1 (good condition) to 5 (poor condition).

**Factors Related to Pipe Structural Condition Modeling**

The factors that could potentially affect the pipe structural condition are presented in Tab. 5.1 which are commonly classified into physical, environmental, and operational categories.

1) **Physical factors:** These include pipe materials, diameter, length, depth, and gradient. According to the findings yielded by extant studies on the influence of sewer size on structural stability, pipes with larger diameters (> 300 mm) are at
a greater risk of structural instability [97]. Empirical evidence also indicates that pipes with high length to diameter ratios are likely to suffer from extensive bending stresses. Although these claims are not supported by sufficient data, frequency of defects is typically deemed to decrease as the depth increases (> 2 m) [97].

2) Environmental factors: This category comprises of soil, ground water, service area, road (using a geographic information system), and traffic type. The type of soil surrounding the pipes, depending on its physical and chemical properties, may provide different level of side support. Similarly, as groundwater may wash soil away, when present in large quantities, it will reduce soil support and can potentially result in infiltration. Different service area types (e.g., industrial or residential) would also determine the characteristics of wastewater collected and passed through the pipe system. Wastewater composition could have different interactions with the internal pipe walls and could potentially cause erosion. Road type will also influence traffic volume, whereby a heavier traffic flow on a main street will increase bending stress in the pipes buried underneath. However, available evidence indicates that truck roads tend to have low defect rate because of stronger road pavements on these locations.

3) Operational factors: Break history, pipe replacement, rehabilitation age, and rehabilitation methods are in this category, as operational aspects play an important role in the analysis of pipe infrastructure failure rates. The break rate can be applied to predict pipe failure probability, whereas maintenance records are critical to the understanding of the current pipe status. However, as most small municipalities do not keep detailed breakage records, analysis of pipe failure rate history is rarely possible.

Factors Related to Pipe Operational Condition Modeling

The factors affecting pipe operation are classified as non-hydraulic and hydraulic, as illustrated in Fig. 5.2. Non-hydraulic problems are generally defined as those deficiencies in pipe performance that are not due to lack of flow capacity, while hydraulic problems occur when a pipe system does not have an adequate size to sustain high flow volume [5]. Since non-hydraulic problems are directly caused by structural condition, in the operational condition modeling, this study focused on
hydraulic problems analysis, considering pipe age, diameter, length, and slope.

**Regression Models**

After identifying the important factors affecting pipe condition and analyzing the statistical features of independent variables, the pipeline condition formula proposed in extant literature [3, 5] was modified in order to obtain equations that can be employed in the assessment of the current state of pipes. Equation 5.1 presents the pipe structural condition regression model, while Eq. 5.2 shows the operational condition regression formulation.

\[
SC = 0.285 - 0.00071 \times Diam + 0.663 \times Depth + 0.03 \times \frac{Age}{5} - 0.384 \times ST + 1.025RT
\]

(5.1)

where \(SC\) refers to the structural condition of a pipe, \(Diam\) represents pipe diameter, \(ST\) indicates service type, and \(RT\) refers to road type.
\[
OC = \left( 0.308 + 0.567 \times \left( \frac{\text{Age}}{\text{Diam}} \right)^{\text{Gradient}} \right)^{\frac{1}{n}}
\]

where \( OC \) refers to operational condition and \( n = 0.11 \).

### 5.1.2 Pipe Failure Consequences

Failure consequence analysis is a more subjective modeling procedure because the determination of factor weights is empiric-based. However, it is quite straightforward and beneficial for asset management by calculating pipe failure consequence index and providing a map with highlighted important area. To consider CoF systematically and more comprehensively, the failure consequences for the Village of Lumby sewage system are classified into economic, operational, environmental, and social groups [48].

**Factors Related to Pipe Failure Consequences**

As indicated by data analysis, the economic impact is primarily associated with the pipe material, diameter, and physical features, as well as buried depth. Pipe material determines the price of replacing or maintaining pipes, while also playing a role in the selection of methods used for underground pipeline inspection. Similarly, previous studies show that the aforementioned costs also increase when pipes of larger diameter are used in the system [48]. Buried depth is important to consider, as it will increase the operational difficulty of rehabilitation and repair, and will thus make the cost of maintenance and/or replacement more prohibitive.

Regarding the operational impact, pipe segmentation and connection type were considered, as if failure in one segment affects the operation of those connected to it, the risk of system failure will increase substantially [98]. Available evidence also suggests that having a greater number of pipe segments connected in a series will exacerbate system failure.

When analyzing the social and environment factors, the GIS technique was employed to locate major institutions and structures, such as schools, hospitals, commercial areas, and rivers. Based on the distance between the pipes and these entities, the performance value of each sub-factor in social and environmental cat-
egories was determined. Additionally, traffic influence was included in the social impact measure because a pipe failure in high traffic flow areas will have a much greater impact than in areas with limited traffic. Finally, since any leakage of industrial wastewater will result in environmental hazard, service area type was also considered as an environmental factor.

**Multi-Criteria Matrix**

Based on the analysis of factors related to pipe failure consequences, more detailed criteria to determine the performance value were established, as reported in Tab. 5.2. Using the multi-criteria matrix and GIS technology, the related data could be collected, visualized, and calculated based on the major factor and sub-factor weights by applying Eq. 5.3.

\[
\text{CoF} = \sum_{i=1}^{4} W_i \left( \sum_{j=1}^{n} W_{ji} PV_{ji} \right)
\]  

(5.3)

The CoF scores were calculated, starting with the scores for the respective entries, whereby the weights were applied to the relevant scores, progressing from the bottom to the top of each table, until all index weights are applied. After CoF score calculation, natural breaks classification was adopted for the pipe failure impact factor ratings, provided by Jenks [99], as shown in Tab. 5.3.

**5.1.3 Risk Matrix**

After the determination of pipe condition and its failure consequence, the risk level of individual pipe could be obtained using a risk matrix. Based on the expression given in Eq. 5.4, the risk index of each pipeline is calculated by combining its condition grade and CoF score. Table 5.4 presents the details of the risk matrix system.

\[
\text{Risk Index} = \text{Condition} \times \text{CoF}
\]  

(5.4)
Table 5.2. Performance values and predetermined weights for the major CoF factors and sub-factors

<table>
<thead>
<tr>
<th>Major factor ((W_i))</th>
<th>Sub-factor</th>
<th>Sub-criteria</th>
<th>Performance value ((PV_{ji}))</th>
<th>Weight (W_{ji})</th>
</tr>
</thead>
<tbody>
<tr>
<td>Economic factors 0.15</td>
<td>Pipe material</td>
<td>PVC CONC/NOAV AC</td>
<td>3 1.5 0</td>
<td>0.4</td>
</tr>
<tr>
<td></td>
<td>Pipe diameter</td>
<td>Diameter &gt;200 100 &lt; Diameter &lt;=200 Diameter &lt;=100</td>
<td>3 1.5 0</td>
<td>0.3</td>
</tr>
<tr>
<td></td>
<td>Pipe depth</td>
<td>Depth &gt;10 3 &lt; Depth &lt;=10 Depth &lt;=3</td>
<td>3 1.5 0</td>
<td>0.3</td>
</tr>
<tr>
<td>Operational factors 0.35</td>
<td>Network connectivity</td>
<td>Pipe No. &gt;1 Pipe No. = 1 Pipe No. = 0</td>
<td>3 1.5 0</td>
<td>1</td>
</tr>
<tr>
<td>Social factors 0.15</td>
<td>Distance to school</td>
<td>Distance &lt;=100 100 &lt; Distance &lt;=200 Distance &gt;200</td>
<td>3 1.5 0</td>
<td>0.25</td>
</tr>
<tr>
<td></td>
<td>Distance to health facility</td>
<td>Distance &lt;=60 60 &lt; Distance &lt;=120 Distance &gt;120</td>
<td>3 1.5 0</td>
<td>0.25</td>
</tr>
<tr>
<td></td>
<td>Distance to recreational area</td>
<td>Distance &lt;=50 50 &lt; Distance &lt;=100 Distance &gt;100</td>
<td>3 1.5 0</td>
<td>0.15</td>
</tr>
<tr>
<td></td>
<td>Distance to commercial area</td>
<td>Distance &lt;=50 50 &lt; Distance &lt;=100 Distance &gt;100</td>
<td>3 1.5 0</td>
<td>0.15</td>
</tr>
<tr>
<td></td>
<td>Road type</td>
<td>Highway major Collector minor Others</td>
<td>3 1.5 0</td>
<td>0.2</td>
</tr>
<tr>
<td>Environmental factor 0.35</td>
<td>Distance to river</td>
<td>Distance &lt;=15 15 &lt; Distance &lt;=30 Distance &gt;30</td>
<td>3 1.5 0</td>
<td>0.5</td>
</tr>
<tr>
<td></td>
<td>Service type</td>
<td>Industrial Commercial Residential</td>
<td>3 1.5 0</td>
<td>0.5</td>
</tr>
</tbody>
</table>
Table 5.3. The final CoF score distribution

<table>
<thead>
<tr>
<th>CoF cut-off scale</th>
<th>CoF rating</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.06-1.66</td>
<td>5</td>
<td>very high</td>
</tr>
<tr>
<td>1.66-1.26</td>
<td>4</td>
<td>high</td>
</tr>
<tr>
<td>1.26-0.86</td>
<td>3</td>
<td>moderate</td>
</tr>
<tr>
<td>0.86-0.47</td>
<td>2</td>
<td>low to moderate</td>
</tr>
<tr>
<td>0.47-0.07</td>
<td>1</td>
<td>low</td>
</tr>
</tbody>
</table>

Table 5.4. Risk matrix [4]

<table>
<thead>
<tr>
<th>Condition</th>
<th>CoF 1 (Low)</th>
<th>CoF 2 (Fair)</th>
<th>CoF 3 (Moderate)</th>
<th>CoF 4 (High)</th>
<th>CoF 5 (Very High)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 (Low)</td>
<td>Low</td>
<td>Low</td>
<td>Fair</td>
<td>Fair</td>
<td>Moderate</td>
</tr>
<tr>
<td>2 (Fair)</td>
<td>Low</td>
<td>Low</td>
<td>Fair</td>
<td>Moderate</td>
<td>High</td>
</tr>
<tr>
<td>3 (Moderate)</td>
<td>Low</td>
<td>Fair</td>
<td>Moderate</td>
<td>High</td>
<td>Very High</td>
</tr>
<tr>
<td>4 (High)</td>
<td>Low</td>
<td>Fair</td>
<td>Moderate</td>
<td>High</td>
<td>Very High</td>
</tr>
<tr>
<td>5 (Very High)</td>
<td>Fair</td>
<td>Moderate</td>
<td>Moderate</td>
<td>Very High</td>
<td>Very High</td>
</tr>
</tbody>
</table>

5.2 Implementation of Risk Analysis Using GIS Techniques

The risk analysis technique described in the preceding section was implemented in practice in the Village of Lumby, British Columbia. The goal was to demonstrate its utility in pipeline asset management. As will be shown, the proposed method can be used as an affordable, flexible, and more comprehensive asset management tool and can help small communities and municipalities to evaluate the risks associated with the pipe system more effectively.

5.2.1 Research Site: The Village of Lumby

The Village of Lumby is located in the southern interior of British Columbia, Canada, approximately 26 km east of the City of Vernon, the largest municipality in the Regional District of North Okanagan [100]. Since its incorporation in 1955, the Village of Lumby has always served as the gateway community to the Monashee Mountains region and the socio-economic activities remained local. However, its population experienced a significant growth in the 1990s. According to the most recent Statistics Canada (2017) data, in 2016, the village had 1,833
residents [101]. While this growth has helped stimulate local socio-economic activities and has brought diversity and opportunities to the community, it has put considerable strain on the operations of basic services. As indicated by the most recent Village of Lumbys Official Community Plan published in 2014, the village decision-makers are seeking ways to create a more sustainable community. In their view, this can be achieved by supporting initiatives such as age-friendly housing, locally operated healthcare services, and other healthy living proposals in order to satisfy the ever-changing and ever-growing service demand.

The Village of Lumby is also a part of the North Okanagan Regional Growth Strategy Bylaw No. 2500, 2011, which seeks to develop an integrated strategy policy framework for long term regional sustainable growth by addressing issues such as compact complete communities, economic development, transportation, infrastructure, environmental concerns, and resilience and prosperity [100]. The Village of Lumby has also developed its own set of guiding principles that include the protection of natural environment and the promotion of environmental stewardship, the development of a diversified economy, the preservation of heritage and cultural resources, the revitalization of downtown, and the implementation of smart growth strategies as well as the creation of vibrant neighbourhoods [100] (pp. 13-14).

As a small community in the interior of British Columbia, the Village of Lumby has experienced a number of challenges in maintaining their aging infrastructure. Many of the underground pipe systems (e.g., water, gas, and sewer pipelines) have exceed their expiry date (by around 32 years on average). While the cost of replacing the aging infrastructure is prohibitive for such a small community, and even monitoring its condition using the traditional intrusive methods (e.g., physical inspection with CCTV) is expensive, the local community cannot afford the environmental risks associated with pipe failures or meet the high costs of the resulting emergency repairs. In Spring 2014, the downtown area was flooded, requiring extensive and costly recovery efforts. In the same year, the Village of Lumby decision-makers explored the new liquid waste management plan in order to develop long-term directions on the collection, treatment, release and reuse of the liquid wastes which are produced within the community [102]. The work presented in this paper is based on the challenges identified by the Village of Lumby that need to be overcome in order to better manage their sewer infrastructure.
5.2.2 Data Collection and Description

Data used in this study was collected from the AssetFinda™ database, a software package used by the government of Village of Lumby for asset management. In this study, a risk-based approach was implemented on sewer pipe system as the case study. According to the data sourced from the AssetFinda™ database, the sewer pipe system used by the Village of Lumby consists of 351 gravity pipes, 10 force mains and 236 service laterals. The data pertaining to the gravity pipes data is detailed and includes pipe diameter (100 – 300 mm) while the data related to service pipes only contains information on pipe length and pipe installation time. Additionally, as force mains comprise only 2.76% of the sewer systems, they are disregarded in the analysis of gravity lines performed as a part of this project.

5.2.3 Data Summarization for Pipe Condition Analysis

Prior to the condition modeling, data considered in the regression analysis was explored and described first. To obtain an overview of scale of every continuous variable and frequency of different factors considered in structural and operational condition modeling, relevant statistical information related to all key variables (maximum, mean, and quartiles, summarized in Tab. 5.5) was obtained. According to the tabulated data, the maximum depth value is an outlier and there are only two pipes in the industrial area, which comprises only 0.57% of gravity sewer pipes.

Table 5.5. Summary of modeled variables

<table>
<thead>
<tr>
<th>Variables</th>
<th>Min.</th>
<th>1st Qu.</th>
<th>Median</th>
<th>Mean</th>
<th>3rd Qu.</th>
<th>Max.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age (years)</td>
<td>9.00</td>
<td>24.00</td>
<td>38.00</td>
<td>35.24</td>
<td>50.00</td>
<td>55.00</td>
</tr>
<tr>
<td>Pipe depth (m)</td>
<td>1.16</td>
<td>5.51</td>
<td>6.66</td>
<td>6.23</td>
<td>7.94</td>
<td>12.98</td>
</tr>
<tr>
<td>Pipe gradient (°)</td>
<td>0.001</td>
<td>0.005</td>
<td>0.015</td>
<td>0.037</td>
<td>0.061</td>
<td>0.233</td>
</tr>
<tr>
<td>Pipe diameter (mm)</td>
<td>100.0</td>
<td>200.0</td>
<td>200.0</td>
<td>203.7</td>
<td>200.0</td>
<td>300.0</td>
</tr>
<tr>
<td>Pipe length (m)</td>
<td>2.23</td>
<td>36.91</td>
<td>58.05</td>
<td>58.65</td>
<td>78.97</td>
<td>166.11</td>
</tr>
<tr>
<td>Pipe material</td>
<td>AC</td>
<td>CONC</td>
<td>PVC</td>
<td></td>
<td></td>
<td>149</td>
</tr>
<tr>
<td>Service type</td>
<td>Commercial</td>
<td>Industrial</td>
<td>3</td>
<td>Residential</td>
<td>257</td>
<td></td>
</tr>
<tr>
<td>Road type</td>
<td>Collector</td>
<td>Highway</td>
<td>36</td>
<td>Other</td>
<td>253</td>
<td></td>
</tr>
</tbody>
</table>

The scatter plots shown in Fig. 5.3 provide an overview of the pairwise relationships between variables using Pearson correlation coefficient. The bottom left of the matrix is the scatter plot between every two variables and the upper right
of the matrix is the corresponding correlation coefficient between variables. As depicted in Fig. 5.3, there is a strong correlation between pipe material and pipe age, while those between service type area, pipe age, road type, and pipe gradient are not as strong (all pairwise correlations are below 0.8).

5.2.4 GIS Techniques for Consequences of Failure Analysis

In the CoF analysis, there are four aspects studied, i.e., economic, operational, social, and environmental. Based on the multi-criteria matrix described in Tab. 5.2 and data collected from AssetFinda™ database, the CoF scores for economic, op-
erational factors were calculated.

Figure 5.4. Social and environmental impact mapping visualization.

Regarding to the social and environment factors, GIS technique was utilized to locate major institutions and structures and to identify the pipes near these areas. Figure 5.4 visualizes the process how the GIS techniques was employed to extract useful information needed for the CoF analysis. The major buildings such as schools, hospitals, commercial areas were first marked with different notations. For instance, the yellow flags indicate schools and the red crosses present health care service places. The green dot means the recreational area, while the blue dots areas are the commercial areas. Additionally, the long curving shape indicates the river area. Based on the criteria given in the Tab. 5.2, the areas at a close distance from these entities were highlighted with red color and cream color so that pipes
near these areas can be identified and, hence, its CoF score of social and environmental factors could be determined.

5.3 Experimental Results

The calculation results of pipe structural and operational condition are presented graphically in Figs. 5.5a and 5.5b respectively. Based on the observation of Figs. 5.5a and 5.5b, it was found that most pipe structural conditions are fair and most operational conditions are good, where 0 indicates good condition while 5 represents poor performance.

Figure 5.6 is the histogram to depict the pipe CoF details with providing the performance value for each different factors, where smaller value indicates lower risk. Most pipe failure will not cause social and environmental hazard as their low score in these two factors. Figure 5.7 provides the results of the CoF final score, which shows that 70 pipes, around 20%, their failure could make a big impact.

The distribution of the risk indices based on our analysis is depicted in Fig. 5.8. According to our calculations, 15 pipes were identified as very high risk and these are the assets that the Village of Lumby should focus on in the near future.

Based on the experimental results, three strategic recommendations were proposed to the Village of Lumby regarding future policy and prioritization on asset management. First, the importance of the pipe system location was highlighted, since the pipes are buried underground, and can thus be affected by soil degradation and underground water erosion. Hence, a monitoring system should be developed to assess condition of pipes close to any larger body of water, such as a river, a creek, or a pond. Local government should also pay attention to the pipe system connectivity. For instance, as multiple pipes are connected to the system using a single joint, they are usually considered high risk, since the system needs to process higher volume of sewage. Consequently, any system failure would have a much greater impact on the community.

Another recommendation would be a greater focus on the infrastructure in the vicinity of major services and businesses (i.e., government offices, clinics, schools, and major businesses in the downtown district). Any failure of the sewer system (and other infrastructure) in these areas would adversely influence the delivery of
main services and the economic activities. Therefore, it would be recommended that the local government conduct the regular service maintenance.

The final recommendation is to address the importance of data management in assessing risk and predicting socio-economic, health, and environmental impacts of infrastructural failures. Closer inspection of the existing dataset revealed some missing and problematic data (e.g., the age of some pipes was reported as being
Figure 5.6. Number of pipes with different scores in terms of their CoF impacts.

Figure 5.7. Number of pipes in different CoF score categories.
over 100 years old). As database cleanup and flagging problematic information to our research partner (the Village of Lumby) was one of the key aspects of our study, an introduction of a systematic report system was proposed, as this would give the local government an opportunity to capture the changes in risk scores over time. While our model provided a thorough risk assessment of the existing pipelines, the findings yielded also prompted the decision-makers to consider additional factors that should potentially be included in prioritizing the maintenance schedule of the underground infrastructure (e.g., sewer, wastewater, water, electricity, gas, etc.). The GIS mapping component of the proposed model may also facilitate updating the risk scores based on the future development activities. Finally, the weighted score system and evaluation matrix can be utilized by the local government to recalculate the risk scores based on their priorities, which may change over time.

### 5.4 Summary

In this work, a number of research techniques was integrated to assess the risk status of the pipeline system in a resource-constrained community. The risk-based

![Risk Index](image)
analysis approach started with pipe condition evaluation based on the age, material, length, diameter, and slope of each individual pipe, as well as connections to other pipes. The GIS mapping technology was employed in order to evaluate the risk level of the pipe system based on social and environmental factors. The final analysis combined all four dimensions with their respective weights, allowing us to calculate the overall risk level of each individual pipe.
Chapter 6

Decision Support for Asset Management: A Cloud-based Framework

This chapter provides a structured approach to data science and practical guidance with a cloud-based platform for solving real-world problems such as asset risk assessment, predictive maintenance, and infrastructure management. In Section 6.1, the architecture of the platform was demonstrated and the data modeling details for the database design were provided. Section 6.2 presents how to deploy the pipe condition prediction methods in the cloud based on the trained model presented in Chapters 3 and 4 and Section 6.3 demonstrates the dashboard designs for risk mapping and asset health monitoring based on the experiments performed in Chapters 4 and 5.

6.1 System Architecture of the Cloud-based Platform

Figure 6.1 presents the overall structure of the cloud-based decision support system based on the cloud computing service model and the MySQL database system. A set of function modules including MySQL, Machine Learning Studio, and Power BI provided by Microsoft Azure serve as the main server for the system. On this cloud-based infrastructure management platform, MySQL is used to store and
manage the structured data including asset information, field data, climate data, GIS data and etc.. After storing the data in MySQL, several applications can be built on the top of the database system. For instance, machine learning based predictive analytics modeling can be developed as a web service Application Programming Interface (API) for users to predict pipe conditions. Furthermore, a decision support based dashboard can be designed for pipe structural health monitoring and infrastructure management. This platform enable users to have comprehensives view on up-to-date detailed information about different aspects of water distribution systems and provide decision making support in infrastructure management through a web-based application.

Figure 6.1. Cloud-based data management framework for decision analytics.

The pipe information model developed in this work is built upon the findings from previous researches described in Chapters 3, 4, and 5, where the required information for integrated pipe management was identified and described. Figure 6.2 shows the data Entity-Relationship (ER) modeling using Unified Modeling Language (UML), where detailed database modeling schema is described. In this database design, the data schema includes information such as pipe physical factors, environmental factors, operational factors, GIS data, sensor data, and etc. to support the potential usages of the data for infrastructure management. Based on the conceptual database design, the pipe data stored in the cloud could be mapped to the schema and then be used by applications supporting the system.
6.2 Pipe Condition Prediction Models Deployment in Cloud

Utilizing data retrieved from MySQL database in the cloud, machine learning models built in Chapters 3 and 4 can be easily deployed in the production by Azure Machine Learning Studio, which simplifies the deployment of machine learning models through an integrated process [103]. After the deployment, the model runs as a web service that can be easily called by points of service from different platforms including servers, laptops, tablets, or smartphones.

Figure 6.3 shows an example of the predictive experiment deployed as a web service, which can be used for cast iron water mains condition prediction. In the web service module, there are two key components, i.e., the original data for data training and trained model, and a new web input and output were added. Data used
for machine learning is reserved for the usage of the schema information such as the number of columns and data type of each columns to interpret the input data entered by users. Based on the user inputs, the trained model will predict the condition and send the results to users through the web service output. Additionally, this web service can be published by an API Key so that it can be invoked in other platforms written by C#, Python, or R \cite{103}.

### 6.3 Decision Analytics Implementation in Cloud

Data analysis and data modeling is essential, but presenting and interpreting the results is just as important. To help users understand the information effectively, data visualization is a powerful technique used to summarize and communicate data by different charts, graphs, diagrams, and maps. Utilizing the data visualization techniques, dashboard offers unique and powerful solutions to an organization’s need for information by displaying several indicators through a common graphic interface \cite{104}. In a dashboard, the most important information will be consolidated and arranged on a single screen so the information can be monitored at a glance \cite{105}. 

---

**Figure 6.3.** The predictive experiment based on saved training model using web service.
In our cloud-based framework, two interactive dashboards were designed for asset risk mapping and health monitoring using Power BI, which provides new ways to visualize data and share results [103]. Figure 6.4 presents the analytical dashboard which provides critical information needed to prioritize asset maintenance and rehabilitation based on risk analysis results in Chapter 5. Histograms and donut charts were used to depict the frequency and proportion of different risk level pipes. Additionally, a table was created in the dashboard with detailed information about ‘high’ and ‘very high’ risk pipes. Finally, the map in the center of the dashboard presents the location of all pipes noted as different circles, where circle size indicates different risk level (the larger the circle is, the higher risk the pipe has) and different color means different material pipes. The graphs and map in the dashboard provide detailed information about pipe performance and risk level, without users needing to learn how to handle data or use specialist visualization software.

Another dashboard, shown as Fig. 6.5, was designed for asset performance monitoring by graphically summarizing sensor data from a field monitoring system in flux. Climate data was integrated and presented in this dashboard to provide more comprehensive views for users and users can choose to observe data in different time ranges by click the ‘Year’ buttons. This dashboard provides users with real time up-to-date detailed information about different aspects of water distribution systems and milieu, and their variation over time.

6.4 Summary

In this chapter, a cloud-based system was proposed for data analysis and infrastructure management. The platform was implemented by Azure cloud service, which enables scalable and flexible data management. In this system, ER modeling was used for database design, where pipe properties, GIS information, sensor data, and etc. could be structured, integrated and managed effectively. Additionally, machine learning based prediction models can be deployed and published easily as a web service and decision analytics can be implemented elegantly through interactive dashboards. The proposed cloud-based framework can facilitate the pipe data management and provide decision making support for human operators.
Figure 6.4. Dashboard design for decision support on asset management. Hitogram in the left: frequency of different condition, failure consequence, and risk level pipes. Map in the upper center: location of different risk level pipe. Donut in the bottom center: proportion of different condition, failure consequence, and risk level pipes. Table in the right: list of ‘high’ and ‘very high’ risk pipes.
Figure 6.5. Dashboard design for pipe structural health monitoring with real-time series plot. Left: sensor network data including pipe strain, soil temperature, soil pressure, and soil water contents. Right: climate condition data including precipitation and rainfall deficit.
Chapter 7

Conclusions

In this thesis, the pipe condition prediction methods and risk assessment for pipe systems were studied. Based on these studies, a cloud-based decision support system was designed for infrastructure monitoring and management. The contributions of this thesis can be summarized as follows.

- With the growing demand for utility services and aging water mains, asset condition evaluation becomes increasingly important. In this thesis, novel data-driven approaches were proposed to improve the pipe performance prediction from different aspects, which is one of the main contributions of this study. The superiority of the proposed methods was demonstrated in two studies: In the first study, the relationship between the soil properties and pipe condition rating was investigated based on the understanding of pipe deterioration failure mechanisms using stacking ensemble based methods. The superiority of the stacking ensemble learning was validated through experiments, as a part of which comparison was made with the state of the art machine techniques. In another study, the pipe deformation was then predicted considering soil and climate data based on the supervised learning techniques, i.e., integrating super learning with random forest-recursive feature elimination algorithm. Experimenting with three historical sensor data sets, the adaptability of super learning for combining information from a set of algorithms to improve the prediction results was proved. By employing the output ensemble models, pipe condition can be predicted, so that the de-
cision for maintenance and rehabilitation of the water mains system can be supported with the derived evidence. Additionally, the findings in these studies yielded can assist municipalities in identifying critical parameters related to water mains performance.

- Another main contribution of this study stems from providing a holistic approach to the assessment of pipe system risks. In this study, a multiple-level risk assessment of pipe networks is presented considering uncertainties and interdependencies among the key factors and sub-factors. Four risk factors were identified to affect the performance of sewer pipelines and networks, which were classified as having physical, environmental, social, and operational impact. Each pipe in our analysis was assigned a score based on a comprehensive risk assessment. Further investigation allowed us to present the distribution of the risk levels among all pipes in the examined system, where the very high risk pipes could also be identified. While it is true that physical inspection (e.g., CCTV recording, flow monitoring, and manhole inspection) remains the best way to monitor pipe condition, the high cost and intrusive nature of such inspections may not be the best choice for smaller and resource-constrained communities such as the Village of Lumby. The risk assessment approach presented in this work provides a useful and effective mechanism for the local government to prioritize the inspection process.

- Finally, in this thesis, a cloud-based platform was proposed which provides a solution for infrastructure management based on the research findings described above. This system demonstrates how to implement the proposed supervised learning techniques and risk analysis methods in a production environment by providing a user-friendly dashboards. The platform can facilitate the infrastructure management and support stakeholder decision making on asset maintenance, renewal and replacement.

This thesis shows that data-driven approaches provide a promising way for pipe condition prediction, which is cost effective, accurate, and scalable. However, there are also some limitations in these studies. The challenges and the potential future studies for the researches in this thesis are discussed and listed as below.
• The data used in the modelling was collected either from research centre or university, which may not be available in many municipalities. A data ecosystem is thus needed to be further researched so that the data used in the modeling can reflect the real world problems. Since the approaches used in this study highly rely on the data, high quality data are required to well-implemented these methods. Hence, future study should focus on the design of more efficient and economic data collection strategies, as well.

• In the risk analysis study, the CCTV reports were excluded from the analyses. Thus, future studies in this field would benefit from incorporating the CCTV assessment, as well as a more comprehensive evaluation that would involve additional condition measurements. For instance, several literature reported that soil acidity and agricultural activities, as well as traffic volume on the road above the pipe system, may impact pipe condition. Thus, in future research, impacts of other activities and physical environment on pipe condition should also be measured.

• Future studies can also gather experts (e.g., site engineers) opinions to ascertain the validity of the weighted parameters in a particular spatial context. Moreover, developing a more comprehensive model that includes hydraulic effects would allow researchers to accurately determine the water flow direction in the assessed network. In sum, collecting actual performance indices from several networks could provide additional ground truth to validate the accuracy of prediction and to determine a suitable value for the parameters.

• The cloud based decision support system is a prototype which hasn’t been tested in a production environment. Therefore, a pilot experiment can be conducted in the future to test the feasibility of the proposed framework for practical use and adoption, and more importantly, to bridge the gap between experimental models and the current utility practice.
Bibliography


88


