www.ijecs.in
International Journal Of Engineering And Computer Science
Volume 15 Issue 02 February 2026, Page No. 28156-28166
ISSN: 2319-7242 DOI: 10.18535/ijecs/v15i02.5390

Comparative Study of Road Transportation Prediction System Using MlPNN and SVM

Umejuru Daniel[1], Promise Enyindah[2]

[1, 2]Department of Computer Science, University of Port Harcourt, Choba, Nigeria
[1]https://orcid.org/0000-0001-6246-7077

Abstract
Vehicle-related Road accidents remain one of the most distressing experiences, often leaving lasting physical, emotional, and economic consequences for victims and society. Frequent road accidents and vehicle collisions are largely attributed to poor road conditions, environmental influences, and driver negligence. Predicting accident severity can assist safety agencies and transportation professionals in understanding key contributing factors and improving preventive strategies. Machine learning techniques provide an effective means of identifying accident patterns and estimating severity levels such as fatal, major, or minor injuries.

This study proposes a machine learning–based approach for predicting traffic accident severity using two algorithms: a multilayer perceptron neural network (MLPNN) and a support vector machine (SVM). Road traffic accident datasets sourced from the Kaggle repository were used for experimentation. The MLPNN model incorporated a decay-based regularization technique to constrain large weights, thereby reducing model complexity and improving generalization. Hyperparameters were carefully optimized to enhance classification performance. Experimental results showed that the MLPNN achieved a prediction accuracy of 98.88%, outperforming the SVM, which recorded an accuracy of 89.0%. Based on these findings, the MLPNN model is recommended for reliable traffic accident severity prediction and effective extraction of critical accident features.

Keywords: Traffic, Prediction, Machine learning Mlpnn, Svm,

1.0 Introduction

Road transportation systems require efficient and effective methods for managing and predicting traffic accidents. It can have a positive influence on the development of new roads, the improvement of safety, the creation of regulations, and the planning of routes to prevent road traffic incidents before they happen. Significant financial, social, and economic damages are brought about by collisions with vehicles for people, families, and the country at large. The frequency of road accidents is alarmingly rising as demonstrated by the World Health Organization (WHO) approximate death rate (Zhou 2019). About 1.2 million people lose their lives and fifty million more are injured every year (Kim et al.,2020). The nations with the highest death rates from traffic accidents per 100,000 people are Zimbabwe (61.90), Liberia, Malawi, Gambia, Togo, Tanzania, Rwanda, Sao Tome, Burkina Faso, and Burundi(Li et al.,2020). These nations exhibit a similar pattern. Traffic accidents remain a major cause of fatalities and serious injuries globally, and the general public continues to suffer from numerous terrible injuries even decades after the terrible event (Wang et al.,2020). There are a number of reasons why accidents happen, such as rollovers, hit-and-run incidents, head-on collisions, multiple vehicle pile-ups, rear-end crashes, side impact collisions, side swap collisions,

and drunk driving. There are two categories of accidents: injuries from major accidents and injuries from small accidents. A major accident injury occurs when there is at least one victim with a serious injury that necessitates hospitalization. The victim or victims have minor injuries from an accident when there is no need for hospitalization.

2.0 Conceptual Review

The huge number of automobiles on the road is a result of citizens owning an ever-increasing number of cars (Li, 2020). This leads to the advancement of road infrastructure and innovations in order to guarantee that transportation safety concerns are consistently given top priority by drivers (Kononen et al., 2021). One of the main causes of casualties is traffic accidents across the world. It has become increasingly important in the modern era since the majority of the poor and middle class use roads for mobility (Li et al., 2021). As a result, a key element needed in the layout of roads for the advancement of transportation is safety on the road. As opposed to where a single organization is meant to be in charge of enhancing road safety, there is a great deal of confusion, fragmentation, and lack of cooperation when it comes to road safety. It may have little to no interaction with the different other organizations that have the ability to impact the state of traffic safety and little to no implementation authority in other domains (Yu, et al., 2020). Only by taking concerted effort to close the gaps in each of the key areas that affect road safety will the issue be adequately addressed. As a result, a key element in creating a safe city is highway safety. This is due to the fact that road accidents are now among the top causes of fatalities and serious injuries (Chen, et al., 2019).

2.1 Accident prediction model

The total number of collisions on major highways, at points of intersection, and on all other structures connected with transportation can be predicted using incident prediction models. The utilization of accident counts is crucial because collisions with vehicles are typically statistically independent random events; that is, accident counts by themselves cannot be utilized for predicting the number of collisions on a particular transportation infrastructure (Hauer, 2019). Predictions concerning the connections between a variable that is dependent and a number of variables are used to create APMs. Predictive models often take the following shape.

$$E(x) = f(x, \beta) \quad 2.1$$

Where E(x) is the number of accidents per unit of time, x is a series of covariates; ranging from x_1, x_2, x_p while B is the coefficients to be estimated given as β_0, β_1, . . . β_p

Hauer (2019) used equation 2.1 in order to forecast the aggregate amount of accidents for each minute of time on a specific transportation facility. It is possible to make use of algorithms to predict accidents based on factors like accident type, severity, or timing. Estimating the coefficients, P, linked to the associated variables, also known as the explanatory variables, is the primary objective of equation 2.1. The methods for determining these values are highly sophisticated, and there are several publications pertaining to models that are not linear, the generalized linear model (GLM), and intra-linear models (Shankar et al., 2019). Generally speaking, GLM uses maximum probability or regression techniques to estimate the correlation coefficients of APMs. There are numerous independent variables that can be found on the right side of equation 2.1, f\), including traffic flow, sight distance, turning lanes, speed limit, road illumination, traffic control, etc. These variables are included in the equations that are suggested. However, the most widely used models often just include traffic flow as an input (Kulmala 2020). There are numerous varieties of modeling forms. The following two components provide some examples of modeling forms for APMs that are utilized for intersection and arterial road segments. These models are available in a variety of formats, from quite basic to highly intricate. Equations 2.2 and 2.3, for instance, provide two distinct kinds of models that forecast collisions on highways:

$$E\{k\} = \boldsymbol{\alpha} L L^{\beta_1} F^{\beta_2} \quad 2.2$$

$$E\{k\} = \boldsymbol{\alpha} L^{\beta_1} F^{\beta_2} e^{\sum_{i=3}^{p} x_i^1 \beta_i} \quad 2.3$$

Where E{k} is the expected number of accidents per unit of time, L: length of highway section, F=Traffic flow on high ways: $\boldsymbol{\alpha}$, $\beta_1, \beta_2, \beta_i$ are the coefficients to be estimated, $\sum_{i=3}^{F}$, $x_i^1 \beta_i$ are the series of independent variables such as sight distance, warning signs, private entrances etc. Fo i= 3 to p

Formula 2.3 presents a highly complex model form that includes many associated variables, whereas equation 2.2 displays a very basic model form. Hauer (2019) argues that when there are enough data points, it is preferable to divide the data into separate groups and build a model for every group as opposed to

building a single model with numerous variables. They contended that because the final result depends on the determinants concurrently, models containing variables that are categorical are typically rigid. They suggested building a set of models that solely required traffic flow as a parameter and grouping them into cells. This APM for signaled junctions with numerous covariates will serve as a demonstration of the proposed methodology.

3.0 Methodology Adopted

This research work will be achieved following the Object-Oriented Analysis and Design Methodology (OOADM). This is aimed at viewing, modeling and implementing the new system as a collection of interacting classes and objects. OOADM is adopted because it is more effective, efficient, reliable, reusable and a faster way of developing systems.

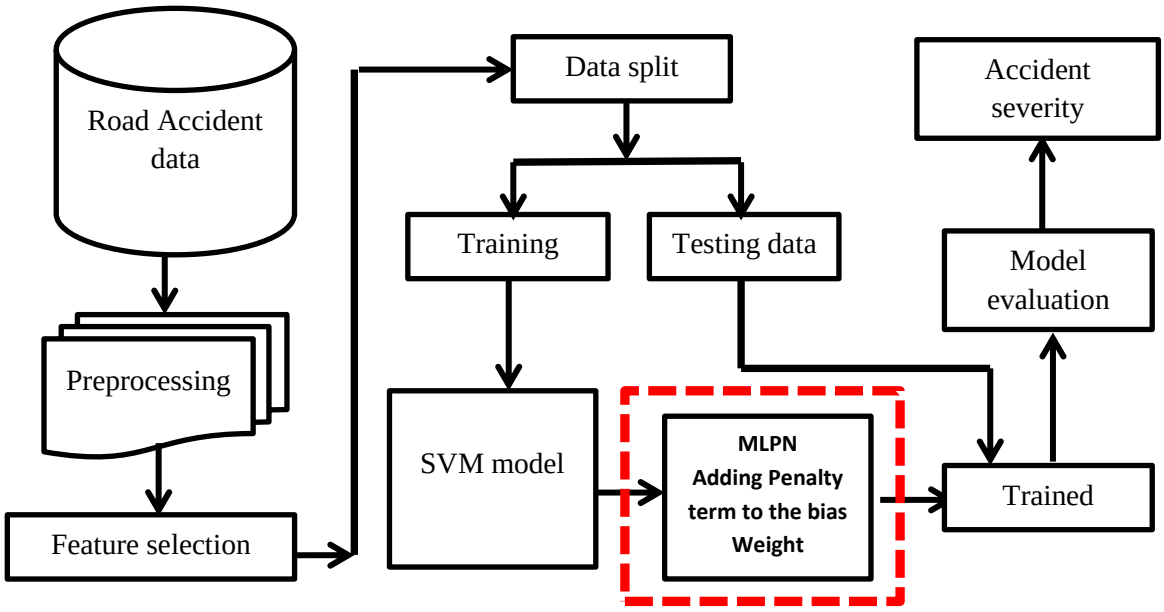

Figure 3.1: Proposed System Architecture

3.1 Explanation of Key Components

The accident data used in this study was obtained from the Kaggle website. A dependable framework is required to communicate with the systems and obtain significant data from outside sources. After that, data must be analyzed to produce information, enabling those responding to complete the difficult task with greater accuracy and precision. Despite our lack of comprehensive understanding of the road accident data collection policy, we adopt the assumption that self-selection bias may have an impact on certain parts of the input features. These input variables include the total amount of vehicles involved, the severity of the accident, the number of casualties, the junction data, and more.

We analyzed the architecture and synthesized the system model in the design in order to meet the fundamental system requirements. The system is made up of software architecture, procedural details (such as algorithms), and modules. One of most prevalent type identified in the medical field is stroke, which is increasing every single year, and we are using a publicly released stroke prediction dataset using a multilayer perceptron neural network (MLPNN) to predict the likelihood of brain stroke disease.

The Back-propagation Neural Network (BPNN) is adopted with multi-layered-**NN**, connected to different layers with no bypassing layers. The BPNN is used to map the network in computing functional relationship of input/output.

The input vector represented as X[a][i]. Where the disease symptom of X/Y/Z and i is the pattern matrixes of the array (i=0,1,2,3, - - - -,5111). The output target-variable is encoded using T[a][i].The learning network is accompanied by a Back Propagation NN with a rule in training the network, described as weight-vectors W_{ij} and W_{jk} and a threshold value. The network accepts input patterns and produces output. The NN uses the equation, given as:

$net_{ai} = \sum W_{ij} O_{ij}$ 3.1

A multilayer classifier model was imported from the sklearn.neural_network class as shown in line stamen 90. We defined the multilayer neural network model with a solver set to 'lbfgs', alpha set to 1e-5 expression for a period of 200 maximum iterations(epoch). The rectified linear unit (Relu) activation function is employed to help activate network neurons at different iterations. The ReLU activation function helps MLPNN solve the diminishing gradient dilemma, allowing the model to learn way quicker and perform better with 10-input, 30-hidden 1-output layers. The components of proposed system are as follows**:**

Feature Extraction: The feature selection process was adopted to determine the correlation between variable or attribute pars based on the level of correlation using a score value. The higher the score value the higher the correlation between attribute pairs. This was used in order to prioritize the features that have the greatest influence on model predictions.

Training and testing data split: Three primary subsets were created from data split: the training set, which is employed to train the algorithm; the validation set, which is used to monitor the parameter settings and prevent over-fitting; and the testing set, which is used to assess the effectiveness of the model on newly collected data. This is taking an 80% random sample of the rows (you can change this) and adding them to our set of training data without replacing them. We include the balance of twenty percent of the sample in our test set. Numerical data are converted into image sets for the purpose of training MLPN, which performs best with image datasets,

Multilayer neural network (MLPN): We are designing a MLPN model to assist with the difficult and time-consuming task of altering weights during each training cycle. The weights included in ordering of inputs to the MLPN constitute the factors that cause its weights to change. The neural net weights vary at each of the layers in addition to the sigma activating function. The activation processes changes with each subsequent cycle since they serve as the data inputs for the subsequent layers. The resulting shift in distribution requires each and every layer to adjust to the changing data inputs, and that is the reason why the deep learning duration for training increases

3.2 Database design and structure

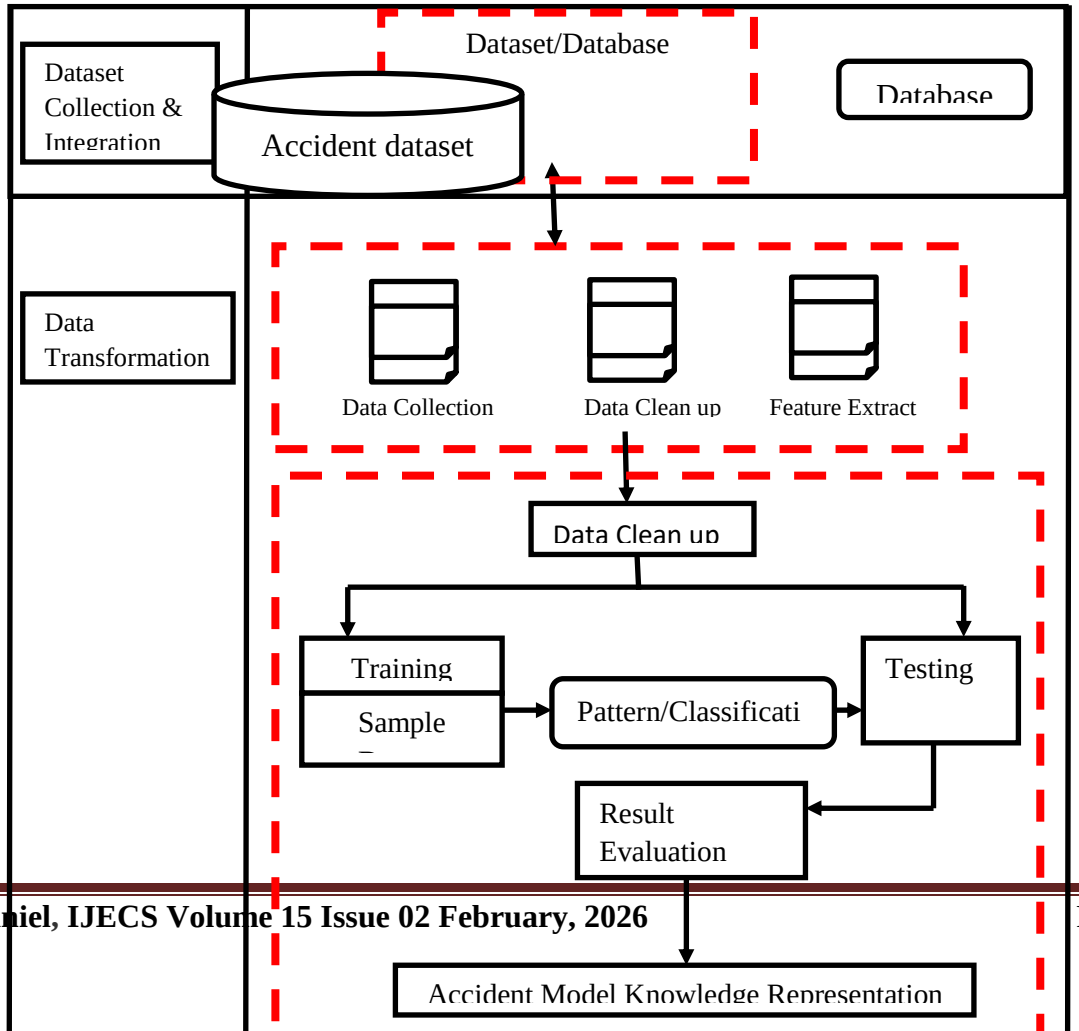

Figure 3.2: Framework of database system design and structure

3.2.1 Database Table definition

Table 4.1 depicts the data types, fields and description which are very important to define data types of a valid field name when designing a relational database. This will help match each of the key features with their data type and field metadata. The most crucial characteristic of a field is its data type, which establishes the kind of data that the field can contain. This is required in order to store and manage both structured and unstructured data, of the accident severity rate. This was utilized in this research for data management, analysis, and storage.

Table 3.1 User registration Table structure

Field	**Data type**	**Description**
UserId	Numeric	A unique user identification number.
Name	Text	User full name of driver
User_address	Text	Contact address of user
Phone	Numeric	The contact number of user
Gender	Boolean	The gender of Driver.
State_of_origin	Text	The state of origin
Passenger	Numeric	Total number of passenger vehicle can contain

Table 3.2 Accident registration Table

Field	**Data type**	**Description**
AcciId	Numeric	A unique user identification number.
AcciScene	Text	Location of accident
Date_of_accident	Date/time	Date and time of accident
Accid_seve	Numeric	Accident severity level
VehicleNo	Text	Vehicle plate umber
Source	Text	Initial take off point
Destinaton	Text	Vehicle Destination
Total_Passenger	Numeric	Total number of passengers
Drivers_name	Text	Full name of vehicle driver
Injured_case	Numeric	Total number of injured victims
Death_cases	Numeric	Total number of death cases

Table 3.3 program Module Specification.

Program module	**Specification**
Sample Dataset	This contains the raw data set as retrieved from Kaggle site
Pre-processing	This is the stage where the raw data is

	cleaned, normalized and transformed
Model Building	This module is where the individual different models are built
Model Training	In this subsystem, the training dataset was divided into k-folds using cross validation value of 5
Model Prediction	The prediction subsystem requires the use of testing set in order to determine accident severity level

4.0 Results and Discussion

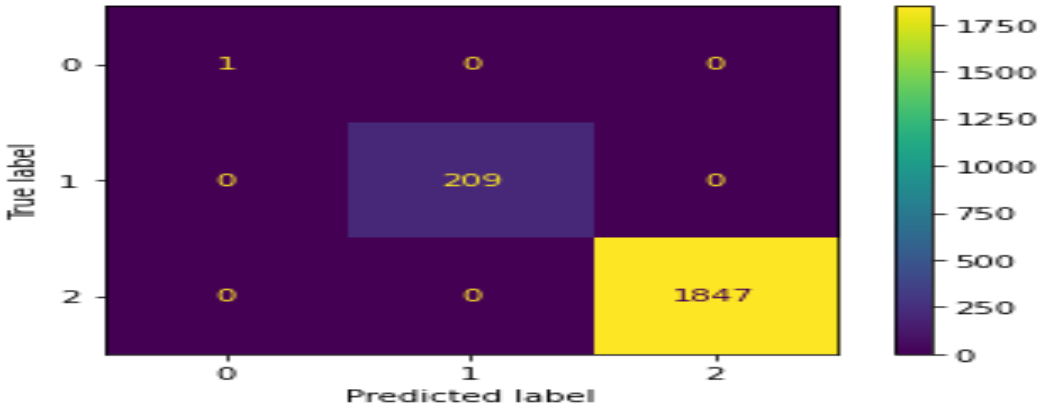

Figure 4.1: Confusion matrix of ANN

Figure 4.1 depicts the confusion matrix used to evaluate the performance of MLPN classification system using the validation set. It displays the type of errors made by the classifier. The MLPN predicted results of accident severity classification problem in road transportation system. The suggested model confusion matrix was created, with accurate predictions displayed at the secondary diagonal and inaccurate predictions noted above and below the main diagonal, or "off-diagonal elements," in that order using the testing dataset. The overall number of correctly predicted values recorded to be 1847+209+1=2057 instances with severity level 1, 2 and three while incorrectly predicted cases yielded 0+0+0=0 cases.

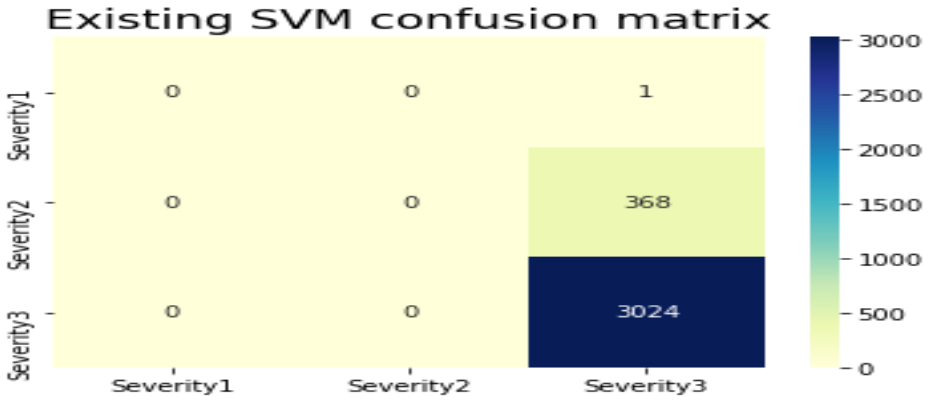

Figure 4.2: Confusion matrix of SVM

Figure 4.2 shows the confusion matrix, which displays a table structure of the various SVM predicted outcomes of a binary-classification task to aid in visualizing its results. This is used to show the predicted and actual values of a classification model. Cell values above and below the main diagonal or off-diagonal

elements showing the incorrectly predicted values, show the total number of correctly predicted values that are equal to the actual or true values. The greater the diagonal value, the more accurate the predicted EV Battery charging duration. According to the confusion matrix, accident severity level 1 had 573 incorrectly predicted cases with zero(no) correct predictions. Level 2 produced 368 incorrectly misclassified cases and level 2 provided 3024 with zero incorrectly predicted values having zero(0) true positive class prediction. 3024 accident severity levels were accurately predicted by the SVM, while 369 accident cases were incorrectly classified.

The ANN Iteration

Epoch 1/50

667/667 [==============================] - 14s 7ms/step - loss: 6.4750 - accuracy: 0.4866 - val_loss: 3.7975 - val_accuracy: 0.6040

Epoch 2/50

667/667 [==============================] - 5s 7ms/step - loss: 5.3162 - accuracy: 0.5431 - val_loss: 2.7403 - val_accuracy: 0.6685

Epoch 3/50

667/667 [==============================] - 4s 6ms/step - loss: 4.7996 - accuracy: 0.5693 - val_loss: 2.6613 - val_accuracy: 0.6550

Epoch 4/50

667/667 [==============================] - 3s 4ms/step - loss: 4.7997 - accuracy: 0.5802 - val_loss: 2.4421 - val_accuracy: 0.7305

Epoch 5/50

667/667 [==============================] - 3s 5ms/step - loss: 4.5660 - accuracy: 0.6045 - val_loss: 2.1527 - val_accuracy: 0.7440

Epoch 6/50

667/667 [==============================] - 4s 6ms/step - loss: 4.2617 - accuracy: 0.6137 - val_loss: 1.7005 - val_accuracy: 0.7790

Epoch 7/50

667/667 [==============================] - 3s 5ms/step - loss: 4.0166 - accuracy: 0.6241 - val_loss: 1.6263 - val_accuracy: 0.7730

Epoch 8/50

667/667 [==============================] - 4s 6ms/step - loss: 3.7516 - accuracy: 0.6417 - val_loss: 1.1500 - val_accuracy: 0.8320

Epoch 9/50

667/667 [==============================] - 3s 5ms/step - loss: 3.4987 - accuracy: 0.6302 - val_loss: 1.0160 - val_accuracy: 0.8080

Epoch 10/50

667/667 [==============================] - 3s 5ms/step - loss: 3.4044 - accuracy: 0.6328 - val_loss: 0.8602 - val_accuracy: 0.8565

Epoch 11/50

667/667 [==============================] - 3s 5ms/step - loss: 3.3323 - accuracy: 0.6775 - val_loss: 1.0960 - val_accuracy: 0.8435

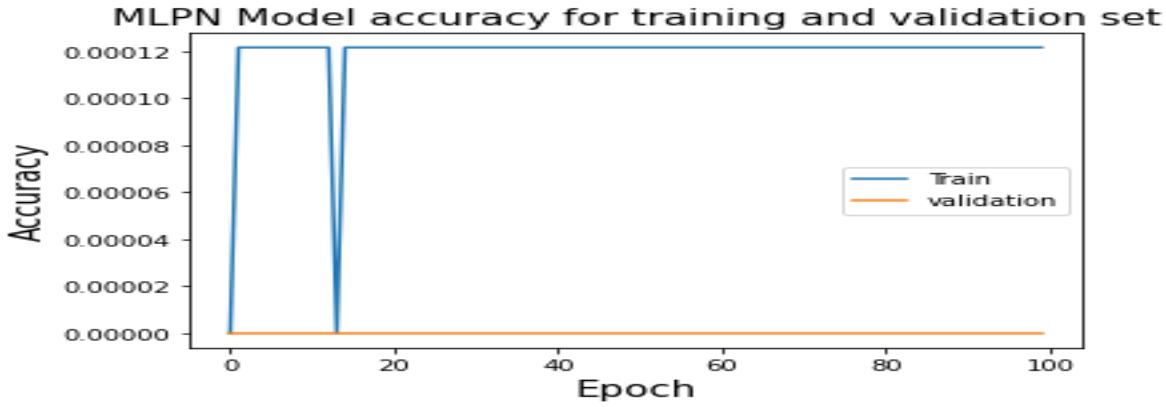

Figure 4.3: Training plot of ANN

The MLPN training and validation curve is shown in Figure 4.3. It illustrates how the predictive algorithm fails to draw valid conclusions from the testing data. The trained model performs well on training samples, but when tested on the validation data set, it performs poorly at the beginning, as the graph illustrates. The training curve steadily increased from 20 to 100 epochs after declining at the beginning and between 0 and 20. Model fitting happens because the artificial neural network (ANN) algorithm in this case was trained for an extended period of time, making it purely too sophisticated for the data. When the loss is gradual and mild, training can be decreased with the intention of terminating early. Over the training loss, the validation curve is bouncing up and down.

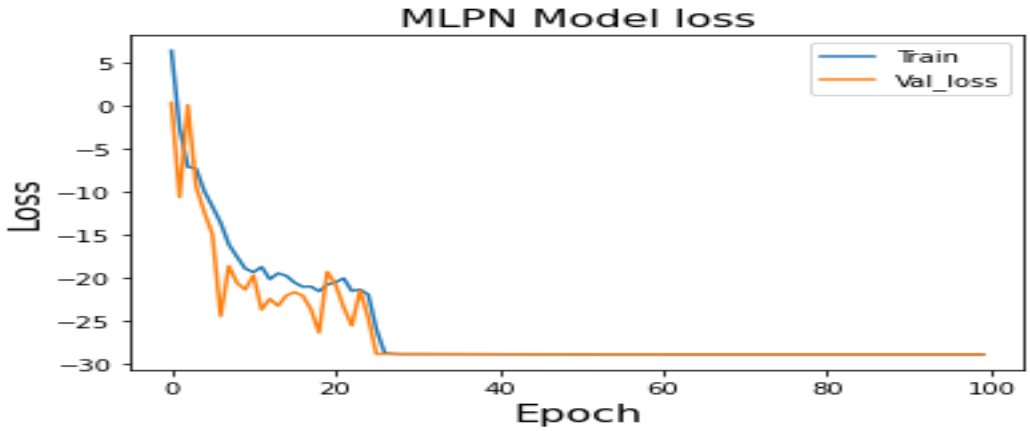

Figure 4.4: Validation plot of ANN

The training against validation loss is shown in Figure 4.4 The training is better and the validation loss is slightly less than the training loss. The validation loss in this case is lower which means the model is converging. The training data is more difficult to model than the validation set because the validation loss is smaller than the training loss, even if both the training and validation losses are decreasing in the plot. The model is receiving new data for both the training and validation sets since there is a noticeable separation between the training and validation losses.

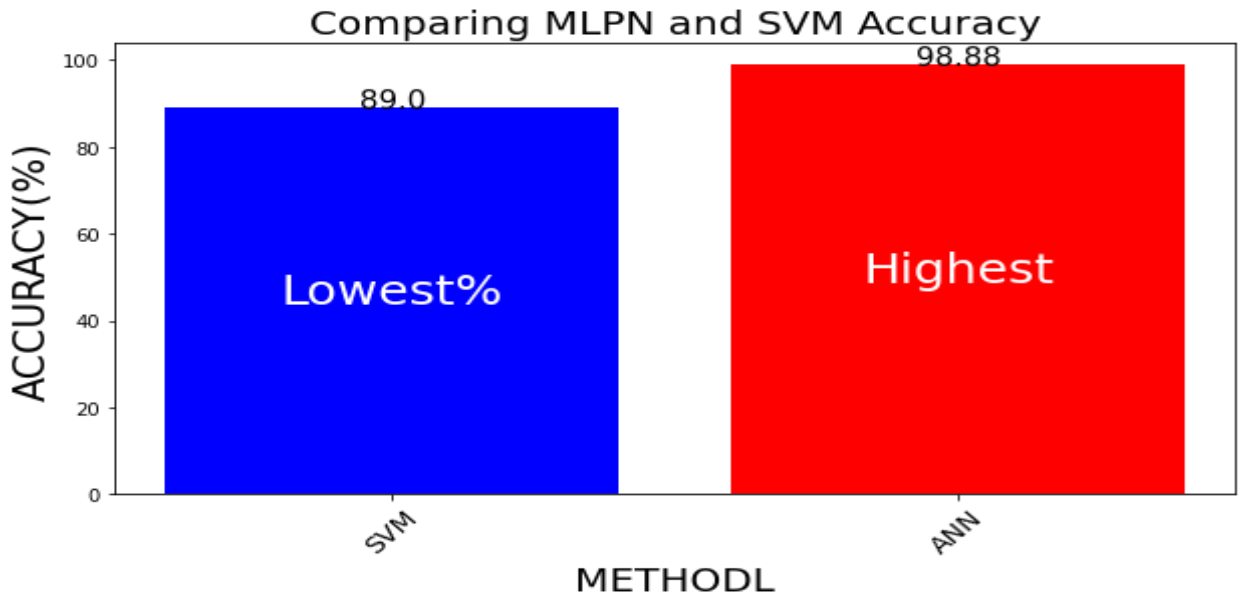

Figure 4.5: Accuracy of SVM and ANN

Figure 4.5 depicts the overall success rate of MLPN, and existing SVM technique. The MLPN algorithm was the best but was improved by incorporating the regularization techniques. The SVM had the lowest prediction accuracy (89.0), followed and MLPNN (98.88). The MLPNN perform exceptionally well, with performance metrics with testing set.

5.0 Conclusion

The effectiveness of two distinct machine learning algorithms for creating trustworthy and accurate classifiers was examined in this study. The usefulness of two different machine learning algorithms was investigated in this work in order to produce reliable and accurate classifiers. SVM and MLPN approaches are among these algorithms. Further research is required to collect pertinent data and analyze the effects of these factors. The MLPN is the most effective model that can be applied in predicting the severity levels of accidents. This helps to categorize several serious accident cases into distinct feature groups within the same target class. The proposed predictive model can be used to swiftly and efficiently identify the primary factor causing crashes in traffic.

This research brings to our knowledge the following contributions:

(a). The SVM and MLPN prediction accuracy was enhanced by the addition of penalty terms(decay concept). This is capable of predicting minor, high and severe accident cases as a multi-classification system.

(b). the adoption of the decay notion gave the user more control over the model weights, which led to smoother system convergence and simpler classification of accident severity types.

(d) Penalizing of models that learn from higher weights was one way to reduce the computational expense.

(e) It makes a significant contribution to literature, practices, and the application of the proposed system adds to knowledge with the concept of knowledge enhanced deep ML pipelining with penalty term

References

1. Chen, T.Y. and Jou, R.C.(2019) Using HLM to investigate the relationship between traffic accident risk of private vehicles and public transportation. *Transp. Res. Part A Policy Pract.*, *119*, 148–161.

2. Flammini, M. G, Prettico G, Julea A, Fulli G, Mazza A, and Chicco G.(2019) Statistical characterisation of the real transaction data gathered from electric vehicle charging stations. Elec Power Syst Res, 24(39), 30-89.

3. Hauer, E., J., Ng, C. N. and Lovell, J.(2019) Estimation of Safety at Signalized Intersections. In Transportation Research Record 1185, TRB, National Research Council, Washington, D.C., 48-61.
4. Kim, M., Lee, S., Lim, J., Choi, J. and Kang, S. G.(2020) Unexpected Collision Avoidance Driving Strategy Using Deep Reinforcement Learning. IEEE Access, 17243–17252.
5. Kononen, D.W.; Flannagan, C.A.C. and Wang, S.C.(2021) Identification and validation of a logistic regression model for predicting serious injuries associated with motor vehicle crashes. Accid. Anal. Prev., 43, 112–122.
6. Kulmala, R.(2020) Safety at Rural Three- and Four-Arm Junctions: Development and Applications of Accident Prediction Models. VTT Publication s 2 33, Technical Research Centre of Finland, Finland, 1-20.
7. Li, R.M.(2020) Traffic incident duration analysis and prediction models based on the survival analysis approach. *IET Intell. Transp. Syst.*;9(4), 351–358.
8. Li, R.M., Pereira, F.C., and Ben-Akiva M.E.(2021) Competing risks mixture model for traffic incident duration prediction. *Accid. Anal. Prev.* 75, 192–201.
9. Li, G., Ma, H. and Guan, T. (2020) Predicting Safer Vehicle Font-End Shapes for Pedestrian Lower Limb Protection via a Numerical Optimization Framework. Int. J Automota Technol, 21, 749–756.
10. Lee, Y., Wei C.H., and Chao K.C.(2017) Non-parametric machine learning methods for evaluating the effects of traffic accident duration on freeways. *Archiv. Transport.* 43(3), 91–104.
11. Ma, T. Y, and Xie, S.(2021) Optimal fast charging station locations for electric ridesharing with vehicle-charging station assignment. Transport Res Transport Environ ;90, 102-682.
12. Motz, M., Huber, J, and Weinhardt, C.(2021) Forecasting BEV charging station occupancy at work places. In: Reussner RH, Koziolek A, Heinrich R, Hrsg, editors. Informatik 2020. Bonn: Gesellschaft für Informatik, 771e81.
13. Shang, Q., Tan, D.R., Gao S., and Feng, L.L.(2019) A hybrid method for traffic incident duration prediction using BOA-optimized random forest combined with neighborhood components analysis. *J. Adv. Transport, 23(40), 1-30*
14. Shankar, V.N., Albin, R. B. Milton, J. C..L.(2019) Mannering. Evaluating Median Cross-Over Likelihoods With Clustered Accident Counts: An Empirical lnquiry Using the Random Effects Negative Binomial Model. In Transportation Research Record 7635, TRB, National Research Council, Washington, D.C., 44- 48.
15. Wang, H., Huang, Y., Khajepour, A., Rasekhipour, Y., Zhang, Y. and Cao, D.(2020) Crash Mitigation in Motion Planning for Autonomous Vehicles. IEEE Transactions on Intelligent Transportation Systems, 1–11.
16. Weng, J.X.; Gan, X.F. and Zhang, Z.Y.(2021) A quantitative risk assessment model for evaluating hazmat transportation accident risk. Saf. Sci., 137, 11
17. Yu, B., Wang, Y.T., Yao, J.B., and Wang J.Y.(2020) A comparison of the performance of ANN and SVM for the prediction of traffic accident duration. *Neural Netw. World.* 26(3), 271–287.