# Accident Severity Prediction: Comparing ANN and Pattern search methods

MSc Research Project
MSC Data Analytics

Jiliya Mathew
Student ID: x20260075
x20260075@student.ncirl.ie

School of Computing
National College of Ireland

Supervisor:Prof. Hicham Rifai

**National College of Ireland**

**MSc Project Submission Sheet**

**School of Computing**

| | |
|---|---|
| **Student Name:** | ……………………Jiliya Mathew ………………………………………………… |
| **Student ID:** | ……………………x20260075……………………………………………………………………… ………………..…… |
| **Programme:** | ………Data Analytics……………………………… **Year:** 2021-2022 |
| **Module:** | …………MSC Research Project………………………………… |
| **Supervisor:** | ………………Hicham Rifai………………………………………………………… |
| **Submission Due Date:** | …………15/08/2022……………………………………………………………… |
| **Project Title:** | ………Accident Severity Prediction: Comparing ANN and pattern searching methods …………… |
| **Word Count:** | 6978 ……………………………………… |

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use theReferencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

**Signature:** …………Jiliya Mathew ……………………………………………………………………………………………

**Date:** ……15/08/2022…………………………………………………………………………………………

**PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST**

| | |
|---|---|
| Attach a completed copy of this sheet to each project (including multiple copies) | □ |
| **Attach a Moodle submission receipt of the online project submission,** to each project (including multiple copies). | □ |
| **You must ensure that you retain a HARD COPY of the project**, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer. | □ |

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

| **Office Use Only** | |
|---|---|
| Signature: | |
| Date: | |
| Penalty Applied (if applicable): | |

# Accident Severity Prediction: Comparing ANN and Pattern search methods

Jiliya Mathew
x20260075

**Abstract**

Vehicle accidents cost the United States thousands of billions of dollars in both monetary and cultural terms every year. A limited number of major incidents are responsible for a large part of the damages. However, preventing road collisions, particularly major accidents, is an ongoing concern. The study's long-term objective has been to a better understanding of the elements that influence the likelihood of a traffic fatality. Accident and severity prediction is critical for the effective deployment of this technique. We may be able to apply systematic responses and improved safe drivingregulations if we can discover the tendencies and hidden layers of how these massive catastrophes occur. The goal of this study was to evaluate the magnitude of road crashes utilizing accident-related factors in pattern search and ANN modeling methodologies, as well as to investigate the impact of weather or other environmental factors on accident occurrence. The most current dataset release may be useful for investigating how COVID- 19 influences accidents and driving habits. The project's goal has been to gain a greater understanding of the factors that influence the likelihood of a highway deadly accident. For feature selection, the particle swarm technique, Gray wolf optimization, Mutation Architecture, and Weight initialization GA are utilized. Classification techniques such as ANN, Boosting, pattern search deep neural network, and Bagging are defined and assessed.

**Keywords:** ANN. Pattern Search Methods, Boosting, Bagging, Accident Severity Prediction, particle swarm approach, Gray wolf optimization, Mutation Architecture, Weight initialization GA

# 1 Introduction

## 1.1 Motivation and Project Background

Highway safety has long been a primary concern in the design of efficient transportation networks. Road accidents will affect people by causing fatalities, congested roads, and contamination, none of which are beneficial for the long-term viability and very well of the transit system. In recent times, as the level of automation in information systems has constantly risen, various public bodies and transportation sector corporations have been devoted to the development of automated driving to encourage the sustainable expansion of transportation. Road accident forecasting is a large and difficult issue in the field of efficient traffic safety control systems; it is crucial for calculating the growing trend of automobile accidents and implementing preventive pre-emptive actions under road network congestion conditions. It is critical to investigate feasible and timely solutions that take into account the gravity of traffic occurrences in order to enhance safe driving management and control. With recent technical and scientific advances, smart mobility technology has reached new heights. Unfortunately, these advanced technologies offer no discernible benefits in terms of accident reduction. HELP LIVES-A Road traffic accidents study says about 1.2 million people died and injure up to 50 million people each year, according to a World Health Organization (WHO) road safety technical package released in 2017. In all age categories, road traffic accidents constitute the seventh leading cause of mortality. Although traffic accidents are common, they may be avoided and handled. As a result, it is each traffic author's responsibility to investigate the reasons for traffic accidents and assist the administration in resolving the issue of reducing the likelihood of automobile accidents. Researchers have evaluated several traffic accident severity analysis models from diverse perspectives throughout the years. The primary goal of this studyis to evaluate several models and select the most precise one for forecasting road accident severity utilizing specified parameters and three degrees of damage severity as dependent and independent variables. This research looks towards modeling approaches that might be used to obtain high expected accuracy. In our research, artificial neural networks were able to capture very nonlinear correlations between the predictor variables crash causes and the target variable injury severity level. This neural network characteristic is especially beneficial when the link between variables is ambiguous or complicated, making statistical analysis difficult. Several studies have demonstrated that neural networks are superior at predicting accident severity. The correlation-based approach attempts to identify effective ways fusion strategies to revisit, reuse, and maintain matching equivalent patterns that occur previously formed in pre-accident, while in an accident, and after the accident pattern sets by picking valuable pattern sets from the dataset. This will allow highway officials to apply road/driver rehabilitation measures whilealso increasing the efficiency of drivers involved in an accident. As a result, fewer traffic injuries are expected in the future.

## 1.2 Research Question

This study is based on the following factors:

- How well can the ANN and pattern matching approaches forecast the severity of an accident?
- Does precipitation or other environmental stimuli effects accident severity?
- Enhance ANN, Boosting, pattern search deep neural network and Bagging algorithms, comparing and critically assessing the outcomes
- How does COVID-19 affect accidents and driving habits?

ANN, Boosting, and Bagging algorithms are used for predicting accident severity. For feature selection, the particle swarm method, Gray wolf optimization, Mutation Architecture, andWeight initialization GA are utilized. Classification techniques such as ANN, Boosting, pattern search deep neural network, and Bagging are defined and assessed. Furthermore, each model is compared to others and critically evaluated. These models are then compared to those that came before them. The goal of this research is to develop a model that can predict the severity of an accident and analyze the impact of precipitation or other environmental stimuli on accident occurrence.

# 2 Related Work

## 2.1 Prediction of accident severity

This part will look at the important studies that have been conducted in the field of accident severity prediction. First, explore the numerous accident severity prediction efforts implemented by different researchers throughout the world. A rigorous review of the comparative work done on machine learning's accident severity prediction follows. Lastly, this will address earlier studies that provided strategies for predicting statistics and accident severity.

(Shangguan et al. (2021)), In order to adequately cope with this challenge, they created a better estimating approach in this study. Using a map coding approach, the information is first translated into sets of data. An enhanced Discrete Synthetic Minority Resampling Method, as well as event samples near to fragmentary values, are presented. Using a convolution neural network, the traffic accident technique was created to predict the severity of an accident. Finally, the solution's efficacy is validated using real-world traffic incidents. The test findings show that the D-SMOTE approach is precise, meaning that it can help with car selections.

(Zheng et al. Ming Zheng and Tong Li (2019) developed an innovative vehicle accident importance forecast neural network model for vehicle accident severity prediction that leverages connections among injury crash data to increase efficiency. The grayscale image approach is given for converting one link of deadly crash information into grayscale images with connections as independent components for a model focused on the weights of road accident characteristics. The suggested strategy for anticipating the intensity of auto-mobile events also excels in tests.

(Li & Jiang (2020))Fangyuan Li and Kun Jiang did studies on the potential contributing elements of 366 important occurrences to study the variables influencing the severity of major car crash in China. To construct algorithms for event data, the fixed-parameter and random-parameter negative binomial approaches were chosen. According to the data, geographical location, accident time of day, road design, climate, and accident type are all related to the severity of the accident. A driver-control system is a great way tolessen the degree of crash damage.

(Fan & Qian (2017))Yu-Bin conducted research. This paper examined data from 261 automotive incidents at the Songjiang District intersection in Shanghai from 2008 to 2016. 80 percent of the data were randomly selected for modeling, and 21 percent of the rest of the data was forecasted. The model of car collision severity evaluation was evaluated using the test for good fit and the parallel line test, and the results show that the approach has abetter fitting effect.

(Shi & Deng (2019))The severity of road accidents was selected as the goal characteristic in this study by Chenjun Shi and YuanChang Deng. The mathematical methodof the interstate road accident site was built using data from over 1300 traffic occurrences on Guangdong urban expressways between January 2018 and August 2018. Based on the results, the facts fit the model nicely. The frequency of automotive accidents was strongly related to the following factors: vehicle type, vehicle safety, and roadway design. (Makarova et al. (2020))Irina Makarova and Polina Buyvol used facts from car crashes in the typical Elabuga town from 2017 to 2018 as a starting place for their research. To differentiate between various types of occurrences, decision trees were utilized. According to the data, different types of occurrences result in a variety of severity levels. Walking and bicycle accidents are the most common types of situations that result in injury. The article emphasizes the most commonly violated traffic laws and crashes resulting from such violations resulted in fatal or serious injury.

(Iveta et al. (2021))Mateja Iveta, Alexander Radovan, and Branko Mihaljevi show how to apply deep learning to predict the effect and likelihood of an accident. Wind speeds, the availability of light, the time of day, and speed limits were among the features observed. Given the context and road circumstances, a multi-classification algorithm is employed to evaluate the severity of the event. The forecast's findings demonstrate a relationship between climatic variables, daylight hours, and the severity of traffic accidents.

(Chong (2004))Xiao Li and Yihui Huang conducted a study and came to the following results: The most key qualities that determine the fatality rate in a vehicle accident are age range, the transmission of accountable automobiles, alcohol concentration, and sight. Because of the diversity of roads, numerous traffic accidents could happen for a range of reasons, making it challenging for our model to lower a mortality rule.

(Parseh et al. (2021)) This research is based on the successful operation of a truly automated vehicle in traffic situations when accidents with other vehicles are possible. A malfunctioning vehicle, delayed barrier recognition, or the existence of an aggressive motorist might all contribute to a potentially dangerous situation. We propose a method for the vehicle'smonitoring system to select the movement that is most likely to cause the least significantharm to the passengers. The technique entails addressing a genuine management issue offline in order to build a number of paths by controlling the wheel speed and halting frequencies within the restrictions of the vehicle.

## 2.2 Comparison of accident severity Prediction Algorithms

(Manzoor et al. (2021),)Mubariz Manzoor and Khwaja Fareed conducted studies. RFCNN models, which mix RF and Convolutional Neural networks, were used to anticipate the severity of car crashes. To determine how well the proposed technique works, it is evaluated to a variety of base learning models. The study made use of accident data from February 2016 to June 2020. The RFCNN outperformed competing theories in this study, with 0.991 accuracies, 0.974 precision, 0.986 recall, and 0.980 F-score using the more thannineteen most significant parameters in predicting collision magnitude.

(Ali & Baiee (2021))The main goal of the research is to employ feature ranking al-algorithms to find a feature subset from the Queensland roads dataset that can be identified by a functionality based on specified qualities. An ANN Feed Forward classifier was used to detect and predict accident rates in automotive crash cases. By comparing the ratings of various FS techniques that employ a focused search. The results show that ANN classification produced extremely precise findings for the car accident dataset.

(Chong (2004)) Bing Nan and Cui Xu investigated the utility of real-time collecting and analyzing data using data mining approaches in this article. They examine and contrast the effectiveness of various classifications. The findings demonstrate that data processing enhances classification performance, and the model might be used to forecast and possibly avoid car crashes. In their work, they used data mining to evaluate Washington road accident records. They focused on real data acquisition, such as current weather situations, and used data analytics techniques to accomplish so. These characteristics might be acquired in real-time and utilized to develop a road collision detection system.

(Labib et al. (2019))Farhan Labib and Ahmed Sady Rifat carried out this research in Bangladesh to evaluate transportation events in greater detail in order to identify the extent of wrecks using data mining approaches. They also emphasize crucialelements that have a direct impact on road accidents and provide some practical suggestions. The intensity of occurrences was classified as Fatal, Serious, Simple Injury, and Vehicle Accident using the Naive Bayes, KNN, AdaBoost, and Decision Tree classifiers. In the end, AdaBoost triumphs.

(Bahiru et al. (2018))Dheeraj Kumar Singh and Tadesse Kebede Bahiru Using previously gathered traffic data, supervised learning techniques were utilized to develop models (classifiers) to identify incident characteristics and predict fatal crash intensity. WEKA is used to develop a data mining decision tree using Nave Bayes classifiers to assess the amount of damage. The efficiency of categorization of all of theseapproaches is compared based on the results. Based on the study's findings, the J48 classifier outperforms all others in terms of accuracy.

(Tambouratzis et al. (2010))A mix of probabilistic neural networks and DT was proposed as a technique for correctly and quickly anticipating accident intensity using the study collision data obtained by the Republic of Cyprus Police in 2005. The following is a combination of the direct, gradual, and effective growth of the probabilistic network andthe completely built data structure partitioning of the decision tree: A range of PNNs have been trained to utilize data partitions produced from the DT's minimum number of highest nodes. The DNN technique outperforms the decision tree in prediction ac- curacy.

(Luo et al. (2021))This study investigates criteria using a fuzzy deep learning algorithmin order to address the issue of road collision estimates. Weather, eyesight, air movement, and moisture were selected as separate factors for modeling, whereas the goal variable was traffic accident incidence. After tests and assessments, this technique can achieve a forecasting accuracy of roughly 78 percent, which is greater than artificial neural networks and k-nearest-neighbor in data mining techniques.

## 2.3 Accident severity prediction performance Analysis

(Assi (2020)) This research approach's aim is to examine how effectively DNN can predict the intensity of an automobile collision utilizing variables that can be obtained quickly on the site. The DNN model's effectiveness was compared to that of the SVM model, which is commonly used to predict the intricacy of car crashes. In terms of predicting RTC severity, DNN surpasses SVM, with predictive accuracy and F1 scores of 95 and 93 %, respectively.

(Beryl Princess et al. (2020))The accident shot was judged as significant data in the research. The performance of the methods is measured using the precision, area under the curve, and F1-score. In comparison to all other models, RF has a high precision rate with an AUC of 0.75. In regards to accuracy rate, the analysis reveals that hybrid features surpass solo aspects.

(Geyik & Kara (2020))Buket Geyik and Medine Kara used the Stats19 dataset in their work, which provides data on car crashes in the United Kingdom from 2010 to 2012. The dataset was divided into three intensities categories: deadly, severe, and smallevents. Based on the statistics, the DT approach is 80.74 percent accurate the NB algorithm is 83.40 percent efficient, the r RF classifier is 85.19 percent efficient, and the MLP model is 86.67 percent accurate.

(Manzoor et al. (2021))In this study, Mubariz Manzoor, Muhammad, and Umer's RF research was utilized to uncover critical factors that are significantly related to the frequency of road collisions. The severity of a crash is influenced by duration, temperature, freezing rain, moisture, sight, and prevailing winds. The study examined incident data from February 2016 to June 2020in the USA. Using the 20 most relevant features, the RFCNN enhanced the judgment method and outperformed earlier models in predicting incident severity with 0.980 F- score, 0.986 recall, and 0.991levels of accuracy,

(Shen & Wei (2020))The purpose of this paper, by Xiaoyan Shen and Shanshan Wei, is to investigate dangerous drug transport accident information from seven Chinese provinces using the XGBoost approach. Due to the rarity of these situations, the categorization performance of different techniques is evaluated using accuracy, recall, F-score, and Area Under Curve (AUC). The results reveal that the proposed XGBoost technique outperforms all others in terms of predicting.

# 3 Research Methodology

## 3.1 Dataset

The 2.8 million accident reports in this collection are samples of the traffic accident data we want to investigate). It depicts the locations and specifics of crashes that occur in the United States. Both the United States Police Department and the Washington State Department of Transportation provide information on all collisions that Traffic Records has been collecting for the past 10 years. The graph depicts the number of road accidents each month from 2016 through 2021. The broken lines show that the number in the first half of the year is frequently lower than the number in the second half. As a result, the traffic collision must have certain undiscovered tendencies. Address type, weather, incident time, collision type, and other criteria are utilized to define the data. This database of vehicle crashes in the United States includes data from all 50 USA states. From February 2016 to December 2021, many APIs which provide real-time traffic collision data were employed to gather traffic incidents. Only a few organizations publish traffic data via these APIs, including US and state transportation authorities, speed cameras, and traffic sensors embedded in road networks. This collection currently has 2.8 million incident records. Moreinformation about this dataset may be found here.

## 3.2 Use Case

Numerous uses of US Accidents include the analysis of casualties and the extraction of cause and effect principles to anticipate vehicle accidents, as well as the investigation of the effects of precipitation or other environmental stimuli on accident incidence. The mostcurrent dataset release might be helpful for researching how COVID-19 affects accidents and driving habits.

## 3.3 Exploratory Data Analysis

### 3.3.1 Pre processing

The dataset includes data categories such as traffic characteristics, residence attributes, climatic attributes, site of interest attributes, and hour of day attributes. Dropping NA values, datatype correction, identifying missing values, and replacing them with suitable ways are all part of data preparation. Data cleaning increases the rate of forecast accuracy.

**Normalization**

Normalization steps include sizing, mapping, and pre-processing. We can derive a new set from an original range. It might be incredibly useful for predicting or prediction goals. As we all know, there are various strategies for predicting or forecasting, but they can all differ greatly. As a result, the Normalization technique is required to make forecasting and predicting similar while retaining their significant variation and entropic reasons.

**Label Encoding**

Encoding is a system control step for algorithmic machine learning when dealing with categorical attributes. Some methods can operate effectively on categorical data. For example, descriptive analysis may be used straight to build a decision tree without the necessity for system integration. Some machine learning algorithms are not able to connect properly with label data. They require that all input and output variables be numerical. In general, this places more constraints on the successful application of machine learning algorithms rather than on the methods themselves.

### 3.3.2 Geographic Analysis

Geographic is the foundation of the pre-processing carried performed in this study. It is fairly demonstrated that the geographical situation, regional rule of law, and environmental backdrop discuss and connect the kind and severity of accidents that occur.

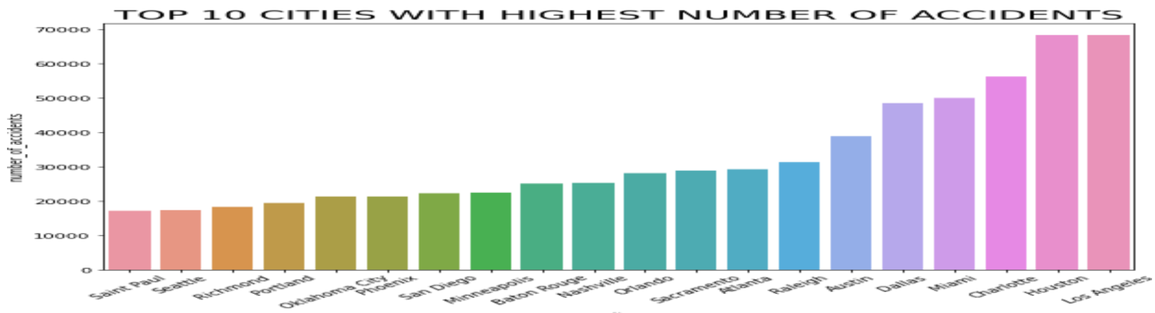



Fig 1. Top 10 cities with highest number of accidents

Top ten cities with the highest rate of crashes in the recent decade. The US geography horizon map also shows the rise in fatalities as you go from east to west. This is partly due to the density of people with recipes in the West being significantly higher than in the eastern horizon areas of the United States. As a result, it will now make contributions to the density of population as well as city accident counts per 1000 people.

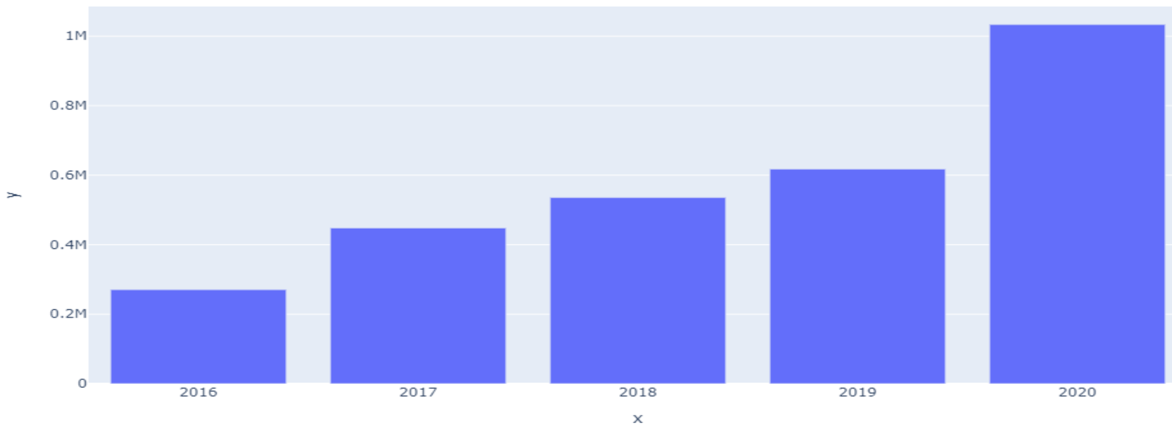



Fig 2 .Density of population as well as city accident counts per 1000 people

This is obvious in the above and below figures that the rise in collisions is in direct proportion to the year of study, as in the previous year from 2016 to 20 we see a linear trend of accident inflation, but in the previous year potential due to the outbreak of covid 19, rules and restrictions to stay at home, panic and dear among citizens to stock up homes and also post covid area, what about some reckless road rage affecting thelinear tendency of increase in the previous year As a future study of this research, it would be quite necessary to comprehend and link the rise in road range in the year 2021.

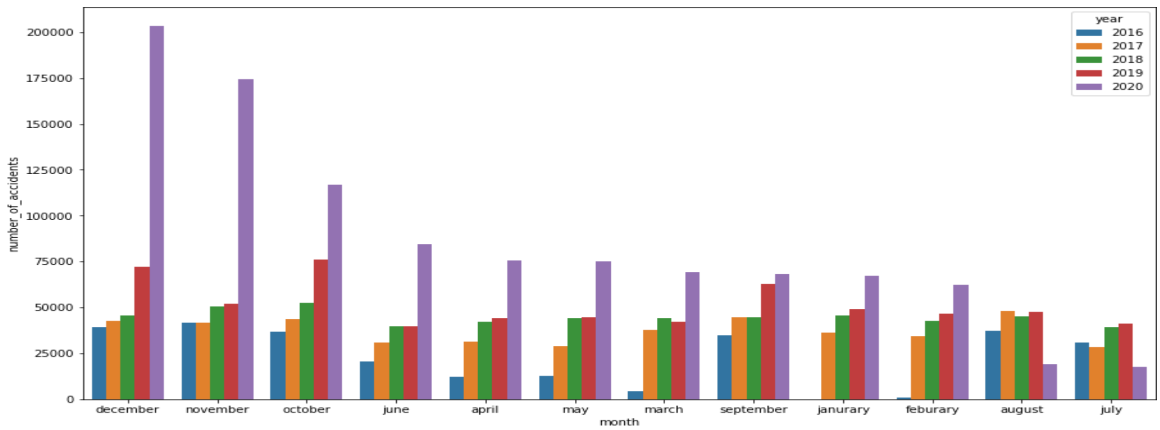



Fig 3. Number of accidents in various months from 2016 to 2020

### 3.3.3 Environmental Factors Analysis

The climate has a direct impact on the amount of road rage incidents and the severity of accidents since driving becomes more difficult when temperature, humidity, and other geographical conditions rise. Furthermore, visibility might diminish exponentially, reducing the care taken when driving safely and raising the chance of carcrashes

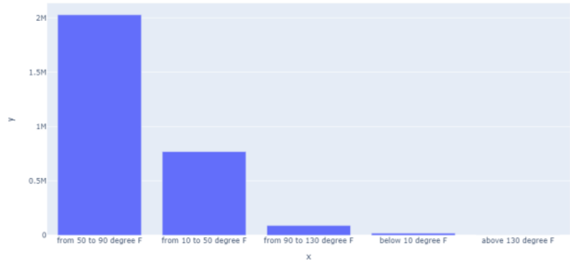



Fig4.Number Accidents with various temperature range

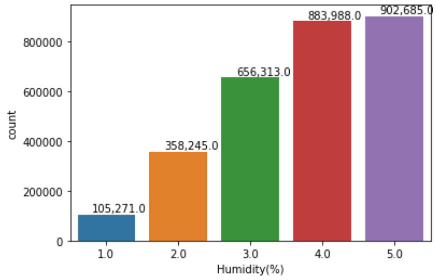
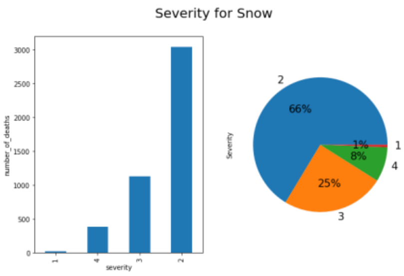
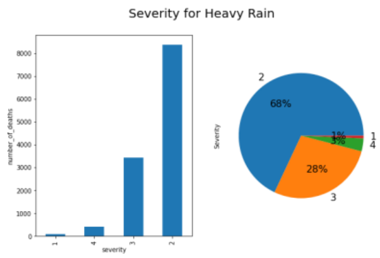



Fig 5 .Number Accidents with various humidity range

The data set's potential distribution is all biased towards the second degree of severity; however, the distribution changes depending on the weather conditions; for example, in snow, 66 of the probability is of the second degree of severity, which climbs to 68in heavy rain and 83in fog. The third degree of most serious accidents drops to roughly 12in fog, 25in snow, and 28in heavy rain. This also affected the fact that in heavy rain, the roads become slick, and the unpredictability of rainy hours not only reduces visibility, but also affects the driver's psyche and lack of preparedness when driving. However, because snow usually falls at night, precautions may be taken before venturingout. The severity the index for category 3 is reduced in slow rather than rain. In snow, the largest incidence of accidents occurs due to a very generous explanation of no or extremely low visibility mixed with the lack of roadblocks.



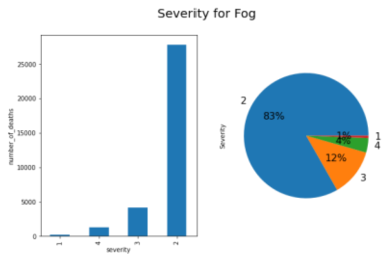




Fig5:Number of deaths with the severity of snow Fig6:Number of deaths with the severity of rain Fig7:Number of deaths with the severity of fog

# 4 Design Specification and Feature Selection

## 4.1 Particle Swarm algorithm

The particle swarm optimization (PSO) method is a demographical search strategy that is focused on simulating the behavioral patterns of birds inside a swarm. The swarm-based concept's main purpose was to find patterns that control a bird's ability to fly synchronously and quickly change direction by reorganizing in an optimum configuration, as well as to visually recreate the graceful and surprising choreography of a bird flock. From this initial premise, the idea evolved into a simple and effective optimization technique. These approaches typically assess the relevance of each item individually, ignoring the intricate interconnections between the elements. These feature selection approaches are far too simplistic to guarantee the quality of the resulting feature subsets. To address these issues, this work will investigate an effective feature selection strategy for boosting the effectiveness of the accident severity analysis model.

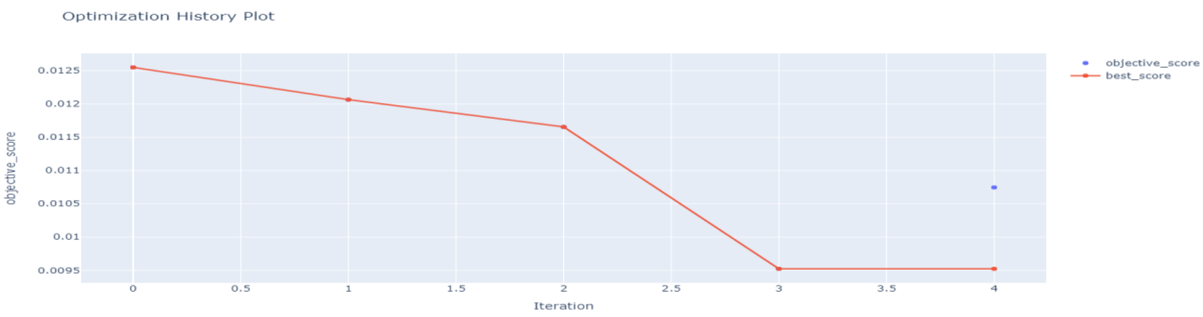



Fig 8 :Optimization history plot PSO

## 4.2 Gray wolf Optimisation

The grey wolf is a member of the Canidae family (Canis lupus). Grey wolves are considered apex predators because they are at the top of the food chain. Grey wolves prefer to live in packs. Their incredibly rigorous global structure and leadership structure are especially attractive. The GWO algorithm simulates the hunting tactics and social structure of grey wolves in the wild. Four different types of grey wolves are utilized to represent the leadership hierarchy, including alpha, beta, delta, and omega. The three major hunting processes looking for prey, encircling prey, and attacking prey are also utilized to optimize. The approach delivers excellent accuracy in classification while using a smaller selection of features and using less time. The search space of the large-scale feature selection issue is considerably reduced by this technique. By narrowing the search area to promising areas

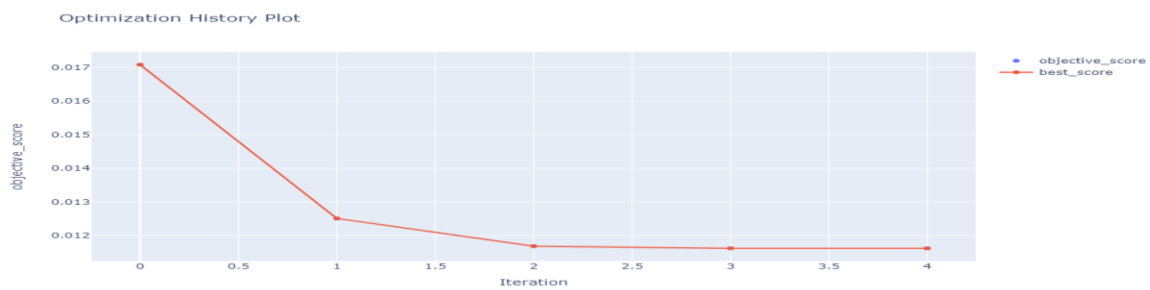



Fig 8 :Optimization history plot GWO

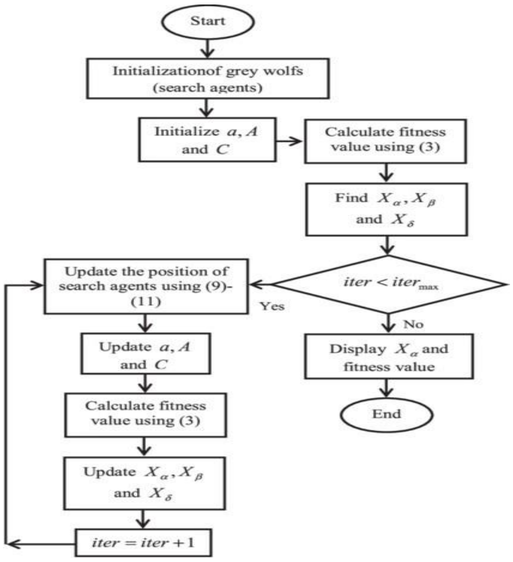



Fig 9:GWO algorithm

### 4.3 Mutation Architecture and Weight initialisation GA

Evolutionary operators are search-based optimization algorithms that discover perfect or almost ideal solutions to search and optimization problems. Optimization issues are defined by maximization or minimization of the objective function. The evolutionary algorithm aims to identify the best or nearly optimal solution to the optimization problem. The genetic algorithm is an optimization approach that makes use of randomness. It is based on Darwin's theory of evolution, which states that all species must follow the "survival of the fittest" concept. It randomly adjusts current solutions in order to discover a solution in order to build new, superior ones. To get a satisfying outcome, evolutionary operators compute the goal functions repeatedly. In other words, the GA development proceeds as follows:

1) Population initialization: The population in GA is simply the chromosomes themselves. When GA is used, it checks a population of 20 or 100 or 5,000 chromosomes at the same time, rather than simply one chromosome at a time. The GA'sinitial values were picked at random from a range of 0 to 255. Bias values between -128and 128 were chosen at random to allow for advancement in either direction.

2) Evaluation: The cost function of a chromosome, also known as the error function

# 5 Implementation

### 5.1 Artificial Neural Network (ANN):

The ANN model, which is a feasible replacement for negative binomial regression, is used to estimate the number of accidents. Another argument for selecting these input variables is comparing the model prediction accuracy of the negative binomial regression and the ANN models, as well as the marginal impactsof each independent variable. Artificial neural networks were able to capture exceedingly nonlinear connections between accident reasons and injury severity level predictor factors. This neural network feature is particularly useful when the relationship between variables is vague or convoluted, making statistical analysis difficult. Several studies have shown that neural networks are better than humans in forecasting accident seriousness.

After the dataset has been created and saved, the selected Programming Languages will be used to acquire summary analysis of the dataset, data visualization, and Advanced Analytic Technique training. After that, the input will be preprocessed before being utilized to generate the ANN. Numpy and Pandas will be utilized in Python for data processing and analysis since Numpy is a descriptive statistical package and the Pandas library provides high-level data structures and algorithms to aid in the data collection and analysis. Because of its elevated graphing capabilities, the Matplotlib package was chosen. Additionally, the Seaborn library, which is built on Matplotlib and offers interesting charting options for analytical graphs, will be used. The dataset will be divided into two sections: a training set and a test set, with the training, set counting for 80percent of the total and the test set representing for 20%. In Python, the sci-kit-learn package was used to split and size data. The library contains implementations of all well-known Machine Learning Algorithms, as well as the data preparation and technique performance evaluation routines. This libraryis used to split, analyze, and calculate the confusion matrix of the implemented Advanced Analytics Technique.

## 5.2 Pattern Search Deep Neural Network(PSDNN) :

Pattern recognition is a critical component of deep neural network and motor learning, allowing the model to assume feature rich mappings between the linked target or dependant variable. Pattern search deep neural networks, or PSDNN, can be used with generative networks. In contrast to machine learning algorithms that employ ephemeral pattern search, the PSDNN al- lows backpropagation and bias setting to synchronize with the supplied waveforms of pattern distribution, allowing the model to establish a harmonic distribution of weights and eventually produce correct results. The DNN may be taught with a large number of training images due to the conditional generative adversarial network architecture. It demonstrated accurate categorization, dependable item identification, and good object recognition in extremely cluttered environments. While thresholding techniques can more efficiently generate classification masks, other DNN designs aimed at semantic segmentation typically ignore essential object attributes. Deep convolutional neural networks, specifically conditional generative adversarial networks, dramatically improved image enhancement, classification, and severity detection accuracy. Deep Neural Networks (DNNs) have been proved to be aviable choice for incorporating into self-driving vehicles, telephones, games, drones, and other elements of daily life. To accelerate DNNs, servers with several computational engines were frequently employed. The rectified linear unit activation function is employed by the DNN classifier’s layers of neurons to conduct classification tasks.

## 5.3 Boosting Techniques

### 5.3.1 XGboost

Drivers are the most vulnerable road users, and they are more likely to be killed in accidents. XGBoost is also used to measure unintentional damage levels and examine how different variables impact the severity of accident injuries. In this work, the benefit of XGBoost is that decision tree-based learning algorithms have no problems with numerical encoding of continuous variables.

### 5.3.2 Adaboost

Adaboost, or adaptive boosting method, is an important step in ensemble methods that uses adaptive learning rate to increase or decrease the potential understanding of difficult samples, increasing by preventing overfitting or necessary/intentional biasing to increase and formulate a robust modeling architecture.

### 5.3.3 GradientBoost

Unlike the adaptive boosting method, gradient boosting employs a stochastic greedy approach in conjunction with stochastic gradient descent, allowing for the generation of dimensional stratification and training notification of problematic data using postulation sampling approaches. The approach might generate self biases that prevented fitting and the acceleration of the same data by a range of weak learners. Gradient boosting may now employ a number of designs, such as xgboost, lightboost, and catboost networks.

## 5.4 Bagging Techniques

### 5.4.1. Random Forest

Random forest frames the whole dataset as a training sample blog that can be used to train multiple tiny and quicker decision tree classifiers before assimilating an aggregated training output. Although this strategy has the potential to declassify the data set, it opens the door to quick learning models and improved variance models. Not only is the design capable of applying entropy-based random forest architecture, but it is also capable of dispersing the log entropy of cross numerous models, allowing dimensional discussion and reduction for rapid training and easy classification viewpoint.

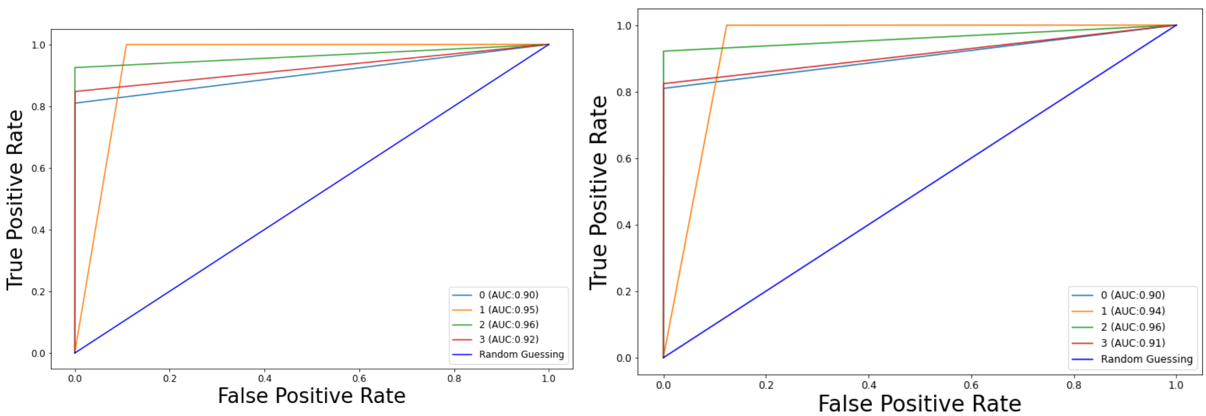
# 6 Evaluation

The findings are divided into three sections: gravel optimization-based machine learning and deep learning architectures, particle swampers optimization-based machine and deep learning results, and whole data set results. A comparison of accuracy, precision-recall, and computation time across all three features election techniques, improved display of findings, and selection of the best performing model.

GWO based Results

| **Model** | **TN** | **FP** | **FN** | **TP** | **Accuracy** | **Precision** | **Recall or Sensitivity** | **F1 Score** | **Specificity** |
|---|---|---|---|---|---|---|---|---|---|
| **XGBClassifier** | 19389 | 16 | 65 | 361 | 0.997378 | 0.97009 | 0.895287 | 0.930332 | 0.972493 |
| **ANN Network** | 19397 | 8 | 75 | 351 | 0.997227 | 0.950507 | 0.888667 | 0.917604 | 0.968906 |
| DecisionTreeClassifier | 19328 | 77 | 62 | 364 | 0.995008 | 0.839218 | 0.893365 | 0.863412 | 0.971286 |
| GradientBoostingClassifier | 19384 | 21 | 309 | 117 | 0.986587 | 0.88984 | 0.597278 | 0.684379 | 0.837417 |
| GaussianNB | 17003 | 2402 | 251 | 175 | 0.850613 | 0.273127 | 0.537984 | 0.246903 | 0.853654 |
| AdaBoostClassifier | 9253 | 10152 | 122 | 304 | 0.73254 | 0.333594 | 0.343816 | 0.216339 | 0.743996 |

XgBoost and artificial network showcase the hegemony over the other deemed algorithms.




(a) XGBoost ROC PLOT (b) ANN ROC PLOT

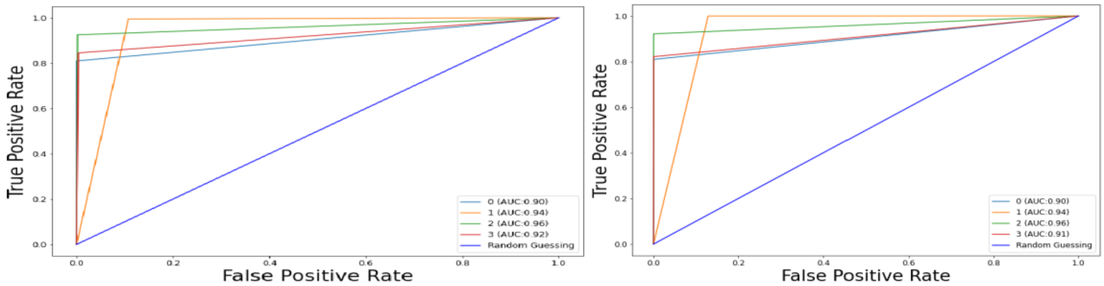
Fig 11: GWO top 2 ROC Plots

PSO based Results

| **Model** | **TN** | **FP** | **FN** | **TP** | **Accuracy** | **Precision** | **Recall or Sensitivity** | **F1 Score** | **Specificity** |
|---|---|---|---|---|---|---|---|---|---|
| **ANN Network** | 19397 | 8 | 76 | 350 | 0.997151 | 0.973204 | 0.888054 | 0.927445 | 0.96759 |
| **DecisionTreeClassifier** | 19327 | 78 | 66 | 360 | 0.994882 | 0.885281 | 0.893416 | 0.889001 | 0.97154 |
| AdaBoostClassifier | 10183 | 9222 | 177 | 249 | 0.754702 | 0.356902 | 0.33888 | 0.244275 | 0.735532 |
| XGBClassifier | 19390 | 15 | 72 | 354 | 0.997075 | 0.966212 | 0.89031 | 0.925683 | 0.969786 |
| GradientBoostingClassifier | 19386 | 19 | 310 | 116 | 0.986385 | 0.905546 | 0.598844 | 0.687485 | 0.833484 |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| GaussianNB | 18303 | 1102 | 322 | 104 | 0.942943 | 0.278475 | 0.449298 | 0.286655 | 0.811738 |

In the selection discussion artificial network and decision tree proved to have the best results in the compatitive analysis with the rest



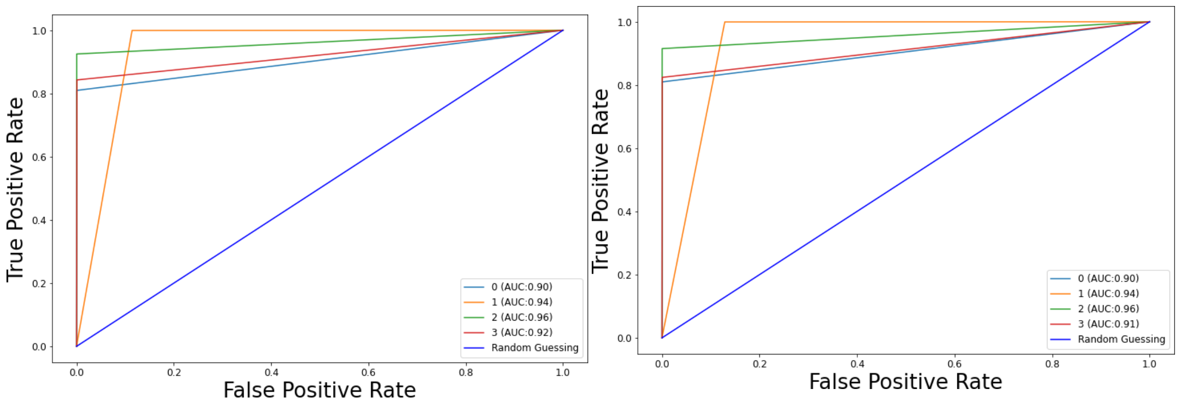

(a) Decision tree (b) ANN network

Fig 12: Top 2 ROC Plots

Full DS based Results

| **Model** | **TN** | **FP** | **FN** | **TP** | **Accuracy** | **Precision** | **Recall or Sensitivity** | **F1 Score** | **Specificity** |
|---|---|---|---|---|---|---|---|---|---|
| **ANN Network** | 19394 | 11 | 75 | 351 | 0.99705 | 0.958725 | 0.886968 | 0.920591 | 0.967539 |
| **DecisionTreeClassifier** | 19338 | 67 | 60 | 366 | 0.995285 | 0.858998 | 0.895461 | 0.876302 | 0.973023 |
| AdaBoostClassifier | 9253 | 10152 | 122 | 304 | 0.73254 | 0.333594 | 0.343816 | 0.216339 | 0.743996 |
| XGBClassifier | 19390 | 15 | 67 | 359 | 0.997277 | 0.869771 | 0.894087 | 0.829485 | 0.871165 |
| GradientBoostingClassifier | 19383 | 22 | 309 | 117 | 0.986738 | 0.90432 | 0.613268 | 0.701384 | 0.839409 |
| GaussianNB | 17177 | 2228 | 272 | 154 | 0.802052 | 0.265049 | 0.501501 | 0.222126 | 0.844412 |

In this features collection provisioning artificial network and decision tree prove to have the hegemony over the exhibition and other similar boosting and bagging techniques




(a) Decision Tree (b) ANN Network

Fig 13: Top 2 ROC Plot

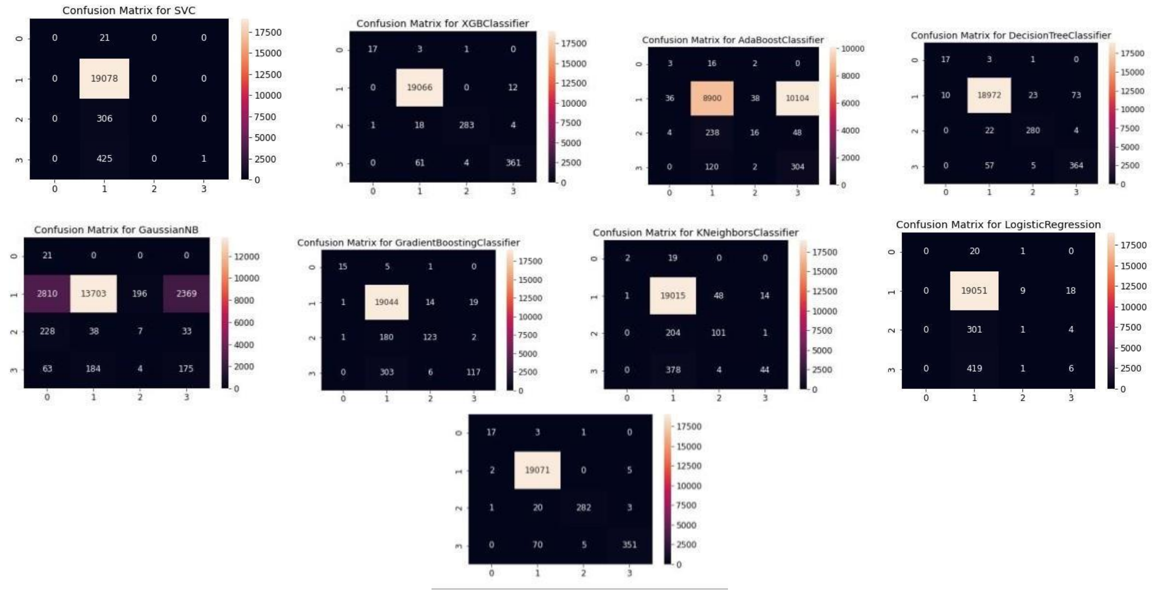



Fig 13: Confusion matrix for SVC, XGBoost, Adaboost, DecisionTree, Gaussian naive Bayes, Gradient Boost, K nearest neighbors, Logistic Regression and ANN network

The above figure discusses the confusion matrix for the full data set for various algorithms, it also showcases how decision tree, xGBoost and type one neural network also called artificial network proves to have a better understanding and feature extrapolation of the set than the other architectures.

Comparison pf different features

| **Feature Selection** | **Accuracy** | **Precision** | **Recall or Sensitivity** | **F1 Score** | **Specificity** | **Computation** |
|---|---|---|---|---|---|---|
| **GWO** | 0.9448927778 | 0.7166937143 | **0.5602544444** | 0.6145325714 | **0.8501743333** | **124** |
| **PSO** | **0.95753** | **0.7319957143** | 0.5494976667 | **0.6289732857** | 0.8436825556 | 194 |
| **FULL_DS** | 0.9395277778 | 0.7215624286 | 0.5579848889 | 0.6155694286 | 0.849234 | 378 |

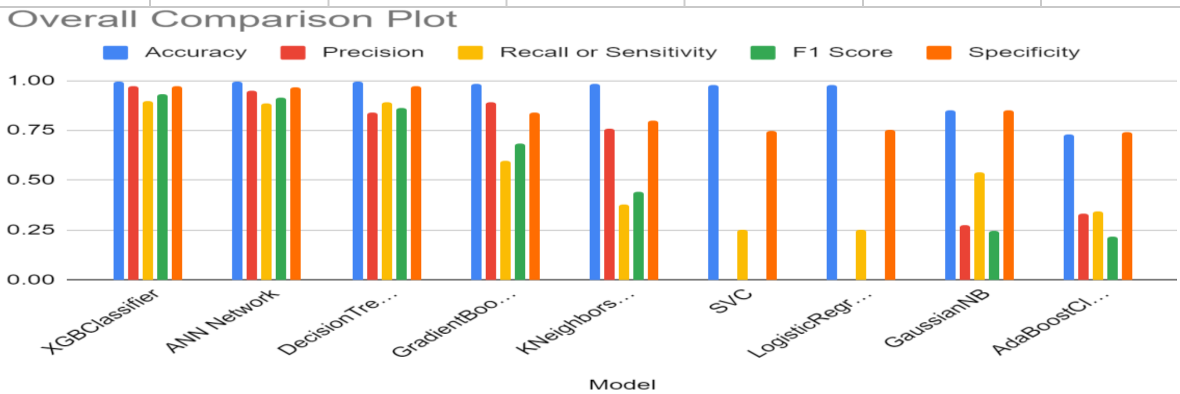


Fig 14 : model overall comaprision plot

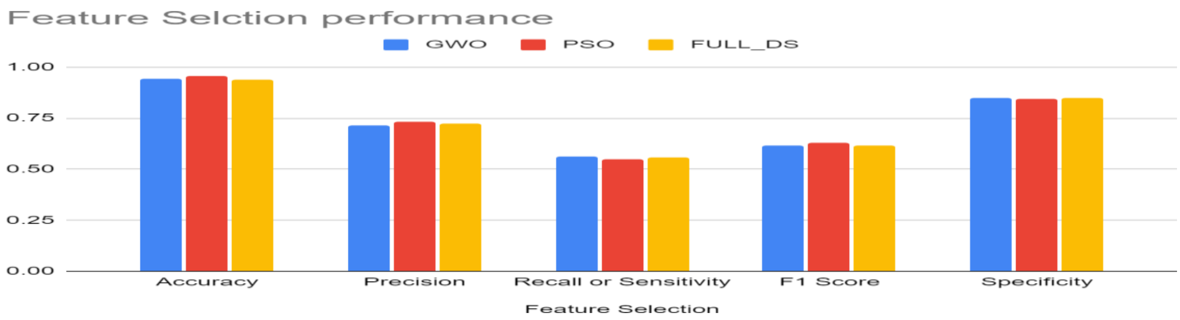



Fig 15:Feature selection performance plot

### 6.1 Discussion

It can be concluded that on the research orientation if the desired result is to increase precision overall affecting the F1 score then particles from the algorithm is probably the best option available. However if the choices have a regression with recall score analysis then the gray wolf optimisation which increases is the mean aggregate by over 1.02.PSO outperformed the rest of the methods with 95 % of accuracy, 73% of precision 62 % of F1 score. In terms of recall or sensitivity and specificity, GWO outperformed the rest methods.56 % of recall and 85 % of specificity gained by GWO. In terms of model comparison, ANN outperformed the rest with accuracy, precision, recall, F1 score, and specificity. XGboosting and Decision would be the second and third choices considering the model performance

## 7 Conclusion and Future Work

The paper aims to understand and correlate the reasoning for accident severity, weather conditions, environmental thresholds, and other attributes. In modeling the severity of accidents between 1 to 5, the paper illustrates the idea of distinguishing feature importance and diagnosis of random entropy attributes in formulating the machine learning conjuncture. Particle swarm optimization and Gray wolf optimization are two examples of nature-inspired algorithms that are used to initialize media exposure networks. Techniques drawn from nature enhance the adulation system in creating better beginning weights,which improves the training cycle. GWO initializer is employed for standard neural networks. When changing the origin weights, existence weight initializers also permit the use of a gradient descent function and improvised back dissemination. The results section demonstrates the differences between various feature selection techniques attributing the enhancement in computation to nature-inspired techniques and demonstrating the class-wise ROC metric of comparison based on the best model.

The final computation chart in the result section showcases artificial neural networks and decision trees to be the first and second choices based on the results. Intuitively showcasing that is strong individual classified is required rather than a bagging or multiple weak learner concept. This is because a random forest combining multiple decision trees does not appropriately shape up with the results due to the absence of multi-data samples and difficult samples which are not correctly classified.

How the covid-19 has influenced accident occurrence, driving behaviors, and population density may be investigated further; this might be the recommended future investigation.

# 8 Acknowledgement

I'd want to convey my heartfelt appreciation to my supervisor, Hicham Rifai, for making this job possible. His tolerance, enthusiasm, and vast expertise aided me throughout the study and creation of my thesis.
In addition, I am grateful to my family, friends, and colleagues for their assistance and encouragement during the course.

# References

Ali, M. H. & Baiee, W. R. (2021), Choosing an appropriate feature selectionmethod to enhance feed-forward ann, *in* '2021 International Conference on Communication Information Technology (ICICT)', pp. 86–91.

AlMamlook, R. E., Kwayu, K. M., Alkasisbeh, M. R. & Frefer, A. A. (2019), Comparison of machine learning algorithms for predicting traffic accident severity, *in* '2019 IEEE Jordan International Joint Conference on Electrical Engineering and Information Technology (JEEIT)', pp. 272–276.

Assi, K. (2020), Prediction of traffic crash severity using deep neural networks: A comparative study, *in* '2020 International Conference on Innovation and Intelligence for Informatics, Computing and Technologies (3ICT)', pp. 1–6.

Bahiru, T. K., Kumar Singh, D. & Tessfaw, E. A. (2018), Comparative study ondata min- ing classification algorithms for predicting road traffic accident severity, *in* '2018 Second International Conference on Inventive Communication and Computational Technolo- gies (ICICCT)', pp. 1655–1660.

Beryl Princess, P. J., Silas, S. & Rajsingh, E. B. (2020), Machine learning approach for identification of accident severity from accident images using hybrid features, *in* '2020International Conference for Emerging Technology (INCET)', pp. 1–4.

Chong, Abraham, A. (2004), Traffic accident data mining using machine learning paradigms, *in* 'Traffic Accident Data Mining Using Machine Learning Paradigms. Fourth International Conference on Intelligent Systems Design and Applications',pp. 415–420.

Fan, F. & Qian, Y.-B. (2017), Analysis of factors affecting the severity of car accidentsat intersections based on cumulative logistic model, *in* '2017 International Conferenceon Computer Systems, Electronics and Control (ICCSEC)', pp. 310–313.

Geyik, B. & Kara, M. (2020), Severity prediction with machine learning methods, *in* '2020 International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA)', pp. 1–7.

Iveta, M., Radovan, A. & Mihaljević, B. (2021), Prediction of traffic accidents sever- ity based on machine learning and multiclass classification model, *in* '2021 44th In- ternational Convention on Information, Communication and Electronic Technology (MIPRO)', pp. 1701–1705.

Labib, M. F., Rifat, A. S., Hossain, M. M., Das, A. K. & Nawrine, F. (2019), Road accident analysis and prediction of accident severity by using machine learning in bangladesh, *in* '2019 7th International Conference on Smart Computing Communica- tions (ICSCC)', pp. 1–5.

Li, F. & Jiang, K. (2020), Application of random-parameter negative binomial modelto examine the relationship between the severity of traffic accident, *in* '2020 IEEE 5th International Conference on Intelligent Transportation Engineering (ICITE)', pp. 351– 354.

Luo, Z., Jin, S., Jiang, Y. & Du, K. (2021), A new traffic accident risk prediction method based on adaptive neural fuzzy inference system, *in* '2021 IEEE International Confer- ence on Emergency Science and Information Technology (ICESIT)', pp. 354–358.

Makarova, I., Buyvol, P., Yakupova, G., Mukhametdinov, E. & Pashkevich, A. (2020), Identification for factors and causes affecting the traffic accident severity, *in* '2020 XII International Science-Technical Conference AUTOMOTIVE SAFETY', pp. 1–6.

Manzoor, M., Umer, M., Sadiq, S., Ishaq, A., Ullah, S., Madni, H. A. & Bisogni, C. (2021), 'Rfcnn: Traffic accident severity prediction based on decision level fusion of machine and deep learning model', *IEEE Access* **9**, 128359–128371.

Parseh, M., Asplund, F., Nybacka, M., Svensson, L. & Törngren, M. (2019), Pre-crash vehicle control and manoeuvre planning: A step towards minimizing collision severity for highly automated vehicles, *in* '2019 IEEE International Conference on Vehicular Electronics and Safety (ICVES)', pp. 1–6.

Shangguan, A., Mu, L., Xie, G., Wang, C., Jing, Y., Fei, R. & Hei, X. (2021), Traffic accident severity prediction based on oversampling and cnn for imbalanced data, *in* '2021 40th Chinese Control Conference (CCC)', pp. 7004–7008.

Shen, X. & Wei, S. (2020), 'Application of xgboost for hazardous material road transport accident severity analysis', *IEEE Access* **8**, 206806–206819.

Shi, C. & Deng, Y. (2019), Analysis of factors influencing severity of urban expresswayaccidents in guangdong province, *in* '2019 5th International Conference on Transport-ation Information and Safety (ICTIS)', pp. 243–246.

Tambouratzis, T., Souliou, D., Chalikias, M. & Gregoriades, A. (2010), Combining prob- abilistic neural networks and decision trees for maximally accurate and efficient accident prediction, *in* 'The 2010 International Joint Conference on Neural Networks (IJCNN)', pp. 1–8.

Zheng, M., Li, T., Zhu, R., Chen, J., Ma, Z., Tang, M., Cui, Z. & Wang, Z. (2019), 'Traffic accident's severity prediction: A deep-learning approach-based cnn network', *IEEE Access* **7**, 39897–39910.

.