



# Development and Comparison of Prediction Models for Sanitary Sewer Pipes Condition Assessment Using Multinomial Logistic Regression and Artificial Neural Network

Daniel (Dan) Atambo, Ph.D, P.E., M. ASCE

[danielogaro.atambo@mavs.uta.edu](mailto:danielogaro.atambo@mavs.uta.edu)

817-793-4554



## Introduction

- Provision of wastewater services is essential for Public health, safety, and social economic development
- 2021 American Society of Civil Engineers (ASCE) Infrastructure Report Card Wastewater Infrastructure – D+ Score
- Regulating agencies demand periodic inspections
- Limited budgets for inspection and condition assessment
- Utilities need pipe condition prediction models to prioritize capital improvements.



## Background (*Cont'd*)

- Large amounts of inspection and condition assessment **data** needed



Pipe Condition  
Source: City of Dallas



## Background (*Cont'd*)

- Sewer pipeline is critical infrastructure.
- Most of sewer pipelines are old and deteriorating (Alegre, 2010).
- **Physical, operational, and environmental factors influence pipe condition.**
- Inspection, condition assessment, and renewal of sewer pipes prevents pipe failures.



## Background (*Cont'd*)

- National Association of Sewer Service Companies (**NASSCO**) Pipeline Assessment Certification Program (**PACP**)
- Standard for **defect coding** and collection of data
- **PACP Pipe Condition Rating grades :**
  - 1-Excellent
  - 2- Good
  - 3- Fair
  - 4- Poor
  - 5- Extremely Poor/Immediate



## Background (*Cont'd*)

- Prediction models are used to determine sewer pipe condition
- **Types of Models: Physical, Artificial, and Statistical**
- **Statistical Models:** Logistic regression, binary regression, linear regression, exponential regression, Markov chain, semi-Markov chain, ordinal regression, and cohort survival
- **Artificial Intelligence Models:** Artificial Neural Network (**ANN**) and genetic algorithms and machine learning



## Background (*Cont'd*)

- Logistic Regression Method

$$\log \left[ \frac{\pi}{1-\pi} \right] Y = \left[ \frac{p(y=1|x_1 \dots x_n)}{1-p(y=1|x_1 \dots x_n)} \right] = a + \beta_1 x_1 + \beta_2 x_2 \dots + \beta_p x_n$$

Where:

Y = dependent variable

i = 1, 2, ..., k-1 correspond to categories of the dependent variable,

n is the number of independent variable

x<sub>i</sub> are independent variables

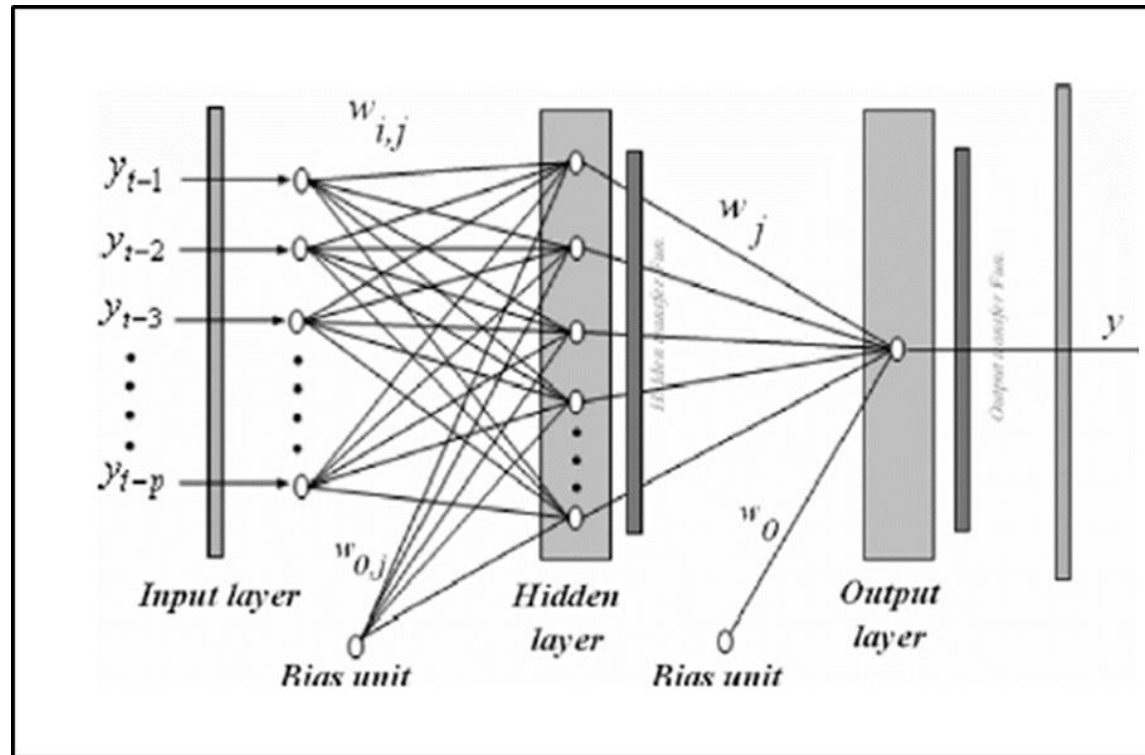
a is intercept parameter

B<sub>p</sub> are regression coefficients associated with p independent variables.

Probability of (y = 1) determined using exponential transformation.

## Background (Cont'd)

- Artificial Neural Network Method (ANN)
- ANN comprised of **input layer**, **hidden layer**, and **output layer**
- ANN Architecture



ANN architecture (Chughtai et al. 2008)



## Problem Statement (*Cont'd*)

- Regulating agencies requirements.
- Limited budget to inspect pipes and prioritize rehabilitation and replacement
- Many studies have been conducted with variables used in prediction models varying from **region to region**
- Need for more research (Mohammadi et al. 2019 and Salman and Salem 2004)
- **Locational attributes** such as soil type, surface condition, water table, etc. need to be incorporated - datasets such as surface condition, soil conditions, slope, and reasons for replacement have not been used in many studies (Vahidi et al. 2016 and Syachrani 2010)



## Problem Statement (*Cont'd*)

- **Most of prediction models** are **statistical methods based** – There is a need of utilizing **artificial intelligence methods** and compare them with the statistical methods.
- Syachrani 2010, in his research study on advanced sewer asset management using dynamic deterioration models stated that there are still some **rooms left for improvement in development of models.**
- **My 7 years** experience of design, planning, construction, and managing water and wastewater pipeline projects with the City of Dallas Water Utilities has made me realize the importance of sewer pipe utilities developing and utilizing prediction sewer pipe condition models.



## Objectives

### **Main Objective:**

To develop Multinomial Logistic Regression and Artificial Neural Network models to predict sanitary sewer pipes condition rating using inspection and condition assessment data.

### **Secondary objectives:**

- i. To identify, evaluate, categorize, and develop relationships of different factors affecting sewer pipes condition.
- ii. To compare the performance of MLR and ANN models for predicting sewer pipe condition.



## Scope of Work

### Included

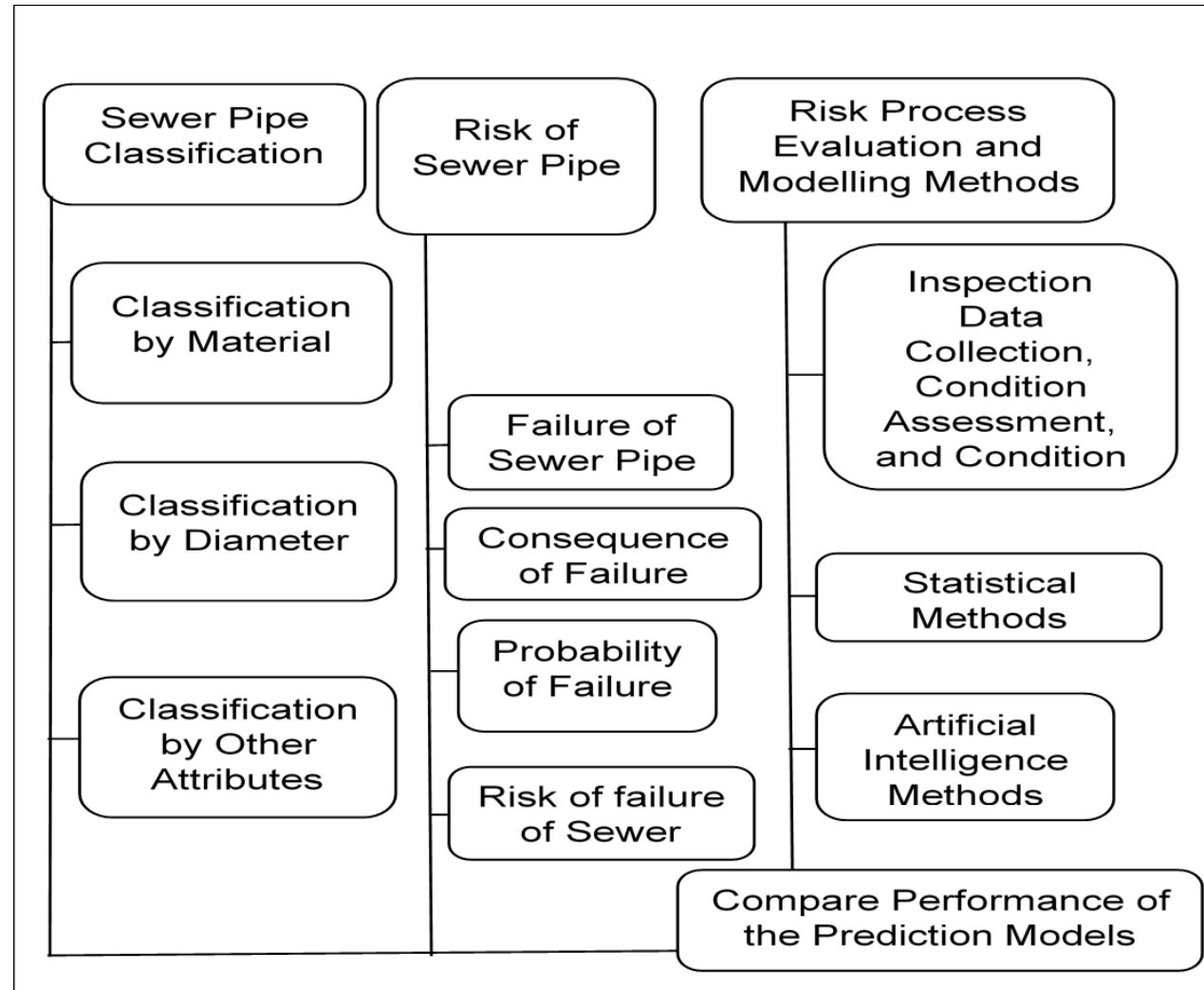
- The condition scores and pipe material, diameter, age, slope, depth, surface condition, soil type, corrosivity concrete, corrosivity steel, and pH variables obtained from condition assessment.
- Data extracted from GIS files for the City of Dallas GIS web/data base.

### Not included

- Pretreatment and Wastewater Treatment Plants (WTPs).
- Financial infrastructure of sewer pipe systems.
- Use, storage, or handling of chemicals.
- Stormwater pipes.
- Force main sewer pipes.



## Literature Review Flow Chart





## Model Development Multinomial Logistic Regression (MLR) Model

- **11 independent variables** were used to the develop the Model
- Data randomly divided into 80% and 20 % for multinomial logistic regression model development and validation.
- Model parameters estimation tables were used to derive one set of model equation broken down into four multinomial logistic regression equations, one for each condition category relative to reference category for sewer pipe condition 5.



## MLR Model (*Cont'd*)

$$g_1(x) = \ln \left[ \frac{\text{Pr}(C=1)}{\text{Pr}(C=5)} \right] = 0.978 * \text{Diameter} + 0.945 * \text{Age} + 1.023 * \text{Slope} + 1.018 * \text{Depth} + 0.999 * \text{Length} + 1.321 * \text{pH} + 1.146 * \text{MaterialCONC} + 1.899 * \text{MaterialPVC} + 0.721 * \text{SurfaceAlley} + 0.771 * \text{SurfaceEasement} + 0.879 * \text{SurfaceHighway} + 0.619 * \text{SoilTypeClay} + 1.037 * \text{SoilTypeLoam} + 0.942 * \text{SoilTypeRock} + 0.962 * \text{CorrosivityConcreteHigh} + 3.653 * \text{CorrosivityConcreteLow} + 1.533 * \text{CorrosivitySteelHigh}$$



## MLR Model (*Cont'd*)

$$\begin{aligned} g_2(x) &= \ln \left[ \frac{\text{Pr}(C = 2)}{\text{Pr}(C = 5)} \right] \\ &= 0.923 * \text{Diameter} + 0.968 * \text{Age} + 0.964 * \text{Slope} + 1.018 * \text{Depth} \\ &+ 0.999 * \text{Length} + 0.598 * \text{pH} + 0.473 * \text{MaterialCONC} + 0.379 \\ &* \text{MaterialPVC} + 1.116 * \text{SurfaceAlley} + 0.961 * \text{SurfaceEasement} \\ &+ 1.133 * \text{SurfaceHighway} + 2.758 * \text{SoilTypeClay} + 3.289 \\ &* \text{SoilTypeLoam} + 1.374 * \text{SoilTypeRock} + 0.572 \\ &* \text{CorrosivityConcreteHigh} + 0.459 * \text{CorrosivityConcreteLow} \\ &+ 0.114 * \text{CorrosivitySteelHigh} \end{aligned}$$



## MLR Model (*Cont'd*)

$$g_3(x) = \ln \left[ \frac{\text{Pr}(C=3)}{\text{Pr}(C=5)} \right] = 0.991 * \text{Diameter} + 0.974 * \text{Age} + 0.922 * \text{Slope} + 0.924 * \text{Depth} + 1.00 * \text{Length} + 1.532 * \text{pH} + 2.090 * \text{MaterialCONC} + 0.518 * \text{MaterialPVC} + 0.660 * \text{SurfaceEasement} + 0.998 * \text{SurfaceHighway} + 1.163 * \text{SoilTypeClay} + 2.168 * \text{SoilTypeLoam} + 1.335 * \text{SoilTypeRock} + 0.507 * \text{CorrosivityConcreteHigh} + 1.029 * \text{CorrosivityConcreteLow} + 2.731 * \text{CorrosivitySteelHigh}$$



## MLR Model (*Cont'd*)

$$\begin{aligned} g_4(x) &= \ln \left[ \frac{\text{Pr}(C = 4)}{\text{Pr}(C = 5)} \right] \\ &= 0.951 * \text{Diameter} + 0.998 * \text{Age} + 0.853 * \text{Slope} + 0.925 * \text{Depth} \\ &+ 1.00 * \text{Length} + 0.663 * \text{pH} + 0.812 * \text{MaterialCONC} + 0.489 \\ &* \text{MaterialPVC} + 1.073 * \text{SurfaceEasement} + 1.503 \\ &* \text{SurfaceHighway} + 0.134 * \text{SoilTypeClay} + 1.298 * \text{SoilTypeLoam} \\ &+ 1.223 * \text{SoilTypeRock} + 0.843 * \text{CorrosivityConcreteHigh} \\ &+ 10.377 * \text{CorrosivityConcreteLow} + 0.719 * \text{CorrosivitySteelHigh} \end{aligned}$$



## MLR Model (*Cont'd*)

- The highest probability value is taken as the predicted respective sewer pipe condition

- $$\Pr(C = 1|x) = \frac{e^{g_1(x)}}{1 + e^{g_1(x)} + e^{g_2(x)} + e^{g_3(x)} + e^{g_4(x)}}$$
- $$\Pr(C = 2|x) = \frac{e^{g_2(x)}}{1 + e^{g_1(x)} + e^{g_2(x)} + e^{g_3(x)} + e^{g_4(x)}}$$
- $$\Pr(C = 3|x) = \frac{e^{g_3(x)}}{1 + e^{g_1(x)} + e^{g_2(x)} + e^{g_3(x)} + e^{g_4(x)}}$$
- $$\Pr(C = 5|x) = \frac{1}{1 + e^{g_1(x)} + e^{g_2(x)} + e^{g_3(x)} + e^{g_4(x)}}$$

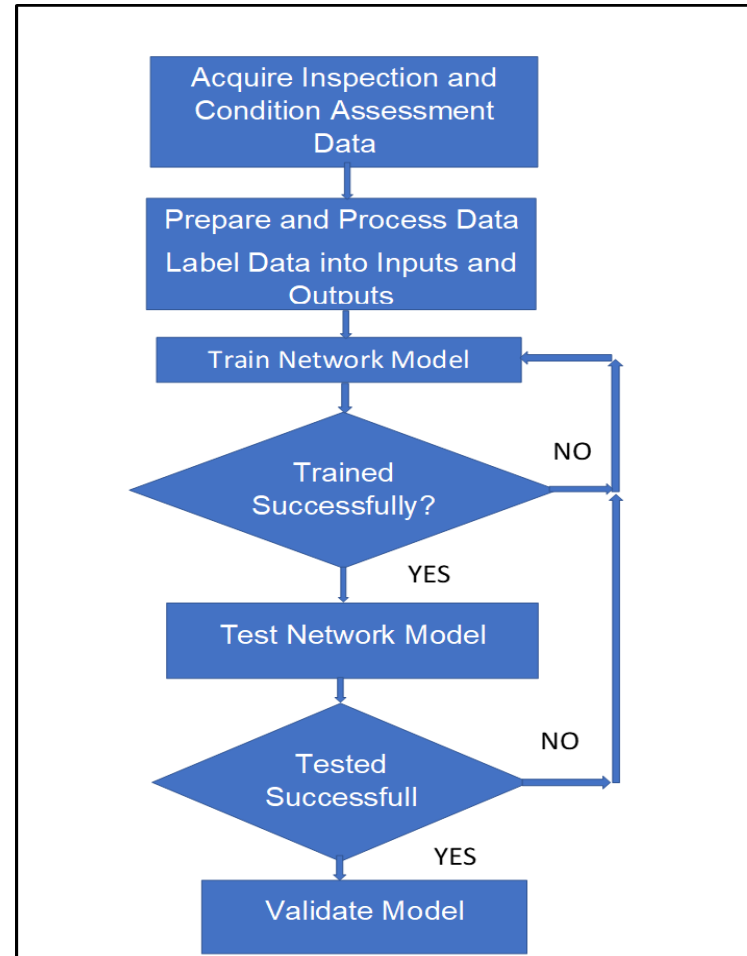


## MLR Model (Classification Table)

Observed Pipe Condition	Predicted Pipe Condition											
	1		2		3		4		5		Percent	
	Correct	Incorrect	Correct	Incorrect	Correct	Incorrect	Correct	Incorrect	Correct	Incorrect	Correct	Incorrect
1	1,846	--	--	--	--	44	--	--	--	11	97%	3%
2	--	104	0	--	--	2	--	--	--	2	0%	100%
3	--	233	--	--	93	--	--	--	--	10	28%	72%
4	--	66	--	--	--	4	3	--	--	7	4%	96%
5	--	138	--	--	--	27	--	--	26	--	14%	86%
Overall Percentage											75%	25%



## ANN Model Development





## ANN Model (*Cont'd*)

- Datasets randomly divided : Training (70%), and Testing (30%) (IBM SPSS Neural network Software)
- Training (85%), and Testing (15%) (Brain Maker Neural network Software, California Scientific Software)
- Backpropagation algorithm was used in training the ANN model



## ANN Model (Cont'd)

Data	Data Evaluation	N	Percent
Sample	Training	1,850	70%
	Testing	764	30%
Valid	--	2,614	100%
Excluded	--	2	--
Total	--	2,616	--



## Database in Netmaker

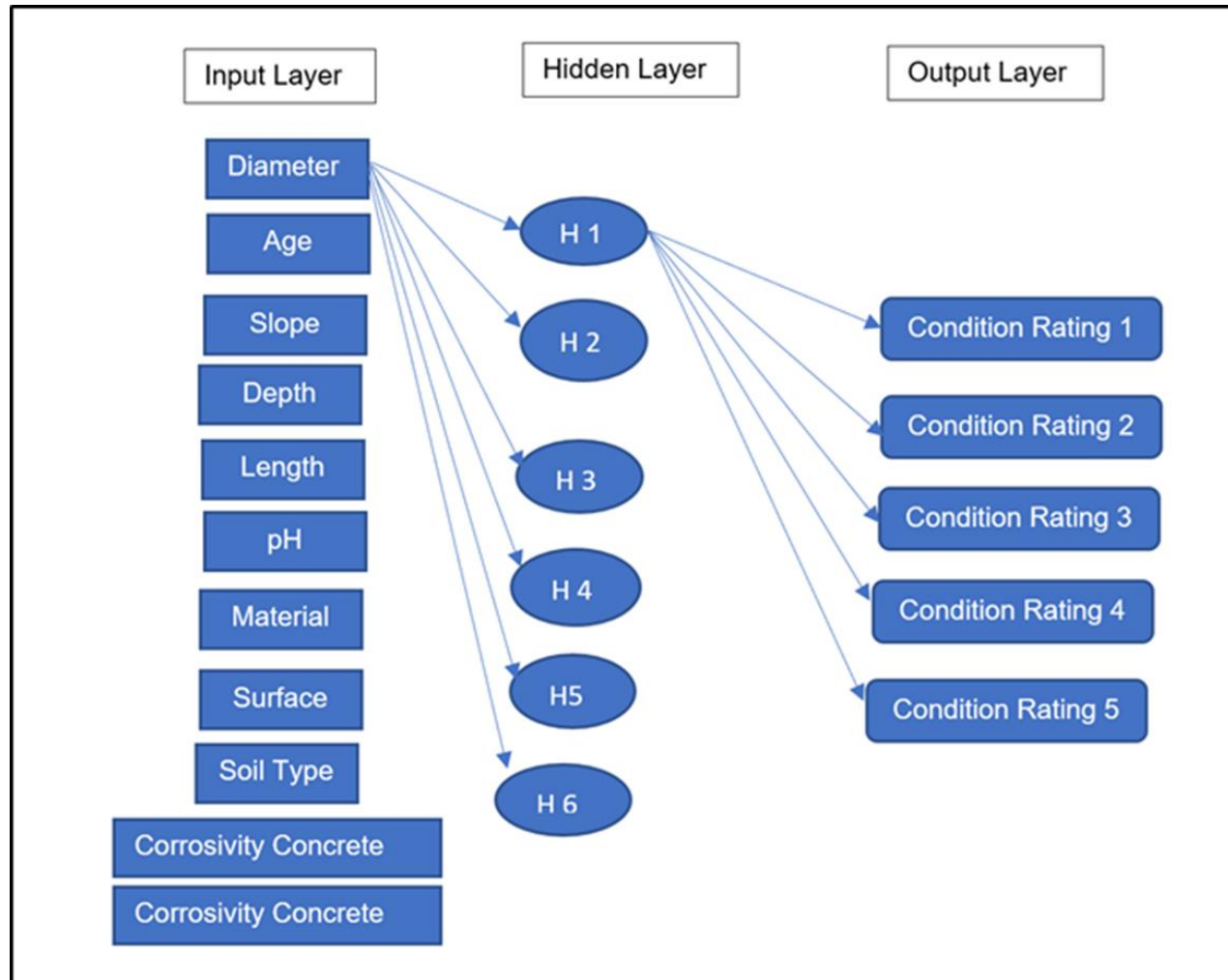
NetMaker - Sewer2.Dat

File Column Row Label Number Symbol Operate

	Annote	Input	Input	Input	Input	Input	Input	Pattern	Input	Input	Input	Input	Input
	ID	Diameter	Age	Slope	Depth	Length	pH	Condition	Material	Surface	SoilType	CorrConc	CorrSteel
1	318	10	59	0.1	6	575.35	7.5	1	Mater3	Surfa4	SoilT3	CorrC3	CorrS2
2	3255	8	22	0.5	7	154.1	7.9	1	Mater2	Surfa4	SoilT3	CorrC1	CorrS2
3	930	8	14	0.61	15	581.22	7.9	1	Mater2	Surfa2	SoilT3	CorrC1	CorrS2
4	1988	12	65	0.5	5	365.97	7.9	3	Mater3	Surfa3	SoilT3	CorrC1	CorrS2
5	2171	8	18	2.3	8	504.53	8.2	1	Mater2	Surfa2	SoilT4	CorrC1	CorrS2
6	2596	8	65	3.6	8	78.94	7.9	1	Mater1	Surfa4	SoilT3	CorrC1	CorrS2
7	1034	10	14	0.25	7	238.38	6.5	1	Mater2	Surfa4	SoilT1	CorrC2	CorrS1
8	1760	8	20	0.6	7	118.03	7.9	2	Mater1	Surfa2	SoilT3	CorrC1	CorrS2
9	1615	8	24	7	6	98.79	8.2	1	Mater2	Surfa2	SoilT3	CorrC1	CorrS2
10	3343	8	13	573	10	375.7	8.2	1	Mater2	Surfa3	SoilT3	CorrC1	CorrS2
11	1597	24	8	0.3	8	551.62	7.9	4	Mater1	Surfa2	SoilT3	CorrC1	CorrS2
12	2511	10	27	1.8	5	129.95	8.2	1	Mater2	Surfa4	SoilT4	CorrC1	CorrS2
13	385	10	60	0.3	7	311.56	8.2	1	Mater1	Surfa2	SoilT3	CorrC1	CorrS2
14	448	15	42	0.54	10	296.51	8.2	1	Mater3	Surfa2	SoilT2	CorrC1	CorrS2
15	3359	54	65	0.2	10	555.85	8.2	3	Mater1	Surfa2	SoilT2	CorrC1	CorrS2
16	258	12	57	0.01	5	372.85	8.2	1	Mater3	Surfa4	SoilT3	CorrC1	CorrS2
17	3188	15	39	0.1	10	752.5	7.9	1	Mater2	Surfa2	SoilT3	CorrC1	CorrS2
18	297	42	57	0.14	8	930.82	8.2	5	Mater1	Surfa4	SoilT2	CorrC1	CorrS2
19	83	8	39	0.4	5	673.57	6.8	1	Mater2	Surfa4	SoilT1	CorrC2	CorrS2
20	1207	8	22	3.8	7	108.3	8.2	1	Mater2	Surfa3	SoilT3	CorrC1	CorrS2
21	----	-	--	--	-	----	--	-	--	-	----	-	--



## Model Network Diagram



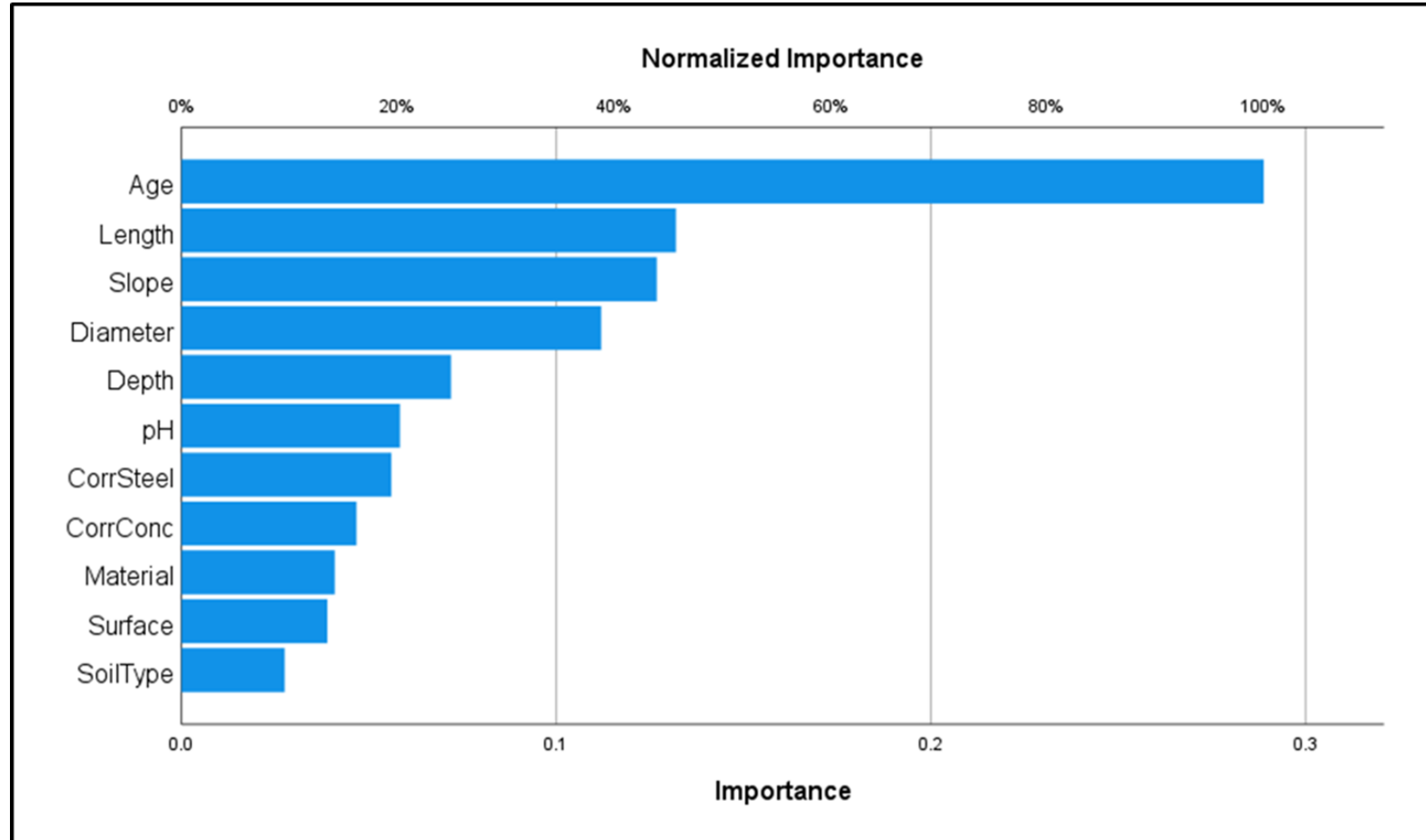


## Network Model Performance

Total Facts	Good	Bad	Tolerance	Average Error	RMS Error
Training Configuration					
2224	1599(72%)	625 (28%)	0.3	0.2519	0.3048
Testing Configuration					
392	334(85%)	58 (15%)	0.3	0.227	0.2823

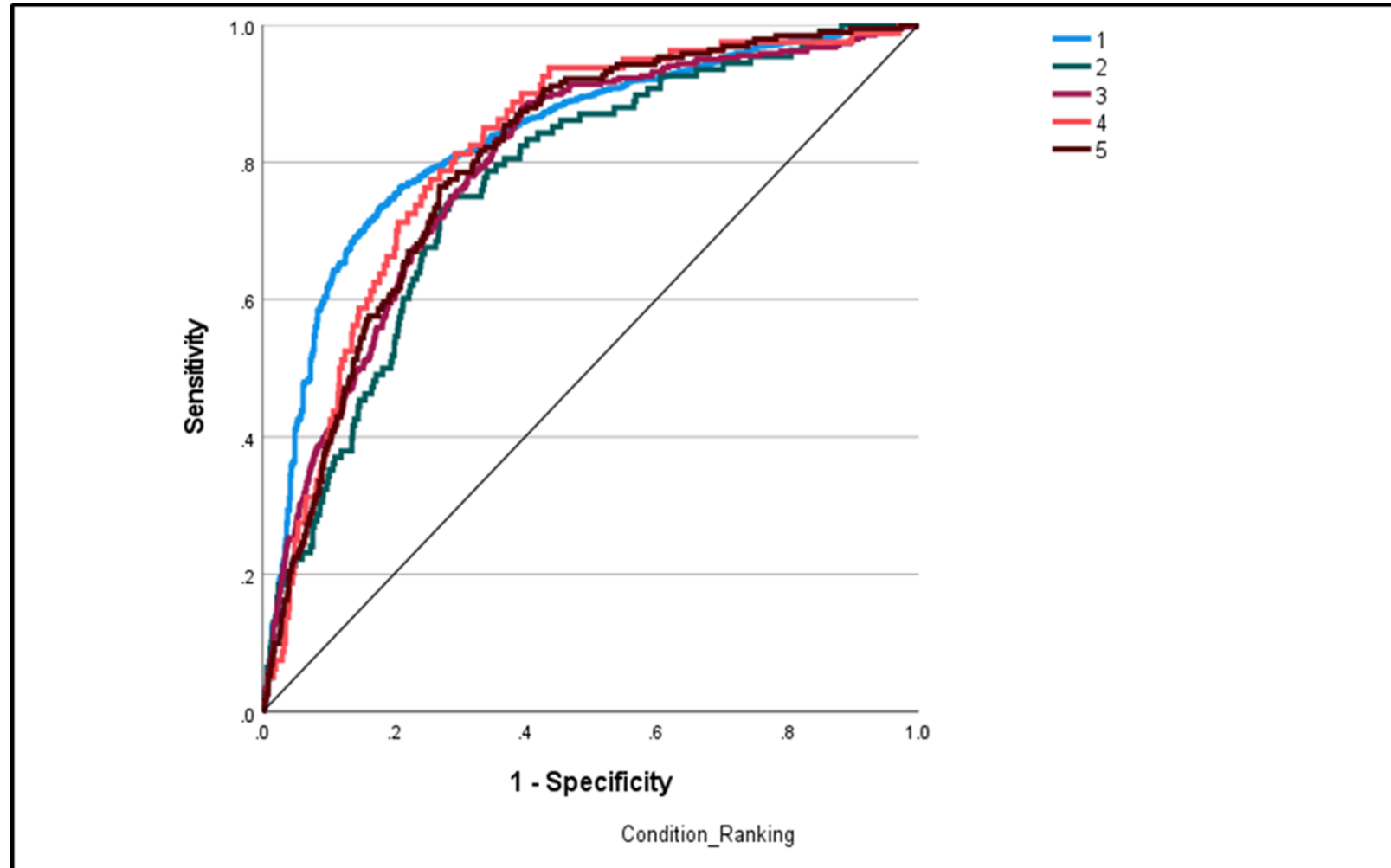


## Independent Variable Importance





## Receiver Operating Characteristic (ROC) Curve





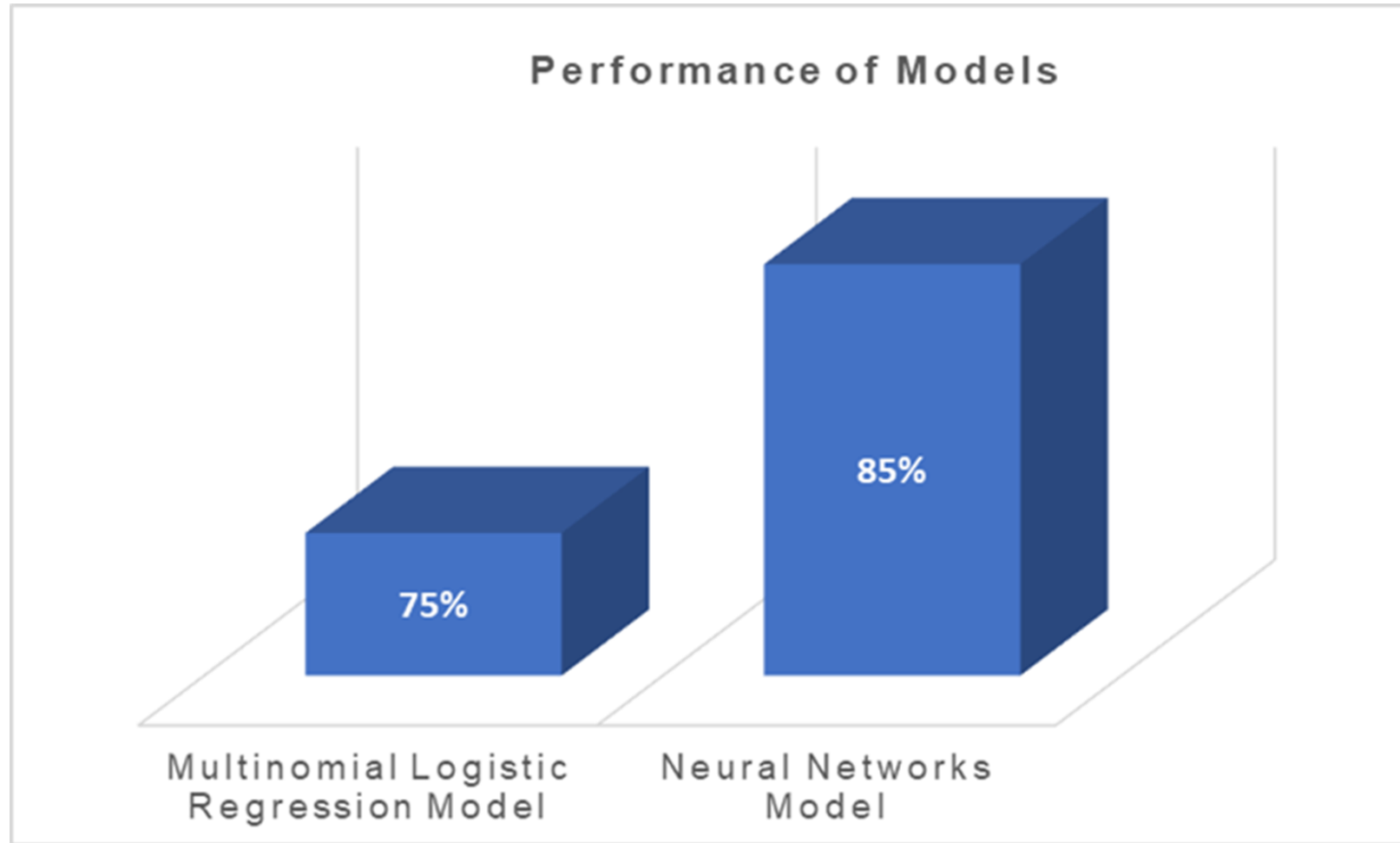
## Area Under Curve

Condition	Area Under Curve
1	0.833
2	0.768
3	0.794
4	0.815
5	0.802

- Area under curve demonstrates Model Performance
- Area close to 1 = Perfect Model
- Area  $>0.7$  demonstrates acceptable model



## Results and Discussions





## Results and Discussions (*cont'd*)

- Influencing Variables
  - Diameter
  - Age
  - Length
  - Pipe material (CONC)
  - Soil type (Loam)
  - Soil type (Clay)
  - Corrosivity concrete (High)
  - Corrosivity concrete (Low)



## Results and Discussions (*cont'd*)

- Non-influencing Variables
  - Depth
  - Slope
  - pH
- Influencing and non-influencing Variables were determined by significance Value ( $P < 0.05$ ) based on 95% Confidence Level.



## Validation of Results

Model	Author	Prediction Accuracy	Dissertation Results	Deviation
Multinomial Logistic Regression	Salman and Salem (2012)	52%	75%	23%
	Malek (2019)	65%		10%
	Laakaso et al. (2018)	62%		13%
	Sousa et al., (2014)	65%		10%
	Chughtai and Zayed (2008)	72%		3%
	Khudair et al., (2019)	55%-90.9%		15.9%-20%
Artificial Neural Network	Sousa et al., (2014)	72%-82%	85%	3% -13%
	Kulandaivel (2004)	84%		1%
	Khudair et al., (2019)	70%-93.6%		8.6%-15%



## Conclusions

- MLR and ANN Models were developed, validated, and tested in the prediction sewer pipe condition scores to prioritize pipes to be rehabilitated and or replaced and or further condition assessment.
- The Accuracy of Performance of the Models

Model	Accuracy
MLR	75%
ANN	85%

- Importance of independent variables: Age (100%), Diameter (80%), Slope (62%), Length (62%), Flow (60%), pH (40%), Corrosivity Steel (38%), Soil Type (36%), Depth (35%), Pipe Material (25%), Surface Condition (22%), and Corrosivity (20%).
- Significant factors influencing Sewer Pipe Condition rating Score
  - Diameter (95% Significance P Value = 0.001<0.05)
  - Age (95% Significance P Value = 0.000<0.05)
  - Length (95% Significance P Value = 0.019<0.05)



## Practical Implications

- Prediction models may be **instrumental** to sanitary sewer utilities managers in the **decision-making process in rehabilitation and replacement** of sewer pipes.
- Sanitary sewer condition assessment and data collection through **CCTV inspection** can be **costly**.
- Due to **inaccessibility** and **inadequate funding**, only about one third of sanitary sewer system are inspected every 5 years.
- Prediction models can **assist in expediting the evaluation of the condition rating** of sewer pipes using independent variables.
- City Engineers can use existing data and use one of the models to predict the condition of sewer pipes underground.
- Using the existing data, the MLR and ANN models can predict the sewer pipes conditions 1 through 5.



## Recommendations for Future Research

- The data for this dissertation was collected from the City of Dallas. Other cities should be included in future studies and results compared with results of this research.
- Excluded pipe material types could be included for further model development.
- The results of this dissertation can be further validated with future pipeline inspection data.
- Data collected for analysis should include more uniformly distributed number of observations in every condition.



## Q&A