

**MODELING AND PREDICTING TRAFFIC CRASH SEVERITY
USING ARTIFICIAL INTELLIGENCE TECHNIQUES**

BY

UMER MANSOOR

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

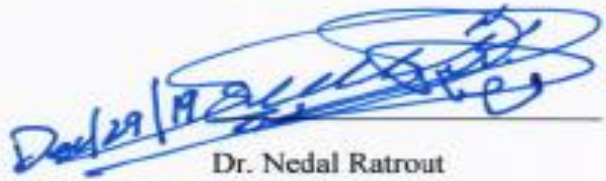
In

CIVIL AND ENVIRONMENTAL ENGINEERING

December 2019

KING FAHD UNIVERSITY OF PETROLEUM & MINERALS
DHAHRAN- 31261, SAUDI ARABIA
DEANSHIP OF GRADUATE STUDIES

This thesis, written by **UMER MANSOOR** 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**

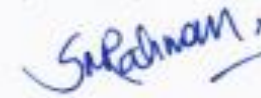

Dr. Nedal Ratrouf
(Advisor)

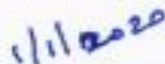

Dr. Salah U. Al-Dulaijan
Department Chairman


Dr. Khaled Assi
(Member)


Dr. Salam A. Zummo
Dean of Graduate Studies




Dr. Syed Masiur Rahman
(Member)


Date

© Umer Mansoor

2019

Dedicated to my Beloved Parents, Brother and Family

ACKNOWLEDGMENTS

I am highly grateful to my advisor, Professor Nedal Ratrouf for his valuable advice and guidance throughout my thesis work. His continuous guidance helped me achieve the objectives of this thesis in the most professional manner. I am also thankful to my thesis Committee Members: Dr. Khaled Assi and Dr. Syed Masiur Rahman for their encouragement, technical support and valuable recommendations.

I am thankful to my parents, brother and family for their support and prayers for my success.

TABLE OF CONTENTS

ACKNOWLEDGMENTS	V
TABLE OF CONTENTS	vi
LIST OF TABLES.....	IX
LIST OF FIGURES.....	XI
LIST OF ABBREVIATIONS.....	XV
ABSTRACT	XVI
ملخص الرسالة	XVIII
CHAPTER 1 INTRODUCTION.....	1
1.1 Background	1
1.2 Significance of Research.....	2
1.3 Objectives of Research	3
CHAPTER 2 LITERATURE REVIEW	4
2.1 Background of Machine Learning	4
2.2 Previous studies	4
2.3 Classification Techniques	10

2.3.1	K-Nearest Neighbor	10
2.3.2	Decision Tree	11
2.3.3	AdaBoost	14
2.3.4	Artificial Neural Network.....	15
2.3.5	Support Vector Machine	17
CHAPTER 3 METHODOLOGY.....		19
3.1	Great Britain’s Crash Dataset.....	20
3.1.1	Data Description and Preparation.....	20
3.1.2	Data Standardization	23
3.1.3	Model Development	23
3.2	Canada’s Crash Dataset	31
3.2.1	Data Description and Preparation.....	31
3.2.2	Model Development	34
CHAPTER 4 RESULTS AND ANALYSIS		41
4.1	Modeling Results for all Locations Great Britain	43
4.1.1	Accuracy of Models for all Locations Great Britain	43
4.1.2	Performance Measures for all Locations Great Britain	50
4.2	Modeling Results for Road Intersections Great Britain	53
4.2.1	Accuracy of Models Road Intersections Great Britain	54
4.2.2	Performance Measures for Road Intersections Great Britain	61
4.3	Modeling Results for all locations Canada	64
4.3.1	Accuracy of Models all Locations Canada	64

4.3.2	Performance Measures for All locations Canada	71
4.4	Modeling Results for Road Intersections Canada	74
4.4.1	Accuracy of Models Road Intersections Canada	75
4.4.2	Performance Measures for Road Intersections Canada	82
4.5	Models Comparison	85
4.6	Simplified Ensemble Model	88
CHAPTER 5 CONCLUSIONS AND RECOMMENDATIONS.....		94
5.1	Conclusions	94
5.2	Future Work	97
REFERENCES.....		98
VITAE		106

LIST OF TABLES

Table 3-1 : Description of Input Features for Great Britain	22
Table 3-2 : Crash Severity Level	23
Table 3-3 : Percent of variance explained by each component	29
Table 3-4 : Rotated Component loadings for PCA of crash data	30
Table 3-5 : Description of Input Features for Canada	33
Table 3-6 : Percent of variance explained by each component	39
Table 3-7 : Rotated Component loadings for PCA of crash data	40
Table 4-1 : Confusion Matrix	41
Table 4-2 : Accuracy of different models all locations GB	43
Table 4-3 : Performance measures for Severe Crashes all locations GB	51
Table 4-4 : Performance measures for Non-Severe Crashes all locations GB	52
Table 4-5 : Accuracy of all Models intersections GB.....	54
Table 4-6 : Performance Measures for Severe Crashes Road Intersections GB.....	61
Table 4-7 : Performance Measures for Non-Severe Crashes Road Intersections GB	62
Table 4-8 : Accuracy of Models for all locations Canada	64
Table 4-9 : Performance measures for Fatal Crashes Canada	72
Table 4-10 : Performance measures for Non-Fatal Crashes Canada	73
Table 4-11 : Accuracy of Models Road Intersections Canada	75
Table 4-12 : Performance Measures Non-Fatal Crashes Road Intersections Canada	82
Table 4-13 : Performance Measures Fatal Crashes Road Intersections Canada.....	83
Table 4-14 : Comparison of Testing Accuracy for Both Datasets.....	85
Table 4-15 : Comparison of performance measures for both Datasets Severe Crashes ...	86

Table 4-16 : Comparison of performance measures for Non-Severe Crashes.....	87
Table 4-17 : Accuracy of Simplified Ensemble Model	89

LIST OF FIGURES

Figure 2-1: Illustration of K nearest neighbor	11
Figure 2-2 : Illustration of Decision Tree	13
Figure 2-3 : Steps for learning process of AdaBoost Model	14
Figure 2-4 : Illustration of Final Prediction by AdaBoost Model.....	15
Figure 2-5 : Typical feed forward neural network.....	16
Figure 2-6 : Maximum Separation Hyperplane	18
Figure 3-1 : Flowchart of Methodology.....	19
Figure 3-2 : Structure for two-layer Ensemble Model.....	28
Figure 3-3 : Scree plot for Principal Components	29
Figure 3-4 : Component Plot in rotated space	31
Figure 3-5 : Structure for two-layer Ensemble Model for Canada Dataset	38
Figure 3-6 : Scree plot for Principal Components	39
Figure 3-7 : Component Plot in rotated space	40
Figure 4-1 : Accuracy of different models all locations GB.....	44
Figure 4-2 : KNN Training CM all locations GB	44
Figure 4-3 : KNN Testing CM all locations GB.....	45
Figure 4-4 : DT Training CM all locations GB	45
Figure 4-5 : DT Testing CM all locations GB	46
Figure 4-6 : AdaBoost Training CM all locations GB.....	46
Figure 4-7 : AdaBoost Testing CM all locations GB	47
Figure 4-8 : ANN Training CM all locations GB	47
Figure 4-9 : ANN Testing CM all locations GB.....	48

Figure 4-10 : SVM Training CM all locations GB	48
Figure 4-11 : SVM Testing CM all locations GB	49
Figure 4-12 : Ensemble Training CM all locations GB	49
Figure 4-13 : Ensemble Testing CM all locations GB	50
Figure 4-14 : Performance measures of Models for Severe Crashes all locations GB.....	51
Figure 4-15 : Performance measures for Non-Severe Crashes all locations GB	52
Figure 4-16 : Accuracy of all Models intersections GB	54
Figure 4-17 : KNN Training CM Road Intersections GB.....	55
Figure 4-18 : KNN Testing CM Road Intersections GB	55
Figure 4-19 : DT Training CM Road Intersections GB	56
Figure 4-20 : DT Testing CM Road Intersections GB.....	56
Figure 4-21 : AdaBoost Training CM Road Intersections GB	57
Figure 4-22 : AdaBoost Testing CM Road Intersections GB	57
Figure 4-23 : ANN Training CM Road Intersections GB.....	58
Figure 4-24 : ANN Testing CM Road Intersections GB	58
Figure 4-25 : SVM Training CM Road Intersections GB	59
Figure 4-26 : SVM Testing CM Road Intersections GB	59
Figure 4-27 : Two-Layer Ensemble Training CM Road Intersections GB	60
Figure 4-28 : Two-Layer Ensemble Testing CM Road Intersections GB	60
Figure 4-29 : Performance Measures for severe crashes Road Intersections GB.....	62
Figure 4-30 : Performance measures for Non-Severe crashes Road Intersections GB	63
Figure 4-31 : Accuracy of Models all locations Canada.....	65
Figure 4-32 : KNN Training CM Canada	65

Figure 4-33 : KNN Testing CM Canada.....	66
Figure 4-34 : DT Training CM Canada	66
Figure 4-35 : DT Testing CM Canada	67
Figure 4-36 : AdaBoost Training CM Canada.....	67
Figure 4-37 : AdaBoost Testing CM Canada	68
Figure 4-38 : ANN Training CM Canada	68
Figure 4-39 : ANN Testing CM Canada.....	69
Figure 4-40 : SVM Training CM Canada	69
Figure 4-41 : SVM Testing CM Canada.....	70
Figure 4-42 : Two-Layer Ensemble Training CM Canada	70
Figure 4-43 : Two-Layer Ensemble Testing CM Canada.....	71
Figure 4-44 : Performance measures for Fatal Crashes Canada	72
Figure 4-45 : Performance measures for Non-Fatal Crashes Canada.....	73
Figure 4-46 : Accuracy of Models Road Intersections Canada	75
Figure 4-47 : KNN Training CM Road Intersections Canada	76
Figure 4-48 : KNN Testing CM Road Intersections Canada.....	76
Figure 4-49 : DT Training CM Road Intersections Canada	77
Figure 4-50 : DT Testing CM Road Intersections Canada	77
Figure 4-51 : AdaBoost Training CM Road Intersections Canada.....	78
Figure 4-52 : AdaBoost Testing CM Road Intersections Canada.....	78
Figure 4-53 : ANN Training CM Road Intersections Canada	79
Figure 4-54 : ANN Testing CM Road Intersections Canada.....	79
Figure 4-55 : SVM Training CM Road Intersections Canada	80

Figure 4-56 : SVM Testing CM Road Intersections Canada.....	80
Figure 4-57 : Ensemble Training CM Road Intersections Canada	81
Figure 4-58 : Ensemble Testing CM Road Intersections Canada.....	81
Figure 4-59 : Performance Measures Non-Fatal Crashes Road Intersections Canada	83
Figure 4-60 : Performance Measures Fatal Crashes Road Intersections Canada	84
Figure 4-61 : Comparison of Testing Accuracy for Both Datasets	86
Figure 4-62 : Comparison of performance measures for both Datasets Severe Crashes..	87
Figure 4-63 : Comparison of performance measures for Non-Severe Crashes	88
Figure 4-64 : Simplified Ensemble training CM all location GB.....	89
Figure 4-65 : Simplified Ensemble testing CM all locations GB	90
Figure 4-66 : Simplified Ensemble training CM road intersections GB	90
Figure 4-67 : Simplified Ensemble testing CM road intersections GB	91
Figure 4-68 : Simplified Ensemble training CM all location Canada	91
Figure 4-69 : Simplified Ensemble testing CM all locations Canada.....	92
Figure 4-70 : Simplified Ensemble training CM road intersections Canada	92
Figure 4-71 : Simplified Ensemble testing CM road intersections Canada.....	93

LIST OF ABBREVIATIONS

ANN	:	Artificial Neural Network
FNN	:	Feedforward Neural Network
SVM	:	Support Vector Machine
KNN	:	K-Nearest Neighbor
DT	:	Decision Tree
PM	:	Performance Measures
GB	:	Great Britain

ABSTRACT

Full Name : UMER MANSOOR

Thesis Title : MODELING AND PREDICTING TRAFFIC CRASH SEVERITY USING
ARTIFICIAL INTELLIGENCE TECHNIQUES

Major Field : TRANSPORTATION ENGINEERING

Date of Degree : [December. 2019]

Road traffic crashes are considered a threat across the globe and are the 8th leading cause of fatality according to World Health Organization (WHO) report. Numerous unfortunate victims in traffic crashes don't get ideal treatment because their injury severity isn't understood at an early stage. Crash severity prediction is beneficial for trauma centers to estimate the potential damage resulting from a traffic crash and provide the appropriate emergency treatment. A two-layer ensemble model is proposed in this study to predict the traffic crash severity. The first layer integrates four basic machine learning classification models: K Nearest Neighbor (KNN), Decision Tree (DT), Adaptive Boosting (AdaBoost), and Support Vector Machine (SVM); the second layer classifies the crash severity based on Artificial Neural Network (ANN) model. The models are developed using Great Britain's traffic crash data for a 6-year period (2011-2016) for all locations collectively and for road intersections specifically. Only the features which can be easily obtained within no time of a crash are used as an input. To simplify the two-layer ensemble model Principal component analysis (PCA) technique is used for dimensionality reduction in the second layer of the model. In model validation, the performance of the two-layer ensemble model is compared with five base models namely; KNN, DT, AdaBoost, SVM, and ANN model. The prediction results reveal that

the ensemble model outperforms the five base classification models based on two performance indicators: accuracy and F1 score. To check the transferability of the developed models, the models are recalibrated and tested using the Traffic crash data for the 3-year period (2014-2016) from the National Collision Database (NCDB) Canada. Model testing for the Canadian dataset reveals that all the models perform in the same line as that for Great Britain. Just like Great Britain's dataset, the two-layer ensemble model outperforms all the other models in terms of accuracy and F1 score for the Canadian dataset. The study also concluded that there is no need of separate models for road intersections, as the models for all locations perform better and are more robust in predicting crash severity. The proposed ensemble technique would be beneficial in predicting crash severity with high accuracy based on limited initial crash information obtained from the crash location. Thus, based on this information the trauma centers would be able to prepare for suitable and prompt medical treatment.

ملخص الرسالة

الاسم الكامل: عمر منصور

عنوان الرسالة: نمذجة وتنبؤ خطورة حوادث حركة المرور باستخدام تقنيات الذكاء الاصطناعي

التخصص: هندسة النقل

تاريخ الدرجة العلمية: ديسمبر 2019

تعتبر حوادث الطرق المرورية من المهددات حول العالم و تأتي في المرتبة الثامنة في الاسباب المؤدية للوفاة طبقاً لتقرير منظمة الصحة العالمية. العديد من ضحايا حوادث الطرق لا يحصلون على التعامل الطبي الأمثل بسبب سوء تقدير شدة إصابتهم في مرحلة مبكرة. التنبؤ بخطورة الحادث المروري مفيد لمراكز علاج الحوادث لتقدير الضرر المحتمل الناتج عن الحادث و تقديم العلاج المناسب لحالات الطوارئ. في هذه الدراسة تم اقتراح نموذج ذو طبقتين (two-layer ensemble) للتنبؤ بخطورة حوادث المرور. تضم الطبقة الأولى اربعة نماذج تصنيف اساسية وهي (K Nearest Neighbor) و (Decision Tree) و (AdaBoost) و (Support Vector Machine). الطبقة الثانية تقوم بتصنيف خطورة الحادث باستخدام نموذج الشبكة العصبية الاصطناعية (Artificial Neural Network). تم تطوير النماذج باستخدام بيانات حوادث المرور في بريطانيا العظمى لفترة 6 سنوات (2011-2016) لجميع المواقع بشكل جماعي وللتقاطعات على وجه التحديد. تم بناء النماذج باستخدام خصائص الحادث التي يمكن الحصول عليها بسهولة وبسرعة كمدخلات. تم استخدام أسلوب تحليل المكون الرئيسي (Principal Component Analysis) لتقليل الأبعاد في الطبقة الثانية من النموذج . تم التحقق من صحة النموذج عن طريق مقارنة أداء نموذج (Two-layer Ensemble) بخمسة نماذج أساسية وهي: (K Nearest Neighbor) و (Decision Tree) و (AdaBoost) و (Support Vector Machine) و (Artificial Neural Network). كشفت نتائج التنبؤ أن نموذج (Two-layer Ensemble) يتفوق على نماذج التصنيف الأساسية الخمسة بناءً على مؤشرين للأداء: مؤشر الدقة ومؤشر تقييم F1. للتحقق من قابلية استخدام النماذج المبنية في أماكن أخرى ، تم إعادة معايرة النماذج واختبارها باستخدام بيانات حوادث حركة المرور لفترة 3 سنوات (2014-2016) من قاعدة بيانات التصادم الوطنية (NCDB) في كندا. أوضح اختبار النماذج لمجموعة البيانات الكندية أن جميع النماذج تؤدي أداءً

مشابهاً لأدائها في بيانات بريطانيا العظمى. وخلصت الدراسة أيضاً إلى أنه لا توجد حاجة إلى نماذج منفصلة لتقاطعات الطرق , حيث أن النماذج لجميع المواقع تعمل بشكل أفضل وأكثر دقة في التنبؤ بخطورة التصادم. النماذج المقترحة في هذه الدراسة مفيدة في التنبؤ بخطورة التصادم بدقة عالية استناداً على خصائص محدودة للحدث المروري. وبالتالي، بناءً على هذه المعلومات، ستكون مراكز الصدمات قادرة على الاستعداد للعلاج الطبي المناسب والسريع.

CHAPTER 1

INTRODUCTION

1.1 Background

Road traffic crashes are considered a threat across the globe resulting in fatalities and injuries, which lead to economic and societal losses. Around 1.25 million people die annually in traffic crashes leading to an annual economic loss of \$260 billion, while non-fatal crashes affect no less than 20-50 million people per year as reported by World Health Organization (WHO) [1]. Although road intersections account for very little proportion of road facilities but a higher number of traffic crashes occur at intersections in urban areas [2]. A great deal of attention is required to minimize the severity and frequency of crash occurrence. The crash data for past years is opportunistic to researchers for identifying the most significant features in traffic crashes, which could prove to be useful in finding suitable solutions for mitigating traffic safety problems in the future. Researchers working on traffic safety take great interest in modeling crash severity. Crash severity models can predict the expected severity of a crash if occurs, which can be helpful for trauma centers in estimating the potential impacts and in providing appropriate and prompt medical treatment.

Since the crash severity resulting from vehicular collision events is random in nature, thus traditionally many parametric techniques like; “ordered probit model” [3-7],

“ordered logit model” [8], the “multinomial logit model” [9, 10] and the “logistic regression model” [11] have been used for predicting crash severity. These parametric techniques although widely used, suffer from few shortcomings [12, 13]. These techniques require a predefined mathematical form while the presence of outliers and missing values in the dataset negatively affect the outcome of the prediction model. In contrast to the parametric techniques machine learning techniques are non-parametric, which can manage outliers and missing values in the dataset. Recently, several machine learning techniques have been employed for extracting useful information from large traffic crash data sets for different road networks [12, 14-19].

1.2 Significance of Research

Numerous unfortunate victims in traffic crashes don't get ideal treatment because their injury severity isn't understood at an early stage. It is important for trauma centers to receive early indication of the injury severity for optimal treatment. A previous study was conducted to predict on-scene crash severity for occupants using conventional regression approach. However, this study includes multiple features that are not readily available from crash sites [20]. The main objective of this research is to aid the trauma centers and emergency units in predicting crash severity with high accuracy based on limited initial crash information. This initial information could be conveyed to the trauma centers through CCTV cameras or through an emergency phone call from the crash location. Thus, based on this information the trauma centers would be able to predict the injury severity of a crash and prepare accordingly for appropriate and prompt medical treatment. Our contribution in this research is the introduction of a two-layer ensemble machine learning model which could predict the crash severity with high accuracy based on crash

features which could be obtained within no time of a crash. The output of this research can play a significant role in enhancing the road traffic safety, as more lives would be saved with early indication of the crash severity.

1.3 Objectives of Research

The objectives of this study are to develop different models which can predict the severity of road traffic crashes for readily available data. Specific objectives are.

1. Perform appropriate data processing to prepare and arrange the available road traffic crash data for further analysis and processing.
2. Conduct an extensive literature review to find the most common models and techniques used worldwide for road traffic crash severity prediction.
3. Determine the important independent variables for predicting crash severity through appropriate objective approach suitable for the available dataset.
4. Develop road traffic crash severity prediction models for readily available data set using artificial intelligence techniques.
5. Develop prediction models for road traffic crash severity at intersections using artificial intelligence techniques.
6. Test the transferability of developed models on different data set by making minor modifications in the models.

CHAPTER 2

LITERATURE REVIEW

2.1 Background of Machine Learning

Machine learning is a subfield of artificial intelligence (AI) which allows the systems to learn and develop from experience without being explicitly programmed. First attempt to use concept of machine learning was conducted by Thomas Ross in 1930's. With the evolution of computers, it is now possible to use machine learning for modeling complex and hidden relationships in data [21]. Machine learning algorithms are divided in to two main categories namely; supervised and unsupervised. In Supervised learning the algorithm is trained for predicting output based on data having both input and output variables e.g. decision trees, Random forest, K-nearest neighborhood, artificial neural network and support vector machine etc. In unsupervised learning algorithm the hidden patterns of data are identified based on original unlabeled data e.g. cluster analysis and greedy layer-wise. Unsupervised learning unlike the supervised learning tries to understand the basic structure of data giving us more intuition of the data [22].

2.2 Previous studies

The importance of traffic safety studies is highlighted by huge economic and societal losses which are not limited to unnecessary delays for road users, property damage and health costs. Previous studies on traffic safety mainly focused on two main aspects i.e. prediction of traffic crash severity and identification of significant factors affecting traffic

crash severity. Traditionally statistical techniques were used for modeling crash severity. A comprehensive literature review related to methods used for the analysis of crash severity was presented [3]. The most commonly used methods are ordered probit model [5, 6, 23, 24], binary logit model [9, 10, 25] and multinomial logit model [11]. The statistical models have a clear mathematical relationship between independent and dependent variables, so the output is easy to interpret, but they suffer from few limitations like they require an assumption about the distribution of data and the linear relationship between dependent and independent variables is assumed. Incorrect factor estimates are produced if any of these assumptions are violated [12-14]. To overcome these shortcomings of statistical methods many machine learning techniques have been employed. Machine learning techniques do not assume any underlying relationship between dependent and independent variables [15, 26-30]. In this section, some of the studies relating to severity prediction are presented.

Many studies compared between parametric and non-parametric techniques for predicting crash severity. A study was conducted to predict and compare the traffic crash severity using machine learning and statistical techniques. Two statistical techniques employed for the severity prediction were ordered probit (OP) model and multinomial logit model (MNL), while four machine learning techniques namely K-Nearest Neighbor, Decision Tree, Random Forest (RF) and Support Vector Machine were also used. The data for traffic flow, crash severity and road geometry were collected at freeway diverge areas in Florida. The study concluded that the machine learning techniques although suffered from over-fitting issues outperformed the classical statistical techniques in terms of prediction accuracy. Among all these six techniques RF showed the highest overall

prediction accuracy while OP had the lowest accuracy [31]. Another study was carried out to model and compare crash severity using a multinomial logit (MNL) model, mixed multinomial logit (MMNL) model and support vector machine (SVM) model. The models were prepared for rear-end crashes using the crash data for California. The comparison of prediction accuracy among three models showed comparable results, although SVM performed a little better than the other two models [32]. Singh et al. [33] modeled the traffic crash severity using a multinomial logit model and two non-parametric techniques namely random forest and decision tree for a data set in Haryana, India. To balance the crashes by severity level, synthetic minority oversampling and randomize class balancing techniques were used which showed a significant improvement in severity prediction. A comparison of results showed that the RF model performed better than the other two techniques for classifying crash severity levels.

Decision trees and Random forests are widely used as classifiers. J48 Decision Tree Classifier, Random Forest (RF), Instance-Based learning with parameter k (IBk) and multinomial logit model were employed for predicting motorcycle crash severity. Five-year motorcycle crash data having four severity levels was collected from the traffic crash database in Ghana. The study revealed that machine learning techniques outperform the statistical technique in terms of accuracy and efficiency. Location type, collision type, day and week of the crash, road surface condition and shoulder condition were some of the factors determining the motorcycle crash severity [34]. Decision trees (J48, ID3, and CART) and Naïve Bayes algorithms were employed using the WEKA software tool for predicting crash severity. J48 decision tree algorithm outperformed the other data mining classification techniques in terms of accuracy. The significant factors affecting crash

severity were found to be speed limit, weather condition, number of lanes, lighting condition, and accident time [35]. Wang and Kim [36] compared discrete models and tree-based models for predicting crash severity using the data for Maryland State for the year 2017. Multinomial logit (MNL) model belonging to discrete models' class and Random forest (RF) model belonging to the tree-based models' class were employed. RF model outperformed the MNL model in terms of accuracy. Similarly, multilayer perceptron (MLP), rule induction (PART) and Simple Cart models were used for predicting motorcycle crash severity in Ghana. The study revealed that the testing accuracy was highest for the Simple Cart model (73.85%) followed by PART (73.45%) and MLP (72.16%) respectively. Location type, crash type, crash time, settlement type and crash partner were some of the significant factors affecting the injury severity [37].

Artificial neural networks (ANN) have also been widely employed for predicting crash severity. ANN model was developed to identify the significant factors affecting crash severity of alcohol-related vehicle crashes using the data of North Carolina. The validation accuracies for three class neural networks and binary class neural networks were found to be 65.33% and 69.65% respectively. The study revealed that vehicle overturning was the main factor contributing to fatal and injury-related crashes while the use of seat belts and deployment of airbags were found to be significant factors for non-injury crashes [38]. Arhin and Gatiba [39] developed ANN models for predicting traffic crash severity of angle crashes at unsignalized intersections in the USA. Around 25 ANN models were developed for a data set of 3301 data points. The architecture of the best ANN model had 5,10 and 5 neurons in three hidden layers respectively, giving an accuracy of 85.62%. A novel rule-based genetic algorithm (GA) was proposed and

compared to ANN, SVM, and DT using 3 years (2011-2013) data for the Tehran province of Iran. The study revealed that GA outperformed the other three classification models with an accuracy of 87% [40].

Ensemble models like random forests and many others are also used for modeling crash severity. A two-layer stacking framework was employed for predicting crash severity. The stacking model is an ensemble model technique consisting of two layers where the first layer integrates three basic classification models which are random forest (RFs), Adaboost (Adaptive boosting) and Gradient boosting decision tree (GBDT) while the second layer classifies the injury severity using logistic regression model. The stacking models when compared with SVM, Neural network and RF showed better classification performance in terms of accuracy and recall [41]. Jiang et al. [42] conducted a study to model highly unbalanced crash injury severity (CIS) data using ensemble models and global sensitivity analysis. The aim of this study was to efficiently model each severity level since most of the existing methods favor the crash categories with maximum observations. Three ensemble models namely random forest, AdaBoost and Gradient Boost were introduced to model unbalanced CIS data. The study concluded that AdaBoost and Gradient Boost produce better results and more balanced prediction accuracies. Also, global sensitivity analysis revealed that grade percentage, driver restraint, crash type, heavy vehicle percentage, and road characteristics are the most important factors influencing the injury severity.

With the advent of deep learning in almost every field in recent times, researchers in traffic safety also started employing deep learning for modeling crash severity. A Deep learning-based convolutional neural network (CNN) was employed for predicting traffic

crash severity. Data for an 8-year period having three levels of severity (fatal, serious and slight) for the city of Leeds, United Kingdom was used for training and testing the neural network. The feature matrix to grey image algorithm (FM2GI) was used for converting the single feature crash relationship to the grey image. The proposed method was compared with several statistical techniques like logistic regression and machine learning techniques like support vector machine and neural network. The study revealed that the CNN model performed better than all the other techniques [43]. Das et al. [44] used deep learning for modeling crash severity of at-fault motorcycle riders. A Deep scooter (deep learning framework) was used to predict severity using five-year data (2010-2014) for Louisiana State. It was found that the model can predict severity with a testing accuracy of 94% while SVM and multinomial logistic regression models for the same data set showed an accuracy of no more than 78%. Gradient boosting trees, deep learning, and Naive Bayes machine learning techniques were used for predicting traffic crash severity using a raw data of 6 years from Spanish traffic agency. The study concluded that deep learning technique outperformed the other two techniques in terms of accuracy and precision [45].

In the preceding section of literature review, we presented several techniques being used for predicting crash severity. Many statistical techniques are presented followed by machine learning techniques. As a general conclusion, we noticed that machine learning techniques like support vector machine, decision tree, and random forest perform better than the statistical techniques like logit and probit models. Also, we noticed that among the machine learning techniques deep learning models generate the most accurate results

for severity prediction however deep learning requires a huge amount of data for better performance.

2.3 Classification Techniques

Several techniques are used for modeling a classifier. In this study, we used five base classification techniques namely K- Nearest Neighbor, Decision Tree, AdaBoost, Artificial Neural Network and Support Vector Machine for modeling traffic crash severity. After developing these base models, a two-layer ensemble model was developed using these base models. A brief conceptual description of all these techniques is presented in this section.

2.3.1 K-Nearest Neighbor

K nearest neighbor (KNN) is the most basic classification technique, which shall be used as the first choice if the information about data distribution is very little or none. It is a supervised learning technique, where the input contains k nearest training examples and output is class membership. KNN works by classifying an observation based on closest k neighbor observations. The new data point is classified based on the majority of k closest observations. A popular metric used for the KNN algorithm is the Euclidean distance method. Two important features to be identified in KNN are k value and distance function. K value is determined by trying many values of k and selecting the one with the best precision accuracy while Euclidean distance (distance function) is determined as a distance between two-dimensional points [46]. Generally, KNN is classified based on Euclidean distance between training samples and the test sample, as shown in the following equation.

$$d(x_i, x_l) = \sqrt{(x_{i1} - x_{l1})^2 + (x_{i2} - x_{l2})^2 + \dots (x_{ip} - x_{lp})^2} \quad (2.1)$$

Here x_i is an input sample with p features $(x_{i1}, x_{i2}, \dots, x_{ip})$; n is the total number of input samples ($i = 1, 2, \dots, n$) and p is total number of features ($j = 1, 2, \dots, p$).

KNN is clearly illustrated in figure 1 where “red star” and “green square” in the figure represent two separate classes of training data. The value of k is a hyperparameter that can be selected through several heuristic techniques. The greater value of k generally minimizes the noise in classification data but the distinction of boundaries between classes is not clear [47]. In Figure 2-1 when “ $K = 3$ ” the test point falls in category A, while for “ $K = 5$ ” the test point is classified the same as class B.

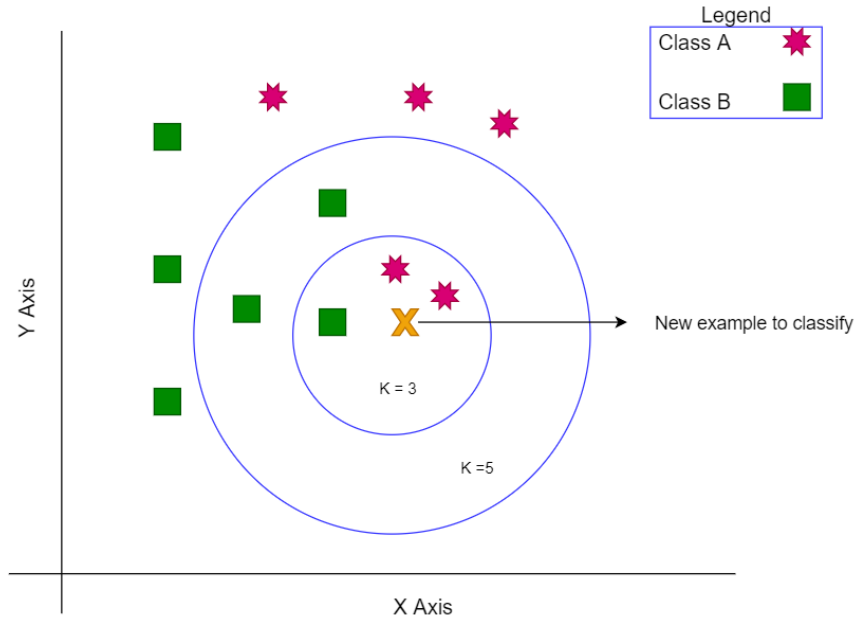


Figure 2-1: Illustration of K nearest neighbor

2.3.2 Decision Tree

The decision tree technique is widely used in machine learning for both regression and classification problems. The goal of the decision tree model is to predict a target value based on many input features as illustrated in (Figure 2-2). Decision tree (DT) is a flow

chart like structure used for classification of data. Decision trees classify the instances by sorting them based on feature values. Each tree has nodes and branches where the node represents the feature to be classified and branch represent the values which a node can assume. The root node is the starting point for the classification of instances in decision tree while instances are sorted based on their feature values. In the decision tree model, the relationship among features is clear, and feature importance is also obvious from the tree [48].

In the root node or first split of a decision tree, all the features are considered, and the training data set is divided into as many groups as the number of features. The feature which is selected first is decided based on the information gain. Information gain is the measurement of information a feature can give about output class. The split with maximum information gain is selected as the first split, and this process continues until the information gain approaches zero. The feature with low information gain does not separate the output classes clearly, while the feature with higher information gain can separate two output classes and bring us closer to the decision [49]. To measure information gain we need to measure the entropy first which is the level of impurity in the training data set. The mathematical relationship for entropy is shown below.

$$Entropy : \sum_i p_i \log_2(p_i) \quad (2.2)$$

$$Entropy (S) = -p_+ \log_2 p_+ - p_- \log_2 p_- \quad (2.3)$$

Where, S is the sample of training examples

p_+ and p_- are proportion of positive and negative training examples respectively

After calculating the entropy, we can calculate the information gain using following equations

$$Gain(S, A) = Entropy(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} \cdot Entropy(S_v) \quad (2.4)$$

Here the set of all possible values for feature A is Values (A), S_v is the subset of S for which feature A has value v. First term in the equation is original entropy while second term is the entropy of children node which is calculated by partitioning S using feature A. Briefly information gain can be shown by following equation.

$$InformationGain = Entropy(Parent\ node) - [AvgEntropy(children\ node)] \quad (2-5)$$

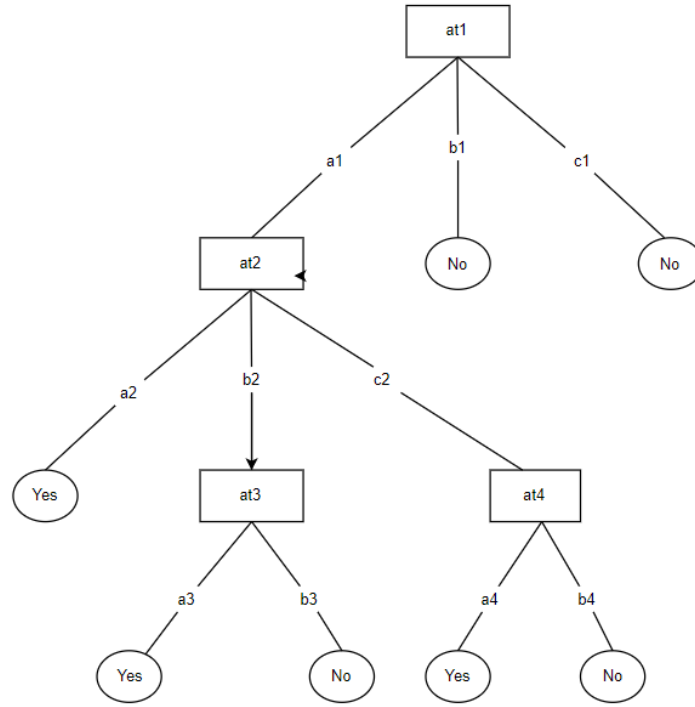


Figure 2-2 : Illustration of Decision Tree

The performance of the DT model can be enhanced by using a technique called pruning. Pruning is cutting off the branches which involve features with little information gain. This way the decision tree becomes simpler and the problem of overfitting is resolved thus enhancing the accuracy of the model [50].

2.3.3 AdaBoost

Adaptive boosting (AdaBoost) is a boosting ensemble model which was first introduced by Freund and Schapire [51]. AdaBoost performs exceptionally well with the decision tree. AdaBoost learns from the misclassification error of pervious data points through many iterations. The main idea of AdaBoost is to train many weak learners using the same training dataset and then construct a strong learner by grouping these weak leaners. AdaBoost has several features: (1) The sampling of dataset remains the same for each iteration and only the distribution of dataset is changed; (2) The change in distribution of dataset depends on the accuracy of classification. In the training dataset D , every sample has a weight w associated with it. The samples which are often misclassified have higher weight while the correctly classified samples have lower weight; (3) Every weak learner in AdaBoost algorithm has a weight represented by vector. The input samples are represented by $[x = (x_1, x_1, \dots, x_i)y_i], i = 1, 2, 3 \dots, m$ while the output is probability of a sample belonging to different severity levels. The steps for learning process and prediction of AdaBoost are shown (Figure 2-3 and Figure 2-4).

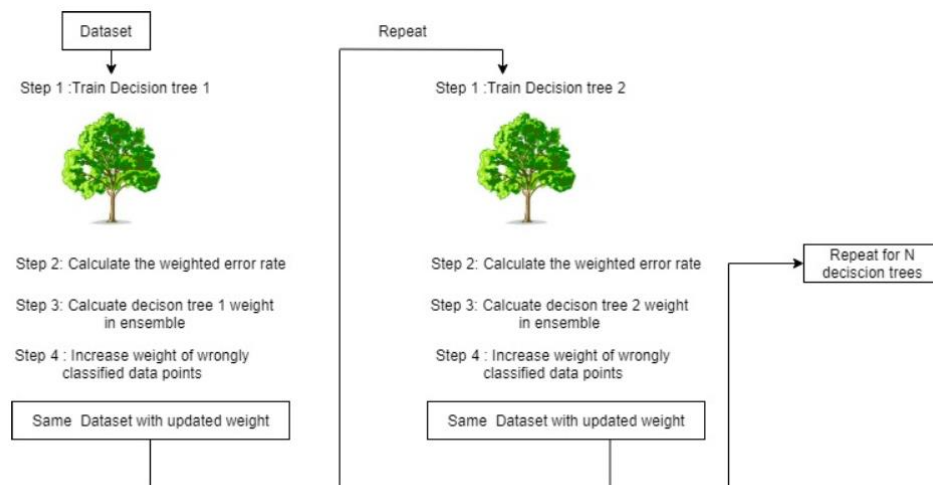


Figure 2-3 : Steps for learning process of AdaBoost Model

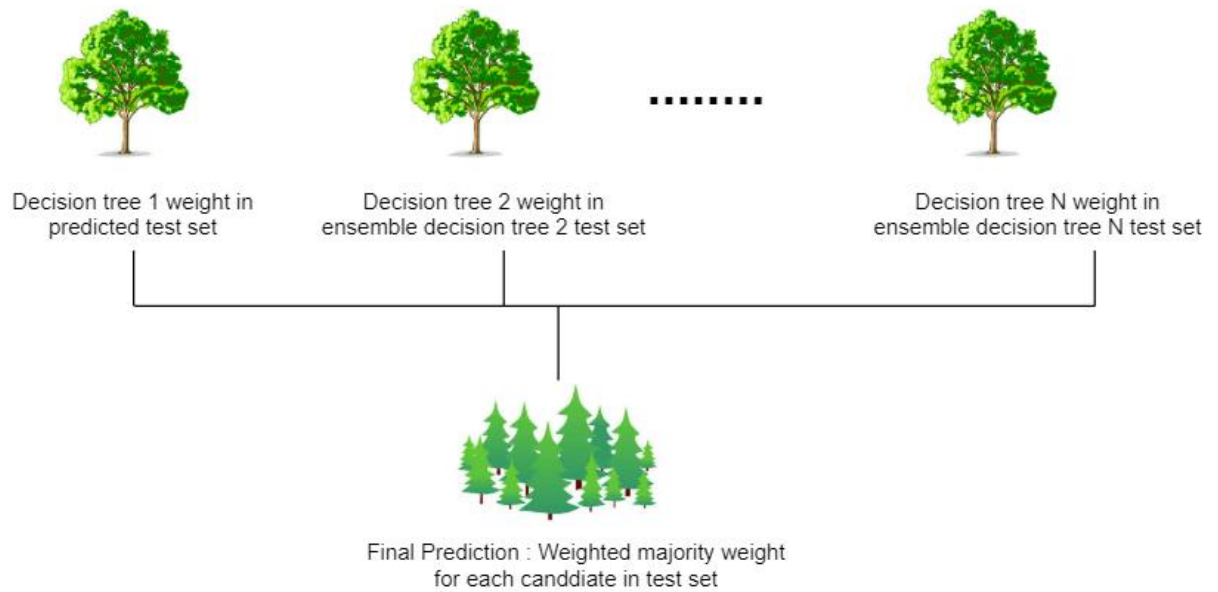


Figure 2-4 : Illustration of Final Prediction by AdaBoost Model

2.3.4 Artificial Neural Network

Artificial neural networks (ANN) are one of the mainly used tools for machine learning which are inspired by human brains. As the name suggests ANN tries to mimic the way a human brain reacts. ANN consists of input, output and in most cases hidden layers. The hidden layers consist of neurons that connect input and output layers by transforming the input layer into something which can be used by the output layer. ANN can find complex patterns in data which are far too complicated to be recognized by human programmer [52, 53].

The basic unit of ANN is an artificial neuron that works by receiving numerical information through several input nodes. After receiving the information neuron processes it internally and generates an outcome. The processing phase consists of two steps; after a linear combination of input variables in the first step, the result is used as an argument for non-linear activation function. Each connection in ANN has a weight term

w_i and a constant bias term Θ . The activation function is a differentiable function which can be either an identity function ($y = x$) or sigmoid function ($y = \frac{1}{1 + e^{-x}}$). In the neural network algorithm, we have some hyperparameters which need to be fixed before the training process starts. Some of the hyperparameters which are determined before training are learning rate, number of hidden layers and batch size. Network architecture is defined by the organization of neurons. For example, in multilayer perceptron (MLP) type, neurons are organized in layers where the same inputs may contribute to neurons of each layer but have no connection with each other. In the feed-forward architecture of ANN, the preceding layer's output is taken as an input for the following layer (Figure 2-5) [52].

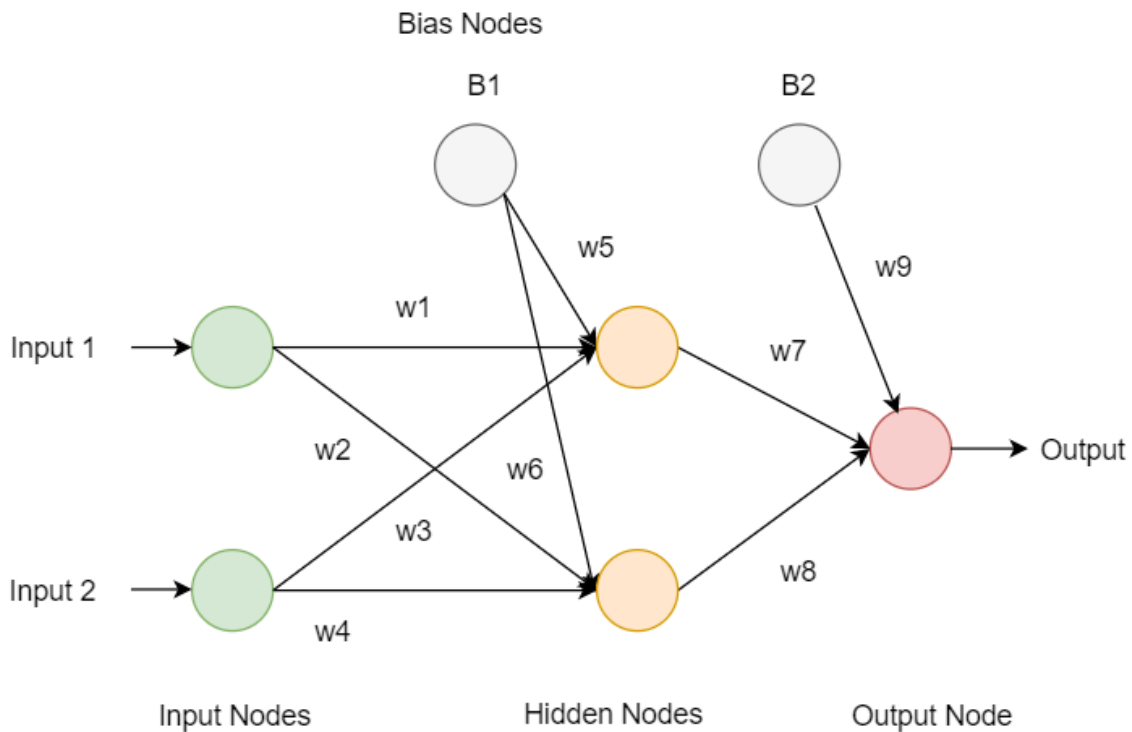


Figure 2-5 : Typical feed forward neural network

2.3.5 Support Vector Machine

Support vector machine (SVM) just like ANN is the most extensively used machine learning technique used for both regression and classification problems. SVM transforms the input data into higher dimension feature space. The main aim of the SVM technique is to find the best hyperplane which maximizes the margin between support vectors as illustrated (Figure 2-6). The points in the data which are nearest to the hyperplane and which if removed would alter the position of hyperplane are called support vectors. If the margin between support vectors is greater, we are more confident about the correct classification of our data by the hyperplane. SVM solves multi-classification problems by finding a hyperplane in high dimensional space for separating points into different groups [54, 55].

In SVM all the crash related input variables are represented by vectors ($x_i \in R^n$) for $i = 1, 2, \dots, N$ and the crash severity which is our training output is represented as ($y_i \in R^n$). The hyperplane which separates our outcome can be formulated as set of point X which satisfy the following equation;

$$W.X - b = 0 \quad (2.6)$$

Where “.” represents dot product and vector W denotes normal vector which is perpendicular to the hyperplane. In binary classification problem SVM needs to be optimized for a given set of input and output variable pairs (x_i, y_i).

$$\min_{w,b,\xi} \frac{1}{2} w^T w + C \sum_{i=1}^N \xi_i \quad (2.7)$$

$$\text{Subject to } y_i(w^T \phi(x_i) + b) \geq 1 - \xi_i, \quad \xi_i \geq 0. \quad (2.8)$$

Where ξ are slack variables that measure misclassification errors and C is the penalty factor to errors that enhance the capacity control of classifier. Lagrange multipliers are

used for minimizing the objective function. Several kernels have been suggested by researchers while radial basis function (RBF) is widely used [54, 56].

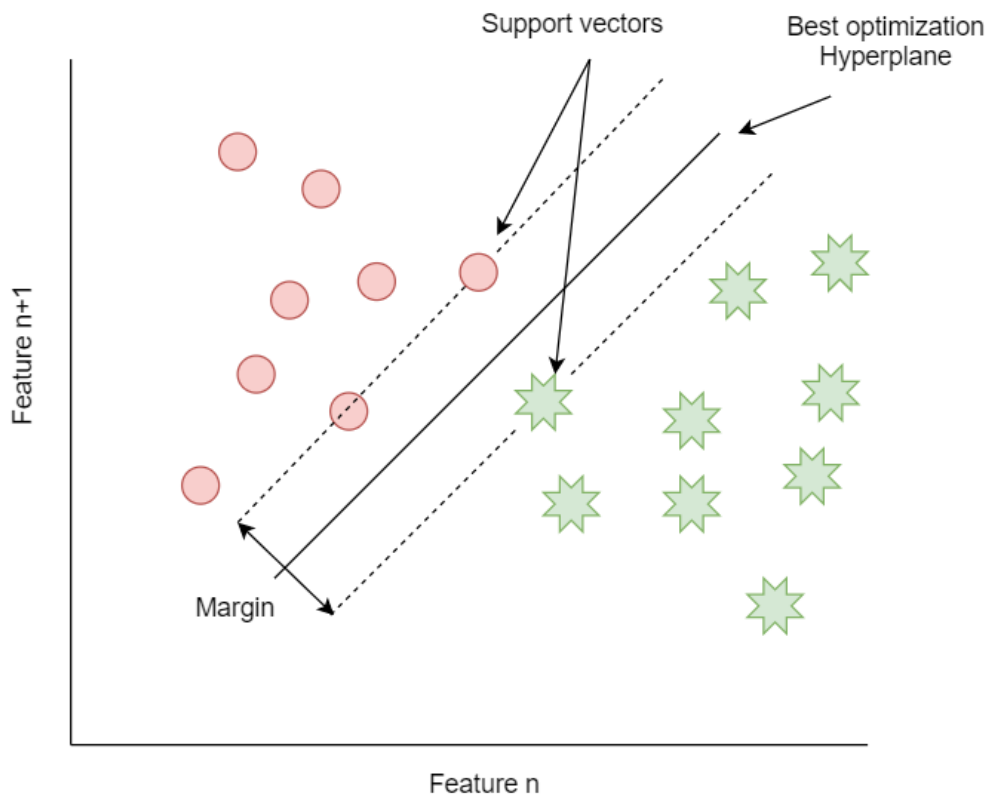


Figure 2-6 : Maximum Separation Hyperplane

CHAPTER 3

METHODOLOGY

A stepwise flow of the methodology for this study is shown in Figure 3-1. A detailed discussion for methodology is provided in this chapter.

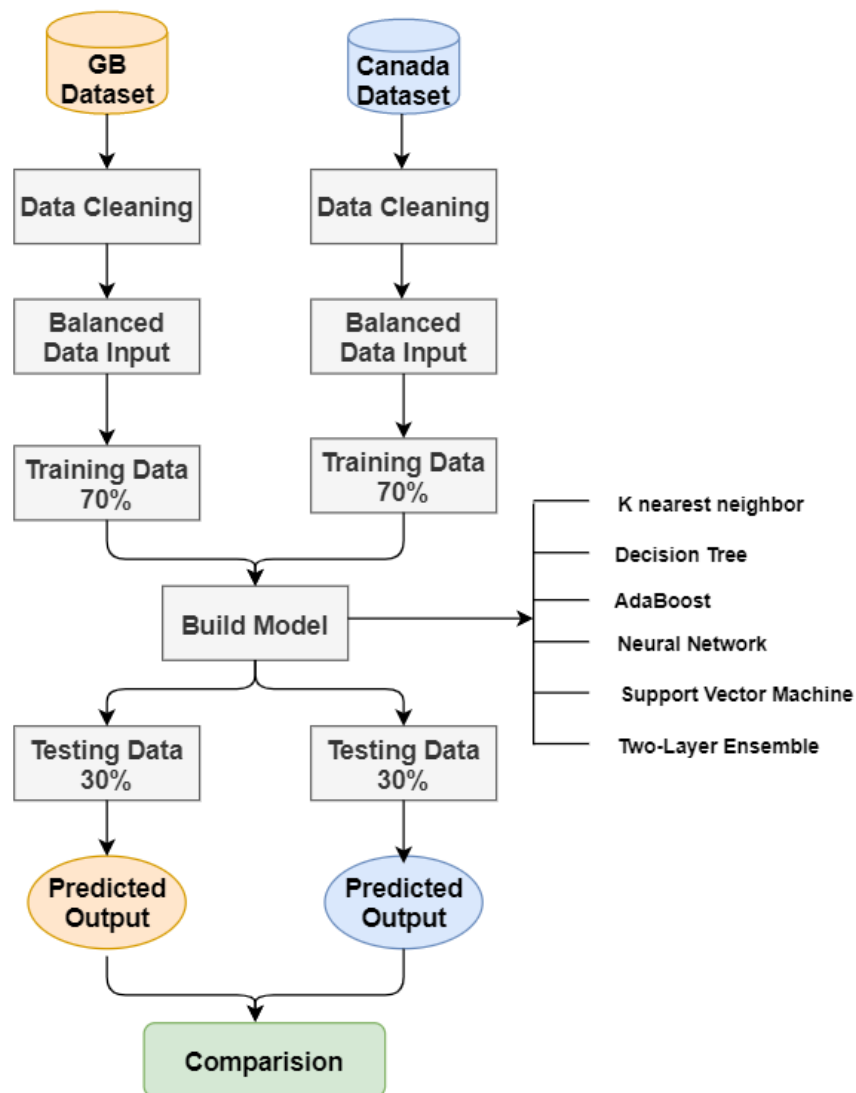


Figure 3-1 : Flowchart of Methodology

3.1 Great Britain’s Crash Dataset

In this part the methodology for model development using the data set for Great Britain is explained in detail.

3.1.1 Data Description and Preparation

Traffic crash models are largely dependent on the availability of crash data, and thus the accuracy of these models relies on the quality of available crash data set. In order to ensure the reliability of crash models, traffic crash data for the 6-year period (2011-2016) from the Department of Transport¹, Great Britain were utilized in this research. Data cleaning was carried out by deleting any duplicate, irrelevant and incomplete data fields. Data cleaning resulted in 440,823 crash records. Features that are instantaneously identifiable at the crash site are used as input so that the limited initial information can be used by trauma centers for predicting crash severity to prepare for the victim’s transport and treatment. Fourteen child features were selected as input from the extracted data set, where every child feature belongs to one of the five parent features. The five parent features are (1) crash feature, (2) roadway feature, (3) environmental feature and (4) vehicle feature and (5) area feature as shown in Table 3-1. A random sample of crashes was extracted from filtered data set using a randomized class balancing procedure [57]. The random sample had 10,000 crash points where equal number of crashes belonged to each severity class. This technique eliminated any possibility of bias towards a specific severity level, which is categorized as slight, serious and fatal in the original data set. Severity level explains the level of injury sustained by individuals involved in a crash. Due to a significantly low percentage of fatal crashes in the data set, fatal and serious

¹ <https://data.gov.uk/dataset/cb7ae6f0-4be6-4935-9277-47e5ce24a11f/road-safety-data>

crashes were categorized as severe crashes. The level of crashes used in the analysis is illustrated in Table 3-2. To fulfill the second objective of this research, the same data was filtered to obtain the data set for road intersections only. A random sample set of 6000 crash points having equal number of crashes for each severity level was picked for analysis. Same input features were used for modeling the traffic crashes at road intersections only.

Table 3-1 : Description of Input Features for Great Britain

Parent Feature	Child Feature	Type	Description
Crash Feature	Number of vehicles involved	Continuous	Minimum number of vehicles involved=1, Maximum number of vehicles involved=5,
	Day of the week	Qualitative	Sunday=1, Monday=2, Tuesday=3, Wednesday=4, Thursday=5, Friday=6, Saturday=7
Roadway Feature	Road surface condition	Qualitative	Dry=1, Wet or damp=2, Snow=3, Frost or ice=4, Flood over 3cm. deep=5, Oil or diesel=6, Mud=7
	Road Type	Qualitative	One-way street=1, Dual Carriageway=2, Single Carriageway=3, Slip road=4, One-way street=5
	1st Road Class	Qualitative	Motorway=1, A(M)=2, A=3, B=4, C=5, Unclassified=6
	Speed Limit	Continuous	Minimum speed limit=20, Maximum speed limit=70
Environmental Feature	Light Condition	Qualitative	Daylight=1, Darkness-lights lit=2, Darkness-lights unlit=3, Darkness no street lighting=4
	Weather Condition	Qualitative	Fine no high winds=1, Raining no high winds=2, Snowing no high winds=3, Fine with high winds=4, Raining with high winds=5, Snowing with high winds=6, For or Mist=7, Others=8
Vehicle feature	Vehicle Type1 Vehicle Type2 Vehicle Type3	Qualitative	Pedal cycle=1, Motorcycle 50cc=2, Motorcycle 125cc=3, Motorcycle over 125cc and up to 500cc=4, Motorcycle over 500cc=5, Taxi/Private hire car=6, Car=7, Minibus 8, Bus or coach =9, Ridden horse=10, Agricultural vehicle=11, Tram=12, Van / Goods 3.5 tonnes mgw or under=13, Goods over 3.5t. and under 7.5t=14, Goods 7.5 tonnes mgw and over=15, Mobility scooter=16, Electric motorcycle=17, Other vehicle=18, Motorcycle - unknown cc=19, Goods vehicle - unknown weight=20
Area Feature	Area Type	Qualitative	Urban=1, Rural=2
	Junction Type	Qualitative	Roundabout=1, T Junction=2, Four arm intersection=3, More than four arms=4
	Junction Control	Qualitative	Authorized person=1, Auto traffic signal=2, Stop sign=3, Give way or uncontrolled=4

Table 3-2 : Crash Severity Level

Crash Severity	Level
Slight	Non-Severe
Serious	Severe
Fatal	

3.1.2 Data Standardization

To achieve accurate results from machine learning models, all the variables utilized in model development shall be on the same scale. Data standardization assists most of the machine learning algorithms to converge by minimizing the loss function. This data scaling is achieved by subtracting the mean value of each variable from the original score of each observation and then dividing it by the standard deviation of the variable. This scaling results in variables transformation having zero mean and unit variance. The standardized value of each variable denoted by Z is calculated using the following equation.

$$Z = \frac{(x-\mu)}{\sigma} \quad (3.1)$$

Where μ is the mean of a variable, σ is the standard deviation and x is the original encoded value of each observation of a variable.

3.1.3 Model Development

The predictive models for all six machine learning techniques were developed using high-level programming language MATLAB. The input features used for model development are shown in Table 3-1. Fourteen features were selected as an input. Only the features which can be easily obtained within no time of a crash are used as an input. The five base models used in this study are K Nearest Neighbor (KNN), Decision Tree

(DT), Adaptive Boosting (AdaBoost), Support Vector Machine (SVM) and Artificial Neural Network (ANN). The sixth model is a two-layer ensemble model developed by stacking the base models. The input crash data were randomly separated into training and testing data sets with a percentage of 70% and 30% respectively.

Base Models

The parameters were carefully set in all the machine learning algorithms to achieve accurate predictive results. Two important features to be identified in KNN are k value and distance function. K value was determined through trial and error procedure. Value of k between 1 and 100 was tried and the predictive performance was checked against every k value. The k value of 25 was finally set in our study while Euclidean distance, which is most commonly used, was selected as a distance function. In the DT model, the parameters are chosen upon consideration of the type of data, sample size and critical interest in the fatal crash. In DT the initially selected feature was decided based on the information gain. The split with maximum information gain was selected as the first split, and this process continued until the information gain approached zero. To overcome the problem of overfitting pruning was carried out by removing the splits with little information gain. In the AdaBoost model, the parameter which needs to be optimized is the number of weak learners/trees to train. A tradeoff is required between model accuracy and computational time during parameter tuning. After performing several experimental tests and looking into extensive literature the number of trees was selected to be 1500, which returned the best testing accuracy.

In the SVM model, two hyperparameters which need to be optimized are penalty factor C and Gamma (γ). C controls the cost of misclassification on training data, while

the γ parameter characterizes how far the impact of single training example extends. A systematic trial and error procedure were followed for determining the values of C and γ in SVM. Finally, the values of C and γ were set to be 130 and 5 respectively. Along with these two parameters kernel function also affects the classification accuracy. Several kernel functions as suggested by many researchers [54] were tried in this study while the most commonly used radial basis function (RBF) was finally chosen based on better results. Eventually, a Artificial Neural Network (ANN) was developed for predicting crash severity. Classification by ANN is an iterative process for adjusting weights and bias based on the provided information. ANN architecture consists of three types of layers namely; the input layer, hidden layers, and output layer. All these layers consist of neurons that are interconnected with the subsequent layer. The input layer in this study consists of 15 neurons where each neuron represents each of the input variables as shown in Table 3-1, while the output layer consists of one neuron which represents the target variable. An iterative searching procedure was followed for setting the number of hidden layers and the number of neurons in each hidden layer until an optimized model was obtained. Several iterations resulted in two hidden layers consisting of 5 and 2 numbers of neurons respectively. In ANN, the Levenberg-Marquardt (LM) training algorithm was selected after testing for several training algorithms as this provided the best predictive accuracy. After setting the parameters for the above machine learning techniques the models were tested on the testing data set and the performance of each model was evaluated based on the confusion matrix (CM).

Two-Layer Ensemble Model

After developing the base models, a stacking framework was applied to integrate five base machine learning models, KNN, DT, AdaBoost, SVM, and ANN for crash severity prediction. Stacking is an ensemble modeling technique for combining multiple models using a meta-classifier [58]. This framework consisted of two layers as illustrated in (Figure 3-2). The four base models: KNN, DT, AdaBoost and SVM were trained in the first layer. In the second layer, the meta-classifier which is ANN in this study was modeled based on the outputs of four base models from the first layer. But using only the outputs of the first layer as an input for layer two, the model was unable to learn properly. To overcome this problem, we came up with the idea to use some features from Table 3-1 as an input along with the output of layer one. Our idea was to simplify the model by using the reduced number of input features in layer two than what we used for the base models. So, to minimize the number of features we performed Principal component analysis (PCA) using the SPSS statistical package.

PCA is a mathematical technique employed for the transformation of high dimensional data set into low dimensional orthogonal feature space. PCA works by transforming several highly correlated features into a smaller number of uncorrelated features while the maximum variance of the dataset is retained. This small number of uncorrelated features is called principal components. The maximum variance in the data is explained by the first principal component followed by each following component explaining the next maximum possible variance. PCA works on the principle of a mathematical technique called Eigen analysis where we solve for the eigenvalue and eigenvector of a square matrix called the covariance matrix. In our analysis the average

eigen value criterion also known as eigenvalue-one rule was followed for selecting the principal components. According to this rule, only the Principal components having eigen value greater than 1.0 are selected [59]. The six principal components had eigenvalue greater than 1.0 as shown by a Scree plot in Figure 3-3 and Table 3-3. These six principal components explain 64.7% of the variance in the data. Extracted principal components can be interpreted based on component loadings. Component loadings represent the correlation between original features and the principal components. For ease of interpreting six principal components, varimax rotation was conducted. The component loading obtained from varimax rotation is shown in Table 3-4. Only the loadings with an absolute value of more than 0.3 are tabulated for ease of interpretation of the principal components [60]. For more information about PCA reader is directed to Rencher [61]. PCA of the entire data set having 14 features resulted in six principal components (Table 3-4) based on eigenvalue-one rule. These six principal components along with the output of layer one was used as an input in layer two of our ensemble model.

The input features in layer two were trained using Artificial Neural Network (ANN). ANN architecture consists of three types of layers namely; the input layer, hidden layers, and output layer. All these layers consist of neurons that are interconnected with the subsequent layer. The input layer of ANN consists of 10 neurons where each neuron represents each of the input features, while the output layer consists of one neuron which represents the target variable. An iterative searching procedure was followed for setting the number of hidden layers and the number of neurons in each hidden layer until an optimized model was obtained. Several iterations resulted in two hidden layers consisting of 10 and 2 numbers of neurons respectively. In ANN, the Levenberg-

Marquardt (LM) training algorithm was selected after testing for several training algorithms as this provided the best predictive accuracy. After setting these parameters the two-layer ensemble model was tested on the test set and results were evaluated based on the confusion matrix.

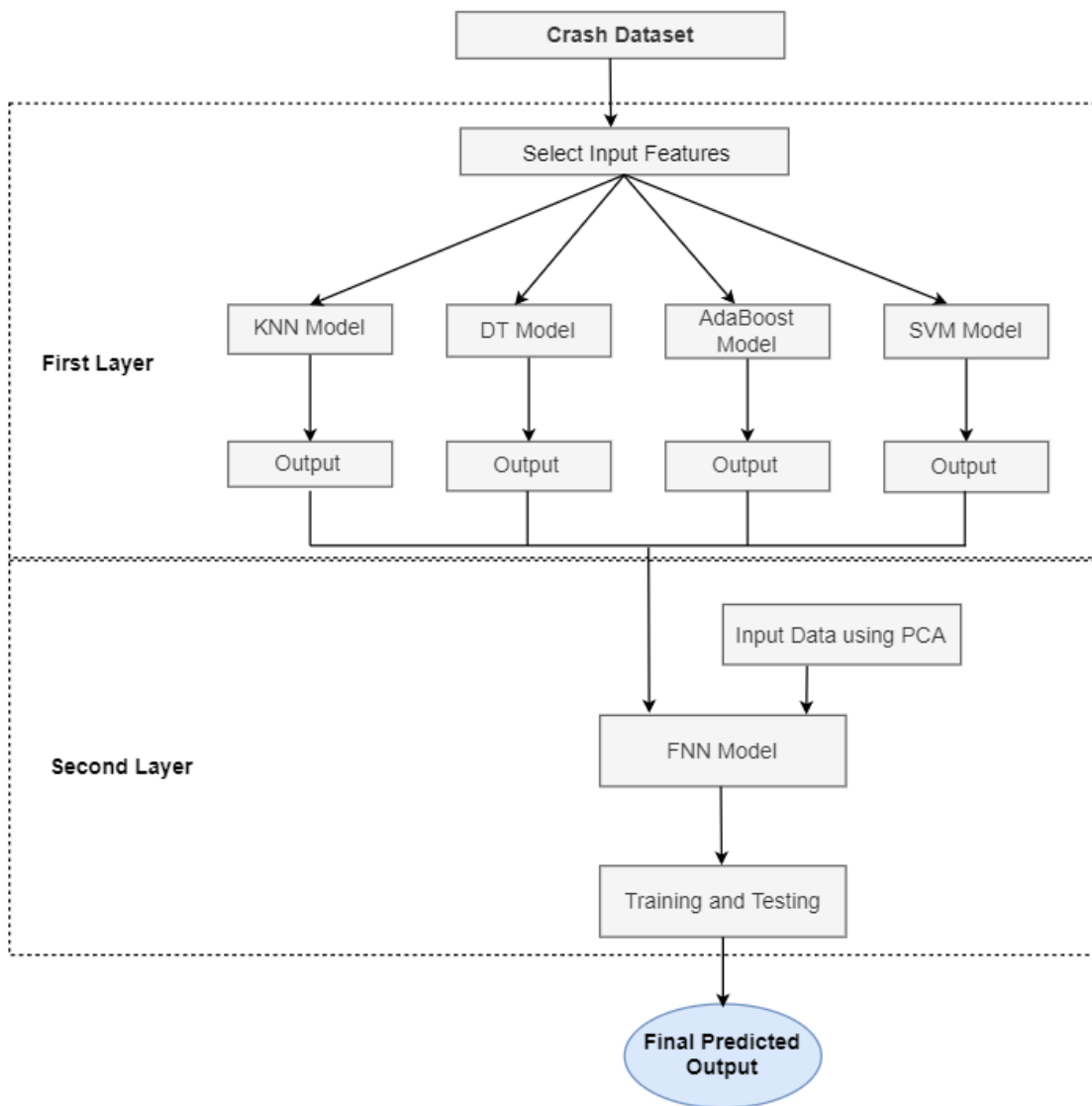


Figure 3-2 : Structure for two-layer Ensemble Model

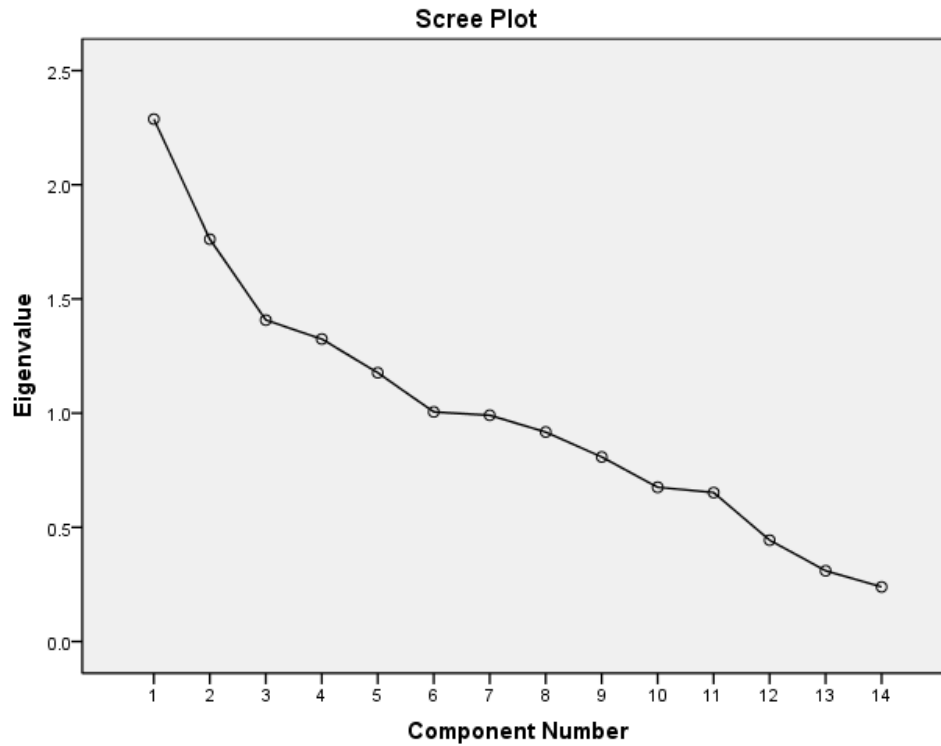


Figure 3-3 : Scree plot for Principal Components

Table 3-3 : Percent of variance explained by each component

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	2.288	16.341	16.341	2.288	16.341	16.341	1.942	13.873	13.873
2	1.761	12.582	28.923	1.761	12.582	28.923	1.798	12.846	26.719
3	1.408	10.054	38.977	1.408	10.054	38.977	1.453	10.381	37.100
4	1.325	9.461	48.438	1.325	9.461	48.438	1.428	10.198	47.298
5	1.177	8.406	56.843	1.177	8.406	56.843	1.330	9.497	56.795
6	1.005	7.181	64.025	1.005	7.181	64.025	1.012	7.229	64.025
7	.991	7.077	71.102						
8	.917	6.552	77.654						
9	.808	5.772	83.426						
10	.675	4.824	88.250						
11	.652	4.659	92.909						
12	.444	3.171	96.081						
13	.310	2.212	98.292						
14	.239	1.708	100.000						

Table 3-4 : Rotated Component loadings for PCA of crash data

Variables	Components					
	1	2	3	4	5	6
Number of Vehicles	.919					
Vehicle2 Type	.739					
Vehicle3 Type	.686					
Speed limit		.882				
Urban or Rural Area		.879				
Road Type			.830			
Junction Detail			.681		-.533	
1st Road Class		-.388	.481		.466	
Road Surface Conditions				.794		
Weather Conditions				.712		
Light Conditions				.526		
Junction Control					.861	
Day of Week						.732
Vehicle1 Type						-.643

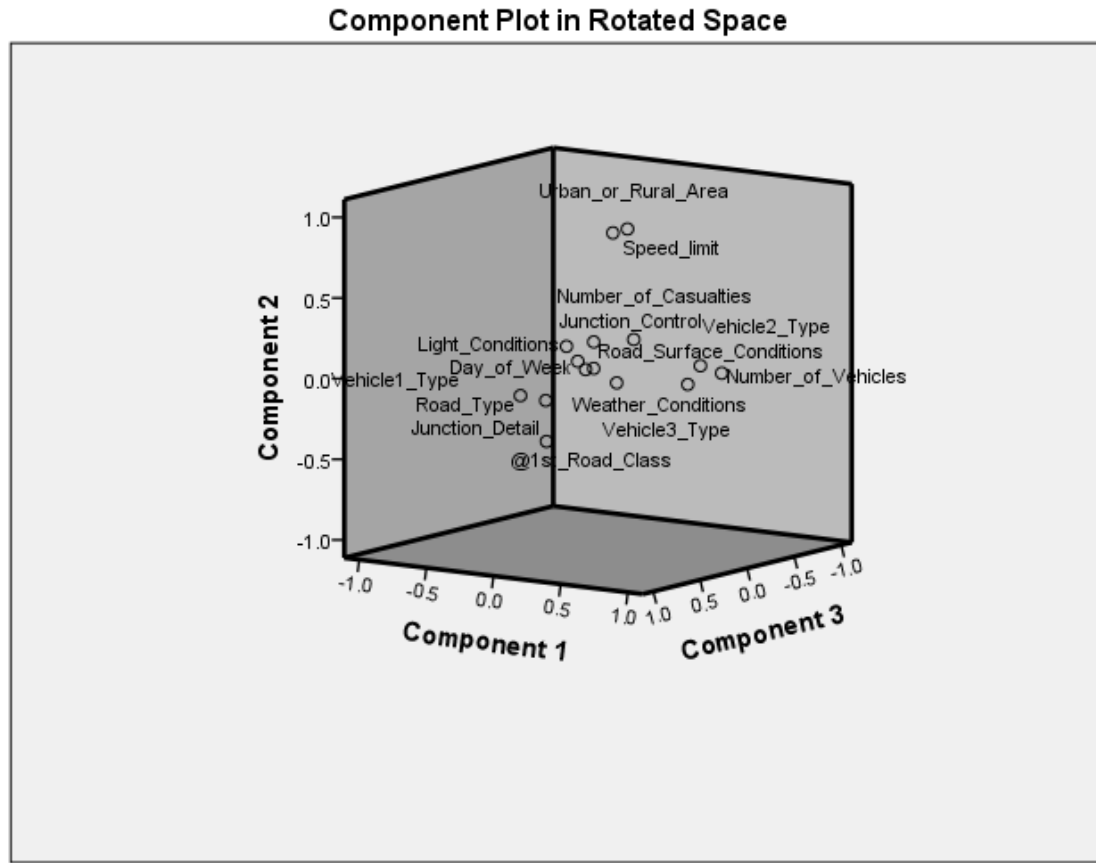


Figure 3-4 : Component Plot in rotated space

3.2 Canada's Crash Dataset

In order to check the validity of models which were originally developed using the data set for Great Britain, a data set from different region was used to ensure that the models work reasonably well for other area.

3.2.1 Data Description and Preparation

Traffic crash data for the 3-year period (2014-2016) from the National Collision Database (NCDB) Canada² was utilized to meet our objective of model's transferability. Data cleaning was carried out by deleting any duplicate, irrelevant and incomplete data fields.

² <https://open.canada.ca/data/en/dataset/1eb9eba7-71d1-4b30-9fb1-30cbdab7e63a>

Data cleaning resulted in 581,923 crash records. Same procedure as Great Britain's dataset was followed for selecting the input features. Twelve child features were selected as input from the extracted data set, where every child feature belongs to one of the six parent features. The six parent features are (1) crash feature, (2) roadway feature, (3) environmental feature and (4) vehicle feature (5) area feature and (6) person feature as shown in (Table 3-5). A random sample of 10,000 crashes was extracted from filtered data set using a randomized class balancing procedure, so each severity class had an equal number of crashes [57]. In this dataset the severity class is categorized as fatal or non-fatal, which is different from the Great Britain's dataset. This technique eliminated any possibility of bias towards a specific severity class. Fatal crash is defined as any crash resulting in at least one fatality. To fulfill the second objective of this research, the same data was filtered to obtain the data set for road intersections only. Same input features were used for modeling the traffic crashes at road intersections only for a sample set of 5000 crash points. The results of models for intersections of Great Britain were compared to the results of intersections for Canadian dataset. For data standardization we used the same procedure as Great Britain.

Table 3-5 : Description of Input Features for Canada

Parent Feature	Child Feature	Type	Description
Crash Feature	Month	Qualitative	January=1, February=2, March=3, April=4, May=5, June=6, July=7, August=8, September=9, October=10, November=11, December=12
	Vehicles involved	Continuous	As indicated
	Day of the week	Qualitative	Sunday=1, Monday=2, Tuesday=3, Wednesday=4, Thursday=5, Friday=6, Saturday=7
	Crash Configuration	Qualitative	Hit a moving object=1, Hit a stationary object=2, Ran off left shoulder=3, Ran off right shoulder=4, Rollover on roadway=5, Rear-end collision=6, Side swipe=7, left turn conflict=8, right turn conflict=9, Head-on collision=10, Approaching side-swipe=11, Left turn across opposing traffic=12, Right angle collision=13, Hit a parked motor vehicle=14
	Crash hour	Qualitative	Indicated time
Roadway Feature	Roadway configuration	Qualitative	Non-intersection=1, intersection of two public roadways=2, private driveway or laneway=3, Railroad level crossing=4, Bridge, overpass, viaduct=5, Tunnel or underpass=6, Passing or climbing lane=7, Ramp=8, Traffic circle=9, Express lane of a freeway system=10, Collector lane of a freeway system=11, Transfer lane of a freeway system=12
	Road Surface	Qualitative	Dry, =1, Wet=2, Snow=3, Slush=4, Icy=5, Debris=6, Muddy=7, Oil=8, Flooded=9
	Road Alignment	Qualitative	Straight and level=1, Straight with gradient=2, Curved and level=3, Curved with gradient=4, Top of hill or gradient=5, Bottom of hill or gradient=6
Environmental Feature	Weather Condition	Qualitative	Sunny=1, cloudy=2, Rain=3, Snow=4, hail=5, fog=6, Strong wind=7
Vehicle feature	Vehicle Type	Qualitative	Light Duty Vehicle=1, Panel/cargo van=2, Other trucks and vans=3, Unit trucks > 4536 KG=4, Road tractor=5, School bus=6, Smaller school bus=7, Urban and Intercity Bus=8, Motorcycle and moped=9, Off road vehicles=10, Bicycle=11, Purpose-built motorhome=12, Farm equipment=13, Construction equipment=14, Fire engine=15, Snowmobile=16, Street car=17
Area Feature	Junction Control	Qualitative	Traffic signals fully operational=1, Traffic signals in flashing mode=2, Stop sign=3, Yield sign=4, Warning sign=5, Pedestrian crosswalk=6, Police officer=7, School guard=8, School crossing=9, Reduced speed zone=10, No passing zone sign=11, Railway crossing with signals=12, Railway crossing with signs only=13
	Road User Class	Qualitative	Motor Vehicle Driver=1, Motor Vehicle Passenger=2, Pedestrian=3, Bicyclist=4, Motorcyclist=5

3.2.2 Model Development

Like Great Britain's crash dataset, the predictive models for all six machine learning techniques were developed using high-level programming language MATLAB. The input features used for model development are shown in (Table 3-5). Twelve features were selected as an input. Only the features which can be easily obtained within no time of a crash are used as an input. The five base models used in this study are K Nearest Neighbor (KNN), Decision Tree (DT), Adaptive Boosting (AdaBoost), Support Vector Machine (SVM) and Artificial Neural Network (ANN). The sixth model is a two-layer ensemble model developed by stacking the base models. The input crash data were randomly separated into training and testing data sets with a percentage of 70% and 30% respectively.

Base Models

The parameters were carefully set in all the machine learning algorithms to achieve accurate predictive results. Two important features to be identified in KNN are k value and distance function. K value was determined through trial and error procedure. Value of k between 1 and 100 was tried and the predictive performance was checked against every k value. The k value of 10 was finally set in our study while Euclidean distance, which is most used, was selected as a distance function. In the DT model just like the Great Britain's data, the parameters are chosen upon consideration of the type of data, sample size and critical interest in the fatal crash. In DT the initially selected feature was decided based on the information gain. The split with maximum information gain was selected as the first split, and this process continued until the information gain approached zero. To overcome the problem of overfitting pruning was carried out by

removing the splits with little information gain. In the AdaBoost model, the parameter which needs to be optimized is the number of weak learners/trees to train. A tradeoff is required between model accuracy and computational time during parameter tuning. After performing several experimental tests and looking into extensive literature the number of trees was selected to be 1000, which returned the best testing accuracy.

In the SVM model, two hyperparameters which need to be optimized are penalty factor C and Gamma (γ). C controls the cost of misclassification on training data, while the γ parameter characterizes how far the impact of single training example extends. A systematic trial and error procedure were followed for determining the values of C and γ in SVM. Finally, the values of C and γ were set to be 100 and 5 respectively. Along with these two parameters kernel function also affects the classification accuracy. Several kernel functions as suggested by many researchers [54] were tried in this study while the most commonly used radial basis function (RBF) was finally chosen based on better results. Eventually, a Artificial Neural Network (ANN) was developed for predicting crash severity. Classification by ANN is an iterative process for adjusting weights and bias based on the provided information. ANN architecture consists of three types of layers namely; the input layer, hidden layers, and output layer. All these layers consist of neurons that are interconnected with the subsequent layer. The input layer in this study consists of 12 neurons where each neuron represents each of the input variables as shown in Table 3-5, while the output layer consists of one neuron which represents the target variable. An iterative searching procedure was followed for setting the number of hidden layers and the number of neurons in each hidden layer until an optimized model was obtained. Several iterations resulted in two hidden layers consisting of 10 and 2 numbers

of neurons respectively. In ANN, the Levenberg-Marquardt (LM) training algorithm was selected after testing for several training algorithms as this provided the best predictive accuracy. After setting the parameters for the above machine learning techniques the models were tested on the testing data set and the performance of each model was evaluated based on the confusion matrix (CM).

Two-Layer Ensemble Model

Again, the two-layer ensemble model was employed using the base models in the same order as for the previous case. The four base models: KNN, DT, RF and SVM were trained in the first layer. In the second layer, the meta-classifier which is ANN in this study was modeled based on the outputs of four base models from the first layer. But using only the outputs of the first layer as an input for layer two, the model was unable to learn properly. We tried to simplify the model by using PCA. SPSS statistical package was used for PCA. The detailed explanation of PCA has been provided in the preceding section of this chapter.

In our analysis, the eigenvalues of first six principal components were greater than 1.0 explaining 64.2% of the variance in the data as shown by a Scree plot in Figure 3-6 and Table 3-6. Extracted principal components can be interpreted based on component loadings. For ease of interpreting six principal components, varimax rotation was conducted. The component loading obtained from varimax rotation is shown in Table 3-7. Only the loadings with an absolute value of more than 0.3 are tabulated for ease of interpretation of the principal components [60]. PCA of the entire data set having 12 features resulted in six principal components (Table 3-7). The first six principal

components were selected to be used in layer two of ensemble model along with the output from layer 1.

The input features in layer two were trained using Artificial Neural Network (ANN). The input layer of ANN consists of 10 neurons where each neuron represents each of the input features, while the output layer consists of one neuron which represents the target variable. An iterative searching procedure was followed for setting the number of hidden layers and the number of neurons in each hidden layer until an optimized model was obtained. Several iterations resulted in two hidden layers consisting of 10 and 2 numbers of neurons respectively. In ANN, the Levenberg-Marquardt (LM) training algorithm was selected after testing for several training algorithms as this provided the best predictive accuracy. After setting these parameters the two-layer ensemble model was tested on the test set and results were evaluated based on the confusion matrix.

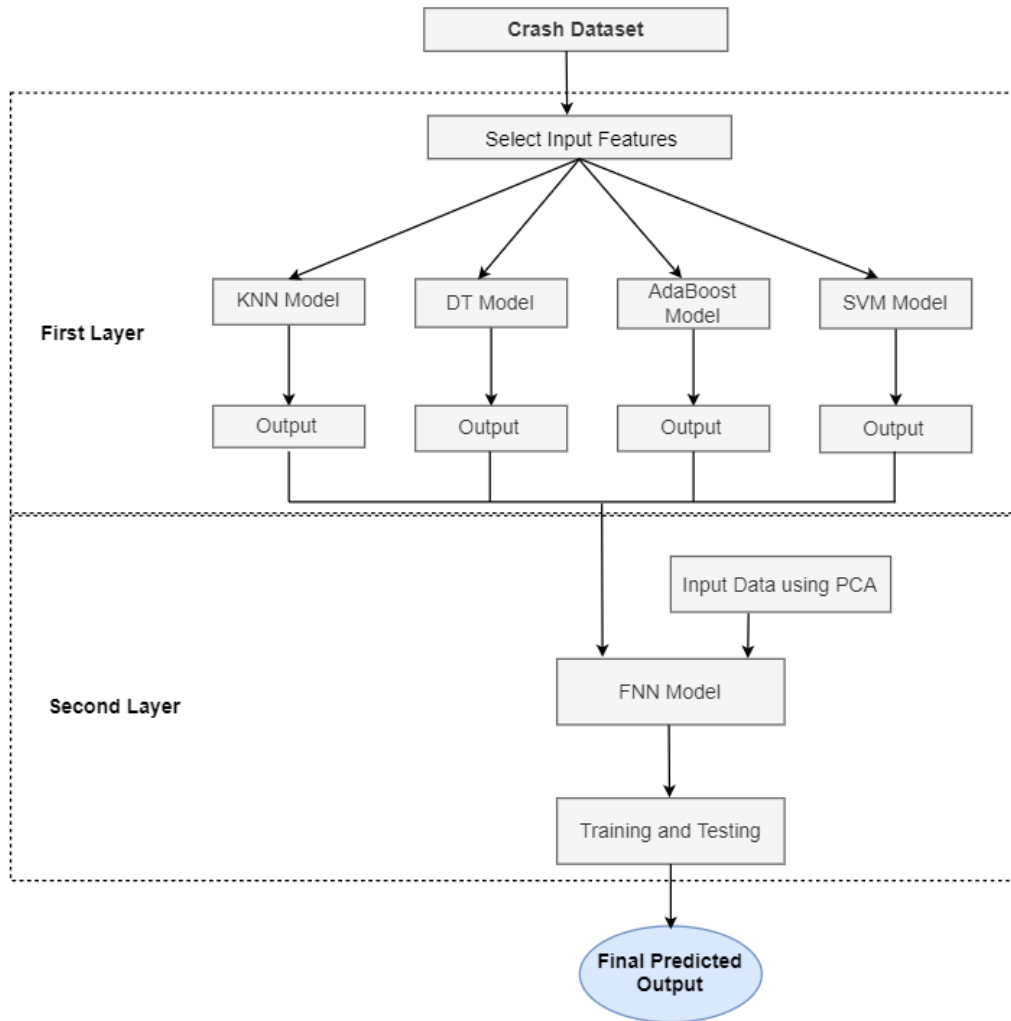


Figure 3-5 : Structure for two-layer Ensemble Model for Canada Dataset

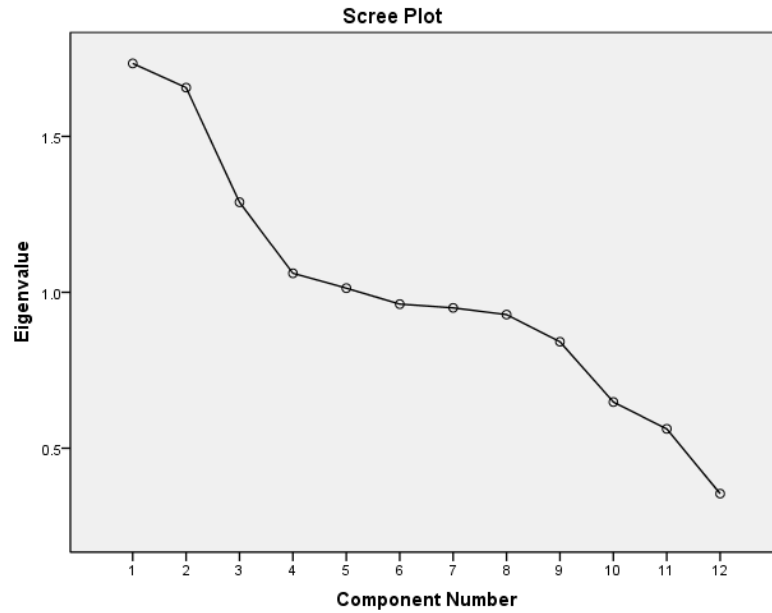


Figure 3-6 : Scree plot for Principal Components

Table 3-6 : Percent of variance explained by each component

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	1.734	14.449	14.449	1.734	14.449	14.449	1.640	13.665	13.665
2	1.656	13.803	28.252	1.656	13.803	28.252	1.527	12.725	26.390
3	1.289	10.740	38.992	1.289	10.740	38.992	1.393	11.607	37.998
4	1.061	8.842	47.834	1.061	8.842	47.834	1.108	9.231	47.229
5	1.013	8.444	56.278	1.013	8.444	56.278	1.039	8.658	55.887
6	1.004	8.018	64.296	1.004	8.018	64.296	1.009	8.409	64.296
7	.950	7.918	72.214						
8	.928	7.737	79.952						
9	.841	7.011	86.963						
10	.648	5.402	92.364						
11	.562	4.683	97.048						
12	.354	2.952	100.000						

Table 3-7 : Rotated Component loadings for PCA of crash data

Variables	Component					
	1	2	3	4	5	6
Month		-.455	.363	.319		
Day of_ week						.919
Collision Hour				.505		
Vehicles involved				.761		
Collision configuration			-.604	.382		
Roadway configuration					.878	
Weather condition		.771				
Road surface		.777				
Road align1ent			.586			
Traffic control			.703			
Vehicle Type	.896					
Road user class	.893					

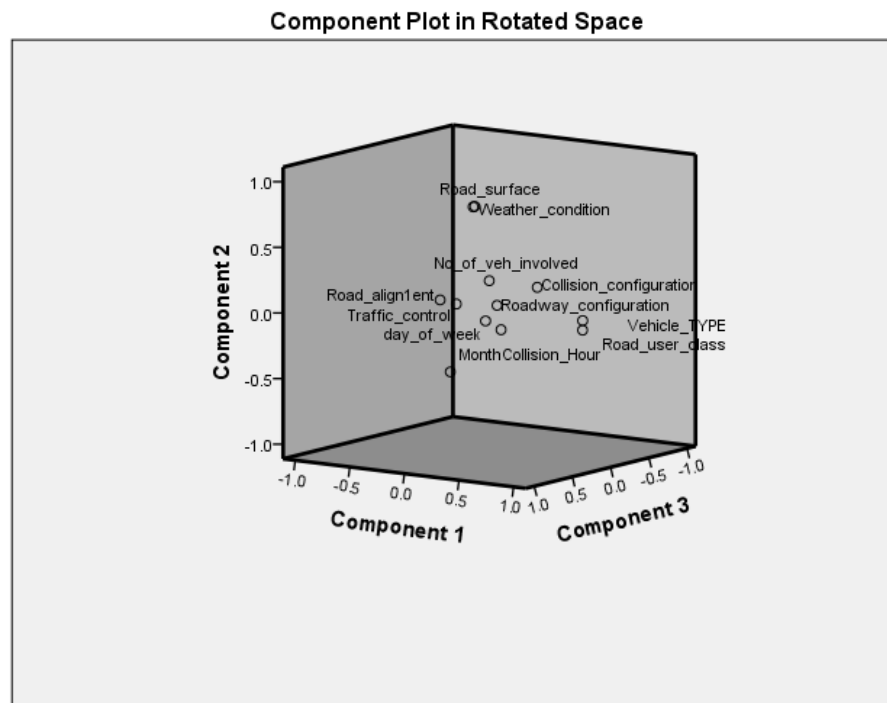


Figure 3-7 : Component Plot in rotated space

CHAPTER 4

RESULTS AND ANALYSIS

In this chapter the results are discussed for the datasets of Great Britain and Canada. For Great Britain the results of the models which are developed for all locations are presented first. After this the results for the intersections separately are provided. At the end the results of the models developed for Canadian dataset are explained. We noticed that the models perform in the same line for both the datasets.

The proposed two-layer ensemble model is compared with five base machine learning models. The data set is randomly classified into training and testing data set with the percentage of 70% and 30% respectively. The performance of each model was evaluated based on the data generated by the confusion matrix (CM). The CM contains the results of the original and predicted classifications provided by a classification model. General representation of a CM for binary output classes is shown in

Table 4-1, where observations of an actual class are shown in the row while observations of the predicted class are represented in the column.

Table 4-1 : Confusion Matrix

Total Instances		Predicted	
		Negative	Positive
Actual	Negative	True Negative (TN)	False Positive (FP)
	Positive	False Negative (FN)	True Positive (TP)

The entities in CM are defined as follows: True Negative (TN) represents the entities that are originally negative and classified correctly as negative while False Negative (FN) represents entities that are positive but incorrectly classified as negative. Similarly, True positive (TP) represents the entities that are originally positive and classified correctly as positive while False Positive (FP) represents entities that are originally positive but incorrectly classified as positive. The observations of the confusion matrix for every model were used to calculate the following performance measures (PM) and evaluate model performance based on these metrics.

- *Accuracy*: It is the proportion of the total number of instances that were classified correctly, shown by the following equation.

$$Accuracy (AC) = \frac{(TP+TN)}{(TP+TN+FP+FN)} \quad (4.1)$$

- *Error rate*: The rate of misclassification of predictions is called error rate.

$$Error\ rate\ (ER) = 1 - Accuracy \quad (4.2)$$

- *Recall*: The proportion of positive instances that were classified correctly, shown by the following equation.

$$Recall\ (R) = \frac{TP}{TP+FN} \quad (4.3)$$

- *Precision*: It is the proportion of the anticipated positive cases that were right, shown by the following equation.

$$Precision\ (P) = \frac{TP}{TP+FP} \quad (4.4)$$

- *F1-Measure*: The performance of the model is measured using the F1 measure which represents the harmonic mean of Recall and Precision. Its value ranges between 0 and

1, where 1 represents the best model while zero represents the worst model. The equation for the F1 score is shown below.

$$F = \frac{2*(R*P)}{R+P} \quad (4.5)$$

4.1 Modeling Results for all Locations Great Britain

The classification accuracy of each model is shown in Table 4-2 and Figure 4-1. The overall training accuracy ranges from 71.3% to 82.1% while the overall testing accuracy ranges from 69.1% to 78.0%. DT performs better in training while the testing accuracy declines. The decline in testing accuracy suggests that the model suffers from overfitting, meaning that the model captures the noise of the data. Among the base models, AdaBoost has the highest testing accuracy while KNN performs the worst among all models. The proposed two-layer ensemble model shows excellent performance in training and testing outperforming all the other models with a training accuracy of 82.1% and testing accuracy of 78.0%. The confusion matrices for all the models are shown in Figure 4-2 to Figure 4-13.

4.1.1 Accuracy of Models for all Locations Great Britain

Table 4-2 : Accuracy of different models all locations GB

S. No	Model	Training Accuracy (%)	Testing Accuracy (%)
1	K Nearest Neighborhood (KNN)	71.3	69.1
2	Decision Trees (DT)	79.7	72.3
3	AdaBoost	78.1	73.6
4	Artificial Neural Network (ANN)	74.3	72.2
5	Support Vector Machine (SVM)	73.9	72.1
6	Ensemble Model	82.1	78.0

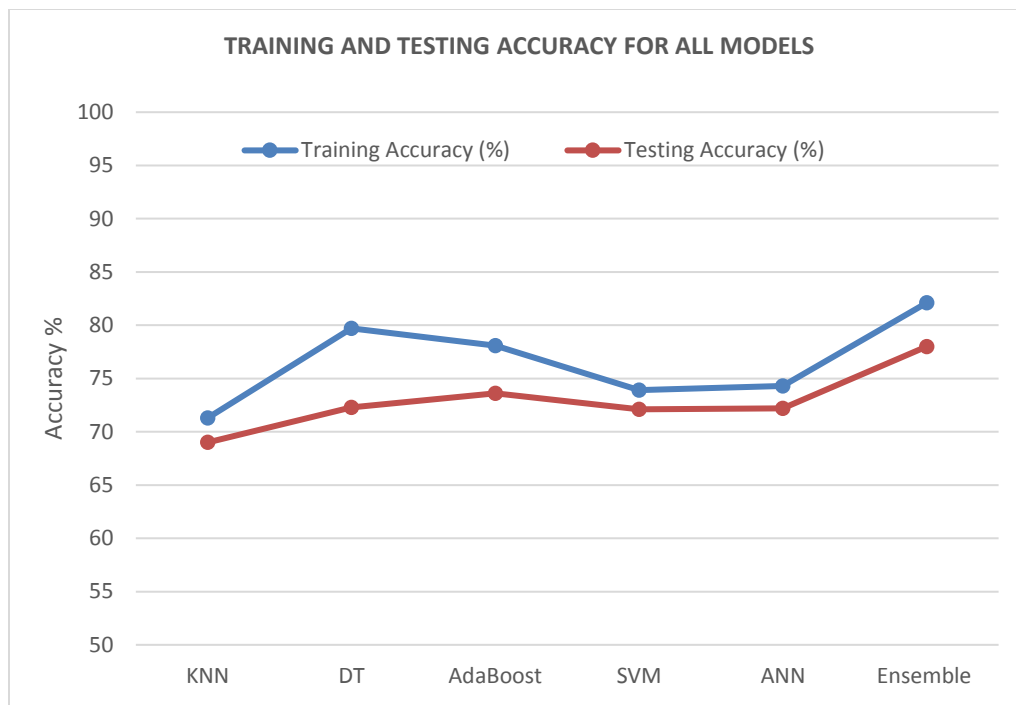


Figure 4-1 : Accuracy of different models all locations GB

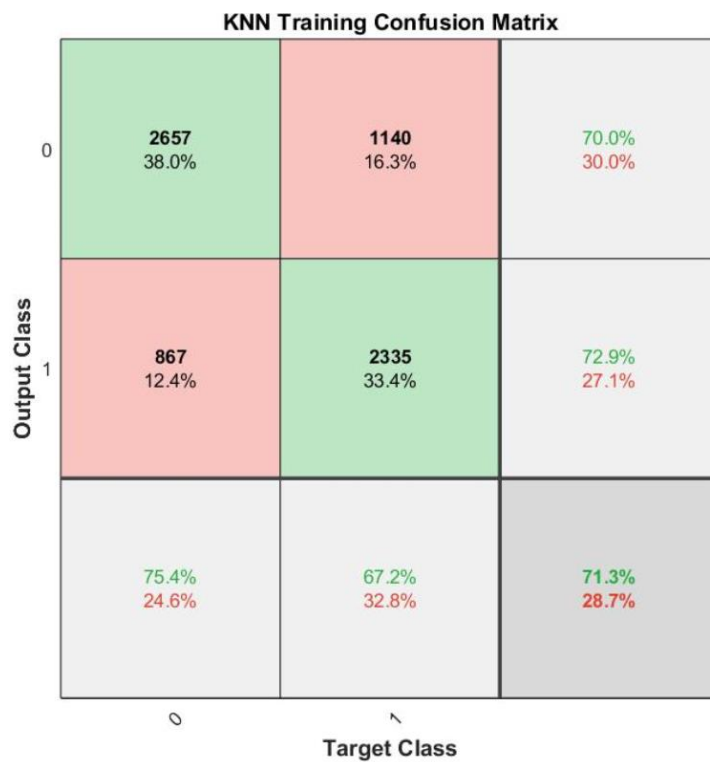


Figure 4-2 : KNN Training CM all locations GB

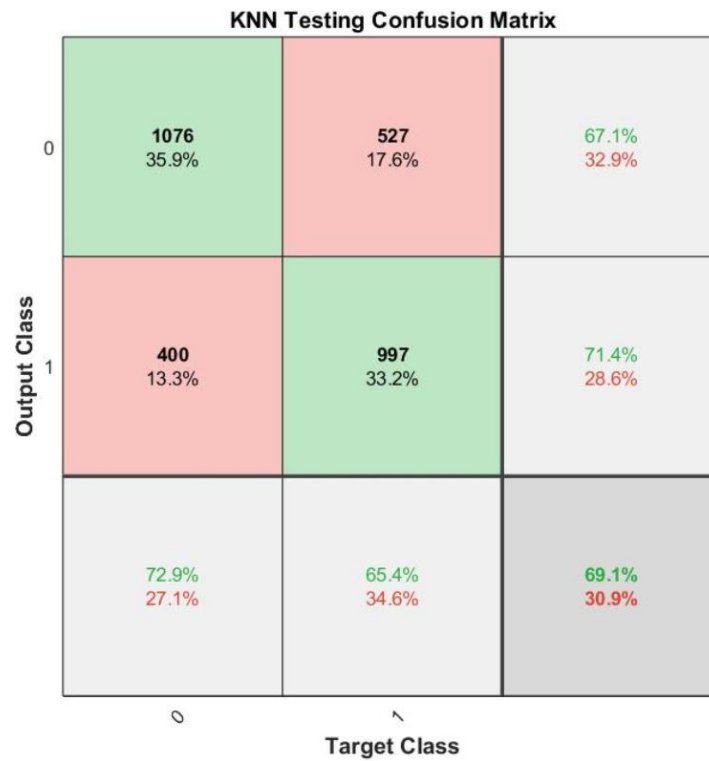


Figure 4-3 : KNN Testing CM all locations GB

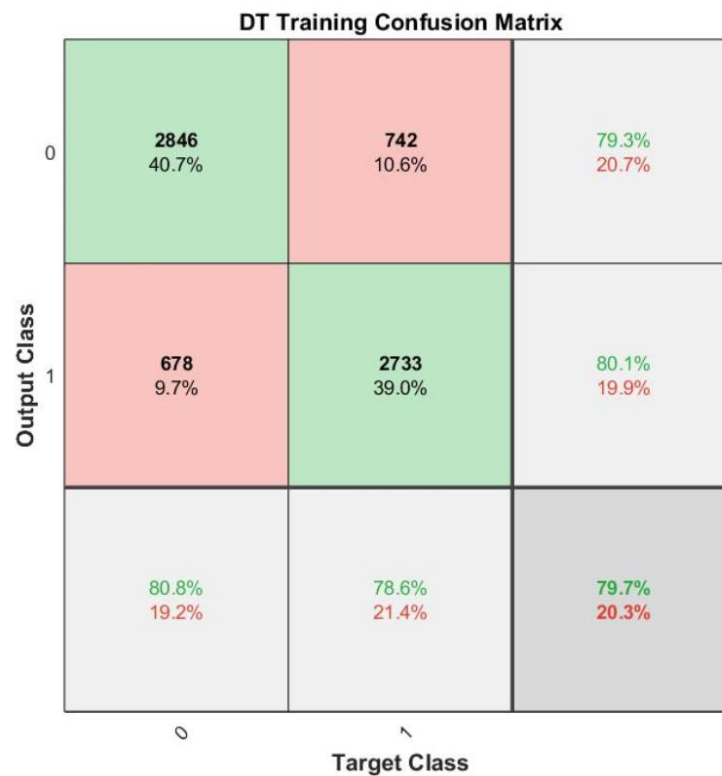


Figure 4-4 : DT Training CM all locations GB

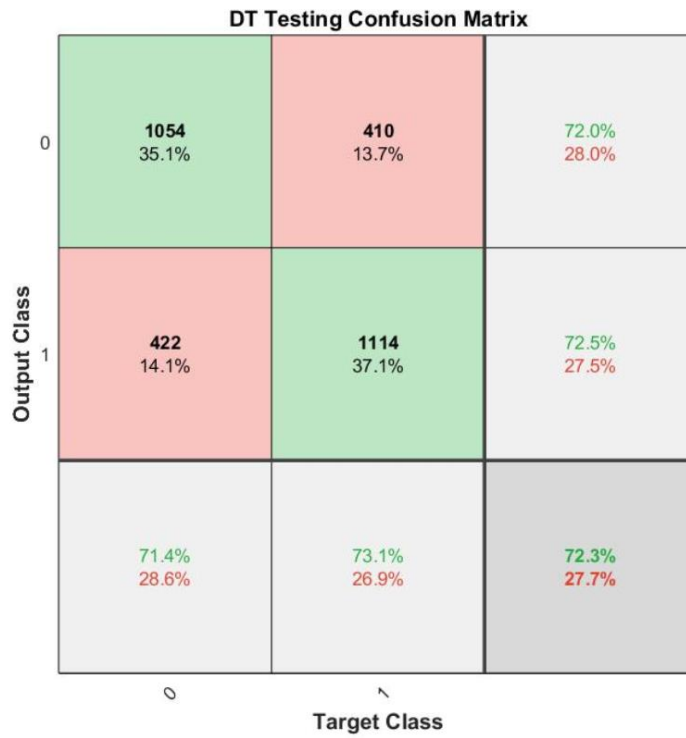


Figure 4-5 : DT Testing CM all locations GB

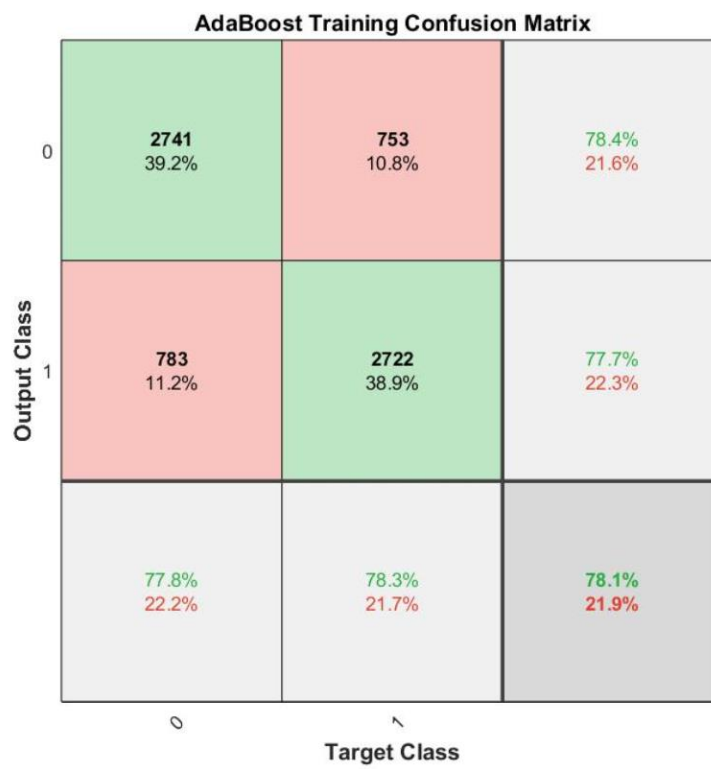


Figure 4-6 : AdaBoost Training CM all locations GB

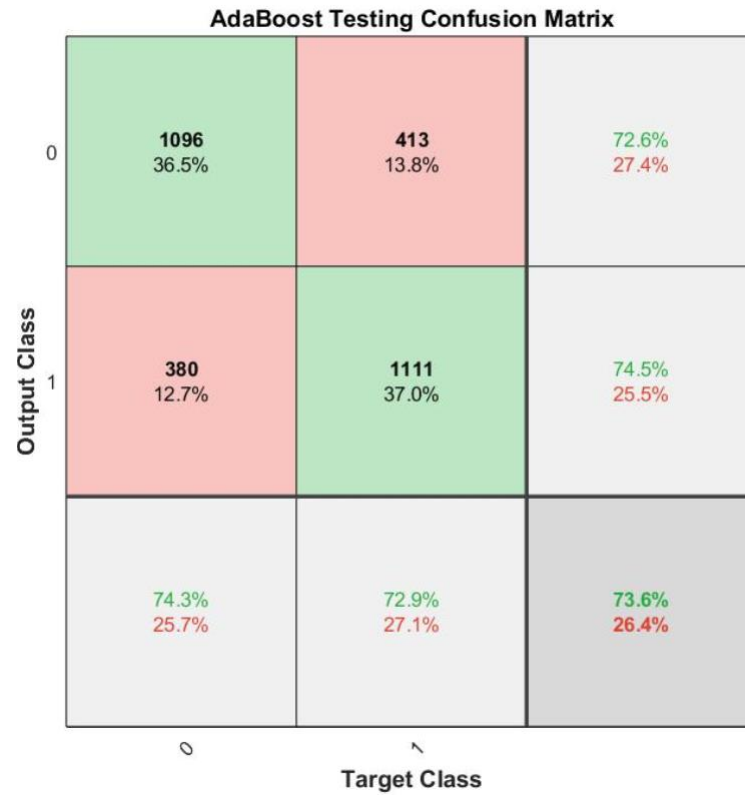


Figure 4-7 : AdaBoost Testing CM all locations GB



Figure 4-8 : ANN Training CM all locations GB



Figure 4-9 : ANN Testing CM all locations GB

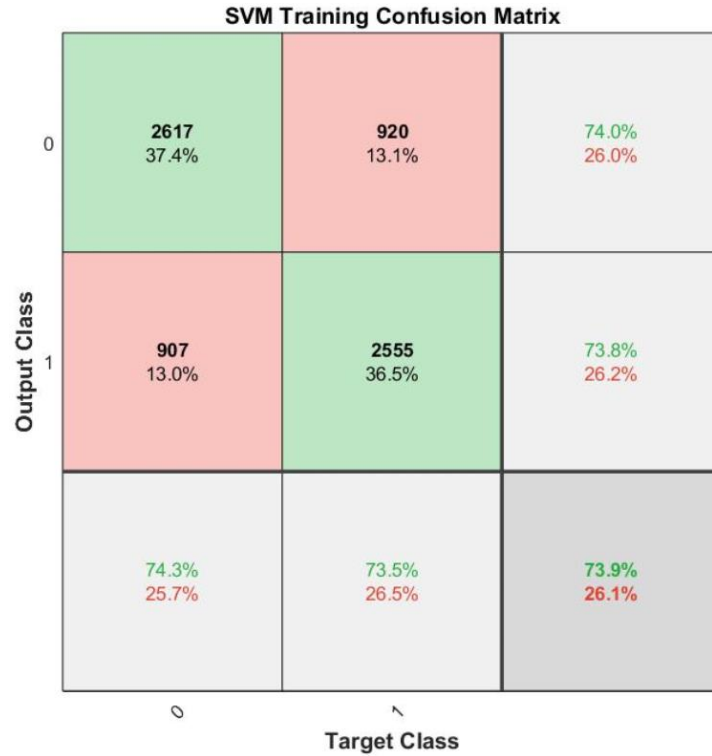


Figure 4-10 : SVM Training CM all locations GB

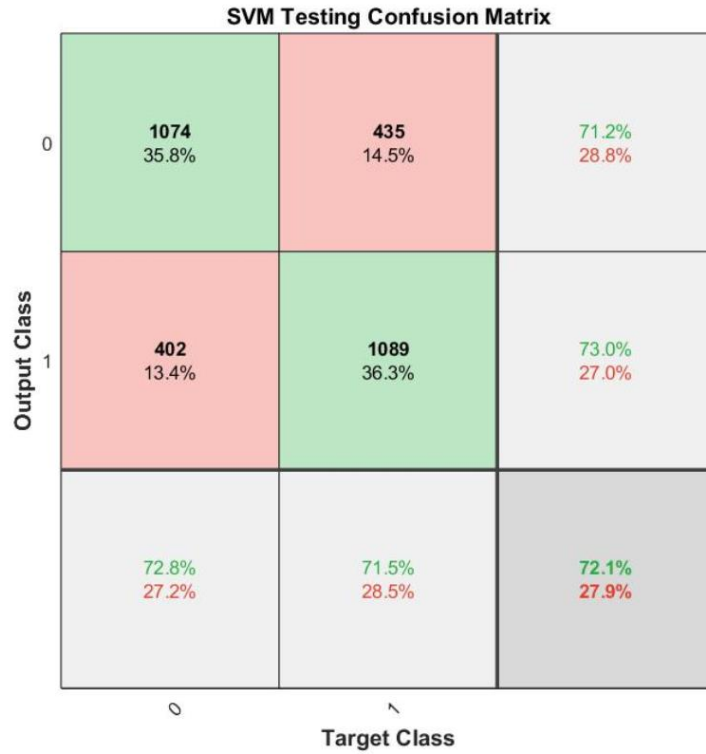


Figure 4-11 : SVM Testing CM all locations GB

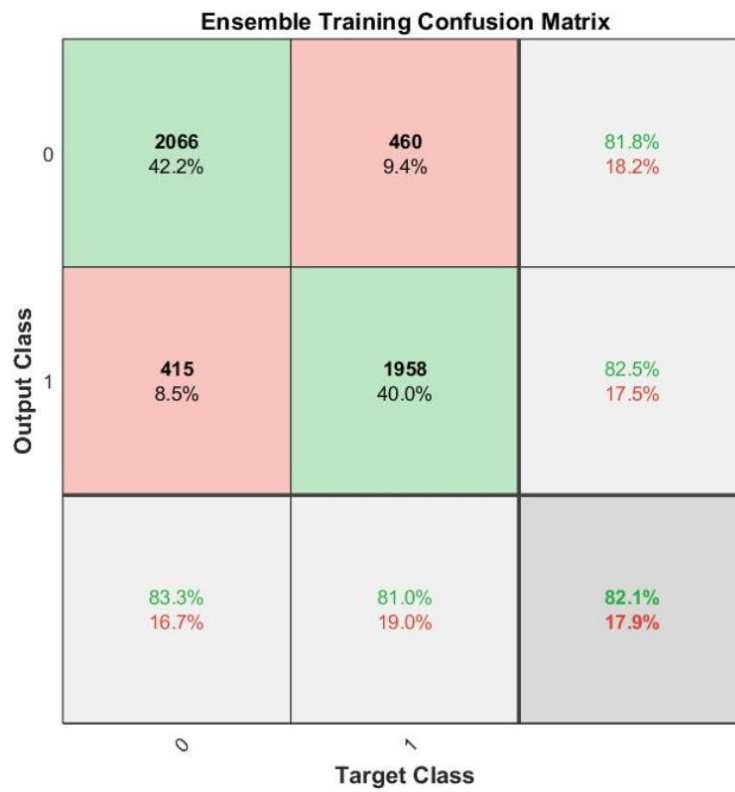


Figure 4-12 : Ensemble Training CM all locations GB

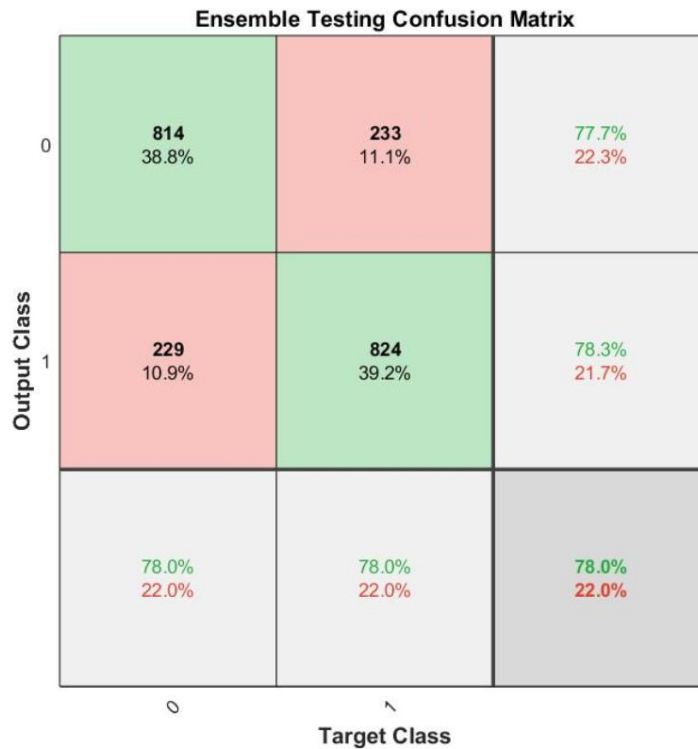


Figure 4-13 : Ensemble Testing CM all locations GB

4.1.2 Performance Measures for all Locations Great Britain

Although accuracy is one of the metrics representing the performance of individual models but relying only on accuracy as a performance measure can be misleading as the model might be biased towards one severity class. So, to overcome these limitations we determined other performance measures like precision, recall and F1 score discussed in detail in this section. These performance measures determine the performance of individual severity levels providing better insight into the performance of models. Results of these performance measures for severe and non-severe crashes are illustrated in Table 4-3 and Table 4-4 and respectively. Figure 4-14 and Figure 4-15 also shed light on the performance of models for severe and non-severe crashes.

Table 4-3 : Performance measures for Severe Crashes all locations GB

Model	Precision	Recall	F1 Score
KNN	0.6542	0.7137	0.6826
DT	0.7060	0.7305	0.7181
AdaBoost	0.7290	0.7451	0.7370
SVM	0.7146	0.7304	0.7224
ANN	0.7329	0.7239	0.7284
Ensemble	0.7796	0.7825	0.7810

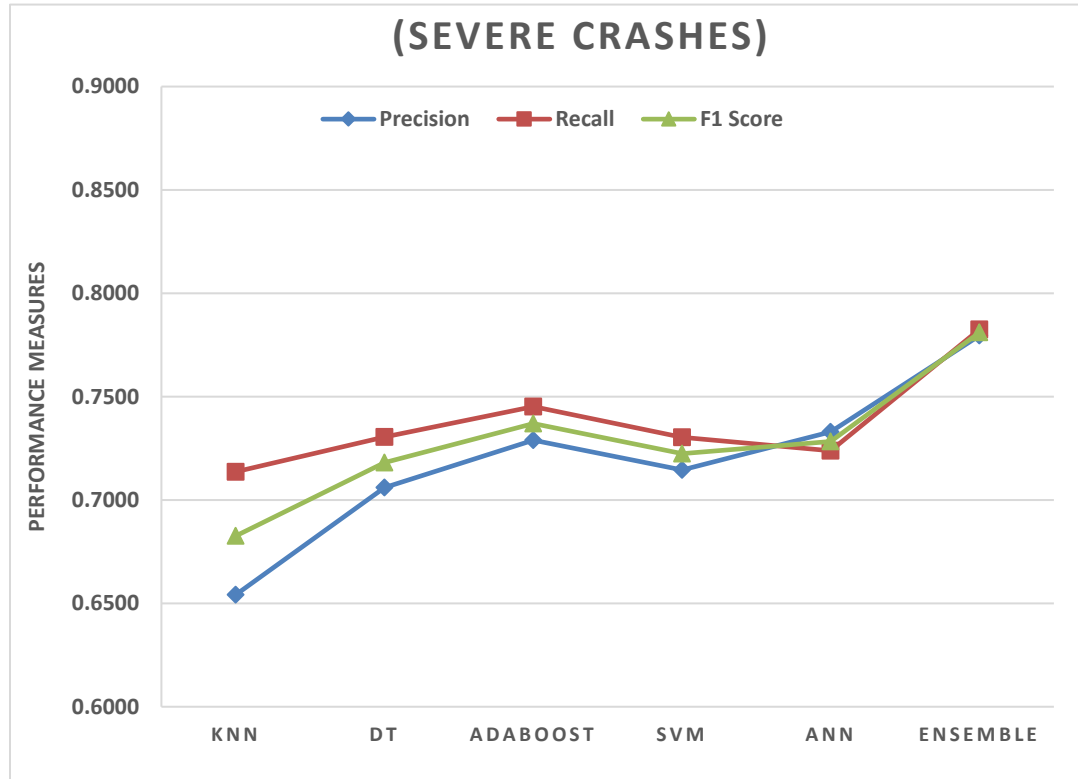
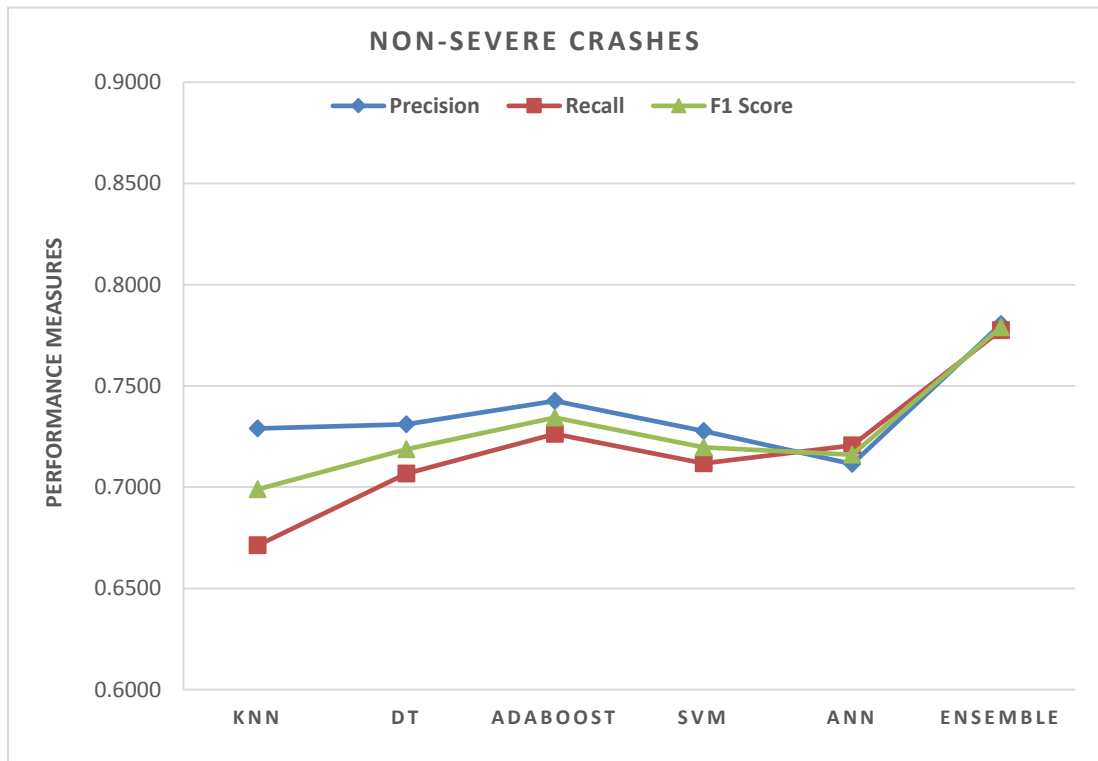


Figure 4-14 : Performance measures of Models for Severe Crashes all locations GB

Table 4-4 : Performance measures for Non-Severe Crashes all locations GB

Model	Precision	Recall	F1 Score
KNN	0.7290	0.6712	0.6989
DT	0.7310	0.7066	0.7186
AdaBoost	0.7425	0.7263	0.7343
SVM	0.7276	0.7117	0.7196
ANN	0.7114	0.7207	0.7160
Ensemble	0.7804	0.7775	0.7789

**Figure 4-15 : Performance measures for Non-Severe Crashes all locations GB**

According to the definition of precision and recall, it is stated that any model which maximizes both performance measures is the best one. F1 score acts as a good performance indicator since it is the harmonic mean of precision and recall. In this study, we notice that each model performs almost equally good for each severity level as

evident from the results of performance measures in Table 4-3 and Table 4-4. Among the base models, the KNN model has the least F1 score while AdaBoost has the maximum score for both the severity levels respectively. Among the individual models, AdaBoost outperforms all other models in terms of accuracy and F1 score while KNN performs the worst among all the models. But there is an abrupt enhancement in the predictive performance with the introduction of the two-layer ensemble model. The test accuracy of the two-layer ensemble model increased to 78.0% while the second-best model which is AdaBoost has a test accuracy of 73.6%. Similarly, the F1 score has improved tremendously for both levels of severity. This enhanced performance of two-layer ensemble model indicates that it is a viable option to be employed for predicting traffic crash severity

4.2 Modeling Results for Road Intersections Great Britain

The training and testing accuracy for all the models are shown in Table 4-5 and Figure 4-16. The overall training accuracy ranges from 67.9% to 81.6% while the overall testing accuracy ranges from 67.1% to 76.7%. DT performs better in training while the testing performance decline to suggest that they suffer from the issue of overfitting. Among the base models, AdaBoost has the highest testing accuracy while KNN performs the worst among all models. The proposed two-layer ensemble model shows excellent performance in training and testing outperforming all the other models with a training accuracy of 81.6% and testing accuracy of 76.7%. The confusion matrices for all the models are shown in Figure 4-17 to Figure 4-28.

4.2.1 Accuracy of Models Road Intersections Great Britain

Table 4-5 : Accuracy of all Models intersections GB

Model	Training Accuracy (%)	Testing Accuracy (%)
K Nearest Neighborhood (KNN)	67.9	67.1
Decision Trees (DT)	79.4	69.2
AdaBoost	75.6	71.4
Artificial Neural Network (ANN)	71.7	69.7
Support Vector Machine (SVM)	71.4	68.8
Two-layer Ensemble Model	81.6	76.7

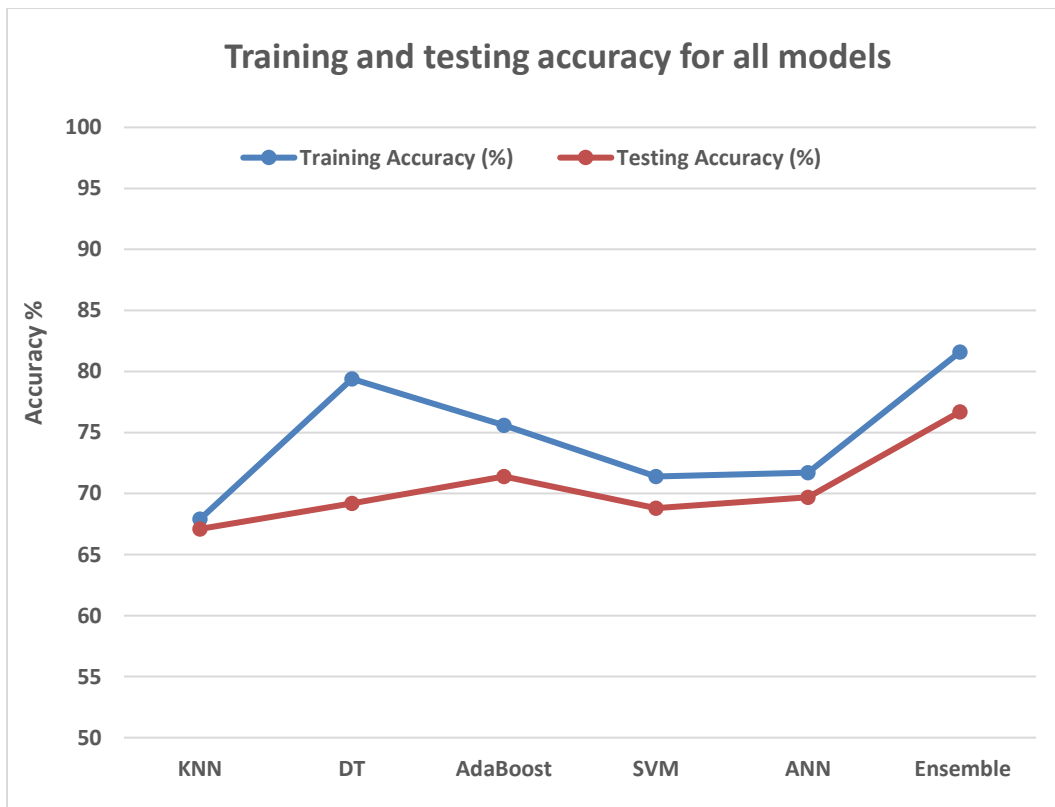


Figure 4-16 : Accuracy of all Models intersections GB

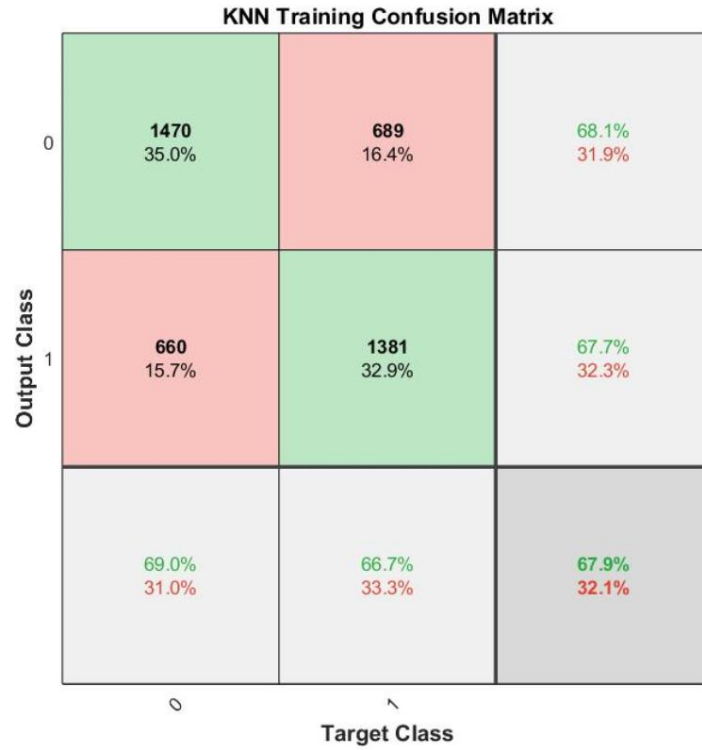


Figure 4-17 : KNN Training CM Road Intersections GB

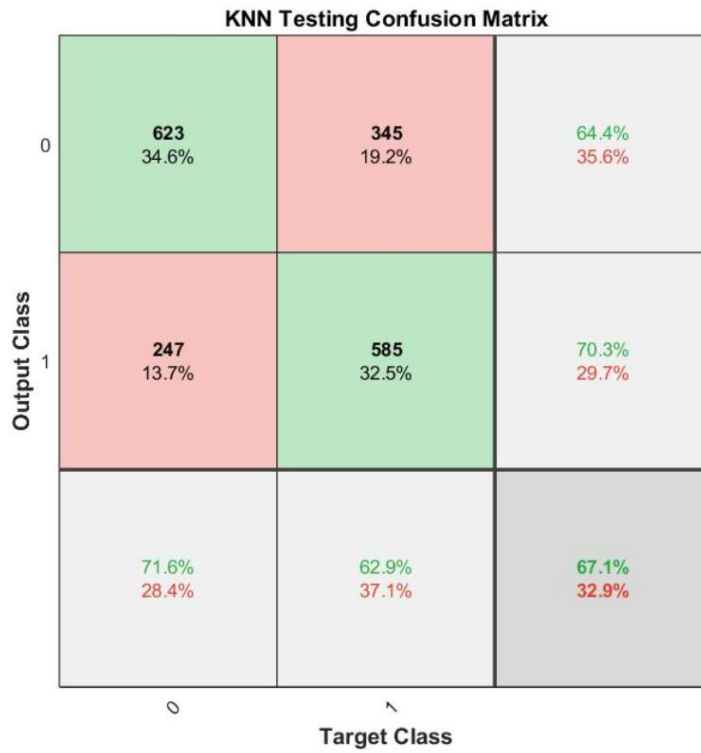


Figure 4-18 : KNN Testing CM Road Intersections GB

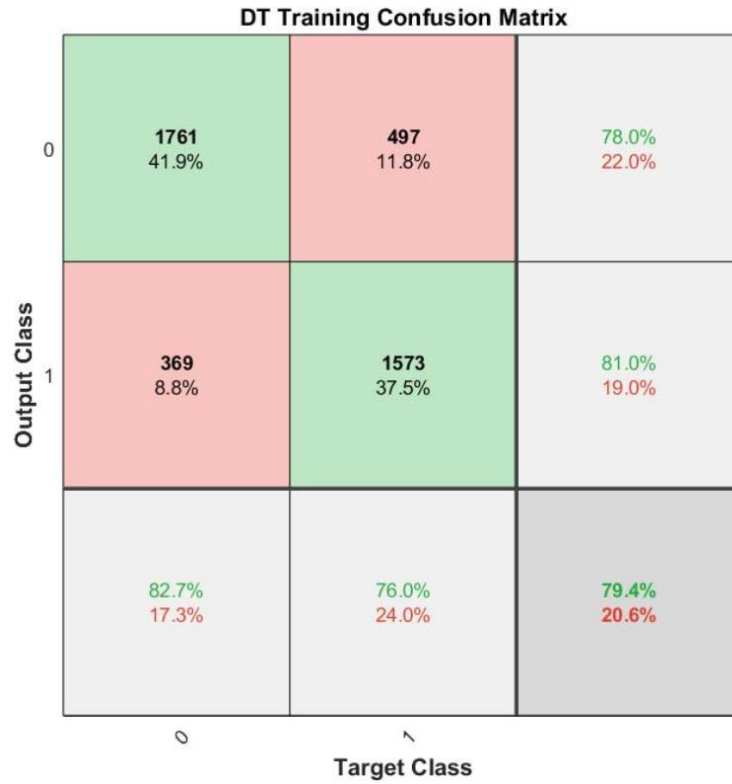


Figure 4-19 : DT Training CM Road Intersections GB

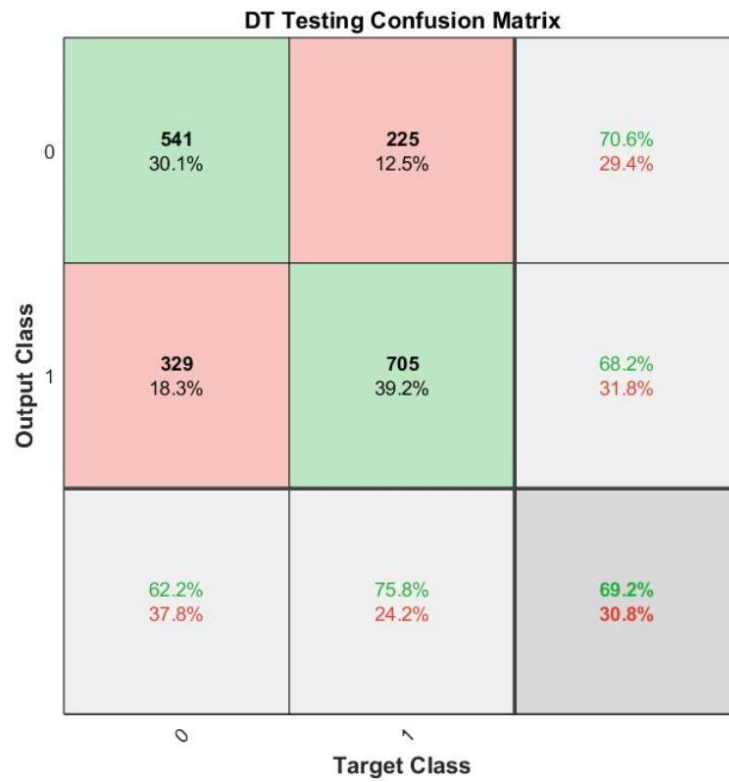


Figure 4-20 : DT Testing CM Road Intersections GB

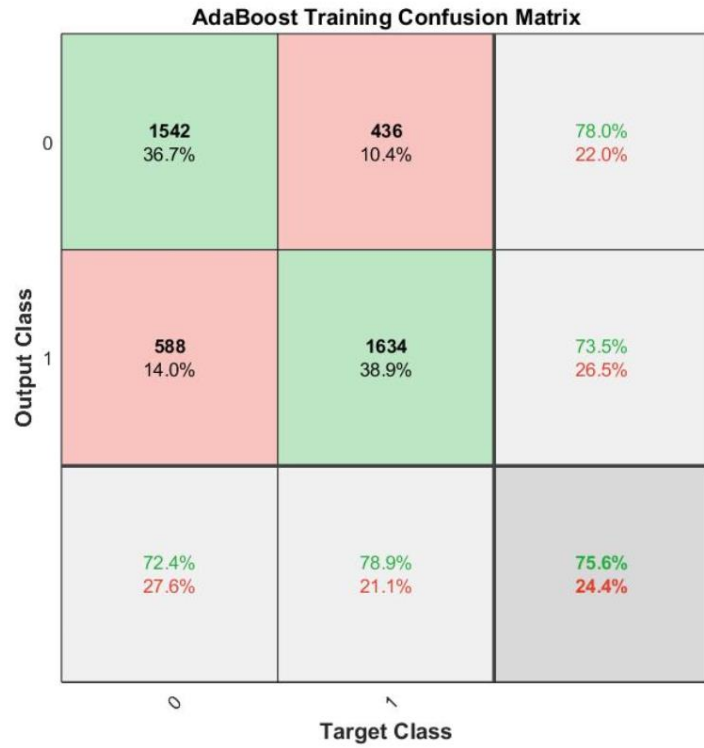


Figure 4-21 : AdaBoost Training CM Road Intersections GB

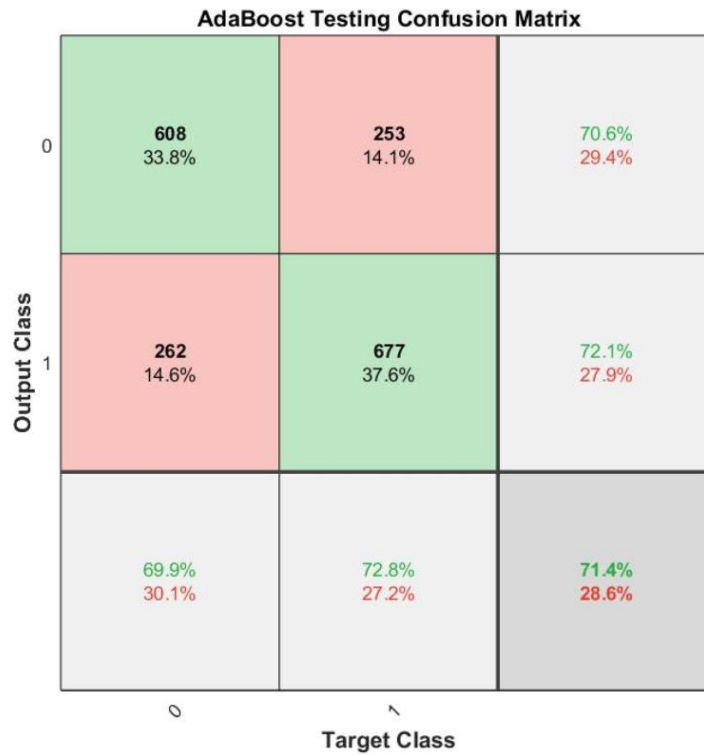


Figure 4-22 : AdaBoost Testing CM Road Intersections GB

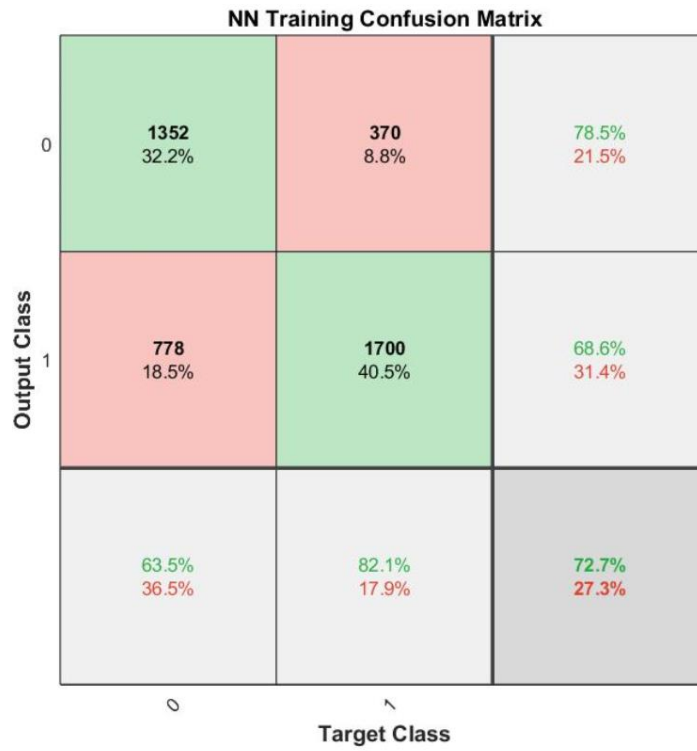


Figure 4-23 : ANN Training CM Road Intersections GB

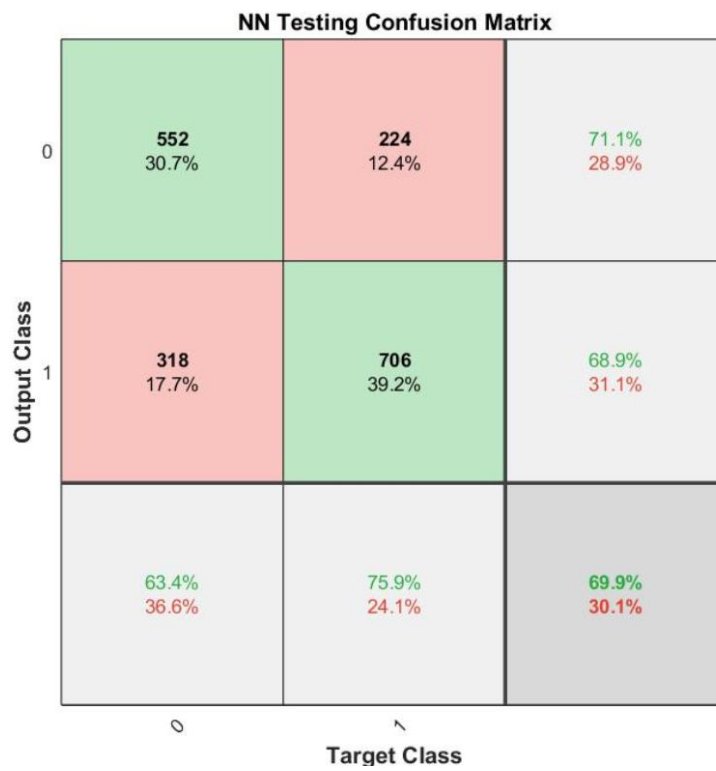


Figure 4-24 : ANN Testing CM Road Intersections GB



Figure 4-25 : SVM Training CM Road Intersections GB

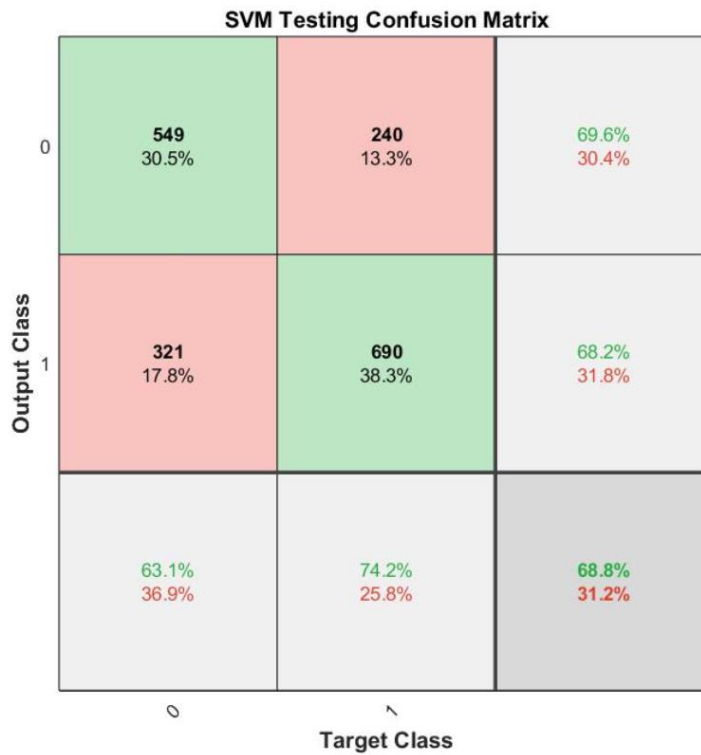


Figure 4-26 : SVM Testing CM Road Intersections GB

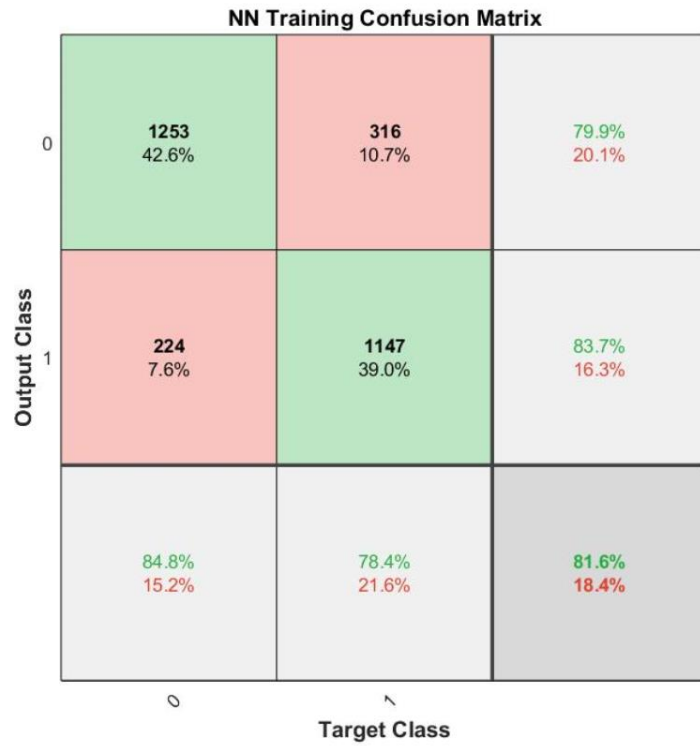


Figure 4-27 : Two-Layer Ensemble Training CM Road Intersections GB

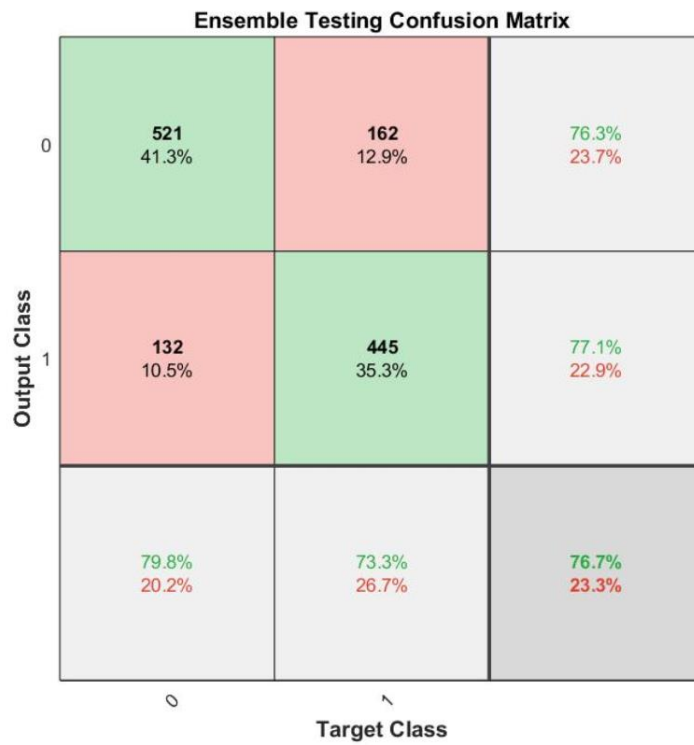


Figure 4-28 : Two-Layer Ensemble Testing CM Road Intersections GB

4.2.2 Performance Measures for Road Intersections Great Britain

Although accuracy is one of the metrics representing the performance of individual models but relying only on accuracy as a performance measure can be misleading as the model might be biased towards one severity class. So, to overcome these limitations we determined other performance measures like precision, recall and F1 score. These performance measures determine the performance of individual severity levels providing better insight into the performance of models. Results of these performance measures for severe and non-severe crashes are illustrated in Table 4-6 and Table 4-7 respectively. Figure 4-29 and Figure 4-30 also shed light on the performance of models for severe and non-severe crashes.

Table 4-6 : Performance Measures for Severe Crashes Road Intersections GB

Model	Precision	Recall	F1 Score
K Nearest Neighborhood (KNN)	0.6290	0.7031	0.6640
Decision Trees (DT)	0.6817	0.7076	0.6944
AdaBoost	0.7280	0.7210	0.7245
Artificial Neural Network (ANN)	0.7032	0.7063	0.7047
Support Vector Machine (SVM)	0.7280	0.6979	0.7126
Two-Layer Ensemble	0.7331	0.7712	0.7517

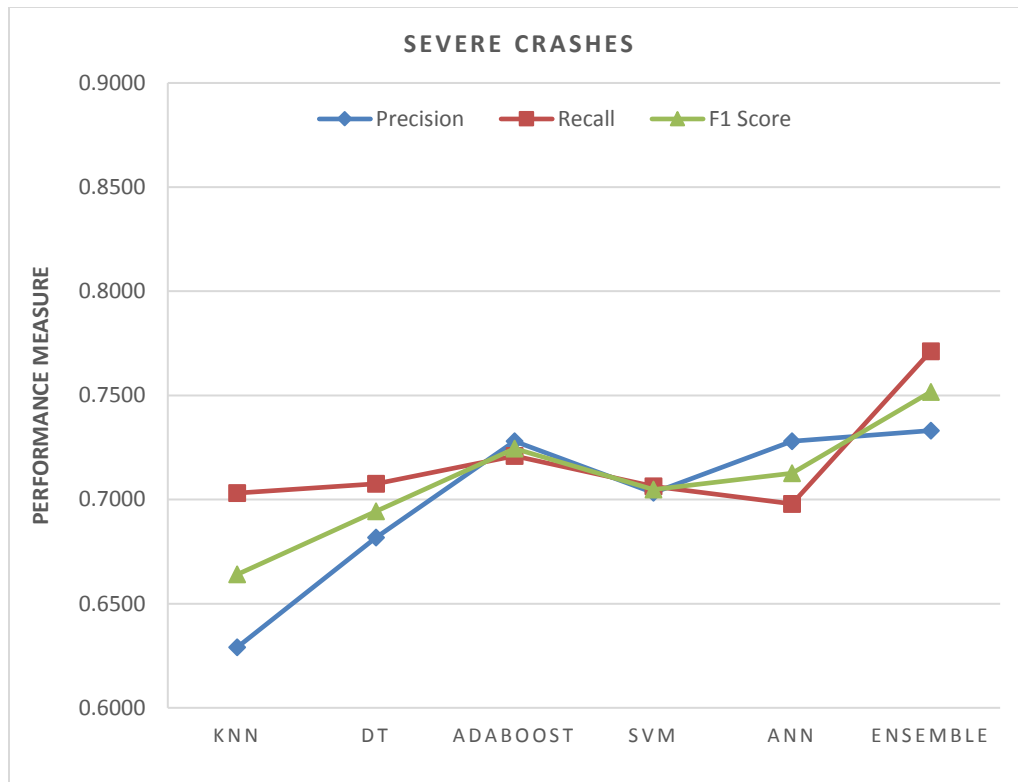


Figure 4-29 : Performance Measures for severe crashes Road Intersections GB

Table 4-7 : Performance Measures for Non-Severe Crashes Road Intersections GB

Model	Precision	Recall	F1 Score
K Nearest Neighborhood (KNN)	0.7161	0.6436	0.6779
Decision Trees (DT)	0.6989	0.6726	0.6855
AdaBoost	0.6989	0.7062	0.7025
Artificial Neural Network (ANN)	0.6874	0.6842	0.6858
Support Vector Machine (SVM)	0.6632	0.6952	0.6788
Two-Layer Ensemble	0.7979	0.7628	0.7799

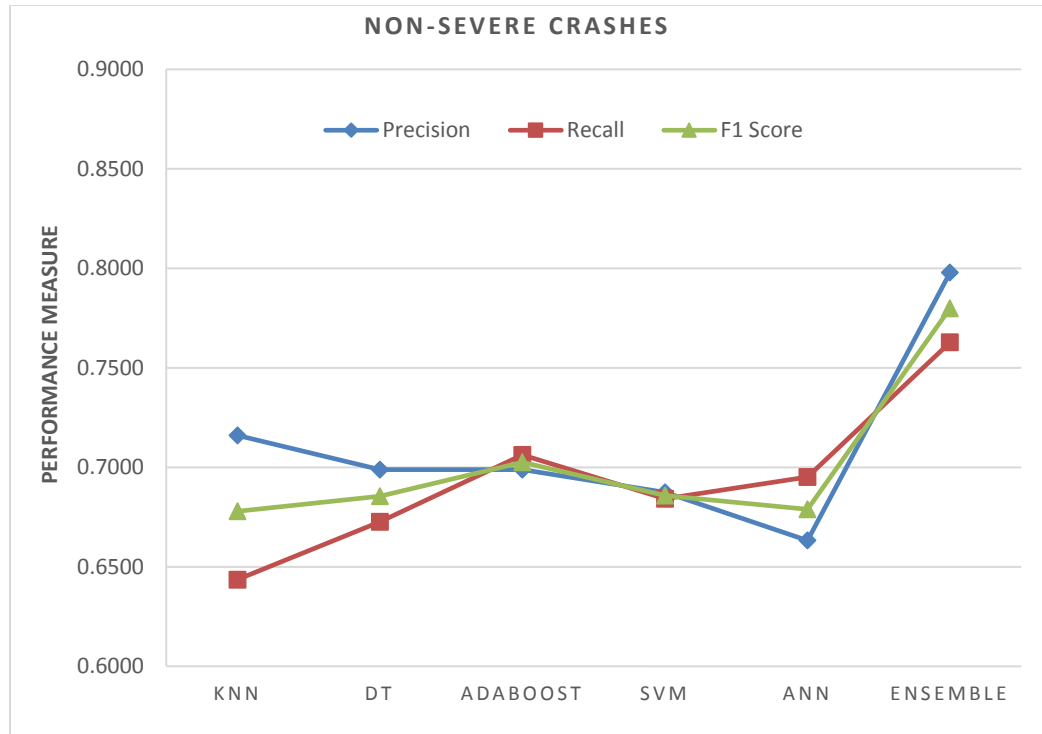


Figure 4-30 : Performance measures for Non-Severe crashes Road Intersections GB

According to the definition of precision and recall, it is stated that any model which maximizes both performance measures is the best one. F1 score acts as a good performance indicator since it uses both precision and recall for interpreting the model's performance. In this study, we notice that each model performed almost equally good for both levels of severity as evident from the results of performance measures in Table 4-6 and Table 4-7. Among the base models, the KNN model has the least F1 score while AdaBoost has the maximum score for both the severity levels. Among individual models, AdaBoost outperforms all other models in terms of accuracy and F1 score while KNN shows the least performance among all the models. But there is an abrupt enhancement in the predictive performance with the introduction of the two-layer ensemble model. The test accuracy of the two-layer ensemble model increased to 76.7% while the second-best model which is AdaBoost has a test accuracy of 71.4%. Similarly, the F1 score has

improved tremendously for both levels of severity. This enhanced performance of two-layer ensemble model indicates that it is a viable option to be employed for predicting traffic crash severity.

4.3 Modeling Results for all locations Canada

The classification accuracy of each model is shown in and Table 4-8 and Figure 4-31. The overall training accuracy ranges from 75.9% to 84.3% while the overall testing accuracy ranges from 70.9% to 82.3%. DT performs better in training while the testing performance decline to suggest that they suffer from the issue of overfitting. Among the base models, AdaBoost has the highest testing accuracy while KNN shows least performance among all models. The proposed two-layer ensemble model shows excellent performance in training and testing outperforming all the other models with a training accuracy of 84.3% and testing accuracy of 82.3%. The confusion matrices for all the models are shown in Figure 4-32 and Figure 4-43.

4.3.1 Accuracy of Models all Locations Canada

Table 4-8 : Accuracy of Models for all locations Canada

S. No	Model	Training Accuracy (%)	Testing Accuracy (%)
1	K Nearest Neighborhood (KNN)	75.9	70.9
2	Decision Trees (DT)	82.6	73.3
3	AdaBoost	81	76.2
4	Artificial Neural Network (ANN)	76.7	73.9
5	Support Vector Machine (SVM)	76.4	73
6	Ensemble Model	84.3	82.3

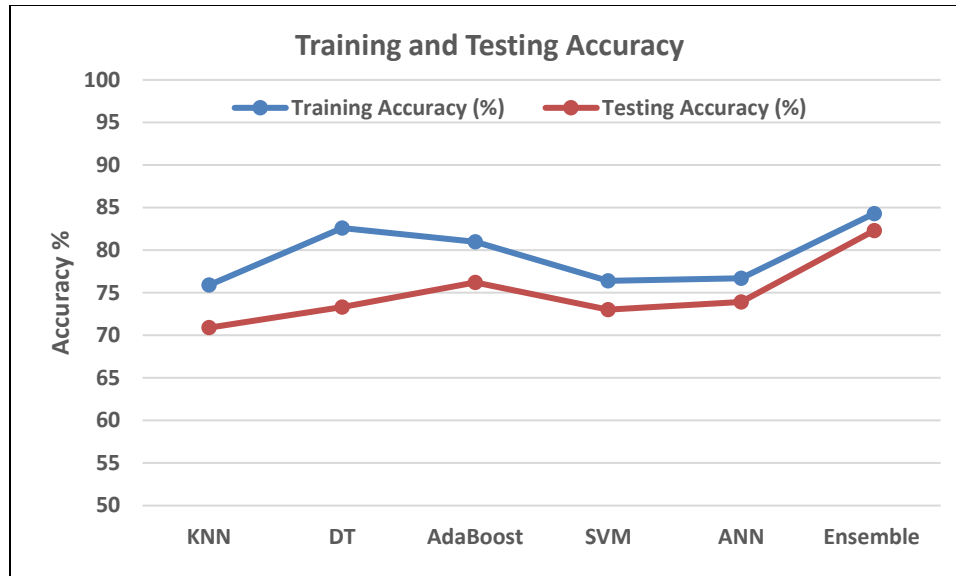


Figure 4-31 : Accuracy of Models all locations Canada

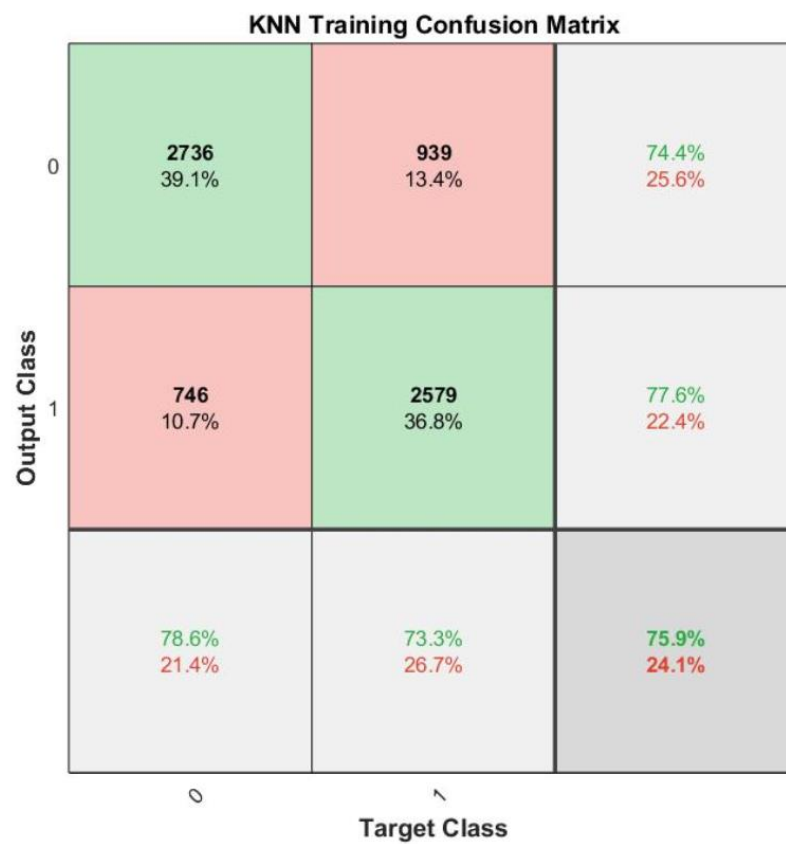


Figure 4-32 : KNN Training CM Canada

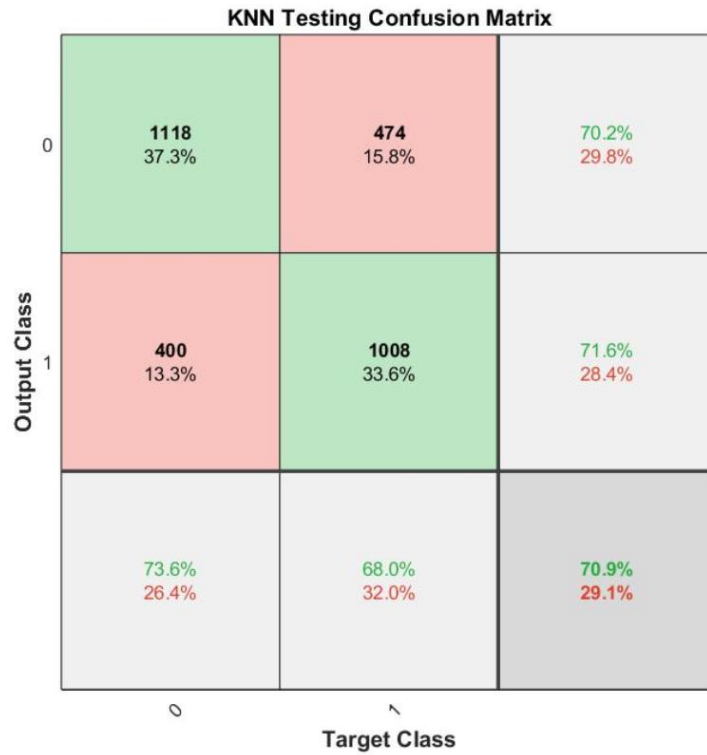


Figure 4-33 : KNN Testing CM Canada

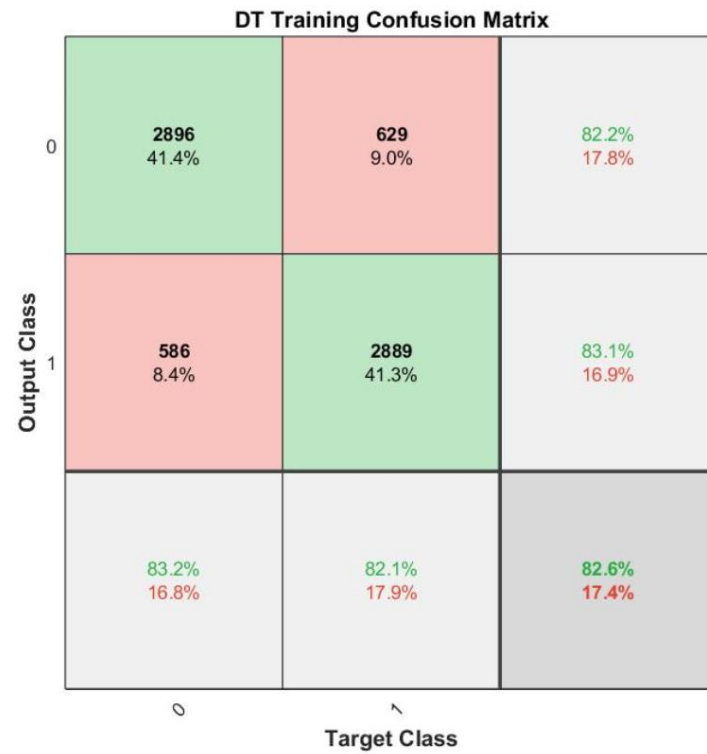


Figure 4-34 : DT Training CM Canada

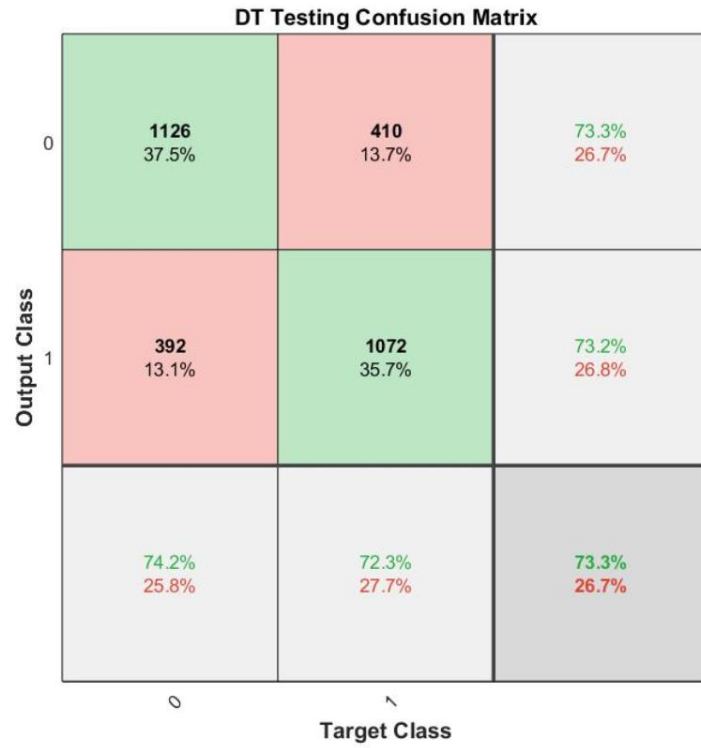


Figure 4-35 : DT Testing CM Canada

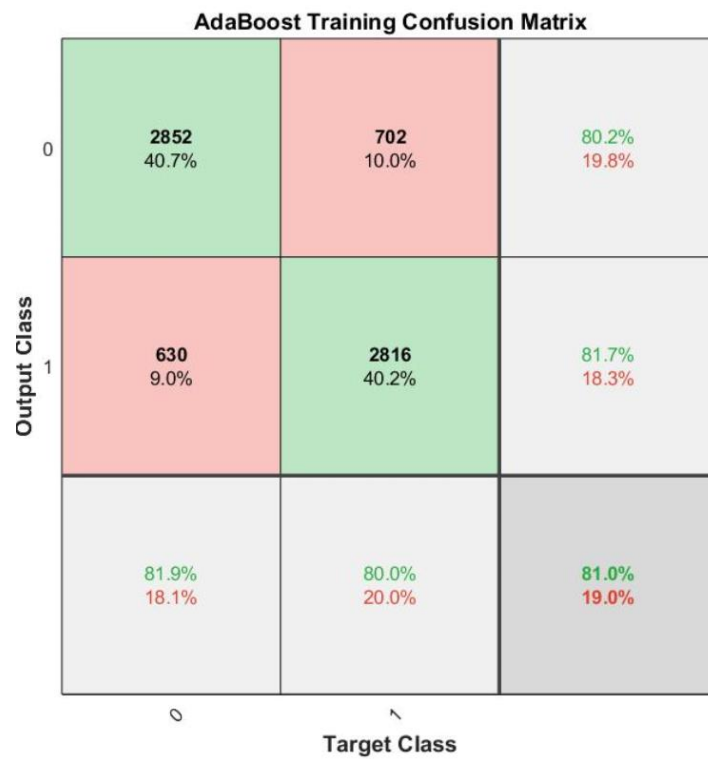


Figure 4-36 : AdaBoost Training CM Canada

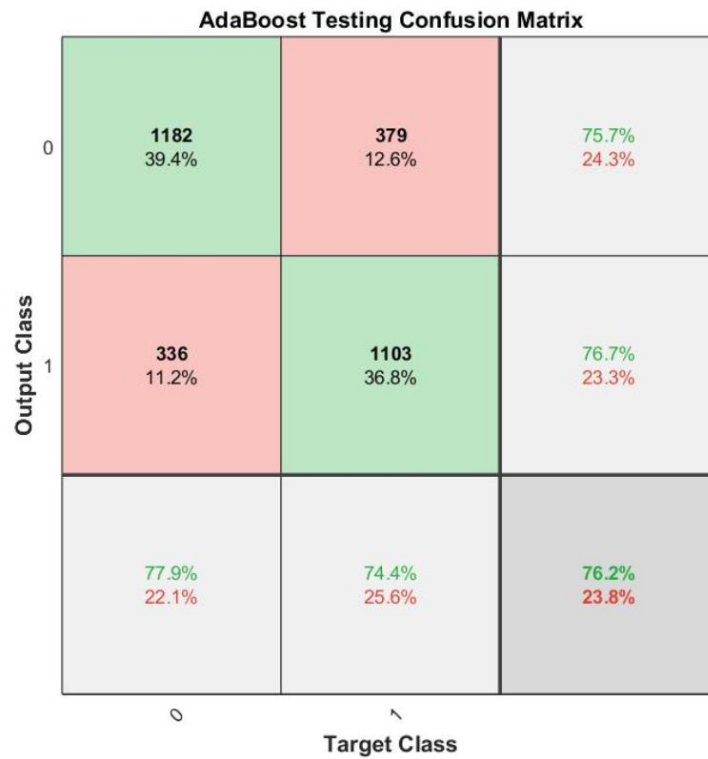


Figure 4-37 : AdaBoost Testing CM Canada

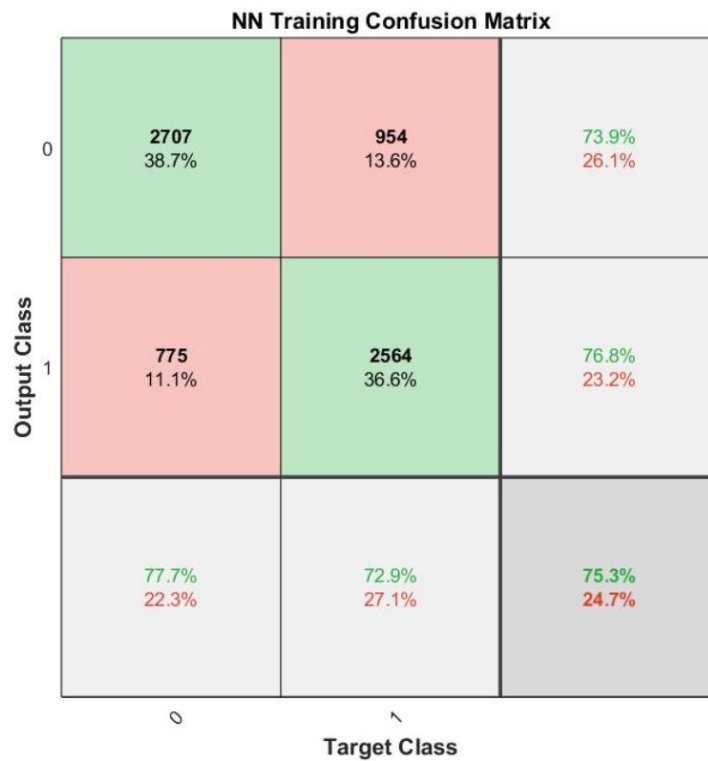


Figure 4-38 : ANN Training CM Canada

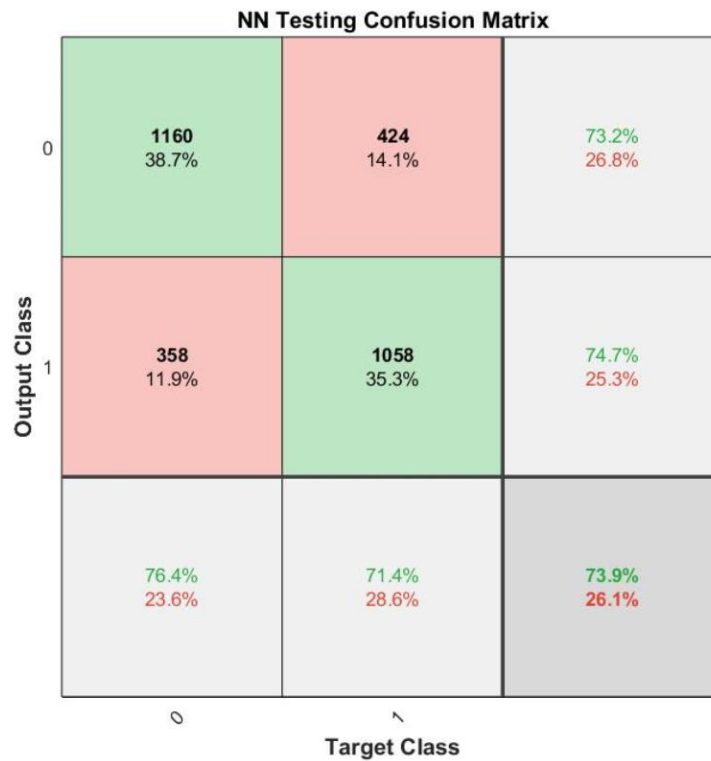


Figure 4-39 : ANN Testing CM Canada

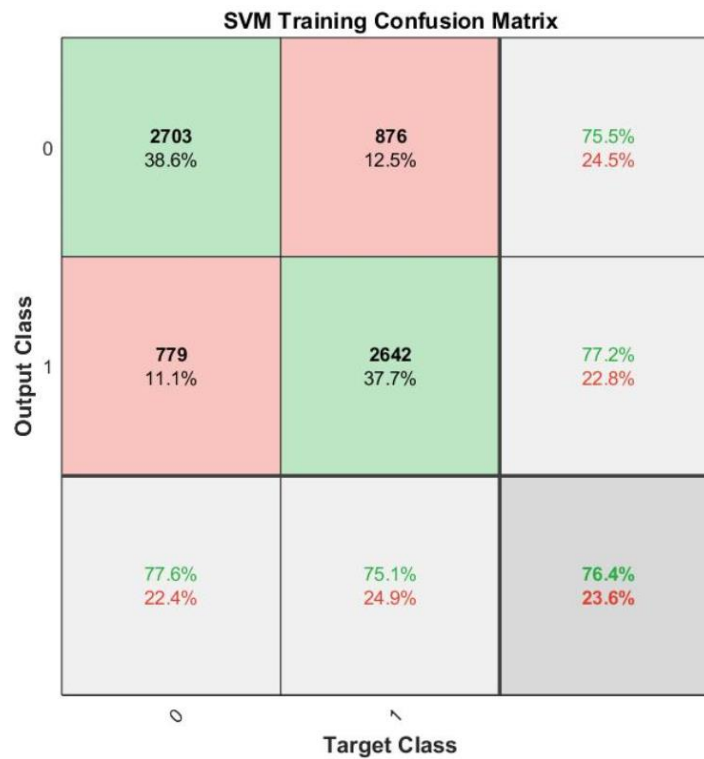


Figure 4-40 : SVM Training CM Canada

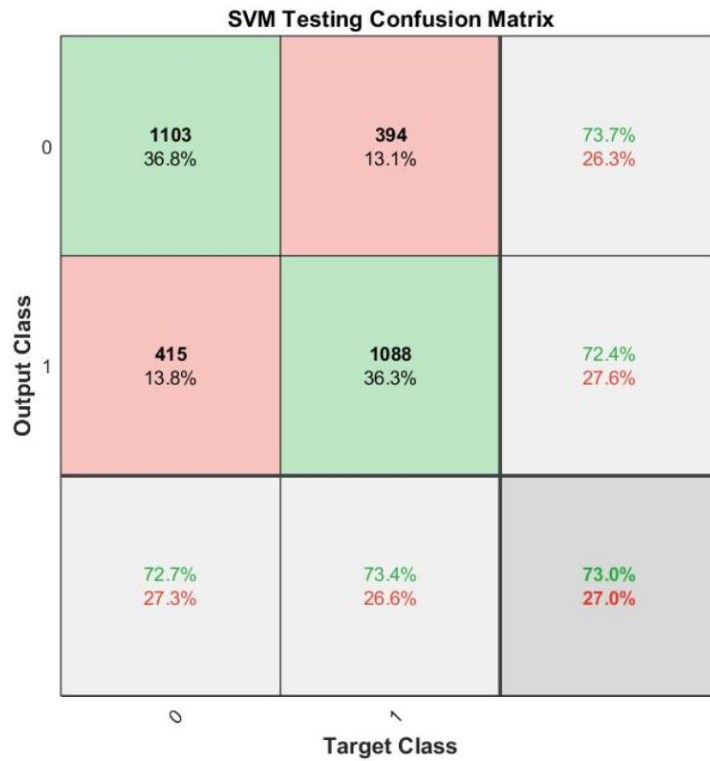


Figure 4-41 : SVM Testing CM Canada

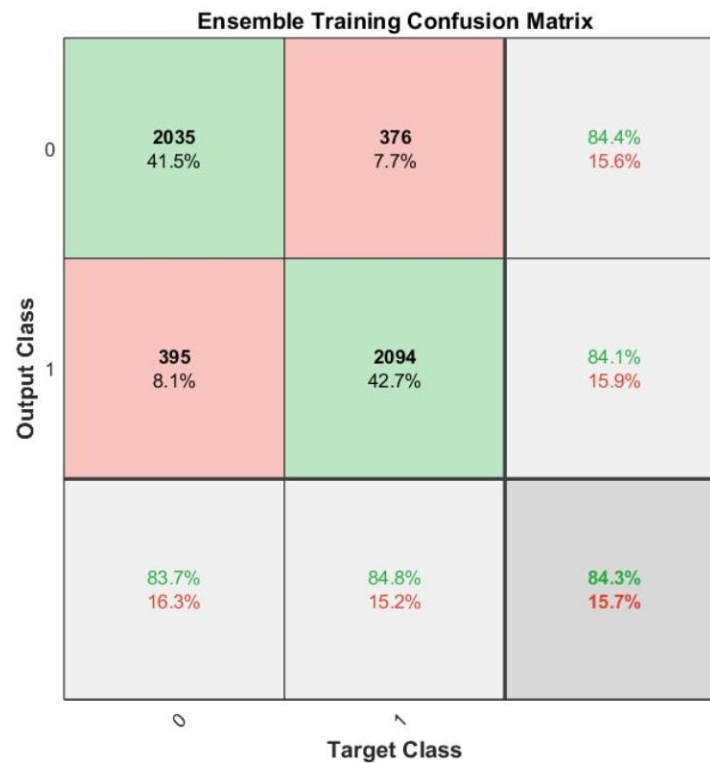


Figure 4-42 : Two-Layer Ensemble Training CM Canada

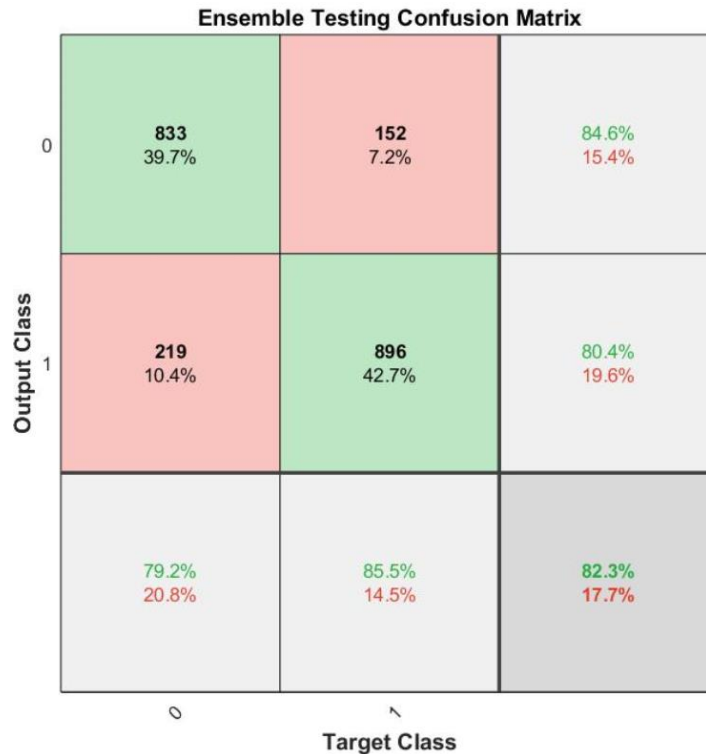


Figure 4-43 : Two-Layer Ensemble Testing CM Canada

4.3.2 Performance Measures for All locations Canada

Although accuracy is one of the metrics representing the performance of individual models but relying only on accuracy as a performance measure can be misleading as the model might be biased towards one severity class. So, to overcome these limitations we determined other performance measures like precision, recall and F1 score. These performance measures determine the performance of individual severity levels providing better insight into the models. Results of these performance measures for fatal and non-fatal crashes are illustrated in Table 4-9 and Table 4-10 respectively. Figure 4-44 and Figure 4-45 also shed light on the performance of models for fatal and non-fatal crashes.

Table 4-9 : Performance measures for Fatal Crashes Canada

Model	Precision	Recall	F1 Score
KNN	0.6802	0.7159	0.6976
DT	0.7233	0.7322	0.7278
RF	0.7443	0.7665	0.7552
SVM	0.7346	0.7239	0.7292
ANN	0.7254	0.7486	0.7368
Ensemble	0.8550	0.8036	0.8285

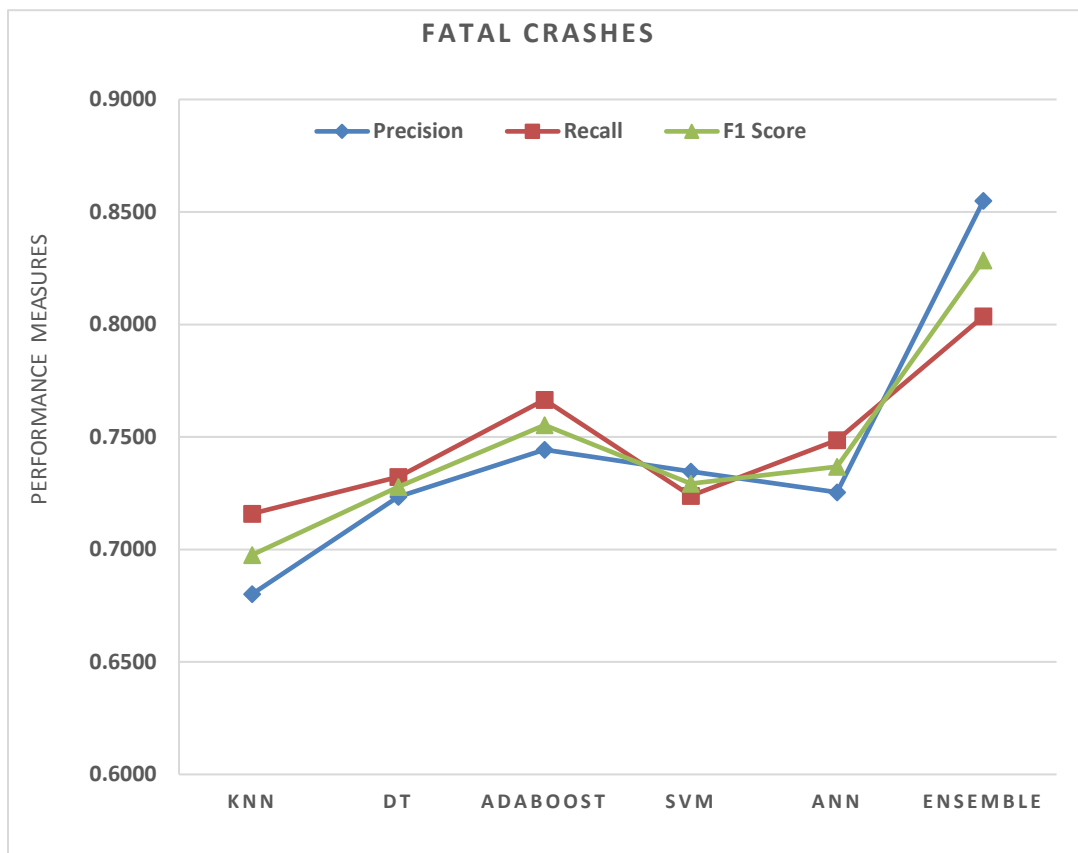
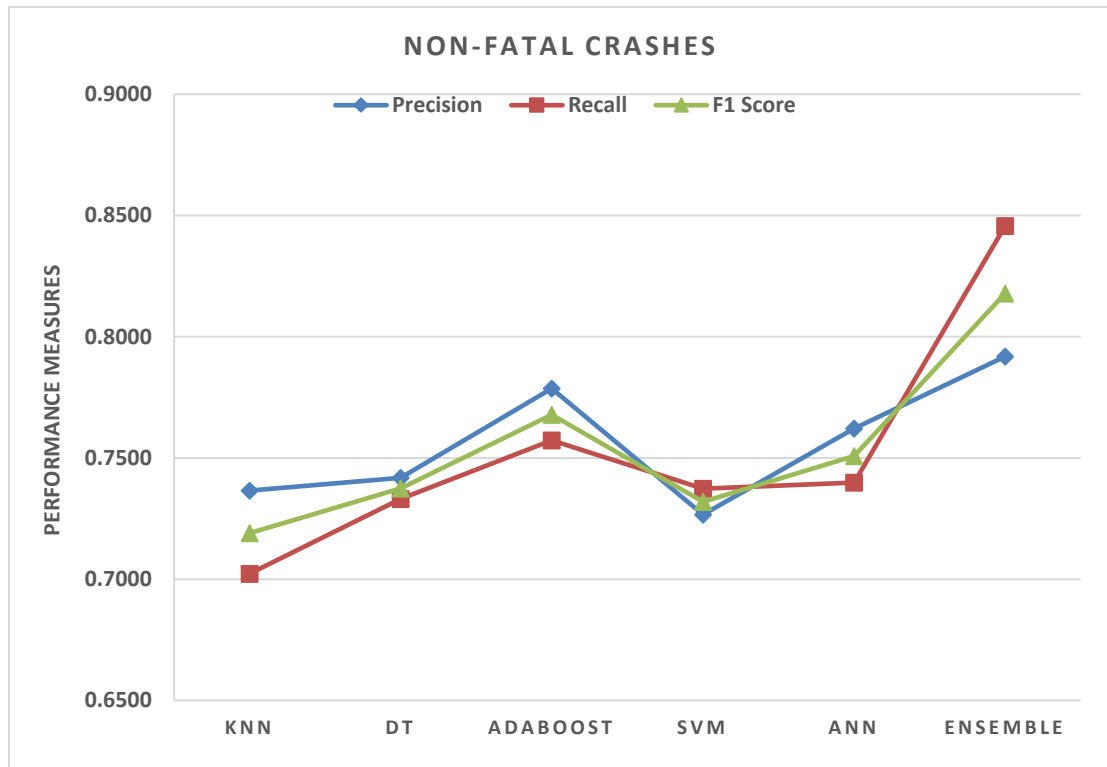


Figure 4-44 : Performance measures for Fatal Crashes Canada

Table 4-10 : Performance measures for Non-Fatal Crashes Canada

Model	Precision	Recall	F1 Score
KNN	0.7365	0.7023	0.7190
DT	0.7418	0.7331	0.7374
AdaBoost	0.7787	0.7572	0.7678
SVM	0.7266	0.7373	0.7319
ANN	0.7622	0.7398	0.7508
Ensemble	0.7918	0.8457	0.8179

**Figure 4-45 : Performance measures for Non-Fatal Crashes Canada**

According to the definition of precision and recall, it is stated that any model which maximizes both performance measures is the best one. F1 score acts as a good performance indicator since it uses both precision and recall for interpreting the model's performance. In this study, we notice that all the models performed almost equally good

for both levels of severity as evident from the results of performance measures in Table 4-9 and Table 4-10. Among the base models, the KNN model has the least F1 score while AdaBoost has the maximum score for both the severity levels. Also, we notice that DT, SVM, and ANN perform almost similarly for both levels of severity. So, among individual models, AdaBoost outperforms all other models in terms of accuracy and F1 score while KNN performs the worst among all the models. But there is an abrupt enhancement in the predictive performance with the introduction of the two-layer ensemble model. The test accuracy of the two-layer ensemble model increased to 82.3% while the second-best model which is AdaBoost has a test accuracy of 76.2%. Similarly, the F1 score has improved tremendously for both levels of severity. This enhanced performance of two-layer ensemble model indicates that it is a viable option to be employed for predicting traffic crash severity.

4.4 Modeling Results for Road Intersections Canada

The training and testing accuracy for all the models are shown in Table 4-11 and Figure 4-46. The overall training accuracy ranges from 72.7% to 83.0% while the overall testing accuracy ranges from 67.3% to 79.3%. DT performs better in training while the testing performance decline to suggest that they suffer from the issue of overfitting. Among the base models, AdaBoost has the highest testing accuracy while KNN performs the worst among all models. The proposed two-layer ensemble model shows excellent performance in training and testing outperforming all the other models with a training accuracy of 83.0% and testing accuracy of 79.3%. The confusion matrices for all the models are shown in Figure 4-47 to Figure 4-58.

4.4.1 Accuracy of Models Road Intersections Canada

Table 4-11 : Accuracy of Models Road Intersections Canada

S. No	Model	Training Accuracy (%)	Testing Accuracy (%)
1	K Nearest Neighborhood (KNN)	72.7	67.3
2	Decision Trees (DT)	78.9	71.6
3	AdaBoost	79.1	73.6
4	Artificial Neural Network (ANN)	72.9	71.9
5	Support Vector Machine (SVM)	75.5	69.9
6	Ensemble Model	83	79.3

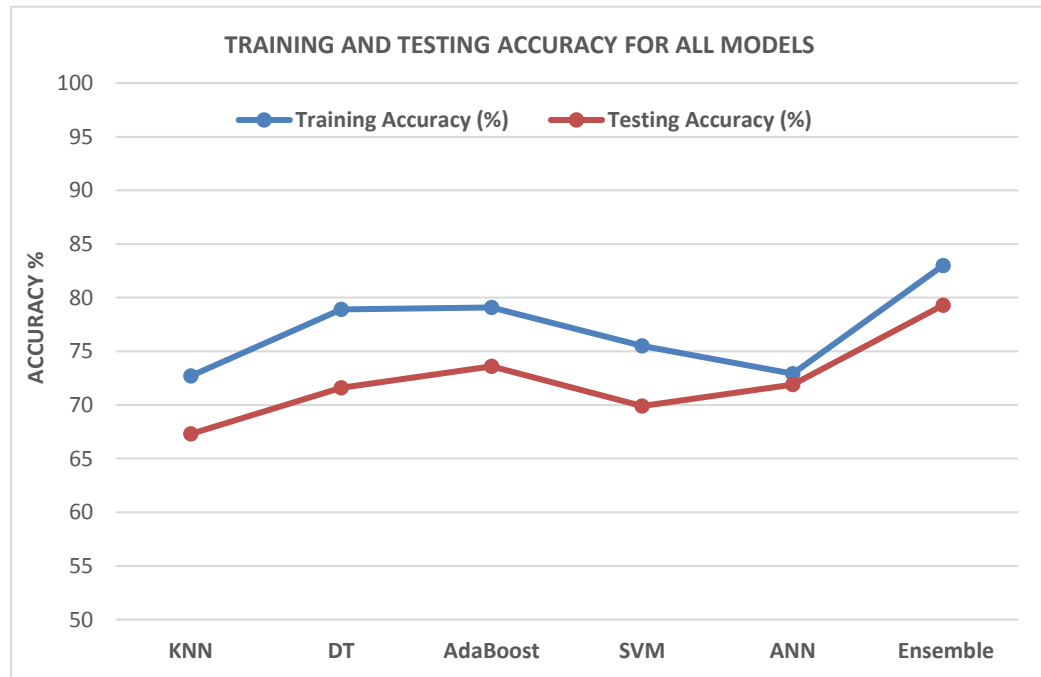


Figure 4-46 : Accuracy of Models Road Intersections Canada

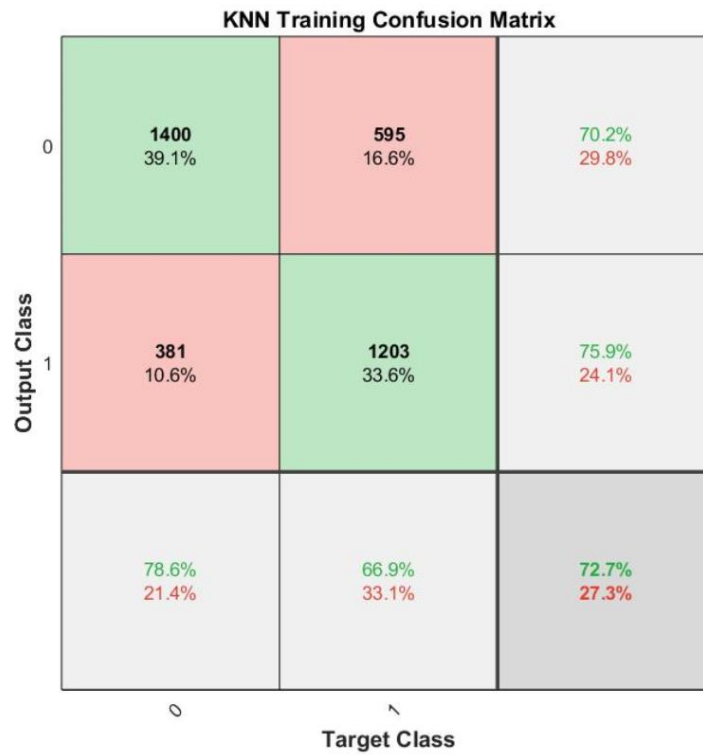


Figure 4-47 : KNN Training CM Road Intersections Canada

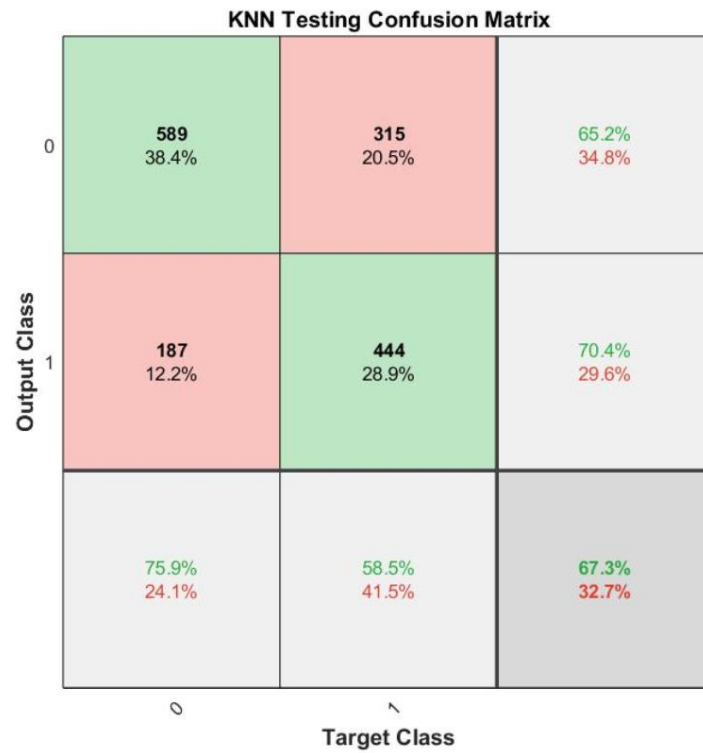


Figure 4-48 : KNN Testing CM Road Intersections Canada



Figure 4-49 : DT Training CM Road Intersections Canada

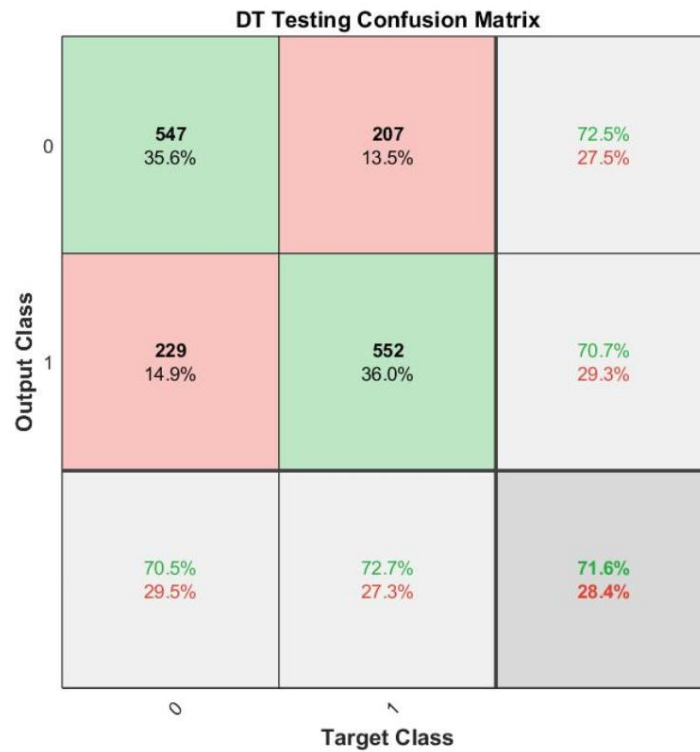


Figure 4-50 : DT Testing CM Road Intersections Canada



Figure 4-51 : AdaBoost Training CM Road Intersections Canada

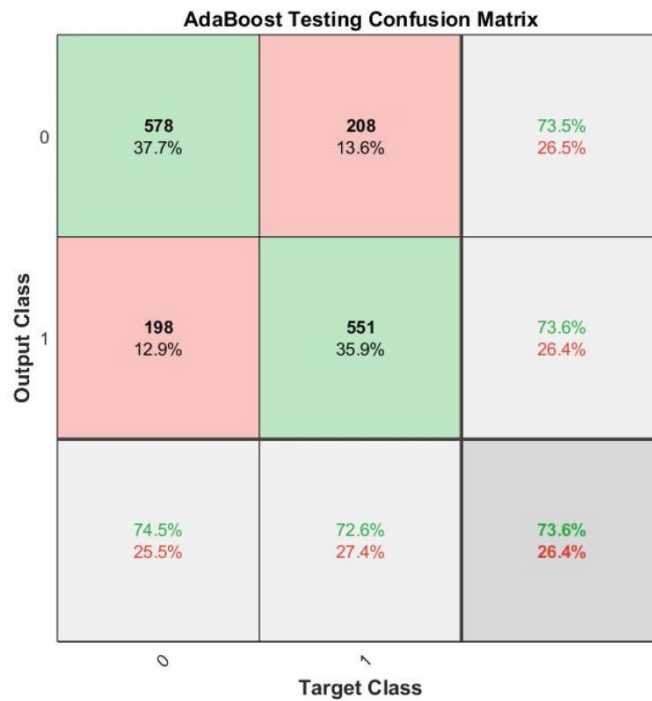


Figure 4-52 : AdaBoost Testing CM Road Intersections Canada

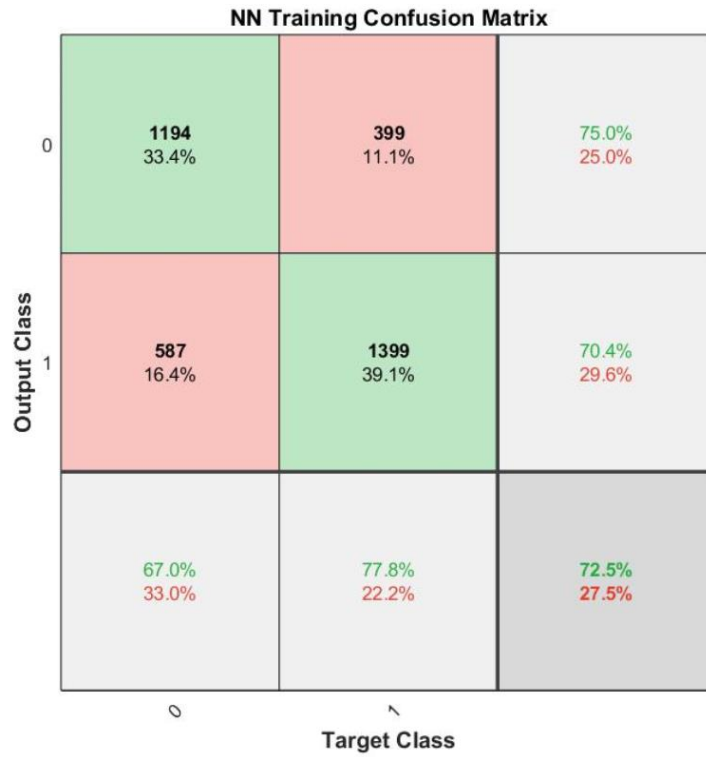


Figure 4-53 : ANN Training CM Road Intersections Canada

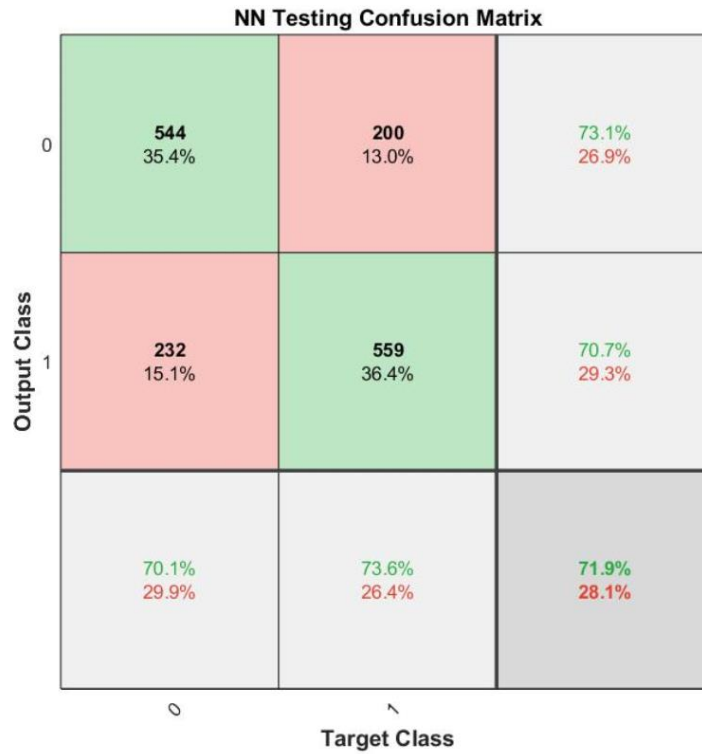


Figure 4-54 : ANN Testing CM Road Intersections Canada

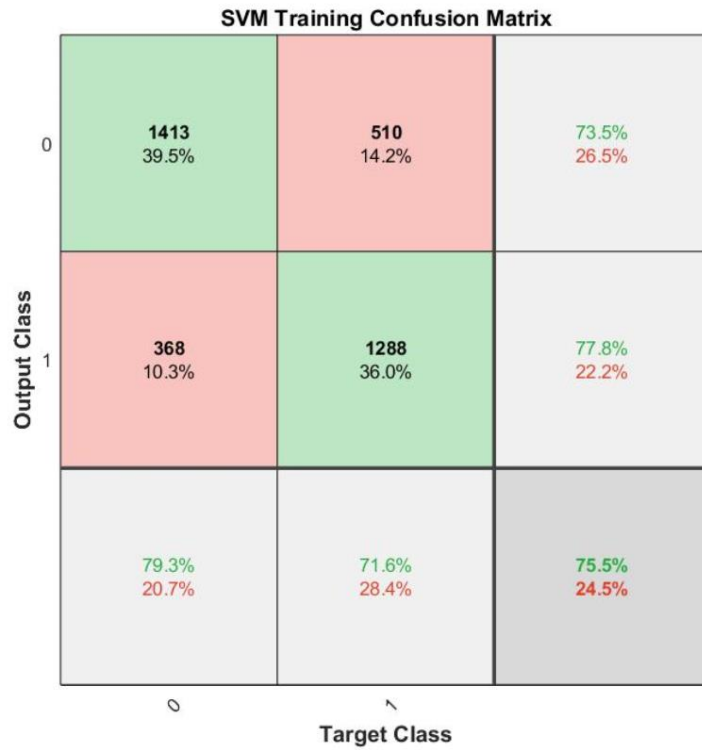


Figure 4-55 : SVM Training CM Road Intersections Canada

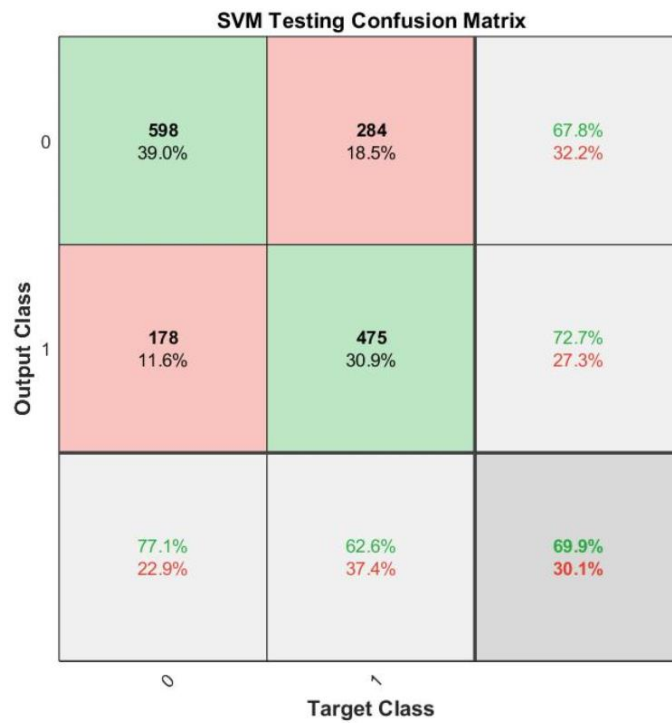


Figure 4-56 : SVM Testing CM Road Intersections Canada

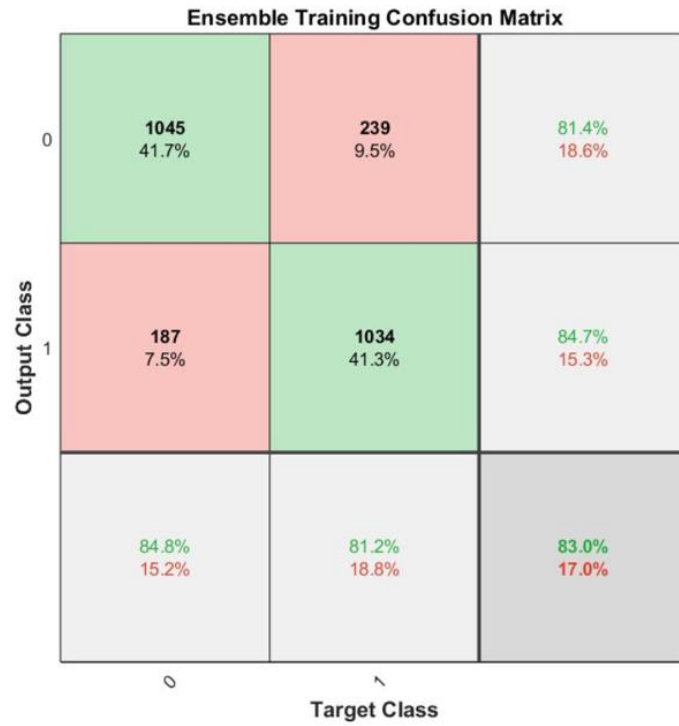


Figure 4-57 : Ensemble Training CM Road Intersections Canada

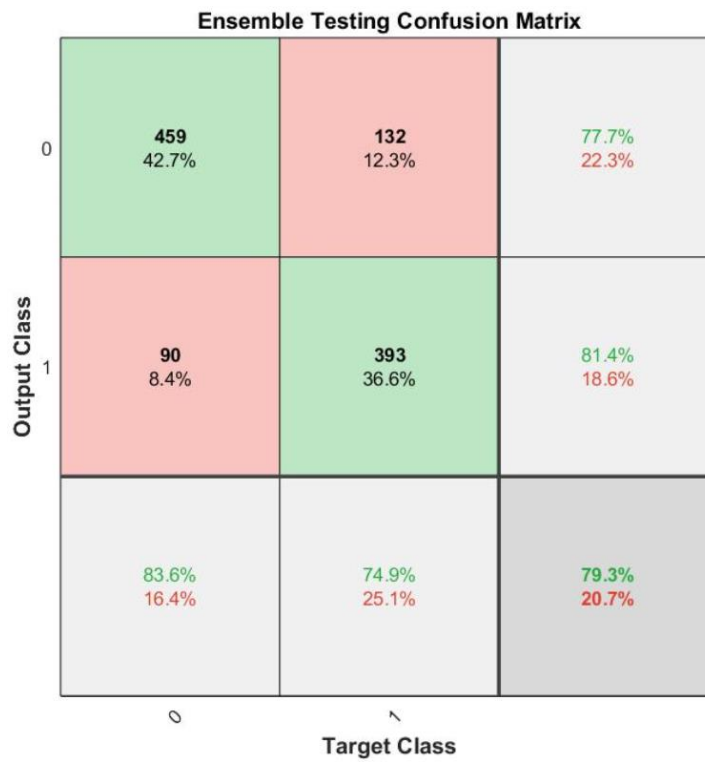


Figure 4-58 : Ensemble Testing CM Road Intersections Canada

4.4.2 Performance Measures for Road Intersections Canada

Although accuracy is one of the metrics representing the performance of individual models but relying only on accuracy as a performance measure can be misleading as the model might be biased towards one severity class. So, to overcome these limitations we determined other performance measures like precision, recall and F1 score. These performance measures determine the performance of individual severity levels providing better insight into the models. Results of these performance measures for fatal and non-fatal crashes are illustrated in Table 4-12 and Table 4-13 respectively. Figure 4-59 and Figure 4-60 also shed light on the performance of models for fatal and non-fatal crashes.

Table 4-12 : Performance Measures Non-Fatal Crashes Road Intersections Canada

Model	Precision	Recall	F1 Score
KNN	0.7590	0.6515	0.7012
DT	0.7049	0.7255	0.7150
AdaBoost	0.7448	0.7354	0.7401
SVM	0.7706	0.6780	0.7214
ANN	0.7019	0.7312	0.7163
Ensemble	0.8361	0.7766	0.8053

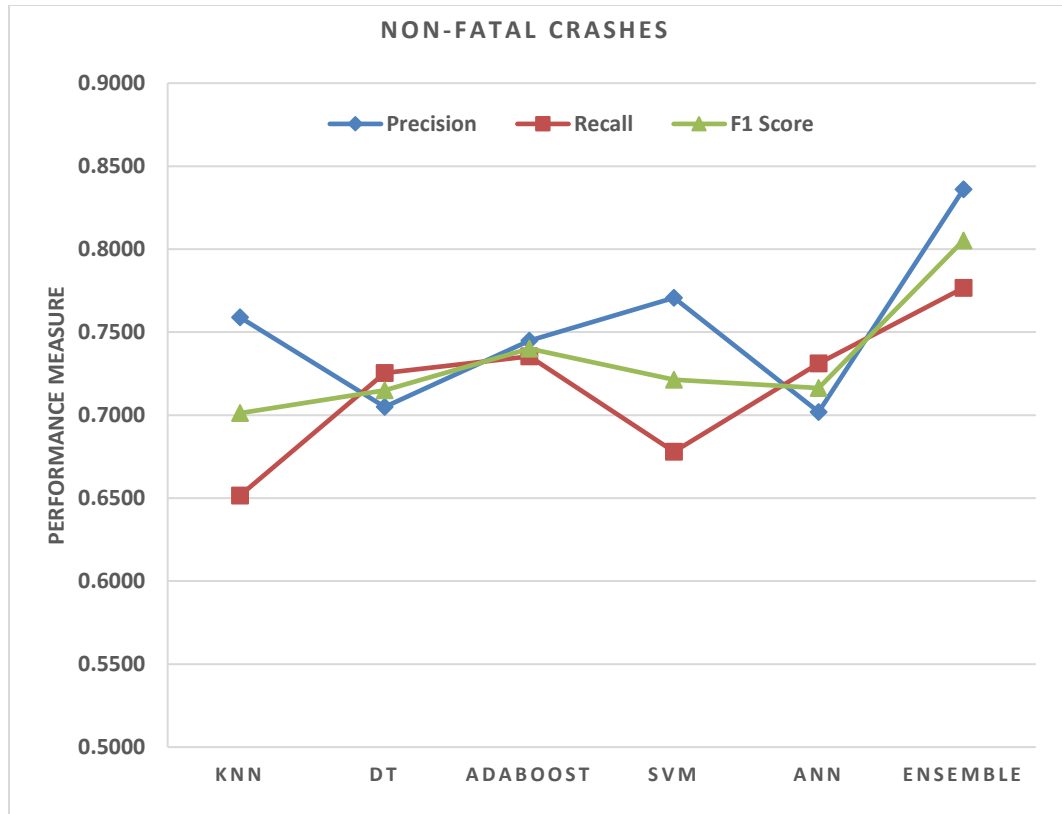


Figure 4-59 : Performance Measures Non-Fatal Crashes Road Intersections Canada

Table 4-13 : Performance Measures Fatal Crashes Road Intersections Canada

Model	Precision	Recall	F1 Score
KNN	0.5850	0.7036	0.6388
DT	0.7273	0.7068	0.7169
AdaBoost	0.7260	0.7356	0.7308
SVM	0.6258	0.7274	0.6728
ANN	0.7365	0.7076	0.7218
Ensemble	0.7486	0.8137	0.7798

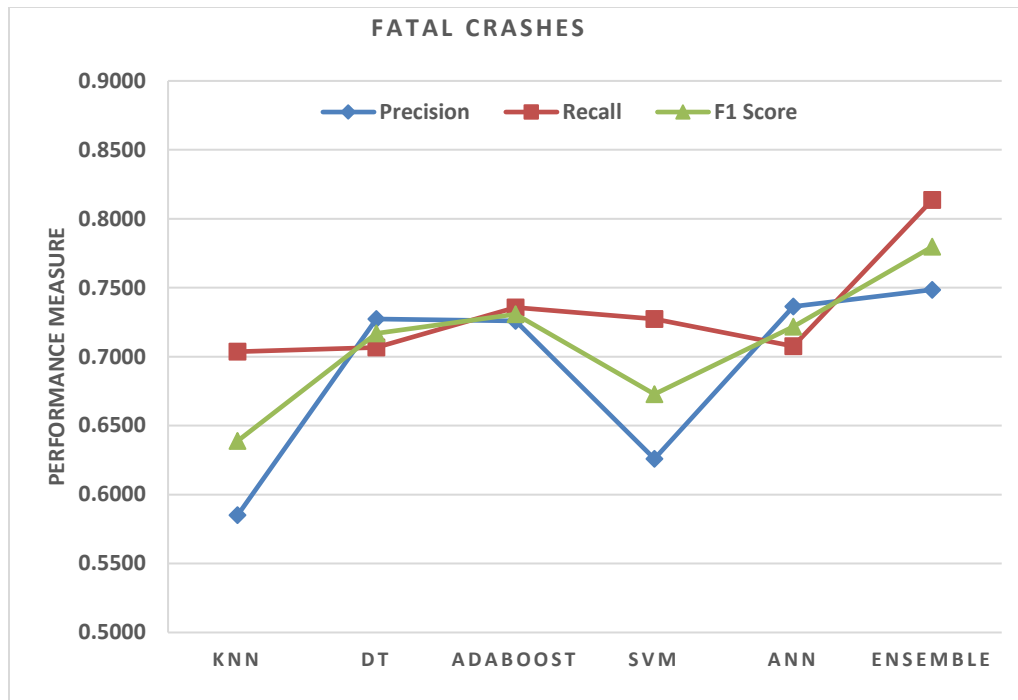


Figure 4-60 : Performance Measures Fatal Crashes Road Intersections Canada

According to the definition of precision and recall, it is stated that any model which maximizes both performance measures is the best one. F1 score acts as a good performance indicator since it uses both precision and recall for interpreting the model's performance. In this study, we notice that all the models performed almost equally good for both levels of severity as evident from the results of performance measures in Table 4-12 and Table 4-13. Among the base models, the KNN model has the least F1 score while AdaBoost has the maximum score for both the severity levels. So, among individual models, AdaBoost outperforms all other models in terms of accuracy and F1 score while KNN performs the worst among all the models. But there is an abrupt enhancement in the predictive performance with the introduction of the wo-layer ensemble model. The test accuracy of the two-layer ensemble model increased to 79.3% while the second-best model which is AdaBoost has a test accuracy of 73.6%. Similarly, the F1 score has improved tremendously for both levels of severity. This enhanced

performance of two-layer ensemble model indicates that it is a viable option to be employed for predicting traffic crash severity.

4.5 Models Comparison

Since we developed models for two crash datasets namely; Crash dataset for Great Britain and Crash dataset for Canada. So, to check how different models perform on these two datasets the models are compared based on the testing accuracy and F1 score as illustrated in Table 4-14 and Figure 4-61. After comparison we notice that each model performs in the same line for each of the four cases. We also notice that the models for all locations perform better compared to the models for intersections. We can say that the models for all locations are robust in understanding the problem.

Table 4-14 : Comparison of Testing Accuracy for Both Datasets

Model	Testing Accuracy			
	All Road Types		Road Intersections	
	Great Britain	Canada	Great Britain	Canada
KNN	69	70.9	67.1	67.3
DT	72.3	73.3	69.2	71.6
AdaBoost	73.7	76.2	71.4	73.6
SVM	72.1	73	68.8	70
ANN	72.2	73.9	69.7	72.6
Ensemble	78	82.3	76.7	79.3

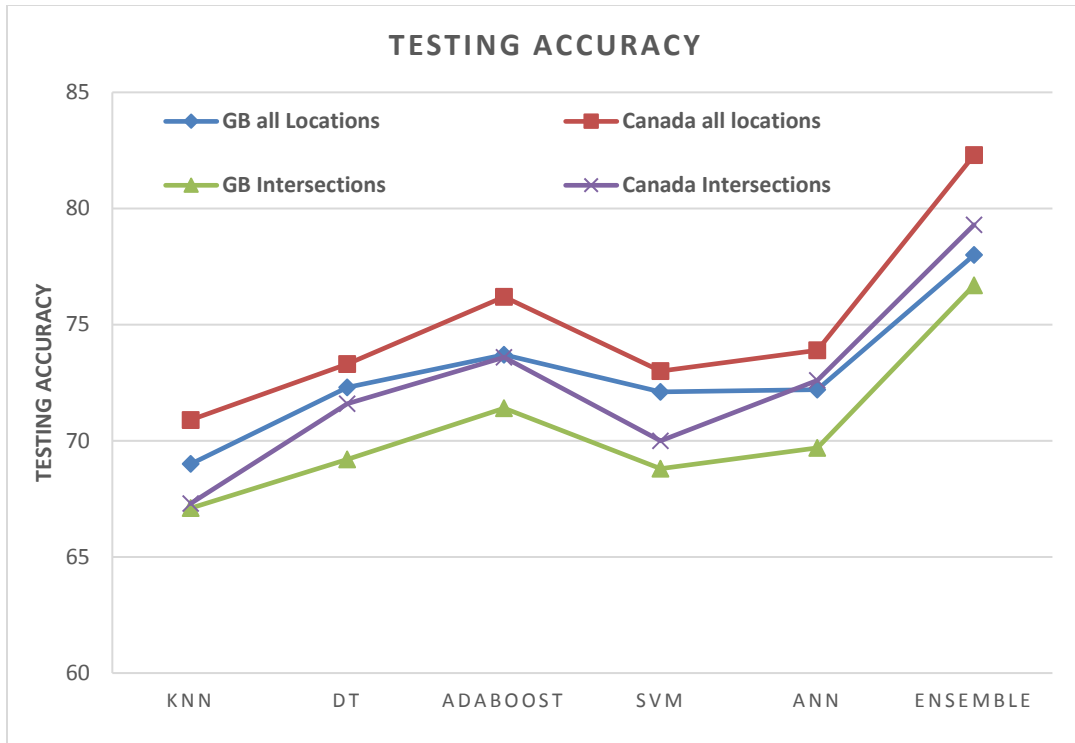


Figure 4-61 : Comparison of Testing Accuracy for Both Datasets

Table 4-15 : Comparison of performance measures for both Datasets Severe Crashes

Severe Crashes F1 Score				
Model	All Road Types		Road Intersections	
	Great Britain	Canada	Great Britain	Canada
KNN	0.6826	0.6976	0.664	0.6388
DT	0.7181	0.7278	0.6944	0.7169
AdaBoost	0.737	0.7552	0.7245	0.7308
SVM	0.7224	0.7292	0.7047	0.6728
ANN	0.7284	0.7368	0.7126	0.7218
Ensemble	0.781	0.8285	0.7517	0.7798

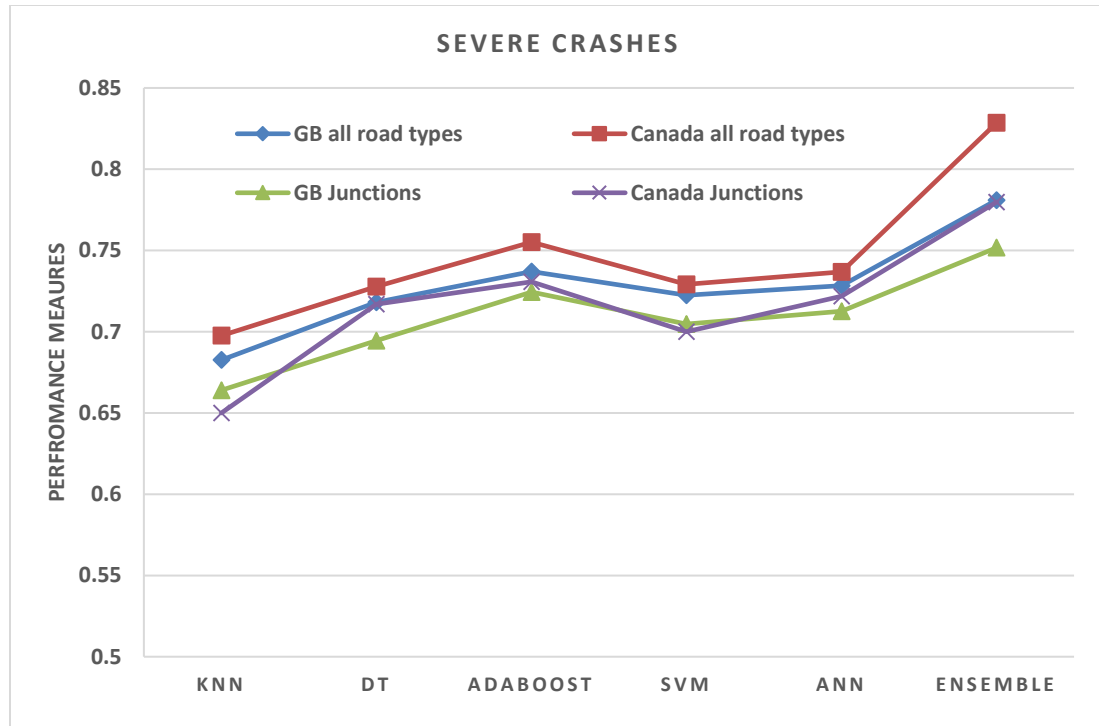


Figure 4-62 : Comparison of performance measures for both Datasets Severe Crashes

Table 4-16 : Comparison of performance measures for Non-Severe Crashes

Non-Severe Crashes F1 Score				
Model	All Road Types		Road Intersections	
	Great Britain	Canada	Great Britain	Canada
KNN	0.6989	0.719	0.6779	0.7012
DT	0.7186	0.7374	0.6855	0.715
AdaBoost	0.7343	0.7678	0.7025	0.7401
SVM	0.7196	0.7319	0.6858	0.7214
ANN	0.716	0.7508	0.6788	0.7163
Ensemble	0.7789	0.8179	0.7799	0.8053

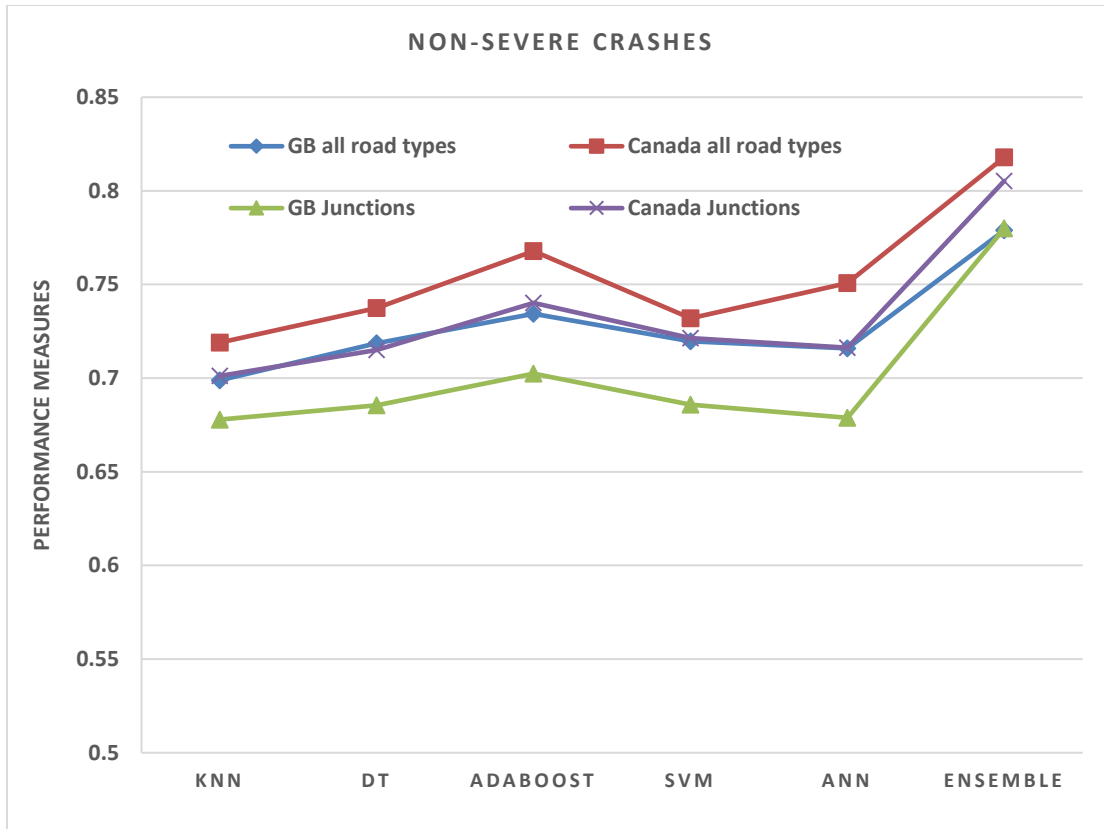


Figure 4-63 : Comparison of performance measures for Non-Severe Crashes

4.6 Simplified Ensemble Model

In order to simplify the two-layer ensemble model, we tried to eliminate each model from the first layer stepwise. We found out that eliminating KNN from the first layer resulted in least performance reduction compared to eliminating other models. Elimination of KNN resulted in an accuracy reduction of around 1 to 2 percent for different cases compared to the original ensemble model. The performance of ensemble model deteriorated abruptly when more than one models are removed from the first layer. This simplified ensemble model can be considered instead of the original model for predicting crash severity. This simplification of the two-layer ensemble model does not result in any computational enhancement and still we are losing a little on accuracy. We recommend

using the original ensemble model instead of the simplified model. The accuracy of the simplified two-layer ensemble model is shown in Table 4-17. The confusion matrices are shown in Figure 4-64 to Figure 4-71.

Table 4-17 : Accuracy of Simplified Ensemble Model

S. No	Simplified Ensemble Model	Training Accuracy (%)	Testing Accuracy (%)
1	Great Britain All Locations	83.7	77.0
2	Great Britain Intersections	81.0	75.6
3	Canada All Locations	82.2	80.1
4	Canada Intersections	80.8	77.7



Figure 4-64 : Simplified Ensemble training CM all location GB



Figure 4-65 : Simplified Ensemble testing CM all locations GB



Figure 4-66 : Simplified Ensemble training CM road intersections GB

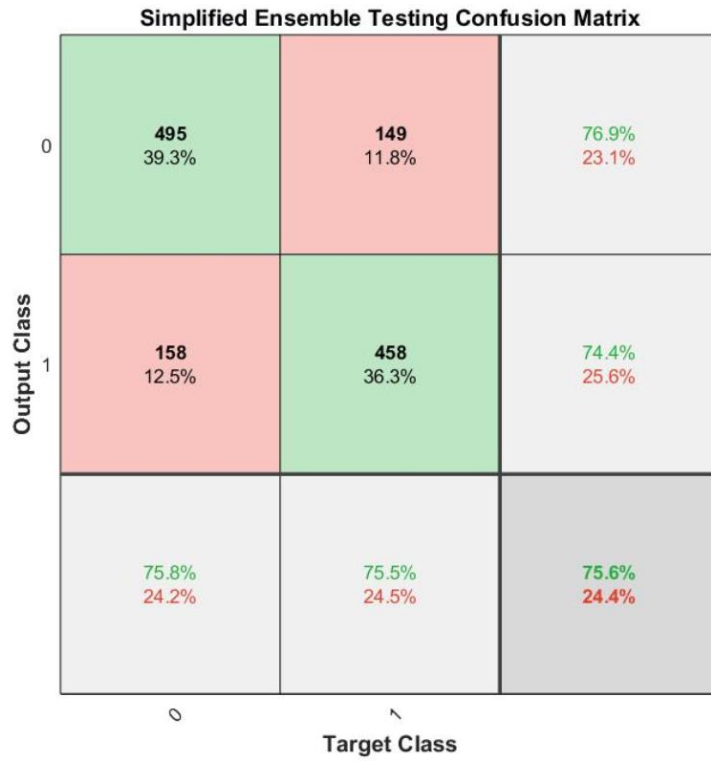


Figure 4-67 : Simplified Ensemble testing CM road intersections GB

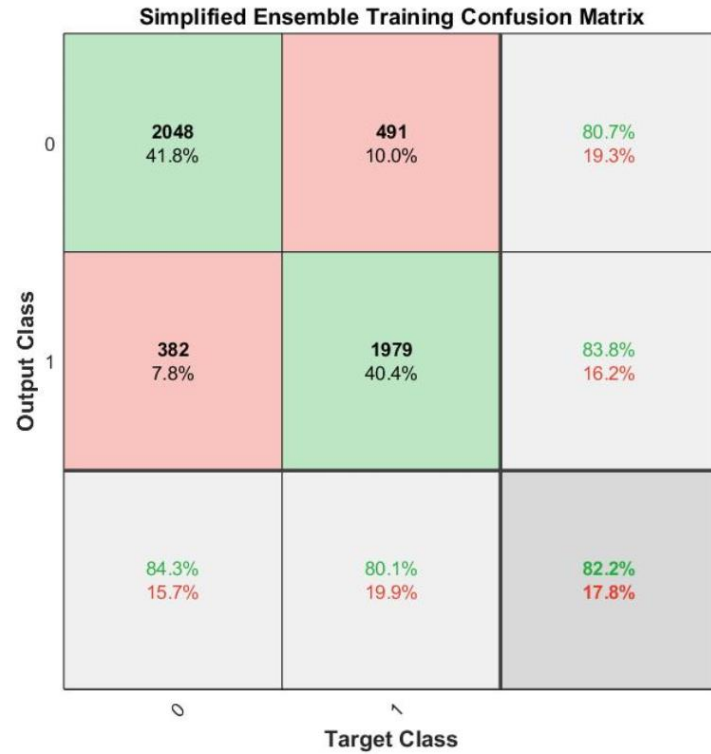


Figure 4-68 : Simplified Ensemble training CM all location Canada

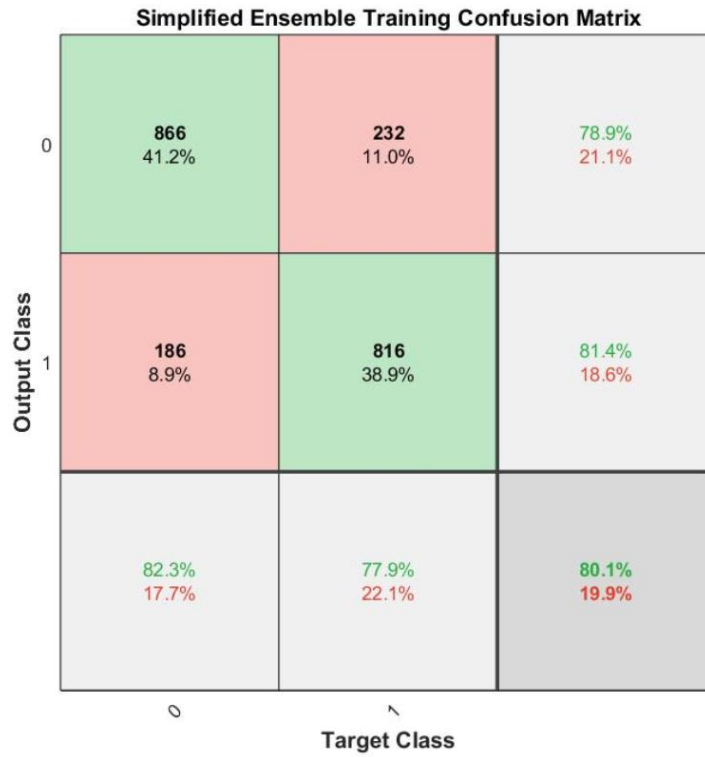


Figure 4-69 : Simplified Ensemble testing CM all locations Canada

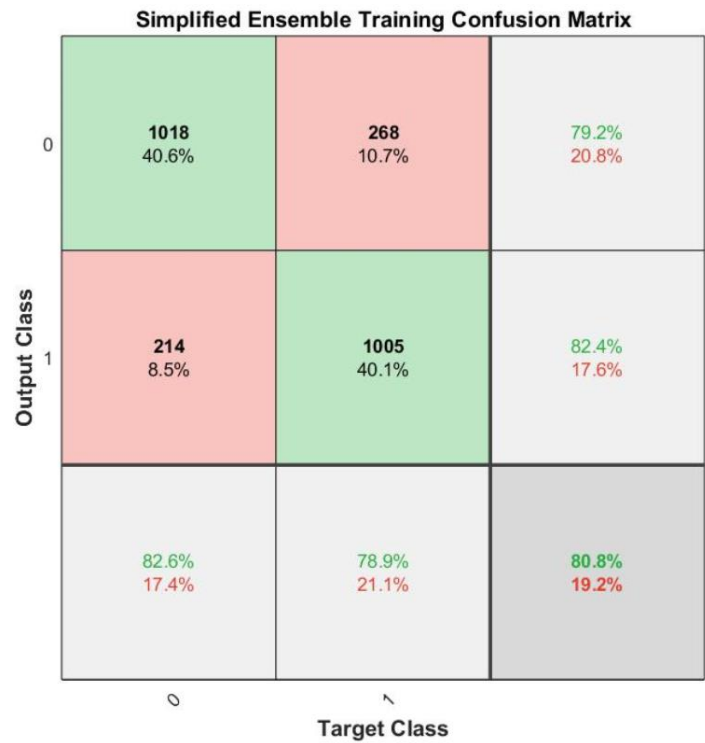


Figure 4-70 : Simplified Ensemble training CM road intersections Canada

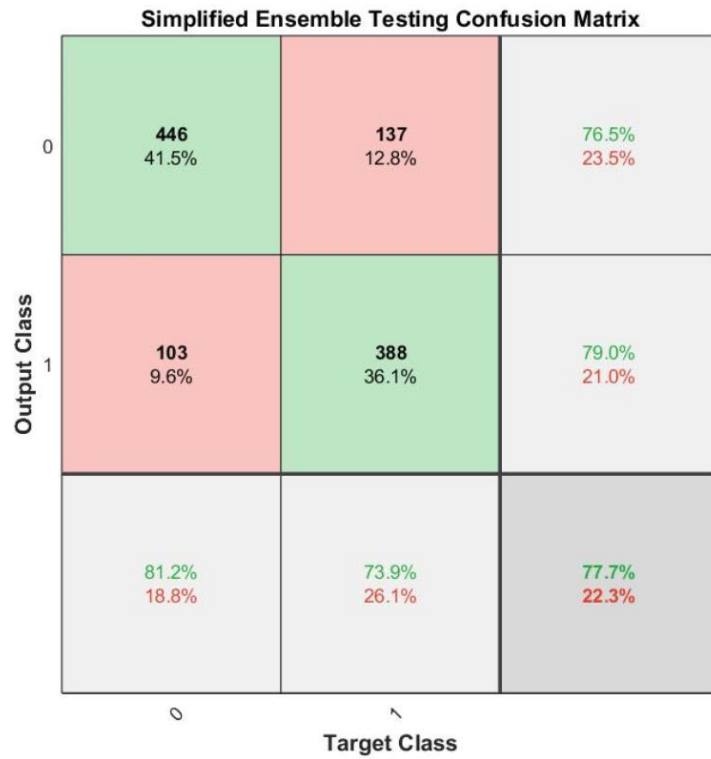


Figure 4-71 : Simplified Ensemble testing CM road intersections Canada

CHAPTER 5

CONCLUSIONS AND RECOMMENDATIONS

5.1 Conclusions

Road traffic crashes are considered a threat across the globe resulting in economic and live losses. Numerous unfortunate victims in traffic crashes don't get ideal treatment because their injury severity isn't understood at an early stage. Crash severity prediction is beneficial for trauma and emergency centers to estimate the potential damage resulting from a traffic crash and accordingly dispatch the proper emergency units to provide appropriate emergency treatment. Several statistical techniques have been used over the years for predicting crash severity. Statistical techniques although widely used, suffer from few shortcomings. These techniques require a predefined mathematical form and the presence of outliers and missing values in the dataset negatively affect the outcome of the prediction model. In contrast to the parametric techniques ML techniques can manage outliers and missing values in the dataset. Recently, several ML techniques have been employed for extracting useful information from large traffic crash datasets for different road networks.

In this research, the performance of various machine learning models was compared for predicting traffic crash severity. Based on the traffic crash data for 6 years (2011-2016) from the Department of Transport, Great Britain, five base models KNN, DT, AdaBoost, SVM, ANN, and a two-layer ensemble model was developed. The two-layer

ensemble model was developed by integrating KNN, DT, AdaBoost, SVM, and ANN in two layers. The models were compared with each other based on the accuracy of the training dataset and testing dataset. The dataset was randomly classified into training and testing dataset with a percentage of 70% and 30% respectively. Since accuracy is not always the recommended performance measure for model interpretation so three other performance indicators precision, recall, and F1 score were calculated to provide better insight into the performance of models. The models were developed for all road locations collectively and for road intersections specifically. After developing the models for Great Britain's dataset, the transferability of models was tested using the Traffic crash data for the 3-year period (2014-2016) from the National Collision Database (NCDB) Canada. The results of the models for two crash datasets were compared based on accuracy and other and F1 score.

Among the base models for both datasets, we noticed that AdaBoost outperformed all the other individual models in terms of accuracy and F1 score. KNN, on the other hand, was the least accurate model having a minimum F1 score for both the severity levels. But the proposed two-layer ensemble model resulted in significant improvement in predicting each crash severity level. The accuracy of the two-layer ensemble model for both training and testing was substantially enhanced. For all locations in Great Britain, the two-layer ensemble model showed better performance than all the base models with a testing accuracy of 78.0% and an F1 score of 0.78 and 0.77 for severe and non-severe crashes respectively. For intersections in Great Britain, the ensemble model outperformed all the base models with a testing accuracy of 76.7% and an F1 score of 0.75 and 0.77 for severe and non-severe crashes respectively.

The transferability of these models was checked on a crash dataset obtained from the online National Collision Database (NCDB) Canada. We found that the models performed in the same line as Great Britain's dataset. The ensemble model developed using the "all locations data" for Canada showed better performance than all the base models with a testing accuracy of 82.3% and an F1 score of 0.82 and 0.81 for fatal and non-fatal crashes respectively. For intersections in Canada, the ensemble model outperformed all the base models with a testing accuracy of 79.3% and an F1 score of 0.77 and 0.80 for fatal and non-fatal crashes respectively. The above results for both the datasets indicate that the models for all locations are superior and more robust in predicting crash severity compared to the models for intersections specifically. This shows that there is no requirement of the specific models for road intersections. Also, the research indicates that a high probability of accurate crash severity prediction is expected, if these models are employed at a global level for any other crash dataset.

To simplify the two-layer ensemble model, a stepwise procedure was followed for eliminating the models in the second layer. We noticed that eliminating KNN from the second layer of original ensemble model resulted in an accuracy reduction of around 1 to 2 percent for different cases compared to the original ensemble model. The elimination of more than one model from the second layer resulted in substantial performance reduction. Although the simplified model is simpler, the original ensemble model is better in accuracy and not expensive in terms of computational time. We recommend using the original ensemble model for predicting crash severity.

5.2 Future Work

This study provides a few recommendations for future research. First, sensitivity analysis can be conducted to provide complete inference on feature importance. Although the selected features for this study are based on extensive literature review but sensitivity analysis will provide a complete understanding of factors contributing to crash severity. Second, other machine learning techniques like deep learning could be considered to compare the predictive performance for crash severity. Third and most important recommendation is to apply these models on data sets for different countries of the world. It is important to note that the transferability of these models using the crash data of other countries can be achieved easily by performing minor modifications. But based on the results of this study we strongly recommend that a standard crash data collection format be used by traffic wardens across the globe. If a standard data format is followed, the transferability and validation of these models will be reasonable and easy.

References

- [1] WHO, *Global status report on road safety 2015*. World Health Organization, 2015.
- [2] T. Neuman *et al.*, "National Cooperative Highway Research Program Report 500: Guidance for Implementation of the AASHTO Strategic Highway Safety Plan, Volume 5: A Guide for Addressing Unsignalized Intersection Collisions," *Transportation Research Board, Washington, DC* <http://safety.transportation.org/guides.aspx>, 2003.
- [3] C. S. Duncan, 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*, vol. 1635, no. 1, pp. 63-71, 1998.
- [4] K. M. Kockelman and Y.-J. Kweon, "Driver injury severity: an application of ordered probit models," *Accident Analysis & Prevention*, vol. 34, no. 3, pp. 313-321, 2002.
- [5] N. A. Siddiqui, X. Chu, and M. Guttenplan, "Crossing locations, light conditions, and pedestrian injury severity," *Transportation research record*, vol. 1982, no. 1, pp. 141-149, 2006.
- [6] K. K. Yau, H.-P. Lo, and S. H. Fung, "Multiple-vehicle traffic accidents in Hong Kong," *Accident Analysis & Prevention*, vol. 38, no. 6, pp. 1157-1161, 2006.
- [7] S. S. Zajac and J. N. Ivan, "Factors influencing injury severity of motor vehicle–crossing pedestrian crashes in rural Connecticut," *Accident Analysis & Prevention*, vol. 35, no. 3, pp. 369-379, 2003.

- [8] C. O'donnell and D. Connor, "Predicting the severity of motor vehicle accident injuries using models of ordered multiple choice," *Accident Analysis & Prevention*, vol. 28, no. 6, pp. 739-753, 1996.
- [9] A. Khorashadi, D. Niemeier, V. Shankar, and F. Mannering, "Differences in rural and urban driver-injury severities in accidents involving large-trucks: an exploratory analysis," *Accident Analysis & Prevention*, vol. 37, no. 5, pp. 910-921, 2005.
- [10] P. Savolainen and F. Mannering, "Probabilistic models of motorcyclists' injury severities in single-and multi-vehicle crashes," *Accident Analysis & Prevention*, vol. 39, no. 5, pp. 955-963, 2007.
- [11] A. S. Al-Ghamdi, "Using logistic regression to estimate the influence of accident factors on accident severity," *Accident Analysis & Prevention*, vol. 34, no. 6, pp. 729-741, 2002.
- [12] D. Delen, R. Sharda, and M. Bessonov, "Identifying significant predictors of injury severity in traffic accidents using a series of artificial neural networks," *Accident Analysis & Prevention*, vol. 38, no. 3, pp. 434-444, 2006.
- [13] L. Mussone, A. Ferrari, and M. Oneta, "An analysis of urban collisions using an artificial intelligence model," *Accident Analysis & Prevention*, vol. 31, no. 6, pp. 705-718, 1999.
- [14] Z. Li, P. Liu, W. Wang, and C. Xu, "Using support vector machine models for crash injury severity analysis," *Accident Analysis & Prevention*, vol. 45, pp. 478-486, 2012.

- [15] J. de Oña, R. O. Mujalli, and F. J. Calvo, "Analysis of traffic accident injury severity on Spanish rural highways using Bayesian networks," *Accident Analysis & Prevention*, vol. 43, no. 1, pp. 402-411, 2011.
- [16] M. Simoncic, "A Bayesian network model of two-car accidents," *Journal of transportation and Statistics*, vol. 7, no. 2/3, pp. 13-25, 2004.
- [17] M. G. Karlaftis and I. Golias, "Effects of road geometry and traffic volumes on rural roadway accident rates," *Accident Analysis & Prevention*, vol. 34, no. 3, pp. 357-365, 2002.
- [18] S. Y. Sohn and H. Shin, "Pattern recognition for road traffic accident severity in Korea," *Ergonomics*, vol. 44, no. 1, pp. 107-117, 2001.
- [19] H. T. Abdelwahab 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*, vol. 1746, no. 1, pp. 6-13, 2001.
- [20] R. Buendia, S. Candefjord, H. Fagerlind, A. Bálint, and B. A. Sjöqvist, "On scene injury severity prediction (OSISP) algorithm for car occupants," *Accident Analysis & Prevention*, vol. 81, pp. 211-217, 2015.
- [21] T. Ross, "The synthesis of intelligence—its implications," *Psychological Review*, vol. 45, no. 2, p. 185, 1938.
- [22] R. Abduljabbar, H. Dia, S. Liyanage, and S. Bagloee, "Applications of artificial intelligence in transport: An overview," *Sustainability*, vol. 11, no. 1, p. 189, 2019.

- [23] C. Lee and M. Abdel-Aty, "Comprehensive analysis of vehicle–pedestrian crashes at intersections in Florida," *Accident Analysis & Prevention*, vol. 37, no. 4, pp. 775-786, 2005.
- [24] Y. Xie, Y. Zhang, and F. Liang, "Crash injury severity analysis using Bayesian ordered probit models," *Journal of Transportation Engineering*, vol. 135, no. 1, pp. 18-25, 2009.
- [25] V. Shankar and F. Mannering, "An exploratory multinomial logit analysis of single-vehicle motorcycle accident severity," *Journal of safety research*, vol. 27, no. 3, pp. 183-194, 1996.
- [26] M. A. Abdel-Aty and H. T. Abdelwahab, "Predicting injury severity levels in traffic crashes: a modeling comparison," *Journal of transportation engineering*, vol. 130, no. 2, pp. 204-210, 2004.
- [27] L.-Y. Chang and H.-W. Wang, "Analysis of traffic injury severity: An application of non-parametric classification tree techniques," *Accident Analysis & Prevention*, vol. 38, no. 5, pp. 1019-1027, 2006.
- [28] Y. Xie, D. Lord, and Y. Zhang, "Predicting motor vehicle collisions using Bayesian neural network models: An empirical analysis," *Accident Analysis & Prevention*, vol. 39, no. 5, pp. 922-933, 2007.
- [29] J. Abellán, G. López, and J. De Oña, "Analysis of traffic accident severity using decision rules via decision trees," *Expert Systems with Applications*, vol. 40, no. 15, pp. 6047-6054, 2013.

- [30] A. Iranitalab and A. Khattak, "Comparison of four statistical and machine learning methods for crash severity prediction," *Accident Analysis & Prevention*, vol. 108, pp. 27-36, 2017.
- [31] J. Zhang, Z. Li, Z. Pu, and C. Xu, "Comparing Prediction Performance for Crash Injury Severity Among Various Machine Learning and Statistical Methods," *IEEE Access*, vol. 6, pp. 60079-60087, 2018.
- [32] A. Ahmadi, A. Jahangiri, V. Berardi, and S. G. Machiani, "Crash severity analysis of rear-end crashes in California using statistical and machine learning classification methods," *Journal of Transportation Safety & Security*, pp. 1-25, 2018.
- [33] G. Singh, S. Sachdeva, and M. Pal, "Comparison of three parametric and machine learning approaches for modeling accident severity on non-urban sections of Indian highways," *Advances in transportation studies*, vol. 45, 2018.
- [34] L. Wahab and H. Jiang, "A comparative study on machine learning based algorithms for prediction of motorcycle crash severity," *PloS one*, vol. 14, no. 4, p. e0214966, 2019.
- [35] T. K. Bahiru, D. K. Singh, and E. A. Tessfaw, "Comparative Study on Data Mining Classification Algorithms for Predicting Road Traffic Accident Severity," in *2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT)*, 2018, pp. 1655-1660: IEEE.
- [36] X. Wang and S. H. Kim, "Prediction and Factor Identification for Crash Severity: Comparison of Discrete Choice and Tree-Based Models," *Transportation Research Record*, p. 0361198119844456, 2019.

- [37] L. Wahab and H. Jiang, "Severity prediction of motorcycle crashes with machine learning methods," *International Journal of Crashworthiness*, pp. 1-8, 2019.
- [38] S. Banerjee and N. K. Khadem, "Factors Influencing Injury Severity in Alcohol-Related Crashes: A Neural Network Approach Using HSIS Crash Data," *ITE Journal*, vol. 45, 2019.
- [39] S. Arhin and A. Gatiba, "Predicting Injury Severity of Angle Crashes Involving Two Vehicles at Unsignalized Intersections Using Artificial Neural Networks," *Engineering, Technology & Applied Science Research*, vol. 9, no. 2, pp. 3871-3880, 2019.
- [40] S. H.-A. Hasheminejad, M. Zahedi, and S. M. H. Hasheminejad, "A hybrid clustering and classification approach for predicting crash injury severity on rural roads," *International journal of injury control and safety promotion*, vol. 25, no. 1, pp. 85-101, 2018.
- [41] J. Tang, J. Liang, C. Han, Z. Li, and H. Huang, "Crash injury severity analysis using a two-layer Stacking framework," *Accident Analysis & Prevention*, vol. 122, pp. 226-238, 2019.
- [42] L. Jiang, Y. Xie, and T. Ren, "Modelling highly unbalanced crash injury severity data by ensemble methods and global sensitivity analysis," in *Proceedings of the Transportation Research Board 98th Annual Meeting, Washington, DC, USA*, 2019, pp. 13-17.
- [43] M. Zheng *et al.*, "Traffic Accident's Severity Prediction: A Deep-Learning Approach-Based CNN Network," *IEEE Access*, vol. 7, pp. 39897-39910, 2019.

- [44] S. Das, A. Dutta, K. Dixon, L. Minjares-Kyle, and G. Gillette, "Using deep learning in severity analysis of at-fault motorcycle rider crashes," *Transportation research record*, vol. 2672, no. 34, pp. 122-134, 2018.
- [45] L. G. Cuenca, E. Puertas, N. Aliane, and J. F. Andres, "Traffic Accidents Classification and Injury Severity Prediction," in *2018 3rd IEEE International Conference on Intelligent Transportation Engineering (ICITE)*, 2018, pp. 52-57: IEEE.
- [46] C. Elkan, "Nearest neighbor classification," 2011.
- [47] B. S. Everitt, S. Landau, M. Leese, and D. Stahl, *Miscellaneous clustering methods Cluster analysis*, 5th ed. John Wiley & Sons, Ltd, Chichester, UK, 2011, pp. 215-255.
- [48] S. Shalev-Shwartz and S. Ben-David, *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.
- [49] J. R. Quinlan, "Induction of decision trees," *Machine learning*, vol. 1, no. 1, pp. 81-106, 1986.
- [50] T. Hastie, "Robert Tibshirani, and Jerome Friedman," *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, vol. 2, 2001.
- [51] Y. Freund and R. E. Schapire, "Experiments with a new boosting algorithm," in *icml*, 1996, vol. 96, pp. 148-156: Citeseer.
- [52] C. M. Bishop, *Neural networks for pattern recognition*. Oxford university press, 1995.
- [53] S. Samarasinghe, *Neural networks for applied sciences and engineering: from fundamentals to complex pattern recognition*. Auerbach publications, 2016.

- [54] C. Cortes and V. Vapnik, "Support-vector networks Machine learning (pp. 237–297), Vol. 20," ed: Boston, MA: Kluwer Academic Publisher, 1995.
- [55] G. Mountrakis, J. Im, and C. Ogole, "Support vector machines in remote sensing: A review," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 66, no. 3, pp. 247-259, 2011.
- [56] N. Cristianini and J. Shawe-Taylor, *An introduction to support vector machines and other kernel-based learning methods*. Cambridge university press, 2000.
- [57] I. H. Witten, E. Frank, M. A. Hall, and C. J. Pal, *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, 2016.
- [58] D. H. Wolpert, "Stacked generalization," *Neural networks*, vol. 5, no. 2, pp. 241-259, 1992.
- [59] S. Valle, W. Li, and S. J. Qin, "Selection of the number of principal components: the variance of the reconstruction error criterion with a comparison to other methods," *Industrial & Engineering Chemistry Research*, vol. 38, no. 11, pp. 4389-4401, 1999.
- [60] C. Xu, Y. Wang, P. Liu, W. Wang, and J. Bao, "Quantitative risk assessment of freeway crash casualty using high-resolution traffic data," *Reliability Engineering & System Safety*, vol. 169, pp. 299-311, 2018.
- [61] A. C. Rencher, "Methods of multivariate analysis, New York: A John Wiley & Sons," *Inc. Publication*, 2002.

Vitae

- Name : UMER MANSOOR
- Nationality : PAKISTANI
- Date of Birth : 12-07-1993
- Email : umerkhan190@gmail.com
- Address : PO BOX 26000, KOHAT, PAKISTAN
- Academic Background : MASTER’S IN CIVIL ENGINEERING
- Thesis Publications: Umer Mansoor, Nedal Ratrou, Syed Masiur Rahman, Khaled Assi. “Crash Severity Prediction Using Two-layer Ensemble Machine Learning Model for the Proactive Emergency Management” (Submitted in Accident Analysis and Prevention on 08-December-2019).