RT&A, Special Issue № 1 (60)
Volume 16, Janyary 2021

COMPARISON OF CRASH PREDICTION MODELS USING MLR AND ANN

Aanal Desai, Dr. L. B. Zala, Amit A. Amin

•

Birla Vishvakarma Mahavidyalaya
Vallabh Vidyanagar, Gujarat, India
aanaldesai735@gmail.com

Abstract

Recently, in crash analysis statistical tools greatly helped in predicting future consequences based on recent data by studying the influence factors in the form of independent variables which can have significant consequences on the dependent variable. Hence, the forecasting of crash prediction models has to be developed. In this study, the stretch of National Expressway-1 has been selected from Ahmedabad to Vadodara because around 1058 crashes and 50 deaths occurred every year. Traffic volume, Speed, and Road characteristics were used to develop the models. Two approaches were chosen for the prediction model namely, the Multiple Linear Regression (MLR) model and Artificial Neural Network (ANN) model. Both methods are used to determine the relationship between the influencing factors of crashes and the frequency of crash occurrence, compared and discovered that the ANN model gives much better results for the prediction of road crash than the MLR model in this study.

Keywords: Crash Prediction Model, Multiple Linear Regression, Artificial Neural Network

I. Introduction

It is estimated that in the World, approximately 1.35 million people died in 2016 and 55 million injured every year due to road crashes [1]. Road traffic injuries cause considerable economic losses to people, their families, and nations as a whole. The safety scenario related to road crashes in India is poor because of the mixed traffic condition. Around 0.15 million of people were killed annually in India due to road crashes [2]. So, road safety is a major concern in the current situation and becomes an issue of national concern. Generally, the crash prediction model was developed by using different statistical techniques in both developed and developing countries. In general, the model development the process is iterative because in this process many models are derived, tested, and built upon until a model fitted in the desired criteria. Also, this process involved several tasks: structural model definition (e.g. linear, exponential, etc.), order (power) evaluation, and parameter estimation. Here, the crash prediction models were developed by using Multiple Linear Regression (MLR) and Artificial Neural Network (ANN). The prediction performance of both models presented in this study by comparing the predicted crashes versus observed crashes for each model.

II. Background

Ahmad, Erwan Sanik [3], studied accident prediction models comparison. The locations were selected in a rural area in Malaysia. Traffic volume, speed, number of access points and gaps data were used to develop the models. These parameters were then used to develop an accident prediction model by using the Artificial Neural Network (ANN) and Multiple Linear Regression (MLR) models. It was found that the MLR gave the best of R2 which was 99.92% validation for the model. Meanwhile, the ANN gave 82.40% which was lower than MLR. So, this was proved that the MLR model was the better model than ANN for this study. Jadaan, Fayyad [4], developed a traffic accident prediction model that was developed using the novel Artificial Neural Network (ANN) simulation to identify its suitability for prediction of traffic crashes under Jordanian conditions. Training, validation and, testing of the network were performed using MATLAB. Four alternative models, with a different number of hidden layers, were considered and Model 4 was found to be the best model with the highest coefficient of determination (R2 =0.992). The model was validated and found to produce good results under Jordanian traffic conditions thus can be used with confidence to predict future traffic crashes on the national road network. Nivea John, Archana [5], in their study, ANN were developed in Python with Keras library. The factors considered in the model are crash data, speed, volume, landuse type, pavement width and condition, shoulder width, number of horizontal curves, vertical curves, intersections, and bus stops. Results show that estimated traffic crashes, based on the input data are close enough to actual road crashes hence it is reliable to predict future crashes in two-lane undivided state highways. The performance of ANN is found to be better than other statistical methods. R.R. Dinu, A. Veeraragavn [6], developed accident prediction models were developed for day-time and night-time crashes over a three-year accident history of nearly 200 km of two-lane undivided highway segment around the city of Chennai in India. The explanatory variables considered for modeling included hourly traffic volume, length of highway segment, the proportion of buses, cars, motorized two-wheeler and trucks in the traffic, driveway density, shoulder width, and horizontal and vertical curvatures. The model coefficient was assumed to be normally distributed and a simulation-based maximum likelihood method was used for parameter estimation. This model helps to safety engineers in computing more realistic ranges of safety benefits of remedial measures. Williams Ackaah, Mohammed [7], Salifu Statistical models have been developed to predict crashes on rural highways in the Ashanti Region of Ghana. Generalized Linear Model (GLM) with Negative Binomial (NB) error structure was used to estimate the model parameters. Two types of models, the 'core' model which included key traffic exposure variables only, and the 'full' model which included a wider range of variables were developed. From the model developed, traffic flow, segment length, junction density, terrain type and presence of village settlement within road links are significant explanatory variables that influence the prediction of injury crashes on the rural highways in the Ashanti Region of Ghana. The study proved that passing a highway through settlement areas, traffic volume, provision of a long straight and flat sections and increasing number of junctions per unit road length tend to worsen road traffic safety. B. Rejoice, L. B. Zala, Amit Amin [8], in this paper they developed the accident prediction model by considering the generalized linear modeling technique. The data used for model development were speed, traffic flow, and road characteristics data and also, they used association rule mining techniques for the identification of accident spots as it can deal with the heterogeneous nature of crashes and help to improve road safety on rural highways.

III. Objective and Scope of the Study

With the help of recent data and statistical tools for analysis of crashes, road crashes analyzed, and predicted. With proper engineering measures, similar crashes can be reduced from happening in the future. In this study two approaches MLR and ANN were chosen for the crash prediction model development. The objective of study is comparison of these two statistical approaches and to determine which is more effective approach for crash prediction in this study.

IV. Study Area and Data Collection

The selected study stretch was Mahatma Gandhi Expressway (NE-1) connecting the cities of Ahmedabad and Vadodara in the state of Gujarat, India. The total length 93.1 km. was selected which is a four-lane divided expressway. For model development variables considered were: Traffic Volume in terms of Hourly Traffic Volume (HTV), Average Speed, Road Condition in terms of straight-section and curve-section and, Carriageway Width as independent variables, and Number of road crashes per hour (average of 3 years i.e., 2016-2018) as a dependent variable. Table 1 shows the descriptive statistics of variables.

Table 1: *Descriptive Statistics*

Variable Code	Variable Description (per hour data)	Minimum	Maximum
HTV	Hourly Traffic Volume (HTV) [in Log Natural (ln) form]	7.27	**8.17**
S	Speed (kmph)	60	**80**
CW	Carriageway Width (m)	11	**11**
R	Road Condition*	1	**2**
Rc	**Road Crash (per hour in a year)**	**28**	**62**

*Subjective rating: 1-Straight section, 2-Curve section

The models developed by considering hourly crashes to find out how traffic volume, average speed and carriageway width will affect road crashes. Generally, the research work carried out as a road crash per km but in this study does not give meaningful image because expressway has limited access points therefore classified volume count and speed for segments are the same. Hence in this research work crashes are considered as per hour based because traffic volume and vehicle speed considerably vary with per hour basis rather than per km based. So, all the variables are converted into hourly basis which is shown in Table 2.

Table 2: *Input Data for Model Development and Validation*

Time (hr)	Ln(HTV)	Speed (kmph)	Carriageway Width (m)	RC*	Road Crash in Hours
0-1	7.493	75	11	1	**40**
1-2	7.311	69	11	1	**44**
2-3	7.394	70	11	1	**42**
3-4	7.277	69	11	2	**39**
4-5	7.422	70	11	1	**41**

5-6	7.585	70	11	1	**39**
6-7	7.729	80	11	2	**62**
7-8	7.953	75	11	1	**51**
8-9	8.040	72	11	2	**52**
9-10	8.176	80	11	1	**60**
10-11	8.090	70	11	2	**51**
11-12	7.936	60	11	1	**42**
12-13	7.969	62	11	1	**38**
13-14	7.996	70	11	1	**39**
14-15	7.990	65	11	1	**47**
15-16	8.054	80	11	2	**56**
16-17	7.963	77	11	1	**54**
17-18	7.967	72	11	1	**43**
18-19	8.116	70	11	1	**42**
19-20	8.092	65	11	1	**34**
20-21	7.974	70	11	1	**30**
21-22	7.841	70	11	1	**33**
22-23	7.726	60	11	2	**28**
23-24	**7.698**	**75**	**11**	**1**	**40**

V. Multiple Linear Regression Analysis

Early models of crash prediction are based on the simple multiple linear regression approach which has more than one explanatory variable X and a scalar dependent variable denoted as Y. The model must be linear and the general form of the linear crash prediction model can be expressed as equation (1).

$$Y/\theta \sim Dist(\theta) \text{ with } \theta = f(X, \beta, \varepsilon) \quad (1)$$

Where,

Y: the dependent variable (i.e. crash frequency),

θ: the crash dataset,

Dist(θ): the model distribution,

X: a vector representing different independent variables (i.e. risk factors),

β: a vector of regression coefficients,

f(.): link function that relates X and Y together,

A similar linear relationship was adopted as per equation (2).

$$R_C = \beta 0 + \beta 1 (\ln(HTV)) + \beta 2 (S) + \beta 3 (R) + \beta 4 (CW) \quad (2)$$

Where,

R_C = Road Crashes,

HTV = Hourly Traffic Volume,

S =Average Speed,

R = Road Condition,

CW = Carriageway Width

The linear regression model developed in Microsoft Excel and the result shown below in Table 3 and The R square value for the model sounds good. It explains that outputs are close to the linear trend line.

Rc (per hour in a year) = 6.954 ln(HTV)+1.036 (Speed) - 8.104 (CW) + 4.125 (RC) (3)

R^2 = 0.984

Table 3: *Regression coefficient of the developed model*

	Coefficients	t	P-value
Intercept	0.000	-	-
Volume	6.954	1.500	0.149
Speed	1.036	4.577	0.002
Carriageway width	-8.104	-2.306	0.032
Road Condition	4.125	1.427	0.169

F-test was performed to find out that there is any variance within samples. It is the ratio of the variance of the two samples. The result is shown in Table 4. F-value is less than F-critical value at a 95% confidence interval, which means the developed model is acceptable. For the validation of the model, a comparison was carried out between Predicted crashes by multiple linear regression model using observed data on same facility and Observed crashes of the same data set used for the model development. Figure 2 shows the correlation between predicted road crashes by the Multiple Linear Regression model and observed road crashes and identified that the R^2 value for the linear regression model was 58.67%. Figure 1 shows the plot of the observed crash per hour and predicted crash by the model per hour.

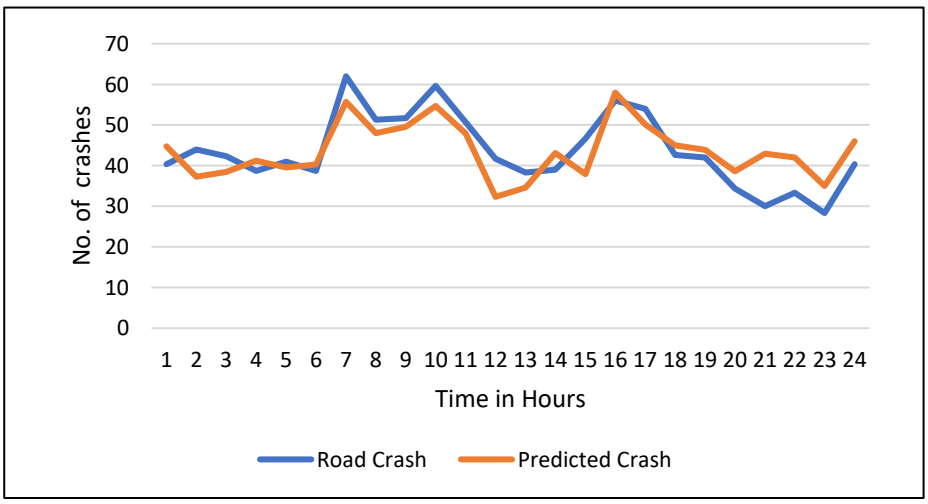


Figure 1: *Observed and Predicted road crashes*

Table 4: *F-test for two-sample for variables*

	Observed Crashes	Predicted Crashes
Mean	43.625	43.616
Variance	77.684	45.432
Observation	24	24
df	23	23
F	1.7099	
P (F<=f) one-tail	0.1029	
F Critical one-tail	2.0144	

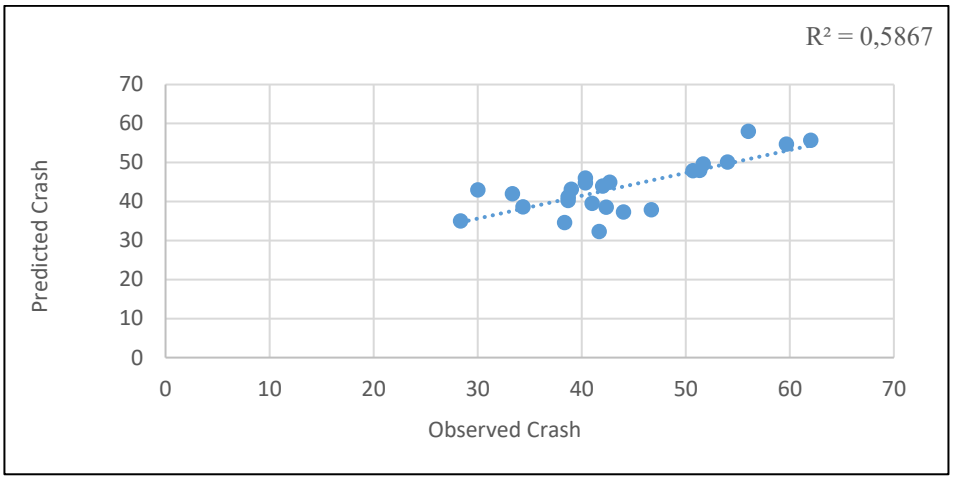


Figure 2: *Observed and Predicted Crashes (MLR Model)*

VI. Artificial Neural Network

Recently, ANN is widely used in solving engineering problems like classification, prediction, and function approximation. This neural network is similar to the human brain in which it composed several series of nodes and weights factors that connect the all nodes together in a systematic style that consists of the input layer, hidden layer, and output layer. This model is now used in road safety as predicting crashes and previous research shows that this model predicts complex observation more precisely than the traditional regression model. As compared with the regression model than ANNs do not form a functional relationship between the dependent variable and independent explanatory variables as established in other statistical approaches.

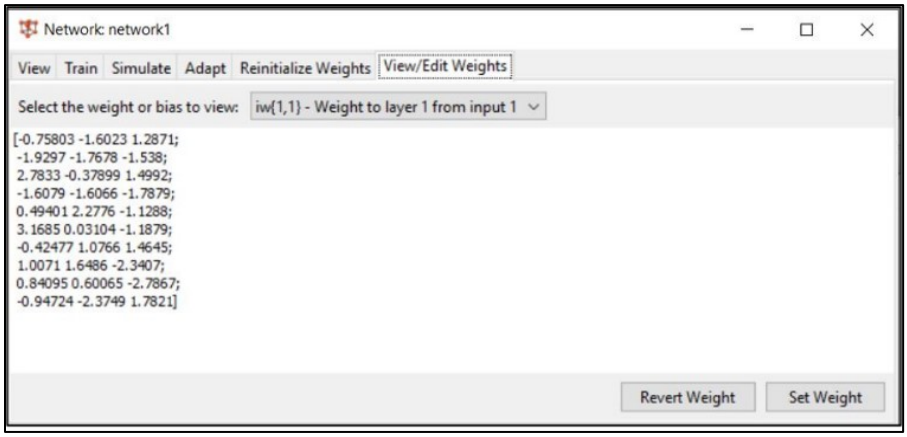
The key element of a neural network as follows:

1. Each neuron has a net-input value which is multiplied by the weights plus the bias in the input layer of the network and also all neurons are highly interconnected in the processing.

2. In the network, the input layer composed the data as is known as independent variables, one or more hidden layers which do the processing and the output layer contains the

observed or desired output value. Every layer of network consists of neurons that connected every other neuron in the previous layer by a link that represents the weights.

3. In the neural network, there is some transfer function which is known as Activation



Function that is used for mapping from input to hidden nodes and from hidden to output nodes respectively.

I. Development of ANN Prediction Model

This model was developed in MATLAB software. The steps for developing the ANN model for prediction of road crashes are as follows:

1. The data imported in a variable window with a title as Input and Target from the Excel sheet. With the use of the "nntool" command the neural network dialogue box will be opened.

2. Also, the data set was divided into three sets, training data (about 70% of the total data set), validation data (about 15% of the total data set), and testing data (about 15% of the total data set).

3. The activation function used for developing model were:
   Input hidden layer: Tan-sigmoid transfer function
   Output hidden layer: Linear transfer function

4. The network is created with 3 hidden layers and 10 neurons. Then setting an initial value for weights and evaluating the output and also measured the errors. The weights continuously to be modified until a minimum error is computed. These are all done in a Training phase shown in figure 3.

Figure 3: *Image of Weight modification screen in MATLAB software*

5. After the training phase, the validation of the network model is carried out by comparing the predicted value with the observed value. In this phase, there is no adjustment carried out in the weights, and this process is carried out with a training process to improve the performance of the model.

6. And the last procedure is the testing phase in which the predicted values are compared with the input values using data that was not used in the training or validation process. Again, no adjustment occurs to the weights.

7. The figure of the Artificial Neural Network model is shown in figure 4.

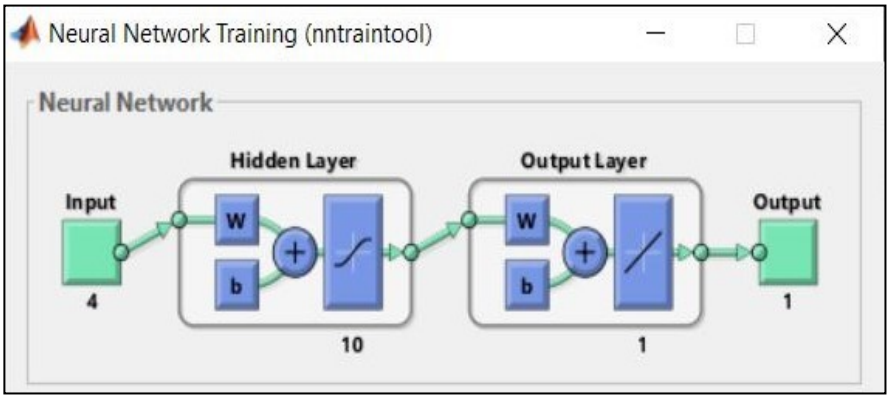


Figure 4: *Architecture of ANN model*

The network also gives the regression graph and R-value for different phases of the model means training phase, validation phase, testing phase. Here figure 5 shows all regression graphs with their R-value. The results found very satisfactory with the value of coefficient of determination (R square 0.8896). Also, F-test was performed between predicted and the observed value of road crashes and the result shown in table 4. Figure 6 and 7 shows the relation between predicted road crashes by the ANN model and observed road crashes per hour with an R^2 value of 88.79%.

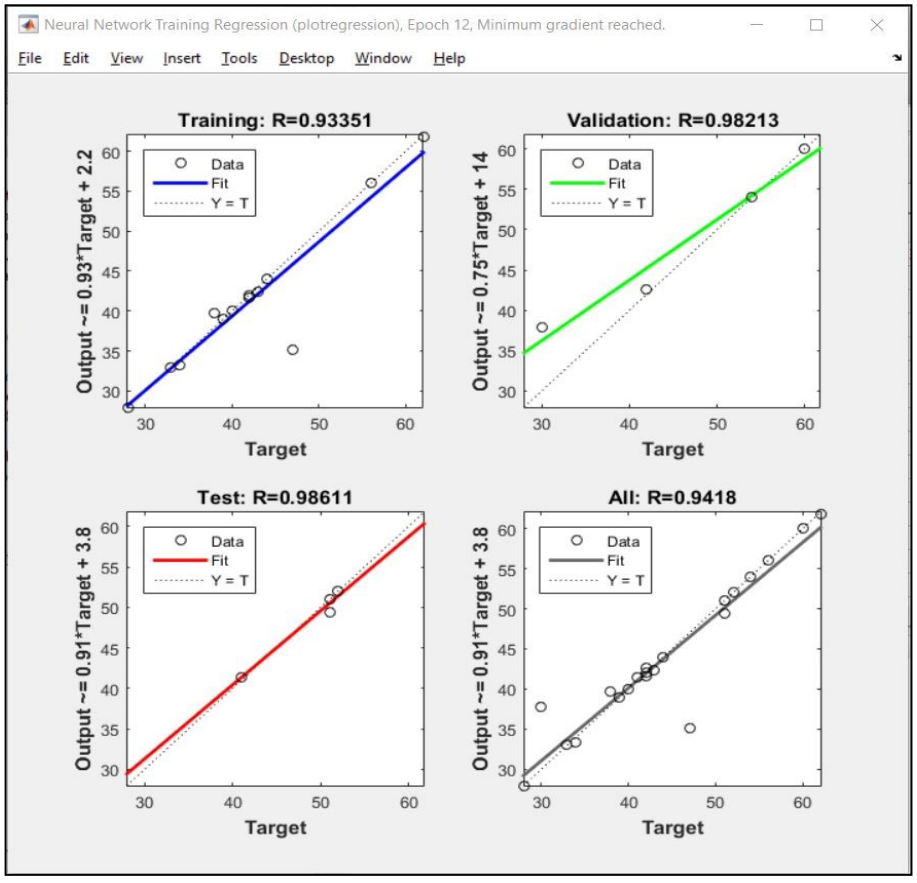


Figure 5: *Correlation between Observed and Predicted Accident Rate (ANN Model)*

Table 5: *F-test for two-sample for variables*

	Observed Crashes	Predicted Crashes
Mean	43.625	43.426
Variance	77.684	73.926
Observation	24	24
Df	23	23
F	1.0508	
P (F<=f) one-tail	0.4532	
F Critical one-tail	2.0144	

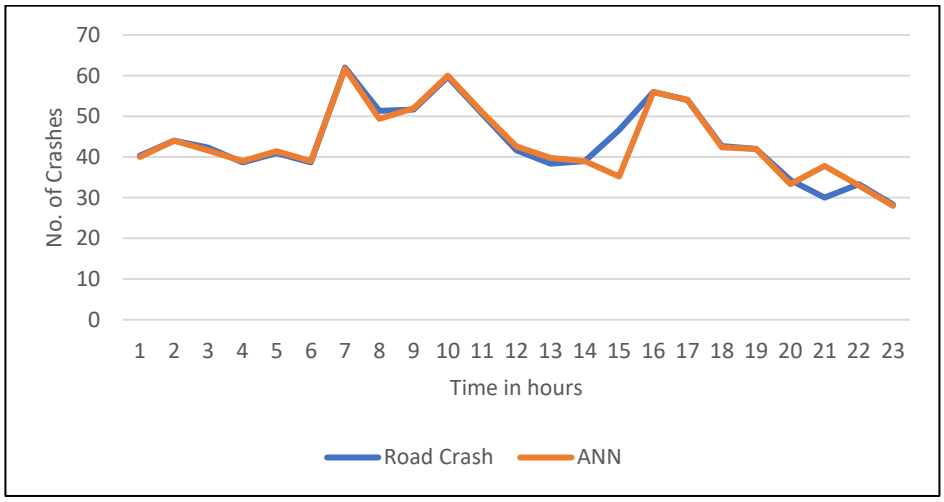


Figure 6: *Observed and Predicted Road Crashes by Artificial Neural Network*

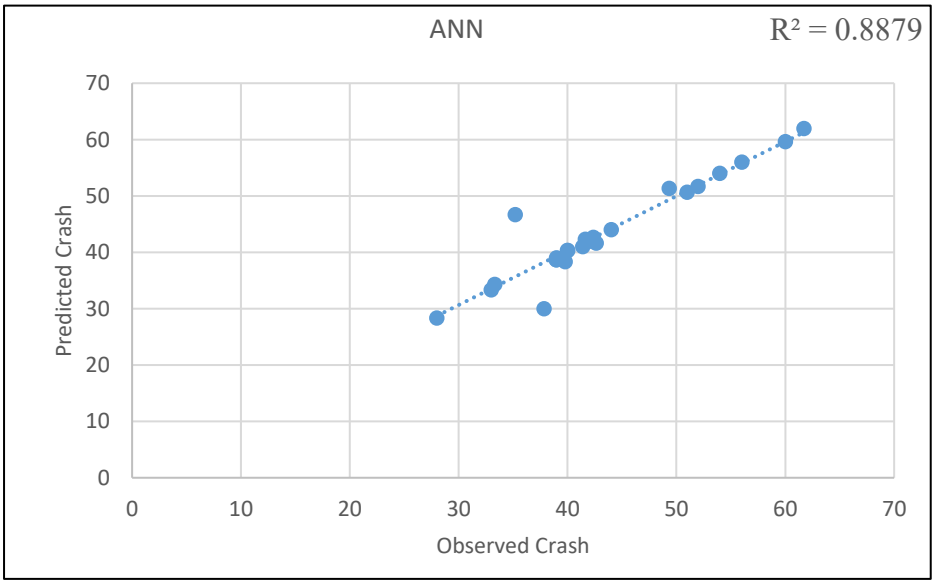


Figure 7: *ANN Model Validation Graph In between Observed and Predicted Accident Crashes*

VII. Conclusion

The study was conducted with an objective of comparison between two models and determined which model give better result. From this research, it can be concluded entirely that:

The comparison between of these models was carried out by validation which is the plotting graph between predicted crashes by the models and observed crash, it was discovered that the ANN gave the better result. The R^2 value of ANN model is 88.79%. Meanwhile, the MLR model gave 58.67% which is lower than ANN model. Clearly, the result shows that the ANN model gives much better result compared to the Multiple Linear Regression model for the prediction of road crashes in this study.

References

[1] Global status report on road safety 2018 – WHO www.who.int › violence_injury_prevention › 2018

[2] Road Crashes in India", Government of India, Ministry of Road Transport and Highways, Transportation Wing, New Delhi, 2018.

[3] Raqib, A., Ghani, A., Sanik, M. E., Aida, R., & Mokhtar, M. (2011). Comparison of Accident Prediction Model.

[4] Jadaan, K. S., Al-fayyad, M., & Gammoh, H. F. (2014). Prediction of Road Traffic Crashes in Jordan using Artificial Neural Network (ANN). 2(2), 92–94.

[5] John, N., & Archana, S. (2019). Crash Prediction Modeling of Two-Lane Undivided Highways Using Artificial Neural Network. 10(5), 326–329

[6] Dinu, R. R., & Veeraragavan, A. (2011). Random parameter models for accident prediction on two-lane undivided highways in India. Journal of Safety Research, 42(1), 39–42.

[7] Ackaah, W., & Salifu, M. (2011). Crash prediction model for two-lane rural highways in the Ashanti region of Ghana. IATSSR, 35(1), 34–40.

[8] Bhavsar, R., Amin, A., & Zala, L. (2020). Development of Model for Road Crashes and Identification of Accident Spots. International Journal of Intelligent Transportation Systems Research.