Evaluation of Statistical and Machine Learning Methods on Predicting One-way Shear Strength of Reinforced Concrete Beams

by

Hongyang Wu

A thesis submitted to the Graduate Faculty of
Auburn University
in partial fulfillment of the
requirements for the Degree of
Master of Science

Auburn, Alabama
May 7, 2022

Keywords: regression, machine learning, reinforced concrete, one-way shear strength, building code, database, accuracy

Approved by

Nedret Billor, Chair, Professor of Mathematics and Statistics
Elvan Ceyhan, Associate Professor of Mathematics and Statistics
Roberto Molinari, Assistant Professor of Mathematics and Statistic

Abstract

It is a well-known fact that predicting shear strength of concrete beams is a very complex topic. The traditional approach is to assume that the shear strength is the sum of the concrete contribution (V_c) and the shear reinforcement contribution (V_s). In recent years, several proposals including Bentz & Collins (2017), Cladera et al. (2017), Frosch et al. (2017), Li et al. (2017), Park & Choi (2017) and Reineck (2017) were published based on different failure mechanisms. The current code ACI318-19 (2019) also updated the one-way shear design specifications with a new set of empirical equations based on Kuchma's research (1998). However, none of these methods can fully agree with each other either. Therefore, methods which are not directly related with failure mechanisms will be valuable to provide alternative ways to predict shear strength of concrete beams as comparisons. Recently, two database including shear database without shear reinforcement (Reineck, Kuchma, & Kim, 2003) and shear database with shear reinforcement (Reineck, Bentz, & Fitik, 2014)were collected and reviewed. With these two databases, several statistical and machine learning (ML) methods can be applied and evaluated as alternative methods to compare with traditional theoretical methods. In this thesis, three theoretical methods (ACI318-19, Frosch et al. and Li et al.) were chosen to be validated by these two databases. Selected statistic (multiple linear regression, LASSO, LARS) and machine learning (random forest, neural network, and support vector machine) methods were applied on the databases to evaluate the efficiency and accuracy of these methods on predicting shear strength. This study is the first study attempting to apply the statistical/ML methods on these databases to predict the shear strength of concrete beams. A sensitivity analysis was conducted on all these different methods at the end of this thesis to evaluate the stability of these methods.

Acknowledgment

I would like to thank my advisor Dr. Billor for her help in my courses and research throughout these two years at Auburn University. I really appreciate her vast knowledge, guidance, and assistance in writing this thesis. I really would like to thank her for giving me a chance to accomplish a master's degree in the department of Mathematics and Statistics in Auburn University. Her patience and passion have encouraged me greatly. Also, I would like to thank the rest of my committee, Dr. Ceyhan, and Dr. Molinari for all the help they have provided.

I also would like to thank my parents for their support and understanding in my entire life. I also appreciate my friends for their friendship and help, especially Victor Aguilar who helped me to come up with ideal in this paper, provided me the database which is used in this thesis.

I would also thank my officemates and all the other fellow graduate students and faculty members. They have offered me a great experience at Auburn University, which I will never forget.

Table of Contents

Abstract ii

Acknowledgment iii

List of Tables vii

List of Figures viii

Chapter 1. Introduction 1

1.1 Motivation for the Research 1

1.2 Scope of work 3

1.3 Organization of thesis 3

Chapter 2 Literature Review 5

2.1 Introduction of shear behavior and shear strength of concrete beam 5

2.1.1 Behavior of beams without shear reinforcement 6

2.1.2 Behavior of beams with shear reinforcement 8

2.2 Database for one way shear in reinforced concrete beams 9

2.3 Statistical/ML methods in structural engineering 10

2.6 Summary 16

Chapter 3 Methodology 17

3.1 Traditional methods (Theoretical methods) 17
3.1.1 ACI 318-19 17
3.1.2 Method by Frosch et al. (2017) 18
3.1.3 Method by Li et al. (2017) 19
3.2 Statistical methods 20
3.2.1 Multiple linear regression and Stepwise Variable Selection Methods 20
3.2.2 Least absolute shrinkage and selection operator (LASSO) 21
3.2.3 Least angle regression (LARS) 22
3.3 Machine learning method 23
3.3.1 Tree based method 23
3.3.2 Neural Networks 26
3.3.3 Support vector machines 28
3.5 Summary 30
Chapter 4 Result and Discussion 31
4.1 Introduction 31
4.2 Beam without shear reinforcement 31
4.2.1 Descriptive analysis 31
4.2.2 Variable selection and Estimation 33
4.2.3 Accuracy 37
4.3 Beam with shear reinforcement 54

4.3.1 Descriptive analysis 54

4.3.2 Variable selection and Estimation 56

4.3.3 Accuracy 60

4.2.4 Sensitivity Analysis 65

4.4 Summary 72

Chapter 5 Conclusion and Future work 73

References 75

List of Tables

Table 4-1: Summary Statistics of variables for beam without shear reinforcement 32
Table 4-2: Correlation Matrix of all the variables for beam without shear reinforcement 33
Table 4-3: Results of variable selection by forward method from MLR for beam without shear reinforcement 34
Table 4-4: Results of variable selection by backward method from MLR for beam without shear reinforcement 34
Table 4-5: Result of variable selection by stepwise method from MLR for beam without shear reinforcement 35
Table 4-6: Results of goodness of fit for all the methods 43
Table 4-7: The details of test dataset for beam without shear reinforcement 45
Table 4-8: Ranges for small test dataset for beam without shear reinforcement 49
Table 4-9: Summary Statistics of variables for beam with shear reinforcement 55
Table 4-10: Correlation Matrix of all the variables for beam with shear reinforcement 56
Table 4-11: Results of variable selection by forward method from MLR for beam with shear reinforcement 57
Table 4-12: Results of variable selection by backward method from MLR for beam with shear reinforcement 57
Table 4-13: Results of variable selection by stepwise method from MLR for beam with shear reinforcement 58
Table 4-14: Results of goodness of fit for all the methods for concrete beam with shear reinforcement 65
Table 4-15: The detail of test dataset for beam with shear reinforcement.................................... 66

List of Figures

Figure 1-1: Primary shear resistance mechanisms (Aguilar, 2020) 2
Figure 2-1: Simple support beam subjected to distributed load (Aguilar, 2020) 5
Figure 2-2: Stress condition on the different location of beam (Aguilar, 2020) 6
Figure 2-3: Sketch of diagonal tension failure of beam (Aguilar, 2020) 7
Figure 2-4: Beam with shear reinforcement (Aguilar, 2020) 9
Figure 2-5: Relationship between Artificial Intelligence, Machine Learning and Machine learning techniques (AGUILAR, WU, & MONTGOMERY, 2021) 13
Figure 2-6: Difference between Traditional Method and Machine Learning Method (AGUILAR, WU, & MONTGOMERY, 2021) 13
Figure 3-1: Sample decision tree based on binary target variable Y (SONG & LU, 2015) 25
Figure 3-2: Neural network diagram with a single hidden layer (Efron & Hastie, 2016) 27
Figure 3-3: Neural network diagram with three hidden layers (Efron & Hastie, 2016) 27
Figure 3-4: Margin maximization of SVMs (Swan, 2017) 29
Figure 4-1: Result of variable selection from LASSO for beam without shear reinforcement 35
Figure 4-2: Result of variable selection from LARS for beam without shear reinforcement (Cross validated MSE as number of steps increases) 35
Figure 4-3: Variable Importance Plot generated by random Forest for beam without shear reinforcement 36
Figure 4-4: Scatter Plot Matrix of selected variables for beam without shear reinforcement 37
Figure 4-5: Comparison between experiment and uncorrected prediction by theoretical methods for beam without shear reinforcement 38
Figure 4-6: Comparison between experiment and corrected prediction by theoretical methods for beam without shear reinforcement 39
Figure 4-7: Comparison between experiment and prediction by statistical methods for beam without shear reinforcement 41
Figure 4-8: Comparison between experiment and prediction by machine learning methods for beam without shear reinforcement 42
Figure 4-9: Sensitivity analysis on bw for concrete beam without shear reinforcement 45
Figure 4-10: Sensitivity analysis on fc_prime for concrete beam without shear reinforcement 46
Figure 4-11: Sensitivity analysis on kap for concrete beam without shear reinforcement 46
Figure 4-12: Sensitivity analysis on low for concrete beam without shear reinforcement 47
Figure 4-13: Sensitivity analysis on ds for concrete beam without shear reinforcement 47
Figure 4-14: Typical boxplot (Galarnyk, 2018) 49
Figure 4-15: Sensitivity analysis on bw for concrete beam without shear reinforcement in small dataset 50
Figure 4-16: Sensitivity analysis on fc_prime for concrete beam without shear reinforcement in small dataset 50
Figure 4-17: Sensitivity analysis on kap for concrete beam without shear reinforcement in small dataset 51

Figure 4-18: Sensitivity analysis on low for concrete beam without shear reinforcement in small dataset 51
Figure 4-19: Sensitivity analysis on ds for concrete beam without shear reinforcement in small dataset 52
Figure 4-20: Result of variable selection from LASSO for beam with shear reinforcement 58
Figure 4-21: Variable Importance Plot generated by Random Forest for beam with shear reinforcement 59
Figure 4-22: Scatter Plot Matrix of selected variables for beam with shear reinforcement 60
Figure 4-23: Comparison between experiment and uncorrected prediction by theoretical methods for beam with shear reinforcement 61
Figure 4-24: Comparison between experiment and corrected prediction by theoretical methods for beam with shear reinforcement 62
Figure 4-25: Comparison between experiment and prediction by statistical methods for beam with shear reinforcement........ 63
Figure 4-26: Comparison between experiment and prediction by machine learning methods for beam with shear reinforcement........ 64
Figure 4-27: Sensitivity analysis on bw for concrete beam with shear reinforcement................ 66
Figure 4-28: Sensitivity analysis on fc_prime for concrete beam with shear reinforcement 67
Figure 4-29: Sensitivity analysis on kap for concrete beam with shear reinforcement................ 67
Figure 4-30: Sensitivity analysis on low for concrete beam with shear reinforcement................ 68
Figure 4-31: Sensitivity analysis on ds for concrete beam with shear reinforcement 68
Figure 4-32: Sensitivity analysis on Asw for concrete beam with shear reinforcement 69
Figure 4-33: Sensitivity analysis on sw for concrete beam with shear reinforcement 69
Figure 4-34: Sensitivity analysis on fyw for concrete beam with shear reinforcement................ 70

Chapter 1. Introduction

1.1 Motivation for the Research

It is difficult to predict one-way shear strength of concrete beam due to its complex failure mechanism. Even today, there is no widespread agreement regarding a shear resistance mathematical model. The traditional assumption is the contribution of concrete and rebars in reinforced concrete beam is separate, that is, they are independent from one another. Nevertheless, consider Figure 1-1, if the primary resistance mechanisms (Kuchma & Collins, 1998) are observed one can notice the external shear is resisted by shear transfer through uncracked concrete, the aggregate interlock along cracks also contributes to the shear strength and the dowel action of longitudinal reinforcement. Moreover, If the beams have an amount of shear reinforcement, the transverse reinforcement also participates in the shear resistance mechanism. On the one hand, the shear reinforcement holds some of the shear force, but also it keeps the cracks from widening; therefore, it has an influence in the concrete contribution coming from aggregate interlock. The quantification of these mechanisms is by no means not straightforward. The shear strength of concrete members is influenced by tensile concrete strength, coarse aggregate size, presence of axial force, slenderness ratio, amount of longitudinal reinforcement, and overall size of the member. Yet, the relative influence of these variables is still debated (Aguilar, 2020). Several proposals including Frosch et al. (2017) and Li et al. (2017) were published based on their thought of failure mechanisms on the concrete beam bending behavior. The empirical formula which was developed by Kuchma (1998) was accepted as new method to *predict* and *design* shear strength of reinforced concrete beam in the current concrete code.

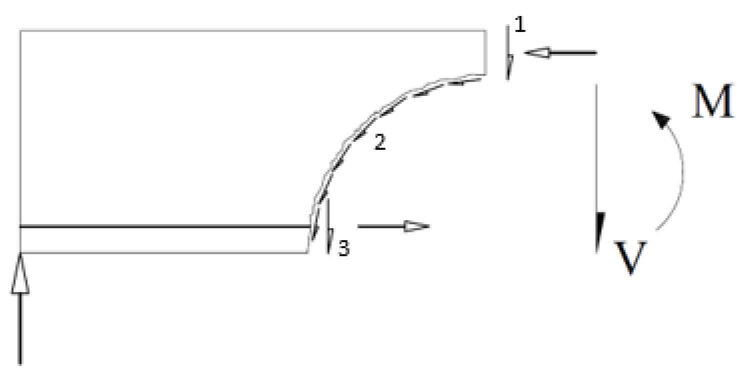

Figure 1-1: Primary shear resistance mechanisms (Aguilar, 2020)

Given the complexity of the topic, there have been numerous experimental efforts to try to understand the shear strength of concrete members. Thus, databases have been assembled and reviewed by the engineering community (Reineck et al. 2003, Reineck et al. 2014). Aguilar (2020) studied the accuracy of those shear design equations and the ACI 318 (2019) building code requirements for structural concrete against the available ACI 445 databases. The result of analysis showed none of these theoretical methods can agree with each other since the assumed shear failure mechanisms they have were different from one another. In addition, the conclusions indicated that the new ACI 318-19 shear design method is the simplest of the methods reviewed, and it gives reasonably accurate predictions for members with and without shear reinforcement. Also, the performance of the proposals from Frosch et al. (2017) and Li et al. (2017) was highlighted. Therefore, the one-way shear equations from ACI 318-19(2019), Frosch et al. (2017), and Li et al. (2017) are selected for further analysis and comparisons in this thesis. Also, in this thesis, with the available experimental databases, statistical (multiple linear regression, LASSO, LARS) and machine learning (Random Forest, support vector machine) method were the first-time attempt to be applied to provide alternative ways to predict the one-way shear strength of concrete beams. Further, *sensitivity analysis* was conducted to evaluate the stability of all these statistical/ML methods compared with the theoretical methods developed by previous researchers (ACI 318-19 (2019), Frosch et al. (2017), and Li et al. (2017)).

1.2 Scope of work

The purpose of this thesis is to attempt to predict one-way shear strength of reinforced concrete beam from ACI 445 database by using statistical/ML techniques and evaluate these methods by conducting sensitivity analysis. Three statistical methods including multiple linear regression, LASSO and LARS and the three machine learning methods including Random Forest, Neural Network and Support Vector Machine were selected in this thesis to be evaluated. The datasets used in this thesis include reinforced concrete members without shear reinforcement (784 samples), and reinforced concrete members with shear reinforcement (87 samples) from the ACI 445 database. To assess the performance of the statistical/ML techniques, three theoretical (mathematical) methods which are ACI318-19 (2019), Frosch et al. (2017) and Li et al. (2017) were also used for comparison purposes in this thesis.

1.3 Organization of thesis

Chapter 1 contains an introduction to define the specific problem that exists in structural engineering in this thesis. The motivation of the research and scope of the work are also discussed in this chapter.

Chapter 2 provides a literature review on shear behavior and shear strength of concrete beam, database used in this thesis and, statistical/ML in civil engineering field.

Chapter 3 provides an overview on the selected statistical/ML and theoretical (mathematical) methods in this thesis.

Chapter 4 presents the details of comparison between experimental results(databases) and results which predicted by all these different methods (theoretical and statistical/ML methods). The results of sensitivity analysis were also presented in this chapter.

Finally, Chapter 5 consists of the conclusion and future work. This chapter will summarize the whole research done in this thesis and provide recommendations for further research.

Chapter 2 Literature Review

2.1 Introduction of shear behavior and shear strength of concrete beam

In structural design, beam is one of the most common elements in the buildings, bridges, stadiums and so on. The major function of beam is to transfer the load to columns then the foundations. Figure 2-1 shows a simple support beam which is subjected to a distributed load and the corresponding cracking due to flexure behavior. When the external force was applied in the beam, the internal force including bending moment and shear force will be generated along beam. Then, due to bending moment, normal compression stress occurs above neutral axis while normal tensile stress occurs below neutral axis. The shear force is the internal force resultant of all the vertical forces acting on beam, at the section under consideration, and it tends to cut or shear the beam. The shear force creates both vertical and horizontal stress.

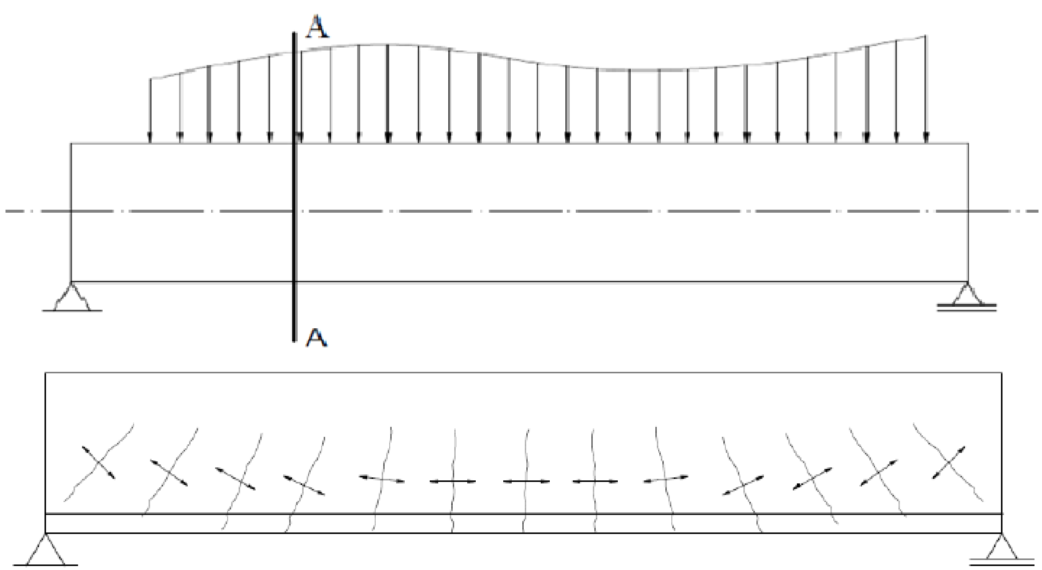

Figure 2-1: Simple support beam subjected to distributed load (Aguilar, 2020)

Figure 2-2 shows the stress conditions on the different location of beam which are above neutral axis, at neutral axis and below neutral axis, respectively. Based on the property of concrete material, The concrete is strong in compression and weak in tension. Therefore, the concrete below neutral axis will start to crack first which is shown in Figure 2-1.

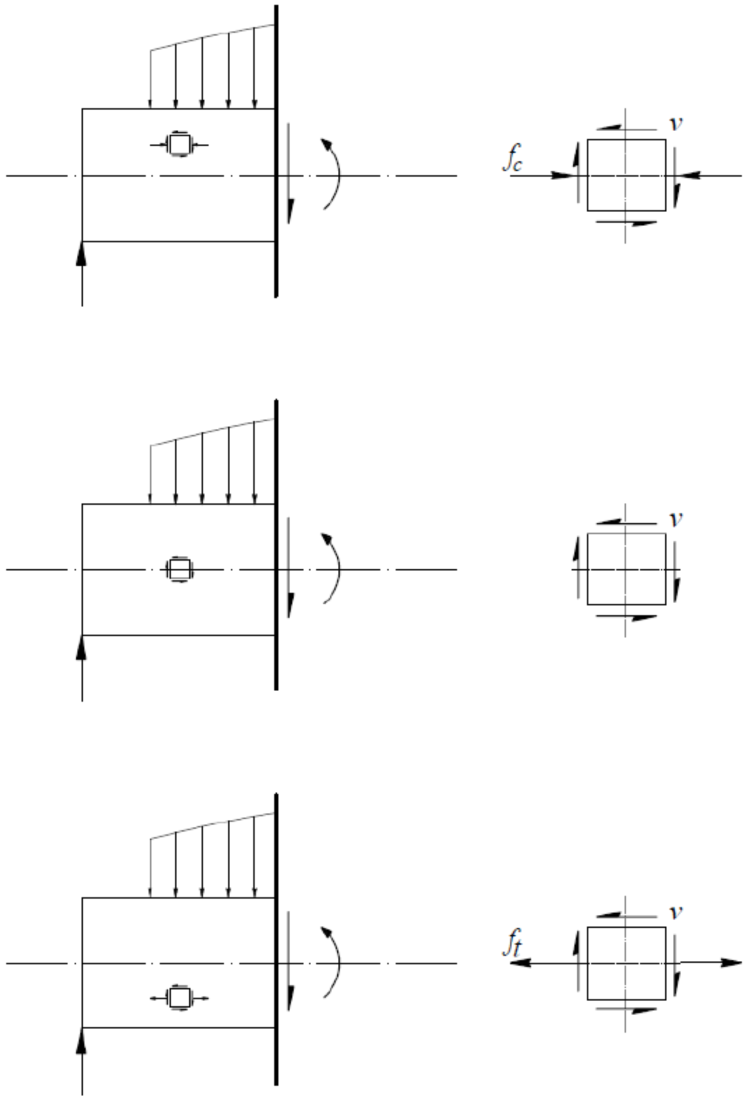

Figure 2-2: Stress condition on the different location of beam (Aguilar, 2020)

2.1.1 Behavior of beams without shear reinforcement

For the simple support beam under vertical uniformed distributed load, the maximum moment occurs in the middle span of the beam while the maximum shear force occurs at the support. When the tensile stress generated by bending moment exceeds the rupture modulus, cracks open almost perpendicular to the axis of beam which was called flexural cracks. When the shear stress generated by shear force reaches the critical stress, the diagonal crack starts to occur. In the location where large shear force occurs, inclined cracks will be developed after those exist flexural cracks which is called flexural-shear cracks.

There are two types of shear failure modes for concrete beam due to its slenderness which is the ratio of the shear span (distance between load and support) to the depth of cross section. When the slenderness is between 1 and 2.5, either shear compression or shear tension failure will happen. This type of failure mode is brittle failure mode with very small deflection which will not be discussed in this thesis. When the slenderness is between 2.5 and 6, the diagonal tension failure will occur. When the slenderness is larger than 6, flexure failure will happen before shear failure. Figure 2-3 shows the sketch of a beam with diagonal tension failure.

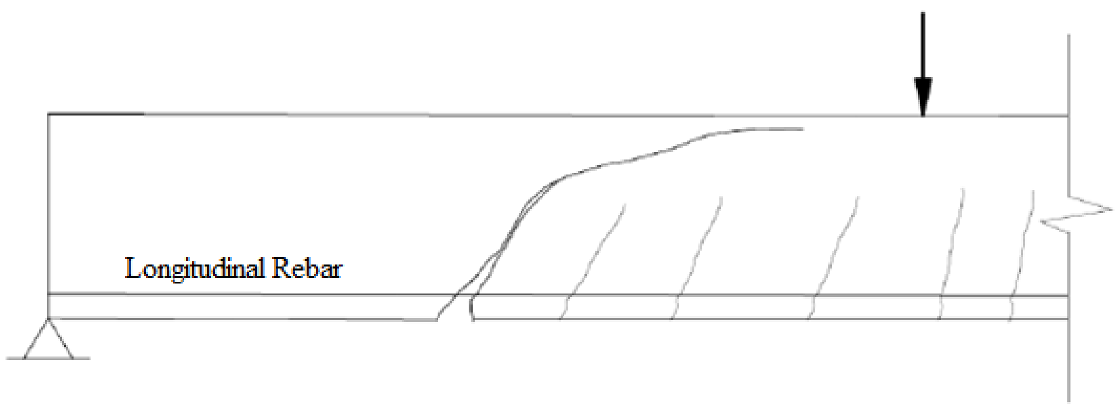

Figure 2-3: Sketch of diagonal tension failure of beam (Aguilar, 2020)

The cracking start to occur in the middle span of beam with some small vertical cracks. Then, more flexural crack shows up along the beam. These flexural cracks will break the bond between concrete and bottom tensile rebars. Then, diagonal flexural-shear crack shows up and widens into a diagonal tension crack which is extended to the top of beam (Wang & Salmon, 1985). Based on Wight’s (Wight & MacGregor, 2012) book, the beam without shear reinforcement will fail quickly after the diagonal tension crack occurs. Therefore, the shear force which causes the inclined crack is considered as the shear strength of concrete beam. Although there are some different shear failure models of beam without shear rebars by different researchers, it was widely accepted that the external shear force was resisted by: (1) shear transfer through uncracked concrete; (2) the aggregate interlock along cracks also contributes to the shear strength; and (3) the dowel action of longitudinal reinforcement. Based on Wang and Salmon's (1985) book, for the rectangular beam without shear rebars, 20-40 percentage of shear is transferred by (1) uncracked concrete, 33-50 percentage of shear is transferred by (2) the aggregate interlock along cracks and 15-25 percentage is transferred by (3) the dowel action of longitudinal reinforcement. According to these major resistance mechanisms, the shear strength of the concrete beam is affected by a lot of variables such as the *size of beam*, *concrete compression strength*, the *aggregate size*, *the longitudinal reinforcement ratio* and so on.

2.1.2 Behavior of beams with shear reinforcement

When the slenderness is between 2.5 and 6 or the bottom longitudinal reinforcement is overly designed, the inclined crack will happen before the beam reaches its flexural capacity which means shear force will control the design in these cases. Therefore, it is important to provide enough shear strength to resist the shear force. When the shear resistance provided by concrete itself is not enough in the beam, shear reinforcement becomes necessary in the beam design. The purpose of shear rebars is to ensure that the flexural strength can be fully achieved to have a ductile failure mode instead of a brittle shear mode for safety purpose. The prime idea of shear reinforcement is to provide steel to cross the diagonal tension cracks to stop them from widening. The most common shape of shear rebar is vertical U-shape. Figure 2-4 shows the beam with shear reinforcement.

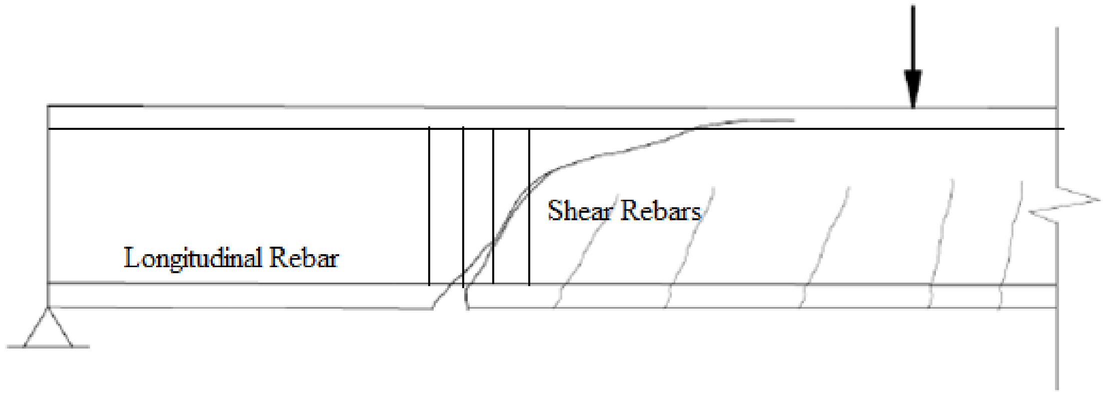

Figure 2-4: Beam with shear reinforcement (Aguilar, 2020)

The shear rebars will not start to be functional until the diagonal tension crack forms. Before that, the stress in the shear rebar is very small. After the diagonal crack forms, the shear rebars start to work to prevent the crack becoming larger and larger. Wang's (1985) book said that if the shar reinforcement is too little, the shear rebars will yield immediately after the formation of inclined crack. If the shear reinforcement is too much, the brittle shear failure will occur. There are three major functions of shear reinforcement in the concrete beams: (1) carry part of shear force, (2) prevent the widening of inclined shear cracks and, (3) hold the longitudinal rebars in the right position to reduce the dowel effect (Wight & MacGregor, 2012). Traditionally, the one-way shear strength is assumed to be the *sum of shear strength* from *concrete* and shear strength from the *shear reinforcement*. Besides the variables which affect the shear strength from concrete, rebar yields strength of shear reinforcement and the space of shear reinforcement are both important which will affect the shear strength from shear rebars.

2.2 Database for one way shear in reinforced concrete beams

To investigate the one-way shear strength of concrete beam, a significant number of efforts were invested in the experimental tests as well. In the past 15 years, a database of slender beams (slenderness ratio larger than 2.4) with clear shear failure was assembled and summarized by the Joint ACI-ASCE Committee 445 and German Committee of reinforced concrete (DAbStb). The dataset with 784 samples was used for the concrete beam without shear reinforcement. There are two datasets which can be used for concrete beam with reinforcement. The large dataset has 170 samples while the small dataset has 87 samples. The small dataset is a subset of large dataset which all the shear rebars were ensured to yield before shear failure occurs under observation. The small dataset was selected to be used in this thesis for the clarification purpose. The description of all the information related with the variables (information of beam geometry, concrete and rebar) in the dataset can be found in the research report from Reineck at el. (Reineck, Kuchma, & Kim, 2003) (Reineck, Bentz, & Fitik, 2014).

2.3 Statistical/ML methods in structural engineering

Before 1990s, linear elastic analysis is very common in the design and analysis practice of structural engineering. A linear elastic analysis is a type of analysis which holds a linear relation between force and displacement. Based on this assumption, this method is suitable for structural engineering problems where all the materials remain in the elastic range during the analysis. For linear elastic analysis, both theoretical and numerical methods can be used. The theoretical methods can be formulated by differential equations or simple calculations based on different concepts while the numerical methods are commonly explained and demonstrated by using matrix method according to all degrees of freedom. In addition, for a linear elastic analysis, the global and local stiffness matrices for systems and elements are constant. Therefore, the corresponding solve process is relatively short and computational inexpensive compared with nonlinear analysis procedures. However, the assumption of linear elastic is not always true in the real world, nonlinear plastic process sometimes is necessary when a more accurate estimate of behavior is needed. A nonlinear analysis is a type of analysis which holds a nonlinear relation between force and displacement. There are three major types of nonlinear effects in structural engineering including geometrical nonlinearity, material nonlinearity, and contact nonlinearity.

Numerical methods are commonly used in the nonlinear analysis. The global and local stiffness matrices cannot remain the same during the analysis opposed to the linear elastic analysis. Therefore, the corresponding solve process is relatively long and computational expensive. Meanwhile, the development of analysis hardware and software provide the possibility for engineers to get results from nonlinear process, for example, finite element analysis. However, to perform a nonlinear analysis, more theoretical and practical knowledge is required to determine all the input parameters and models including the effect of different types of nonlinearities which can easily be inappropriate. Incorrect selection of parameters and assumptions will cause significant error in the analysis while correct selection of parameters and assumption can obtain reasonable results. Therefore, model validation with experiments or field test is often needed to make final decision. In conclusion, linear elastic analysis is simple and relatively inaccurate while nonlinear plastic analysis is complicated and relatively accurate.

Meanwhile, in the past several decades, structural data generation has increased dramatically due to the development of technology of measuring, recording and storing data. These data come not only from reliable nonlinear analysis but also from laboratory and field tests. By using these growing volumes of databases, opportunities are opened to engineers and scientists to come up with simple and accurate ways to predict structural behavior including beam behavior, column behavior, joint behavior and so on. Artificial intelligence is one of them. The research of artificial intelligence has been developed since 1956, when the term “Artificial Intelligence, AI” was used at the meeting hold in Dartmouth College (Lu, Chen, & Zheng, 2012).

AI is the simulation of human intelligence processes by machines, especially computer systems. These processes include learning (the acquisition of information and rules for using the information), reasoning (using rules to reach approximate or definite conclusions) and self-correction. There are several ways to achieve artificial intelligence, machine learning is one of the most popular approaches. Figure 2-5 illustrates the relationship between Artificial Intelligence, ML and ML techniques (AGUILAR, WU, & MONTGOMERY, 2021). Based on Figure 2-5, artificial intelligence is the broadest term which includes the whole machine learning term. Tree-based method, Neural Networks and Support Vector Machine are three different techniques for implementing machine learning approach. These three machine learning techniques have been used in some structural problems which will be emphasized in this literature review.

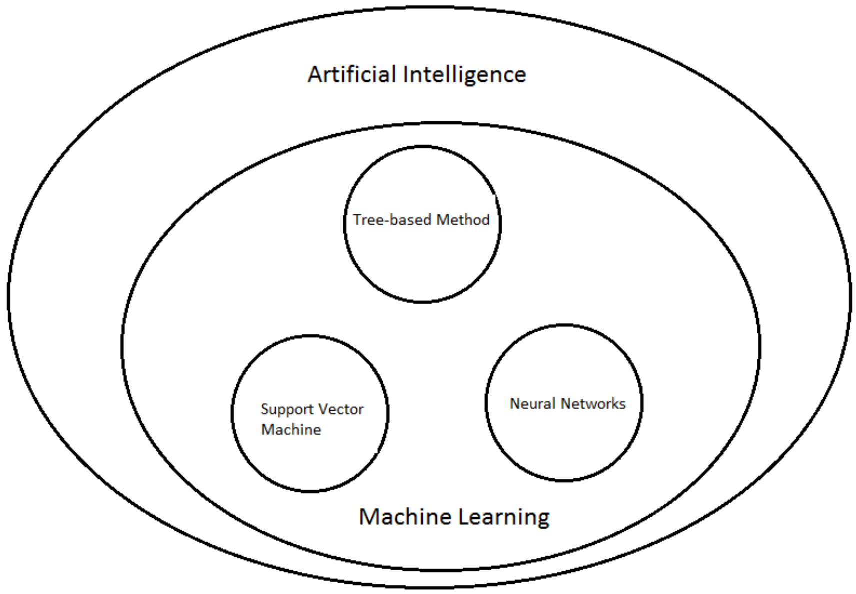

Figure 2-5: Relationship between Artificial Intelligence, Machine Learning and Machine learning techniques (AGUILAR, WU, & MONTGOMERY, 2021)

Practical applications of machine learning are very different from traditional or empirical approaches in structural engineering. Figure 2-6 shows the differences of analysis procedures between traditional approach and machine learning approach.

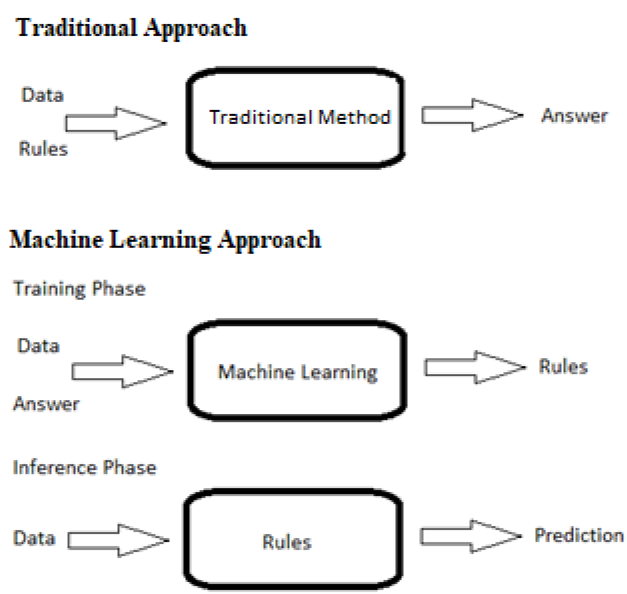

Figure 2-6: Difference between Traditional Method and Machine Learning Method (AGUILAR, WU, & MONTGOMERY, 2021)

The definition of traditional methods is very broad since all the methods which don't belong to artificial intelligence approaches in the past several decades will be considered as traditional methods including laboratory test, field test, finite element analysis, empirical methods and so on. There are several common characteristics among these traditional methods. First, the input data information is preselected and reasonably simplified which means the input values are normally required to be preprocessed before the test or analysis. Then, all the test and analysis procedures and methods which are described as "Rules" in Figure 2-6 need to be well understood and properly applied to physical tests or computer analysis. Otherwise, the test or analysis results might not be reliable and valuable. Last but not the least, the relationship between answer or prediction and input valuable values are clear and interpretable since all the rules are known and well applied which is different from machine learning approaches. For machine learning approaches, the procedures of methods are completely different from traditional methods. First, the raw data sets are not necessarily to be pre-processed before applying machine learning approach. Second, the theories of machine learning approaches need to be well understood. However, this is different from the "Rules" referred in the traditional method, the "rules" are created by machine learning approaches in training phase. The most different part between traditional methods and machine learning methods is the interpretation section. Although the "Rules" are created by machine learning approaches which are well understood and carefully selected, the "Rules" itself is developed by computer in a "Black box "which means there is no direct relationship between input values and answers or predictions which can be interpreted. The main goal of machine learning approaches is to provide an answer or prediction as accurate as possible in the inference phase. There are a lot of machine learning techniques which have been developed in the past several decades. Among all these techniques, Tree-based method, Neural Networks and Support Vector Machine are three different techniques which have been frequently used in structural engineering recently.

Similar with the machine learning methods, the traditional statistic methods can also be used to solve some complex structural engineering problems. The process of statistical methods is very similar with machine learning methods. However, the statistical methods also have a couple of differences compared with machine learning approaches. The statistical methods will also be applied on the observed data first. Different from machine learning methods, different statistical methods will require to do different data transformation to get reasonable output ("Rule"). Then, these "Rules" are not developed in the "Black box". The statistical models will be developed to describe these "Rules". In the end, these "Rules" which are made of statistical models can also be used to do prediction. The statistical methods including multiple linear regression, LASSO and LARS were used in this thesis.

2.6 Summary

In this chapter, the bending behavior of concrete beams with simple support was briefly introduced. Then, the process of failure mechanism was detailly explained for the concrete beams with shear reinforcement and without shear reinforcement, respectively. One-way shear strength of concrete beam without shear rebars is contributed by concrete only while, for concrete beam with shear reinforcement, the shear strength is the contribution of both concrete and shear rebars. The database which was used in this thesis was also briefly introduced in this chapter. Finally, the reason why statistic and machine learning methods are valuable to be used to solve structural engineering problems were mentioned. The difference of basic concept between traditional methods and statistical/ML methods were explained in the end.

Chapter 3 Methodology

An overview of traditional and statistical/ML methods to predict shear strength and a comparison between the statistical/ML techniques and the traditional methods are given in this chapter. First, three traditional methods from structural engineering point of view are introduced. These are ACI318-19 (2019), and the methods proposed by Frosch et al. (2017) and Li et al. (2017). Then, three statistical methods were introduced including simple multiple regression, LASSO and LARS. Finally, the basic concepts of three machine learning methods were presented.

3.1 Traditional methods (Theoretical methods)

3.1.1 ACI 318-19

Based on the method given by ACI 318-19 (2019), for beams without shear reinforcement and slabs, the estimate of contribution of concrete is shown below:

$$\boldsymbol{V_c = [8\lambda_s \lambda \rho_w^{\frac{1}{3}} \sqrt{f'_c} + \frac{N_u}{6A_g}] b_w d} \qquad \textbf{Equation 3-1}$$

$$\boldsymbol{\lambda_s = \frac{1.4}{\sqrt{1+\frac{d}{10}}} \leq 1.0} \qquad \textbf{Equation 3-2}$$

where:

V_c= shear strength contributed by concrete

λ_s= size effect factor which is defined in Equation 3-2

λ = modification factor to reflect the reduced mechanical properties of light-weight concrete relative to normal-weight concrete of the same compressive strength. It is always 1.0 in this thesis since all the concretes in this thesis are normal weight concrete.

ρ_w= longitudinal reinforcement ratio

$\boldsymbol{f'_c}$= compressive strength of concrete

$\boldsymbol{N_u}$= Axial force in the beam. It is always 0 in this thesis since axial force in the beam is not considered.

$\boldsymbol{A_g}$= gross cross section area

$\boldsymbol{b_w}$= web width of beam

$\boldsymbol{d}$ = distance from extreme compression fiber to centroid of longitudinal tension steel.

For the beam with shear reinforcement, the shear strength contributed by concrete is provided in Equation 3-3. The first line is the traditional shear equation while the second line is Equation 3-1 without size effect factor, which is adopted in this thesis.

$$V_c = Either\ of \begin{cases} [2\lambda\sqrt{f'_c} + \frac{N_u}{6A_g}]b_w d] \\ 8\lambda\rho_w^{\frac{1}{3}}\sqrt{f'_c} + \frac{N_u}{6A_g}]b_w d \end{cases} \qquad \textbf{Equation 3-3}$$

The shear strength which is provided by shear reinforcement is given by.

$$V_s = A_v f_{yt}\left(\frac{d}{s}\right) \qquad \textbf{Equation 3-4}$$

where:

$\boldsymbol{V_s}$= shear strength contributed by shear reinforcement,

$\boldsymbol{A_v}$= area of shear reinforcement within spacings,

$\boldsymbol{f_{yt}}$= yield strength of shear reinforcement,

S= center-to-center space between shear rebars.

The total shear strength $\boldsymbol{V_n}$ of beam with shear reinforcement is given below.

$$V_n = V_c + V_s \qquad \textbf{Equation 3-5}$$

3.1.2 Method by Frosch et al. (2017)

Based on Frosch et al. (2017), for beams with or without shear reinforcement and slabs, the estimate of contribution of concrete is shown below:

$$V_c = (5\lambda\sqrt{f'_c}b_w c)\gamma_d \qquad \textbf{Equation 3-6}$$

$$\gamma_d = \frac{1.4}{\sqrt{1+\frac{d}{d_0}}} \leq 1.0 \qquad \textbf{Equation 3-7}$$

$$d_0 = \begin{cases} 10'' & if\ A_v < A_{vmin} \\ 100'' & if\ A_v \geq A_{vmin} \end{cases} \qquad \textbf{Equation 3-8}$$

where:

$\boldsymbol{c}$= distance from extreme compression fiber to neutral axis

$\boldsymbol{\gamma_d}$= size effect factor which is defined by Equation 3-7

$\boldsymbol{d_0}$**=** factor which is defined in Equation 3-8, for beam without shear reinforcement, it is equal to 10''. For beam with shear reinforcement, it is equal to 100''.

$\boldsymbol{A_{vmin}}$= minimum shear reinforcement which is required within space s

$\boldsymbol{V_c}$, b_w, and f_c are the same as defined before.

The shear strength which is provided by shear reinforcement is same with ACI318-19 method which is shown in Equation 3-4. The total shear strength $\boldsymbol{V_n}$ of beam with shear reinforcement was also same with ACI318-19 method which is shown in Equation 3-5.

3.1.3 Method by Li et al. (2017)

Based on Li et al. (2017), for beams with or without shear reinforcement and slabs, the estimate of contribution of concrete is shown below:

$$V_c = 17\lambda\left(\frac{V_u d}{M_u}\right)^{0.7}\sqrt{f'_c}\,b_w c\,\frac{1}{\sqrt{1+\frac{h}{11.8}}}\ . \qquad \textbf{Equation 3-9}$$

where:

$\boldsymbol{V_u}$= factored shear force at section

$\boldsymbol{M_u}$= factored moment at section

$\boldsymbol{h}$= depth of beam

a= shear span, distance from center of the concentrated load to center of support for simple support beam.

This equation in Equation 3-9 is very unstable because of the term of $\frac{\boldsymbol{V_u d}}{\boldsymbol{M_u}}$. Therefore, in this evaluation, it is replaced by d/a.

The shear strength which is provided by shear reinforcement is the same with Equation 3-4. The total shear strength $\boldsymbol{V_n}$ of beam with shear reinforcement is also the same with Equation 3-5.

3.2 Statistical methods

3.2.1 Multiple linear regression and Stepwise Variable Selection Methods

Multiple linear regression (MLR) is a statistical technique and attempts to model the relationship between two or more explanatory (predictors) variables and a response variable by fitting a linear equation to observed data. A multiple linear regression model with m predictor variables X_1, X_2, X_m and a response variable y as

$$\boldsymbol{y = \beta_0 + \sum_{j=1}^{m} X_j \beta_j + \varepsilon} \qquad \textbf{Equation 3-10}$$

If there are n observations on the m+1 predictor variables, and the MLR model can be written as

$$\boldsymbol{y_i = \beta_0 + \sum_{j=1}^{m} X_{ij} \beta_j + \varepsilon_i} \quad , \boldsymbol{i = 1, \dots, n}. \qquad \textbf{Equation 3-11}$$

where,

$\boldsymbol{y_i}$= dependent variables

$\boldsymbol{X_{ij}}$= explanatory variables

$\boldsymbol{\beta_0}$= y-intercept (constant term)

$\boldsymbol{\beta_j}$= slope coefficient for each explanatory variable

$\boldsymbol{\varepsilon_i}$= residuals with assumptions of 0 mean and σ^2variance

To estimate regression coefficients parameters, $\boldsymbol{\beta_0}$, $\boldsymbol{\beta_j}$s, least -squares method was employed which aimed at finding regression coefficients minimizing the objective function, residual sum of squares, given by

$$\boldsymbol{RSS(\beta) = \sum_{i=1}^{n}(y_i - \beta_0 + \sum_{j=1}^{m} X_{ij}\beta_j)^2} \qquad \textbf{Equation 3-12}$$

here:

The resulting fitted equation can be written as

$$\boldsymbol{\widehat{y_i} = \widehat{\beta_0} + \sum_{j=1}^{m} X_{ij}\widehat{\beta_j}\ , i = 1, \dots, n.} \qquad \textbf{Equation 3-13}$$

Three variable selection procedures were used to find the best model in this thesis including forward selection, backward selection, and stepwise selection. Forward selection begins with no candidate variables in the model. Then, select the variable that has the highest R-Squared. At each step, select the candidate variable that increases R-Squared the most. Stop adding variables when none of the remaining variables are significant. The backward selection starts with all candidate variables in the model. At each step, the variable that is the least significant is removed. This process continues until no nonsignificant variables remain. Stepwise regression is a combination of the forward and backward selection techniques. It is a modification of the forward selection so that after each step in which a variable was added, all candidate variables in the model are checked to see if their significance has been reduced below the specified tolerance level. If a nonsignificant variable is found, it is removed from the model (NCSS, 2021).

3.2.2 Least absolute shrinkage and selection operator (LASSO)

Least absolute shrinkage and selection operator (LASSO) is a regularized regression which can tackle the problem of overfitting by adding penalty term to the least-squares objective function to control complexity (Tibshirani, 1986). The purpose of this shrinkage is to prevent overfit arising due to either collinearity of the covariates or high dimensionality. LASSO is L1 penalized estimation technique that shrinks the estimates of the regression coefficients toward zero relative to least squares estimates. In LASSO penalized regression, the objective (cost) function is

$$\boldsymbol{J(\beta) = RSS(\beta) + \lambda \sum_{j=1}^{m} \left|\beta_j\right|}$$ **Equation 3-14**

When $\boldsymbol{\lambda}$ is 0 (i.e., no shrinkage), it will lead to the least squares estimates for all regression coefficients associated with predictors in a standard multiple regression. Intercept term is not regularized in this method. When λ increases, the bias increases.

The most common mode of selecting penalty parameters is using k-fold cross-validation (CV). Presence of outliers in data can exert undue influence over the coefficient estimates and the active set selection of a Lasso fit. However, outliers also can affect the amount of sparsity induced in the model. This influence is most notable in CV-selected penalty parameters, even in low-dimensional settings. Also, as multicollinearity becomes more serious, it also appears as though cross validation tends to select larger models in the presence of higher correlation between predictors (Kirtland, 2017).

3.2.3 Least angle regression (LARS)

Least angle regression (LARS) is an algorithm used in regression for high dimensional dataset. It is like forward stepwise regression. The general model is same with multiple regression model which is shown in Equation 3-10. The LARS algorithm is given below (EFRON, HASTIE, JOHNSTONE, & TIBSHIRANI, 2004):

1) Start with all $\beta_i = 0$

2) Find the predictor, X_{i1}, most correlated with **y**

3) Increase β_i from 0 by a series of small steps in the direction of the sign of its correlation with **y.** Along this way, compute residuals, $r = y - \hat{y}$. Stop when some other predictor, X_{j2}, has as much correlation with r as X_{i1} has.

4) Increase (β_i, β_j) in their joint least square direction until some other predictor, x_k, has as much correlation with the residual, r.

5) Continue until all predictors are in the model.

As penalty parameter values are generally unknown a priori, the LARS algorithm works very similarly like CART (Breiman, Friedman, Stone, & Olshen, 1984). Standard tree-building method: begin with a model with no predictors and gradually add in predictors until the model is "full". The path may then be "prunned" to an optimal size by setting the penalty parameter to an appropriate value.The LARS algorithm has also been shown to be stable with multicollinearity among the predictors (Hebiri & Lederer, 2013).

3.3 Machine learning method

3.3.1 Tree based method

Tree-based methods partition the feature space into a set of rectangles, and then fit a simple model (like a constant) in each one. They are conceptually simple yet powerful (Hastie, Tibshirani, & Friedman, 2017). There are two major types of decision trees in tree-based machine learning approaches. One is classification tree analysis which means the predicted results are the class to which the data belongs. The other one is regression tree analysis which means the predicted results are real number, for example, beam column joint moment capacity, roof displacement and so on. Figure 3-1 illustrates a simple decision tree model that includes a single binary target variable Y (0 or 1) and two continuous variables, x1 and x2, that range from 0 to 1. The main components of a decision tree model are nodes and branches and the most important steps in building a model are splitting, stopping, and pruning (SONG & LU, 2015). Classification and Regression Tree (CART) analysis covers both classification tree and regression tree which is initially proposed by Breiman et al. in 1984 (BREIMAN, FRIEDMAN, OLSHEN, & STONE, 1984). It is non-parametric decision tree. It produces regression or classification tree depends on dependent variable's type either numeric or categorical respectively. The CART method addresses the classification and regression problem by building a binary decision tree according to some splitting rule based on the predictor variables. In this way, the space of predictor variables is partitioned recursively in a binary fashion. The partition is intended to increase within-node homogeneity, where homogeneity is determined by the response variable in the problem. The partitioning is repeated until a node is reached for which no split improves the homogeneity, whereupon the splitting is stopped, and this node becomes a terminal node. Prediction is determined by terminal nodes and takes the form either of a class level in classification problems, or the average of the response variable in least squares regression problems (BREIMAN, FRIEDMAN, OLSHEN, & STONE, 1984).

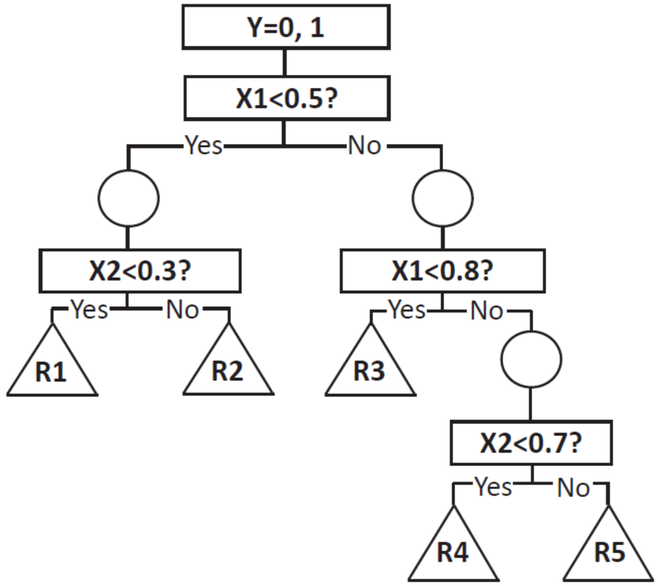

Figure 3-1: Sample decision tree based on binary target variable Y (SONG & LU, 2015)

The CART method only constructs one non-parametric decision tree which leads either a weak or an amazing prediction. However, there are a lot of more complicated decision tree methods which can construct more than one decision trees to make predictions be more accurate and reliable, for instance, random forest and boosting.

Random forest (Breiman L. , 2001) is an ensemble learning method of classification, regression and other tasks which construct decision tree at training time and predicts the class at output time. It grows many deep regression trees to randomized of the training data and average them which overcomes the over fitting problem of decision trees. Here “randomized” is a wide-ranging term and includes bootstrap sampling and/or subsampling of the observations, as well as subsampling of the variables. (Efron & Hastie, 2016)

Boosting is one of the most powerful learning ideas introduced in the last twenty years. It was originally designed for classification problems, but also was extended to regression as well. The motivation for boosting was a procedure that combines the outputs of many "weak" classifiers or predictions to produce a powerful "committee" (Hastie, Tibshirani, & Friedman, 2017). It repeatedly grows shallow trees to the residuals, and hence build up an additive model consisting of a sum of trees. The basic mechanism in random forests is variance reduction by averaging. Each deep tree has a high variance, and the averaging brings the variance down. In boosting the basic mechanism is bias reduction, although different flavors include some variance reduction as well. Both methods inherit all the good attributes of trees, most notable of which is variable selection (Efron & Hastie, 2016).

3.3.2 Neural Networks

In the mid-1980s, neural networks (NNs) were first introduced, and they marked a shift of predictive modeling towards computer science and machine learning. A neural network is a highly parametrized model, inspired by the architecture of the human brain, that was widely promoted as a universal approximator—a machine that with enough data could learn any smooth predictive relationship (Efron & Hastie, 2016). Figure 3-2 shows a simple example of a neural network diagram with a single hidden layer. Based on this neural network diagram, the neurons can be separated into three layers, including input layer, hidden layer and output layer. In addition, the hidden layer can contain one or more layers which is shown in figure 3-3.

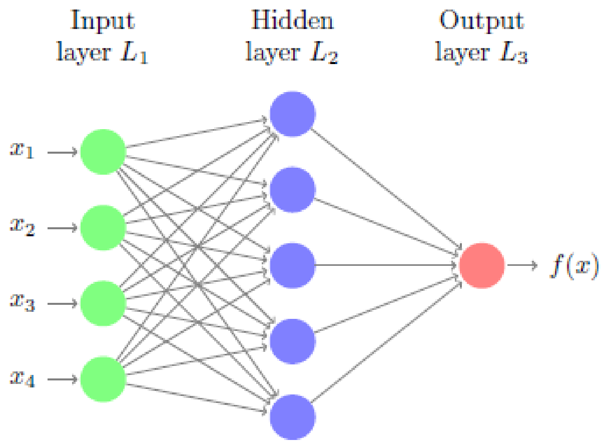

Figure 3-2: Neural network diagram with a single hidden layer (Efron & Hastie, 2016)

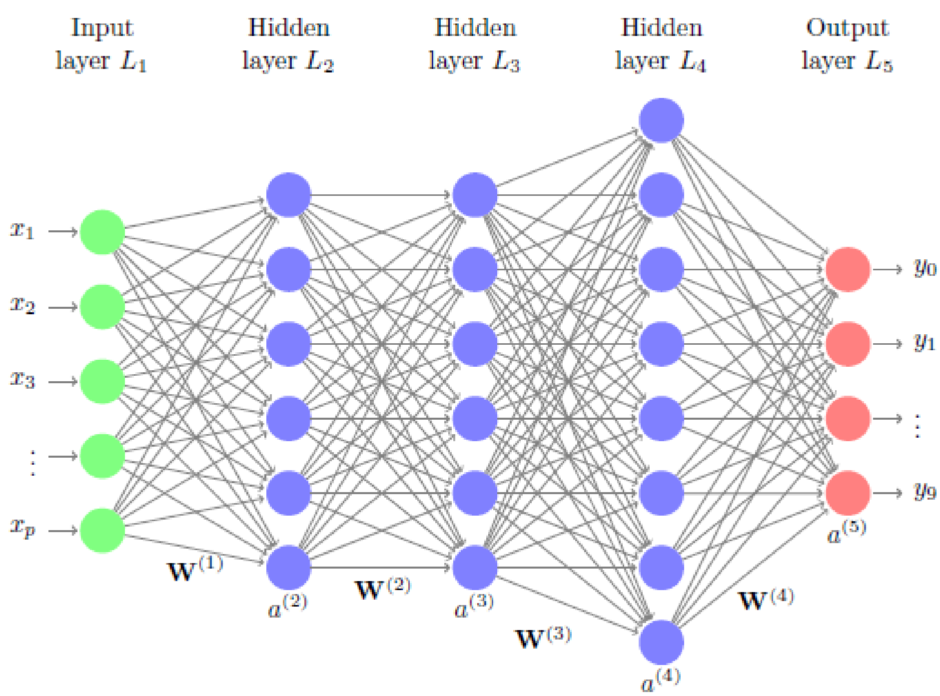

Figure 3-3: Neural network diagram with three hidden layers (Efron & Hastie, 2016)

Neural networks are biologically inspired systems consisting of a massively connected network of computational "neurons," organized in layers. By adjusting the weights of the network, NNs can be "trained" to approximate virtually any nonlinear function to a required degree of accuracy. NNs typically are provided with a set of input and output exemplars. A learning algorithm would then be used to adjust the weights in the network so that the network would give the desired output, in a type of learning commonly called supervised learning (ISHAK & TRIFIRÒ, 2007). There are lots of neural network training techniques can be used based on different approaches, including backpropagation, quick propagation, conjugate gradient descent, projection operator, Delta-Bar-Delta and so on. The back-propagation algorithm is one of the most frequently used methods for training a multilayer network with onward connections. In this method, random values are used to set initial weights and biases. The NNs is then processed for the entire set of input data and known outputs, measuring the error or difference between the target output and the computed output. This error is then propagated backwards to modify and update the weights and biases, again processing the network with the new values and obtaining a new error. This process is repeated until reaching a minimum error (or until a maximum number of iterations, or epochs, is reached). At this point, the weights and biases are fixed, and the NNs can be used to make predictions (Aguilar, Sandoval, Adam, Garzon-Roca, & Valdebenito, 2016). Other neural network training techniques can be performed as well to select a better solution.

3.3.3 Support vector machines

Support Vector Machines (SVMs) were first described by Boser, Guyon, & Vapnik (1992). It is a supervised machine learning algorithm which can be used for both classification and regression . However, it is mostly used in classification problems. The goal of SVMs algorithms is to find a separating hyperplane among two or more classes in a high-dimensional space. The individual data points are called vectors. The vectors closest to the separating hyperplane are called support vectors and they are the only ones which define the hyperplane. Support vectors are like swing state voters in a tight election: they are the only vectors whose vote "counts". Thus, if the support vectors are altered, deleted, or replaced, the hyperplane will be altered as well. A separating hyperplane is considered optimal when the distance from the support vectors to the plane, called the margin, is maximized (Swan, 2017). Figure 3-4 shows that all the lines (a, b, c, d) are potential separating hyperplanes while line b is the best one since it maximizes the distance between support vectors and decision boundary (line b).

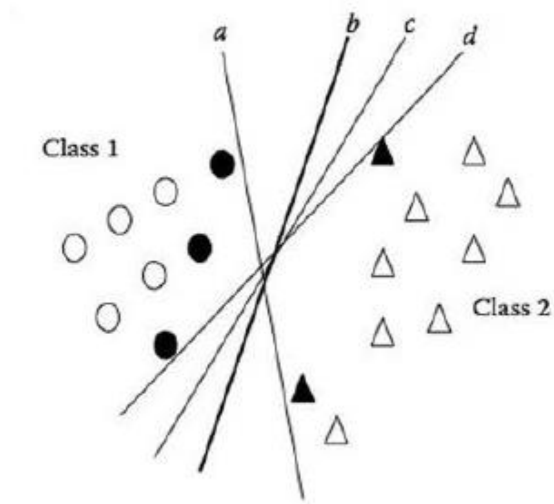

Figure 3-4: Margin maximization of SVMs (Swan, 2017)

Besides the straight hyperplanes, SVMs can learn a smooth, curved boundary by remapping the raw data into a higher dimensional space (Tso & Mather, 2009). There are three advantages to use SVMs. First, as a non-parametric method, SVMs make no assumptions about the distribution of their input data (Wilson, 2008). In the real world, most of data is not normally distributed while lots of methods were developed based on normal distribution assumption. Therefore, SVMs can handle datasets which are not normally distributed well. Second, SVMs have been proven to generalize well even with a small amount of training data (Foody & Mathur, 2004). This is a big advantage since collecting raw data is always expensive and time consuming and most of machine learning approach often requires a relatively big data set. Finally, SVMs is good at handling high dimensional dataset. It is also valuable since we might have lots of parameters in the structural analysis.

3.5 Summary

In this chapter, all the methods which were used in this thesis were introduced. The equations of three traditional methods which were developed by ACI318-19 (2019), Frosch et al. (2017) and Li et al. (2017) were presented first. The notation of all the variables in these equations were also included in the same section. Then, three statistical methods including multiple regression, LASSO and LARS were introduced. Finally, a brief introduction of basic concepts of three machine learning methods were presented in the end.

Chapter 4 Result and Discussion

4.1 Introduction

The result and discussion in this chapter were separated into two sections. One is the result and discussion of concrete beam without shear reinforcement and the other one is that of concrete beam with shear reinforcement. Each result and discussion will include the descriptive analysis, variable selection, comparison of accuracy among all the methods and the sensitivity analysis to evaluate the stability of all the methods.

4.2 Beam without shear reinforcement

4.2.1 Descriptive analysis

First, a descriptive analysis of dataset of beam without shear reinforcement was conducted. There are 13 predictors including beam width (b), beam web width (bw), beam depth (h), moment to shear ratio (kap) which is equivalent to $\frac{M_u}{V_u d}$, effective depth of beam (ds) which is distance from extreme compression fiber to centroid of longitudinal tension steel, average diameter of longitudinal rebars (dst), area of longitudinal rebars (As), geometrical reinforcement ratio (rhos) which is equivalent to $\frac{As}{b*ds}$, geometrical reinforcement ratio related to web width (rhosw) which is equivalent to $\frac{As}{bw*ds}$, yield strength of longitudinal steel (fsy), maximum diameter of aggregates (diaa), mechanical reinforcement ratios (oms) and compression strength of concrete (fc_prime). The one-way shear strength of concrete beam without shear reinforcement (Vu_Rep) is the dependent variable. The summary statistics of these variables are and the correlation matrix are given in Tables 4-1 and 4-2. The mean values of all the variables are closer to the minimum value compared with maximum value because most concrete beams for experiment are relatively small since they are less costly. The correlation coefficients between b and bw and h and ds are significant since they are geometrically related. As is highly related with b, bw, h, and ds since larger cross section usually contains more longitudinal reinforcement. The response variable, shear strength (Vu_Rep), is strongly related with b, bw, h, ds, As, as expected since they are also significant variables for predicting shear strength in the theoretical problems as well. However, rhosw and fc_prime are not highly correlated with shear strength which we expected.

Table 4-1: Summary Statistics of variables for beam without shear reinforcement

Variables	unit	Mean	Standard deviation	Maximum	Minimum
b	mm	241.7	211.5	3005.0	50.0
bw	mm	218.6	207.1	3005.0	50.0
h	mm	390.8	322.3	3140.0	76.2
kap	-	3.5	0.99	8.1	2.4
ds	mm	345.5	303.3	3000.0	57.2
dst	mm	20.4	7.2	40.0	6.0
As	mm^2	1410.5	1682.2	18252.0	56.5
rhos	%	1.9	1.05	6.64	0.14
rhosw	%	2.2	1.14	6.64	0.14
fsy	MPa	449.9	153.6	1779.3	174.5
fc_prime	MPa	36.9	21.1	136.6	10.5
diaa	mm	17.8	7.1	51.0	2.5
oms	-	0.25	0.1	0.95	0.03
Vu_Rep	kN	98.1	124.0	1308.4	7.2

Table 4-2: Correlation Matrix of all the variables for beam without shear reinforcement

	b	bw	h	kap	ds	dst	As	rhos	rhosw	fsy	fc_prime	diaa	oms	Vu_Rep
b	1	0.922	0.321	-0.035	0.323	0.198	0.628	-0.299	-0.137	0.028	0.02	-0.073	-0.289	0.804
bw	0.922	1	0.387	-0.104	0.388	0.219	0.678	-0.203	-0.292	0.095	0.075	0.028	-0.201	0.875
h	0.321	0.387	1	-0.139	0.999	0.417	0.701	-0.24	-0.331	0.097	-0.004	0.154	-0.205	0.639
kap	-0.035	-0.104	-0.139	1	-0.143	0.058	-0.048	0.148	0.239	0.022	-0.109	0.019	0.242	-0.158
ds	0.323	0.388	0.999	-0.143	1	0.405	0.692	-0.255	-0.344	0.099	-0.009	0.152	-0.216	0.637
dst	0.198	0.219	0.417	0.058	0.405	1	0.525	0.394	0.312	-0.24	0.082	0.142	0.241	0.382
As	0.628	0.678	0.701	-0.048	0.692	0.525	1	0.051	-0.024	-0.04	0.112	0.078	-0.043	0.888
rhos	-0.299	-0.203	-0.24	0.148	-0.255	0.394	0.051	1	0.765	-0.233	0.26	0.076	0.616	-0.116
rhosw	-0.137	-0.292	-0.331	0.239	-0.344	0.312	-0.024	0.765	1	-0.334	0.152	-0.123	0.42	-0.203
fsy	0.028	0.095	0.097	0.022	0.099	-0.24	-0.04	-0.233	-0.334	1	0.177	-0.048	-0.067	0.068
fc_prime	0.02	0.075	-0.004	-0.109	-0.009	0.082	0.112	0.26	0.152	0.177	1	-0.232	-0.391	0.166
diaa	-0.073	0.028	0.154	0.019	0.152	0.142	0.078	0.076	-0.123	-0.048	-0.232	1	0.217	0.042
oms	-0.289	-0.201	-0.205	0.242	-0.216	0.241	-0.043	0.616	0.42	-0.067	-0.391	0.217	1	-0.209
Vu_Rep	0.804	0.875	0.639	-0.158	0.637	0.382	0.888	-0.116	-0.203	0.068	0.166	0.042	-0.209	1

4.2.2 Variable selection and Estimation

There are 13 explanatory variables in the dataset which might not be needed since the sample size is not large, only 784 samples. What's more, based on the correlation matrix in Table 4-2, there are some explanatory variables are not significantly related with shear strength. Therefore, the variable selection and estimation was conducted by using multiple linear regression, LASSO, LARS and random forest method first in this thesis. Tables 4-3, 4-4 and 4-5 show the results of forward, backward, and stepwise methods, respectively. Figures 4-1, 4-2 and 4-3 show the variable selection result from LASSO, LARS and random forest, respectively.

Table 4-3: Results of variable selection by forward method from MLR for beam without shear reinforcement

Selection Summary

Step	Variable Entered	R-Square	Adj. R-Square	C(p)	AIC	RMSE
1	bw	0.7584	0.7580	1562.5170	5953.8238	55.1323
2	As	0.9167	0.9164	184.0869	5372.1373	32.3978
3	ds	0.9257	0.9253	108.0103	5311.7951	30.6346
4	fc_prime	0.9320	0.9315	55.1153	5265.2873	29.3352
5	kap	0.9350	0.9344	30.5082	5242.1357	28.6961
6	h	0.9370	0.9363	15.3166	5227.2482	28.2835
7	b	0.9377	0.9369	11.6037	5223.5192	28.1621
8	diaa	0.9380	0.9371	10.5464	5222.4283	28.1088
9	fsy	0.9382	0.9372	10.5975	5222.4490	28.0841
10	rhosw	0.9385	0.9373	10.6954	5222.5103	28.0606
11	oms	0.9386	0.9374	11.0811	5222.8596	28.0445

Table 4-4: Results of variable selection by backward method from MLR for beam without shear reinforcement

Elimination Summary

Step	Variable Removed	R-Square	Adj. R-Square	C(p)	AIC	RMSE
1	b	0.9388	0.9374	12.0039	5223.7552	28.0424
2	oms	0.9387	0.9374	10.6377	5222.4052	28.0329
3	dst	0.9386	0.9374	9.5373	5221.3266	28.0303

Table 4-5: Result of variable selection by stepwise method from MLR for beam without shear reinforcement

Stepwise Selection Summary

Step	Variable	Added/ Removed	R-Square	Adj. R-Square	C(p)	AIC	RMSE
1	bw	addition	0.758	0.758	1562.5170	5953.8238	55.1323
2	As	addition	0.917	0.916	184.0870	5372.1373	32.3978
3	ds	addition	0.926	0.925	108.0100	5311.7951	30.6346
4	fc_prime	addition	0.932	0.931	55.1150	5265.2873	29.3352
5	kap	addition	0.935	0.934	30.5080	5242.1357	28.6961
6	h	addition	0.937	0.936	15.3170	5227.2482	28.2835
7	b	addition	0.938	0.937	11.6040	5223.5192	28.1621
8	diaa	addition	0.938	0.937	10.5460	5222.4283	28.1088

```
14 x 1 sparse Matrix of class "dgCMatrix"
                        1
(Intercept) -26.17476230
b             .
bw            0.31153919
h             .
kap          -0.64693573
ds            0.03379833
dst           .
As            0.03113260
rhos          .
rhosw         .
fsy           .
fc_prime      0.08739116
diaa          .
oms           .
```

Figure 4-1: Result of variable selection from LASSO for beam without shear reinforcement

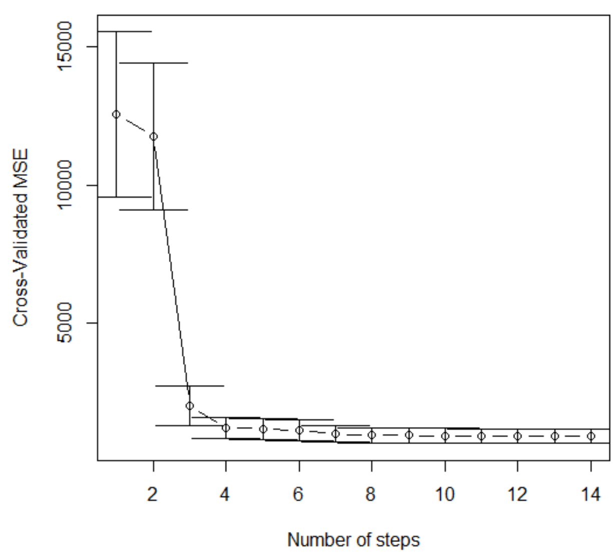

Figure 4-2: Result of variable selection from LARS for beam without shear reinforcement (Cross validated MSE as number of steps increases)

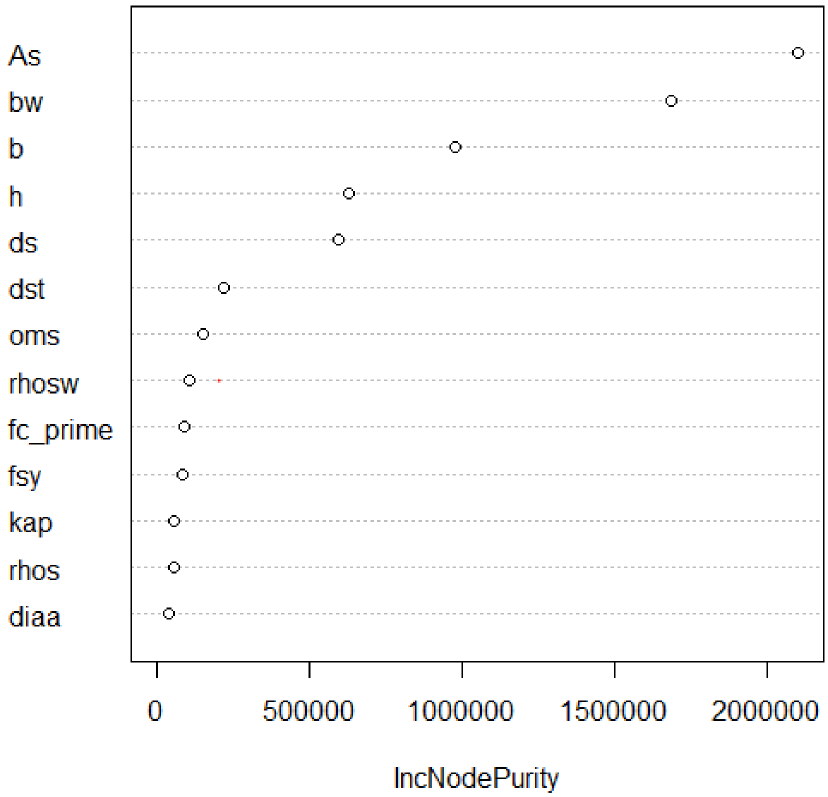

Figure 4-3: Variable Importance Plot generated by random Forest for beam without shear reinforcement

The results of variable selection from statistical methods (MLR, LASSO, LARS) were similar with each other. Bw, As and ds are very significant to keep. Fc_prime and kap are also significant but not as much as previous three variables. The remaining variables are not significant at all. The results of variable selection from statistical methods are exactly same with the variables in the traditional methods. For the result of random forest, bw, As, ds are same with the results from statistical methods which are highly statistically significant. However, kap and fc_prime is not significant in the random forest. Based on the results of variable selection and knowledge of subject matter, bw, As, ds, kap and fc_prime were decided to keep in the following study.

4.2.3 Accuracy

A scatter plot matrix was created in Figure 4-4 to visually show the relationship between all the variables. Based on Figure 4-4, bw, As and ds have a very strong positive correlation with shear strength while this trend is not significant for kap and fc_prime (fp). The KDD densities show that distributions of bw and ds were skewed right which means that the dataset at hand does not contain many big concrete beams.

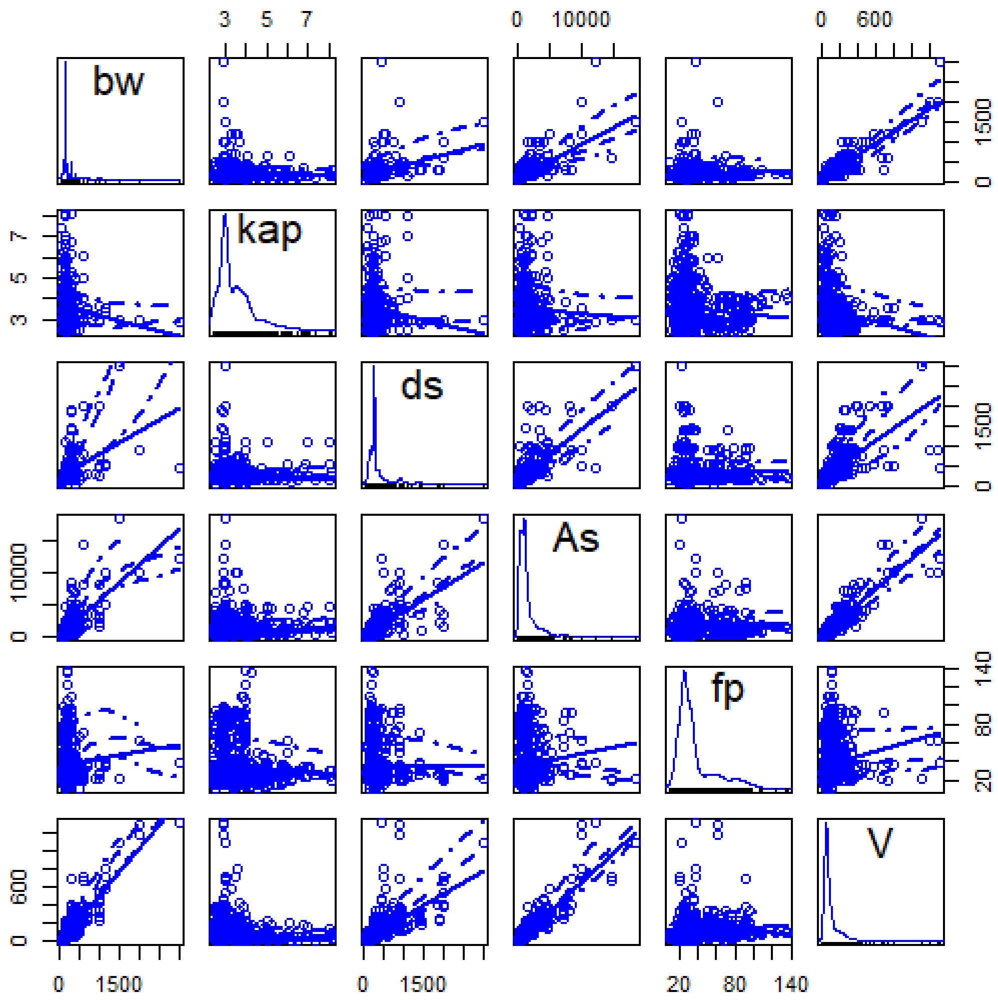

Figure 4-4: Scatter Plot Matrix of selected variables for beam without shear reinforcement

For theoretical methods, the equations to calculate the shear strength of beam without shear reinforcement were provided in chapter 3. One thing needs to be noticed here is that the theoretical methods usually underestimate the shear strength for safety purpose. Figure 4-5 shows the comparison between experimental shear strength and predicted shear strength using three theoretical methods. If the predicted values are exactly same with the experimental values, all the observations should be on the solid line. The dashed lines in Figure 4-5 correspond to 20% error lines. It was clear that an embedded factor was applied to these theoretical methods to make sure that the most of predicted values were less than the experimental values. Therefore, a correction factor was added in the calculation. These correction factors were calculated by using simple regression to shift the results to have best fit with accuracy line. They are 1.23 for ACI318-19, 1.49 for Frosch et al. and 1.31 for Li et al.

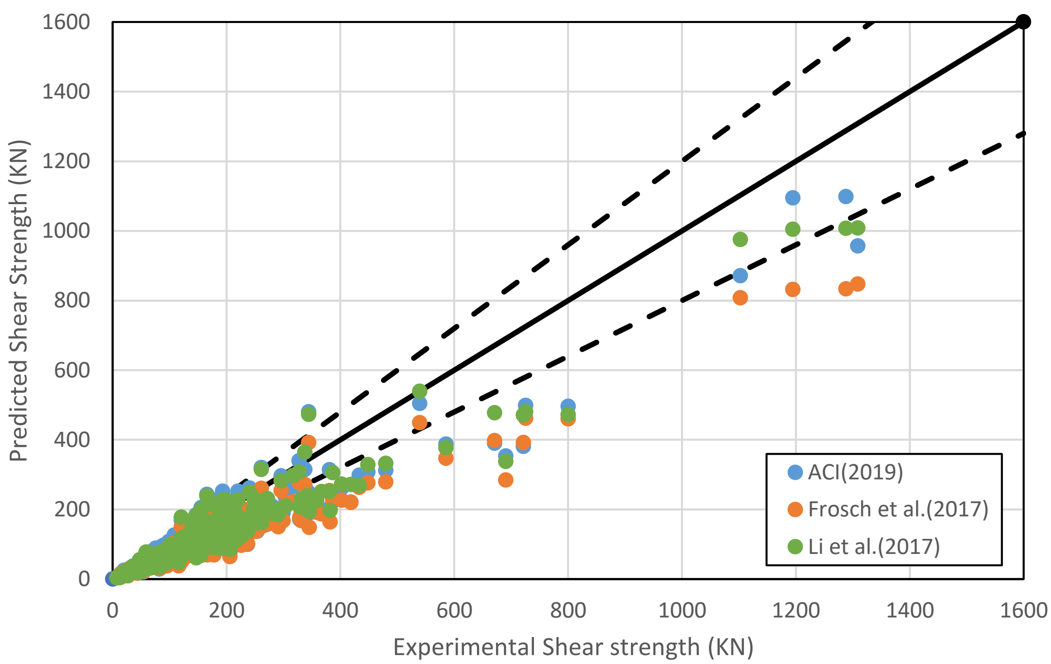

Figure 4-5: Comparison between experiment and uncorrected prediction by theoretical methods for beam without shear reinforcement

Figure 4-6 shows the comparison of corrected predicted shear strength and experimental shear strength by using theoretical methods.

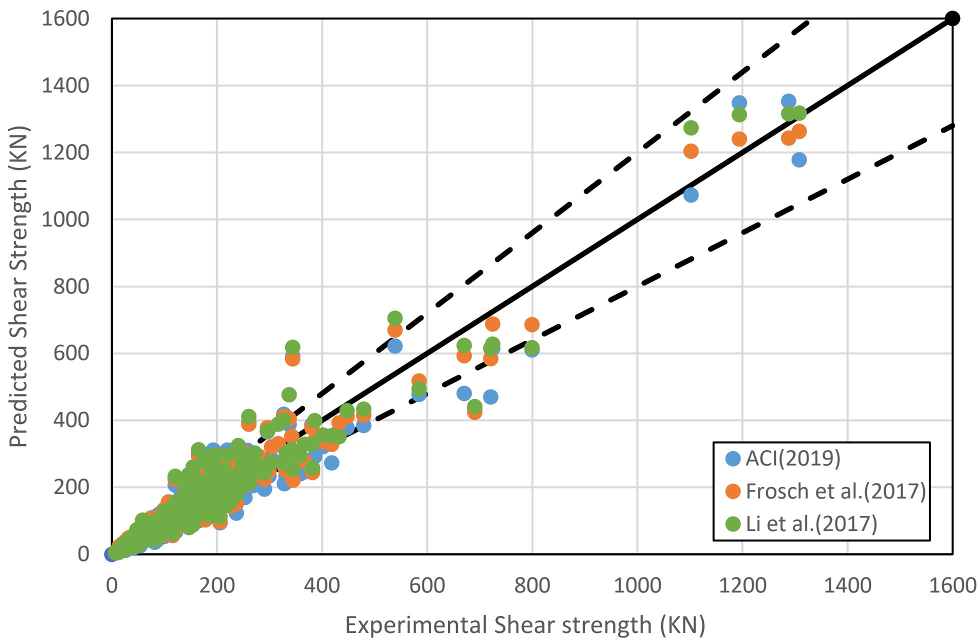

Figure 4-6: Comparison between experiment and corrected prediction by theoretical methods for beam without shear reinforcement

Statistical models were achieved by using MLR, LASSO and LARS which were shown in Equations 4-1, 4-2 and 4-3, respectively.

For MLR method after use of variable selection methods, the final model is the following:

$$\boldsymbol{Vu\_Rep = -23.69 + 0.31bw - 7.11kap + 0.051ds + 0.032As + 0.431fc\_prime} \quad \textbf{Equation 4-1}$$

For LASSO method,

$$\boldsymbol{Vu\_Rep = -21.6 + 0.29bw - 1.53kap + 0.034ds + 0.032As + 0.121fc\_prime} \quad \textbf{Equation 4-2}$$

For LARS method,

$$\boldsymbol{Vu\_Rep = -23.69 + 0.31bw - 7.11kap + 0.051ds + 0.032As + 0.431fc\_prime} \quad \textbf{Equation 4-3}$$

The MLR model and LARS model were exactly same which make sense. That's because the way how it works. The first step of LARS method is to identify the predictor variable which is most correlated with the dependent variable. Instead of fitting this predictor completely, LARS moves the coefficient of this predictor continuously towards its least squares value. Then the second predictor will join the procedure. The process continues till all the predictors are in the model and ends at the full least-squares fit. Least angle regression only enters as much of a predictor as it deserves. However, in this section, variable selection has been done before. All the variables in this reduced dataset were significant. Therefore, the LARS method will have full least squares fit like MLR. The result of LARS will not be shown in this chapter. Figure 4-7 shows the comparison of predicted shear strength and experimental shear strength by using statistical methods. The accuracy of prediction is decent. Most of predictions were within the 20% error.

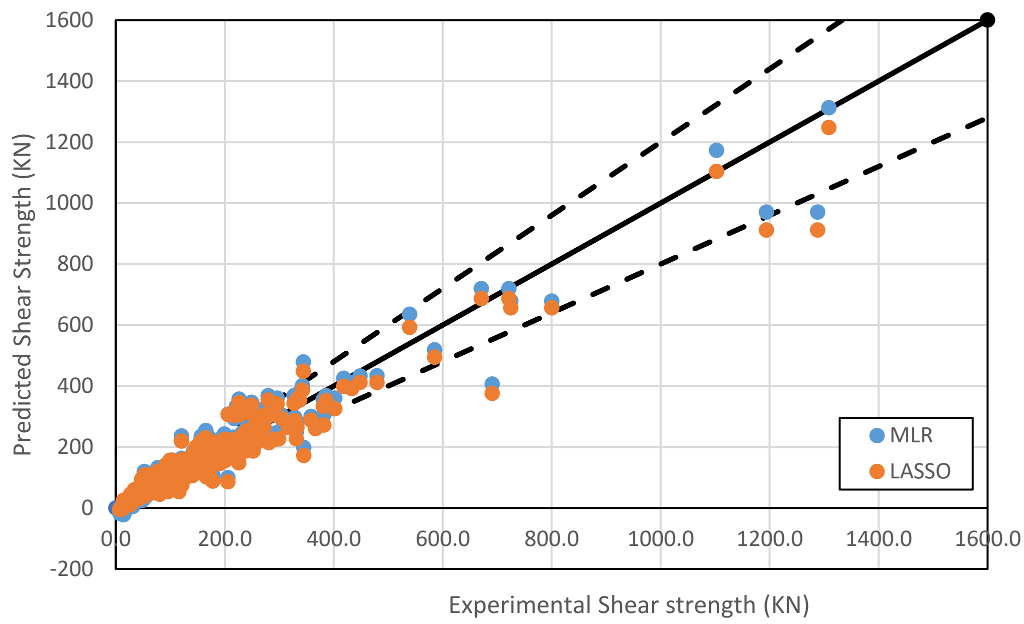

Figure 4-7: Comparison between experiment and prediction by statistical methods for beam without shear reinforcement

Then the machine learning methods were applied on the reduced dataset as well. The comparison of predicted shear strength and experimental shear strength by using machine learning methods was shown in Figure 4-8. 500 trees (ntree=500) were selected in the random forest method while 5 hidden layers (hidden=5) was chosen for the NN method. The accuracies of prediction of random forest and NN are also decent. Most of the predictions fall with 20% error line. For support vector method, it performs well for the smaller beams (shear strength is relatively small) while it severely underestimates shear strength for the larger beams.

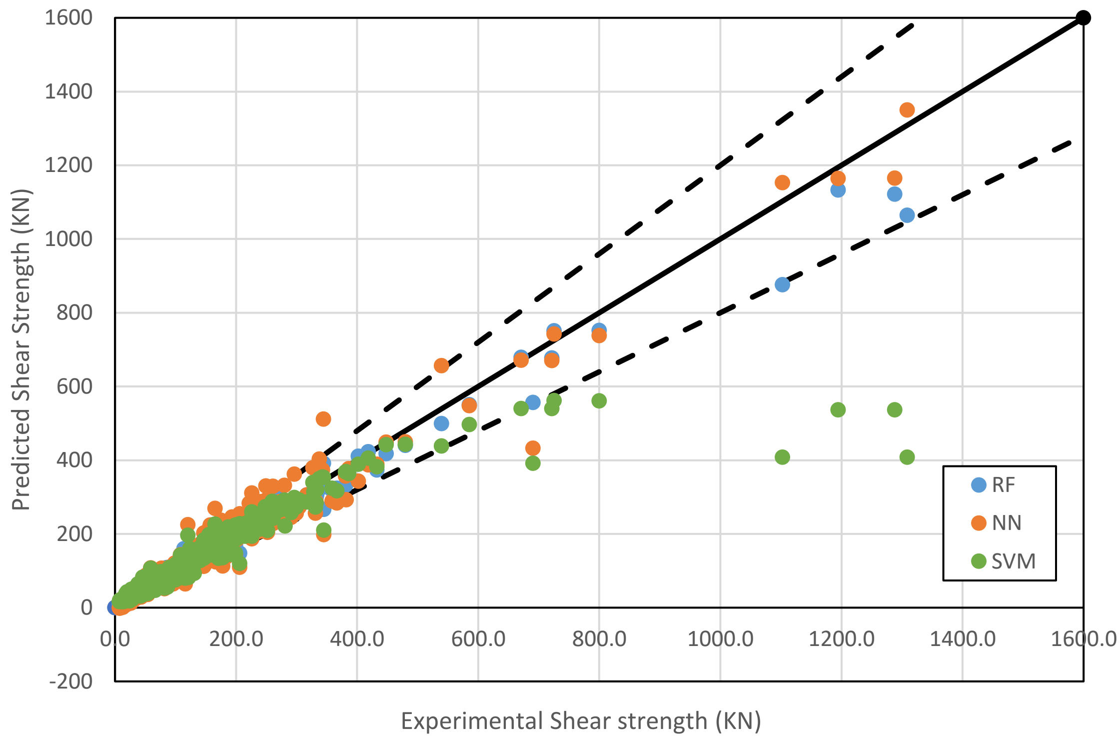

Figure 4-8: Comparison between experiment and prediction by machine learning methods for beam without shear reinforcement

To reduce the bias and uncertainty of statistical/ML methods, in other words, to measure the quality of our models, one of the widely used method is employed to validate the models which is the 10-fold cross validation for the assessment of the quality of the predicted models obtained based on different methods. The dataset will be partitioned into 10 folds. 9 out of 10 folds will be used as a training dataset while the remaining fold will be used as a test dataset. The statistical/ML methods will be applied to the training dataset first. Then, the test dataset will be used to see the goodness of fit for each method by using the criteria called root of mean square error (RMSE), which a metric to summarizes predicted model quality. This process will be repeated by 10 times since each fold will be used as test dataset once. It is unnecessary to apply the 10-fold cross validation for the theoretical methods since the equations for theoretical methods are fixed which will not change based on different training dataset. Table 4-6 shows the results for the 10-fold cross validation of statistical/ML methods compared with theoretical methods. The results show that all statistical/ML methods perform well except SVM. As known before, SVM method is not successful for predicting the big beams since the dataset contains a small set of large beam samples.

Table 4-6: Results of goodness of fit for all the methods

Group	Method	RMSE
Theoretical	ACI318-19 (2019)	36
	Frosh et al. (2017)	27
	Li et al. (2017)	31
Statistical	MLR	31
	LASSO	34
Machine Learning	RF	40
	NN	32
	SVM	75

4.2.4 Sensitivity Analysis

The sensitivity analysis was applied on all these methods except LARS regression since the LARS based model is exactly the same as multiple linear regression-based model. The purpose of sensitivity analysis is to check the *stability of the performance* of all these methods when one predictor variable changes. In other words, sensitivity analysis determines how different values of a predictor variable affects a particular dependent variable under a given set of assumptions. New test datasets were created based on the experimental dataset. The range of five selected predictor variables was determined based on their maximum and minimum values in Table 4-1. Table 4-7 shows the range and median value of each variable used in test dataset. The longitudinal reinforcement ratio (low) which is ρ_w in the theoretical methods can be calculated by As, bw and ds, $\rho_w =$

$\frac{A_s}{bw*ds}$. Therefore, only low was listed in the table, the corresponding As will be calculated in the analysis. If the sensitivity analysis is applied for bw, the bw in the test dataset will increase from 50mm to 3050mm while the other variables will remain median value. The baseline shear strength will be calculated as well when all the variables are equal to their median value. Then, the predicted shear strength will be divided by the baseline shear strength which is shear strength ratio. Finally, a shear strength ratio vs. bw plot can be generated to see the trend of shear strength ratio as the widths of concrete beams increase among all different methods. The same procedure will be repeated by five times for each explanatory variable. Figures 4-9, 4-10, 4-11, 4-12 and 4-13 show the sensitivity analysis for bw, fc_prime, kap, low and ds, respectively. The trend of theoretical method can give a reference to compare with the statistical/ML methods since the theoretical methods were published with peer review.

Table 4-7: The details of test dataset for beam without shear reinforcement

Variables	unit	median	Maximum	Minimum
bw	mm	1550.0	3050.0	50.0
Kap	-	5.0	8.0	2.4
ds	mm	1550.0	3050.0	50.0
low	-	0.0286	0.0542	0.0030
fc_prime	Mpa	50.0	80.0	20.0

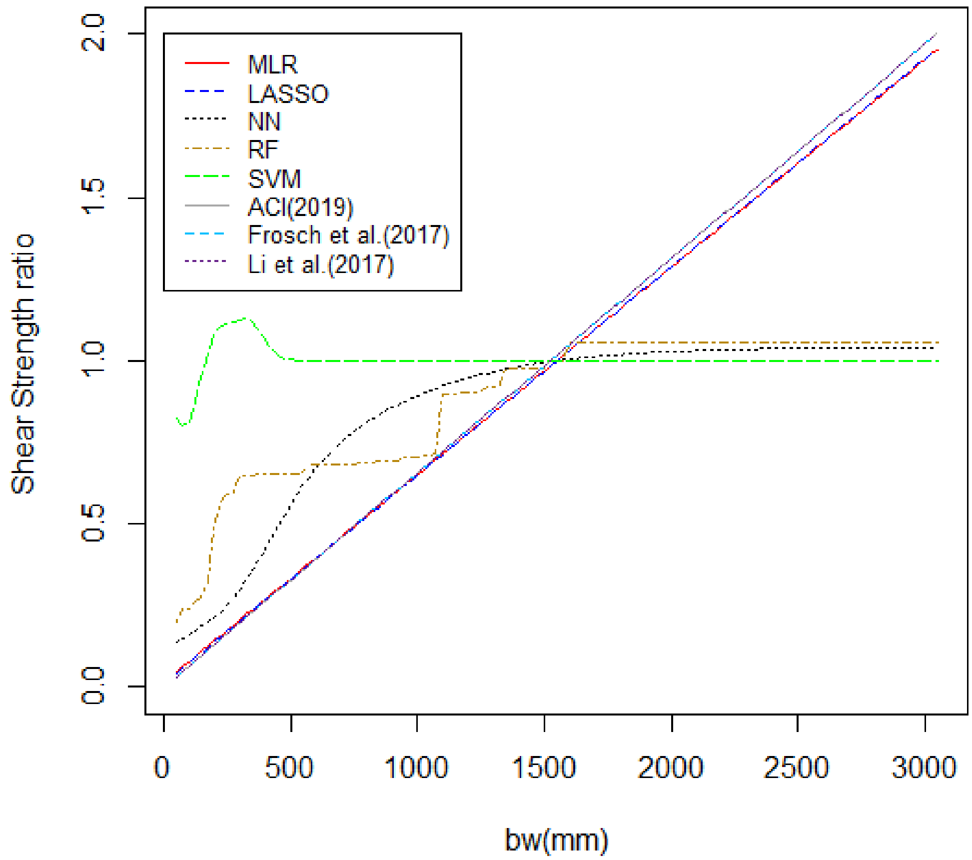

Figure 4-9: Sensitivity analysis on bw for concrete beam without shear reinforcement

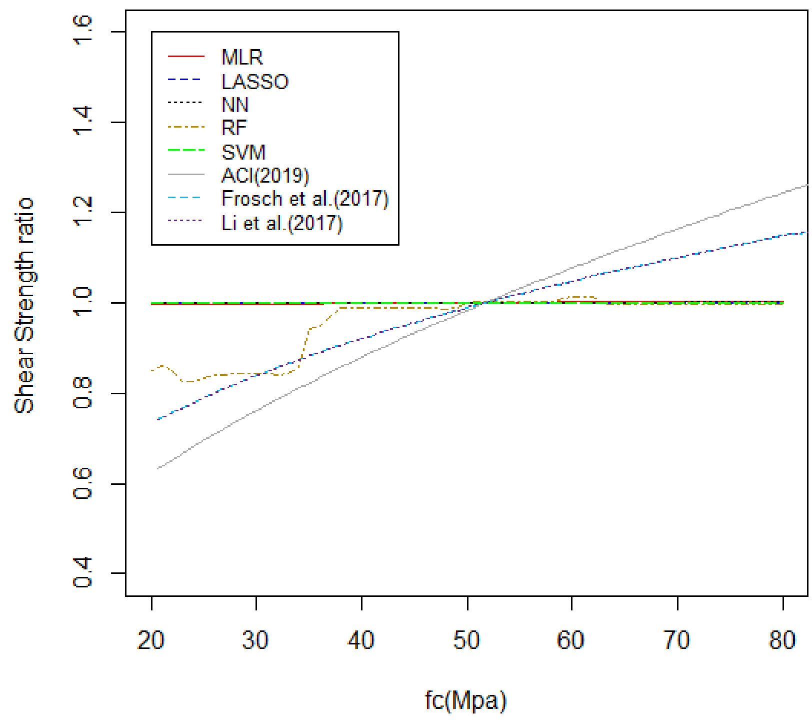

Figure 4-10: Sensitivity analysis on fc_prime for concrete beam without shear reinforcement

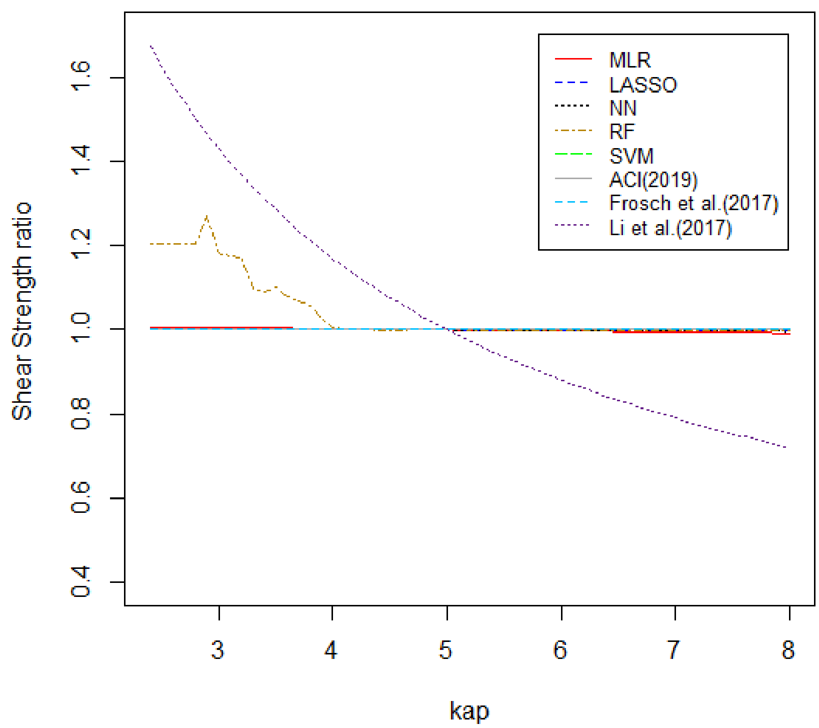

Figure 4-11: Sensitivity analysis on kap for concrete beam without shear reinforcement

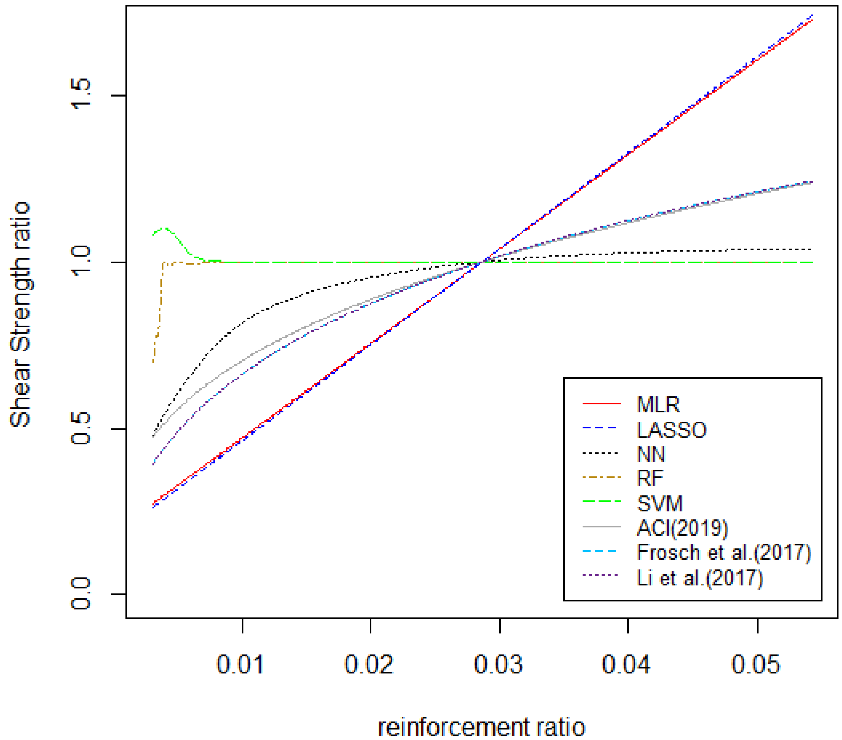

Figure 4-12: Sensitivity analysis on low for concrete beam without shear reinforcement

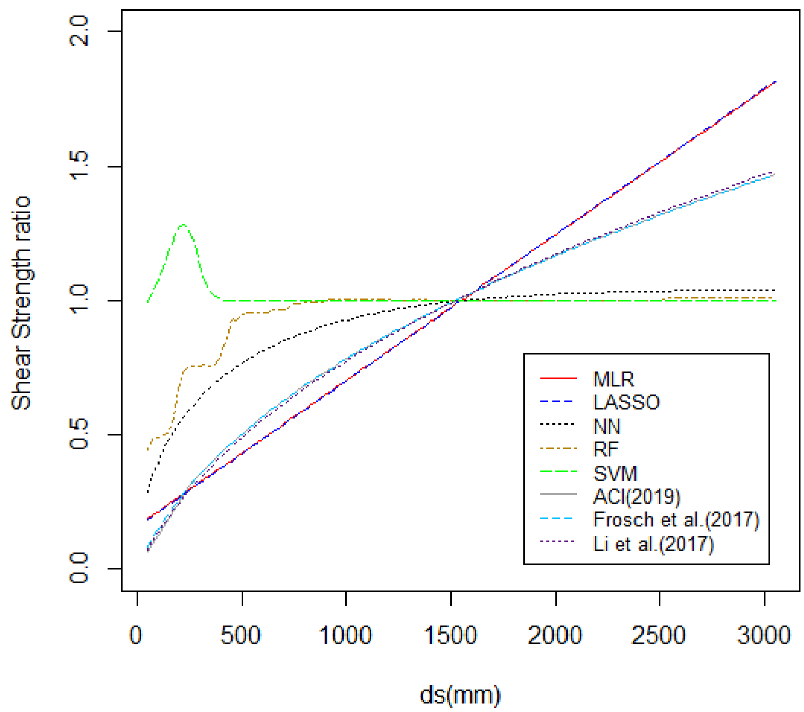

Figure 4-13: Sensitivity analysis on ds for concrete beam without shear reinforcement

For sensitivity analysis on bw (Figure 4-9), the trends of shear strength ratio from theoretical methods and statistical methods are very similar to each other. The machine learning methods only have an increasing trend at very beginning.

For sensitivity analysis on fc_prime (figure 4-10), none of the statistical/ML methods have a similar trend with theoretical methods.

For sensitivity analysis on kap (Figure 4-11), ACI and Frosch methods don't have any trend since kap is not considered at all. MLR, LASSO, NN and SVM didn't show any trend along with kap as well. Random forest has a slightly similar trend with Li et al. method, but the trend diminished when kap became larger.

For sensitivity analysis on low (Figure 4-12), statistical methods have an increasing trend while the machine learning methods are not in agreement with theoretical methods at all.

For sensitivity analysis on ds (Figure 4-13), the statistical methods resulted in similar trends as the theoretical methods. Neural network and random forest methods are in agreement to each other, but not to the theoretical methods when ds measurements are less than the median value. SVM method cannot reflect the change of shear strength along with ds at all.

Overall, statistical methods can capture the increasing or decreasing trend, but it cannot capture any curved trend since both of two methods are based on linear regression. The machine learning methods do not perform well, especially when the predictor variables become larger and larger. That's because there are more small beams than the large beams in the experimental dataset. After the dataset was trained by different methods, the results are good at predicting the small concrete beams. In the test dataset, the shear strength of baseline concrete beam might not be predicted properly. Therefore, a smaller test dataset was established based on the boxplot of each variable. Figure 4-14 shows a typical boxplot. The range of each variable in the smaller test dataset will be between Minimum and Maximum in their boxplot. Table 4-8 shows the range and median value of each variable which were used in small test dataset.

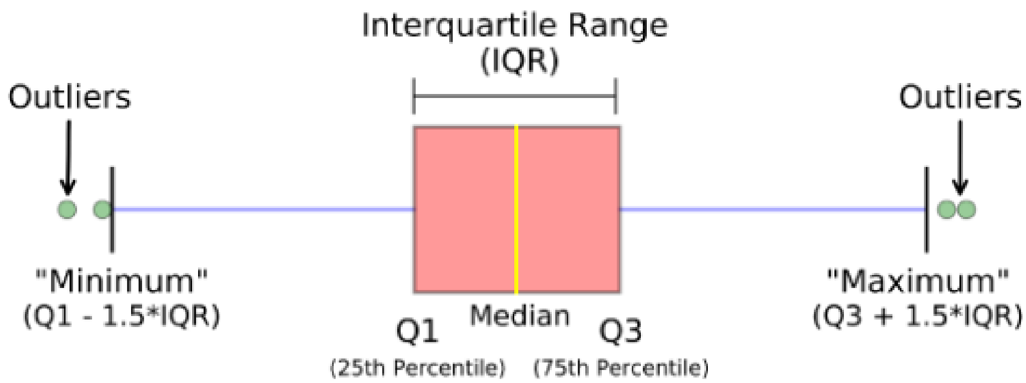

Figure 4-14: Typical boxplot (Galarnyk, 2018)

Table 4-8: Ranges for small test dataset for beam without shear reinforcement

Variables	unit	median	Maximum	Minimum
bw	mm	205.0	360.0	50.0
Kap	-	4.2	6.0	2.4
ds	mm	325.0	600.0	50.0
low	-	0.0255	0.0500	0.0010
fc_prime	Mpa	40.0	70.0	10.0

The same process of sensitivity analysis was applied for each predictor variable given in the Table 4-8 by using small test dataset which contained the ranges of predictors as depicted in table 4-8. Figures 4-15, 4-16, 4-17, 4-18 and 4-19 show the sensitivity analysis for bw, fc_prime, kap, low and ds by using small test dataset, respectively.

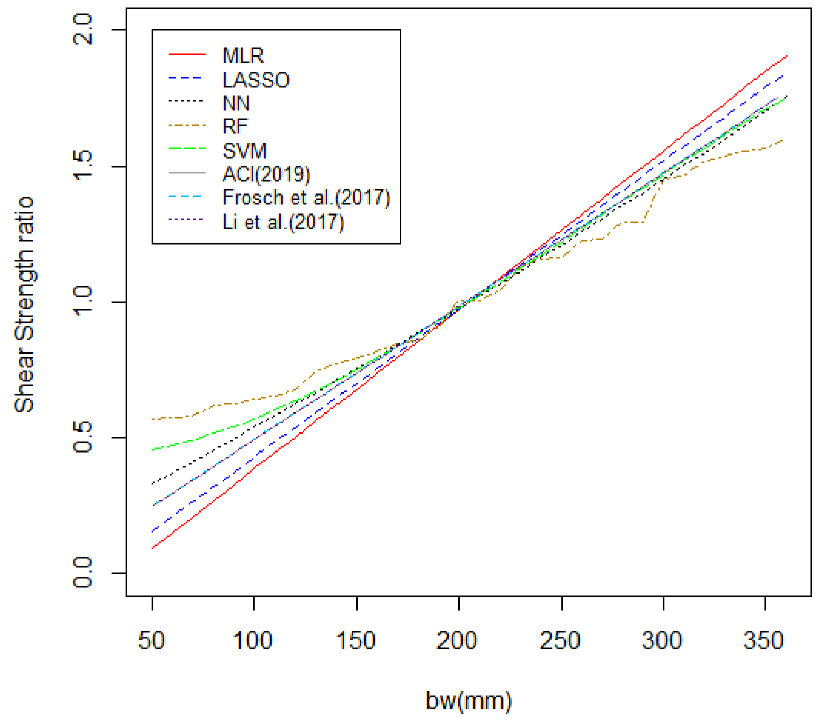

Figure 4-15: Sensitivity analysis on bw for concrete beam without shear reinforcement in small dataset

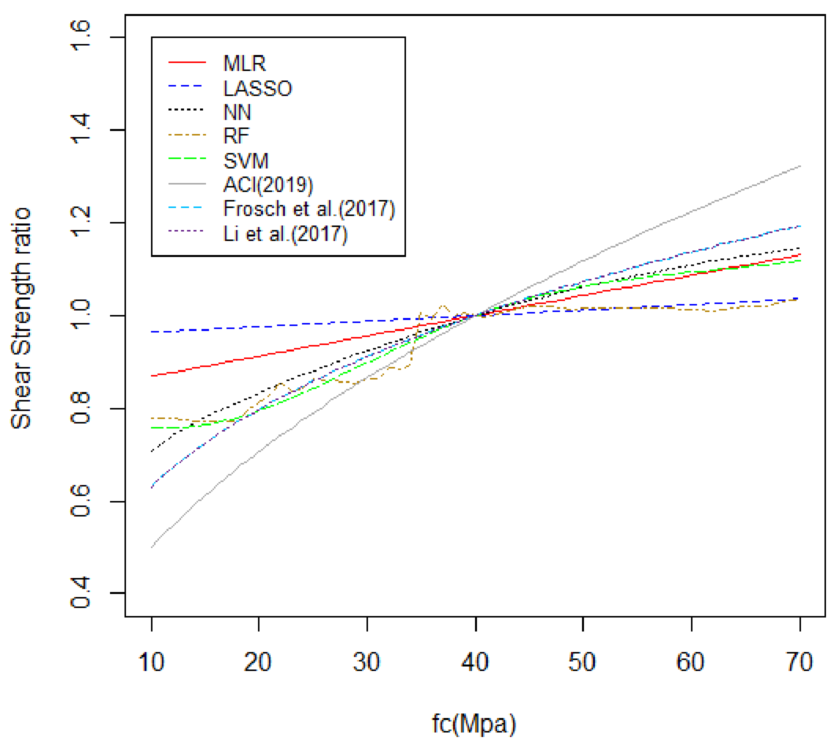

Figure 4-16: Sensitivity analysis on fc_prime for concrete beam without shear reinforcement in small dataset

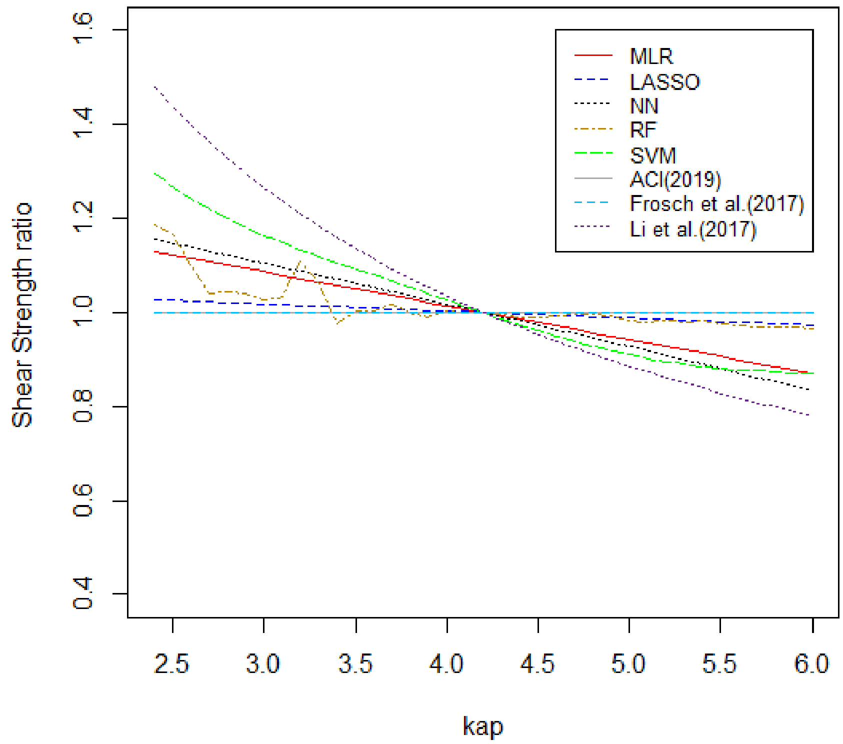

Figure 4-17: Sensitivity analysis on kap for concrete beam without shear reinforcement in small dataset

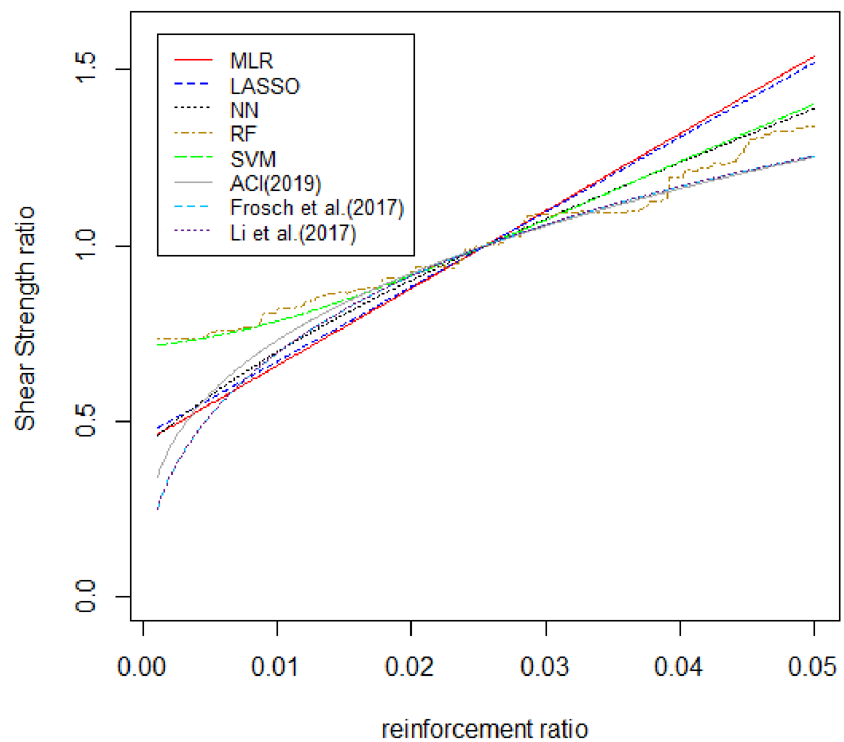

Figure 4-18: Sensitivity analysis on low for concrete beam without shear reinforcement in small dataset

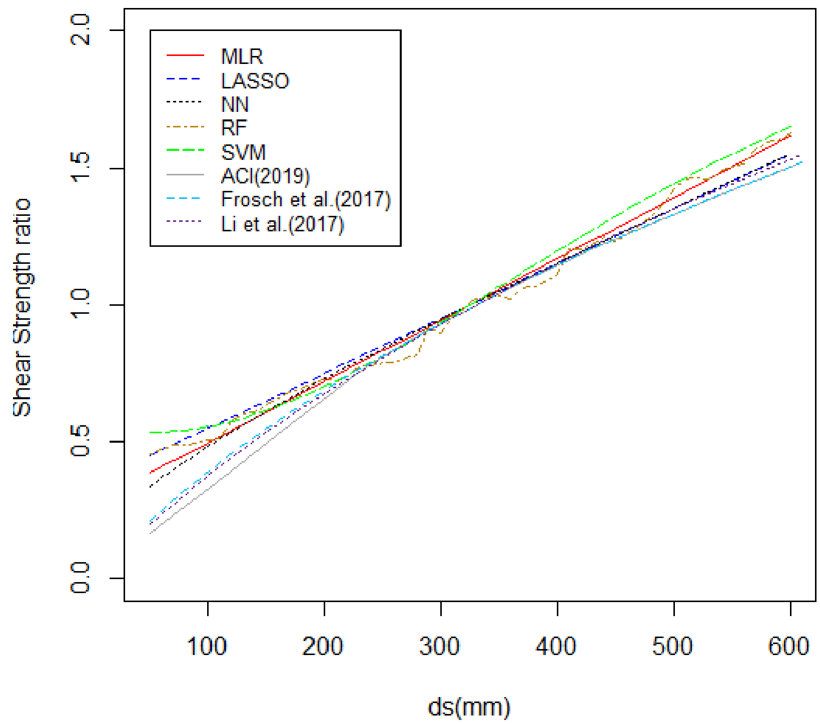

Figure 4-19: Sensitivity analysis on ds for concrete beam without shear reinforcement in small dataset

For sensitivity analysis on bw (Figure 4-15), the trend of shear strength ratio from theoretical methods and statistical methods are very similar. The machine learning methods have also similar trend. The shear strength ratio-based sensitivity analysis from machine learning methods gives some different result compared with those from theoretical methods when bw is small. The results of NN methods is closest and most smooth among all the machine learning methods. (What does this mean?? Make it clear!!!The result of random forest is not very stable.

For sensitivity analysis on fc_prime (Figure 4-16), the result of LASSO regression is worst since it didn’t capture the changes along with fc_prime. MLR and random forest performs somewhat fine but not very accurate. The random forest method- based model is not stable, which is like the previous result. The SVM and NN gives very similar result as Frosh and Li's methods, especially NN method. ACI method is a little bit different from the other theoretical methods.

For sensitivity analysis on kap (Figure 4-17), ACI and Frosch methods don't have any trend since kap is not considered at all. LASSO only reflects a small change of shear strength ratio along with kap. The other methods have similar trends as Li et al.'s method but not as significant as Li's method. SVM has the most similar trend to Li's method. The result of random forest was observed to be not stable again.

For sensitivity analysis on low (Figure 4-18), two statistical methods provide very similar trends and closer to theoretical methods than random forest and SVM when reinforcement ratio is smaller than median value. Random forest and SVM are more accurate when reinforcement ratio is larger than median value. NN performs well all the time.

For sensitivity analysis on ds (Figure 4-19), all the methods are similar to each other. Similar with Previous result, random forest is not very stable. NN method produces the result the closest to theoretical methods.

Overall, the statistical/ML methods performed better in the small test dataset than big test dataset. MLR, NN and SVM are more accurate compared with LASSO and random forest based on sensitivity analysis. NN method has the best performance among all these statistical/ML method. All the statistical/ML methods are closer to Frosh and Li et al's methods when they have disagreement with ACI method on sensitivity analysis with fc_prime.

4.3 Beam with shear reinforcement

4.3.1 Descriptive analysis

First, a descriptive analysis of dataset of beam with shear reinforcement was conducted. There are two datasets established by researchers including a large dataset with 170 samples and a small dataset with 87 samples which was used in this thesis. The small dataset is a subset of large dataset where there is certainty that the shear rebars yielded when the concrete failed. All the explanatory variables except diaa and oms which are used in the concrete beam without shear reinforcement is kept in the analysis of concrete beam with shear reinforcement. Besides these explanatory variables, another three variables which are related with shear reinforcement were added in the dataset. They are yield strength of shear reinforcement (fyw), center-to-center space between shear rebars (sw) and area of shear reinforcement with spacing sw. Same with concrete beam without shear reinforcement, the one-way shear strength of concrete beam without shear reinforcement (Vu_Rep) is the dependent variable. The summary statistics of these variables and the correlation matrix are shown in Table 4-9 and 4-10. The contribution of concrete and the contribution of shear reinforcement is considered as explanatory with each other. All three added variables have positive correlation with shear strength of concrete beam which is a little bit unusual. The purpose of adding shear reinforcement is to increase the shear strength of concrete beam. Therefore, increase the area and yield strength of shear rebars will increase the shear strength. However, increase the space of shear rebars which is equivalent to reduce the density of shear reinforcement should decrease the shear strength. A negative correlation is expected between sw and Vu_Rep. The parameter of sw needs to be concerned in the statistical model later. In addition, the correlation between b and bw is very small which is usual as well since they have high positive correlation in the concrete beam without shear reinforcement dataset before.

Table 4-9: Summary Statistics of variables for beam with shear reinforcement

Variables	unit	Mean	Standard deviation	Maximum	Minimum
b	mm	470.4	280.7	1500.0	125.0
bw	mm	193.3	90.3	457.2	50.0
h	mm	531.9	260.7	1250.0	250.0
kap	-	3.3	0.7	7.1	2.4
ds	mm	468.5	244.5	1200.0	198.0
dst	mm	25.7	6.3	36.0	15.9
As	mm²	3465.5	3169.7	14137.2	603.2
rhos	%	1.8	1.1	4.7	0.5
rhosw	%	4.3	3.1	15.6	0.5
fsy	MPa	483.9	117.7	990.0	271.0
fc_prime	MPa	47.0	26.2	122.9	13.3
Asw	mm^2	97.4	60.3	232.0	24.6
sw	mm	147.0	57.7	325.0	63.5
fyw	Mpa	456.9	135.9	820.0	270.0
Vu_Rep	kN	363.1	288.2	1330.0	94.0

Table 4-10: Correlation Matrix of all the variables for beam with shear reinforcement

	b	bw	h	kap	ds	dst	As	rhos	rhosw	fsy	Asw	sw	fyw	fc_prime	Vu_Rep
b	1	-0.42	0.377	0.317	0.363	-0.233	0.513	-0.619	0.641	0.101	0.717	-0.033	-0.06	-0.382	0.516
bw	-0.42	1	0.207	-0.256	0.193	0.602	0.119	0.439	-0.573	-0.333	-0.21	0.399	0.041	0.421	0.108
h	0.377	0.207	1	0.058	0.998	0.309	0.786	-0.23	0.299	-0.166	0.589	0.594	0.107	-0.025	0.803
kap	0.317	-0.256	0.058	1	0.054	-0.383	0.206	-0.257	0.335	0.058	0.27	0.152	-0.525	-0.349	0.049
ds	0.363	0.193	0.998	0.054	1	0.294	0.763	-0.244	0.291	-0.152	0.573	0.595	0.108	-0.041	0.784
dst	-0.233	0.602	0.309	-0.383	0.294	1	0.346	0.518	-0.077	-0.507	-0.094	0.197	0.446	0.525	0.348
As	0.513	0.119	0.786	0.206	0.763	0.346	1	-0.006	0.617	-0.166	0.743	0.318	0.105	0.194	0.944
rhos	-0.619	0.439	-0.23	-0.257	-0.244	0.518	-0.006	1	-0.179	-0.334	-0.364	-0.172	0.265	0.722	-0.073
rhosw	0.641	-0.573	0.299	0.335	0.291	-0.077	0.617	-0.179	1	0.085	0.651	-0.169	0.147	-0.017	0.565
fsy	0.101	-0.333	-0.166	0.058	-0.152	-0.507	-0.166	-0.334	0.085	1	0.08	-0.235	0.044	-0.056	-0.076
Asw	0.717	-0.21	0.589	0.27	0.573	-0.094	0.743	-0.364	0.651	0.08	1	0.269	-0.079	-0.127	0.742
sw	-0.033	0.399	0.594	0.152	0.595	0.197	0.318	-0.172	-0.169	-0.235	0.269	1	-0.156	-0.195	0.255
fyw	-0.06	0.041	0.107	-0.525	0.108	0.446	0.105	0.265	0.147	0.044	-0.079	-0.156	1	0.37	0.216
fc_prime	-0.382	0.421	-0.025	-0.349	-0.041	0.525	0.194	0.722	-0.017	-0.056	-0.127	-0.195	0.37	1	0.253
Vu_Rep	0.516	0.108	0.803	0.049	0.784	0.348	0.944	-0.073	0.565	-0.076	0.742	0.255	0.216	0.253	1

4.3.2 Variable selection and Estimation

The variable selection is conducted by using multiple linear regression, LASSO, LARS and random forest method on the small dataset in this thesis first. Table 4-11, 4-12 and 4-13 show the result of forward, backward, and stepwise methods from multiple regression, respectively. Figures 4-20 and 4-21 show the variable selection result from LASSO, and random forest, respectively

Table 4-11: Results of variable selection by forward method from MLR for beam with shear reinforcement

Selection Summary

Step	Variable Entered	R-Square	Adj. R-Square	C(p)	AIC	RMSE
1	As	0.8913	0.8900	106.0323	1044.2912	95.5721
2	kap	0.9135	0.9115	69.3923	1026.3968	85.7523
3	rhos	0.9257	0.9230	50.2347	1015.2062	79.9693
4	fc_prime	0.9424	0.9395	23.2709	995.1295	70.8684
5	h	0.9448	0.9414	20.9423	993.2903	69.7485
6	sw	0.9496	0.9459	14.6098	987.3860	67.0662
7	Asw	0.9533	0.9492	10.1404	982.7121	64.9498
8	fyw	0.9548	0.9502	9.5512	981.8907	64.3134
9	rhosw	0.9574	0.9524	7.0947	978.8093	62.8667

Table 4-12: Results of variable selection by backward method from MLR for beam with shear reinforcement

Elimination Summary

Step	Variable Removed	R-Square	Adj. R-Square	C(p)	AIC	RMSE
1	fsy	0.9586	0.9512	13.0613	984.3884	63.6740
2	ds	0.9585	0.9517	11.2421	982.6063	63.3216
3	b	0.9581	0.952	9.7921	981.2662	63.1370
4	dst	0.9579	0.9523	8.3075	979.8800	62.9419
5	bw	0.9574	0.9524	7.0947	978.8093	62.8667

Table 4-13: Results of variable selection by stepwise method from MLR for beam with shear reinforcement

Stepwise Selection Summary

Step	Variable	Added/ Removed	R-Square	Adj. R-Square	C(p)	AIC	RMSE
1	As	addition	0.891	0.890	106.0320	1044.2912	95.5721
2	kap	addition	0.914	0.911	69.3920	1026.3968	85.7523
3	rhos	addition	0.926	0.923	50.2350	1015.2062	79.9693
4	fc_prime	addition	0.942	0.940	23.2710	995.1295	70.8684
5	h	addition	0.945	0.941	20.9420	993.2903	69.7485
6	sw	addition	0.950	0.946	14.6100	987.3860	67.0662
7	Asw	addition	0.953	0.949	10.1400	982.7121	64.9498

```
15 x 1 sparse Matrix of class "dgCMatrix"
                      s0
(Intercept)  23.96259878
b             .
bw            .
h             0.19761255
kap         -21.46025351
ds            .
dst           .
As            0.06222774
rhos        -22.63219158
rhosw         .
fsy           .
Asw           0.56275579
sw           -0.23544248
fyw           0.09940803
fc_prime      1.39608741
```

Figure 4-20: Result of variable selection from LASSO for beam with shear reinforcement

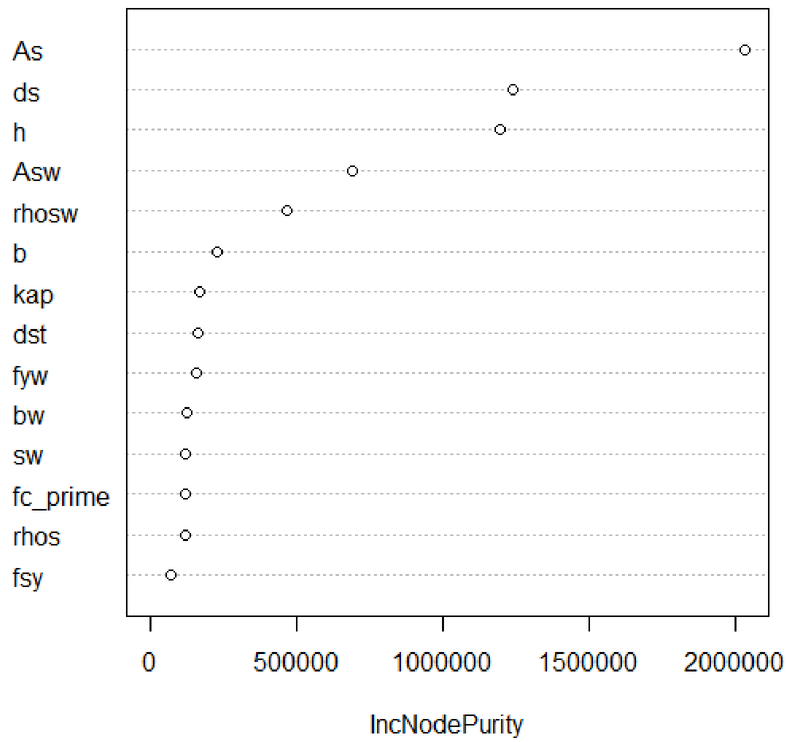

Figure 4-21: Variable Importance Plot generated by Random Forest for beam with shear reinforcement

The results of variable selection from statistical methods (MLR, LASSO) were similar with each other. As, kap, rhos (related with As, b, ds) are highly statistically significant. fc_prime, h (related with ds), sw, Asw and fyw is also significant but not as much as previous variables. The remaining variables are not significant at all. The results of variable selection from statistical methods are very similar with the variables in the theoretical equations. For the result of random forest, As, ds and Asw are highly statistically significant as well as rhosw (related with bw, ds, As). However, kap, fc_prime, sw and fyw are not significant in the random forest. Based on the results of variable selection and knowledge of subject matter, bw, As, ds, kap, fc_prime, Asw, sw and fyw were decided to keep in the following study.

4.3.3 Accuracy

A scatter plot matrix was created in Figure 4-22 to visually show the relationship between all the variables. Based on Figure 4-22, ds, As and Asw have a very strong positive correlation with shear strength while the trend is not significant for the remaining selected variables. The KDD densities show that distributions of bw and ds were skewed right which means that the dataset at hand does not contain many big concrete beams.

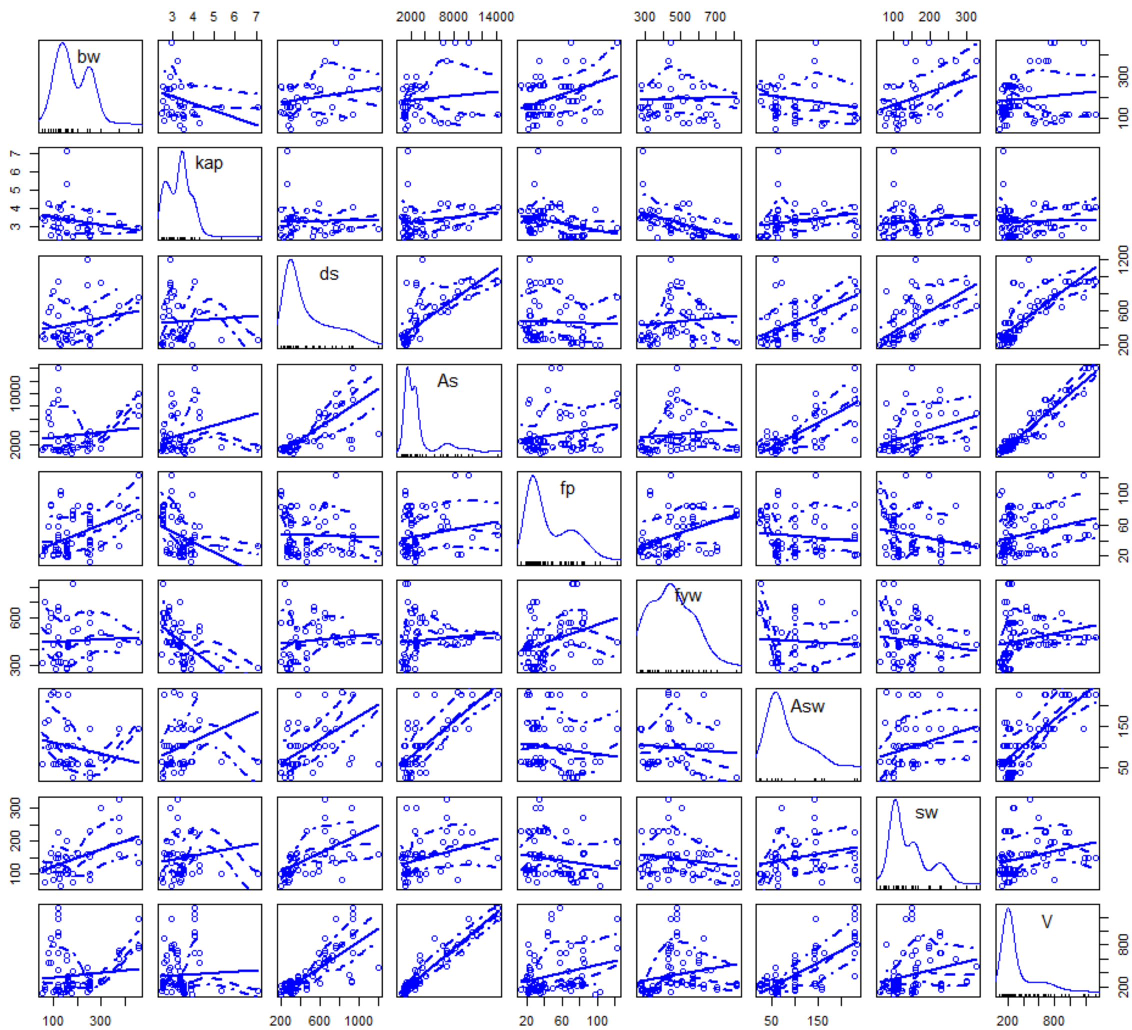

Figure 4-22: Scatter Plot Matrix of selected variables for beam with shear reinforcement

For theoretical methods, the equations to calculate the shear strength of beam with shear reinforcement were provided in chapter 3. Same with the concrete beam without reinforcement, the theoretical methods usually underestimate the shear strength for safety purpose. Figure 4-23 shows the comparison between experimental shear strength and predicted shear strength using three theoretical methods. Then, a correction factor was added in the calculation. These correction factors were calculated by using simple regression to shift the results to have best fit with accuracy line. They are 1.16 for ACI318-19, 1.14 for Frosch et al. and 1.27 for Li et al.

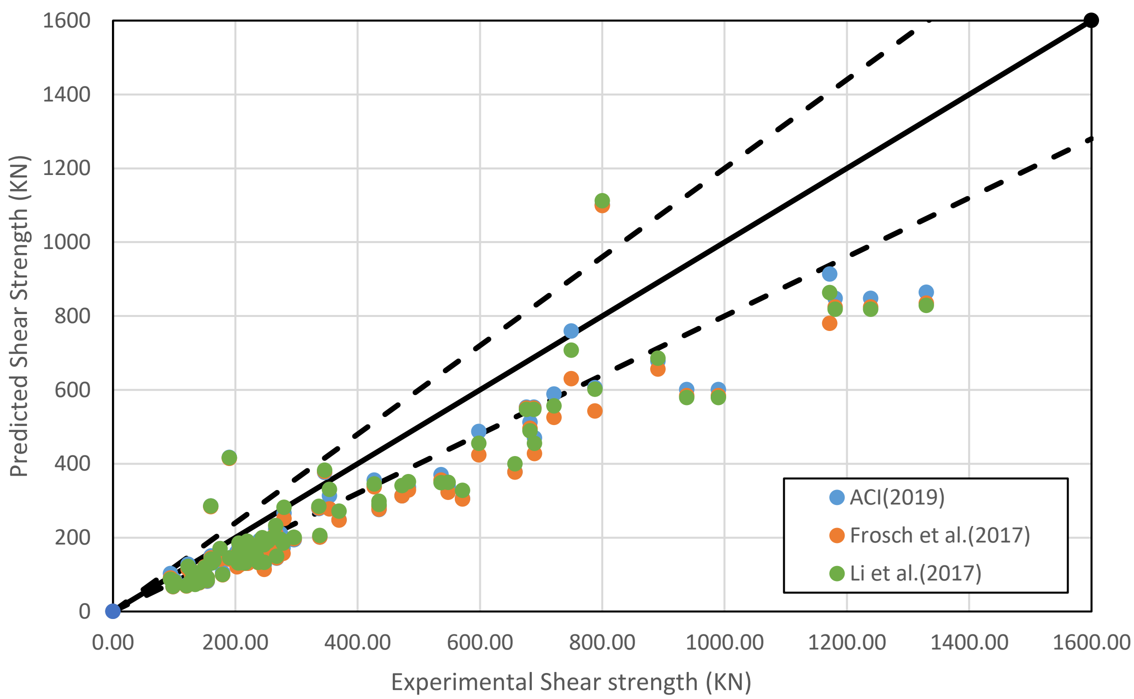

Figure 4-23: Comparison between experiment and uncorrected prediction by theoretical methods for beam with shear reinforcement

Figure 4-24 shows the comparison of corrected predicted shear strength and experimental shear strength by using theoretical methods. After correction, some cases are overestimated which are not conservative anymore.

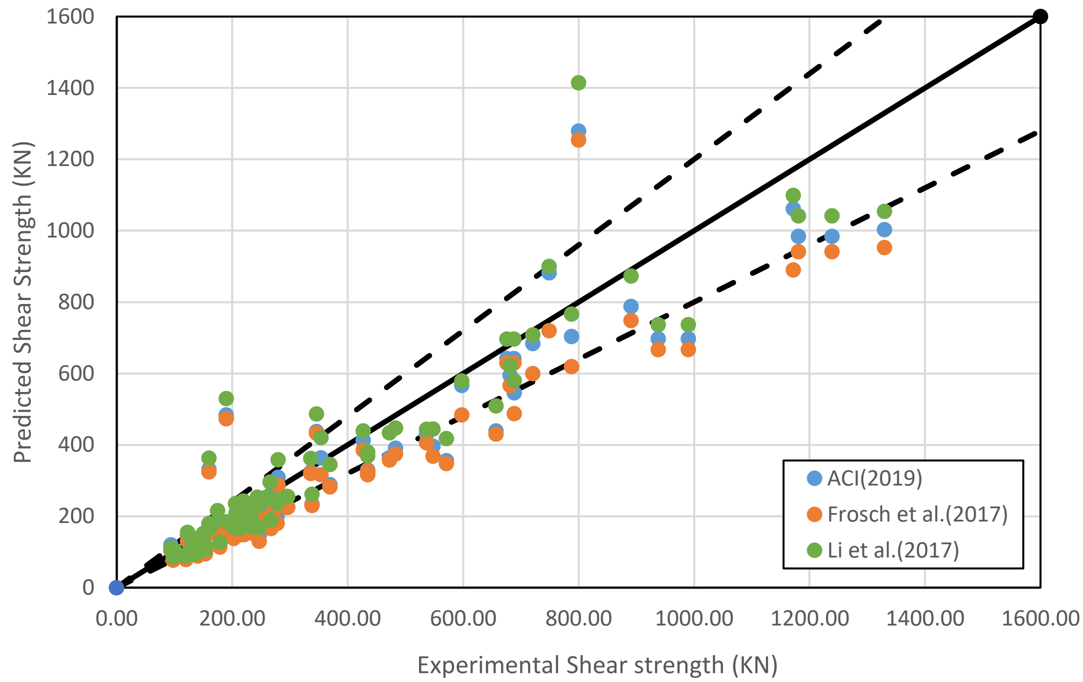

Figure 4-24: Comparison between experiment and corrected prediction by theoretical methods for beam with shear reinforcement

Statistical models were achieved by using MLR, LASSO and LARS which were shown in Equations 4-4, 4-5 and 4-6, respectively.

For MLR method after use of variable selection methods, the final model is the following:

$$\boldsymbol{Vu_{Rep} = -1.40 + 0.042bw - 26.12kap + 0.33ds + 0.055As + 1.12fc_{prime} + 0.097fyw - 0.64sw + 0.96Asw} \quad \textbf{Equation 4-4}$$

For LASSO method,

$$\boldsymbol{Vu_{Rep} = 5.32 - 10.05kap + 0.14ds + 0.064As + 0.44fc_{prime} + 0.082fyw + 0.45Asw} \quad \textbf{Equation 4-5}$$

For LARS method,

$$\boldsymbol{Vu_{Rep} = -1.40 + 0.042bw - 26.12kap + 0.33ds + 0.055As + 1.12fc_{prime} + 0.097fyw - 0.64sw + 0.96Asw} \quad \textbf{Equation 4-6}$$

The MLR model and LARS model were exactly same again. One thing needs to be noticed here is the parameter of bw and sw are zero in LASSO method which means they are considered as insignificant variables in LASSO regression. The result of LARS will not be shown in this chapter. Figure 4-25 shows the comparison of predicted shear strength and experimental shear strength by using statistical methods. The accuracy of prediction is decent. Most of predictions were within the 20% error.

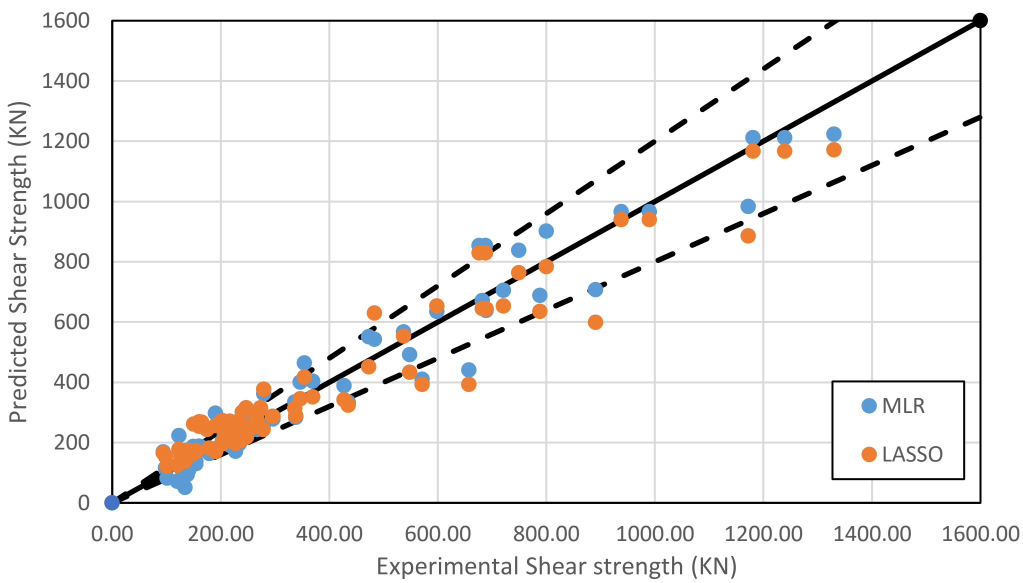

Figure 4-25: Comparison between experiment and prediction by statistical methods for beam with shear reinforcement

Then the machine learning methods were applied on the reduced dataset as well. The comparison of predicted shear strength and experimental shear strength by using machine learning methods was shown in Figure 4-26. 500 trees (ntree=500) were selected in the random forest method while 10 hidden layers (hidden=10) was chosen for the NN method. The accuracies of prediction of random forest and SVM are decent. Most of the predictions fall with 20% error line. Neutral network performed bad in this dataset. The sample size might be the reason why NN is not good here.

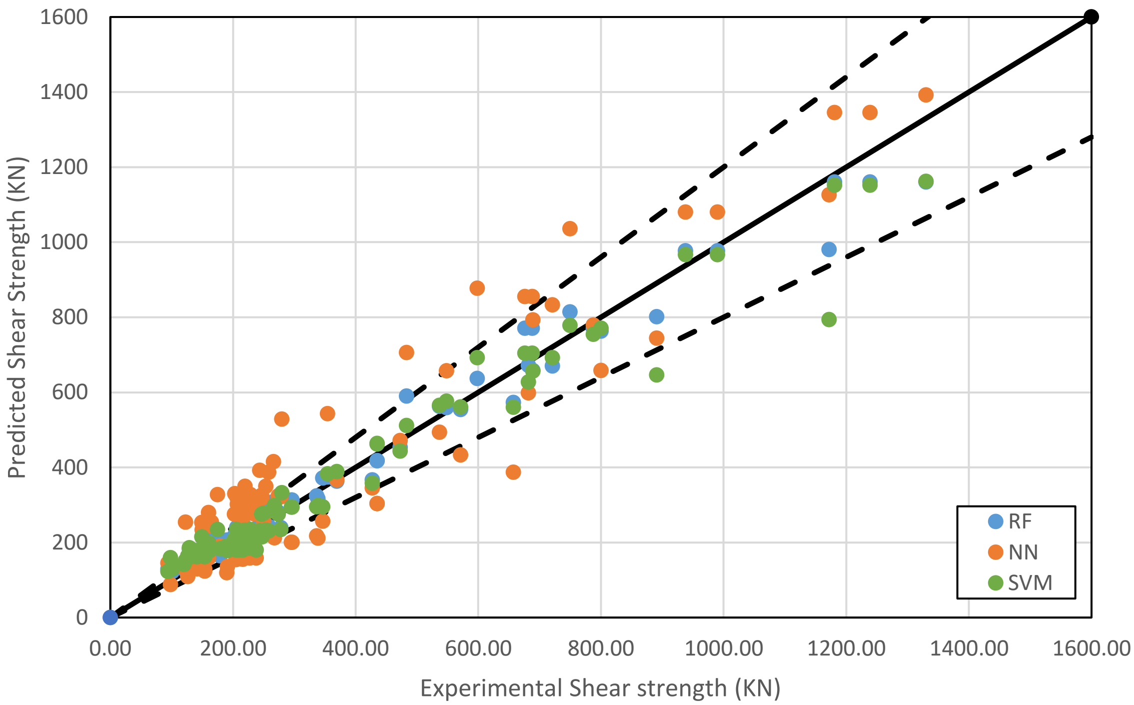

Figure 4-26: Comparison between experiment and prediction by machine learning methods for beam with shear reinforcement

To reduce the bias and uncertainty of statistical and machine learning methods, leave-one-out cross validation was applied on the dataset by using these different methods. Each observation will be used as a test dataset while the remaining fold will be used as a training dataset. The statistical/ML methods will be applied on the training dataset first. Then, the test dataset will be used to see the goodness of fit for each method by using the criteria called root of mean square error (RMSE), which a metric to summarizes predicted model quality. This process will be repeated by 87 times since each observation will be used as test dataset once. Table 4-14 shows the result of cross validation of statistical/ML methods compared with theoretical methods. The result shows that all statistical/ML methods are doing well except NN which was reflected in Figure 4-26 as well. Compared with the prediction of concrete beam without shear reinforcement, the accuracy of all the methods drops a lot, especially NN method.

Table 4-14: Results of goodness of fit for all the methods for concrete beam with shear reinforcement

Group	Method	RMSE
Theoretical	ACI318-19 (2019)	107
	Frosch et al. (2017)	102
	Li et al. (2017)	111
Statistical	MLR	81
	LASSO	86
Machine Learning	RF	81
	NN	467
	SVM	87

4.2.4 Sensitivity Analysis

The sensitivity analysis was applied on all these methods except LARS regression since the LARS based model is the same as multiple linear regression-based model. A test dataset was established based on the boxplot of each variable. The range of each variable in the test dataset will be between Minimum and Maximum in their boxplot. Table 4-15 shows the range and median value of each variable which were used in test dataset. Figures 4-33, 4-34, 4-35, 4-36, 4-37, 4-38, 4-39 and 4-40 show the sensitivity analysis for bw, fc_prime, kap, low, ds, Asw, sw and fyw, respectively. The trend of theoretical method can give a reference to compare with the statistical/ML methods since the theoretical methods were published with peer review.

Table 4-15: The detail of test dataset for beam with shear reinforcement

Variables	unit	median	Maximum	Minimum
bw	mm	205	360	50
Kap	-	4.2	6	2.4
ds	mm	325	600	50
low	-	0.0255	0.05	0.001
fc_prime	Mpa	40	70	10
Asw	mm^2	128.5	232	25
sw	mm	163	262.5	63.5
fyw	Mpa	545	820	270

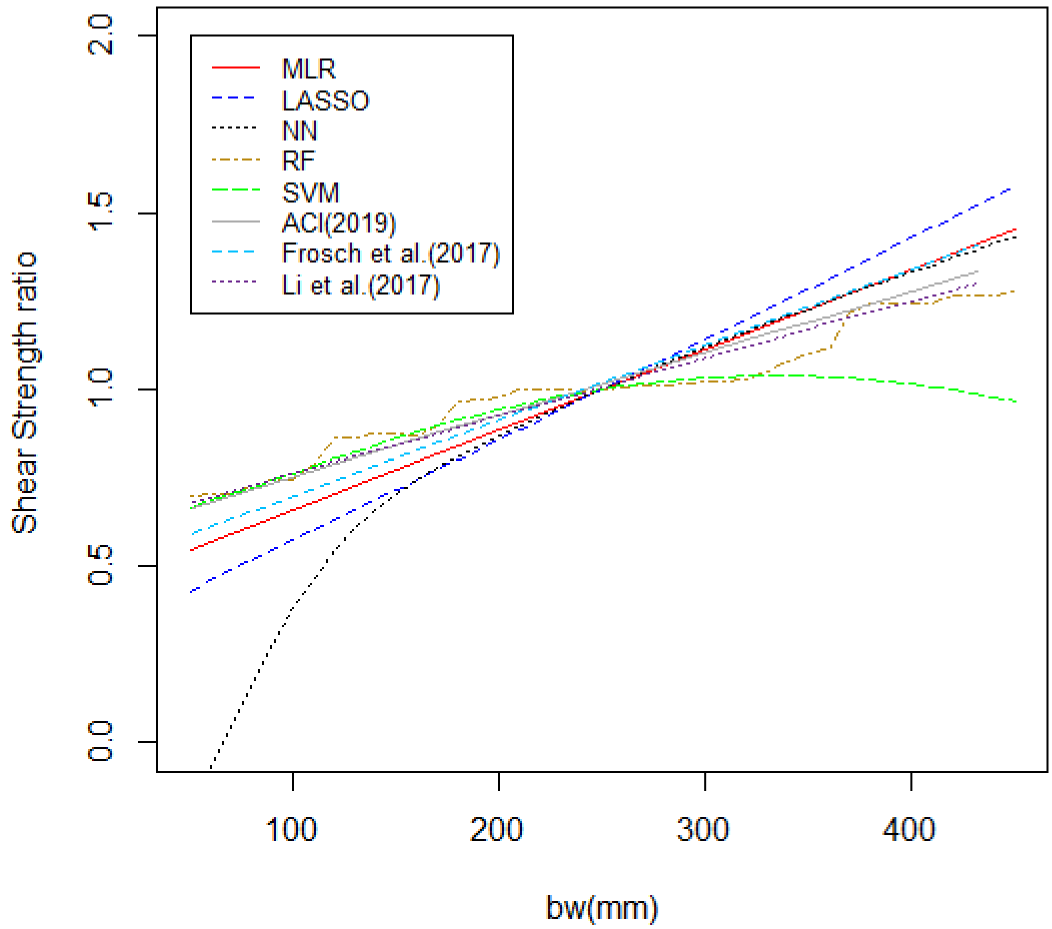

Figure 4-27: Sensitivity analysis on bw for concrete beam with shear reinforcement

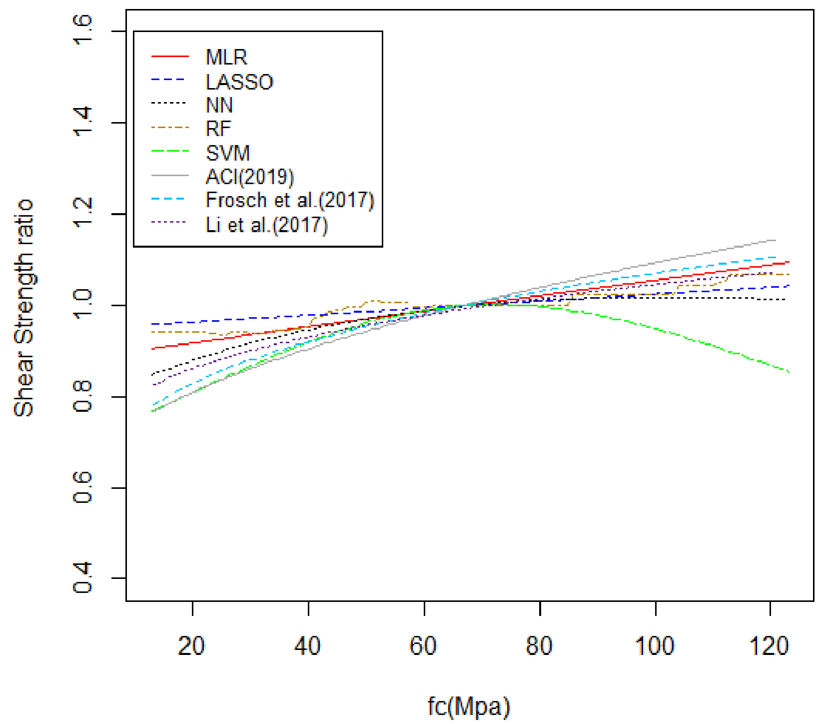

Figure 4-28: Sensitivity analysis on fc_prime for concrete beam with shear reinforcement

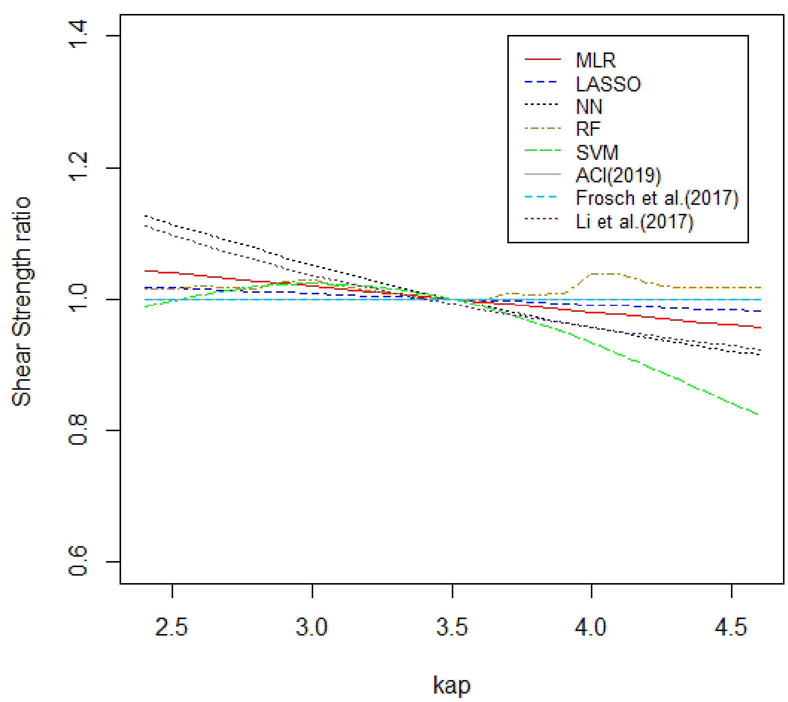

Figure 4-29: Sensitivity analysis on kap for concrete beam with shear reinforcement

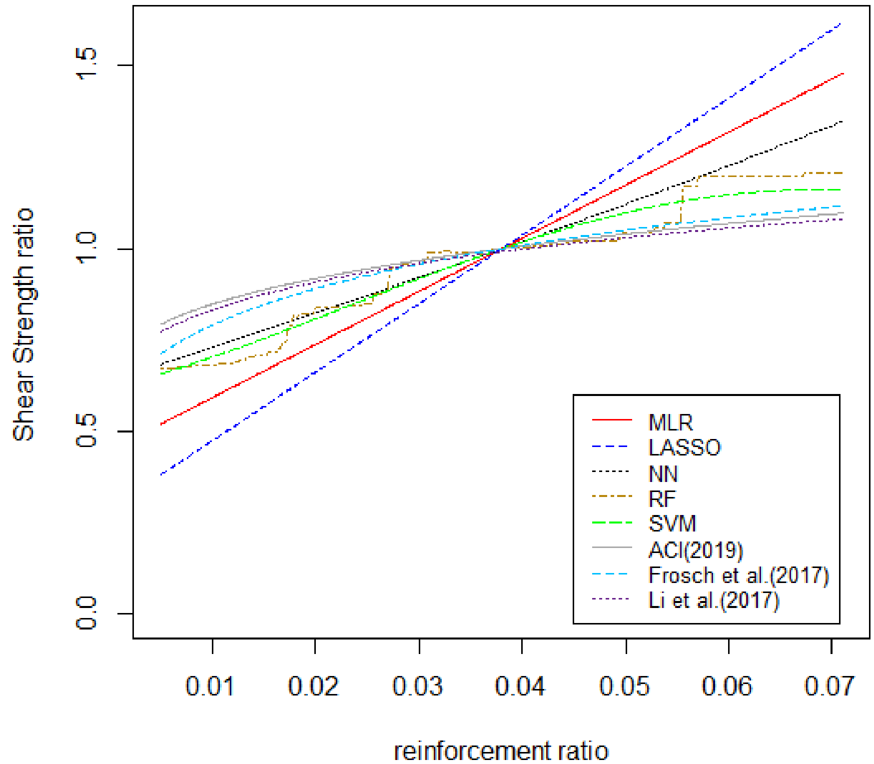

Figure 4-30: Sensitivity analysis on low for concrete beam with shear reinforcement

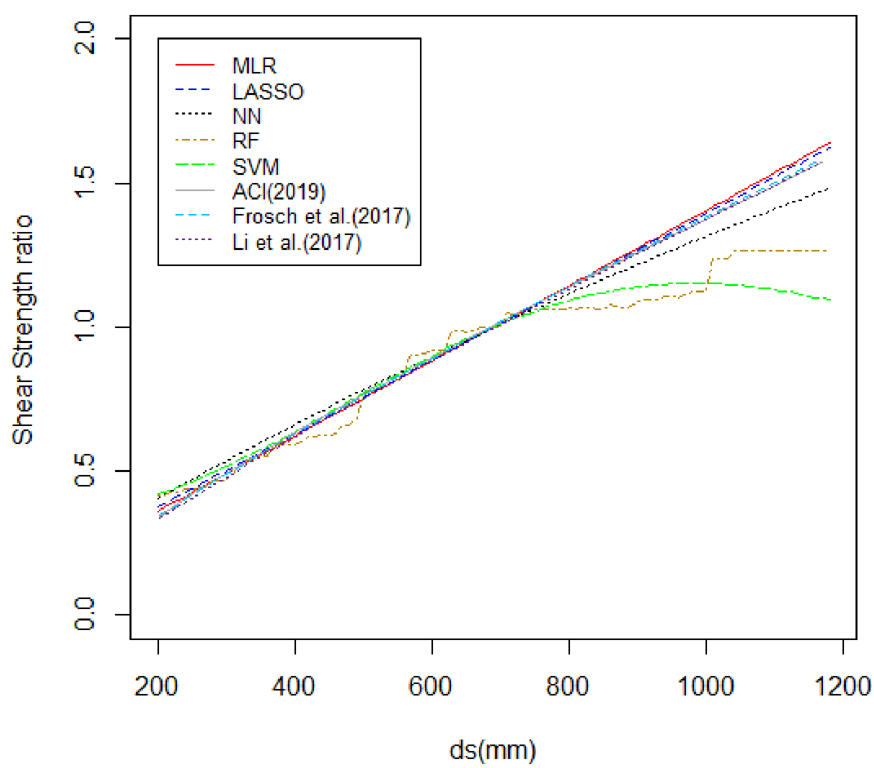

Figure 4-31: Sensitivity analysis on ds for concrete beam with shear reinforcement

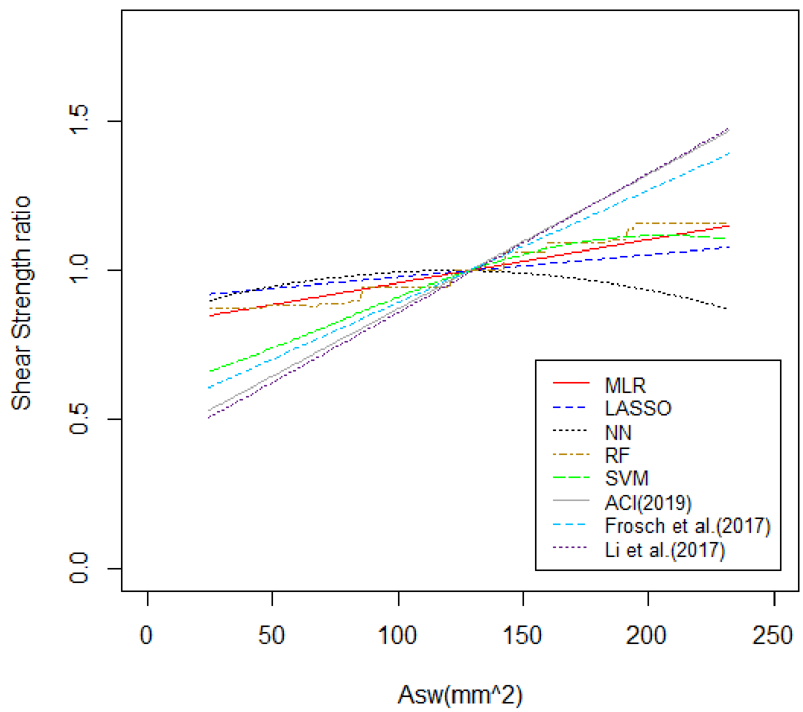

Figure 4-32: Sensitivity analysis on Asw for concrete beam with shear reinforcement

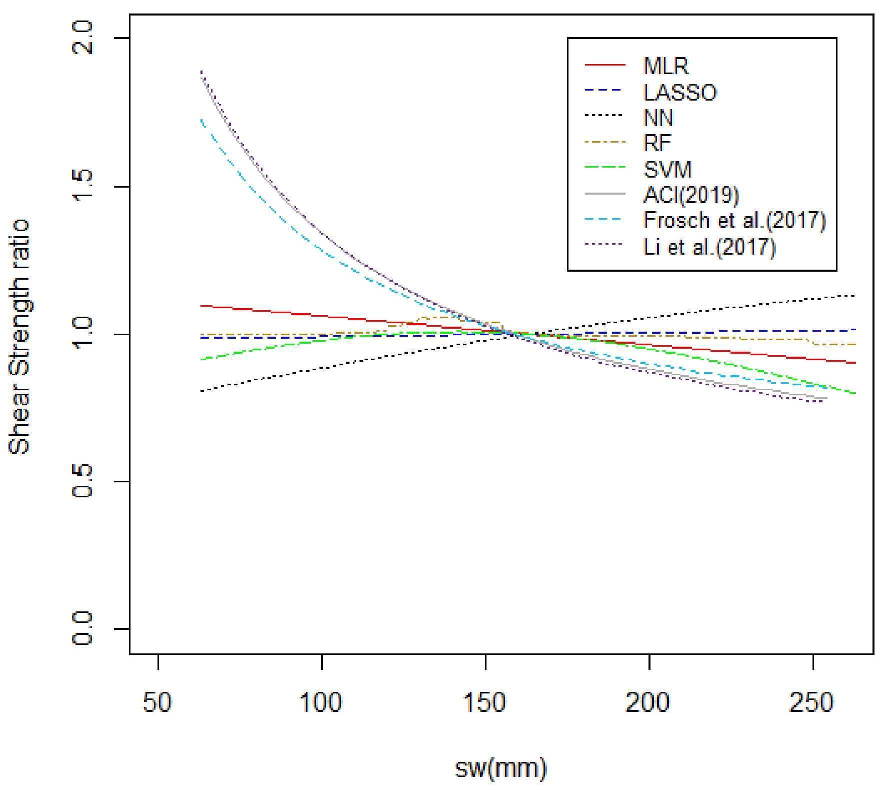

Figure 4-33: Sensitivity analysis on sw for concrete beam with shear reinforcement

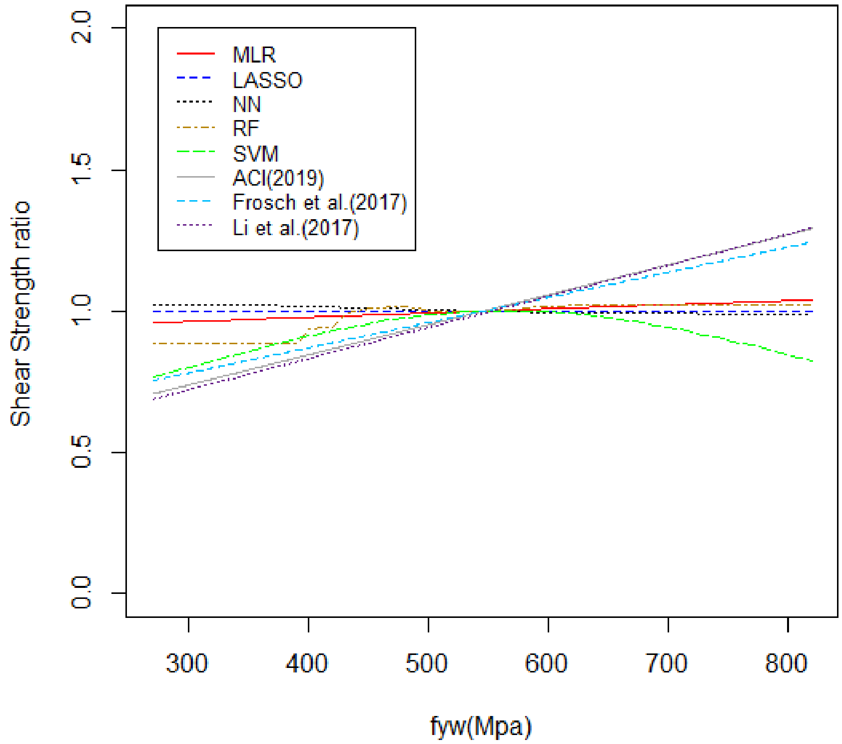

Figure 4-34: Sensitivity analysis on fyw for concrete beam with shear reinforcement

For sensitivity analysis on bw (Figure 4-27), the trend of shear strength ratio from theoretical methods and statistical methods are very close to each other. The random forest method has similar trend as well, but it is very unstable. The SVM method performs well only when bw is less than median value while NN method behaves well only when bw is larger than median value.

For sensitivity analysis on fc_prime (Figure 4-28), the result of LASSO regression is worst since it didn't capture the changes along with fc_prime. MLR and random forest have the similar trend compared with theoretical methods but not very accurate. Similar with previous result, the random forest method is not stable. The SVM and NN are very close to theoretical methods when fc_prime is less than median value, especially NN method.

For sensitivity analysis on kap (Figure 4-29), ACI and Frosch methods don't have any trend since kap is not considered at all. Only NN method has similar trend with Li's method. The other statistical/ML methods didn't reflect significant change of shear strength along with kap.

For sensitivity analysis on low (Figure 4-30), all the statistical/ML methods have the positive trend along with the increasement of longitudinal reinforcement ratio, but none of them is close to theoretical methods.

For sensitivity analysis on ds (Figure 4-31), the statistical methods and NN method are close the theoretical methods. Random forest and SVM method perform well when ds is less than median value.

For sensitivity analysis on Asw (Figure 4-32), none of the statistical/ML methods except SVM method is close to theoretical methods. The SVM method only performs well when Asw is less than median value.

For sensitivity analysis on sw (Figure 4-33), none of the statistical/ML methods is close to theoretical methods. The NN method even has the opposite trend compared with the theoretical methods.

For sensitivity analysis on fyw (Figure 4-34), none of the statistical/ML methods except SVM method is close to theoretical methods. The SVM method only performs well when fyw is less than median value.

Overall, the statistical/ML methods performed bad in the sensitivity analysis. The main reason of this is because the experimental dataset is too small for concrete beam with shear reinforcement. The performance of sensitivity analysis on the properties of shear reinforcement is even worse than that on properties of concrete, especially on Asw and sw. That’s because the Asw and sw are not exactly “continuous” which can be found in the scatter plot (Figure 4-22). The designers would like to pick the similar Asw and sw for shear reinforcement in the concrete beam which means a lot of beams in the dataset have exactly same Asw and sw.

4.4 Summary

In this chapter, data description and variable selection were applied on the concrete beam with and without shear reinforcement, respectively. Statistical models by using MLR, LASSO and LARS regression were established as well in this chapter. Since variable selection was conducted first, the MLR and LARS model were exactly same with each other. Then, goodness of fit was checked by RMSE for all the methods on concrete beam with and without shear reinforcement dataset, respectively. For concrete beam without shear reinforcement, all the methods performed well. however, for beam with shear reinforcement, the NN didn’t have a good performance. Finally, the sensitivity analysis was applied on the concrete beam with and without shear reinforcement dataset, respectively. For concrete beam without shear reinforcement, the statistical/ML methods performed decent in the small test dataset. However, for concrete beam with shear reinforcement, the statistical/ML methods performed bad because the training dataset which is the experimental dataset for beam with shear reinforcement is too small.

Chapter 5 Conclusion and Future work

The goal of this thesis is to evaluate statistical/ML methods for predicting one-way shear strength of reinforced concrete beam with and without shear reinforcement. Three statistical methods including multiple linear regression, LASSO and LARS and the three machine learning methods including Random Forest, NN and SVM were selected in this thesis for the evaluation. A dataset with 784 observations was used to evaluate the concrete beam without shear reinforcement while a dataset with 87 observations was used to evaluate the concrete beam with shear reinforcement. Three theoretical methods including ACI318-19 (2019), Frosch et al. (2017) and Li et al. (2017) were also employed in this thesis to compare with the statistical/ML methods. The statistical models and the "Rules" of machine learning methods were accomplished by using the experimental datasets. Goodness of fit and sensitivity analysis were used to evaluate all statistical/ML methods in this thesis. The results of theoretical methods were used as a reference to evaluate whether the statistical/ML methods performed well or not.

For concrete beam without shear reinforcement, all the statistical/ML methods performed well in the goodness of fit test. In the sensitivity analysis, the statistical/ML methods did not perform well in the large test dataset which contains a lot of large beams. The reason for this poor performance was that the training dataset with 784 observations does not have enough samples of large beams. However, the statistical/ML methods performed better in the small test dataset. MLR, NN and SVM are more accurate compared with LASSO and random forest. NN method has the best performance among all these statistical/ML method.

For concrete beam with shear reinforcement, all the statistical/ML methods performed well except NN method in the goodness of fit test. In the sensitivity analysis, the statistical/ML methods have not performed well at all, especially the sensitivity analysis on the properties of shear reinforcement. The main reason for this is because the experimental dataset is too small for concrete beam with shear reinforcement. In addition, the variables which were related with shear reinforcement are identical in the structural design which means a lot of beams in the dataset have exactly same Asw, sw and fyw. All the drawback of the concrete beam with shear reinforcement dataset had a negative impact on the accuracy of prediction by using statistical/ML methods.

Overall, the statistical/ML methods can be used to evaluate the one-way shear strength of concrete beam if the experimental dataset is sufficient and well designed. The prediction will be less accurate or just bad if the properties of beams are significantly beyond the range of properties of beams in training dataset. More experimental work needs to be done to increase the size of dataset which can make prediction more accurate. With the development of computer hardware and software, the finite element analysis can be used to simulate the shear failure of concrete beam with much less cost compared with experiment. When the sufficient and well-designed datasets are generated, the statistical/ML methods might be more accurate compared with theoretical methods since they can capture more nonlinear behavior during the simulation.

References

ACI. (2014). *Building code requirements for structural concreteconcrete: (ACI 318-14);and commentary (ACI 318R-14).* Farmington Hills, MI: American Concrete Institute.

ACI. (2019). *Building code requirements for structural concrete: (ACI 318-19); and commentary (ACI 318R-19).* Farmington Hills, MI: American Concrete Institute.

Aguilar, V. (2020). *Reliability of Shear-Critical Reinforced Concrete Members.* Auburn. AL: Auburn University.

Aguilar, V., Sandoval, C., Adam, J. M., Garzon-Roca, J., & Valdebenito, G. (2016). Prediction of the shear strength of reinforced masonry walls using a large experimental database and artificial neural networks. *Structure and Infrastructure Engineering*.

AGUILAR, V., WU, H., & MONTGOMERY, J. (2021). Machine Learning in Foundation Design and More. *TRB Centennial Circular Number E-C273*, 103-118.

Bentz, E. C., & Collins, M. P. (2017). Updating the ACI Shear Design Provisions. *Concrete International, V. 39, No. 9*, 33-38.

Boser, B., Guyon, I., & Vapnik, V. (1992). A training algorithm for optimal margin classifiers. *Proceedings of the Fifth Annual Workshop on Computational Learning Theory* (pp. 144–152). ACM.

Breiman, L. (2001). Random Forests. *Machine Learning 45(1)*, 5-32.

BREIMAN, L., FRIEDMAN, J. H., OLSHEN, R., & STONE, C. J. (1984). *Classification and Regression Trees.* Belmont, CA: Wadsworth.

Breiman, L., Friedman, J., Stone, C. J., & Olshen, R. A. (1984). *Classification and regression trees.* CRC Press.

Cladera, A., Marí, A., & Bairán, J.-M. (2017). One-Way Shear Design Method Based on a Multi-Action Model. *Concrete International, V. 39, No. 9*, 40-46.

Efron, B., & Hastie, T. (2016). *Computer Age Computer Age Inference Algorithms, Evidence , and Data Science.* california: Cambridge University Press.

EFRON, B., HASTIE, T., JOHNSTONE, I., & TIBSHIRANI, R. (2004). LEAST ANGLE REGRESSION. *The Annals of Statistics Vol. 32, No. 2*, 407-499.

Foody, G. M., & Mathur, A. (2004). A relative evaluation of multiclass image classification by support vector machines. *IEEE Transactions on Geoscience and Remote Sensing 42* (pp. 1335–1343). Piscataway, NJ: IEEE.

Frosch, R. J., Yu, Q., & Cusatis, G. (2017). A Unified Approach to Shear Design. *Concrete International, V. 39, No. 9*, 47-52.

Galarnyk, M. (2018). *Understanding Boxplots*. Retrieved from towards data science: https://towardsdatascience.com/understanding-boxplots-5e2df7bcbd51

Hastie, T., Tibshirani, R., & Friedman, J. (2017). *The Elements of Statistical Learning Data Mining, Inference, and Prediction Second Edition.* California: Springer.

Hebiri, M., & Lederer, J. (2013). How correlations influence lasso prediction. . *Information Theory 59(3)* (pp. 1846-1854). IEEE.

ISHAK, S., & TRIFIRÒ, F. (2007, January). Neural Networks. *Artificial Intelligence in Transportation*, pp. 17-32.

Kirtland, K. M. (2017). *Outlier Detection and Multicollinearity in Sequential Variable Selection: A Least Angle Regression-Based Approach.* Auburn, Alabama: Auburn University.

Kuchma, D. A., & Collins, M. P. (1998). Advances in understanding shear performance of concrete structures. *Progress in Structural Engineering and Materials, V. 1, No. 4*, 360-369.

Li, Y.-A., Hsu, T. T., & Hwang, S.-J. (2017). Shear Strength of Prestressed and Nonprestressed Concrete Beams. *Concrete International, V. 39, No. 9*, 53–57.

Lu, P., Chen, S., & Zheng, Y. (2012). Artificial Intelligence in Civil Engineering. *Mathematical Problems in Engineering Volume 2012.* Hindawi Publishing Corporation.

NCSS. (2021). *Stepwise Regression.* NCSS, LLC.

Park, H.-G., & and Choi, K.-K. (2017). Unified Shear Design Method of Concrete Beams Based on Compression Zone Failure Mechanism. *Concrete International, V. 39, No. 9*, 59-63.

Reineck, K.-H. (2017). Proposal for ACI 318 Shear Design. *Concrete International, V. 39, No. 9,*, 65–70.

Reineck, K.-H., Bentz, E., & Fitik, B. (2014). ACI-DAfStb Databases for Shear Tests on 7 Slender Reinforced Concrete Beams with Stirrups. *ACI Structural Journal, V. 111, No. 8* .

Reineck, K.-H., Kuchma, D. A., & Kim, K. S. (2003). Shear Database for Reinforced 4 Concrete Members without Shear Reinforcement. *ACI Structural Journal, V. 100, No. 5 2*.

SONG, Y.-y., & LU, Y. (2015). Decision tree methods: applications for classification and prediction. *Shanghai Archives of Psychiatry, 2015, Vol. 27, No. 2*, 130-135.

Swan, B. T. (2017). *Adaptation and Comparison of Machine Learning Methods for Geomorphological Mapping and Terrace Prediction: A Case Study of the Buffalo River, Arkansas.* Auburn, AL: Auburn University.

Tibshirani, R. (1986). Regression Shrinkage and Selection via the lasso. *Journal of the Royal Statistical Society. Series B (methodological). Wiley. 58 (1)*, 267-288.

Tso, B., & Mather, P. (2009). *Classification Methods for Remotely Sensed Data, 2nd ed.* Boca Raton: CRC Press.

Wang, C.-K., & Salmon, C. G. (1985). *Reinforced concrete design.* New York: Harper & Row Publishers.

Wight, J. K., & MacGregor, J. G. (2012). *Reinforced concrete: mechanics and design, 6th edition.* New Jersey: Pearson Education.

Wilson, M. (2008). Support Vector Machines. *Jørgensen, S.E., Fath, B.D. (Eds.), Encyclopedia of Ecology* (pp. 3431-3437). Oxford: Academic Press.