

**A STEP TOWARDS GENERALIZED CRASH
SEVERITY PREDICTION MODEL – A DEEP
LEARNING APPROACH COUPLED WITH
INTERNATIONAL TRAFFIC CRASH DATA**

BY

OSAMA JAMAL MOH'D MOSAMEH

A Thesis Presented to the
DEANSHIP OF GRADUATE STUDIES

KING FAHD UNIVERSITY OF PETROLEUM & MINERALS

DHAHRAN, SAUDI ARABIA

In Partial Fulfillment of the
Requirements for the Degree of

MASTER OF SCIENCE

In

CIVIL ENGINEERING

DECEMBER, 2021

KING FAHD UNIVERSITY OF PETROLEUM & MINERALS

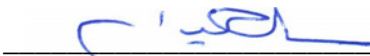
DHAHRAN- 31261, SAUDI ARABIA

DEANSHIP OF GRADUATE STUDIES

This thesis, written by **OSAMA JAMAL MOH'D MOSAMEH** under the direction of his thesis advisor and approved by his thesis committee, has been presented and accepted by the Dean of Graduate Studies, in partial fulfillment of the requirements for the degree of **MASTER OF SCIENCE IN CIVIL ENGINEERING**.



Dr. SHAMSAD AHMAD
Department Chairman

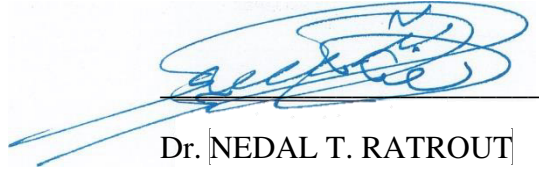


Dr. SULIMAN S. AL-HOMIDAN
Dean of Graduate Studies

Date



Dr. KHALED ASSI
(Advisor)



Dr. NEDAL T. RATROUT
(Member)



Dr. SYED MASIUR RAHMAN
(Member)



© Osama Mosameh

2021

To the great man and kind lady who brought me to this life, raised me the best they could
and kept encouraging me to push through thesis to finish my master's degree.

To my lovely wife and kids who stood by my side and supported me in my hard times
and were patient with me and foremost never lost the hope in me.

For everybody who believed in me and helped me down the way.

Thank you!

Most gratefully,

Osama |

ACKNOWLEDGMENTS

I am full of appreciation and gratitude to my advisor, Dr. Khaled Assi, who extended every support possible to help me get through my thesis process achieving the goals behind it with most satisfactory results, guided me when I was lost, or things went wrong. I also would extend warm thanks to my thesis respectful committee members: Prof. Nedal Ratrouf as we had a great run together whether in classes or in thesis work and never spared any advice, he could offer nor had he lost the enthusiasm he had from day one for me getting degreed, Dr. Syed Masiur Rahman who was the technical back bone of the project showed me the way getting things done with the programing.

Lastly, for the faculty members who taught me and gave me the knowledge and got to this stage.

A big thank you with all my heart |

TABLE OF CONTENTS

ACKNOWLEDGMENTS	V
TABLE OF CONTENTS	VI
LIST OF TABLES.....	X
LIST OF FIGURES.....	XI
LIST OF ABBREVIATIONS.....	XIII
ABSTRACT	XIV
ملخص الرسالة	XVI
CHAPTER 1 INTRODUCTION.....	1
1.1 Background	1
1.2 Significance of Research	3
1.3 Objective of Research	4
1.4 Outline	4
CHAPTER 2 LITERATURE REVIEW	6
2.1 Previous studies and research	6
2.1.1 Traditional Methods	6
2.1.2 Machine learning models	9
2.2 Comparison between methods.....	18
2.2.1 Comparison between conventional and ML methods	19
2.2.2 Comparison of ML methods	20
2.3 Summary.....	24

CHAPTER 3 DATA	25
3.1 Raw data from sources	25
3.1.1 Australia Data.....	25
3.1.2 Canada Data	27
3.1.3 Great Britain Data	28
3.1.4 United States Data	30
3.2 Data preparation/Processing	33
3.2.1 Finding the common/ most relevant variables	33
3.2.2 Uniting the different classes of a variable under a new generalized classes arrangement	34
3.2.3 Cleansing and compiling the four countries data in one sheet.....	35
3.3 International data	35
3.4 Summary.....	42
CHAPTER 4 METHODOLOGY.....	43
4.1 Introducing ML topic	43
4.2 Classification Problem	44
4.3 System Overview	45
4.4 Data	45
4.4.1 Data Preprocessing.....	46
4.4.2 Final Data	49
4.5 ML Models	55
4.5.1 K-Nearest Neighbors (K-NN)	55
4.5.2 Support Vector Machine (SVM)	57
4.5.3 Random Forest (RF)	59
4.5.4 Decision Tree (DT)	61
4.5.5 Naïve Bayes (NB)	63

4.5.6 Neural Network (NN)	65
4.5.7 Deep learning Model (DL)	67
4.6 Summary.....	69
CHAPTER 5 MODEL DEVELOPMENT AND RESULTS	70
5.1 Data	71
5.2 Tuning hyperparameters.....	71
5.3 Evaluation metrics	74
5.4 Results	79
5.4.1 K-Nearest Neighbors (K-NN).....	79
5.4.2 Support Vector Machine (SVM).....	83
5.4.3 Random Forest (RF).....	87
5.4.4 Decision Tree (DT)	91
5.4.5 Naïve Bayes (NB).....	95
5.4.6 Neural Network (NN)	99
5.4.7 Deep learning Model (DL)	103
5.4.8 Deep learning Model (DL) for individual countries	107
5.5 Summary.....	116
CHAPTER 6 COMPARISON AND DISCUSSION	117
6.1 Accuracy	117
6.2 Recall	118
6.3 F1 Score	118
6.4 Average Processing time	119
6.5 Summary.....	120
CHAPTER 7 CONCLUSION AND FUTURE WORK	121

7.1 Conclusion.....	121
7.2 Future work.....	122
REFERENCES.....	124
VITAE.....	129

LIST OF TABLES

Table 1 Comparison between conventional and machine learning methods	19
Table 2 Comparison of different machine learning techniques in literature	20
Table 3 Most relevant crash data for Australia	26
Table 4 Most relevant crash data for Canada.....	27
Table 5 Most relevant crash data for GB	29
Table 6 Most relevant crash data for USA.....	31
Table 7 International version of variables	35
Table 8 Variables considered in final data.....	50
Table 9 Pros and Cons of K-NN	57
Table 10 Pros and cons of SVM	59
Table 11 Pros and cons of RF	61
Table 12 Pros and cons of DT.....	62
Table 13 Example on NB.....	63
Table 14 Pros and cons of NB	64
Table 15 Pros and cons of NN	67
Table 16 Pros and cons of DL.....	69
Table 17 Possible outcomes of a classification problem	74
Table 18 Tuning Parameters for K-NN	79
Table 19 Tuning Parameters for SVM.....	83
Table 20 Tuning Parameters for RF.....	87
Table 21 Tuning Parameters for DT	91
Table 22 Tuning Parameters for NB	95
Table 23 Tuning Parameters NN	99
Table 24 Tuning Parameters for DL	103
Table 25 Comparison of accuracies for the 7 studied models	117
Table 26 Comparison of recall for the 7 studied models	118
Table 27 Comparison of F1 Scores for the 7 studied models	119
Table 28 Comparison of Avg. processing time for the 7 studied models.....	119

LIST OF FIGURES

Figure 1 System Overview.....	45
Figure 2 KNN concept graphical representation	56
Figure 3 Euclidean distance between points	56
Figure 4 SVM Hyperplane.....	58
Figure 5 RF structure and process flow	60
Figure 6 Structure of DT/CART	62
Figure 7 NN structure sample	65
Figure 8 Feed forward and back-propagation in NN	66
Figure 9 Example of DL NN	68
Figure 10 Confusion Matrix sample	77
Figure 11 ROC Plot	78
Figure 12 Confusion matrix for K-NN	80
Figure 13 ROC Curve for K-NN	81
Figure 14 KNN Feature Importance	82
Figure 15 Confusion matrix for SVM.....	84
Figure 16 ROC Curve for SVM.....	85
Figure 17 SVM Feature Importance	86
Figure 18 Confusion matrix for RF	88
Figure 19 ROC Curve for RF.....	89
Figure 20 RF Feature Importance	90
Figure 21 Confusion matrix for DT	92
Figure 22 ROC Curve for DT	93
Figure 23 DT Feature Importance.....	94
Figure 24 Confusion matrix for NB.....	96
Figure 25 ROC Curve for NB	97
Figure 26 NB Feature Importance	98
Figure 27 Confusion matrix for NN.....	100
Figure 28 ROC Curve for NN.....	101
Figure 29 NN Feature Importance	102
Figure 30 Confusion Matrix for DL.....	104
Figure 31 ROC Curve for DL	105
Figure 32 DL Feature Importance.....	106
Figure 33 Confusion Matrix DL - Australia	107
Figure 34 ROC Curve DL - Australia.....	108
Figure 35 Feature Importance DL - Australia.....	108
Figure 36 Confusion Matix DL - Canada	109
Figure 37 ROC Curve DL - Canada	110

Figure 38 Feature Importance DL - Canada	110
Figure 39 Confusion Matix DL - GB.....	111
Figure 40 ROC Curve DL - GB.....	112
Figure 41 Feature Importance DL - GB.....	112
Figure 42 Confusion Matix DL - USA	113
Figure 43 ROC Curve DL - USA	114
Figure 44 Feature Importance DL - USA	114

LIST OF ABBREVIATIONS

WHO	:	World Health Organization
RTC	:	Road traffic crash
ML	:	Machine Learning
DL	:	Deep Learning
K-NN	:	K-Nearest Neighbor
SVM	:	Support Vector Machine
RF	:	Random Forest
DT	:	Decision Tree
NB	:	Naïve Bayes
NN	:	Neural Networks
ROC	:	Receiver Operator Characteristic
AUC	:	Area under curve
FDR	:	False Discovery Rate
FNR	:	False Negative Rate

ABSTRACT

Full Name : [OSAMA JAMAL MOH'D MOSAMEH]

Thesis Title : [A STEP TOWARDS GENERALIZED CRASH SEVERITY
PREDICTION MODEL – A DEEP LEARNING APPROACH
COUPLED WITH INTERNATIONAL TRAFFIC CRASH DATA]

Major Field : [TRANSPORTATION ENGINEERING]

Date of Degree : [December,2021]

Road traffic crash (RTC) is one of the critical problems worldwide. According to WHO, millions of people are killed and injured in RTCs. The global mean fatalities due to RTCs was 18.2 deaths per hundred-thousand people in 2016, with lower-income countries suffering the most. About 1.35 million people face death in traffic crashes annually meaning a daily average exceeding 3,290 people. Casualties involving injury of impairment coming to a sum between 20 to 50 million. Traffic crashes come in the 8th place as a death leading cause, summing up to 2.2% of death counts worldwide. Countries of low and middle income owning below 50% of the vehicles of the world contribute with more than 90% of the crash fatalities. The accurate prediction of traffic crash severity contributes to generate crucial information which can be used to adopt appropriate measures reducing the aftermath of crashes. Many scientific studies developed different kinds of machine learning (ML) models for crash severity predictions. However, there is a need of the model using relevant data reasonably covering the world which can be customized for new datasets with limited training data. It will pave the path towards the development of a global solution to traffic crash severity prediction. This study aims to build the solution through the synergy between big data and deep learning. The initial step is to develop an international database of traffic crash severity consists of publicly available resources from relevant authorities and academic publications. Different machine learning models along with many learning and regularization approaches will be investigated to develop an international model which will address the bias and variance issues adequately. It is assumed that the deep learning model will provide a high-level understanding of the features. Therefore, the model is expected to perform well even with limited training

datasets for new locations. It will also be compared with conventional machine learning models. The output of this research will mainly assist proper actions and remedies required to enhance traffic safety in the developing countries due to the limited availability of traffic crash data. Moreover, the output of this research can be utilized by researchers globally and international entities such as World Health Organization, United Nations.

ملخص الرسالة

الاسم الكامل: اسامه جمال محمد مسامح

عنوان الرسالة: خطوةً باتجاه نموذج عام للتنبؤ بفداحة الحوادث المرورية – مقارنة باستخدام التعلم العميق مقروناً بمعلومات حوادث مرورية دولية

التخصص: هندسة مدنية

تاريخ الدرجة العلمية: ربيع الثاني 1443 هـ

تعد الحوادث المرورية أحد أهم المشاكل عالمياً. وفقاً لمنظمة الصحة العالمية، فإن ملايين البشر يقضون ويصابون إصاباتٍ بليغةً في الحوادث المرورية. حيث بلغت معدلات الوفيات العالمية الناتجة عن الحوادث 18.2 وفاة لكل مئة ألف نسمة للعام 2016 مع معاناة الدول الفقيرة بشكلٍ أكبر. يبلغ معدل الوفيات السنوي بسبب الحوادث ما يقارب 1.35 مليون وفاة أي ما يعادل 3290 شخصاً بمعدل يومي، بينما تتسبب بإصابات بالإعاقة تتراوح ما بين 20 و50 مليون إصابة سنوياً. تمتلك الدول الفقيرة ومتوسطة الدخل أقل من نصف التعداد العالمي من المركبات بنما تسهم بأكثر من 90% من معدل وفيات الحوادث عالمياً. يشكل التنبؤ الدقيق بفداحة (شدة) الحوادث المرورية مصدرًا مهمًا للمعلومات التي يمكن أن تستخدم لتبني المعايير المناسبة للتخفيف من تبعات الحوادث. هناك الكثير من الدراسات العلمية التي تم تطويرها باستخدام أنواع مختلفة من نماذج التعلم الآلي للتنبؤ بفداحة الحوادث المرورية. على الرغم من هذا، هناك حاجة ملحة لنموذج يستطيع التعامل مع المعلومات بمنطقية واعتدال لتغطية العالم ولديه مرونة للتعديل بقوائم مدخلات جديدة وكمية محدودة من المعلومات للتدريب. سيمهد هذا النموذج الطريق نحو تطوير حل عالمي للتنبؤ بفداحة الحوادث. تهدف هذه الدراسة لبناء حل من خلال الموائمة بين البيانات الضخمة وتقنية التعلم العميق. تكمن الخطوة الأولى بتطوير قاعدة بيانات دولية لفداحة الحوادث المرورية من خلال مصادر البيانات المفتوحة من السلطات المعنية والأبحاث الأكاديمية. ثم استقصاء أنواع مختلفة من التعلم الآلي باستخدام طرق تعلم وتسوية لتطوير نموذج دولي يقوم بحل مشاكل التحيز والفروق في المدخلات بشكل مناسب. ويفترض أن نموذج التعلم العميق سيوفر مستوىً عاليًا من فهم خصائص المدخلات. لذلك يتوقع أن يعمل بشكلٍ جيد حتى بوجود شُحٍّ في مدخلات التدريب لمواقع جديدة. وسيتم مقارنة أدائه مع نماذج التعلم الآلي الأخرى. ستساعد حصيلة هذا البحث باتخاذ الإجراءات والعلاجات اللازمة لتحسين الأمن والسلامة المروريين للبلدان النامية بسبب محدودية توفر بيانات الحوادث المرورية. أضف إلى ذلك إمكانية استخدام مخرجات هذا البحث عالميًا من باحثين آخرين وكياناتٍ دوليةٍ مثل منظمة الصحة العالمية والأمم المتحدة.

CHAPTER 1

INTRODUCTION

1.1 Background

According to statistics by World Health Organization (WHO)[1], about 1.25 million people face death in traffic crashes annually meaning a daily average exceeding 3,290 people. Casualties involving injury of impairment coming to a sum between 20 to 50 million. Traffic crashes come in the 9th place as a death leading cause, summing up to 2.2% of death counts worldwide. Countries of low and middle income owning below 50% of the vehicles of the world contribute with more than 90% of the crash fatalities. A certain financial liability accompanies these crashes costing individual countries a loss of 1-2% of annual GDP and a bill worth 518 billion dollars in global terms. Lastly, if the traffic crashes continue their growth without limiting actions being taken, they are expected to become the 5th death leading cause by 2030.

The frequency of traffic crashes has been on the upward trajectory all over the world with the increase of population and motorization. Conventional measures for reducing crashes include improvement in geometric design, demand management techniques and better driver education. Various factors associated with traffic collisions have a major impact on one another and thus make it harder for transportation safety designers to consider any of the parameters individually in explaining the traffic crash severity. Many transportation

professionals take interest in the modeling and analysis of traffic crash severity. The primary objective of traffic crash severity modeling is to understand the connection between traffic injury, collision severity and several factors involved such as human attributes, environmental factors, road characteristics and vehicle characteristics. The model helps decision makers to predict the crash severity for possible future conditions and thus implement appropriate countermeasures for reducing the traffic crash severity and taking necessary actions at the time of traffic crashes.

Over the years various statistical methods have commonly been used to estimate the magnitude of traffic collisions. Of these statistical models are commonly used the ordered probit modeling [2-6], the ordered logit modeling [7], the multinomial logit modeling [8, 9] and the logistical regression model [10]. While these standard models have been used widely over the years to forecast the severity of crashes, they struggle of some obvious limitations. For instance, the assumptions relating to data distribution and linearity, functional form between explanatory and dependent variables which do not always hold true leading to unreliable conclusions [11, 12]. In order to tackle those constraints a number of non-parametric techniques and machine learning models have been proposed for analysis of traffic crash severity [13]. Such techniques comprise the Bayesian structure of the network [14, 15], the framework of regression tree carts [16, 17] and the artificial neural networks [11, 18].

This study aims to move a step forward towards a generalized prediction solution through the synergy between international data from several countries and deep learning models. The initial step is to develop a database of traffic crash severity consists of publicly available resources from relevant authorities, data then is preprocessed to overcome any

flaws as outliers, gaps and most importantly imbalance between classification classes. Several conventional ML models are tuned to reach their optimal performance. A deep learning (DL) model is then developed and optimized. It is anticipated that the deep learning model will provide a high-level understanding of the features. Therefore, the model is expected to perform well even with limited training datasets for new locations. It will also be compared with conventional machine learning models. The output of this research will mainly assist proper actions and remedies required to improve traffic safety in the growing countries due to limited availability of traffic crash data.

1.2 Significance of Research

Crash severity is a booming field of research as scholars feel responsible to provide their brothers in humanity a better world to live in. Unfortunately, none of the previous study efforts tried combine data from several resources in the pursuit of opening a new door of developing a generalized model that can serve the less fortunate.

As many developed nations have reliable systems for traffic crash recording and forecasting offering an abundant resource for data to be used for the future crash avoidance, a lot of the developing countries do not possess such records or acquire them over discrete time intervals and these countries participate by the majority of world crashes recorded per annum [1].

This research aims to set the corner stone for the development of a generalized solution to traffic crash severity prediction. This is achieved through the synergy between data from several countries and deep learning models.

1.3 Objective of Research

The specific objectives of this study are as following:

- 1- Conduct an extensive literature review to find the most common models and techniques used worldwide for road traffic crash severity prediction
- 2- Develop a database of traffic crash severity consists of publicly available resources from relevant authorities and academic publications, using data from four countries and preprocess for any flaws
- 3- Use conventional ML models after tuning them to reach their best performance to fit and test the final data set.
- 4- Develop a DL model which will be optimized then carry out the data set fitting and testing.
- 5- Conduct comparative analysis between DL model and the conventional ML models.

1.4 Outline

As this chapter comes to an end at this point, we wanted to give the reader a glance of what to expect next; the next chapter two will cover a thorough literature review, followed by chapter three which stem lines the data from its' original sources to the point where we have uniform data in their classes and attributes, afterwards in chapter four we dive deeper in the methodology and discuss the models in greater detail, later in chapter five we get down to business in model development, tuning and results, chapter seven is where we

discuss the results and compare models' performance and the final chapter eight is the place to draw the conclusion and future works. |

CHAPTER 2

LITERATURE REVIEW

2.1 Previous studies and research

This section synthesizes literatures on modeling and prediction techniques pertaining to traffic crash severity research. General modelling strategy of severity of crashes includes statistical modeling that considers severity as an outcome variable whereas road, vehicle, driver, and environmental conditions are explanatory variables.

2.1.1 Traditional Methods

Conventionally, regression tools such as logit or even probit were utilized to simulate the severity of a traffic crash [19, 20]. Several researches employed these same models in binary mode [2, 21, 22] to model two crash severity levels when others modeled multiple severity degrees using logit and probit multinomial models [23, 24]. A number of innovative models were created in order to consider normal behavior, diversity and within-crash correlation in traffic crash severity data. These techniques Incorporate Bayesian hierarchical [25], ordered probit (OP) and logit models [3, 26], nested logit models [24, 27] and combinations of these models [28, 29].

A model was developed using the ordered probit modeling approach for predicting the roadway crashes severity [30]. Multiple driver, road and vehicle-related contributing parameters were investigated assess their impact on the level of injury. The study concluded that urban area crashes are less severe compared to the rural area crashes, while

females have higher probability of severe injuries compared to males. Moreover, dry surface, light vehicle travel and two-way roads lead to more severe crashes compared to wet surface, heavy vehicle travel and one-way roads. Kockelman and Kweon [3] built probit models for all forms of single-vehicle crashes and two vehicle which revealed that sports cars and pickups involve in more severe crashes evaluated by comparing to passenger vehicles, when younger men users are operating at lower velocity cause less serious crashes in new vehicles. Hu et al [31] developed a logit model to investigate essential crash severity causes at railroad grade crossings. Three-year period data for Taiwan was used to analyze the crash severity level which was classified in to four classes (0-3), with higher severity level showing more severe crash. The study revealed that the daily number of trucks, trains, the system for monitoring obstructions and the incoming passing points were the key significant factors statistically affecting the severity level of crashes.

A comparative analysis between the regression modeling and the Bayesian network was performed to predict traffic crash severity. A data set of 2,246 cases for China containing the information about crash severity, crash characteristics, environment characteristics and road characteristics was used for analysis, with 50% of the data used for training while the other 50% used for testing. The study revealed that Bayesian network having higher goodness of fit outperforms regression modeling in crash severity prediction [32]. A study was conducted to investigate the essential parameters influencing driver injury severity and vehicle destruction by developing a Bayesian hierarchical binomial logistic model at signalized intersections using the data for Singapore [25]. This study revealed that injury severity level was affected by time of the day, type of intersection, nature of lanes, presence

of camera at signals, street lighting and pedestrian involved. The study concluded that crashes taking place in peak time and Reasonable state of street lighting are less severe while those occurring in the evening, at T/Y intersection, on extreme right lane and having cameras with red traffic signal appears to be more serious. The driver age and vehicle condition also affect the severity.

In Riyadh, Saudi Arabia, Al-Ghamdi [10] applied logistic regression to classify variables that led to the severity of the crash. The considered factors include crash time, location, cause, driver age and nationality. The really critical parameters which determine the intensity of the collision found to be location and cause of the collision. Multiple logistic regression (MLP) model was used to forecast the crash severity on Thailand highways [33]. The study came to a result that the velocity on the road section is the main important aspect influencing degree of severity. To investigate the effect of roadway and environmental factors on traffic crash severity and predict its level in the USA a logistic regression model was created [34]. Posted speed, condition of lighting, geometry of the road, road function class, crash location and road surface status were found to be significant factors affecting traffic crash severity. A lognormal regression was used to investigate the effect of various parameters such as weather conditions and crash type on the traffic crash severity in Greece [35]. The authors developed separate severity prediction models for each type of vehicles.

Zhu and Srinivasan [36] modeled occupant level injury severity (OIS) for crashes involving large trucks. The highest injury severity (HIS) models predict the overall highest crash severity while (OIS) models predict the injury sustained by each individual involved in the crash. The use of drugs and inattention while driving was associated to highly severe crashes, while use of seat belt and availability of airbags tend to lower the crash severity.

Driver's familiarity with road and vehicle also has significant effect on crash severity. A study in Netherlands [37] revealed that risk of crash occurrence is associated with drug use but not enough evidence is found to demonstrate the effect of drugs on level of crash severity.

Nested logit models and zero-inflated count models were used to study the roadside factors effect on run-off collisions frequency and intensity [38]. The study revealed that the frequency and severity of such crashes can be minimized by avoiding cut slopes, reducing space between external shoulder edge and guardrail, decreasing the count of disconnected trees on the roadside and expanding the space between outer shoulder boundary and lamp posts. Abdel-Aty [39] analyzed the driver injury severity in Central Florida using OP model. Results showed that injury severity was affected by age, use of seat belt and car speed. Random effects generalized ordered probit model was used to find the factors affecting the injury severity of rollover crashes in Malaysian Federal roads [40]. It was found that darkness, rainy weather, number of lanes, speed limit and traffic volume are positively associated with highly severe rollover crashes.

2.1.2 Machine learning models

Without the assumptions of statistical models, a number of machine learning models have been developed for predicting traffic crash severity. These machine learning algorithms are able to simulate the non - linear interaction between the magnitude of the crash and relevant variables. Sameen and Pradhan [41] forecasted the intensity of traffic injuries and fatalities using recurrent neural networks (RNNs). Compared to multilayer perceptron neural network (MLP) and Bayesian logistic regression (BLS), the validity of the model RNN was measured. It was observed that RNN is outstanding with 71.77 percent accuracy trailed

by MLP and BLS with 65.48 percent and 58.30 percent accuracy respective. Comparison was made of the results of artificial neural network and OP model for estimating the degree of traffic crash severity [42]. Based on the presumption that crash had happened, Fuzzy adaptive resonance theory (ART) and MLP were used to analyze the harm severity rates. The research suggested that MLP scored higher with 73.6 percent accuracy rate and ART displayed 70.6 percent prediction performance however OP model findings were less reliable with 61.7 percent accuracy. In addition, it was concluded that crash severity level is affected by gender, vehicle speed, seat belt use, type of vehicle, type of crash and area type.

Artificial neural network (ANN) was used for predicting the intensity of collisions for Abu Dhabi using 6 years crash data [18]. The input attributes were sixteen and output variable (i.e. severity) had four levels. The study concluded that the developed ANN can predict crash severity with reasonable accuracy using the k-means algorithm. ANN 's findings were contrasted with the OP model as ANN surpassed the OP model in respect of precision [43]. Another study was conducted to investigate the use of ANN in predicting the traffic injury severity at signalized intersection in Central Florida, USA [18]. The authors conducted a comparison between two types of Neural Networks namely, MLP and ART neural networks based on the prediction accuracy. It is found that MLP had outrun the ART.

Along with neural network, many other machine learning models were developed to predict the crash severity. A study employed ANN, support vector machine (SVM), decision tree (DT) and logistic regression models to forecast intensity [44]. It displayed that the prediction with highest accuracy was carried out by SVM followed by DT and ANN,

respectively. The logistic regression model was the least accurate amongst the four types of models. Use of seat belt, manner of collision and drug involvement proved to be top predictors for injury severity. A multi-objective genetic algorithm was used to predict the severity of road crashes in Tehran, Iran [45]. The main objective was to compare the accuracy for the proposed method, with other techniques such as ANN, SVM and DT. The study concluded that results for severity prediction using the proposed technique are far more precise contrasted with DT, SVM and ANN techniques. Zhang et al. [46] conducted a study to predict and compare the traffic crash injury severity using machine learning and statistical techniques. Two statistical techniques employed for the severity prediction namely OP model and multinomial logit model (MNL), while four machine learning techniques namely, K-Nearest Neighbor, DT, Random Forest (RF) and SVM were also used. The study concluded that although the machine learning techniques suffered from over-fitting problems, they outpaced conventional statistical methods in contexts of forecast consistency. Among all these six techniques, RF showed the highest overall prediction accuracy while OP had the lowest accuracy.

Likewise, an analysis was carried out to forecast and assess the magnitude of traffic collisions using ANN, genetic algorithm (GA), integrated genetic algorithm and pattern search approaches [47]. The study revealed that artificial neural network surpassed the other three approaches with an R value of 0.87 and an R value of 0.79 was determined for genetic algorithm and pattern search methods. Li et al. [48] produced a hybrid model of ANN and GA to evaluate the major reasons that influence road crashes intensity in Washington, USA. The researchers merged GA and ANN structure to increase the efficacy of the quest for the major parameters. Using the proposed method, it was found that the

significant variables affecting the severity of road crashes are Operator behavior, age, road pavement situation, driver discipline and vehicle performance. Deep learning derived convolutional neural network (CNN) has been used to forecast the magnitude of traffic collisions. CNN has been compared to a variety of statistical methods such as logistic regression and machine learning techniques like SVM and ANN. The study showed that CNN model outperformed all the other techniques [49].

Moomen et al. [50] used a negative binomial model to assess the parameters that impacted severity level as well as frequency of truck crashes in downhill of Wyoming mountainous terrains. The analysis depicted a direct proportional relation between the daily numbers of trucks, width of shoulder, the sloping segment length and the truck crash intensity and magnitude. On the other hand, access point numbers and length of horizontal curve have an inverse relation with the severity of the crash.

Lin et al. [51] utilized Random forest (RF) and XGBoost models with label, one hot encoders to create four models that aimed to stem the major sources of young drivers collisions in Texas west rural areas. The authors concluded that high speeds, inability to regulate speed, inability to grant the right of way, driver lack of concentration and exhaustion (organized descending). Severity, however, was a function of; the class of road, posted speed, first damaging incident, number of lanes, form of traffic management and age of individual. Comparison of the four models revealed a superiority of XGBoost with label encoder computing speed and accuracy wise.

Mondal et al.[52] studied the relationship between crash severity and weather conditions for the state of Connecticut. By using two ML models; Bayesian additive regression tree

(BART) and RF. The results shown higher potential in favor of RF with 73 percent crash prediction capabilities relevant to weather conditions whereas BART was inferior with only 61 percent. Additionally, the authors considered two significant parameters that influence rate of severity; collision manner and weather situation.

Assi et al. [53] checked the applicability of machine learning techniques by developing sample models AdaBoost, KNN, SVM, ANN, DT and 2-layer ensemble model applied for a 5-year dataset from GB. Then they applied the same model to a 3-year data set from Canada with reasonable accuracy compared to the base training and testing set from GB. In terms of accuracy ensemble model came in first place followed by AdaBoost with KNN being the last.

A study [54] of rickshaw crashes severity using the data between years 2017 and 2019 in the city of Rawalpindi in Pakistan employing deep learning algorithms, the models were decision tree, random forest, and decision jungle. Parameters including youth drivers, high speed facilities, bright weather, off-peak timing, and weekday were contributing to worse rickshaw crash severity. Results showed a prediction accuracy of 83.7% for DJ outperforming both DT and RF.

A study [55] for severity of motorbike crashes in Ghana used bike crashes database during the time span between years (2011-2015). Four injury categories were considered: damage-only, injured, hospitalized and fatal. Instance based learning calibrated to K parameter, random forest, decision tree and multinomial logit models were used to analyze crash severity of bikes. The results depicted a superiority of ML models over MNLM in terms of effectiveness and accuracy whereas RF were best outcomes explanatory model among

all. Parameters including road separation, crash time type of location, weekday, collision type, condition of road shoulder, time of the day and colliding partner were highly correlation to the degree of severity.

A combination of improved particle swarm optimization and artificial neural network was developed [56] to forecast the crash severity on Washington state highway I5 with a data set of 6 years was used for input and training. A comparison to NN revealed that the combined model produced better explanation for the outcomes. Parameters related to vehicle were found to be of greater significance compared to road characteristics.

A study [57] to find out the primary inputs affecting the injury severity of bike pillion passengers. The data set from Iran for 4 consecutive years was introduced in classification-regression tree model to perform analysis. A dramatic improvement of 74 percent was observed in the precision of model forecasting as sixteen parameters were used in the architecture. Elements as the body part hurt, land use and sort of area were of most significance to passengers' fatality.

Work zones crash severity in Louisiana during the period (2014-2018) was studied [58] and analyzed using a new deep learning technique with a special loss function. The method considered factors pertaining to vehicle, road, and user as influencers to crash severity. The DL model displayed a superior forecasting capability of collision severity in death and injury collisions both in recall and accuracy.

A comparison [59] of PBNN, OL and convex combination modified NN was studied on 2 datasets from the year 2006 for Florida for crash severity. The outcomes revealed a quicker

training and convergence in CCNN with a near accuracy to full BPNN and OL coming in the last place in terms of data interpretation and forecast precision.

The development of ensemble model (CSSV-AGX) including extreme gradient boosting, decision tree gradient boosting and adaptive boosting to minimize the crash severity in smart vehicles in emergency cases [60]. Datasets for crashes between 1995 and 2015 were employed. The outcomes revealed that CSSV-AGX came in first place followed by SVM then MLP with RF in last place depending on forecast accuracy. Also, vehicle weight and velocity as well for passenger age are important parameters in determining crash severity for smart vehicles.

An investigation [61] in the parameters affecting severity of heavy vehicles was conducted using three machine learning models: adaptive boosting, random forest, and gradient boost decision tree GBDT, in addition to a conventional mixed logit model. The dataset of four consecutive years for crashes in Texas was obtained and used. GBDT came in the first place both by forecast precision and capability of finding explanatory parameters that define the damage level of a crash. Fatigue and use of alcohol/drugs while driving was found to be of a major impact of heavy vehicle crash severity.

Using classification & regression tree model CART to study the light condition effect on the severity level of crashes at work zones on highways in Tennessee [62], the dataset contained crashes between the years 2003 and 2015. Illumination status in work zones were in 3 main categories: night with lighting, night without lighting and day light. The study concluded with the significance of lighting conditions when observing crash severity. And

indicated a proportional relation between crash severity and lanes number during daytime, on the contrary an inverse relation during night.

A study for motor bike crash severity in mountainous terrains was conducted using a Wyoming 10-year crash data set [63]. The study tested different neural networks including recurrent, deep belief, single layer and multilayer. The results indicated a better output of DBN, MLN compared to SLN in severity projection while RNN came in the first place overall with a lower error margin.

A study of multi vehicle crash severity with consideration to temporal steadiness was carried out on a Washington highway crash data set between 2016 and 2018 [64]. The employed models included: RF and RPL (Random Parameters Logit. The study used four Collision-related parameter categories for models. Outcomes displayed RF outperformed RPL in terms of global precision. But RPL performed better when it came to prediction presentation looking towards collision costs.

Depending on crash data between 2014 and 2016 in Ghana, the researchers [65] developed a Bayesian Network simulation, an ensemble approach combined with mass weighted minority oversampling for overcoming the training data imbalance. The results revealed the suggested approach was efficient to locate the utmost abundant predictors for the level of crash damage. Lastly, the study proposed high likelihood interpretations of incidents involving several lethal wounds.

A study for rural highway crash severity in Saudi Arabia [66] used a data set of 3 years for 15 different locations with two modes of severity as deadly or non-deadly severity. The developed model feedforward neural network was used with BP algorithm for training. The

model outcomes were tested against different indicators including global precision, specificity and sensitivity which proofed an adequate model performing. Significant variables were found to be weather conditions, traffic flow, sort of vehicle and road, travel speeds, presence of pedestrians and field damage conditions.

The study [67] suggested a random under sampling of mass category (RUMC) approach to solve the problem of imbalanced crash datasets and give an improved forecasting for marginal category. Using models: RF, logistic regression, KNN, and RT for testing the imbalanced data set from one angle and RUMC adjusted data set from the other. Results shown an improvement in identifying lethal collisions and harm causing ones when using RUMC. Depending on global precision LR came in first place followed RF then RT with KNN in the last place.

Using the traffic flow and crash data from Florida in 3 different highway locations over a span of eight years, the study [68] sought to produce a real time forecasting model to dispatch the appropriate number of emergency vehicles to the crash location. The study employed big data mining and considered 5 different models for severity forecasting including: RF, KNN, DT, LR, and Naïve Bayes NB. The minimal calculation time was reached through DT and NB however with the lowest precision. KNN showed high precision but was most time expensive on the other hand. Leaving RF in the lead both in result precision and calculation time savings.

With a 5-year crash data set from Dade County, Miami. Researchers [69] aimed to forecast collision severity in work zones and influencing variables involved, the proposed method had 2 main models in comparison one was SVM and the other was mixed logit (traditional).

Improvements to SVM action were done by introducing meta heuristic algorithms: whale optimization WOA, harmony search HS, and particle swarm optimization PSO. Outcomes showed a clear superiority for SVM over ML and HS algorithm to provide best results in combination with SVM. Crash severity influencer variables at work zone included: left rear crashes, morning peak hours, interstate and highway road types and areas of activity and termination.

A study [70] to forecast the crash severity for motorized bikes in the district of Chennai, India obtained crash data of 3 years in the same district. Researchers proposed the use of three models: RF, ordered probit (OP), and conditional inference forest (CIF). The results placed CIF in first place followed by RF and OP. Important parameters affecting the bikes crash severity were: firstly, the kind of collided vehicle with the major influence then crash kind, age of driver, and road visibility.

Using a 4-year-suburban crash data of 8 provinces in Iran, researchers [71] targeted to forecast the parameters affecting severity in crashes related to heavy vehicles. The utilized methods included SVM and RPBM (random parameter binary logit). Following scenarios were considered when developing models: truck related multivehicle crashes with truck drivers mistake, truck related multivehicle crashes with truck driver not responsible, solo vehicle truck crashes. The parameters of exhaustion as well for divergence to left were considered of significant affect causing lethal collisions when it is truck driver's mistake.

2.2 Comparison between methods

In this section two comparison tables are provided (one between parametric and non-parametric models, the other includes most important machine learning models contained in literature), to make it easier for the reader to ingest the broad ideas and motivations

behind different models and comprehend the reason that drives researchers to innovate novel models and prediction techniques.

2.2.1 Comparison between conventional and ML methods

Table 1 Comparison between conventional and machine learning methods

Conventional (Parametric)	Machine Learning (Non-parametric)
Simple (easy application without previous experience)	Complex (requires someone with programming knowledge)
Fast (simple mathematics)	Usually, time expensive compared to traditional
Requires less data for training and getting results	Data hungry needs huge datasets for training
Rigid as the function would have a single form and no manipulations are allowed	Flexible deals with any number of functional and data amount
Has preset biases that affect the results accuracy	High solving capabilities as no underlying biases or intrinsic assumptions
Usually, single function with expected outcomes	Performs well with higher orders of functions and shows improvements with combinations
Poor fitting	Better fitting in most cases but may suffer under/ overfitting if no adequate parameter calibration

2.2.2 Comparison of ML methods

Table 2 Comparison of different machine learning techniques in literature

Reference	Dataset	Records (Training/Testing) size	Attributes	ML model(s)	Output of ML	Best result(s)
[41]	Crashes in Petaling Jaya, Malaysia (2009-2015)	1130 crashes (80% for training and 20% for validation)	9	LSTM-RNN	3-classes (PDO, Evident Injury, and Disabling Injury)	RNN accuracy is 71.77%
[44]	NASS GES database (2011-2012)	279,470 crashes (10-fold cross-validation)	29	NNs, SVMs, DTs and logistic regression	2-classes (low-level-of-injury (possible injury and minor non-incapacitating injury) or high-level-of-injury (incapacitating injury and fatality))	SVM accuracy is 90.41%
[46]	Crashes in Florida	5538 crashes (training and testing with a ratio of 4:1)	18	KNN, DT, RF, and SVM	5-classes (no-injury, possible/invisible injury, no-incapacitating injury crashes, incapacitating injury, and fatal)	RF and KNN accuracies are 53.9% and 52.9% respectively
[49]	Crashes in Leeds City (2009-2016)	18,727 crashes (80% for training and 20% for testing)	12	TASP-CNN	3-classes (slight, serious, and fatal)	Average Micro-F1 score is 0.84
[51]	Crashes in Texas (2010-2018)	8859 crashes (75% for training and 25% for testing)	17	Encoding (hot coding or Label Encoder) and predictive	5-classes (Killed, Incapacitating injury, Non-incapacitating injury, Possible injury, Not injured)	Average MAE is 0.7140

				algorithm (RF or XGBoost)		
[53]	Crashes in Great Britain (2011-2016)	(7000 crashes for training and 3000 crashes for testing)	15	FNN, SVM, FNN-FCM, and SVM-FCM	2-classes (severe crashes and non-severe crashes)	Accuracy of SVM combined with FCM is 74.2%
[54]	Crashes in Rawalpindi city, Pakistan (2017-2019)	2,743 crashes (stratified 10-fold cross-validation)	12	DJ, RF, and DT	4-classes (No Injury, Minor, Severe, and Fatal)	DJ accuracy is 83.7%
[55]	Crashes in Ghana (2011-2015)	8516 crashes (10-fold cross validation)	13	J48 DT, RF and IBk	4-classes (fatal, hospitalized, injured, and damage-only)	RF accuracy is 73.91%
[56]	Crashes in Washington State (2011-2015)	Crashes from 2011-2014 for training (17839), and the testing set (4984) from 2015	20	PSO + ANN	3-classes (Injury, Non – injury, Incapacitating – Injury)	NN with PSO accuracy is 73.6%
[58]	Crashes in Louisiana (2014-2018)	6336 crashes (80 % for training and 20 % for testing)	12	CNN based deep learning algorithm	3-classes (no injury (PDO), injury, and fatal)	Recall is 0.67 for fatal, and 0.84 and 0.64 for PDO and Injury
[59]	Crashes in Florida (2006)	Training set is 85,971 and testing set is 21,493	17	NN with CC algorithm, and modified N2PFA algorithm	4-classes (No injury, Possible injury, No-incapacitating injury, and Incapacitating injury/fatality)	Optimized CC accuracy is 54.91%
[60]	Crashes in US provided by	28071 crashes (training set and testing set with a ratio of 7:3)	17	CSSV-AGX with AdaBoost, GBDT, and XGBoost	3-classes (no-injury, slight injury, and serious injury)	Accuracy of CSSV-

	(NASS-CDS) (1995-2015)					AGX is 50.96%
[62]	Crashes in Tennessee work zones (2003- 2015)	19941 crashes	15	CART	2-classes (injury and PDO)	Accuracy is 72.2%
[63]	Crashes in Wyoming (2006- 2016)	2,430 crashes (80% for training and 20% for testing, 10 fold cross validation)	40	deep belief network, RNN, multilayer neural network, and single-layer neural network	5-classes (fatal, incapacitating injury, non- incapacitating injury, possible injury, and PDO)	RNN accuracy is 73.76%
[64]	Crashes in Washington, United States (2017-2018)	13,478 crashes	18	RPL and RF	3-classes (PDO, injury, and fatal)	RF accuracy is 67.16%
[65]	Crashes in Ghana (January 2014 and December 2016)	27,387 records (10-folds cross validated)	16	EbVR	2-classes (non-injury and fatal injury)	Accuracy is 89.75%
[66]	Crashes in KSA (2017–2019)	12,566 crashes (80% for training, 10% for validation and 10% for testing)	21	FFNN	2-classes (fatal or non-fatal injury)	Accuracy is 77.5%
[67]	Crashes in York, Great Britain, (2005-2018)	6515 crash records (70% for training and 30% for testing)	17	RUMC with random tree, KNN, logistic regression, and RF	2-classes (fatal and injury crashes, and PDO crashes)	Accuracy for RUMC- based KNN is 57.2%
[68]	Crashes in Florida (2008- 2015)	55188 crashes (80% for training and 20% for validation)	13	RF, NB, KNN, LC, and DTC	2-classes (slight and fatal)	RFs accuracy is 83%

[69]	Crashes in Florida Signal Four (2013-2017)	6,865 crashes ((a) 80% for training and 20% for testing and (b) 90% for training and 10% for testing)	13	SVM	2-classes (PDO, Injury)	Optimized SVM accuracy is 83.5%
[70]	Crashes in Chennai (2016-2018)	7654 crashes (75% for training and 25% for testing)	8	RF and Cforest	4-classes (fatal injury, grievous injury, simple injury, and non-injury)	Test error is 46%
[57]	Crashes in Iran (2011-2014)	8390 crashes (10-fold stratified cross validation)	9	RPBL and SVM	2-classes (non-fatal crashes and fatal crashes)	Average AUC of SVM is 73.5%

2.3 Summary

It can be observed that the research community has been developing road traffic crash severity prediction using machine learning models mainly focusing on a specific dataset. The proposed study will develop a deep learning (DL) model using several countries database based on the publicly available resources. It is assumed that the DL model will provide a high-level understanding of the features. Therefore, the model is expected to perform well even with limited training datasets for new locations. This study will report a unique approach to solve the local problems using the global knowledge related to traffic crash severity. The challenges associated with the global knowledge of traffic crash severity is the data from several countries which will be coupled with the DL model.

CHAPTER 3

DATA

In this chapter we dive in the details of data sources, characteristics considered in our study, variables adopted under each characteristic, unification of data from different sources and the international data to be further developed then be used as an input in the modeling stage.

3.1 Raw data from sources

As this study states, an objective of building a crash severity model for multinational use was set, in addition to paving the path for international application of such a model, we collected the data available through open sources for four different countries: USA, Canada, GB, and Australia. Following subsections will address the data of each country in further details.

3.1.1 Australia Data

The data for crashes in Victoria, Australia was made available through a governmental open-source website. Victoria is a state located in the southeast of Australia with Melbourne as its' capital and a population of nearly 6.7 million people. Collected data dates from Jan 2006 to Aug 2019, almost 14 years with a database size of approximately 143,053 samples and nearly 43 variables. *Table 3* below outlines the chief crash variables and their types of Australia.

Table 3 Most relevant crash data for Australia

Characteristic	Most relevant variables	Type & number of classes of variable
Crash factor	Accident No	Numeric
	Accident Date	Numeric
	Accident Time	Numeric
	Accident Type	Nominal (9 classes)
	Day of the week	Nominal (7 classes)
	Severity	Nominal (3 classes)
	Number of vehicles involved	Nominal (2 classes)
User factor	Sex of driver	Nominal (3 classes)
	Age group of driver	Nominal (13 classes)
	Injury level	Nominal (4 classes)
	Number of persons injured	Numeric
	Number of persons killed	Numeric
Vehicle factor	Vehicle manufacturing year	Numeric
	Vehicle type	Nominal (24 classes)
Environment factor	Light Condition	Nominal (9 classes)
	Atmosphere Condition	Nominal (8 classes)
Roadway factor	Road Geometry	Nominal (9 classes)

Speed zone	Nominal (13 classes)
Surface Condition	Nominal (6 classes)
Road surface type	Nominal (4 classes)
Traffic control type	Nominal (17 classes)

3.1.2 Canada Data

The data for crashes in Canada was made available online through the national collision database. Collected data dates from Jan 2014 to Dec 2016, nearly 3 years with a database size of approximately 581,922 samples and nearly 22 variables. *Table 4* shows the focal crash variables for Canada.

Table 4 Most relevant crash data for Canada

Characteristic	Most relevant variables	Type & number of classes of variable
Crash factor	Collision Year	Nominal (3 classes)
	Collision Month	Nominal (13 classes)
	Collision day of week	Nominal (8 classes)
	Collision hour	Nominal (25 classes)
	Collision Severity	Nominal (3 classes)
	Collision Configuration	Nominal (19 classes)
	Number of vehicles involved	Numeric

User factor	Person sequence number	Numeric
	Person Sex	Nominal (3 classes)
	Person Age	Numeric
	Person Position	Nominal (14 classes)
	Person Injury severity	Nominal (4 classes)
	Person safety used	Nominal (9 classes)
	Person classification	Nominal (6 classes)
Vehicle factor	Vehicle sequence number	Numeric
	Vehicle model year	Numeric
	Vehicle type	Nominal (20 classes)
Environment factor	Weather Condition	Nominal (9 classes)
Roadway factor	Roadway configuration	Nominal (14 classes)
	Road surface	Nominal (11 classes)
	Road alignment	Nominal (8 classes)
	Traffic control	Nominal (20 classes)

3.1.3 Great Britain Data

The data for crashes in Great Britain was made available through a governmental open-source website. Great Britain is composed of England, Scotland and Wales, without North Ireland. Collected data dates from 2011 to 2016, nearly 6 years with a database size of

approximately 439,302 samples and nearly 50 variables. *Table 5* displays most relevant crash variables for Great Britain.

Table 5 Most relevant crash data for GB

Characteristic	Most relevant variables	Type & number of classes of variable
Crash factor	Accident Index number	Numeric
	Accident Severity	Nominal (3 classes)
	Number of vehicles	Numeric
	Number of casualties	Numeric
	Day of week	Nominal (7 classes)
	Time	Numeric
User factor	Journey purpose of driver	Nominal (7 classes)
	Sex of driver	Nominal (3 classes)
	Age band of driver	Nominal (10 classes)
Vehicle factor	Vehicle type	Nominal (28 classes)
	Vehicle Maneuver	Nominal (18 classes)
	Vehicle location	Nominal (11 classes)
	Junction Location of vehicle	Nominal (9 classes)
	Skidding and overturning	Nominal (6 classes)
	First point of impact	Nominal (5 classes)

	Age of Vehicle	Numeric
Environment factor	Light condition	Nominal (5 classes)
	Weather Condition	Nominal (9 classes)
Roadway factor	Road class	Nominal (6 classes)
	Road Type	Nominal (7 classes)
	Speed limit	Nominal (6 classes)
	Junction detail	Nominal (9 classes)
	Junction control	Nominal (5 classes)
	Pedestrian cross human control	Nominal (3 classes)
	Pedestrian cross facility	Nominal (6 classes)
	Road surface condition	Nominal (7 classes)
	Special condition at site	Nominal (8 classes)
	Carriageway hazards	Nominal (8 classes)
	Urban or rural area	Nominal (2 classes)

3.1.4 United States Data

The data for crashes in Allegheny County, Pennsylvania was made available through public open-source data. Allegheny is Pennsylvania's second most populous county with nearly 1,250,278 residents (2020) located in the state southwest side and encompassing the city of Pittsburgh. Collected data dates from Jan 2004 to Dec 2017, nearly 14 years with a

database size of approximately 170,358 samples and nearly 190 variables. The US significant crash variables are summarized in *table 6*.

Table 6 Most relevant crash data for USA

Characteristic	Most relevant variables	Type & number of classes of variable
Crash factor	Crash record number	Numeric
	Crash year	Nominal (14 classes)
	Crash Month	Nominal (12 classes)
	Day of week	Nominal (7 classes)
	Hour of day	Nominal (25 classes)
	Collision type	Nominal (10 classes)
	Vehicle count	Numeric
	Fatal count	Numeric
	Injury count	Numeric
	Max severity level	Nominal (7 classes)
User factor	Drinking Driver	Nominal (2 classes)
	Driver age { 16yr, 17yr, 65_74 year, 75plus }	Nominal (2 classes/ variable)
	Aggressive driving	Nominal (2 classes)
	Cell phone	Nominal (2 classes)

	Fatigue asleep	Nominal (2 classes)
	Impaired driver	Nominal (2 classes)
Vehicle factor	Vehicle type (Multi variable) {bicycle, bus, commercial vehicle, heavy truck, motorcycle, pedestrian. Van) count	Numeric/ each variable
Environment factor	Illumination	Nominal (9 classes)
	Weather	Nominal (9 classes)
Roadway factor	Road condition	Nominal (10 classes)
	Relation to road	Nominal (9 classes)
	Intersect type	Nominal (12 classes)
	Traffic control device type	Nominal (10 classes)
	Urban Rural	Nominal (4 classes)
	Location type	Nominal (10 classes)
	Roadway surface type code	Nominal (9 classes)
	Traffic control device state	Nominal (6 classes)
	Work zone type	Nominal (3 classes)
	Work zone location	Nominal (6 classes)
	Speed limit	Nominal (16 classes)

3.2 Data preparation/Processing

As observed from the previous section we have divided the variables into five major factors: crash, user, vehicle, environment, and road. With diversity of data collection location (some were county level, some state level and others country wide level), also the fact that each country database has its unique way of expressing data whether in terminology or in the weight – number of variables - it places on each of the factors, the difference in years data was collected, the class length variations as well for the different variable numbers and ways of presentation all of these pose a challenge, we shall tackle in this section to achieve the international dataset as a result.

3.2.1 Finding the common/ most relevant variables

As mentioned earlier after setting the five broad categories with the huge variance in the variables number from one country to the other, one can notice the repetition of certain types of variables in the different country sets, which is an indicator of the vitality and significance of such elements, and we shall recognize them in this subsection.

- Crash factor: {Accident ID, Accident Year, Accident Month, Day of week, Accident time, Accident type, Severity level, Vehicle count}, {Deaths count*, Injuries count*}.
- User factor: {Age of Driver},{Sex of Driver*}.
- Environment factor: {Weather Type}, {Light condition*}.
- Vehicle factor: {Vehicle Type}, {Model of Vehicle*}.
- Road factor: {Road surface condition, Traffic control type, Road geometry}, {Speed Limit*}.

The selected variables are twenty variables, fourteen of which are common in all the four countries, as for the remaining six variables*, they are added considering their high relevance and being mutual between three countries out of four as the last will be substituted as unknown.

3.2.2 Uniting the different classes of a variable under a new generalized classes arrangement

After finding out our twenty common variables, it is the time to unify the variable classes for all countries for a logical data use and to achieve consistency, to prepare for the next step which is data refining and merging.

The process is quite simplistic and includes two steps for checking and tuning, perhaps providing examples will make it easier for the reader to grasp:

- **Check 1: Number of classes.** Classes of road surface condition 6, 11, 6 and 10 for Australia, Canada, GB and USA respectively. In such case we adopt the classification that allow a wider range class with a lower number of classes to be able to fit the smaller range classes, meaning we end up having 6 classes adding one for the “unknown” class we get a total of 7 classes
- **Check 2: Compatibility of variable type throughout data.** For instance, driver age is nominal both in Australia and GB, numeric in Canada, and represented by not just by one variable in the US but with different variables keeping independent count for each. In such a case the best solution that comes to mind is to categorize these variables in ranges (make them nominal) and deal with one variable only instead several to make it easier for the model to process and the user to

comprehend. Once again, we produce a unified variable with six classes and book the seventh class for “unknown”.

It is worth mentioning that the considered non-sever (Class-0) would express accidents involving property damage only, or accidents involving minor personal injuries like bruises (does not require extended medical attention). While severe (class-I) expresses accidents involving fatalities, losing limbs, impairment and bone fracture (requires extended medical attention 3 days onwards).

3.2.3 Cleansing and compiling the four countries data in one sheet

In preparation for data consolidation, we take each data set independently and start eliminating outliers from each country set individually, an outlier to define is any data point that does not fall within the natural distribution for the population (per set). Moving forward towards compiling the data and to avoid any underling biases we take an equal size but random data from each country. For the purpose of this study the data size per country is 69,000 considered for merging meaning the international data set contains 276,000 data point with each individual point comprising of 20 variables. Also, the near time span is kept in consideration, so we tried to make the years difference between the four countries marginal by choosing the most common years possible.

3.3 International data

This section depicts in *table 7* the broad lines for the data included in the international version the variables, their definitions and coding.

Table 7 International version of variables

#	Column Name	Definition	Code	Description
---	-------------	------------	------	-------------

1	ACCIDENT_ID	Accident record number		
2	ACCIDENT_YEAR	Year when the accident occurred	20XX	Australia (2011-2019), Canada (2014-2016), Great Britain (2011-2016), and USA (2011-2018)
3	ACCIDENT_MONTH	Month when the accident occurred	1	January
			2	February
			3	March
			4	April
			5	May
			6	June
			7	July
			8	August
			9	September
			10	October
			11	November
			12	December
4	DAY_OF_WEEK	Day of the Week code when accident occurred	1	Sunday
			2	Monday
			3	Tuesday
			4	Wednesday
			5	Thursday

			6	Friday
			7	Saturday
5	ACCIDENT_TIME	The time of day when the accident occurred	HH	Hours
6	ACCIDENT_TYPE	Collision category that defines the accident	1	Collision with vehicle
			2	Struck Pedestrian/animal
			3	Collision with a fixed object
			4	Vehicle overturned and Fall from
			5	Other accident and Collision with some other object
			6	Not given
7	SEVERITY_LEVEL	Severity level of the accident	0	non-severe (slight injury and no injuries)
			1	Severe (fatal, major and moderate injury)
8	WEATHER_TYPE	Code for the weather type at time of accident	1	Clear and sunny
			2	Raining
			3	Snowing
			4	Visibility limitation e.g. drifting snow,

				fog, smog, dust, smoke, mist
			5	Freezing rain, sleet, hail
			6	Strong winds
			7	Other
			8	Not given
9	ROAD_SURFACE_CONDITION	Roadway Surface Condition Code	1	Dry
			2	Wet
			3	Sand/ mud/ dirt/ oil/ or gravel
			4	Snowy
			5	Icy
			6	Other
			7	Not given
10	TRAFFIC_CONTROL_TYPE	Code that defines the Traffic Control	1	Traffic signals in flashing mode
			2	Traffic signals fully operational
			3	Stop sign
			4	Yield sign
			5	School crossing
			6	Pedestrian crosswalk
			7	Reduced speed zone and No passing zone sign

			8	Railway crossing
			9	Police officer or flagman
			10	No control
			11	Other types of traffic control
			12	Not given
11	LIGHT_CONDITION	Code that defines lighting at accident scene	1	Day
			2	Dusk/Dawn
			3	Dark Street lights on
			4	Dark Street lights off
			5	Dark No street lights
			6	Dark Street lights unknown
			7	Other
			8	Not given
12	ROAD_GEOMETRY	Code that defines the road and intersection type	1	Cross intersection
			2	T/Y intersection
			3	Multiple intersection
			4	Not at intersection
			5	Private property, Dead end and Road closure

			6	Roundabout
			7	Other
			8	Not given
13	VEHICLE_TYPE	Code that defines the type of the involved vehicle	1	Car
			2	Bicycle/Motorcycle
			3	Bus
			4	Truck
			5	Sport Utility Vehicle/VAN
			6	Others
			7	Not given
14	VEHICLE_COUNT	Total amount of all vehicles involved	Number	number of involved vehicles
			X	Not given
15	SEX_OF_DRIVER	It represents the sex of the driver for the main vehicle	M	Male
			F	Female
			U	Not given
16	AGE_OF_DRIVER	It represents the age of the driver for the main vehicle	1	less than 17
			2	from 17 to 21

			3	more than 21 and to 50
			4	more than 50 and to 64
			5	more than 64 and to 74
			6	more than 74
			7	Not given
17	MODEL_OF_VEHICLE	It represents the model of the main vehicle	1	less than 1990
			2	from 1990 and less than 2000
			3	from 2000 and less than 2005
			4	from 2005 and less than 2010
			5	from 2010 and less than 2015
			6	equal 2015 or more than 2015
			7	Not given
18	SPEED_LIMIT	Speed limit of the road when the accident occurred	1	less than 40
			2	from 40 and to 60
			3	more than 60 and to 80
			4	more than 80 and to 100

			5	more than 100 and to 120
			6	others
			7	Not given
19	DEATHS_COUNT	Total amount of fatalities involved	Number	number of killed persons
			X	Not given
20	INJURIES_COUNT	Total amount of fatalities involved	Number	number of injured persons
			X	Not given

3.4 Summary

Throughout this chapter we have traveled with the data starting from their sources, to identifying their characteristics as raw data, then we had a walk down through the data preparation and processing stage making it to final stage of international data which is only the start for the next chapter of modeling where we will make data into a more valuable outcome that can be compared and measured.

CHAPTER 4

METHODOLOGY

In this chapter we go through the methodology applied in this study, starting by an introduction to the machine learning topic, followed by a discussion about classification problems and the binary type we have, then take an overview of the system. After that we discuss the data used and the phases, they go through to get them ready to be fed to the ml model. Finally, we get a closer look on each of the selected seven ML models.

4.1 Introducing ML topic

The concept of machine learning relies on the execution of artificial intelligence (AI) and it basically enables a system to perform self-learning and evolve by experience without a need for being directly programmed. ML is concerned of developing computer applications that can read data and employ it for self-education.

The beginning of learning process can start with direct experience, or data, or instruction, or observations, then locate the underlying relations and patterns in data to make enhanced verdicts in the future based on instances we offer. Automatic computer learning and action adjustment accordingly without human interference or assistance is the main goal.

ML can be used for prediction in two different approaches: one is regression and the other is classification. As our problem of crash severity represents a classification problem where we have two possible outcomes, i.e., severe or non-severe we will be focusing on classification using ML in the following paragraphs.

4.2 Classification Problem

Classification [72] is model of prediction where a class label is forecasted for a given instance of input data. For instance, infected or uninfected, which has two outcome classes. In our case the output classes are severe or non-severe.

As a requirement for modeling, classification requires a rich data set with many examples of inputs and outputs to achieve model learning. Usually, the training set will be used by the model in calculating how best fit the input data instances in specific class labels. Given that the data set used for training shall be sufficient to represent the problem and provide several instances for every class label.

String (nominal) values are linked to numeric values prior to being fed to an algorithm for simulation in the process known as label encoding, where each class label receives a unique integer, as: 'severe' = 1, 'non-severe' = 0.

The classification models are classified depending on their expected number of outcomes into four main types: binary, multi-class, multi-label, and imbalanced classification. In our study we have a case of binary classification as we have two outcome classes, also our data set is imbalanced meaning that an input example normal class is dominant over the other. It is worthy to say that imbalanced classification is considered a special case of binary classification.

Binary classification is better recognized by the presence of two class labels, like:(server or not). Normally this kind of classification sets one class as the standard state and the other as the irregular state. For example, severe can be considered as the abnormal state and denoted by the integer 1 and non-severe as normal with denoting of a 0 integer.

Commonly, the representation of binary classification problem is expressed through a Bernoulli distribution of probability which is a discrete one. Algorithms that best fit binary classification include: SVM, DT, NB, K-NN and LR.

4.3 System Overview

The following *figure 1*, displays the system in which data is entered, pre-processed, fed to the ML model until result is reached.

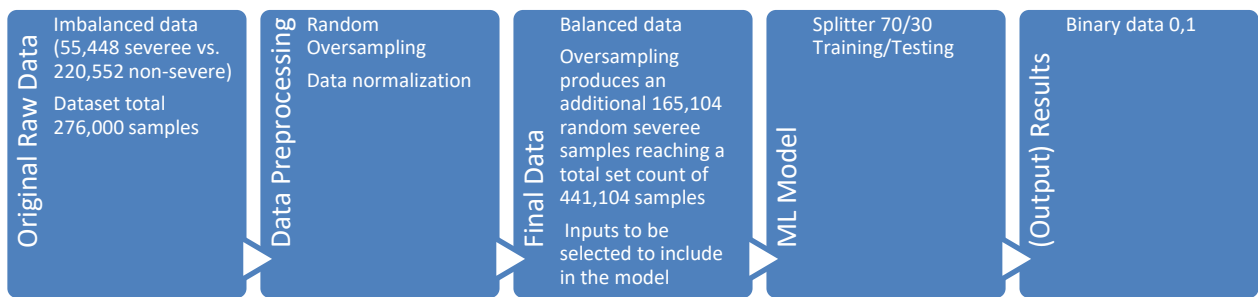


Figure 1 System Overview

4.4 Data

As we previously mentioned in chapter 3, we are using the data from four different source (countries), in order to eliminate biases that may result from oversampling of source over the others, we set a quota of individual countries to 69,000 each resulting of a total sample size of 276,000 data points for merged international with each data pint being comprised of 20 variables. Also, the near time span is kept in consideration, so we tried to make the years difference between the four countries marginal by choosing the most common years -possible-.

However, the data still suffers of a shortfall which is being imbalanced, as we have mentioned in the previous section 2 of this chapter that the data is skewed towards one

class over the other. In the following paragraphs we will explore the imbalanced data and how to treat it for classification purposes.

4.4.1 Data Preprocessing

When working with a classification problem involving imbalanced data the examples distribution among classes would be skewed or biased, the skew might be mild or severe, such that the minority group examples would be limited to a single digit figure whereas those of majority would be in three, four or even six digits. This type of data skew can be challenging to ML predication models, as they usually consider an even distribution of examples among different classes. As a result, using untreated imbalanced data will lead to a faulty prediction act, especially for the minority group which in most cases is the one of highest significance for the researcher, for instance, the severe class in our case.

Many reasons are involved in the case of imbalanced data but usually they fall under two main categories, these are: domain properties and data sampling. A probable example of imbalance cause is the method of data collection or sampling which may have biases or errors during either sampling or collection processes like using a small geographical area or short span of time.

A technique, known to be naïve yet efficient in terms of output accuracy is resampling of imbalanced data which can be either under sampling or oversampling. We considered **random oversampling** for this study.

Oversampling to start with, is a process of randomly cloning the samples of the minority class in the training set to match their count with number of samples in the majority group, in other word our data used for model fitting becomes balanced data. One thing to keep in

mind, is that the testing data set does not get affected by the resampling process, i.e., minority and majority class portions remain identical to that of original data set before resampling.

In our case knowing that non-severe crash is a majority group with 220,552 samples and severe crashes are a minority with only 55,448 samples using the oversampling technique an additional 165,104 random clones are required to reach balance point. Doing so, will get our balanced dataset size to 441,104 samples which will be considered for splitting.

The last procedure to be done before getting the final data ready to be fed to the splitter then to the ML model is **normalization** [73]. Normalization is a process of arranging data following a normal distribution giving them similar appearance for different records and fields.

Normalization employs different machines, two of which are standard scaler and minmax scaler. Minmax reflects the value if each entry in original data relatively to a range between 0 and 1, this can be useful when treating data feature as it sustains the shape of the set. Whereas standard scaler reflects the values of entries with respect to a mean of zero and a standard deviation of 1, this is best advised to use pre-knowing the data follows normal distribution.

$$MinMax = \frac{x - Min}{Max - Min} \quad (4.1)$$

$$Standard\ scaler\ z = \frac{x - \bar{X}}{Standard\ Deviation} \quad (4.2)$$

One of the well renowned validation methods of ML models is K-fold cross validation. Let us have an overlook on it in the following paragraphs.

The statistical procedure of resampling used for ML models' assessment on a limited dataset is referred to as cross validation. This procedure relies on a single parameter called k which corresponds to the groups number which a dataset will be broken into. Which explains the name k -fold cross validation where k is a real number (usually an integer).

Principally, the cross validation is used to test how well the ML learning performs on unobserved data, which basically tries to comprehend how the model will act in forecasting the data which were not included in the training set. The reason behind the popularity of cross validation is its simplicity, easy comprehension and it gives lower bias margin compared to other methods as naïve train/test division.

The general arrangement of the method goes as follows:

- An arbitrary shuffle of dataset
- Followed by the breaking of data set into k -number of groups
- Then for every distinctive group:
 - a) Consider that group as test data set (held off)
 - b) Remaining groups are considered a training dataset
 - c) Model fitting on the training set then use the test set for model evaluation
 - d) A record is kept of the assessment score and abandon the model
- The model skill is finally reported as a summary of assessment scores.

During the procedure each data point in the general dataset remains in the same individual group it is assigned to meaning it is used for testing once and for training as $k-1$ times.

Setting the K value

High attention shall be paid while assigning a value for the factor k , a poor selection of the value may lead to misrepresentation for the model skill inference. For example, a high variance score (has a chance of extreme changes when changing the input data for model fitting) or high bias score (provides an overrated model skill score).

Normally three strategies are common for the k value selection:

- a) Representative: The value is picked such that each data stratum is sufficiently large to represent -statistically- the global set it is derived from.
- b) $K=10$, empirically chosen the value of 10 showed a balanced model skill score with low bias and variance values.
- c) $K=n$, setting the k value to the sample size, this results in a high representative model skill score as it considers each element as a hold stratum, but is more time consuming compared to the other methods. This type is called leave out cross validation.

4.4.2 Final Data

As established in the first part of this section the final dataset size to be used for splitting to achieve model fitting is 441,104 samples.

To take this a step further we revise the number of variables to remove redundant/ repeated inputs, taking a look into the variables we can infer that, [Accident ID, Accident year, Country] have a minimal impact to the crash severity value. [Injuries count, Deaths count]

express the same idea of severity being non-severe/severe which implies their redundancy. All of these five attributes will be neglected from this point going forward leaving another 16 variables to be considered including the dependent variable “severity”, the following *table 8* outlines the variables

This procedure not only refines the model inputs to achieve better prediction quality, but also makes it easier for researchers that want to benefit from this study for future work as it keeps the most significant variables and does not require a high number of variables making it applicable to growing countries known for data scarcity and limited numbers of variable that can serve prediction.

Table 8 Variables considered in final data

#	Column Name	Description
1	ACCIDENT_MONTH	January
		February
		March
		April
		May
		June
		July
		August
		September
		October
		November
		December

2	DAY_OF_WEEK	Sunday
		Monday
		Tuesday
		Wednesday
		Thursday
		Friday
		Saturday
3	ACCIDENT_TIME	Hours
4	ACCIDENT_TYPE	Collision with vehicle
		Struck Pedestrian/animal
		Collision with a fixed object
		Vehicle overturned and Fall from
		Other accident and Collision with some other object
		Not given
5	SEVERITY_LEVEL	non-severe (slight injury and no injuries), marked as 0
		Severe (fatal, major and moderate injury), marked as 1
6	WEATHER_TYPE	Clear and sunny
		Raining
		Snowing
		Visibility limitation e.g. drifting snow, fog, smog, dust, mist
		Freezing rain, sleet, hail
		Strong winds

		Other
		Not given
7	ROAD_SURFACE_CONDITION	Dry
		Wet
		Sand/ mud/ dirt/ oil/ or gravel
		Snowy
		Icy
		Other
		Not given
8	TRAFFIC_CONTROL_TYPE	Traffic signals in flashing mode
		Traffic signals fully operational
		Stop sign
		Yield sign
		School crossing
		Pedestrian crosswalk
		Reduced speed zone and No passing zone sign
		Railway crossing
		Police officer or flagman
		No control
		Other types of traffic control
		Not given
9	LIGHT_CONDITION	Day
		Dusk/Dawn

	Dark Street lights on
	Dark Street lights off
	Dark No street lights
	Dark Street lights unknown
	Other
	Not given
10 ROAD_GEOMETRY	Cross intersection
	T/Y intersection
	Multiple intersection
	Not at intersection
	Private property, Dead end and Road closure
	Roundabout
	Other
11 VEHICLE_TYPE	Not given
	Car
	Bicycle/Motorcycle
	Bus
	Truck
	Sport Utility Vehicle/VAN
	Others
12 VEHICLE_COUNT	Not given
	number of involved vehicles
13 SEX_OF_DRIVER	Male

	Female
	Not given
14	AGE_OF_DRIVER
	less than 17
	from 17 to 21
	more than 21 and to 50
	more than 50 and to 64
	more than 64 and to 74
	more than 74
	Not given
15	MODEL_OF_VEHICLE
	less than 1990
	from 1990 and less than 2000
	from 2000 and less than 2005
	from 2005 and less than 2010
	from 2010 and less than 2015
	equal 2015 or more than 2015
	Not given
16	SPEED_LIMIT
	less than 40
	from 40 and to 60
	more than 60 and to 80
	more than 80 and to 100
	more than 100 and to 120
	others
	Not given

4.5 ML Models

4.5.1 K-Nearest Neighbors (K-NN)

A machine learning algorithm known for being simplistic and straight forward in execution and falling under the category of supervised learning, applicable both to classification and regression problems.

The supervised term in the definition implies that the algorithm acts as a child that learns from a parent or guardian meaning in order for it to come up with the correct results for raw data without labeling it depends on receiving input data with labeling to learn from and create a function in the first place.

The *figure 2*, below explains the principle upon which KNN operates, as can be noticed the model considers a graphical presentation of the problem, -in our case in classification- the value of the star depends on the number of surrounding elements that lie within the boundary line set by the value of **K**, assuming the value is equal to 4 in case of the solid line we count the nearest four neighbors to the star and they are: one circle, two triangles, and one square. After doing so the star assumes the value of the neighbor with most repetitions in this case a triangle and this is called a collective voting. Whereas, if we consider a value of $k = 6$ (dotted line), the nearest six neighbors become: one circle, two triangles, and three squares, obviously the result is concluded to a square.

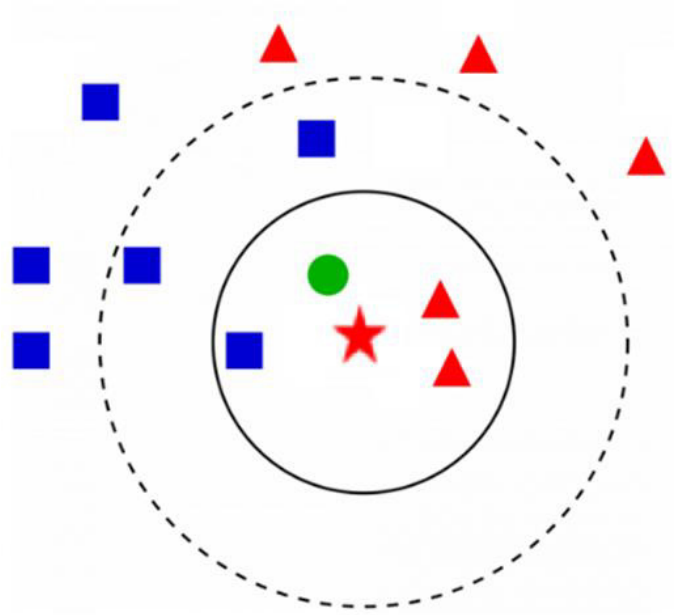


Figure 2 KNN concept graphical representation

The way KNN operates can be presented using two mathematical equations [74]:

1. Euclidean distance, to measure the distance between any respective points assuming point A is the one in question and point B is where the nearest neighbor is located as shown in *figure 3* below.

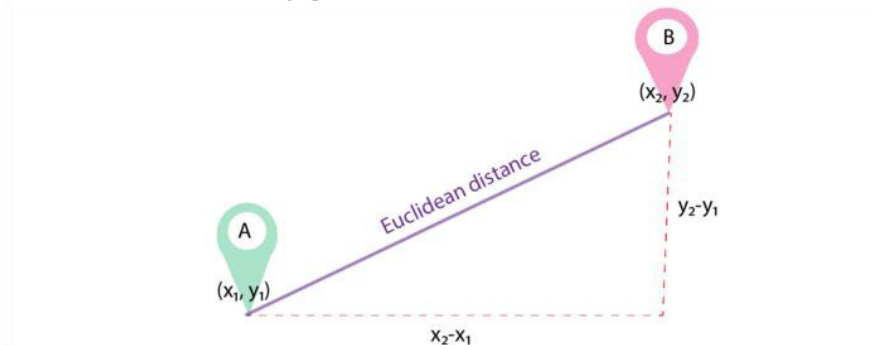


Figure 3 Euclidean distance between points

$$\text{Distance (A, B)} = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2} \quad (4.3)$$

2. The probability of the point in question to be assigned to the nearest neighbor with the highest number of points (repetitions) is expressed as:

$$P(y = j|X = x) = \frac{1}{K} \sum_{i \in A} I(y^{(i)} = j) \quad (4.4)$$

Where,

$P(y=j|X=x)$, is the probability to assign y the value of j given that x of the point is equal to that of neighbor

K, is the preset value for the number of nearest neighbors

i, number of elements

A, actual number

I, iteration

Lastly, we go through the pros and cons of K-NN in *table 9*

Table 9 Pros and Cons of K-NN

Pros	Cons
Easy to comprehend	Saves whole training data set which can be time consuming in calculation
Works with nonlinear data as it has no underlying biases	Relevant to previous point high storage capacity required
Multifunctional; can be used both for regression and classification problems	Slow computation when presented with large sample size
Predicts with great accuracy	Data scale sensitive

4.5.2 Support Vector Machine (SVM)

Also, a supervised machine learning technique usually present in the problems of regression and classification in addition to outliers' detection. This model makes a transfer

of raw data to higher dimension feature space. The target behind SVM is to reach the optimal hyperplane maximizing support vectors margin.

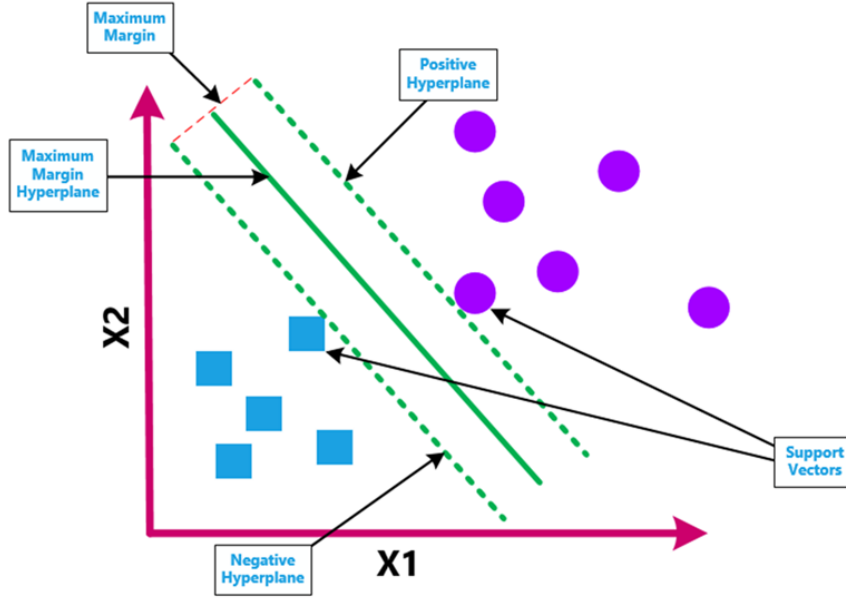


Figure 4 SVM Hyperplane

As shown in *figure 4*, data points coinciding to hyperplane, are called support vectors these are recognized by alterations to the hyperplane if they -for any reason- are removed or changed. The confidence in hyperplane correct data classification is increased with a greater margin between support vectors. Also, it can be inferred that SVM uses data grouping separation approach for classification purposes.

Relevant to our study subject all the crash parameters are explained by vectors ($x_i \in R_n$) for $[i = 1, 2, \dots, n]$, where the severity level which is the output in quest can be written as ($y_i \in R_n$). The hyperplane formula [75] with respect to set of point X can be expressed as

$$W \cdot X - b = 0 \quad (4.5)$$

which is a dot product with the normal vector W orthogonal to the hyperplane.

SVM optimization is required for binary classification problems with a provided array of input and output variable pairs (x_i, y_i) . Misclassification errors [75] are measured using slack variables ξ , the penalty factor C is employed for classifier capacity control enhancement errors.

$$\min_{w, b, \xi} \frac{1}{2} w^T w + C \sum_{i=1}^N \xi_i \quad (4.6)$$

$$y_i(w^T \phi(x_i)) + b \geq 1 - \xi_i; \quad \xi_i \geq 0 \quad (4.7)$$

Lastly, we go through the pros and cons of SVM in *table 10*.

Table 10 Pros and cons of SVM

Pros	Cons
Successful in high dimensional spaces	No explicit probability estimates are provided by SVM instead the employment of five-fold cross validation is required additionally
Efficient even when the dimensions outnumber the samples	
Low memory consumption as it uses a division of training points in the decision function	Might get into over fitting in case the features outnumber the samples. Therefore, the correct choice of regularization term and kernel function is essential
Has versatility in terms of kernel functions	

4.5.3 Random Forest (RF)

Another instance of supervised machine learning, and one of the commonly used techniques in classification and regression problems. The RF is an ensemble learning algorithm meaning it improves its prediction accuracy by stacking a considerable number of classifiers. The classifiers are simply the decision trees (DT) in this case which are

grouped together to create a forest then most of these DTs outcomes are taken yielding better prediction results.

The following *figure 5*, shows the structure of a random forest and how it operates, which to mention basically starts by selecting data points from the complete set randomly, then using them to construct the DTs with the selection of a number for those DTs, then getting an outcome from each individual DT, perform a collective voting for trees results and lastly assigning a class using the result with the highest votes.

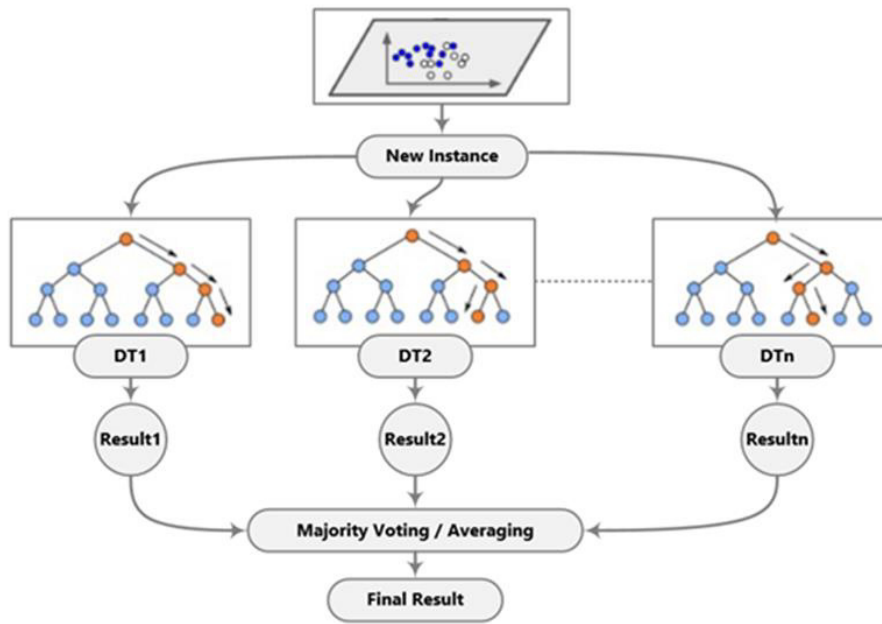


Figure 5 RF structure and process flow

RF can be expressed mathematically as follows [76]:

$$RF = \underset{j \in \{1,2,\dots,n\}}{\operatorname{argmax}} \sum_{i=1}^i DT_{i,j} \quad (4.8)$$

Where.

Argmax, is the function max value -majority vote-

i , is the DT number starting at 1 and ending with i th DT

j , the data number of classes

Lastly, we go through the pros and cons of RF in *table 11*.

Table 11 Pros and cons of RF

Pros	Cons
High precision as it depends on a huge number of DTs	Time expensive as each input has to go through a long process including individual DT prediction then voting
Marginal to none overfitting as the voting/averaging action cancels biases	
Works with data gaps (missing values)	Not as easy as a single DT to explain and follow through
Good tool for finding the relative significance of features	

4.5.4 Decision Tree (DT)

Also known as classification and regression trees, CART for short. As the name we just mentioned suggests the model works both for classification and regression, and is considered among the supervised ML. To understand the way this model works, we assume that we have three sorts of fruits, i.e., apples, bananas, and kiwis to start classifying them in three groups we set rules, take shape as the first one and it has three possible outcomes roundish, elongated, and ovoid respectively, then we can make it more interesting by adding the rule of color with possible results being [red, green, yellow] , [green, yellow], and [brown] respectively. Lastly, we consider a rule for skin texture with smooth and fuzzy as possible outcomes. These rules are equal to decision nodes (DN) in this algorithm with the outcomes being either other DNs if they have outcomes below them or leaves if they are the last outcome. *Figure 6*, shows the general structure and process of DT.

DT can be expressed mathematically as follows [77]:

$$\text{Gini Impurity } GI = 1 - \text{Gini} = 1 - \sum_{i=1}^n p_i^2 \quad (4.9)$$

Where, p_i is the portion of item in i th class.

To calculate best split weighted summation of GI for both child nodes and this shall be done to all possible splits then the one with lowest GI is said to be the best split.

Another mathematical expression of DT is entropy also.

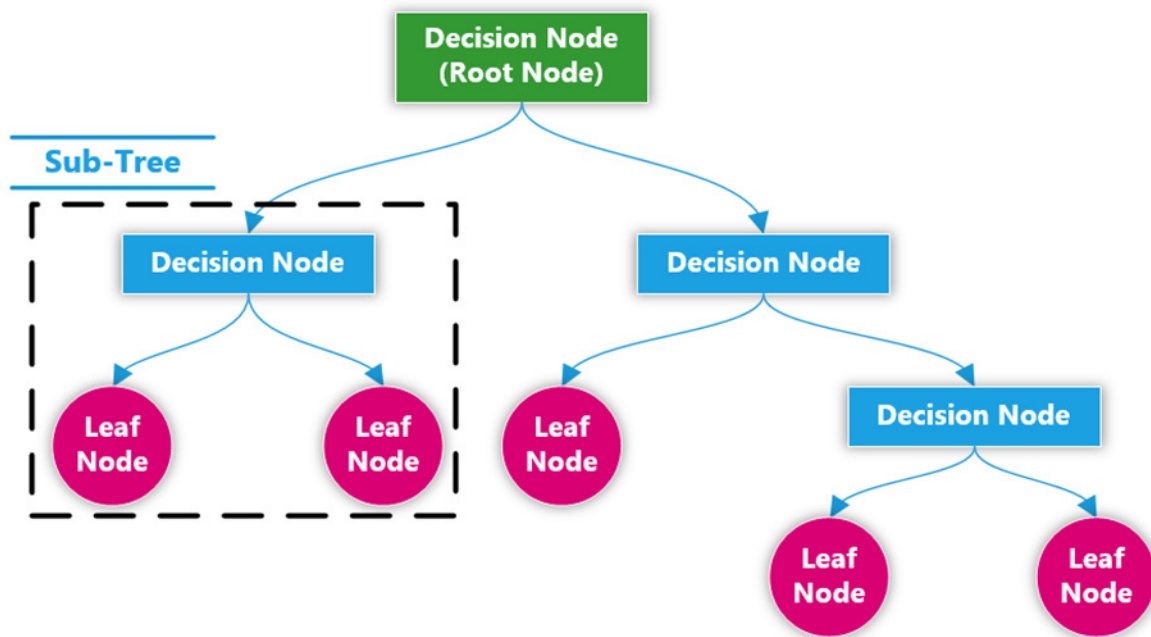


Figure 6 Structure of DT/CART

Lastly, we go through the pros and cons of DT in *table 12*.

Table 12 Pros and cons of DT

Pros	Cons

Easy to understand and simple	Can run in overfitting meaning a creation of over compound DT which are not able to generalize data
Performs feature selection	Can experience noticeable variance with small data changes
Works well with categorial and numerical data	If a class in data is a majority DT learners tend to create biased algorithms

4.5.5 Naïve Bayes (NB)

A supervised ML technique of classification derived from Bayes' Theorem with a condition of predictors independence meaning a presence of a certain feature is not dependable on the presence or absence of others. To understand how it works, we shall have an example, assume we have a weather forecast for ten days as in *table 13* below

Table 13 Example on NB

Weather	Action
Cloudy	Go for picnic
Windy	Stay home
Clear	Go for picnic
Sunny	Go for picnic
Haze	Stay home
Stormy	Stay home
Rainy	Stay home
Clear	Go for picnic
Cloudy	Go for picnic
Sunny	Go for picnic

What is done next is creating a histogram or frequency table for each weather condition and related action for instance, cloudy/go for picnic gets a frequency of 2, windy/stay home gets a frequency of 1 and so forth. After this has been done the individual frequency for each category from the previous ones is calculated for cloudy/go... will be 0.2 on so on. Lastly, we apply NB equation [75] to calculate posterior for each class, and the one with highest probability is assigned as the result class.

$$P(c|x) = \frac{P(x|c)P(c)}{P(x)} \quad (4.10)$$

Where,

$P(c|x)$ Posterior probability of class c given attribute x

$P(x|c)$ likelihood of attributed x for the class c

$P(c)$ individual probability of class c

$P(x)$ probabilty of attribute x

Lastly, we go through the pros and cons of NB in *table 14*.

Table 14 Pros and cons of NB

Pros	Cons
Low training data requirement	Known to be a bad estimator
Faster in prediction than most other methods	

4.5.6 Neural Network (NN)

A deep learning segment of ML, as the name suggests, are interconnected perceptron existing in layers, making them the building units similar to the neurons in the human brain, making the architecture use a node, link pattern. These units are in fact individual algorithms that use nonparametric functions to explain the relations between the data set elements. They are constructed mainly from an input (raw data receiver), output (outcome displayer) and hidden neuron layers to connect the first two. *Figure 7* below shows an example for a single hidden layer NN structure.

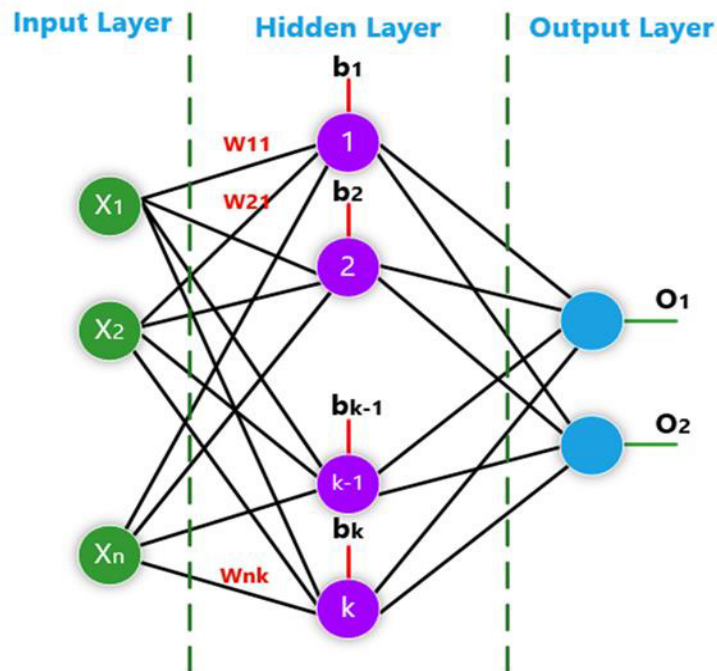


Figure 7 NN structure sample

In the hidden layer the raw data received from input neurons is process in each neuron individually on two phases; first being linear combination if input variables, then a non-linear activation function is exploiting the argument resulting from first phase. Every link

in ANN has a set bias term Θ and a weight term W_i . Activation function may be either a sigmoid function [78]

$$y = 1/(1 + e^{-x}) \quad (4.11)$$

or an identity function [78]

$$y = x \quad (4.12)$$

ANN algorithm pre-training hyperparameters include rate of learning, hidden layers number and batch size which require optimization, this determination is prior to the training process.

Also, it is noteworthy that there are two directions in which data are treated one forward that produces prediction and a reverse backward one to get feedback and adjust weights and bias terms as in *figure 8* [79].

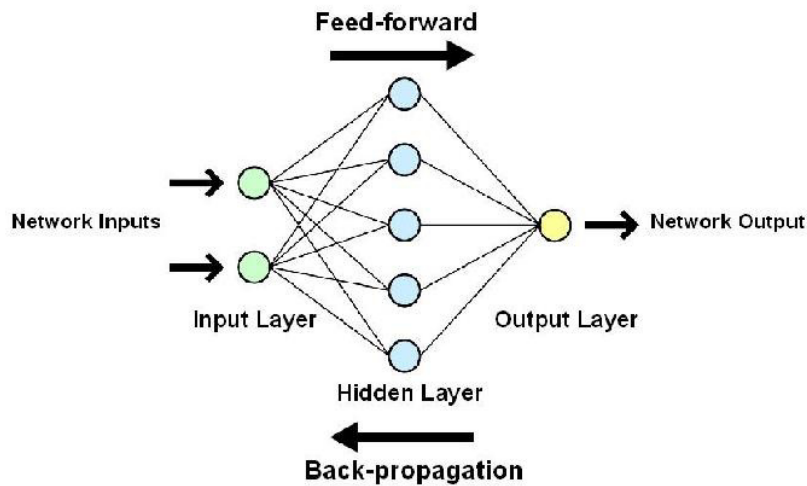


Figure 8 Feed forward and back-propagation in NN

Lastly, we go through the pros and cons of NN in *table 15*.

Table 15 Pros and cons of NN

Pros	Cons
After learning NN provides quick predictions	High training data requirement
Works with different types of data, discrete continuous	Requires high end computers to achieve quick training/ time expensive
Overcomes noise in training data	Black boxes no explanation behind any answer provided
	Requires time to be developed

4.5.7 Deep learning Model (DL)

A DL model is a special case of higher order neural networks, similar to an ANN the DL network consists of neurons, weights, biases, activation functions and synapses and falls under the ML family. What makes the DL model special is the number of internal (hidden) layers which act progressively together to achieve a deep feature interpretation compared to a single layer NN, the *figure 9* below shows a sample of multilayer NN.

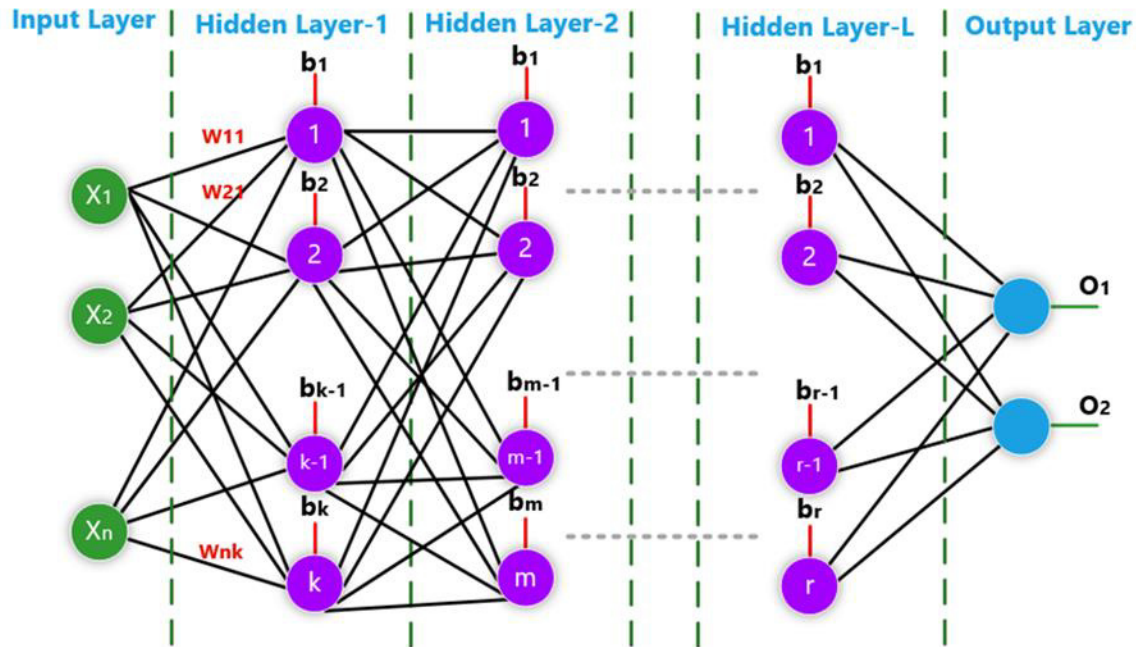


Figure 9 Example of DL NN

The fact that each neuron in the hidden layer processes the input data then passes the result it attains to the neuron in the next layer -of hidden- with a different bias term to do a reprocessing and process continues till inputs reach nth layer neuron then to the output layer, this all makes DL a more intelligent tool capable of explaining deeper features in data than a simple ANN, taking an image for example the ANN will assist by detecting edges only while DL WL explore deeper features as the pixels inside image and color weights, shapes and patterns, etc....

The most important parameters that affect the DL model performance are:

- 1- Number of hidden layers: we will mention how to assign optimal number of layers in chapter 5.
- 2- Number of neurons per hidden layer: this also optimized used a special search technique as point one with will be unveiled in chapter 5.

- 3- Batch size: refers to the number of samples fed to the model in each iteration, this size is important when considering the processing time.
- 4- No of epochs: these refer to the iterations that the model needs to reach an optimal solution and it should be noteworthy that the model usually builds up on each epoch result for the next meaning they are dependent on each other.
- 5- Optimizer type: considering two optimization types one is RMSprop and the other is Adam.

Lastly, we go through the pros and cons of NN in *table 16*.

Table 16 Pros and cons of DL

Pros	Cons
Flexible architecture	High training data requirement to get superior performance
The interchangeable use of the same NN approach for variant applications and data types	Requires high end computers to achieve quick training/ time expensive
Automatic feature deducing and tuning to achieve optimal outcomes	Acts as a black box with no traceability of a function how results are reached

4.6 Summary

In this chapter we dismantled the mechanism behind our study to make the theory a reality, we first started by introducing the machine learning and how it depends on the computer to develop an understanding like a human child, then we specified that the problem at hand is a classification problem which requires setting a label or a class for the outcome, then a schematic of the system was displayed to better understand the phases of data preparation and standardization we applied to make it adequate for ML model feeding (being

oversampling and normalization), lastly we spoke in detail about our seven ML model (KNN, SVM, RF, DT, NB, NN, and DL), mentioned the philosophy and equations behind each and finished by each model strengths and weaknesses.

CHAPTER 5

MODEL DEVELOPMENT AND RESULTS

After introducing the seven different models we considered for this study in previous chapter, it is time to dig deeper in this chapter by talking about the tuning of models’

hyperparameters and take a closer look at the evaluation metrics. After that we display the individual model results after drawing a fine line of used data.

5.1 Data

As discussed in chapter 4 the final data considered for feeding the ML model consisted of 16 variables including the predicted variable of severity, and we used the oversampled data set of 441,104 towards model training fitting and filter the testing split for any duplications and eliminated them, both followed by data normalization.

Training/testing split. As simple yet efficient method of ML deliverables assessment. The process is basically to take a data set and split it into two groups. First group used to fit the ML model and called training set, while the second is used to compare the real observations it has, to those predicted by ML model and known as testing set. In our case we are using 70/30 split for training and testing respectively. **Also in this stage, we assured that there are no duplicate copies from the same sample in the testing set.** Hence, the training set consists of 308,772 samples and the testing set consists of 132,332 samples.

5.2 Tuning hyperparameters

Configuration points which enable the ML model customization to perform a certain task are called hyperparameters [73], which can also be expressed as developer specified model argument to lead the process of learning for a certain dataset and set manually.

Hyperparameters not to be confused with parameters which are internal constants selected by means of training/optimization of model on a training dataset and are acquired automatically.

Basically, the common sense of the model is influenced by hyperparameters given the fact that no clear way is known to set them for a given set of data, in an additional complication these hyperparameters may interact in non-linear relations. As a result, a search is usually needed to discover the hyperparameter set which gives best model performance for a set of data better known as hyperparameter tuning (search).

Optimization process includes the definition of a search space. The best way to reflect this space is considering a volume of n-dimension, where each different dimension is represented by a hyperparameter and the amount of dimension as the possible values that may be assigned to that hyperparameter. The ultimate objective of the optimization process is to reach the vector producing best model performance following learning.

Many algorithms for the purpose of optimization are available, but two of the common and most robust procedures are random and grid searches[78]. Random search considers the search space as constrained field of hyperparameter values and takes sample points arbitrarily in that field. Where, grid search considers the search space as network of hyperparameter values and assesses every spot of that network.

Generally, grid search best fits discrete search space where discrete points can be placed on a log scale to achieve spot check of the combinations known for well model performance. For these reasons we are choosing grid search for this study.

Hyperparameters not to be confused with parameters which are internal constants selected by means of training/optimization of model on a training dataset and are acquired automatically.

Basically, the common sense of the model is influenced by hyperparameters given the fact that no clear way is known to set them for a given set of data, in an additional complication these hyperparameters may interact in non-linear relations. As a result, a search is usually needed to discover the hyperparameter set which gives best model performance for a set of data better known as hyperparameter tuning (search).

Optimization process includes the definition of a search space. The best way to reflect this space is considering a volume of n-dimension, where each different dimension is represented by a hyperparameter and the amount of dimension as the possible values that may be assigned to that hyperparameter. The ultimate objective of the optimization process is to reach the vector producing best model performance following learning.

Many algorithms for the purpose of optimization are available, but two of the common and most robust procedures are random and grid searches[78]. Random search considers the search space as constrained field of hyperparameter values and takes sample points arbitrarily in that field. Where, grid search considers the search space as network of hyperparameter values and assesses every spot of that network.

Generally, grid search best fits discrete search space where discrete points can be placed on a log scale to achieve spot check of the combinations known for well model performance. For these reasons we are choosing grid search for this study.

By using grid search we also avoided model overfitting, which occurs when the data fluctuations in the training set is learned by the model as normal trend in data then continues to implement them as a rule on new data it receives.

5.3 Evaluation metrics

As literature suggests when dealing with a classification problem of imbalanced data threshold metrics are best fit to measure the prediction errors. The design of these metrics represents the fraction of a predicted class that does not match the observation in the holdout set.

Threshold metrics include accuracy, precision, error, recall, false discovery rate, false negative rate and F1 score. Following we will introduce their equations, but first let's have a look on *table 17* below representing possible outcomes of a classification model.

- To make it simple the **true positive** and the **true negative** are when the model predicts a class correctly (predicts severe when the actual class is severe and non-severe when non-severe)
- While the **false positive** is when the model says the class is severe will it is non severe, and **false negative** is when the actual class is severe, and the model predicts it as non-severe.

Table 17 Possible outcomes of a classification problem

	True Prediction Class-I (Severe)	False Prediction Class-0 (Non-severe)

True Observation Class-I (Severe)	True Positive (TP)	False Negative (FN)
False Observation Class-0 (Non-severe)	False Positive (FP)	True Negative (TN)

Accuracy is the portion of the correct predictions the model achieves (whether positive or negative) divided by all the predictions made by the model; the equation (5.1) is for the average accuracy.

$$Accuracy = \frac{TP + TN}{TP + FN + FP + TN} \quad (5.1)$$

While the individual class accuracy can be found through the division of true prediction of that individual class by the sum of true and false of it, which is the **precision equation (5.2) [78]** below for positive (severe) class. An optimal accuracy of 1 coincides with a skillful classifier.

$$Precision = \frac{TP}{TP + FP} \quad (5.2)$$

Recall, sensitivity also known as the true positive rate is found as in equation (5.3) [72] by dividing the true positive prediction portion by the sum of true positive, and false negative

(which is positive in reality is positive, but the model predicted as negative falsely), the optimal value for it is 1 and reflects a good model performance.

$$Recall (Sensitivity) = \frac{TP}{TP + FN} \quad (5.3)$$

$$Error = 1 - Accuracy \quad (5.4)$$

$$FDR = 1 - Precision \quad (5.5)$$

$$FNR = 1 - Recall \quad (5.6)$$

F1 Score is the harmonic average of the recall and precision and can be found through equation (5.7)[78] below and the significance of this measure is that it poises importance of recall and precision also helps in cases of imbalanced data.

$$F1 Score = \frac{2 * Percision * Recall}{Percision + Recall} \quad (5.7)$$

A confusion matrix containing these values will be provided for each model individually.

Figure 10 below shows a sample of a confusion matrix [80].

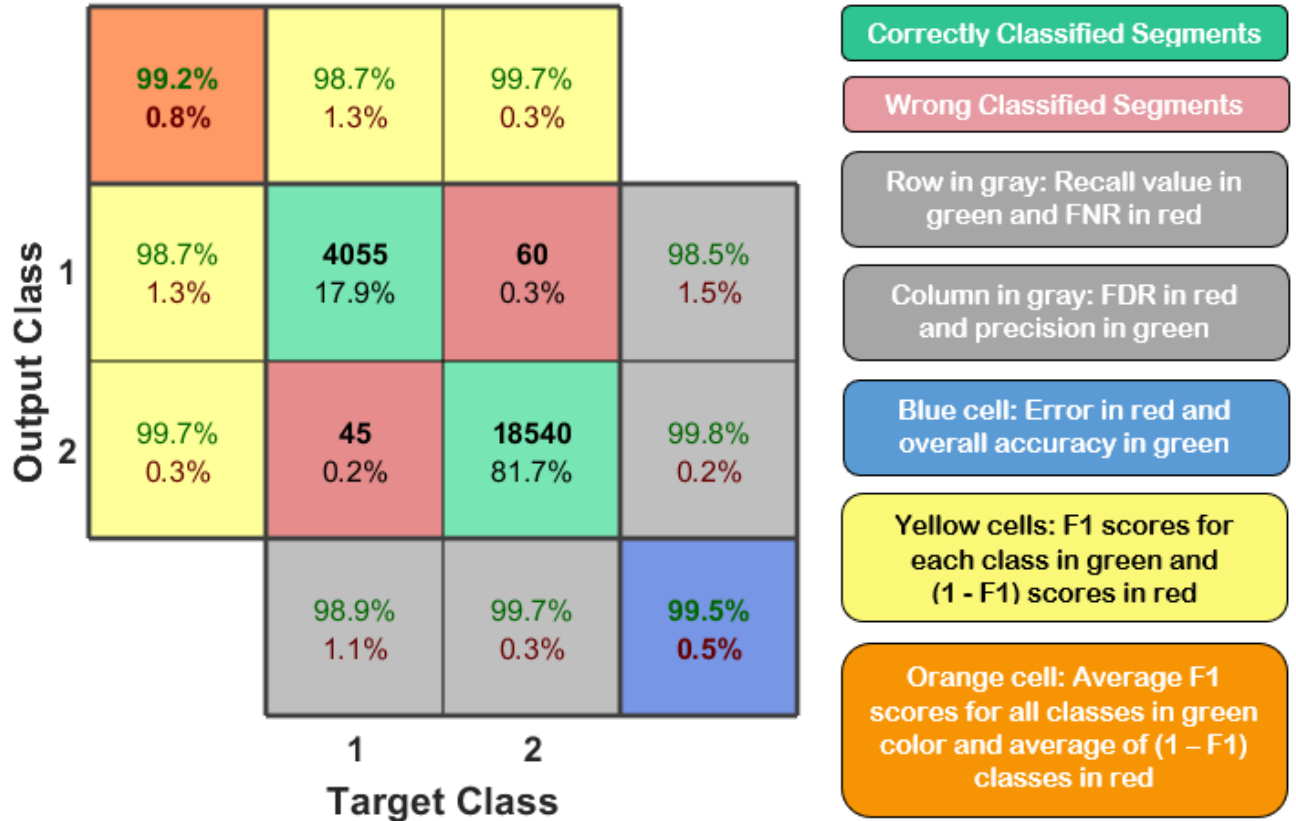


Figure 10 Confusion Matrix sample

Another important evaluation metric we consider in this study for each model is **receiver operator characteristic curve ROC curve** [74] for short which is a graphical plot of model performance between true positive rate (recall) in y-axis and false positive in x-axis. A diagonal line drawn from the bottom left (origin) towards upper right represents the threshold or baseline/ no skill model. When the model scores points below the threshold line – lower right triangle- we say it is less skilled or performing worse than the baseline while when it scores points above threshold lie – upper left triangle- we say it is more skillful, the closer it gets to y-axis the higher skillful it is considered. *Figure 11* below shows an ROC plot [81]. We use micro-average ROC and macro-average ROC to describe the output of the ML model.

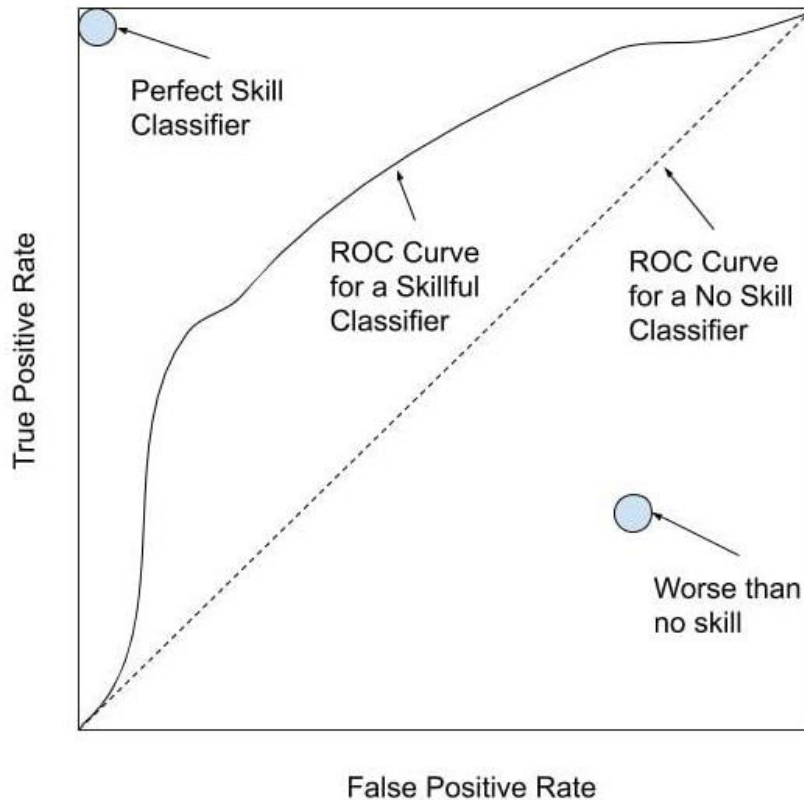


Figure 11 ROC Plot

An interconnected measure to ROC is the area under ROC curve **ROC AUC** for short which is a representation of probability function with values ranging from 0 to 1 corresponding to worst predictor and best predictor respectively. Assuming the baseline or no skill scenario the AUC value will be equal to 0.5.

The main difference between macro and micro averages is that the macro-average evaluates the metric independently for each class (0 and 1) and then find the average value. While the micro-average aggregates the contributions/probabilities of class 0 and class 1 to evaluate the average metric.

5.4 Results

5.4.1 K-Nearest Neighbors (K-NN)

1. Tuning:

Below *table 18*, shows the ranges for each parameter and the best values for each reached by grid search technique, for highest accuracy and least processing time and other parameter not mentioned here are considered as default.

Table 18 Tuning Parameters for K-NN

Parameter	Range	Best value
n_neighbors	Range (1, 200, 10)	50
weights	['uniform', 'distance']	'uniform'
metric	['euclidean', 'manhattan', 'minkowski']	'euclidean'

2. Confusion Matrix:

Below *figure 12*, shows the confusion matrix which can provide the following:

- True negative is equal to 44,418 samples
- True positive is equal to 64,599 samples
- False negative is equal to 1,942 samples
- False positive is equal to 21,373 samples
- Accuracy for non-severe class is 95.8%
- Accuracy for severe class is 75.1%
- Average class accuracy is 82.4%
- Recall for non-severe is 67.5%
- Recall for severe is 97.1%
- F-1 score for non-severe 79.2%
- F-1 score for severe 84.7%

- Average F-1 score 82.0%

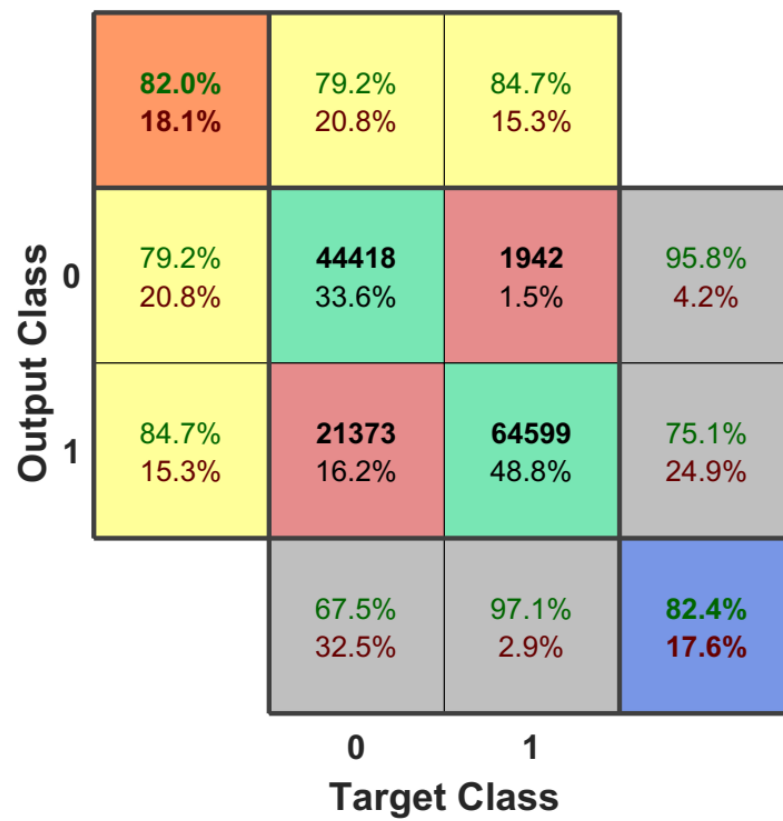


Figure 12 Confusion matrix for K-NN

3. ROC:

In case of K-NN and as shown in *figure 13*, the AUC-ROC for class 0 is 97.4% and for class 1 is 97.4%. Also, the micro-average AUC-ROC for the two classes is 92.0% and the macro-average AUC-ROC is 97.0%

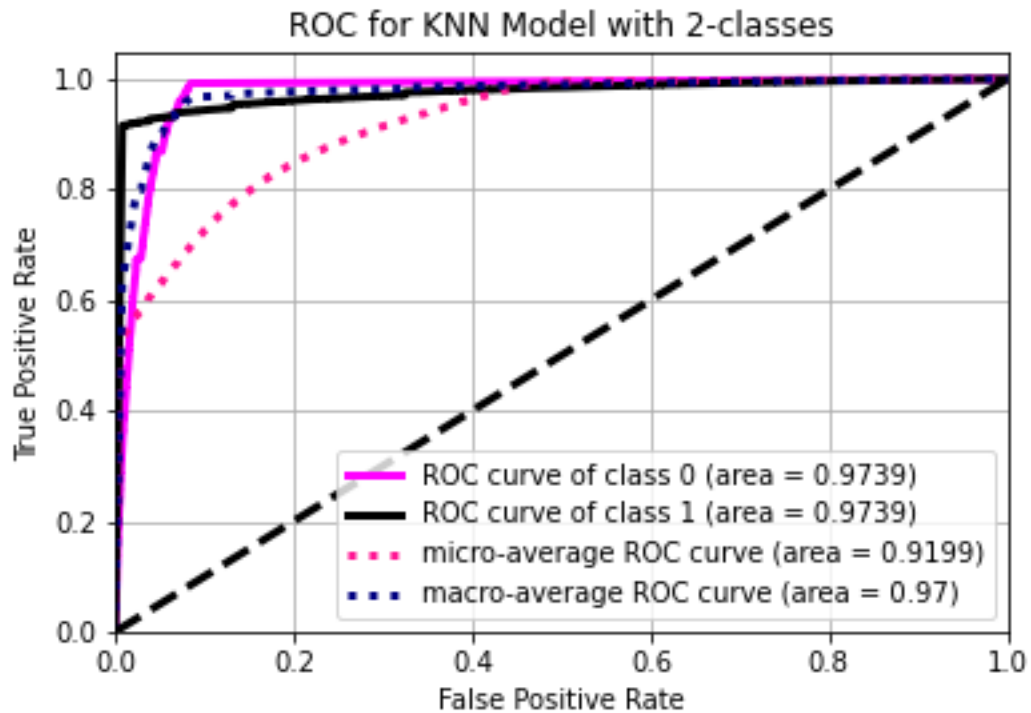


Figure 13 ROC Curve for K-NN

4. Average processing time to test one sample:

Average Processing time to test one sample = 0.003500824 secs

5. Feature Importance:

In case of K-NN and as shown in *figure 14*, the most important feature was accident time followed by accident month then day of the week and vehicle model equally.

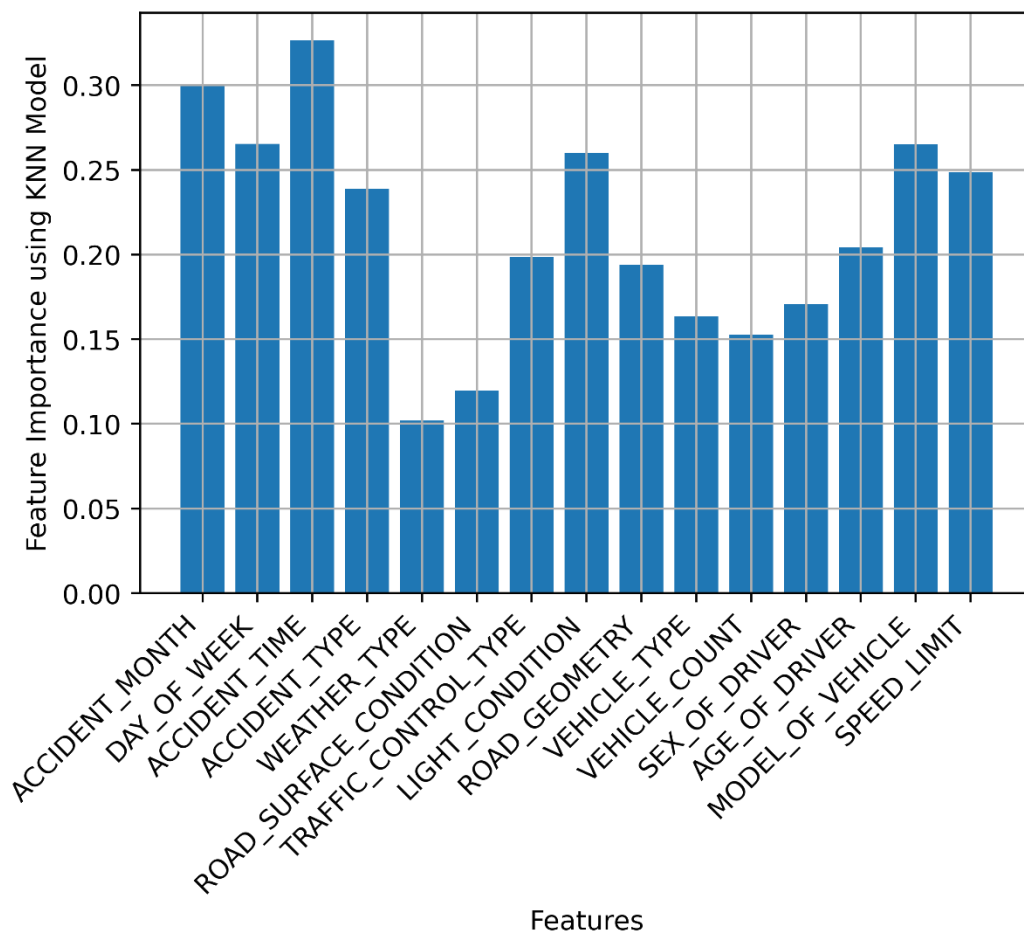


Figure 14 KNN Feature Importance

5.4.2 Support Vector Machine (SVM)

1. Tuning:

Below *table 19*, shows the ranges for each parameter and the best values for each reached by grid search technique, for highest accuracy and least processing time and other parameter not mentioned here are considered as default.

Table 19 Tuning Parameters for SVM

Parameter	Range	Best value
kernel	['poly', 'rbf', 'sigmoid']	'rbf'
C	[100,50,10,1.0, 0.1]	100
gamma	[0.1, 0.01,0.001,0.0001]	0.1

2. Confusion Matrix:

Below *figure 15*, shows the confusion matrix which can provide the following:

- True negative is equal to 50,968 samples
- True positive is equal to 61,764 samples
- False negative is equal to 4,777 samples
- False positive is equal to 14,823 samples
- Accuracy for non-severe class is 91.4 %
- Accuracy for severe class is 80.6 %
- Average class accuracy is 85.2 %
- Recall for non-severe is 77.5 %
- Recall for severe is 92.8 %
- F-1 score for non-severe 83.9 %
- F-1 score for severe 86.3 %
- Average F-1 score 85.1 %

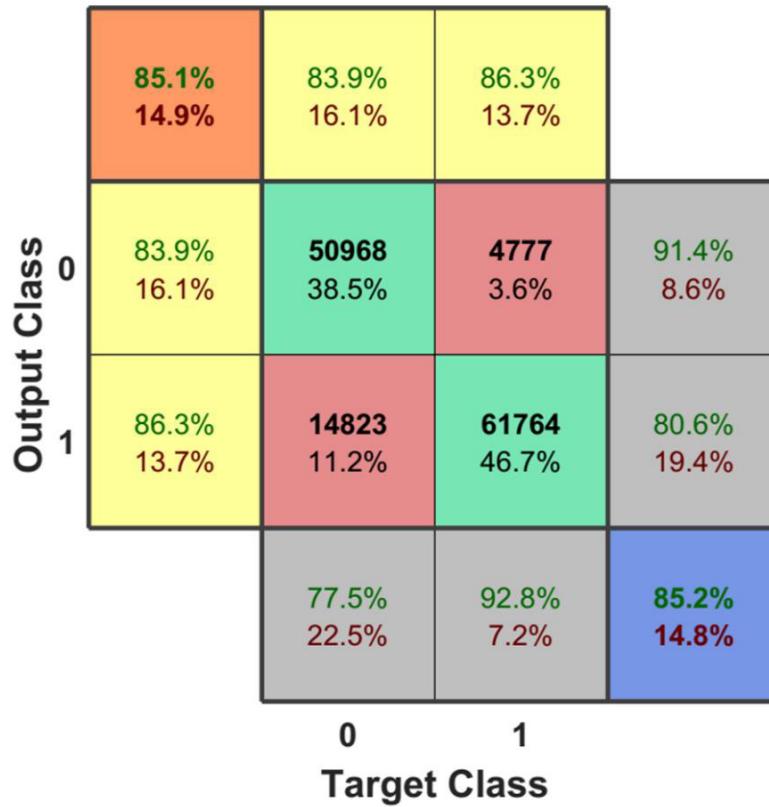


Figure 15 Confusion matrix for SVM

3. ROC:

In case of SVM and as shown in *Figure 16*, the AUC-ROC for class 0 is 89.2 % and for class 1 is 89.2 %. Also, the micro-average AUC-ROC for the two classes is 89.8 % and the macro-average AUC-ROC is 89.0 %

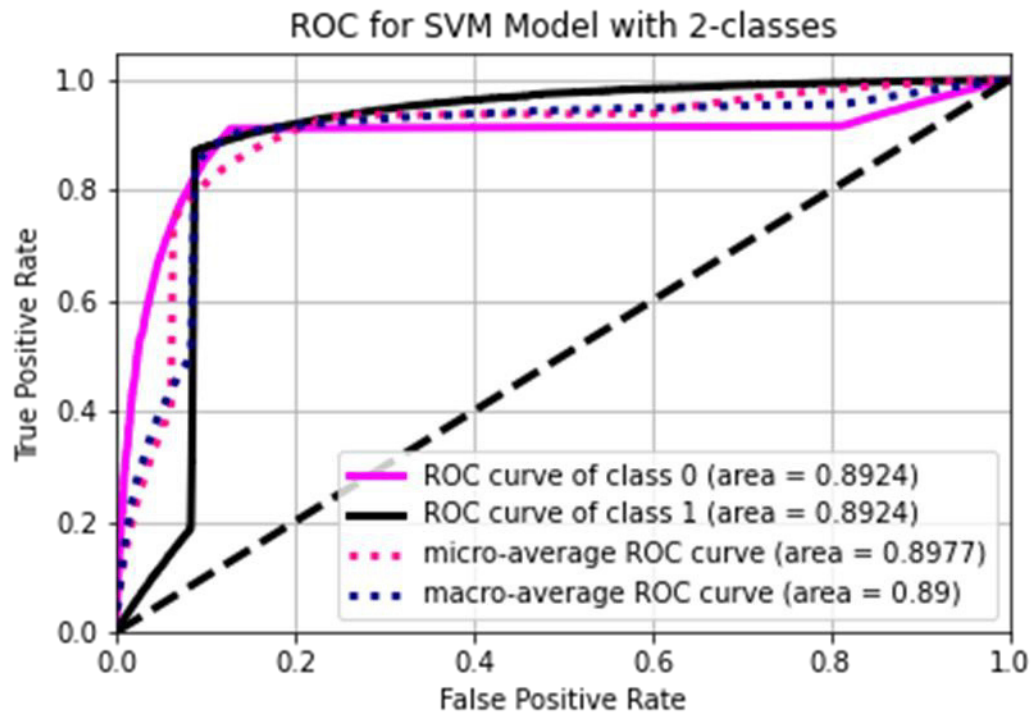


Figure 16 ROC Curve for SVM

4. Average processing time to test one sample:

Average Processing time to test one sample = 0.008482715 secs

5. Feature Importance:

In case of SVM and as shown in *figure 17*, the most important feature was accident time followed by accident month then day of the week and light condition equally.

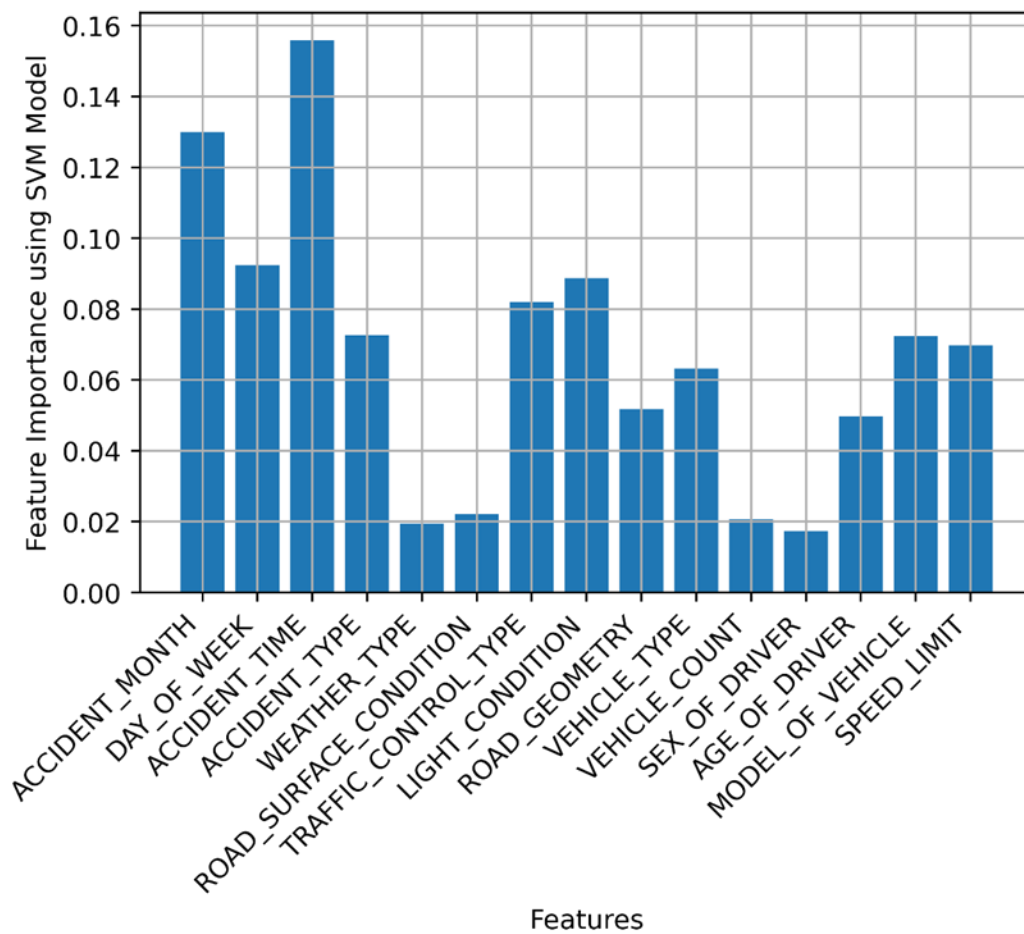


Figure 17 SVM Feature Importance

5.4.3 Random Forest (RF)

1. Tuning:

Below *table 20*, shows the ranges for each parameter and the best values for each reached by grid search technique, for highest accuracy and least processing time and other parameter not mentioned here are considered as default.

Table 20 Tuning Parameters for RF

Parameter	Range	Best value
n_estimators	[50,100,200,300,500]	100
max_features	['auto', 'sqrt', 'log2']	'auto'
max_depth	[20,25,30,35,40]	25
criterion	['gini', 'entropy']	'gini'

2. Confusion Matrix:

Below *figure 18*, shows the confusion matrix which can provide the following:

- True negative is equal to 52,031 samples
- True positive is equal to 57,784 samples
- False negative is equal to 8,757 samples
- False positive is equal to 13,760 samples
- Accuracy for non-severe class is 85.6%
- Accuracy for severe class is 80.8%
- Average class accuracy is 83.0%
- Recall for non-severe is 79.1%
- Recall for severe is 86.8%
- F-1 score for non-severe 82.2%
- F-1 score for severe 83.7%

- Average F-1 score 82.9%

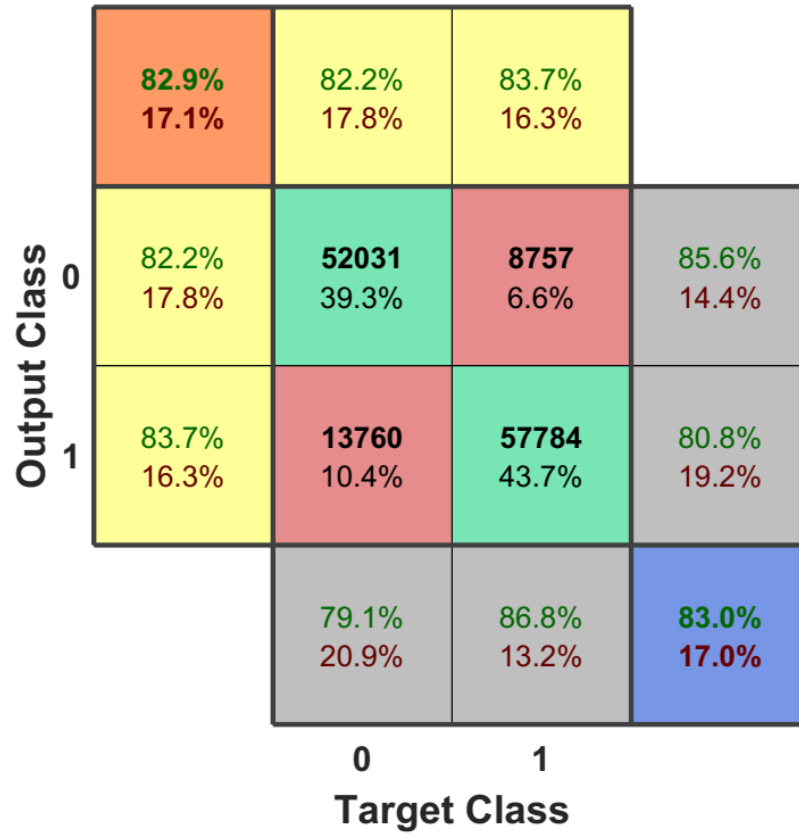


Figure 18 Confusion matrix for RF

3. ROC:

In case of RF and as shown in *Figure 19*, the AUC-ROC for class 0 is 91.6% and for class 1 is 91.6%. Also, the micro-average AUC-ROC for the two classes is 91.3% and the macro-average AUC-ROC is 92.0%

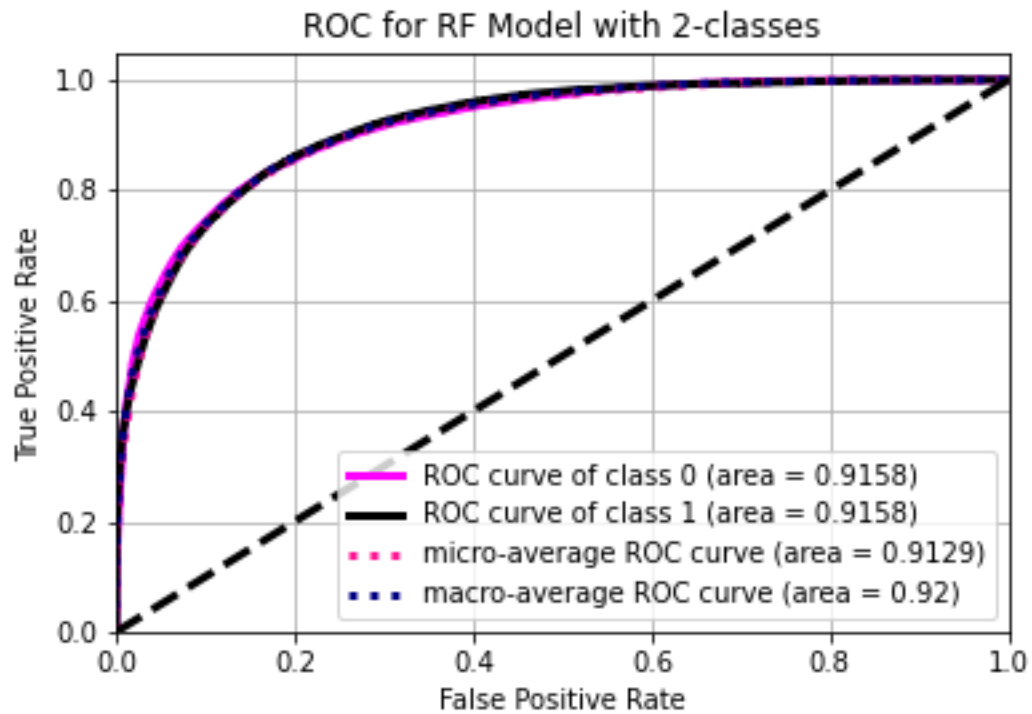


Figure 19 ROC Curve for RF

4. Average processing time to test one sample:

Average Processing time to test one sample = 0.009608444 secs

5. Feature Importance:

In case of RF and as shown in *figure 20*, the most important feature was accident time followed by accident month then day of the week.

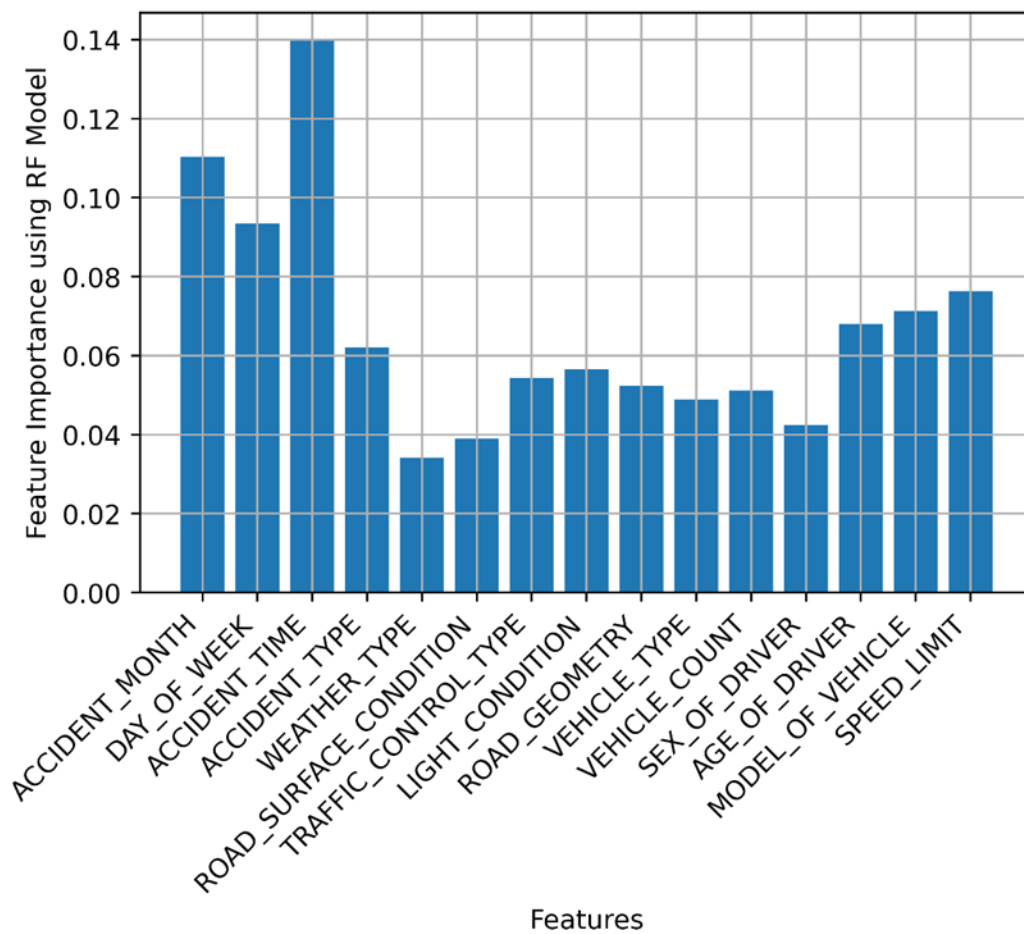


Figure 20 RF Feature Importance

5.4.4 Decision Tree (DT)

1. Tuning:

Below *table 21*, shows the ranges for each parameter and the best values for each reached by grid search technique, for highest accuracy and least processing time and other parameter not mentioned here are considered as default.

Table 21 Tuning Parameters for DT

Parameter	Range	Best value
criterion	['gini','entropy']	'gini'
max_depth	range(30, 70, 2)	25
min_samples_split	[2, 3, 4]	2

2. Confusion Matrix:

Below *figure 21*, shows the confusion matrix which can provide the following:

- True negative is equal to 48,329 samples
- True positive is equal to 59,955 samples
- False negative is equal to 6,586 samples
- False positive is equal to 17,462 samples
- Accuracy for non-severe class is 88.0%
- Accuracy for severe class is 77.4%
- Average class accuracy is 81.8%
- Recall for non-severe is 73.5%
- Recall for severe is 90.1%
- F-1 score for non-severe 80.1%
- F-1 score for severe 83.3%
- Average F-1 score 81.7%

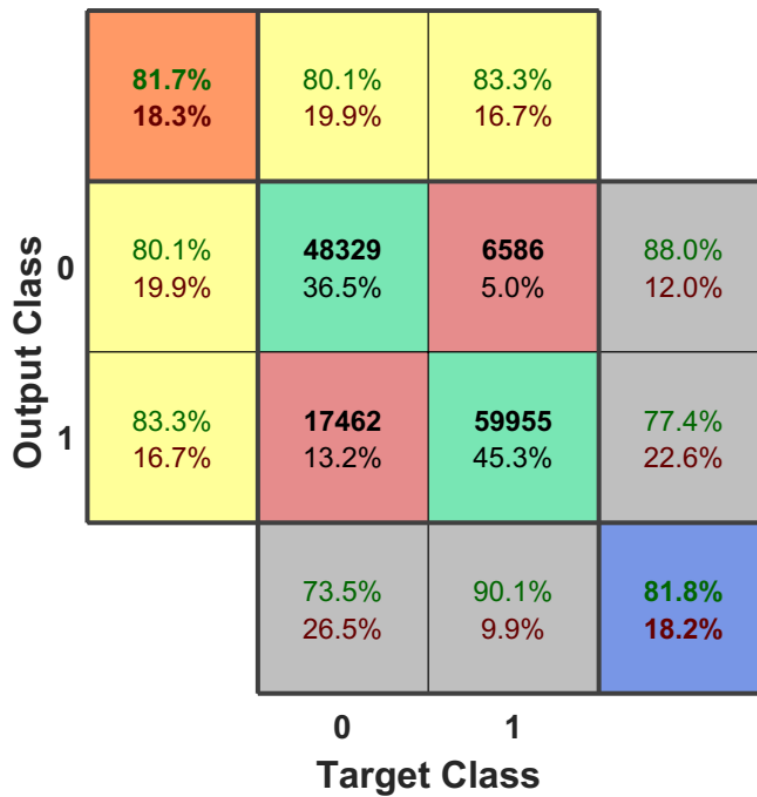


Figure 21 Confusion matrix for DT

3. ROC:

In case of DT and as shown in *Figure 22*, the AUC-ROC for class 0 is 86.0% and for class 1 is 86.0%. Also, the micro-average AUC-ROC for the two classes is 86.7% and the macro-average AUC-ROC is 86.0%

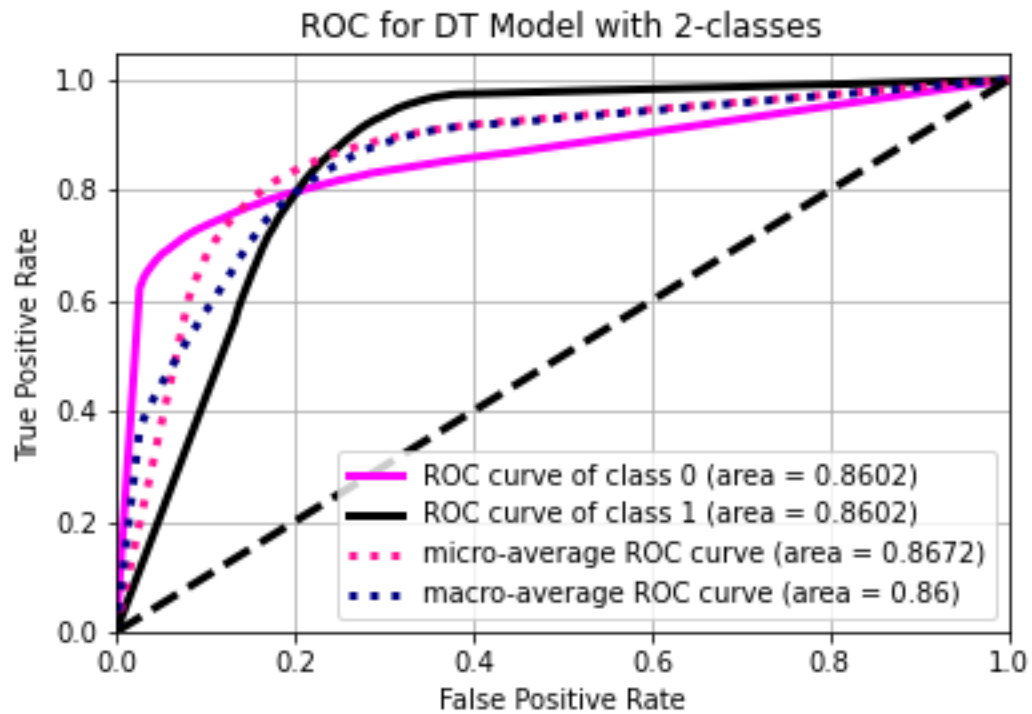


Figure 22 ROC Curve for DT

4. Average processing time to test one sample:

Average Processing time to test one sample = 0.000072569 secs

5. Feature Importance:

In case of DT and as shown in *figure 23*, the most important feature was accident time followed by accident month then day of the week.

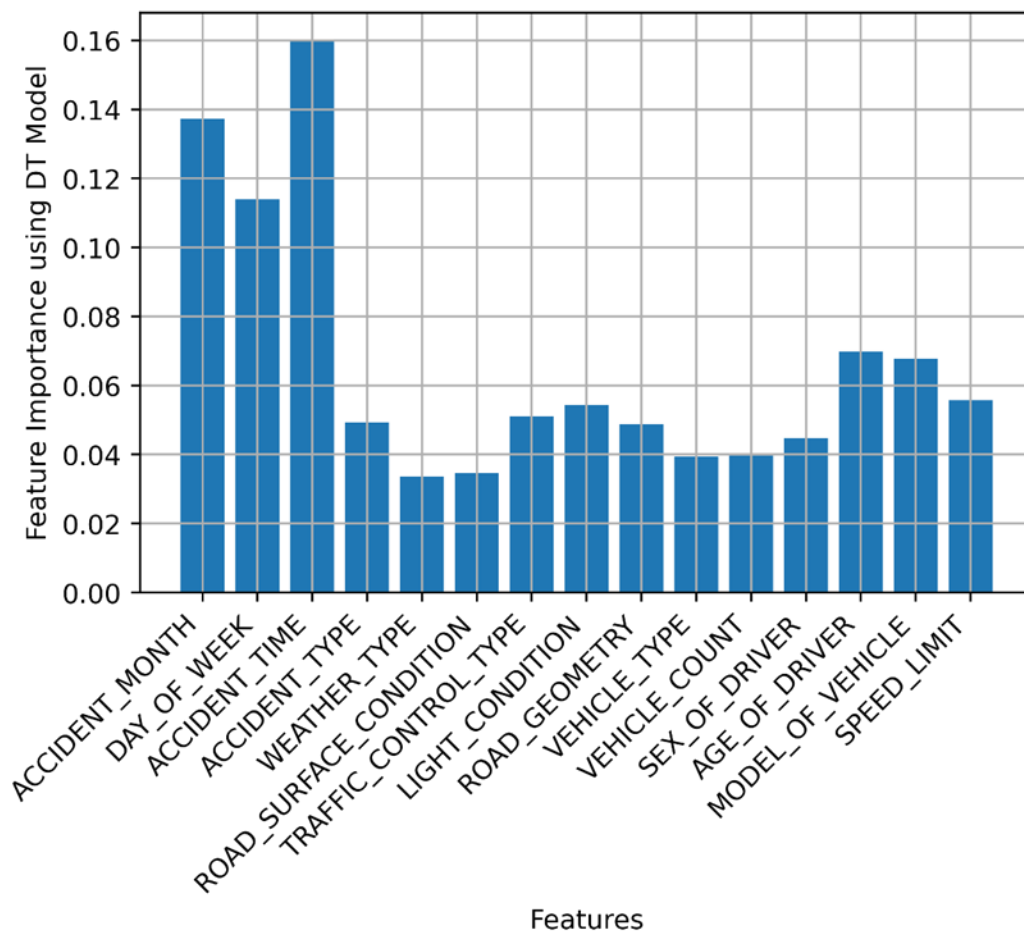


Figure 23 DT Feature Importance

5.4.5 Naïve Bayes (NB)

1. Tuning:

Below *table 22*, shows the ranges for each parameter and the best values for each reached by grid search technique, for highest accuracy and least processing time and other parameter not mentioned here are considered as default.

Table 22 Tuning Parameters for NB

Parameter	Range	Best value
alpha	[1, 0.1, 0.01, 0.001, 0.0001, 0.00001]	1

2. Confusion Matrix:

Below *figure 24*, shows the confusion matrix which can provide the following:

- True negative is equal to 42,634 samples
- True positive is equal to 42,997 samples
- False negative is equal to 23,544 samples
- False positive is equal to 23,157 samples
- Accuracy for non-severe class is 64.4%
- Accuracy for severe class is 65.0%
- Average class accuracy is 64.7%
- Recall for non-severe is 64.8%
- Recall for severe is 64.6%
- F-1 score for non-severe 64.6%
- F-1 score for severe 64.8%
- Average F-1 score 64.7%

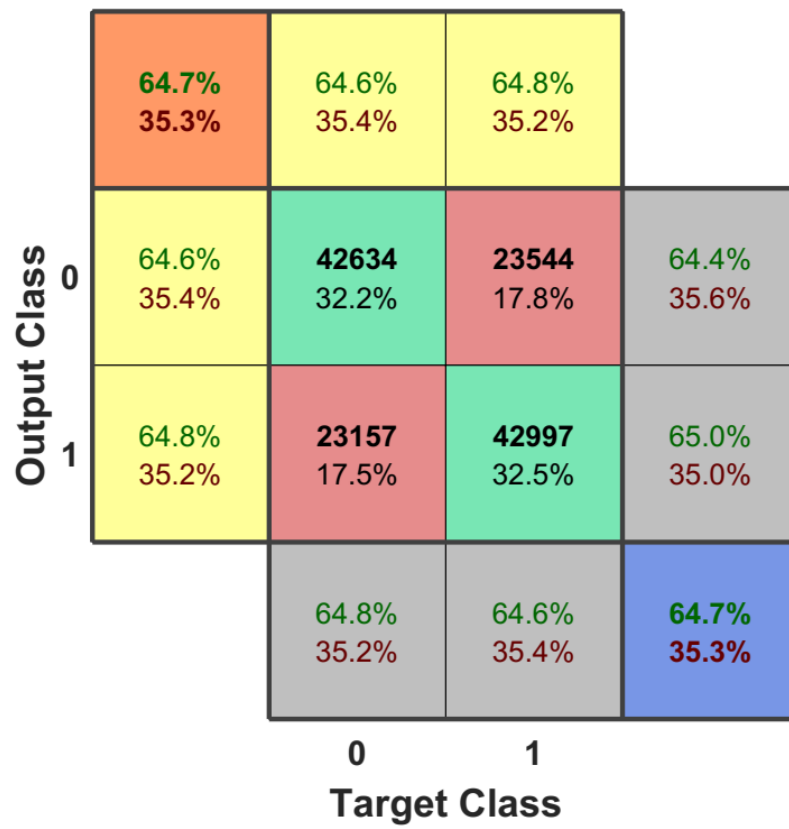


Figure 24 Confusion matrix for NB

3. ROC:

In case of NB and as shown in *Figure 25*, the AUC-ROC for class 0 is 69.8% and for class 1 is 69.8%. Also, the micro-average AUC-ROC for the two classes is 69.8% and the macro-average AUC-ROC is 70.0%

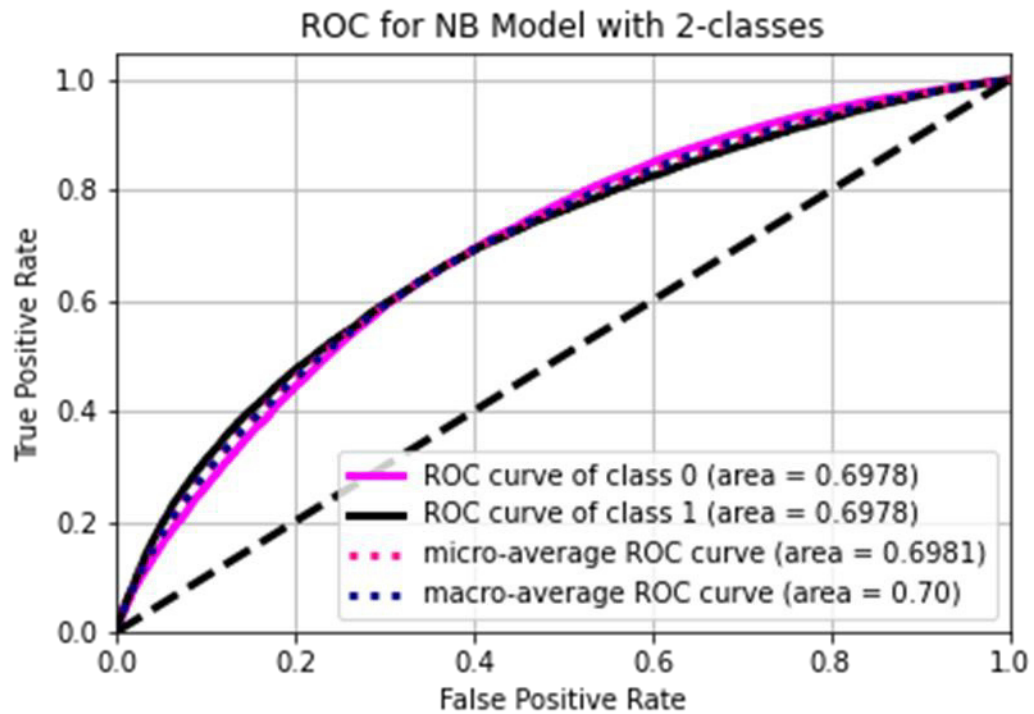


Figure 25 ROC Curve for NB

4. Average processing time to test one sample:

Average Processing time to test one sample = 0.001618518 secs

5. Feature Importance:

In case of NB and as shown in *figure 26*, the most important feature was light condition followed by vehicle model then speed limit and traffic control type equally.

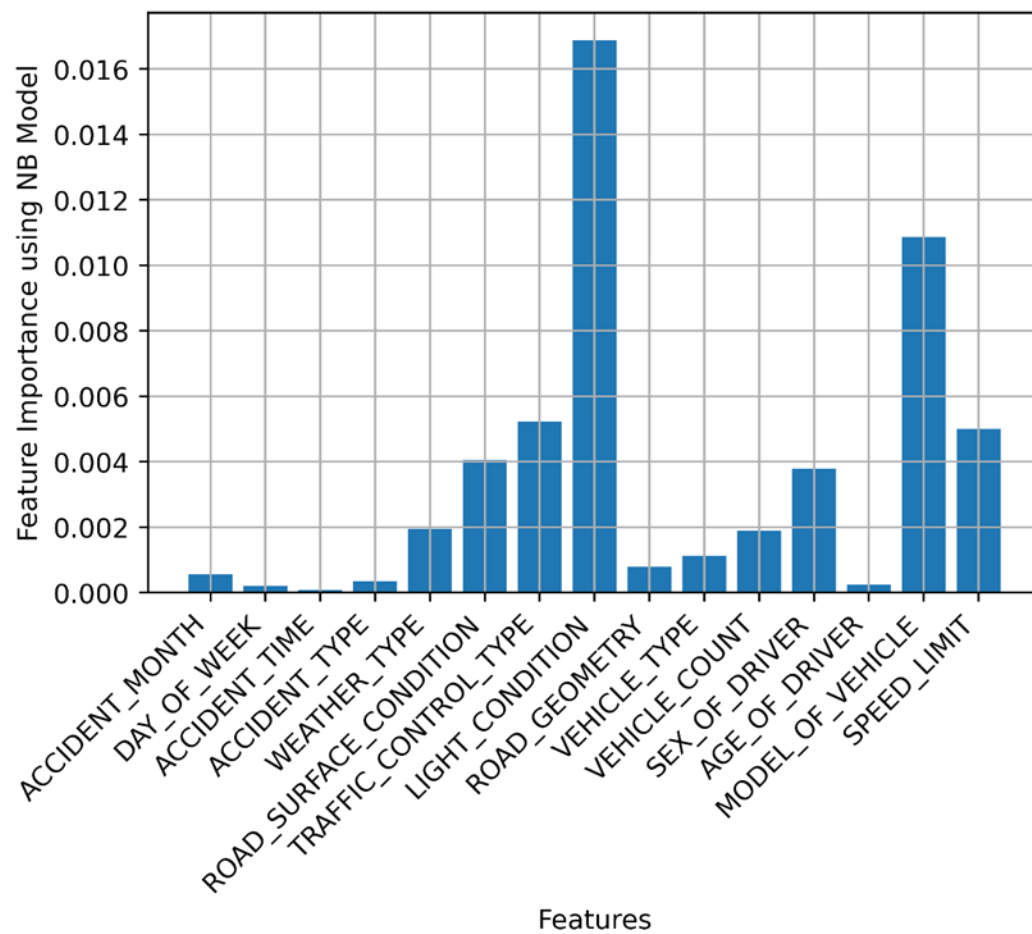


Figure 26 NB Feature Importance

5.4.6 Neural Network (NN)

1. Tuning:

Below *table 23*, shows the ranges for each parameter and the best values for each reached by grid search technique, for highest accuracy and least processing time and other parameter not mentioned here are considered as default.

Table 23 Tuning Parameters NN

Parameter	Range	Best value
neurons	range (100, 1000, 100)	800
batch_length	[50,100,150,200]	150
number_epoch	[500,1000,1200,1500]	1000
optimizer	['RMSprop', 'Adam']	'adam'

Remaining parameters

inner_activation_fun = 'relu'

outer_activation_fun = 'sigmoid'

optimizer_loss_fun = 'binary_crossentropy'

hidden_layers = 1

2. Confusion Matrix:

Below *figure 27*, shows the confusion matrix which can provide the following:

- True negative is equal to 46,006 samples
- True positive is equal to 45,114 samples
- False negative is equal to 21,427 samples
- False positive is equal to 19,785 samples
- Accuracy for non-severe class is 68.2%
- Accuracy for severe class is 69.5%

- Average class accuracy is 68.9%
- Recall for non-severe is 69.9%
- Recall for severe is 67.8%
- F-1 score for non-severe 69.0%
- F-1 score for severe 68.6%
- Average F-1 score 68.8%

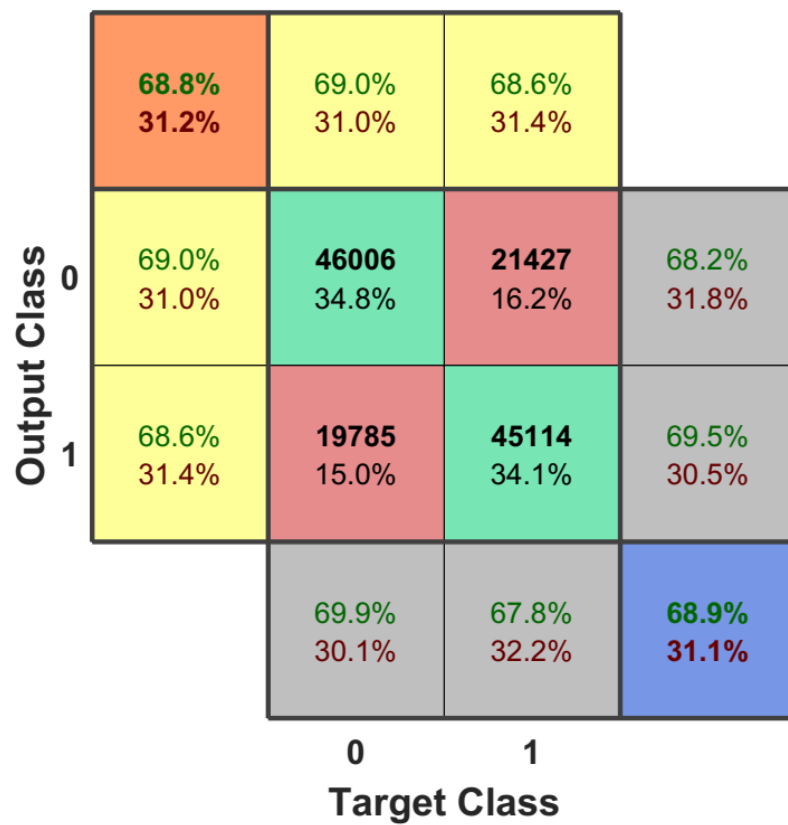


Figure 27 Confusion matrix for NN

3. ROC:

In case of NN and as shown in *Figure 28*, the AUC-ROC for class 0 is 75.4% and for class 1 is 75.4%. Also, the micro-average AUC-ROC for the two classes is 75.4% and the macro-average AUC-ROC is 75%

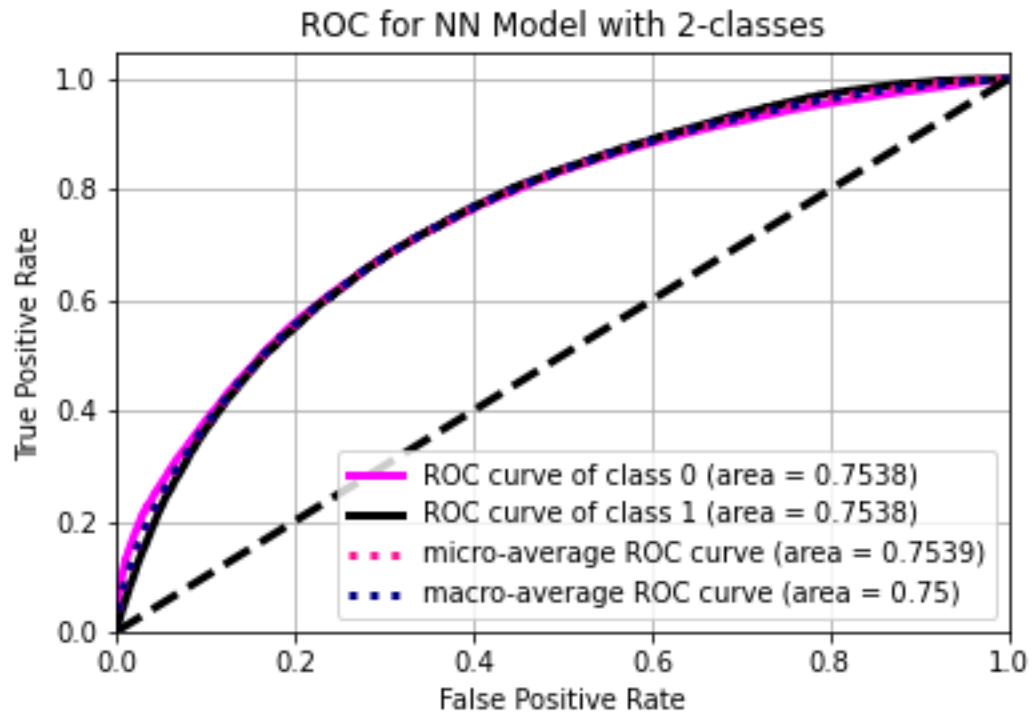


Figure 28 ROC Curve for NN

4. Average processing time to test one sample:

Average Processing time to test one sample = 0.000336223 secs

5. Feature Importance:

In case of NN and as shown in *figure 29*, the most important feature was traffic control type followed by sex of driver then speed limit after that was accident type and vehicle type..

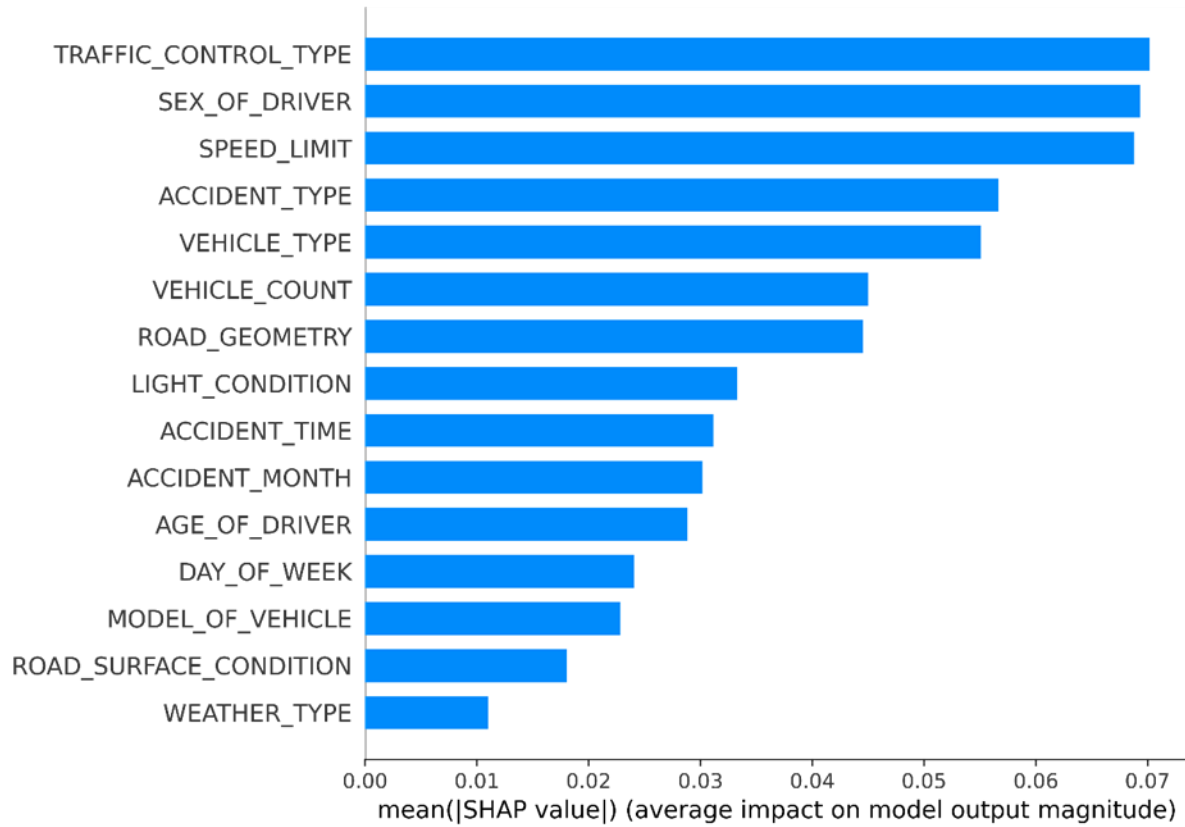


Figure 29 NN Feature Importance

5.4.7 Deep learning Model (DL)

1. Tuning:

Below *table 24*, shows the ranges for each parameter and the best values for each reached by grid search technique, for highest accuracy and least processing time and other parameter not mentioned here are considered as default.

Table 24 Tuning Parameters for DL

Parameter	Range	Best value
neurons	range(100, 1000, 100)	600
batch_length	[100,150,200,250]	200
number_epoch	[400,600,800,1000,1200]	600
optimizer	['RMSprop', 'Adam']	'adam'
hidden_layers	[2,3,4,5,6,7]	3

Remaining parameters

inner_activation_fun = 'relu'

outer_activation_fun = 'sigmoid'

optimizer_loss_fun = 'binary_crossentropy'

2. Confusion Matrix:

Below *figure 30*, shows the confusion matrix which can provide the following:

- True negative is equal to 51,897 samples
- True positive is equal to 62,804 samples
- False negative is equal to 3,737 samples
- False positive is equal to 13,894 samples
- Accuracy for non-severe class is 93.3%
- Accuracy for severe class is 81.9%

- Average class accuracy is 86.7%
- Recall for non-severe is 78.9%
- Recall for severe is 94.4%
- F-1 score for non-severe 85.5%
- F-1 score for severe 87.7%
- Average F-1 score 86.6%

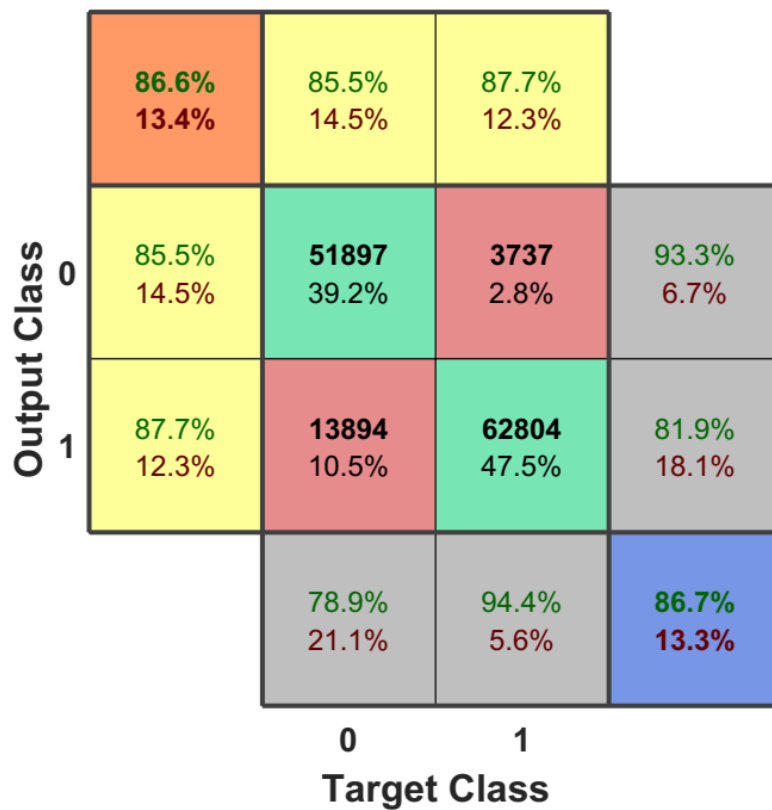


Figure 30 Confusion Matrix for DL

3. ROC:

In case of DL and as shown in *Figure 31*, the AUC-ROC for class 0 is 89.8% and for class 1 is 89.8%. Also, the micro-average AUC-ROC for the two classes is 89.8% and the macro-average AUC-ROC is 90.0%

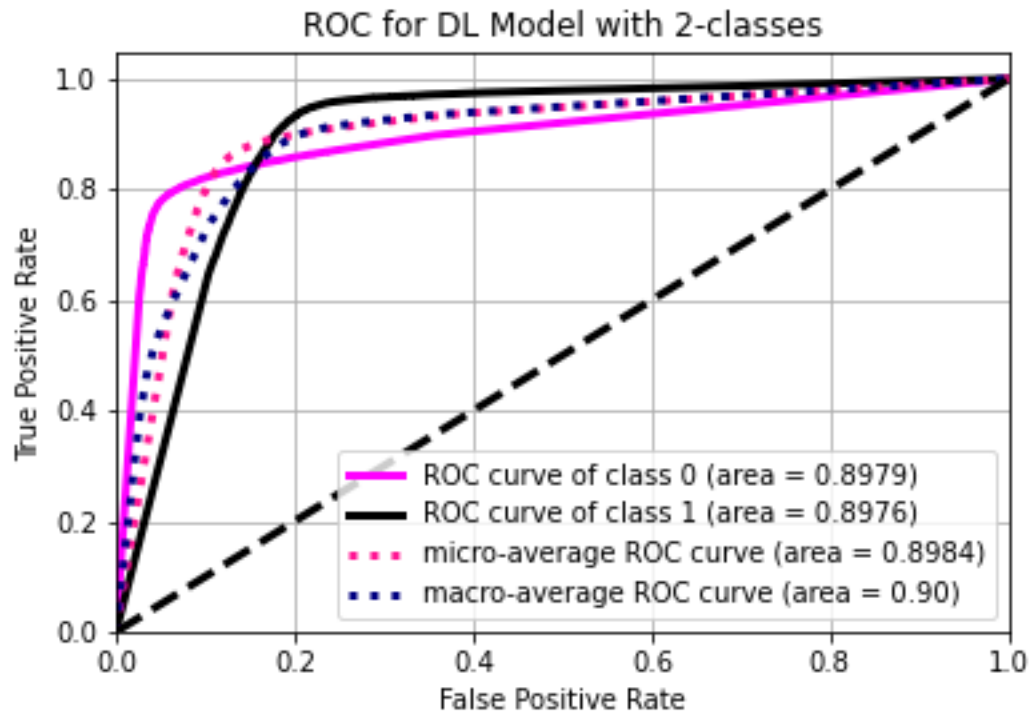


Figure 31 ROC Curve for DL

4. Average processing time to test one sample:

Average Processing time to test one sample = 0.000468575 secs

5. Feature Importance:

In case of DL and as shown in *figure 32*, the most important feature was accident type followed by vehicle count then light condition, accident time and speed limit

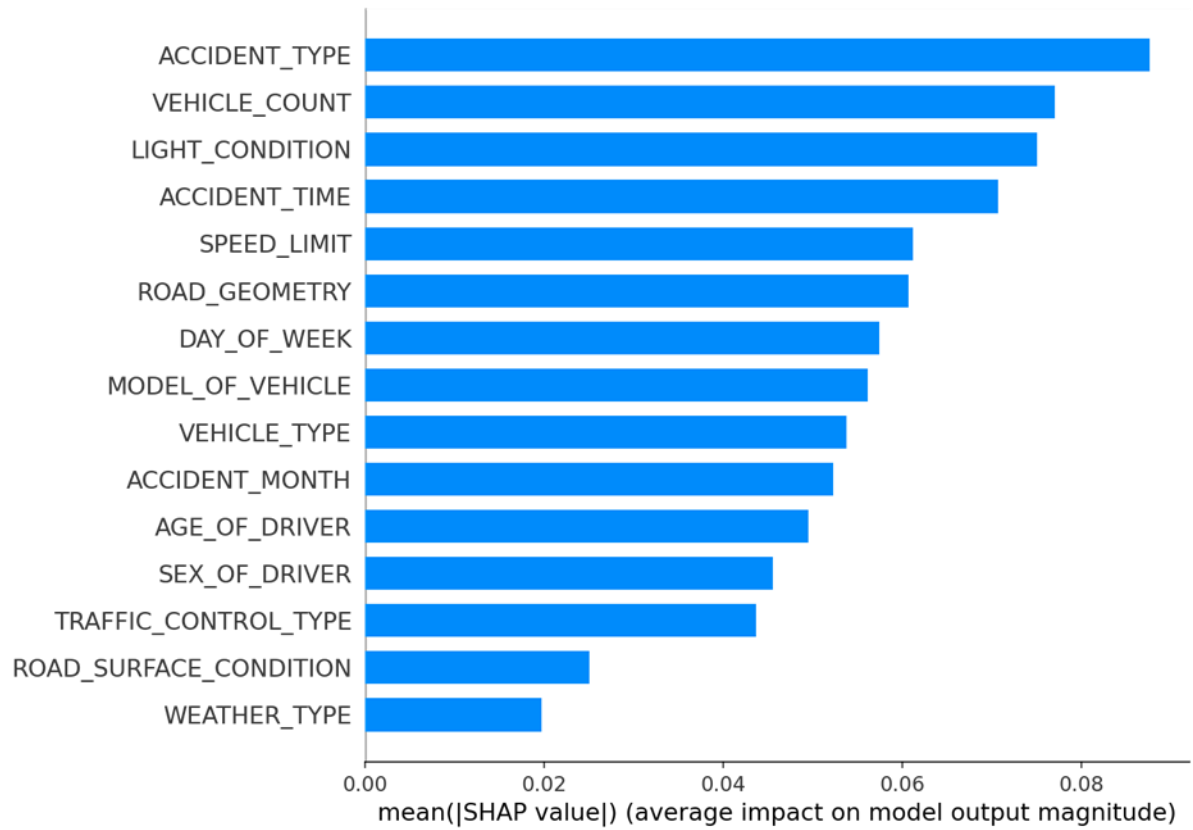


Figure 32 DL Feature Importance

5.4.8 Deep learning Model (DL) for individual countries

To further inspect our model performance and its variance with the input data quality, we made a run for the four countries only individually this time

1. Australia

A. Data:

$X_{\text{train}} = (60334, 15)$, $X_{\text{test}} = (25858, 15)$, $Y_{\text{train}} = (60334, 1)$, and $Y_{\text{test}} = (25858, 1)$

B. Confusion Matrix:

Output Class				
		0	1	
0	0	68.5% 31.4%	66.4% 33.6%	70.7% 29.3%
	1	66.4% 33.6%	8023 31.0%	3223 12.5%
1	0	70.7% 29.3%	4873 18.8%	9739 37.7%
	1		62.2% 37.8%	75.1% 24.9%
		0	1	
		Target Class		

Figure 33 Confusion Matrix DL - Australia

C. Processing Time:

Average Processing time to test one sample = 0.000439455 secs.

D. ROC:

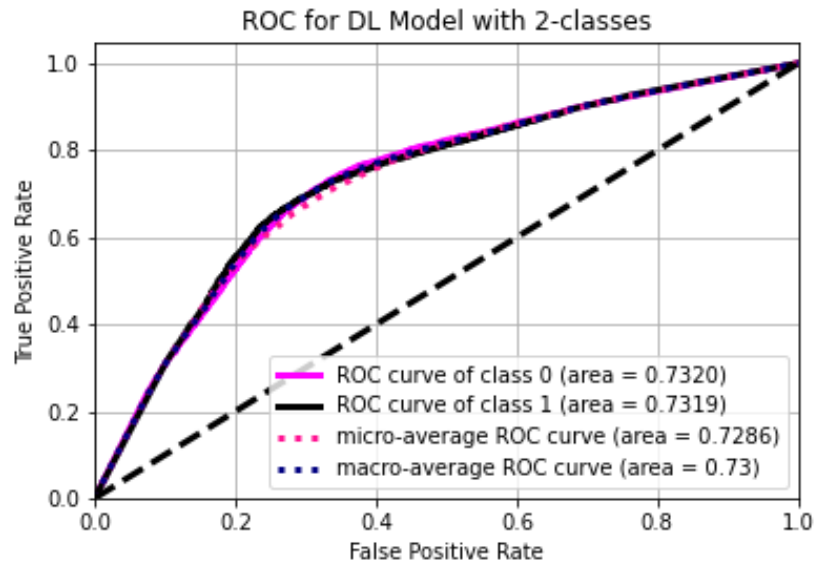


Figure 34 ROC Curve DL - Australia

E. Feature Importance:

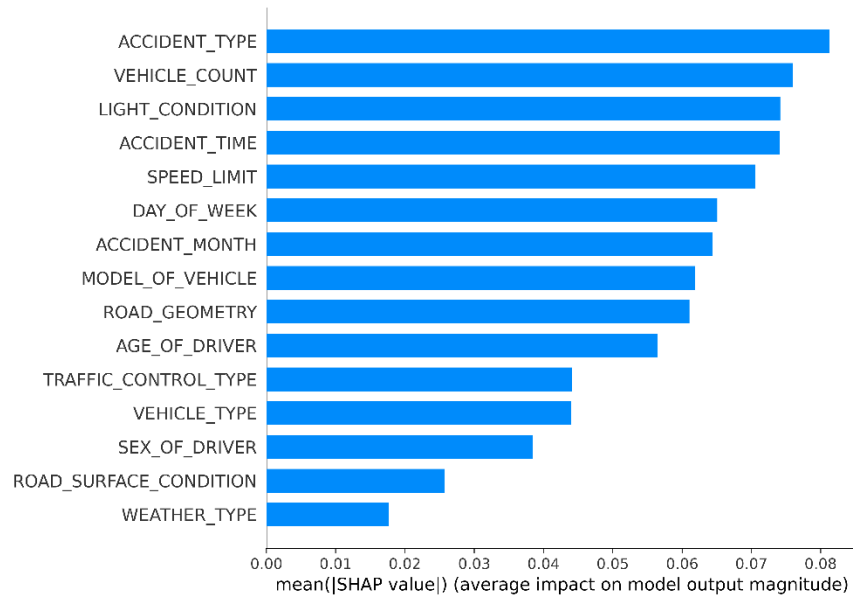


Figure 35 Feature Importance DL - Australia

We notice from *figures 33,34 and 35* a drop in the prediction accuracy compared to the values attained in the previous subsection. However, it is also noticeable the important features are the same to what was recorded earlier.

2. Canada

A. Data:

$X_{\text{train}} = (85044, 15)$, $X_{\text{test}} = (36448, 15)$, $Y_{\text{train}} = (85044, 1)$, and $Y_{\text{test}} = (36448, 1)$

B. Confusion Matrix:

Output Class	0	95.4% 4.6%	95.2% 4.8%	95.6% 4.4%	
		95.2% 4.8%	16645 45.7%	81 0.2%	99.5% 0.5%
	1	95.6% 4.4%	1591 4.4%	18131 49.7%	91.9% 8.1%
				91.3% 8.7%	99.6% 0.4%
		0	1	Target Class	

Figure 36 Confusion Matrix DL - Canada

C. Processing Time:

Average Processing time to test one sample = 0.000405647 secs.

D. ROC:

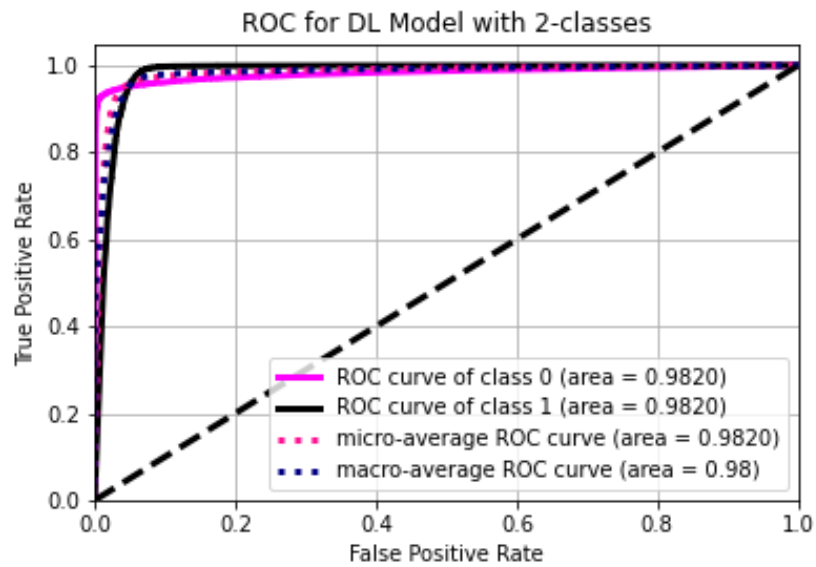


Figure 37 ROC Curve DL - Canada

E. Feature Importance:

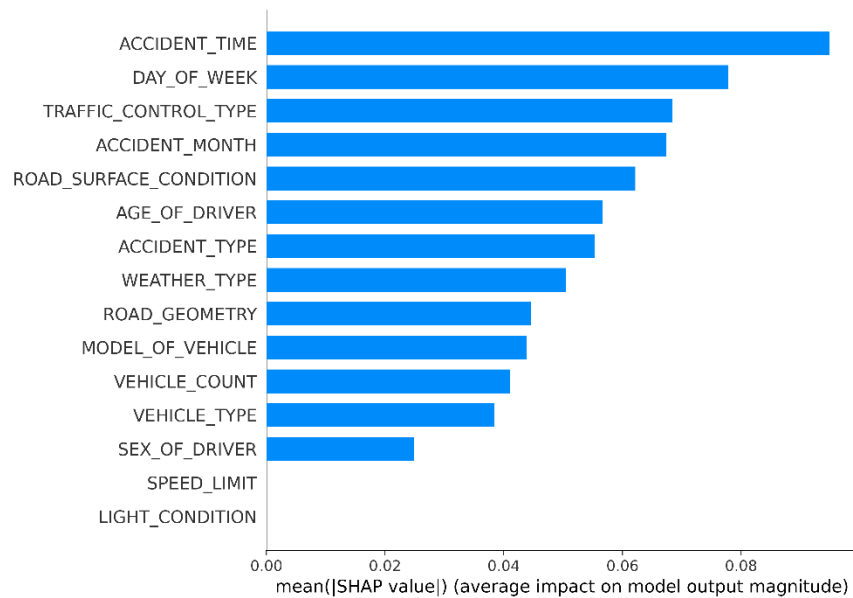


Figure 38 Feature Importance DL - Canada

We notice from *figures 36,37 and 38* an enhanced prediction accuracy compared to the values attained in the previous subsection. However, the important features vary from those recorded earlier.

3. Great Britain

A. Data:

X_train = (80336, 15), X_test = (34430, 15), Y_train = (80336,1), and Y_test = (34430,1)

B. Confusion Matrix:

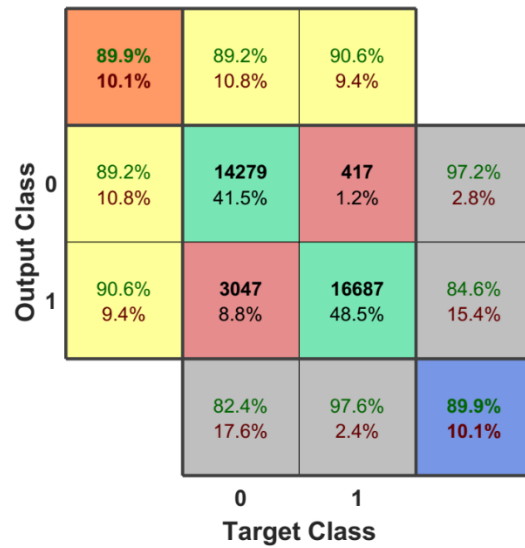


Figure 39 Confusion Matrix DL - GB

C. Processing Time:

Average Processing time to test one sample = 0.000445445 secs.

D. ROC:

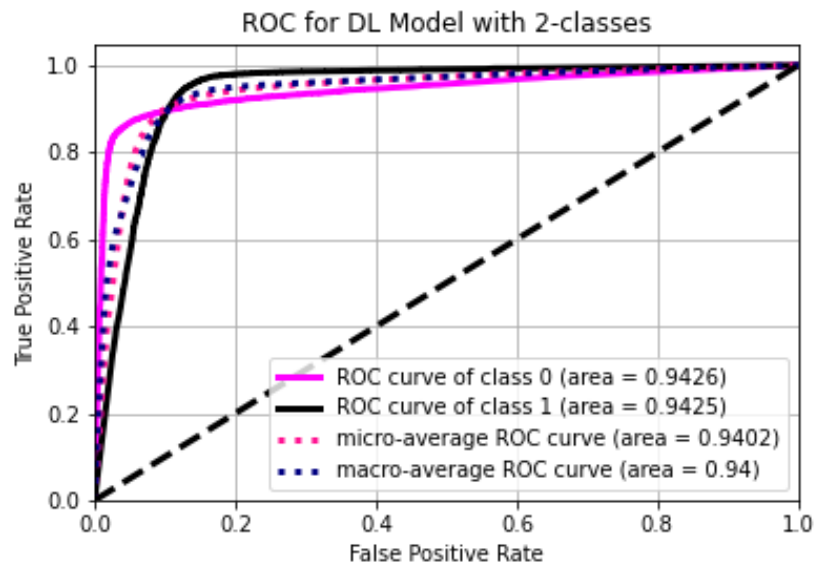


Figure 40 ROC Curve DL - GB

E. Feature Importance:

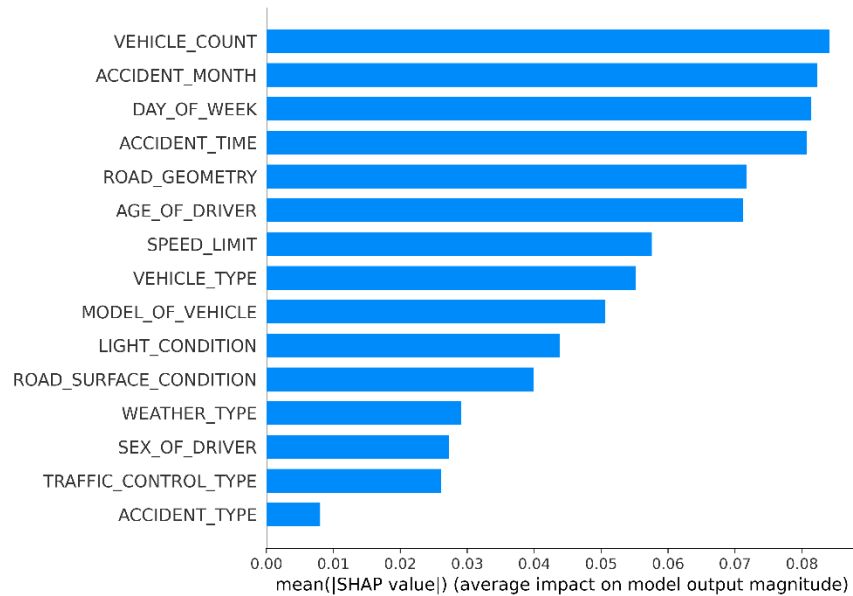


Figure 41 Feature Importance DL - GB

We notice from *figures 39,40 and 41* a slightly enhanced prediction accuracy compared to the values attained in the previous subsection. However, the important features vary from those recorded earlier.

4. USA

A. Data:

$X_{\text{train}} = (83057, 15)$, $X_{\text{test}} = (35597, 15)$, $Y_{\text{train}} = (83057, 1)$, and $Y_{\text{test}} = (35597, 1)$

B. Confusion Matrix:

Output Class					
		0	1	2	3
	0	90.1% 9.9%	89.4% 10.6%	90.7% 9.3%	
	1	89.4% 10.6%	14904 41.9%	629 1.8%	96.0% 4.0%
	2	90.7% 9.3%	2895 8.1%	17169 48.2%	85.6% 14.4%
	3		83.7% 16.3%	96.5% 3.5%	90.1% 9.9%
		Target Class			

Figure 42 Confusion Matrix DL - USA

C. Processing Time:

Average Processing time to test one sample = 0.000428412 secs.

D. ROC:

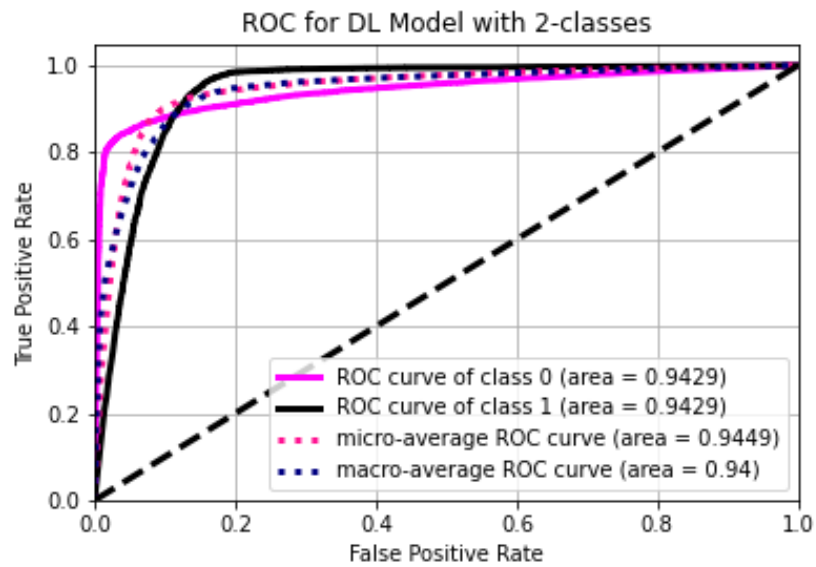


Figure 43 ROC Curve DL - USA

E. Feature Importance:

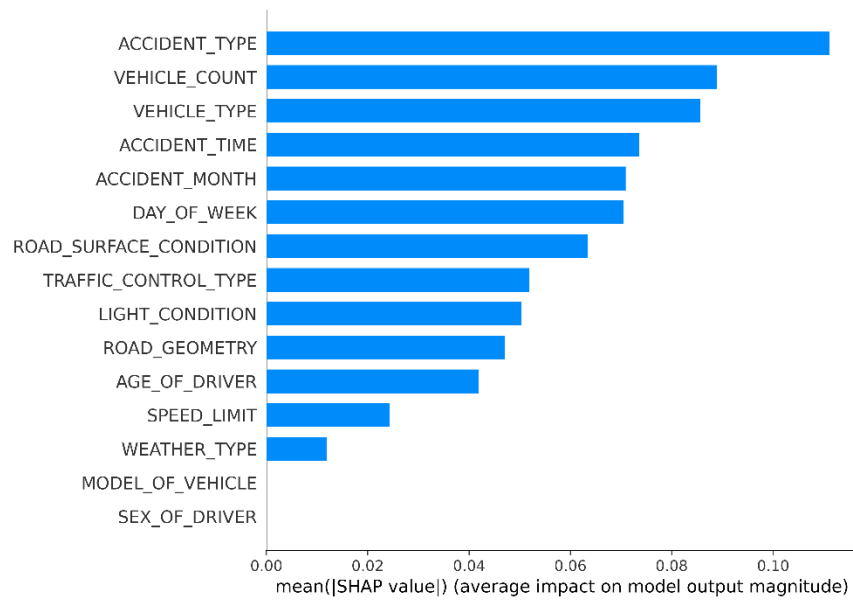


Figure 44 Feature Importance DL - USA

We notice from *figures 41,43 and 44* a slightly enhanced prediction accuracy compared to the values attained in the previous subsection. Also, the important features match those recorded earlier.

In a nutshell, if we assess the model performance based on the individual countries, we find it underperforming for Australia, better performing for Canada and consistent for GB and the USA. This can be caused by many factors varying from one country to the other and requires its' own analysis to get to the bottom why the model performed well in one place but did not in the other. Plus, analyzing the underlying conditions that made the model rearrange feature importance -in some cases-. Average sample processing time was fairly near to original average with marginal savings. Finally, three out of four is a success score.

5.5 Summary

In this chapter we understood the last procedure on data which was splitting 70/30, then we defined the hyperparameters and went through the techniques for tuning and chose grid search, after that we spoke about the evaluation metrics and defined some of them and provided their equations and a confusion matrix that combines all of them and showed ROC with explanation. Lastly, we displayed the tuned parameters, results being presented by the confusion matrix and ROC curve in addition to processing time of a sample, these last three will act as precursors for the comparison in the next chapter.

|

CHAPTER 6

COMPARISON AND DISCUSSION

After presenting the results for the seven different models it is time that we compare these results and discuss them with respect to different measures.

6.1 Accuracy

Accuracy is one of the most important measures and can be divided into three categories: positive class accuracy, negative class accuracy, and the average of both provides avg. accuracy. *Table 25* below shows the accuracies for the seven classifiers.

Table 25 Comparison of accuracies for the 7 studied models

Model	Non-severe Accuracy	Severe Accuracy	Avg. Accuracy
K-NN	95.8%	75.1%	82.4%
SVM	91.4%	80.6%	85.2%
RF	85.6%	80.8%	83.0%
DT	88.0%	77.4%	81.8%
NB	64.4%	65.0%	64.7%
NN	68.2%	69.5%	68.9%
DL	93.3%	81.9%	86.7%

It can be noted that the model with the highest accuracy with respect to non-sever class is K-NN (95.8%), followed by DL (93.3%), then SVM (91.4%) while NB came in last (64.4%). However, with regards to severe class DL (81.9%) came in first place, second was RF (80.8%), third was SVM (80.6%) and NB kept the last place (65.0%). Lastly, on

terms of avg. accuracy best was DL (86.7%), followed by SVM (85.2%), then RF (83.0%) and NB (64.7%) was the worst.

6.2 Recall

Recall is also another important measure which is as significant as accuracy, it is split into two categories depending on the classes, as shown in *table 26* below.

Table 26 Comparison of recall for the 7 studied models

Model	Non-severe Recall	Severe Recall
K-NN	67.5%	97.1%
SVM	77.5%	92.8%
RF	79.1%	86.8%
DT	73.5%	90.1%
NB	64.8%	64.6%
NN	69.9%	67.8%
DL	78.9%	94.4%

According to non-severe class RF (79.1%) comes first, followed marginally by DL (78.9%), then SVM (77.5%) and NB (64.8%) came last. But, when severe class is in question KNN (97.1%) takes the lead, followed by DL (94.4%), then SVM (92.8%) while again NB (64.8%) is in last place.

6.3 F1 Score

Perhaps, one of the most -if not the most- important measures, especially in the situation if imbalanced data as it is a combination of recall and accuracy. It can be seen from *table 27* below that F1 score is split into three categories similar to that of accuracy.

Table 27 Comparison of F1 Scores for the 7 studied models

Model	Non-severe F1 Score	Severe F1 Score	Avg. F1 Score
K-NN	79.2%	84.7%	82.0%
SVM	83.9%	86.3%	85.1%
RF	82.2%	83.7%	82.9%
DT	80.1%	83.3%	81.7%
NB	64.6%	64.8%	64.7%
NN	69.0%	68.6%	68.8%
DL	85.5%	87.7%	86.6%

It is notable that the model with the highest F1 score with respect to non-sever class is DL (85.5%), followed by SVM (83.9%), then RF (82.2%) while NB came in last (64.6%). Also, with regards to severe class DL (87.7%) came in first place, second was SVM (86.3%), third was K-NN (84.7%) and NB kept the last place (64.8%). Lastly, on terms of avg. F1 score best was DL (86.6%), followed by SVM (85.1%), then RF (82.9%) and NB (64.7%) was the worst.

6.4 Average Processing time

A very vital tool that every researcher considers when carrying out a study and performing a model simulation is the time cost attached with each model in his study. *Table 28* below gives the average sample process time.

Table 28 Comparison of Avg. processing time for the 7 studied models

Model	Avg. Sample Processing Time (Micro sec)
K-NN	3500.8
SVM	8482.7

RF	9608.4
DT	72.6
NB	1618.5
NN	336.2
DL	468.6

We can say that according to avg. sample processing time the best three (least time expensive) models are DT (72.6) first, NN (336.2) second, and DL (468.6), all times in microseconds. While the worst three (most time consuming) are RF (9,608.4) absolute worst, then SVM (8,482.7) second worst, after that K-NN (3,500.8) as third worst. NB (1,618.5) stands in the midway.

6.5 Summary

In this chapter we compared models from four different perspectives, with some underlying categories in some of them which would change the game in most cases. Any scholar with adequate logic will seek a model that combines a relatively high value in evaluation measures and a moderate processing time.

CHAPTER 7

CONCLUSION AND FUTURE WORK

7.1 Conclusion

Crash severity prediction is a vital tool, that if employed properly by the law makers and active enforcement agencies can save lives and limit injuries and property losses in traffic crashes, this basically gives governmental bodies an upper hand by providing them a proactive preventive position rather than a reactive one. In this study we collected data from four countries, combined it, then preprocessed it to be fed to ML models. On the other hand, we had six base models that we tuned the hyperparameters for each. Also developed a seventh DL model and tuned its' hyperparameters as well, then subjected the testing set output for all the seven model to evaluation metrics and compared their results together.

DL model coming in first places for severe and average accuracies at 81.9% and 86.7% respectively, also, coming in second place for recall in non-severe and severe at 78.9% and 94.4% respectively, then keeping the first place for the three categories (non-severe, severe, and avg.) for F1 score at (85.5%, 87.7%, and 86.6%) respectively, with an avg. sample process time of only 469 microsecond suggests how practical and robust this model is, as it does not compromise the output quality for processing like NN or NB, nor does it propose intolerant time frames in exchange of better output quality like with SVM and RF.

NB model proved itself o be a bad classifier as it anchored its' results at last place whether in avg. accuracy, severe recall, or avg. F1 score 64.7%, 64.6%, and 64.7% respectively all

that with an avg, sample process time of 1,619 microsec, which is relatively higher than other predictors that provided better results.

Lastly, the results suggest that this approach of generalized prediction is valid and can be adopted for wider application, especially if legislative bodies of various countries agree on passing a law bill to share and use such valuable data with researchers and educational institutions, who in return will contribute to a better transportation environment worldwide.

7.2 Future work

Crash severity prediction field has taken the attention of researchers, for it has an abundance of topics to publish in and touches a lot of aspects of our daily lives, adding to that deep learning subject which is still relatively young and mystic there is a lot of spaces for new creativity.

We take a stand to emphasize that this study is just a corner stone in the great build of generalized crash severity prediction, this humble yet novel effort may change the paths future researchers walk.

One suggestion is to work on adding more countries' crash data and publicize the process, because the more countries' data the more credible the model and its' outputs are. Another suggestion is to incorporate growing (third world) countries as they are in dire need for such models to avoid disasters that occur on daily basis (remember WHO referred in its' 2013 report that more than 90% crashes are contributed by growing countries), for the sake of mentioning this study team tried to incorporate crash data from a fifth country in the middle east region and established a connection with concerned authorities but the request

of data disclosure was met with rejection, however someone with a right set of connections might be able to extract such data easily.

On terms of enhancing future road user safety, our model can be further inspected and upgraded by doing sensitivity analysis to identify the critical factors that led in the past to severe crashes, like (light condition, traffic control type,...etc.) and eliminate or enhance those factors to give the user higher safety and reliability. Another idea is prompt prediction of crash severity for a crash that occurred instantly, using simple data from crash and warning the rescue teams simultaneously to allocate necessary resources based on severity need.

From a programmer's point of view in order to increase reliability of models, by investigating the cases where models produce wrong results individually and comparison among the models for the same cases one model gets correct answers while the other gets wrong ones. Additionally, a spatial distribution is recommended for incorrect results to develop a better understanding.

Lastly, DL specifically and ML in general are a great ocean with endless limits a tiny variation might do, or a different combination might land you in a whole different area.

References

1. Organization, W.H., *Global status report on road safety 2013: supporting a decade of action: summary*. 2013, World Health Organization.
2. Duncan, C.S., A.J. Khattak, and F.M. Council, *Applying the ordered probit model to injury severity in truck-passenger car rear-end collisions*. Transportation Research Record, 1998. **1635**(1): p. 63-71.
3. Kockelman, K.M. and Y.-J. Kweon, *Driver injury severity: an application of ordered probit models*. Accident Analysis & Prevention, 2002. **34**(3): p. 313-321.
4. Siddiqui, N.A., X. Chu, and M. Guttenplan, *Crossing locations, light conditions, and pedestrian injury severity*. Transportation research record, 2006. **1982**(1): p. 141-149.
5. Yau, K.K., H.-P. Lo, and S.H. Fung, *Multiple-vehicle traffic accidents in Hong Kong*. Accident Analysis & Prevention, 2006. **38**(6): p. 1157-1161.
6. Zajac, S.S. and J.N. Ivan, *Factors influencing injury severity of motor vehicle–crossing pedestrian crashes in rural Connecticut*. Accident Analysis & Prevention, 2003. **35**(3): p. 369-379.
7. O'donnell, C. and D.H. Connor, *Predicting the severity of motor vehicle accident injuries using models of ordered multiple choice*. Accident Analysis & Prevention, 1996. **28**(6): p. 739-753.
8. Khorashadi, A., et al., *Differences in rural and urban driver-injury severities in accidents involving large-trucks: an exploratory analysis*. Accident Analysis & Prevention, 2005. **37**(5): p. 910-921.
9. Savolainen, P. and F. Mannering, *Probabilistic models of motorcyclists' injury severities in single-and multi-vehicle crashes*. Accident Analysis & Prevention, 2007. **39**(5): p. 955-963.
10. Al-Ghamdi, A.S., *Using logistic regression to estimate the influence of accident factors on accident severity*. Accident Analysis & Prevention, 2002. **34**(6): p. 729-741.
11. Delen, D., R. Sharda, and M. Bessonov, *Identifying significant predictors of injury severity in traffic accidents using a series of artificial neural networks*. Accident Analysis & Prevention, 2006. **38**(3): p. 434-444.
12. Mussone, L., A. Ferrari, and M. Oneta, *An analysis of urban collisions using an artificial intelligence model*. Accident Analysis & Prevention, 1999. **31**(6): p. 705-718.
13. Li, Z., et al., *Using support vector machine models for crash injury severity analysis*. Accident Analysis & Prevention, 2012. **45**: p. 478-486.
14. De Oña, J., R.O. Mujalli, and F.J. Calvo, *Analysis of traffic accident injury severity on Spanish rural highways using Bayesian networks*. Accident Analysis & Prevention, 2011. **43**(1): p. 402-411.
15. Simoncic, M., *A Bayesian network model of two-car accidents*. Journal of transportation and Statistics, 2004. **7**(2/3): p. 13-25.
16. Karlaftis, M.G. and I. Golias, *Effects of road geometry and traffic volumes on rural roadway accident rates*. Accident Analysis & Prevention, 2002. **34**(3): p. 357-365.

17. Sohn, S.Y. and S.H. Lee, *Data fusion, ensemble and clustering to improve the classification accuracy for the severity of road traffic accidents in Korea*. Safety Science, 2003. **41**(1): p. 1-14.
18. Abdelwahab, H.T. and M.A. Abdel-Aty, *Development of artificial neural network models to predict driver injury severity in traffic accidents at signalized intersections*. Transportation Research Record, 2001. **1746**(1): p. 6-13.
19. Kashani, A.T. and A.S. Mohaymany, *Analysis of the traffic injury severity on two-lane, two-way rural roads based on classification tree models*. Safety Science, 2011. **49**(10): p. 1314-1320.
20. Mujalli, R.O. and J. de Ona, *Injury severity models for motor vehicle accidents: a review*. in *Proceedings of the Institution of Civil Engineers-Transport*. 2013. Thomas Telford Ltd.
21. Fan, W.D., et al., *Identifying and quantifying factors affecting vehicle crash severity at highway-rail grade crossings: Models and their comparison*. 2016.
22. Shibata, A. and K. Fukuda, *Risk factors of fatality in motor vehicle traffic accidents*. Accident Analysis & Prevention, 1994. **26**(3): p. 391-397.
23. Malyschkina, N.V. and F.L. Mannering, *Empirical assessment of the impact of highway design exceptions on the frequency and severity of vehicle accidents*. Accident Analysis & Prevention, 2010. **42**(1): p. 131-139.
24. Shankar, V., F. Mannering, and W. Barfield, *Statistical analysis of accident severity on rural freeways*. Accident Analysis & Prevention, 1996. **28**(3): p. 391-401.
25. Huang, H., H.C. Chin, and M.M. Haque, *Severity of driver injury and vehicle damage in traffic crashes at intersections: a Bayesian hierarchical analysis*. Accident Analysis & Prevention, 2008. **40**(1): p. 45-54.
26. Khattak, A.J., P. Kantor, and F.M. Council, *Role of adverse weather in key crash types on limited-access: roadways implications for advanced weather systems*. Transportation research record, 1998. **1621**(1): p. 10-19.
27. Yasmin, S. and N. Eluru, *Evaluating alternate discrete outcome frameworks for modeling crash injury severity*. Accident Analysis & Prevention, 2013. **59**: p. 506-521.
28. Christoforou, Z., S. Cohen, and M.G. Karlaftis, *Vehicle occupant injury severity on highways: An empirical investigation*. Accident Analysis & Prevention, 2010. **42**(6): p. 1606-1620.
29. Eluru, N. and C.R. Bhat, *A joint econometric analysis of seat belt use and crash-related injury severity*. Accident Analysis & Prevention, 2007. **39**(5): p. 1037-1049.
30. Garrido, R., et al., *Prediction of road accident severity using the ordered probit model*. Transportation Research Procedia, 2014. **3**: p. 214-223.
31. Hu, S.-R., C.-S. Li, and C.-K. Lee, *Investigation of key factors for accident severity at railroad grade crossings by using a logit model*. Safety science, 2010. **48**(2): p. 186-194.
32. Zong, F., H. Xu, and H. Zhang, *Prediction for traffic accident severity: comparing the Bayesian network and regression models*. Mathematical Problems in Engineering, 2013. **2013**.

33. Ratanavaraha, V. and S. Suangka, *Impacts of accident severity factors and loss values of crashes on expressways in Thailand*. IATSS research, 2014. **37**(2): p. 130-136.
34. Wang, Y. and W. Zhang, *Analysis of roadway and environmental factors affecting traffic crash severities*. Transportation research procedia, 2017. **25**: p. 2119-2125.
35. George, Y., T. Athanasios, and P. George, *Investigation of road accident severity per vehicle type*. Transportation research procedia, 2017. **25**: p. 2076-2083.
36. Zhu, X. and S. Srinivasan, *Modeling occupant-level injury severity: An application to large-truck crashes*. Accident Analysis & Prevention, 2011. **43**(4): p. 1427-1437.
37. Smink, B.E., et al., *Drug use and the severity of a traffic accident*. Accident Analysis & Prevention, 2005. **37**(3): p. 427-433.
38. Lee, J. and F. Mannering, *Impact of roadside features on the frequency and severity of run-off-roadway accidents: an empirical analysis*. Accident Analysis & Prevention, 2002. **34**(2): p. 149-161.
39. Abdel-Aty, M., *Analysis of driver injury severity levels at multiple locations using ordered probit models*. Journal of safety research, 2003. **34**(5): p. 597-603.
40. Anarkooli, A.J., M. Hosseinpour, and A. Kardar, *Investigation of factors affecting the injury severity of single-vehicle rollover crashes: a random-effects generalized ordered probit model*. Accident Analysis & Prevention, 2017. **106**: p. 399-410.
41. Sameen, M.I. and B. Pradhan, *Severity prediction of traffic accidents with recurrent neural networks*. Applied Sciences, 2017. **7**(6): p. 476.
42. Abdel-Aty, M.A. and H.T. Abdelwahab, *Predicting injury severity levels in traffic crashes: a modeling comparison*. Journal of transportation engineering, 2004. **130**(2): p. 204-210.
43. Alkheder, S., M. Taamneh, and S. Taamneh, *Severity prediction of traffic accident using an artificial neural network*. Journal of Forecasting, 2017. **36**(1): p. 100-108.
44. Delen, D., et al., *Investigating injury severity risk factors in automobile crashes with predictive analytics and sensitivity analysis methods*. Journal of Transport & Health, 2017. **4**: p. 118-131.
45. Hashmienejad, S.H.-A. and S.M.H. Hasheminejad, *Traffic accident severity prediction using a novel multi-objective genetic algorithm*. International Journal of Crashworthiness, 2017. **22**(4): p. 425-440.
46. Zhang, J., et al., *Comparing Prediction Performance for Crash Injury Severity Among Various Machine Learning and Statistical Methods*. IEEE Access, 2018. **6**: p. 60079-60087.
47. Kunt, M.M., I. Aghayan, and N. Noii, *Prediction for traffic accident severity: comparing the artificial neural network, genetic algorithm, combined genetic algorithm and pattern search methods*. Transport, 2011. **26**(4): p. 353-366.
48. Li, Y., et al., *Identification of significant factors in fatal-injury highway crashes using genetic algorithm and neural network*. Accident Analysis & Prevention, 2018. **111**: p. 354-363.

49. Zheng, M., et al., *Traffic Accident's Severity Prediction: A Deep-Learning Approach-Based CNN Network*. IEEE Access, 2019. **7**: p. 39897-39910.
50. Moomen, M., et al., *Predicting injury severity and crash frequency: Insights into the impacts of geometric variables on downgrade crashes in Wyoming*. Journal of Traffic and Transportation Engineering (English Edition), 2020. **7**(3): p. 375-383.
51. Lin, C., et al., *Factor Identification and Prediction for Teen Driver Crash Severity Using Machine Learning: A Case Study*. Applied Sciences, 2020. **10**(5): p. 1675.
52. Mondal, A.R., M.A.E. Bhuiyan, and F. Yang, *Advancement of weather-related crash prediction model using nonparametric machine learning algorithms*. SN Applied Sciences, 2020. **2**(8): p. 1372.
53. Assi, K., et al., *Predicting Crash Injury Severity with Machine Learning Algorithm Synergized with Clustering Technique: A Promising Protocol*. International Journal of Environmental Research and Public Health, 2020. **17**(15): p. 5497.
54. Ijaz, M., et al., *A comparative study of machine learning classifiers for injury severity prediction of crashes involving three-wheeled motorized rickshaw*. Accident Analysis & Prevention, 2021. **154**: p. 106094.
55. Wahab, L. and H. Jiang, *A comparative study on machine learning based algorithms for prediction of motorcycle crash severity*. PLoS One, 2019. **14**(4): p. e0214966.
56. Zhang, C., et al., *A Crash Severity Prediction Method Based on Improved Neural Network and Factor Analysis*. Discrete Dynamics in Nature and Society, 2020. **2020**: p. 4013185.
57. Tavakoli Kashani, A., R. Rabieyan, and M.M. Besharati, *A data mining approach to investigate the factors influencing the crash severity of motorcycle pillion passengers*. Journal of Safety Research, 2014. **51**: p. 93-98.
58. Rahim, M.A. and H.M. Hassan, *A deep learning based traffic crash severity prediction framework*. Accident Analysis & Prevention, 2021. **154**: p. 106090.
59. Zeng, Q. and H. Huang, *A stable and optimized neural network model for crash injury severity prediction*. Accident Analysis & Prevention, 2014. **73**: p. 351-358.
60. Liu, L., et al. *An Ensemble of Multiple Boosting Methods Based on Classifier-Specific Soft Voting for Intelligent Vehicle Crash Injury Severity Prediction*. in *2020 IEEE Sixth International Conference on Big Data Computing Service and Applications (BigDataService)*. 2020.
61. Zhu, X. and S. Srinivasan, *A comprehensive analysis of factors influencing the injury severity of large-truck crashes*. Accident; analysis and prevention, 2011. **43**: p. 49-57.
62. Wei, X., et al., *Analyzing Traffic Crash Severity in Work Zones under Different Light Conditions*. Journal of Advanced Transportation, 2017. **2017**: p. 5783696.
63. Rezapour Mashhadi, M.M., S. Nazneen, and K. Ksaibati, *Application of deep learning techniques in predicting motorcycle crash severity*. Engineering Reports, 2020. **2**.
64. Song, X., et al., *Determinants and Prediction of Injury Severities in Multi-Vehicle-Involved Crashes*. International Journal of Environmental Research and Public Health, 2021. **18**(10): p. 5271.

65. Yahaya, M., et al., *Ensemble-based model selection for imbalanced data to investigate the contributing factors to multiple fatality road crashes in Ghana*. Accident Analysis & Prevention, 2021. **151**: p. 105851.
66. Jamal, A. and W. Umer, *Exploring the Injury Severity Risk Factors in Fatal Crashes with Neural Network*. International Journal of Environmental Research and Public Health, 2020. **17**(20): p. 7466.
67. Fiorentini, N. and M. Losa, *Handling Imbalanced Data in Road Crash Severity Prediction by Machine Learning Algorithms*. Infrastructures, 2020. **5**(7): p. 61.
68. Alnami, H.M., I. Mahgoub, and H. Al-Najada. *Highway Accident Severity Prediction for Optimal Resource Allocation of Emergency Vehicles and Personnel*. in *2021 IEEE 11th Annual Computing and Communication Workshop and Conference (CCWC)*. 2021.
69. Mokhtarimousavi, S., et al., *Improved Support Vector Machine Models for Work Zone Crash Injury Severity Prediction and Analysis*. Transportation Research Record, 2019. **2673**(11): p. 680-692.
70. Panicker, A.K. and G. Ramadurai, *Injury severity prediction model for two-wheeler crashes at mid-block road sections*. International Journal of Crashworthiness, 2020: p. 1-9.
71. Hosseinzadeh, A., A. Moeinaddini, and A. Ghasemzadeh, *Investigating factors affecting severity of large truck-involved crashes: Comparison of the SVM and random parameter logit model*. Journal of Safety Research, 2021. **77**: p. 151-160.
72. Flach, P., *Machine learning: the art and science of algorithms that make sense of data*. 2012: Cambridge University Press.
73. Murphy, K.P., *Machine learning: a probabilistic perspective*. 2012: MIT press.
74. Kuhn, M. and K. Johnson, *Applied predictive modeling*. Vol. 26. 2013: Springer.
75. Barber, D., *Bayesian reasoning and machine learning*. 2012: Cambridge University Press.
76. Pal, M., *Random forest classifier for remote sensing classification*. International journal of remote sensing, 2005. **26**(1): p. 217-222.
77. Rokach, L. and O. Maimon, *Decision Trees*, in *Data Mining and Knowledge Discovery Handbook*, O. Maimon and L. Rokach, Editors. 2005, Springer US: Boston, MA. p. 165-192.
78. Géron, A., *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems*. 2019: O'Reilly Media.
79. Güllü, M. and I. Yilmaz. *Application of Back Propagation Artificial Neural Network for Modelling Local GPS/Levelling Geoid Undulations: A Comparative Study*. 2011.
80. Nemer, I., et al., *RF-Based UAV Detection and Identification Using Hierarchical Learning Approach*. Sensors, 2021. **21**(6): p. 1947.
81. Brownlee, J. *How to use learning curves to diagnose machine learning model performance*. Machine Learning Mastery. 2019 [cited 2021 Nov 27]; Available from: <https://machinelearningmastery.com/wp-content/uploads/2019/10/Depiction-of-a-ROC-Curve.jpg>.

Vitae

Name	:OSAMA J. MOSAMEH
Nationality	:JORDANIAN
Date of Birth	:4/18/1990
Email	:Ojm.mosameh@gmail.com
Address	:11947 Amman, JO
Academic Background	:MSc in CIVIL ENGINEERING
Publications	:“An Interpretable Deep Machine Learning Model for Crash Severity Prediction – Use of International Big Data”, Journal of Accident Analysis and Prevention, Elsevier (Submitted Dec 29,2021)