Chapter One

Introduction

1.1 Background to the Study

Accidents involving motor vehicles pose a significant risk to both human life and the security of residential areas. According to a report published by the World Health Organization (WHO), road traffic accidents account for approximately 1.35 million deaths worldwide each year and cause between 20 and 50 million people to sustain non-fatal injuries[1]. In addition to the number of people who are killed and injured, road traffic accidents also result in an economic burden for the victims and their families. This includes the costs of treating injuries, the loss of productivity that occurs when someone is disabled or killed, as well as the loss of resources.

It has been reported that Nigeria, closely followed by Kenya, has the highest rates and incidences of road accidents in the world, with more than 29.1 deaths per 100 000 people[2]. In Nigeria, a total of 3,037,301 people were injured or killed in motor vehicle accidents between the years 1990 and 2012. 28.6% of these road-related accidents resulted in fatalities, 44.7% of them were considered to be serious accidents, and the remaining 26.7% were considered to be minor accidents[3]. About one quarter of this number resulted in fatalities, while the remaining accidents left victims with varying degrees of injuries[3].

Recent reports indicate that there were more than 11,800 people killed or injured in traffic accidents in Nigeria during the fourth quarter of 2021[4]. About 10.2 thousand of those were treated for injuries, while only 1.7 thousand were officially counted as fatalities. The most recent accounting period revealed that there were approximately

8.8 thousand people injured and approximately 1.4 thousand people killed as a direct result of road traffic accidents in the country[4]. One source claims that the majority of traffic collisions that take place in Nigeria are considered to be serious[5,6].

Road traffic accidents can be expressed as events or accidents like a car accident, vehicle collision, vehicle crash, pedestrian, road debris, animal, or other obstruction like a pole, building, or tree. Accidents are primarily quantifiable in losses like fatality, injury, resource in number, or money. Automobile manufacturers have made a significant number of efforts, primarily focused on both active and passive safety systems, to reduce the number of people killed or injured in traffic accidents[7]. These initiatives have been successful in improving traffic safety, resulting in a significant drop in the number of people killed on the roads

Techniques for accurate road traffic prediction and administration are absolutely necessary for Intelligent Transportation Systems (ITS), which enable the management and control of the transportation system to function in an efficient manner. For instance, traffic monitoring and management of operations, the prevention of traffic collisions, the regulation of traffic flow, and the issuance of warnings for potentially hazardous conditions such as curves, departures, and prone areas[8].

On the basis of the characteristics of road traffic accidents and information related to accidents, one can make suggestions regarding analyses, classifications, and forecasts of road traffic accidents. Several studies have presented potential solutions to the problem of road traffic accidents in terms of classification, analysis, and prediction[9,10,11,12]. Before and after the accident are the two times that the contributions can be divided into their respective categories. The works that were

done after the accidents were based on actual hand accident cases in order to analyze, classify, and predict losses caused by accidents.

For instance, the categorization of the severity level of an accident as minor, serious, or fatal, as well as the prediction of the amount of time an accident will last. On the other hand, the works referred to as "pre-accident" are based on real-time road traffic attributes, and their purpose is to forecast potential losses associated with an accident in advance. Work on identifying the location of hazards or the conditions under which they exist can also be included in pre-accident prediction work.

Results from pre-accident are significantly more important for maintaining a healthy human population and a habitable environment than post-accident results. Accidents involving vehicles on the road can be caused by a variety of factors related to traffic. The attributes such as the weather, pedestrians, the driver's experience (including age), the status of the vehicle (type, number of wheels, size, age, and speed), the amount of time spent traveling, the day of the week, the traffic flow, the road status, the condition of the lights, and the area (urban, rural, or junction)[13].

The influence that these characteristics will have on the accident will vary, particularly for the activities that pertain to pre-accident prediction. Road traffic accidents can be caused by a wide variety of factors, which are collectively referred to as traffic attributes. The primary actor who can control or manage the occurrence of an accident is the driver of the vehicle. Conducting a situation analysis and sounding an alarm about the potential consequences, such as collisions with other vehicles, can be useful as a method or instrument for assisting drivers. The traffic offices can also track vehicles that are at a high risk for being involved in a traffic accident in real

time, which makes them a secondary actor in this phenomenon. Therefore, the application of such technologies could potentially save the lives of people who are in danger (such as pedestrians, drivers, or roadside workers)[13].

Researchers have presented potential solutions for both pre and post-accident contexts in terms of safety, resource management, and control. Previous studies demonstrated what researchers had suggested as potential solutions for safety strategy (as pre-accident) and resource control (as post-accident)[11,12]. The post-accident solutions for traffic reflect how accidents in traffic can be managed and controlled in an efficient and effective manner based on an existing accident and the characteristics it possesses. For instance, categorizing the level of seriousness of a traffic accident, estimating the amount of time it will take for emergency services to arrive at the scene of an accident, and determining the extent to which traffic congestion is caused by traffic accidents[14].

When it comes to managing or resolving road traffic accidents, it is helpful to classify them so that the causes can be identified based on traffic attributes such as the type of accident, the condition of the lights, the type of road, and the characteristics of the road. There are five distinct categories that have been assigned to accidents involving vehicles[15]. Specifically, an accident that took place near or inside a curve road resulted in one injury, an accident that took place on a straight road resulted in one injury, an accident that took place on a straight road resulted in two injuries, an accident that took place on roads that were mostly not highways resulted in two injuries[15].

The categories will be useful in investigating situations and coming up with solutions so that accidents don't happen[15]. When attempting to characterize the occurrence and

losses that result from an accident, it is helpful to classify or predict traffic accidents within the context of their real-time implications.

For instance, a model for classifying accidents according to their severity, based on data collected from traffic accidents, was recently presented. This model divides accidents into two categories: damage only and severity (including injury and fatality)[10]. A significant factor in the occurrence of an accident will be a driver's personal characteristics, such as their level of aggression and aggressive behavior. In a model that was proposed for classifying drivers based on information about them, such as their aggressiveness and traffic violations, the model's goal was to categorize drivers according to their individual risk levels using the violation types. In a similar vein, a model for predicting vehicle accidents based on a driver's past infractions of traffic laws as well as their individual characteristics was proposed[16].

In addition, a work was done on a future driving risk prediction model for collisions, which was based on drivers' previous infractions of the traffic law[12]. Accidents caused by traffic (Level of Seriousness), congestion in the traffic (Level of traffic state), emergency Service (Response time). Identifying the factors that lead up to traffic accidents is another essential step, particularly for activities involving road maintenance and traffic control. Analysis of accident risk factors was suggested to be done in a study that was done on the prediction of traffic accidents[11].

Machine learning is a field of artificial intelligence (AI) whose aim is to comprehend how data is structured and model it so that it can be utilised by people[17]. Computers can train data using machine learning algorithms and employ statistical tools to produce values contained in a specified range. Machine learning facilitates the

development of models from available data that will enable decisions to be made based on these data inputs[18]. The commonest machine learning methods are supervised learning and unsupervised learning, even though reinforcement learning and semi-supervised learning methods exist[19].

The former methods are generally dependent on how data/information is received or how feedback is returned to the system being built[20]. In the supervised learning approach, the computer program is fed sample inputs already prelabeled with the desired outputs. Hence, the ML algorithm is trained by contrasting its real output with the initial data to discern errors. This action finetunes the model appropriately. Because a third party is required to "supervise" the computer programme, classification is classified as supervised learning. From the data miner's perspective, classification is synonymous with prediction and forecasting because it employs the same techniques.

Supervised learning methods aims to construct a model distribution of the response variable class in terms of sampled predictor labels[20]. The resultant classifier can now be employed to designate class values to proposed situations where the predictor labels are given, but the corresponding class feature is not known. Based on this, various machine learning classification methods have been developed in AI.

Unsupervised learning is used to describe the use of Artificial Intelligence algorithms to discover and learn patterns in datasets which have not been labeled or classified[21].

The three main types of classification are binary, multi-label, and multi-classification. Of the three, binary classification is employed the most, as most real-life tasks are

based on two discerning groups[22]. The efficiency of a machine learning system is governed by data quality and also the choice of representation features used to train it.

Though the usefulness of features depends on the task, it is generally assumed that certain features or sets of features are representative of a dataset and should be used as input for classification.

1.2 Statement of the Problem

Accidents on the world's roadways are responsible for a significant number of fatalities each year. The alarmingly high number of people killed and injured every year as a result of motor vehicle collisions is one indicator of the widespread issue that exists in the area of road safety. Traffic accidents can be managed and controlled differently for different purposes like traffic congestion, response time estimation, traffic accident duration, and level of accident seriousness. Some studies have been proposed solutions such as pre-accident solution: for example, predict the occurrence of traffic accidents[11,12]. On the other hand, these solutions did not address the severity of vehicle accidents based on factors such as vehicle speed, impact angle, and vehicle type, which could help emergency responders prioritize their responses and more effectively allocate resources. This work tends to predict the severity of any road traffic accident. An additional crucial aspect of the study is the evaluation of multiple traffic accident attributes for the purpose of achieving a better prediction performance using two algorithms (the Random Forest model and the Decision Tree Classifier model).

1.3 Aim and Objectives of the Study

The aim of this study is to develop a model to predict the severity of vehicle accidents based on traffic accident factors or attributes using Machine Learning. The specific objectives are to:

i. preprocess and encode the dataset feature using Machine Learning Algorithm
ii. train and test the dataset in (ii) using Random Forest and Decision Tree Classifier models and optimizing using hyperparameter tuning
iii. evaluate the models' performance using (a) precision, recall, F1-score metric, confusion matrix and ROC curve (b) accuracy, sensitivity, and specificity.

1.4 Significance of the Study

Predicting the severity of vehicle accidents based on traffic accident factors or attributes using Machine Learning will play a critical role in securing lives during accidents cases. This project provides solution and helps to enhance the current technology of security and safety issues. It will help emergency responders prioritize response and allocate resources more effectively during accident. Additionally, the proposed design will be able to provide detailed information about vehicle. These details can aid incident management decision-making. There has been a recent uptick in the number of self-driving vehicles either in production or in use.

A prediction model of this kind can be incorporated with other driver support systems in order to reduce the number of accidents and assist the vehicle (vehicle). By using it as a support system, drivers have the ability to get information about the potential outcomes of an accident in advance, in the context of a real-time scenario that includes human factors (such as the driver) and natural factors (such as the type of

road, the lighting condition, and the weather condition). This means that drivers have the ability to adjust their involvement (such as their speed and their focus to the ahead) prior to the occurrence of an accident. Using this model, traffic offices will also be able to monitor individuals who are in potentially hazardous situations in real time.

Academically, the study will contribute to the body of knowledge and proffer intelligent solutions to issues relating to response to accidents. The findings of this study will also serve as a reference for computer science students, lecturers, and researchers, as well as serve as a catalyst for further research on the subject. Additionally, findings may result in the development of new theories regarding using artificial intelligence for traffic monitoring and accident predictions.

1.5 Scope of the Study

The purpose of this thesis is to develop a model to predict the severity of vehicle accidents using machine learning. Prior to implementation, the developed model will be evaluated by weighing the benefits and limitations of various designs. This thesis considers a pre-accident prediction task using supervised learning algorithm. In order to provide assistance to individuals or groups, such as drivers and traffic offices, a solution in the form of a traffic accident prediction model was proposed as a solution. This model would support individuals or groups by classifying the number of casualties based on a variety of human and natural road traffic factors. This work also covers how an accident prediction model can be helped by predicting the severity of an accident and its uncertainties regarding a predicted value of new observation. Specifically, this work focuses on how this can be accomplished. This encompasses the following five encompassing areas: road traffic accidents and the factors that

contribute to them, feature selection, a machine learning model, analysis of the model, and interpretation.

The developed design will be evaluated based on precision, recall, and F1-score metrics. The confusion matrix will also be used to visualize the models' performance on each class, and the ROC curve will be used to evaluate the models' overall performance. The models were also evaluated based on their accuracy, sensitivity, and specificity. The findings will be presented and interpreted descriptively.

1.6 Limitation of Research

Despite the valuable insights gained from the research and the promising results achieved, there are some limitations:

i. Dataset Limitations: The research's conclusions heavily rely on the quality and representativeness of the dataset used for training and testing the models. If the dataset is small, unbalanced, or contains biases, it may not fully capture the complexities of the real-world problem, leading to potential over-fitting or under-performance on unseen data.

ii. Availability of Real-time Data

iii. Changing Real-world Conditions: The performance of machine learning models can be influenced by changes in the underlying data distribution or external factors. Models that perform well during the research period might become less effective over time due to shifts in the environment or user behavior.

1.7 Definition of Operational Terms

Intelligent Transportation Systems (ITS): Enable the management and control of the transportation system to function in an efficient manner. For instance, traffic monitoring and management of operations, the prevention of traffic collisions, the regulation of traffic flow, and the issuance of warnings for potentially hazardous cotions such as curves, departures, and prone areas

Machine Learning (ML): Machine learning is a field of artificial intelligence (AI) whose aim is to comprehend how data is structured and model it so that it can be utilised by people, facilitates the development of models from available data that will enable decisions to be made based on these data inputs.

Neural Network: Neural Network are computational model utilised in the domains of problem-solving and machine learning. Neural networks (NNs) have been widely applied to real-world problems in various domains such as business, education, economics, and other areas of life.

Post-Accident: Post accidents reflect how accidents in traffic can be managed and controlled in an efficient and effective manner based on an existing accident and the characteristics it possesses.

Pre-Accident: Pre-accidents are based on real-time road traffic attributes, and their purpose is to forecast potential losses associated with an accident in advance. They work on identifying the location of hazards or the conditions under which they exist.

Reinforcement Learning: Reinforcement learning is a feedback-driven process whereby an artificial intelligence agent, which is a software component,

autonomously explores its environment through trial and error. It takes actions, learns from its experiences, and enhances its performance.

Road Traffic Accidents (RTA): Road traffic accidents can be expressed as events or accidents like a car accident, vehicle collision, vehicle crash, pedestrian, road debris, animal, or other obstruction like a pole, building, or tree.

Semi-Supervised Learning Methods: Semi-supervised learning is a machine learning algorithm that occupies an intermediate position between supervised and unsupervised learning algorithms. These algorithms leverage both labeled and unlabeled datasets during the training phase

Supervised Learning Methods: In supervised learning, machines undergo training using a dataset that has been labeled, and subsequently utilize this training to generate predictions. The labeled data denotes that certain inputs have already been mapped to their respective outputs..

Traffic Attributes: Traffic Attributes are the wide variety of factors causes accidents

Unsupervised Learning: Unsupervised machine learning involves training a machine using an unlabeled dataset, whereby the machine is capable of predicting output without any form of supervision. The models are trained using unclassified and unlabeled data, and subsequently operate on this data in an unsupervised manner.

Endnotes

[1]M.B Rabbani, M.A Musarat, W.S Alaloul, A. Maqsoom, H. Bukhari, & W. Rafiq. *Road Traffic Accident Data Analysis and its Visualization*. **Civil engineering and architecture.** 2021;9(5):1603-14.

[2]A.G Salaudeen. *Risk Factors and Safety Measures for Road Traffic Crashes Among Inter-City Commercial Drivers in Kwara State, Nigeria*, **Doctoral Dissertation, University Of Ilorin**. 2018

[3]A.A Audu, O.F Iyiola, A.A Popoola, B.M Adeleye, S Medayese, C Mosima, & N Blamah. *The Application of Geographic Information System as an Intelligent System Towards Emergency Responses in Road Traffic Accident in Ibadan*. **Journal of Transport and Supply Chain Management**. 2021 Mar 4;15:17

[4]C.A Eneh, A Okosun, M.C Oloto, V Emenuga, C.P Ehiogu, C.I Eneonwo & O.C Eneh. *A Comparative Analysis of Road and Vehicle Qualities as Factors of Road Traffic Carnage in Nigeria*.DOI: https://doi.org/10.21203/rs.3.rs-2393243/v1.2022

[5]A.O Oyetubo, O.J Afolabi & M.E Ohida. *Analysis of Road Traffic Safety in Minna Niger State, Nigeria.* ***Logistics, Supply Chain, Sustainability and Global Challenges***. 9(1): 2018, 23-38.

[6]C Uzondu, S Jamson & F Lai. *Exploratory Study Involving Observation of Traffic Behaviour and Conflicts in Nigeria Using the Traffic Conflict Technique*. **Safety science**. 1;110, Dec 2018:273-84

[7]N. Lubbe, H Jeppsson, A Ranjbar, J Fredriksson, J Bärgman, & M Östling. *Predicted road traffic fatalities in Germany: The Potential and Limitations of Vehicle Safety Technologies from Passive Safety to Highly Automated Driving*. **In Proceedings of IRCOBI conference. Athena, Greece,** Sep 2018

[8]L.A Klein. "S*ensor and Data Fusion for Intelligent Transportation Systems."* **Society of Photo-Optical Instrumentation Engineers**, 2019.

[9]M Zahid, Y Chen, S Khan, A. Jamal, M Ijaz & T Ahmed. *Predicting Risky and Aggressive Driving Behavior among Taxi Drivers: do Spatio-Temporal Attributes Matter?* **International Journal of Environmental Research and Public Health**. 17(11), Jun 2020: 3937.

[10]XL Xia, B. Nan & C. Xu. *Real-time Traffic Accident Severity Prediction using Data Mining Technologies*. **In 2017 International Conference on Network and Information Systems for Computers (ICNISC)** Apr 14, 2017, pp. 242-245. IEEE.

[11]A. Finogeev, M. Deev & I. Kolesnikoff. *Proactive Big Data Analysis for Traffic Accident Prediction*. **In 2020 5th International Conference on Innovative**

Technologies in Intelligent Systems and Industrial Applications (CITISIA), IEEE, Nov 25 2020, pp. 1-9..

[12]H Wen, X Zhang & Q Zeng. *Predicting Future Driving Risk of Crash-Involved Drivers Based on a Systematic Machine Learning Framework*. **International Journal of Environmental Research and Public Health**. 16(3), 2019;334-52.

[13]J.A Andeta. *Road-traffic Accident Prediction Model: Predicting the Number of Casualties*. **Master Degree Thesis in Informatics. ECTS Spring Term** 2021

[14]S Haynes, P.C Estin, S Lazarevski, M Soosay & A.L Kor. *Data Analytics: Factors of Traffic Accidents in the UK,* **In 2019 10th International Conference on Dependable Systems, Services and Technologies (DESSERT) IEEE**, Jun 5 2019, pp. 120-126..

[15]P.A Nandurge & N.V Dharwadkar. *Analyzing Road Accident Data Using Machine Learning Paradigms*. **In 2017 International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC) IEEE**, Feb 10, 2017 (pp. 604-610)..

[16]D.H Kim, L.M Ramjan & K.K Mak. *Prediction of Vehicle Crashes by Drivers' Characteristics and Past Traffic Violations in Korea using a Zero-Inflated Negative Binomial Model*. **Traffic injury prevention.** 17(1): Jan 2 2016; 86-90.

[17]J Alzubi, A Nayyar & A Kumar. *Machine Learning from Theory to Algorithms: An Overview*. **In Journal of Physics: Conference Series** (Vol. 1142, p. 012012). **IOP Publishing**, Nov 2018.

[18]K.R Varshney. *Trustworthy Machine Learning and Artificial Intelligence*. *XRDS: Crossroads*, **the ACM Magazine for Students**. 25(3):26-9, Apr 10, 2019.

[19]I.H Sarker. *Machine Learning: Algorithms, Real-World Applications and Research Directions*. **SN computer science.** (3):160, May 2 2021.

[20]A.A Jamali, R Ferdousi, S Razzaghi, J Li, R Safdari & E Ebrahimie. *DrugMiner: Comparative Analysis of Machine Learning Algorithms for Prediction of Potential Druggable Proteins*. **Drug Discovery Today**. 21(5):718-24, May 1 2016.

[21]A.S Heinsfeld, A.R Franco, R.C Craddock, A Buchweitz, & F Meneguzzi. *Identification of Autism Spectrum Disorder Using Deep Learning and the ABIDE Dataset*. **NeuroImage: Clinical**; 17:16-23 Jan 1 2018.

[22]J Hagenauer & M Helbich. *A Comparative Study of Machine Learning Classifiers for Modeling Travel Mode Choice*. Expert Systems with Applications, 78:273-82. Jul 15 2017.

[23]S Tangirala. *Evaluating the Impact of GINI Index and Information Gain on Classification Using Decision Tree Classifier Algorithm*. **International Journal of Advanced Computer Science and Applications**. 11(2):612-9. 2020;

Chapter Two
Literature Review

2.1 Conceptual Review

2.1.1 Road Traffic Accident

A road traffic accident can be characterized as an incident that is unpredictable and influenced by multiple factors, typically occurring after one or more road users have demonstrated an inability to navigate the road conditions[1,2]. Road traffic accidents encompass a range of incidents, including but not limited to collisions between vehicles, crashes involving pedestrians, encounters with road debris, and obstructions such as poles, buildings, or trees. The quantification of accidents is primarily based on losses, such as fatalities, injuries, resources in number, or monetary value. Scholars have put forth potential remedies for both pre-accident and post-accident situations with respect to safety, resource allocation, and regulation[3]. The management and control of traffic accidents can be optimized through post-accident solutions that take into account the specific attributes of the accident in question[3]. Instances of potential applications of the proposed methodology include the categorization of the severity level of a vehicular collision, the anticipation of the duration of an emergency service's arrival in response to a collision, and the evaluation of the extent of traffic congestion resulting from vehicular accidents[3].

The classification of road traffic accidents is a valuable tool for identifying the underlying causes of accidents. This is achieved by analyzing various traffic attributes

such as accident type, light condition, road type, and road characteristics. The resulting insights can be used to effectively manage and resolve such incidents[4].

Traffic accidents have been classified into five different categories[5]. Namely, the incidents took place in various locations, including curved roads resulting in one injury, straight roads resulting in one or two injuries, mostly non-highway roads resulting in two injuries, and straight or slightly curved roads resulting in mostly one injury[3].

According to research, the utilization of categories can aid in the examination of circumstances and the development of preemptive solutions to prevent accidents. The classification or prediction of traffic accidents in real-time scenarios is a useful approach for characterizing the incidence and associated damages resulting from an accident[6,7]. A model for classifying accident severity was introduced in a study for utilizing traffic accident data to differentiate between incidents resulting in solely property damage and those involving injury or fatality. The personal behavior and level of aggressiveness exhibited by drivers are factors that can significantly contribute to the occurrence of an accident[3].

Previous research presented a classification model that utilized driver-specific data, including traffic violations and aggressive behavior, to categorize individuals into varying levels of risk. The model's objective was to classify individuals based on the type of violation committed[8]. In another study that proposed a vehicle crash prediction model that takes into account a driver's historical traffic violations and personal characteristics[9]. Another study has proposed a prospective model for predicting driving risks in the future, which is based on drivers' past traffic

violations[10]. The identification of factors contributing to traffic accidents is a crucial aspect, particularly in relation to road maintenance and traffic control operations.

Public health professionals globally acknowledge the existence of a worldwide epidemic of road traffic accidents (RTA).

The prevalence of the aforementioned phenomenon is comparatively greater in developing nations. In 2011, the World Health Organization reported that a disproportionate number of road traffic fatalities, specifically 92%, occurred in low- and middle-income countries despite these countries only having 53% of registered vehicles. The incidence of road traffic accidents (RTAs) in Nigeria has been increasing, leading to a corresponding rise in injuries and fatalities. This trend is particularly concerning as RTAs are currently the leading cause of mortality on the African continent. In Nigeria, road accidents rank as the third highest contributor to overall mortality, the primary cause of trauma-related fatalities, and the prevailing cause of disability[11,12].

According to World Health Organization's report, the nation exhibits a yearly average of 1042 fatalities per 100,000 vehicles, which is among the highest global rates of road accidents[11]. Statistics indicates an increasing prevalence of road traffic accidents (RTAs) in Nigeria and other developing nations, which have significant negative impacts on both physical and socioeconomic aspects. Nevertheless, a comprehensive and integrated strategy to address this issue has not yet been developed. To facilitate the development of effective interventions, it is essential to initiate the process by formulating a clear and concise statement of research questions utilizing the Problem identification, interventions, comparisons, and outcome (PICO) model[11].

The incidence of injuries and fatalities arising from Road Traffic Accidents (RTA) in Nigeria has been observed to be increasing, thereby ranking as the third most prevalent cause of mortality in the country[12]. Additionally, RTA is the primary cause of trauma-related deaths and the most frequent cause of disability.

The situation in Nigeria is particularly challenging due to inadequate traffic infrastructure, poor road design, insufficient enforcement of traffic regulations, a burgeoning population, and a consequent increase in the number of individuals operating motor vehicles.

It is anticipated that the increase in Nigeria's economy will result in a corresponding rise in traffic volume, with projections indicating a surge from 8 million vehicles in 2013 to 20-40 million vehicles by 2023. During fourth quarter of 2021, Nigeria recorded over 11,800 incidents of road traffic casualties. Out of the total number, approximately 10.2 thousand cases were recorded as injuries, whereas 1.7 thousand cases were documented as registered deaths. During the preceding quarter, the nation recorded an estimated 8.8 thousand injuries and 1.4 thousand fatalities as a consequence of vehicular accidents[13]. According to the source, a majority of road accidents that take place in Nigeria are categorized as severe. The Road Traffic Accidents (RTA) have multifaceted implications that encompass physical, social, emotional, and economic dimensions.

The demographic groups that are most impacted by road accidents in terms of fatalities, physical disability, and morbidity are typically those who are young and within the economically productive age range[12]. Individuals who have survived traumatic events frequently experience a reduction in their overall standard of living due to physical deformities and disabilities, post-traumatic stress disorder, and a

decrease in personal income. This is particularly true in a nation where rehabilitation services are not widely recognized for their exceptional quality[12]. The rest of the population experience a continuous and widespread sense of apprehension when it comes to transportation due to their lack of perceived safety while on the roadways.

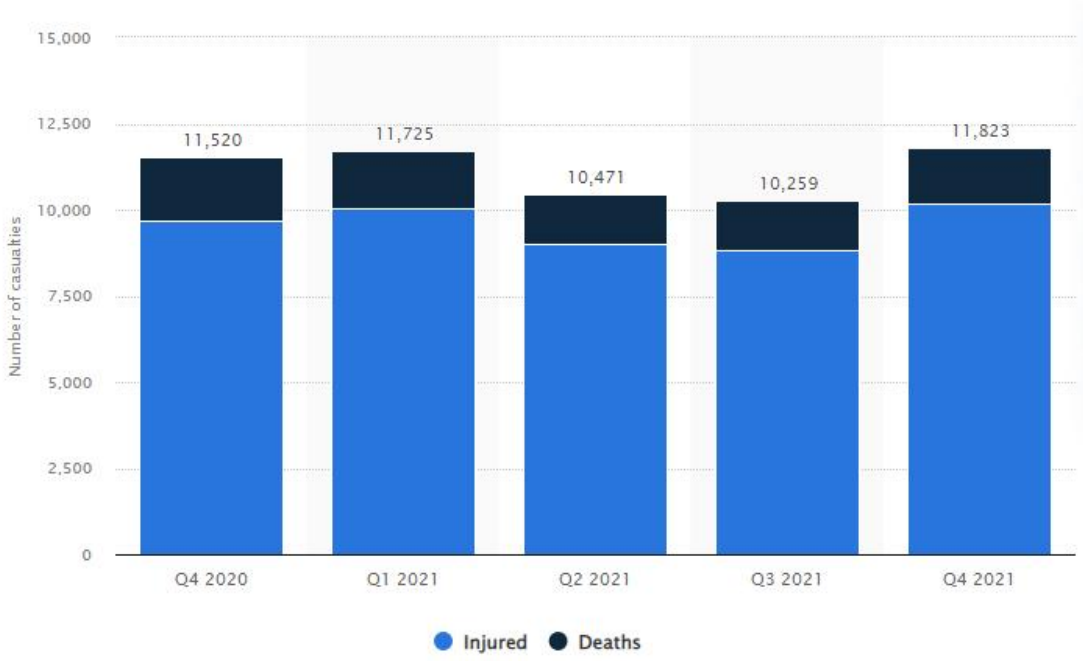

Figure 2.1: Number of Road Traffic Injuries and Deaths in Nigeria from Q4 2020 to Q4 2021[13].

The cumulative effects of these injuries constitute significant social, economic, and psychological losses. These losses are on a large scale. The direct economic cost of RTA was estimated to be 518 billion US dollars per year across the globe in 2003, with 100 billion US dollars of that amount occurring in economically developing nations[11]. The World Health Organization (WHO) places the cost of RTA at a national level somewhere between 1% and 3% of the gross domestic product. RTA suffers losses of approximately 80 billion Naira every year in Nigeria[14]. This economic cost takes into account the cost of repairing any damaged property or public

amenity, the cost of any necessary medical treatment, and the cost of lost productivity as a direct result of the accident. This is a significant loss for the economy, particularly for a nation that already struggles with high levels of poverty.

2.1.2 Machine Learning

Machine learning is a subfield of AI that focuses on the study of computer algorithms designed to learn and improve automatically. Supervised learning, unsupervised learning, and reinforcement learning are the three categories that make up machine learning[15,16]. The goal of supervised learning is to produce a model that can predict an output by using historical observations of that output that have been labeled. Additionally, regression and classification are the two types of supervised learning that are categories based on the output value, which can be continuous or discrete[3]. The process of regression involves fitting a model to the data that is provided, and it produces a continuous output. On the other hand, classification sorts the input data in order to produce the most useful output and generates discrete (or what is more commonly referred to as a class) output[3,17].

An unlabeled feature serves as the input for unsupervised learning, which then returns newly organized data by grouping, clustering, or organizing based on similarity measures such as distances[3,18]. Reinforcement learning is again different from both supervised and unsupervised learning. It learns by continuously optimizing an unknown reward function and updating its internal state based on some performance criterion[19].

2.1.2.1 Types of Machine Learning

Machine learning is divided into mainly four types, which are:

i. Supervised Machine Learning

ii. Unsupervised Machine Learning

iii. Semi-Supervised Machine Learning

iv. Reinforcement Learning

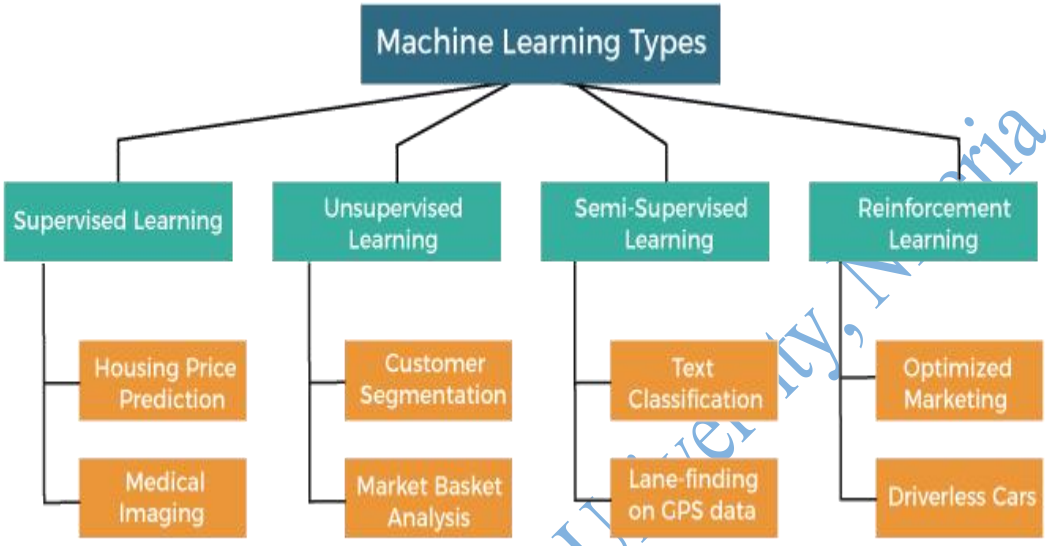

Figure 2.2: Types of Machine Learning[19]

Supervised Machine Learning

In the context of supervised learning, machines undergo training using a dataset that has been labeled, and subsequently utilize this training to generate predictions. The labeled data denotes that certain inputs have already been mapped to their respective outputs. Initially, the machine is subjected to training using input and corresponding output data. Subsequently, the machine is tasked with predicting the output by utilizing the test dataset. The primary objective of the supervised learning methodology is to establish a correlation between the input variable (x) and the output variable (y)[17].Several practical implementations of supervised learning include Risk Assessment, Fraud Detection, and Spam Filtering, among others.

There are two distinct categories of problems in supervised machine learning, as follows:

i. Classification
ii. Regression

Classification: Classification algorithms are utilised to address classification problems where the output variable is categorical in nature, such as binary categories like "Yes" or "No," or nominal categories like "Male" or "Female[18]. The classification algorithms are utilised to make predictions regarding the categories that are present within a given dataset. Instances of classification algorithms in practical applications include Spam Detection and Email Filtering, among others. Several widely used classification algorithms include the Random Forest Algorithm, Decision Tree Algorithm, Logistic Regression Algorithm, and Support Vector Machine Algorithm[18].

Regression: Regression algorithms are commonly employed to address regression problems that exhibit a linear correlation between input and output variables. Regression models are utilised to forecast continuous output variables, such as market trends, weather patterns, and other related phenomena[17]. Several commonly used regression algorithms include the Simple Linear Regression Algorithm, Multivariate Regression Algorithm, and Lasso Regression.

Supervised Learning are used in various applications in diverse fields such as Image Segmentation, Medical Diagnosis, Fraud Detection, Spam Detection, and Speech Recognition[17,18].

Unsupervised Machine Learning

Unsupervised machine learning involves training a machine using an unlabeled dataset, whereby the machine is capable of predicting output without any form of supervision. The models are trained using unclassified and unlabeled data, and subsequently operate on this data in an unsupervised manner. The primary objective of the unsupervised learning algorithm is to cluster or classify the unstructured dataset based on similarities, patterns, and dissimilarities. The machines are programmed to identify concealed patterns within the input dataset[17].

Unsupervised Learning can be categorized into two distinct types, as follows:

i. Clustering

ii. Association

Clustering: The clustering methodology is employed to identify the intrinsic clusters within the dataset. Cluster analysis is a method of categorizing objects into groups based on their similarities, with the aim of ensuring that objects within a group share the most similarities while having fewer or no similarities with objects in other groups[18]. One instance of the utilization of a clustering algorithm is the categorization of customers based on their purchasing patterns. Several widely used clustering algorithms include the K-Means Clustering algorithm, Mean-shift algorithm, DBSCAN Algorithm, Principal Component Analysis, and Independent Component Analysis.

Association: Association rule learning is an unsupervised machine learning methodology that aims to discover significant associations between variables in a vast dataset. The primary objective of this particular learning algorithm is to identify the

interdependence between two data items and subsequently establish a correlation between these variables, thereby optimising the potential for profit generation. Several widely used Association Rule Learning algorithms include the Apriori Algorithm, Eclat, and FP-growth algorithm. Unsupervised Learning finds various applications such as Network Analysis, Recommendation Systems, Anomaly Detection, and Singular Value Decomposition (SVD)[18].

Semi-Supervised Learning

Semi-supervised learning is a machine learning algorithm that occupies an intermediate position between supervised and unsupervised learning algorithms. Semi-supervised learning algorithms occupy a middle ground between supervised learning, which utilities labeled training data, and unsupervised learning, which operates without labeled training data[18,19]. These algorithms leverage both labeled and unlabeled datasets during the training phase. Semi-supervised learning can be considered as an intermediary approach between supervised and unsupervised learning techniques, wherein the algorithm operates on a dataset that contains a limited number of labeled instances.

However, it is noteworthy that the majority of the data in semi-supervised learning is unlabeled. Due to their high cost, corporations may opt to utilise a limited number of labels for their business needs. In contrast to supervised and unsupervised learning, which rely on the availability or lack of labels, this approach exhibits a distinct dissimilarity[18]. The concept of Semi-supervised learning has been introduced as a means of addressing the limitations of both supervised and unsupervised learning algorithms. The primary objective of semi-supervised learning is to optimise the

utilisation of all available data, as opposed to solely relying on labeled data as in the case of supervised learning. The first step involves clustering comparable data using an unsupervised learning algorithm[18]. This process subsequently facilitates the labelling of previously unlabeled data as labeled data. The reason for this is that obtaining labeled data is a more costly process compared to acquiring unlabeled data.

Advantages of Semi-supervised Learning

i. It is simple and easy to understand the algorithm.

ii. It is highly efficient.

iii. It is used to solve drawbacks of Supervised and Unsupervised Learning algorithms.

Disadvantages

i. Iterations results may not be stable.

ii. Cannot be applied these algorithms to network-level data.

iii. Accuracy is low.

Reinforcement Learning

Reinforcement learning is a feedback-driven process whereby an artificial intelligence agent, which is a software component, autonomously explores its environment through trial and error. It takes actions, learns from its experiences, and enhances its performance. The reinforcement learning agent is incentivized to optimise its performance by receiving positive reinforcement for favourable actions and negative

reinforcement for unfavourable actions, with the ultimate objective of maximising its cumulative reward[19,20].

Reinforcement learning is a type of machine learning that differs from supervised learning in that it does not rely on labeled data. Instead, agents acquire knowledge solely through their experiences[20]. The process of reinforcement learning bears resemblance to that of human learning, whereby a child acquires knowledge through experiential encounters in their daily routine. Reinforcement learning can be exemplified through gameplay, wherein the game serves as the environment, the actions taken by an agent at each step determine the states, and the objective of the agent is to attain a high score.

Agents are subject to feedback mechanisms that involve both punishment and rewards. Reinforcement learning has been utilised in various domains, including but not limited to Game theory, Operation Research, Information theory, and multi-agent systems, owing to its distinctive mode of operation. The formalisation of a reinforcement learning problem can be achieved through the utilisation of a Markov Decision Process (MDP)[19]. Within the framework of Markov Decision Processes (MDP), the agent engages in ongoing interactions with the environment by executing actions. Following each action, the environment responds by generating a new state.

Reinforcement learning is categorized mainly into two types of methods/algorithms:

i. Positive Reinforcement Learning: Positive reinforcement learning specifies increasing the tendency that the required behaviour would occur again by adding something. It enhances the strength of the behaviour of the agent and positively impacts it.

ii. Negative Reinforcement Learning: Negative reinforcement learning operates in a manner that is diametrically opposed to that of positive reinforcement learning. By avoiding the negative condition, the likelihood of the specific behaviour recurring is heightened.

Reinforcement Learning has been applied in various domains such as video games, resource allocation, robotics, and text analysis.

2.1.3 Classification of RTA Using Machine Learning

Classification is a supervised learning technique utilized in statistics and machine learning for predicting the category of a set of provided data.

In addition, the process of classification modeling involves the estimation of a function's mapping *(f)* from a given input value *(x)* and its discrete output value *(y)*, as shown in Equation (1) below[20]. The process of classification involves the categorization of a particular dataset into distinct classes or targets. The process of categorizing data can be applied to datasets that are either structured or unstructured. The initial step involves predicting the target of a specific data point. The fundamental concept of classification involves the identification of the specific class or target to which a given data point belongs. There exist four distinct types of classification, which are widely recognized in the field, including binary, multi-label, multi-class, and imbalanced.

Multi-class classification is employed in studies in which the variable *(y)* consists of more than two targets or classes.

y=f(x), where y=class or target output

This research only used supervised machine learning for the classification of road traffic accident. Hence, there are several supervised learning classification algorithms some are discussed below;

2.1.3.1 Gradient Boosting Machines (GBoost)

A GBoost is a popular machine learning algorithm that builds an ensemble of shallow trees sequentially. Each current tree learns and improves on the previous one, where each new tree in the order tries to fix up the error made on the last one, shown in Figure2.2 [23,21].

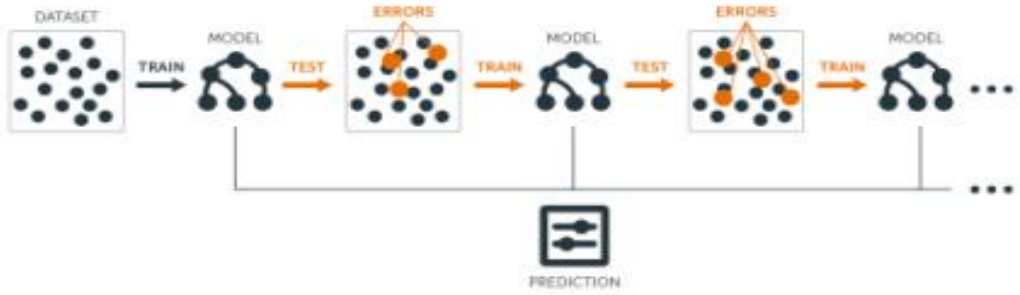

Figure 2.3: Sequential Ensemble Approach[21]

GBoost iteratively improves the predictions of y from x with respect to L by adding base learners or new weak that improve upon the previous ones, founding an additive ensemble model of size M[22]:

$$g_o(x) = c, g_i(x) = g_{i-1}(x) + \gamma_i h_i(x), \quad i = 1, \ldots, M \qquad 2$$

Where i the iteration index; h_i is the i^{th} base model, for example, a decision tree; γ_i is the weight or the coefficient of the i^{th} base model. GBoost has been used for several regression tasks and achieved better performance than alternative algorithms[3,22].

2.1.3.2 Support Vector Regression (SVR)

Support vector machine (SVM) solves binary classification problems by framing a convex optimization which involves finding the maximum margin separating the hyperplane[23]. The SVM is employed for either regression or two-group classification problems, though it is mainly used in classification[24]. In this method, each data point is plotted in n-dimensional space, with each coordinate representing the value of a given feature. Here, classification is done by locating the line that clearly segregates both classes. An optimal hyperplane would be the linear decision function with the highest definitive boundary between vectors of both groups[3].

If the need for an additional feature arises, the SVM uses the kernel algorithm to change a low-dimensional input space into a high-dimensional one. In other words, it transforms problems that seem inseparable into integratable ones. If there is no error in this separation, then, the Expected value of error is given as

$$E[Pr\ Pr\ (error)\ ] \leq \frac{E[number\ of\ support\ vectors]}{[number\ of\ training\ vectors]} \quad (3)$$

The decision function will be given as

$$D(x) = w\emptyset(x) + b \quad (4)$$

which is the best line that integrates the training data, w and b are parameters of the SVM, and $\emptyset(x)$ is the function which transforms the data into the new M dimension

$$\frac{D(x)}{||w||} \quad (5)$$

represents the line, which is the distance between item x and the hyperplane.

The parameters of the linear decision function that will maximize M are:

$$w = \sum_k \quad a_k y_k x_k \tag{6}$$

$$b = (y_k - w * x_k) \tag{7}$$

The principal function of the above training algorithms is to solve the equation quadratically[3].

$$J = \left(\frac{1}{2}\right) ||w|| \tag{8}$$

There are two types of SVM:

Linear SVM which is obtainable when the data is integratable in linear form. This implies that the data points can be clearly integratable by a single straight line[25].

Non-Linear SVM which covers when data is not linearly separable, and advanced techniques (kernel tricks) are applied[25]. It is important to choose a kernel to work with, and this is largely dependent on the dataset at hand. If linear, then a linear kernel function must be adopted. Starting with the hypothesis that the data is linear is ideal, then working through other kernels to compare performance metrics.

SVM performs better on a linear dataset, and its efficiency is enhanced in high-dimensional data. SVM is very robust as it is non-sensitive to outliers[26]. SVM generalization to SVR is accomplished by presenting an ε-insensitive area around the function, named the ε-tube[27,28]. This tube redevelops the optimization problem to find the tube that best approximates the continuous-valued function. SVR is framed as an optimization issue by first defining a convex ε-insensitive loss function to minimize and discover the flattest tube that comprises most of the training instances[23]. A general idea of the SVR is shown in Figure 2.3 below[23].

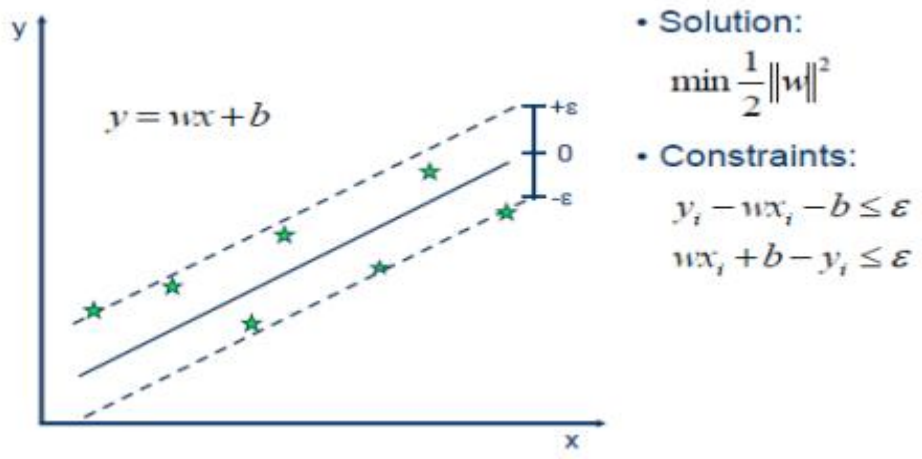

Figure 2.4: Support Vector Regression[23]

SVR has achieved a better performance in terms of performance than an artificial neural network[31]. However, SVR may not fit with more than 10000 observations. Instead, a LinearSVR version (SVR with its linear kernel) can handle a prediction task with a large dataset of observations.

2.1.3.3 Extreme Learning Machines (ELM)

ELM is a training method whose training speediness is very fast, in a single hidden feed-forward artificial neural network (SLFN)[31]. The main advantage of the ELM algorithm is that it allocates the weights and thresholds between the input layer and the hidden layer randomly[32]. Once these values are assigned, the ELM does not need to adjust these random parameters during the whole learning process that helps to complete the training process extremely fast. The general structure of the ELM algorithm is depicted in the Figure below[31].

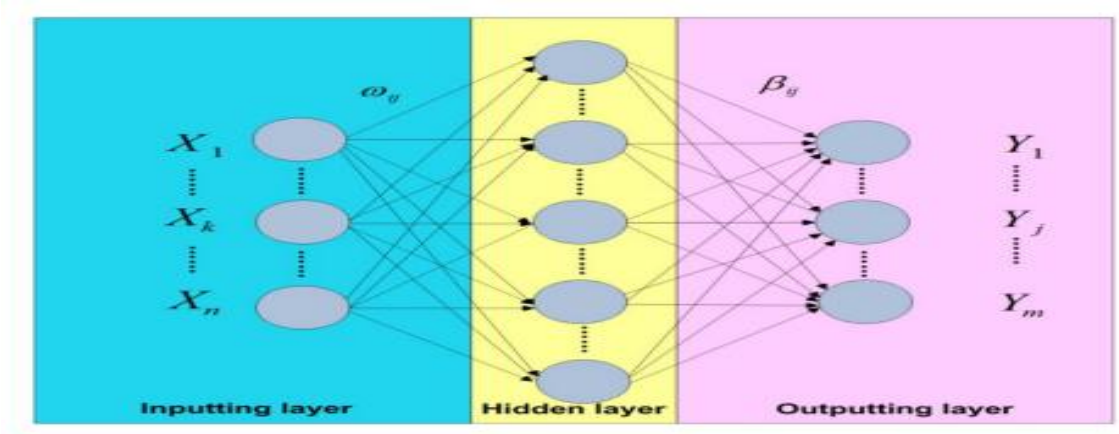

Figure 2.5: Architecture of ELM[23].

The main parameters of the ELM algorithm are described as follow

$$\omega = [\omega_{11}\ \omega_{12}\ \omega 1_n\ \omega_{21}\ \omega_{22}\ \omega 2_n\ \omega_l 1\ \omega_l 2\ \omega l_n\ ] \quad (9)$$

Where ω is the network weight between the input and the hidden layers ωij is the weight between the i^{th} input node of the input layer and the j^{th} hidden node of the hidden layer l is the number of input nodesin the input layer.

2.1.3.4 K-Nearest Neighbour (KNN)

The KNN approach estimates the conditional distribution of Y given X, then assigns an observation to the modal class; the class with the highest probability[33].

In order to predict the outcome for an observation X=x, the observations closest to x are located. Then X is set to the class to which these observations belong. This "closeness" is determined by distance metrics such as Euclidean and Minkowski[34]. Given that the original measurement values of the predictors will affect the distance results, it is pertinent to scale and centre all predictor values to eliminate bias towards predictors with higher scale and give all predictors an equal chance while calculating Euclidean distance.

KNN is a non-parametric method, which requires that all observations in the experiment be higher than the number of chosen predictors.

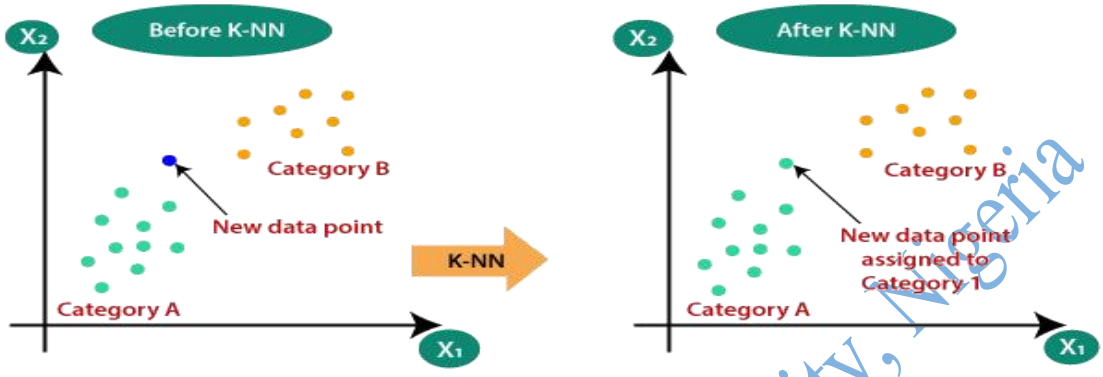

Figure 2.6: Operation of K-NN Algorithm[33].

Assume a training set D made up of X_i, $i \in [1, |D|]$ samples, where F described the number of features and assumed normality. Each output was labeled with class label $y_j \in Y$. The aim was to classify an unknown variable q. For every $x_i \in D$, the distance between q and x_i was computed thus:

$$d(q, x_i) = \sum_{f \in F} \quad w_f \partial(q_f, x_{if}) \tag{10}$$

and the equation, which measured distance in both discrete and continuous cases was written as: $\partial(q_f, x_{if}) = \{0 \quad f\ discrete\ and\ q_f = x_{if}\ 1 \quad f\ discrete\ \ and\ q_f \neq x_{if}\ |q_f - x_{if}| \qquad f\ discrete$ (11)

upon which the k nearest neighbours are chosen. The class of q is ideally selected by assigning it to the majority class among the chosen nearest neighbours. It will often make sense to assign more weights are assigned to the nearest neighbours, and a

voting system is implemented where the neighbours decide the class of q, based on the inverse of their distance from it.

$$Vote(y_i) = \sum_{c=1}^{k} \frac{1}{d(q,x_c)^n} 1(y_j, y_c) \quad (12)$$

$1(y_j, y_c)$ results in 1, if the class labels match and 0, if otherwise. n, is usually 1, even though greater values can be applied to minimise the effect of further distant neighbours. The commonest distance metrics used in KNNs are Euclidean distance – which measures distance as the straight line between two points

$$- \sqrt{\sum_{k=1}^{n} (x_{ik} - x_{jk})^2}, \quad (13)$$

and Minkowski distance - which uses the distance vector length which must be non-negative.

$$MD_p(q, x_i) = \left(\sum_{f \in F} |q_f - x_{if}|^p\right)^{\frac{1}{p}} \quad (14)$$

2.1.3.5 Gaussian Naive Bayes (GNB)

Naive Bayes (NB) refers to a set of supervised learning models used for predictive purposes. They are simple and effective models which learn the probabilities of certain features belonging to a given group[35].

The name arises as a result of the assumption that the occurrence of an event is not dependent on the occurrence of other events. A Naive Bayes model is rooted in the Bayes theorem and the Bayesian rule, which computes the probability that a feature belongs to a specific class. The Naive Bayes classifier uses two assumptions[35]:

i. Given the class label, features are conditionally independent of each other and contribute equally to the process.

ii. No latent feature affects the class label prediction process.

Suppose a vector $(x_1, x_2, \ldots x_n)$ represents the n features of X. Let θ represent the class label of X. The naïve theorem describes the conditional probability of observing X given the class label θ, $P(X)$ as a product of several simpler probabilities as shown below[35]:

$$P(f_1, f_2, \ldots f_n) = \frac{P(\theta)P(f_1, f_2, \ldots, f_n|\theta)}{P(f_1, f_2, \ldots f_n)} \quad (15)$$

Where $(f_1, f_2, \ldots f_n)$ represents the features of vector X. Following the assumption of independence, the probability can be expressed as;

$$P(f_1, f_2, \ldots, f_{i-1}, f_{i+1}, \ldots f_n) = P(f_i|\theta) \quad (16)$$

For all i

$$P(f_1, f_2, \ldots f_n) = P(\theta)\frac{\prod_{i=1}^{n} P(f_i|\theta)}{P(f_1, f_2, \ldots f_n)} \quad (17)$$

Assuming every feature is constant, the classification rule follows below:

$$P(f_1, f_2, \ldots f_n) \alpha\, P(\theta) \prod_{i=1}^{n} P(\theta), \quad (18)$$

And,

$$\hat{\theta} = arg P(\theta) \prod_{i=1}^{n} P(\theta) \quad (19)$$

It is worth noting that for the GNB model, the probability of feature occurrence follows a Gaussian distribution

$$P(\theta) = \frac{1}{\sqrt{2\pi\sigma_\theta}} exp\left(\frac{-(f_i-\mu_\theta)}{2\sigma_\theta^2}\right) \quad (20)$$

whose parameters σ_θ and μ_θ are computed using maximum likelihood.

2.1.3.6 Multilayer Perceptron (MLP)

The perceptron is the most basic type of neural network used to classify specific kinds of linearly separable patterns[37]. The system is made up of one neuron with adaptable weights and biases. Perceptron model will converge, and the decision surface will be positioned as a hyperplane between the classes if the patterns (vectors) used in training them are taken from linearly separable classes. This proof is popularly dubbed the perceptron convergence theorem[37].

The single-layer perceptron, which has just one neuron can only conduct binary classification i.e., distinguish between two classes.

$$v(x) = \sum_{i=0}^{n} w_i x_i = w^T x \qquad y = \{1,\ v \geq 0\ 0,\ v < 0 \quad (21)$$

A neuron's processing unit is generally expressed as

$$a = \emptyset\left(\sum_i w_i x_i + b\right) \quad (22)$$

where x_i are the inputs, w_i are the sample weights, b is the model bias, $\emptyset$ is the nonlinear activation function, and a specifies the activation function.

The MLP is a simple form of feed-forward neural network. In feed-forward networks, all units are placed in a series of layers, with every layer containing certain identical units[37].

When all units in each layer are interconnected with every unit in the following layer; the network is termed "fully connected". The input layer is the first layer, and its units are the input features. The middle layer(s) denotes the hidden units which perform

activations and computations. The output layer is the last layer, which specifies output values. The depth of the network is defined by the number of layers, while the width, describes the number of units.

Assuming we denotes input units as x_i, hidden units as h_i, and output units as y. It is then assumed that the system is fully connected, and every unit and every layer receives connections from every unit and layer. This then implies that every unit has its corresponding error and a specific weight for each pair of units in subsequent consecutive layers. The fully connected network is then expressed as:

$$h_i^1 = \emptyset^1\left(\sum_j \quad w_{ij}^1 x_j + b_i^1\right)$$

$$h_i^2 = \emptyset^2\left(\sum_j \quad w_{ij}^2 h_j^1 + b_i^2\right)$$

$$y_i = \emptyset^3\left(\sum_j \quad w_{ij}^3 h_j^2 + b_i^3\right) \tag{23}$$

The activation function is then represented as a vector of units, and weights as a matrix. The resulting vector is:

$$h^1 = \emptyset^1(W^1 X + b^1)$$

$$h^2 = \emptyset^2(W^2 h^1 + b^2)$$

$$y = \emptyset^3(W^3 h^2 + b^3) \tag{24}$$

Finally, transpose the vectors into a single matrix H, for computational ease.

$$H^1 = \emptyset^1(XW^{(1)T} + 1b^{(1)T})$$

$$H^2 = \emptyset^2(H^1 W^{(2)T} + 1b^{(2)T})$$

$$Y = \emptyset^3(H^1 W^{(3)T} + 1b^{(3)T}) \tag{25}$$

BackPropagation

Backpropagation is a network process that permits the MLP to repeatedly tune the network weights in a bid to minimise the average loss of the training dataset[38]. For the backpropagation process to be effective, the weighted sum, and the threshold function, e.g. (ReLU) must be differentiable. The optimization function in MLP is the Gradient Descent which has a bounded derivative.

Bagging (Bootstrap Aggregating)

Bagging is the most straightforward ensemble method. It fits each base classifier on randomised subsets of the original data, drawn with replacement, and then piles up the separate predictions (by averaging or voting) to produce a final classifier[40]. It helps in the reduction of variance in a classification model, by adding randomness to its building. Bagging is an important concept in machine learning because it avoids overfitting data[41]. It is commonly applied to decision trees but can also be used on other classifiers. In a study that compared the classification accuracy amongst seven classifiers, their results showed that stacking, bagging, J48, NB, and linear SVM performed exceptionally with an accuracy of 100%, while KNN (k=3) and Ada boost trailed at 98.6% and 98.3%[42]. In bagging, the final prediction is undertaken by majority voting.

Steps in the execution of bagging

i. Considering n instances and m input variables in the training dataset, a random sample is selected without replacement.
ii. A subset of m features is randomly chosen from the sample observations to make a model.

iii. The input feature producing the best results is used in the split of the nodes.

iv. The decision tree is regrown to obtain the best root nodes.

v. The steps are repeated n times, to sum the results of all decision trees to produce the best predictive value.

Advantages of Bagging

i. Overfitting of data is minimised

ii. The accuracy of the model is usually improved

iii. Huge volumes of data are efficiently dealt with

The hyperparameters of a bagging algorithm are base_estimator, n_estimators, max_samples, max_features, bootstrapbool, bootstrap_featuresbool, oob_scorebool, e.t.c. Another work pointed out that bagging hardly improves *K*NN because it can produce accurate classifiers through bootstrap resampling; which is very efficient in unstable methods like neural networks and decision trees.

Boosting

Boosting is an ensemble method that aims to create a better classification model from a combination of serial weak classifiers. In the boosting method, weights are assigned to the training datasets at every iteration[43]. After every iteration, higher weights are assigned to mislabeled observations and lesser weights to the correctly labeled ones. Boosting is important in dealing with variance and bias.

Examples of boosting techniques are extreme gradient boosting (XGboost), Adaptive boosting (Adaboost), gradient boosting, light gradient boosting machine (Light GBM), and CATboost (Category boosting)[43].

Adaptive Boosting: AdaBoost creates a strong learner through multiple iterations of adding a weak learner in each cycle[44]. The weight vector is also tweaked to account for previously misclassified sample points. Hence, the resulting classifier has higher accuracy. Adaboost is not robust to outliers and noise. The hyperparameters in Adaboost are learning_rate, n_estimators, and base_estimator.

Gradient Boosting: Gradient boosting is an improvement of Adaboost, which aims to minimize the loss function by using the gradient descent optimization method while combining weak learners[45]. The loss function estimates the best model depending on the problem task. Weak learners are added based on the additive model component. Gradient boosting is notably improved as it inculcates subsampling, which encourages randomness of the model, shrinkage reduces the impact of each added learner, and the size of added steps, thus penalising consecutive iteration.

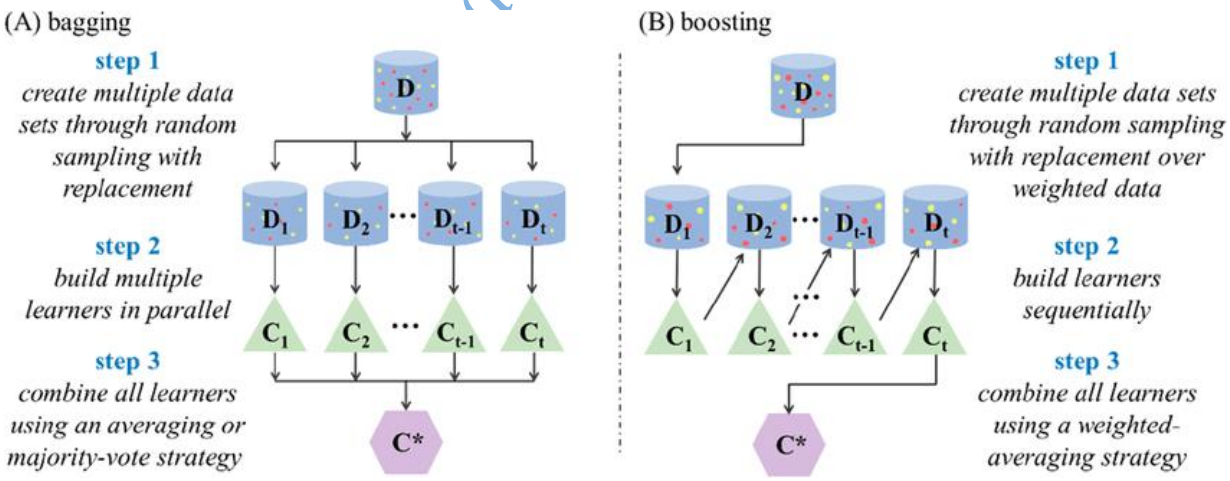

Figure 2.7 : An Illustration of Bagging and Boosting Methods[46].

Extreme Gradient Boost (XGboost): XGBoost is an improvement of the gradient boost designed to improve speed and performance. This technique employs regularized learning for smoothing and shrinkage to reduce the impact of each tree

and make room for future trees to aid improvement, and feature subsampling to prevent over-fitting[46]. All these features, speed up the run time of the algorithm. Parameters of XGboost which can be tuned by the user are eta, gamma, max_depth, seed, eval_metric etc.

Light Gradient Boosting Machine: LightGBm uses two gradient-based methods: exclusive feature bundling (EFB) and gradient-based onside sampling (GOSS)[47].

GOSS operates by excluding the portion of the dataset with relatively small gradients and then uses the remaining data to compute the overall information gain. The EFB uses the mutually exclusive features in the dataset, and non-zero values simultaneously to minimise the number of features. This enhances the overall accuracy of the split point.

Stacking: As opposed to the previous methods which use homogenous weak models, stacking employs heterogeneous weak learners. Learning occurs simultaneously, and then they are combined by training a meta-learner, which then makes predictions using the various models' predictions.

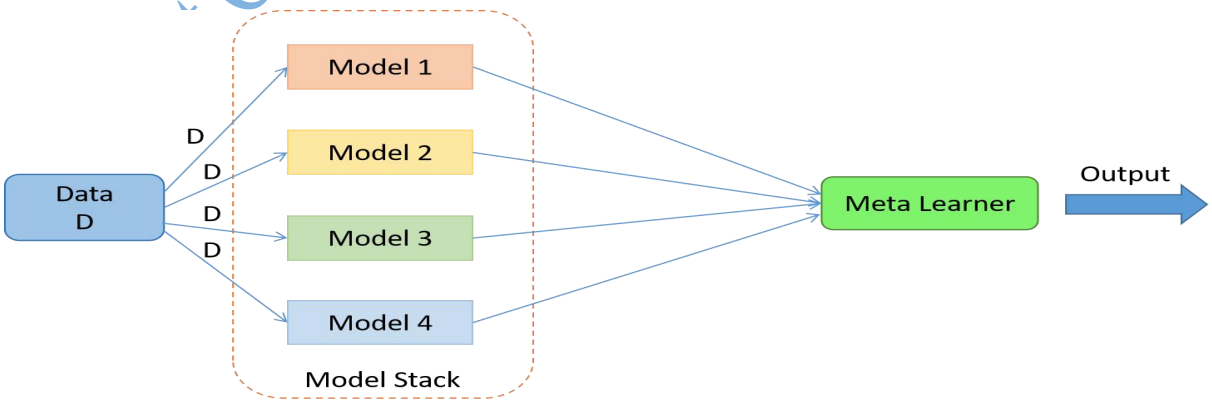

Figure 2.8: Stacking Workflow[47].

2.1.4 Neural Networks

Currently, the utilisation of Artificial Neural Networks (ANNs) has gained significant traction across diverse domains of human necessities. Numerous organisations are currently allocating resources towards the implementation of neural networks as a means of addressing issues within diverse fields, including the economic sector. These areas have traditionally been within the purview of operations research.

Artificial intelligence is notable for its application in data analysis across various disciplines, including social science and arts, in addition to its established utility in science and engineering.

This versatility is attributed to the extensive range of applications of artificial intelligence. In the present day, there has been a widespread implementation of artificial intelligence (AI) in various domains such as industrial production, petroleum exploration, and business environments, primarily for the purpose of optimising processes. One notable benefit of utilising artificial neural networks (ANNs) is their ability to enhance the usability and precision of models derived from intricate natural systems featuring extensive inputs[48]. The Artificial Neural Network (ANN) is a contemporary and advantageous computational model utilised in the domains of problem-solving and machine learning.

The ANN, or artificial neural network, is a model for managing information that bears resemblance to the functioning of the human brain's biological nervous system[48]. In recent times, there has been a significant surge in global research attention towards the functionality of the brain. Artificial neural networks (ANNs) are designed to

emulate the cognitive processes of the human brain in order to perform specific tasks. The human brain is characterised by its significant size and remarkable efficiency[48]. The human brain can be likened to a sophisticated information-processing device that is capable of executing a range of intricate signal computing operations, which can be effectively synchronised to achieve a specific objective.

A prototypical instance of a neural network function is exemplified by the human brain, which is interlinked to transmit and receive signals for the purpose of human action. The independence of neural network (NN) layers is evidenced by the fact that a given layer can accommodate a variable number of nodes[48]. The node that is assigned an arbitrary numerical value is commonly referred to as the bias node. The bias nodes are consistently assigned a value of one.

Analogously, it can be observed that the bias nodes bear resemblance to the offset in linear regression, which is expressed as $y=ax + b$. Here, "a" represents the coefficient of the independent variable "x", while "b" is commonly referred to as the slope[48].

The primary purpose of a bias in a neural network is to furnish a node with a trainable constant value, which supplements the standard inputs received by the node. The bias value holds significant importance as it allows for the displacement of the activation function towards either the right or left, thereby contributing to the analytical aspect of achieving successful training of Artificial Neural Networks (ANNs). When utilising the neural network as a classifier, the input and output nodes will correspond to the input features and output classes, respectively.

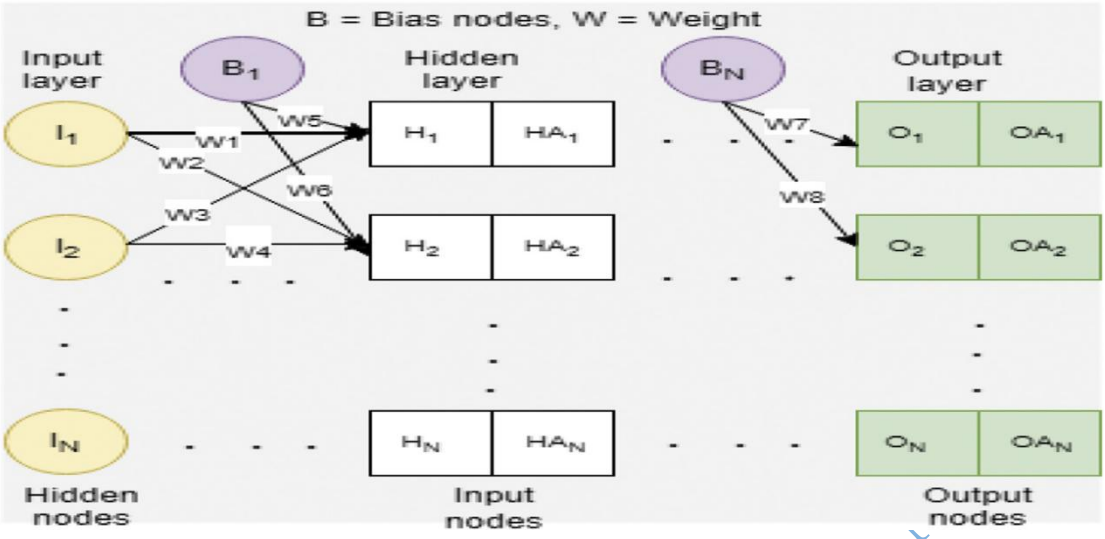

Figure 2.9: Framework for Artificial Neural Networks Classification[48].

However, when the NN is used as a function approximation, it generally has an input and an output node. However, the number of designed hidden nodes essential greater than those of input nodes.

2.1.5 Applications of Neural networks

Neural networks (NNs) have been widely applied to real-world problems in various domains such as business, education, economics, and other areas of life.

This is due to their ability to function effectively and efficiently, as well as their practical applications and uses. Neural networks have been found to be useful in the fields of intrusion detection and data classification using optimisation methods[48]. This has been demonstrated in previous research studies. The utilisation of machine learning (ML) methodologies has been prevalent among researchers in addressing classification problems. Neural networks are proficient in recognising trends and patterns within data, making them well-suited for the purposes of prediction and forecasting[48].

2.1.6 Classification of ANN

ANN can be classified as depicted in figure 2.10. A feed forward neural network (FFNN) is a machine learning classification algorithm that made up of organized in layers that are similar to human neuron processing units.

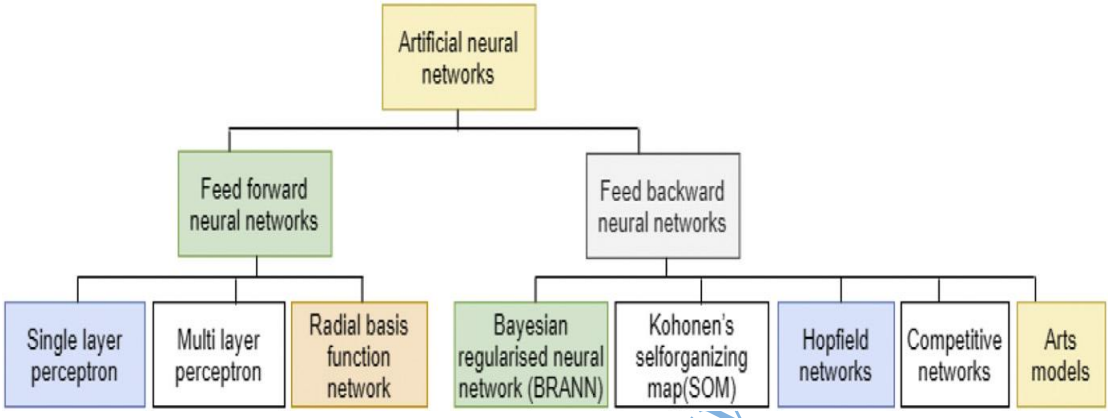

Figure 2.10: Framework For Artificial Neural Networks Classification[48].

In a feedforward neural network (FFNN), each unit within a layer is interconnected with all other units in the same layer. The connections between layers comprising units are not uniformly equivalent, as each connection may possess a distinct weight or level of potency. The magnitudes of the network connections' weights serve as an indicator of the network's potential knowledge capacity. NN units are commonly referred to as nodes. The process of information processing within a network entails the initial input of data from input units, which subsequently traverses through the network, progressing from one layer to the next, until it ultimately reaches the output units. This process has been documented in literature [48,50].

In the normal operation of a neural network as a classifier, inter-layer feedback is absent. The feedforward neural network (FFNN) is characterised by unidirectional information flow, whereby data is transmitted solely from the input nodes to the hidden nodes, if present, and subsequently to the output nodes. These neural networks

are referred to as feedforward due to their behaviour. Instances of feedforward neural networks (FFNNs) include the single-layer perceptron and the multilayer perceptron. Illustrated in Figure 2.11 are the components of a two-layered network, consisting of three input units, four hidden layer units, and five output layer units, represented by circular nodes[50].

The applications of FFNN are categorised into two distinct areas, namely the control of dynamical systems and spaces where conventional machine learning techniques are utilised[50].

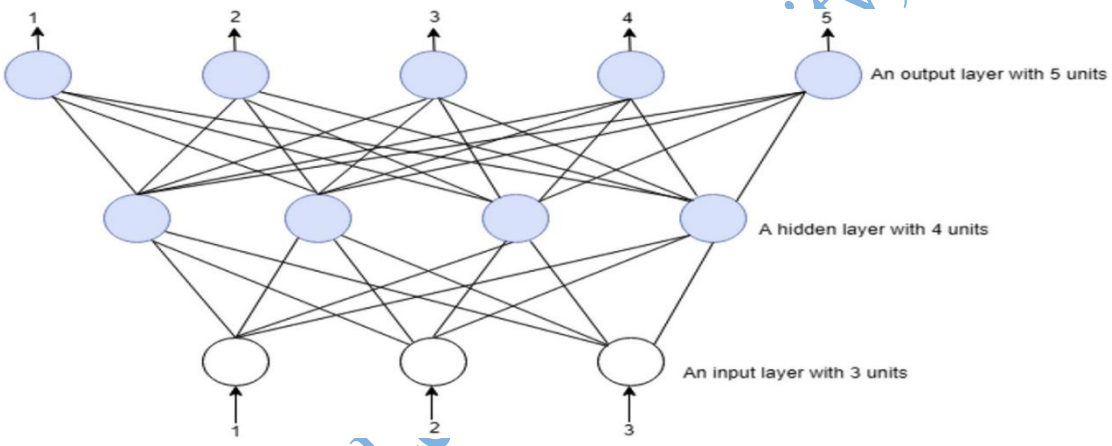

Figure 2.11. Two-Layered Feedforward Neural Network[48].

NNs with two or more hidden layers are called deep networks because the network has become complex with more than 1 hidden layer.

Unlike FFNN, the feed-backward neural network (FBNN) can use internal state "memory" (store information) to process sequence of data inputs[50]. That means FFNN can logically handle task according to first come first serve bases of inputs. Feed-backward NN can applied to tasks like un-segmentation, and pattern recognition (connected handwriting recognition). Feed-backward neural network application areas include mathematical proofs, seismic data fitting, medicine, science, engineering, classification, function estimation, and time-series prediction, etc[50].

An architecture of FBNN illustrated in Figure 2.11. In feedback NNs or backpropagation, connections between nodes produced a coordinated graph in sequence. The coordinated graph in sequence allows feedback NNs to demonstrate dynamic terrestrial behaviour for a time sequence. Examples are Kohonen's self organizing map and recurrent neural network (RNN)[50].

RNN referred to a standard kind of neural network which extended over time, with edges that feed into the next time step rather than feeding into the next layer concurrent time of step[50]. RNN is constructed to sequences recognition, for instance, a text or a speech signal. It has cycles within that indicates presence of short-memory in the net. Unlike a recurrent neural network, an RNN is like a hierarchical network where the input need processing hierarchically in the form of a tree because there is no time to the input sequence.

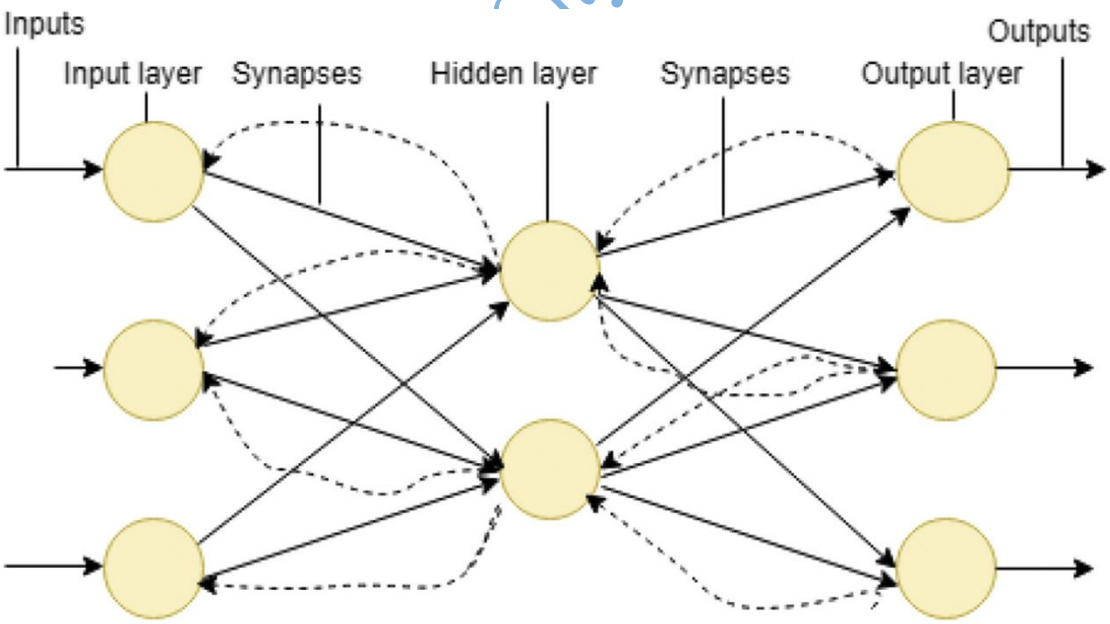

Figure 2.12. Feed-Backward Neural Network[50].

2.2 Methodological Framework

This section gives a theoretical background of the main classification algorithms used in this study.

2.2.1 Random Forest (RF) Algorithms

Random forest is a supervised ensemble method that uses a collection of numerous decision trees to make predictions[51]. Random Forest is a classifier consisting of a set of tree-structured classifiers with identically distributed independent random vectors and each tree casting a unit vote at input x for the most popular class[52]. A random vector that is independent of the previous random vectors of the same distribution is generated and a tree is generated using the training test , an upper bound is extracted for Random Forests to get the generalization error in terms of two parameters Exactitude and interdependence of individual classifiers[53,54].

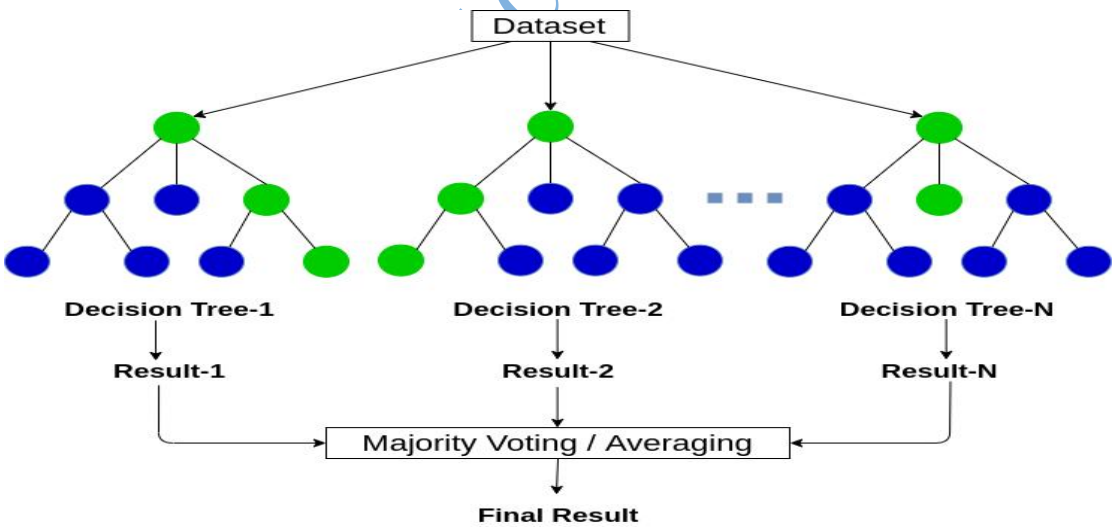

Figure 2.13: Random Forest Flow Chart[54].

To get multiple subsets of samples, it implements the bootstrap method, creates a Decision Tree utilizing each subset of samples, and combines several Decision Trees into a Random Forest[55]. When the sample to be classified is reached, the final

outcome of the classification is decided by a vote on the Decision Tree[56]. Generally, scholars increase the precision of the classifier starting from the classifier and reduce the association between classifiers[57].

Random Forest algorithm in the classification process, where the effects of the classification of each base classifier have a common distribution of errors, the final reduction of the classification effect is accomplished[58]. Takes the test characteristics and uses the rules of each randomly generated Decision Tree to forecast the result and store the expected result (target). Determine the votes for each predicted goal. Consider the predicted high-voted goal as the final prediction from the Random Forest algorithm[59,60].

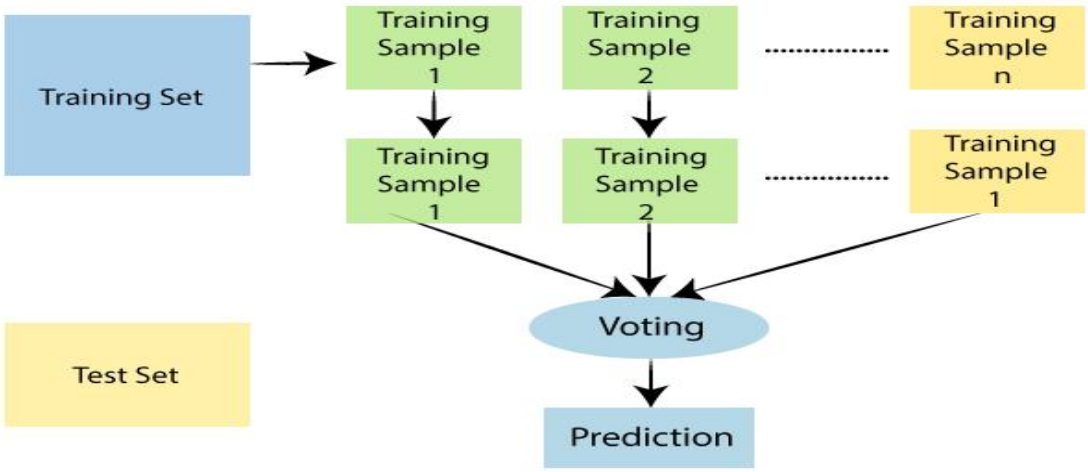

Figure 2.14: Random Forest training Flow Chart[54].

The Random Forest's basic algorithm steps are as follows: In the Random Forest algorithm, there are two steps, one is Random Forest formation, and the other is to make a guess from the first step of the Random Forest classifier[61].

1. Select "K" features at random from the complete "m" features, where k << m.

2. Calculate the node "d" among the "K" features using the best split point.

3. Using the best division to divide the network into daughter nodes.

4. Repeat measures from 1 to 3 until the number of nodes 'n' has been reached.

5. Develop a forest to build the "n" number of trees by repeating steps 1 to 4 for "n" number of times

The RF algorithm is very efficient, as it handles datasets that contain continuous variables, as well as categorical variables robustly. An RF classifier contains subsets of various tree classifiers $\{h(x, \Theta_k), k = 1,2, ...\}$ where the Θ_k are independently and identically distributed random vectors, with each tree being able to specify the modal class at input x[62]. The performance index, which solely approximates the confidence interval (CI) of the RF model is given as

$$mg(x, y) = av_k I(h_k(x, \Theta_k) = y) - av_k I(h_k(x, \Theta_k) = j) \quad (26)$$

where $I(.)$ denotes an indicator function, and $av(.)$, the average value. It is observed that as the margin increases, the confidence level also increases. The generalisation error becomes

$$PE^* = P_{x,y}(mg(x, y) < 0), \quad (27)$$

where $P(.)$ denotes probability. With an increase in trees for all sequences Θ_k, PE^* converges to

$$P_{x,y}(P_\Theta(h(x, \Theta) = y) - P_\Theta(h(x, \Theta) = j < 0) \quad (28)$$

Convergence of this generalisation error proves that the RF model does not overfit as more trees are introduced. The upper bound for the generalisation error is given as

$$PE^* \leq \frac{\underline{\rho}(1-s^2)}{s^2} \ , \tag{29}$$

where $\underline{\rho}$ is the average correlation value, s is the strength of each tree in the model. An increased strength of individual trees and a low correlation between them produces more accurate prediction results.

Advantages of RF Algorithm

i. There is greater accuracy. Effective in working with large databases
ii. It manages thousands of input variables quickly and effectively.
iii. Provides information on variables that are important and are not in the Classifying.
iv. Provides techniques to estimate incomplete data.
v. Deals with lost details without losing accuracy.
vi. Prototypes are used to provide data or meta data on the relationship between different factors
vii. Permits the analysis of variable relationships

Disadvantages

i. One of the main problems found is over-fitting a single data set, especially in the tasks of regression
ii. Random Forests have trouble dealing with multi-valued and multivalue attributes Multi-dimensionally. They prefer multi-level categorical variables.

2.2.2 Decision Tree

Decision trees are one of the powerful methods commonly used in various fields, such as machine learning, image processing, and identification of patterns[63]. DT are a

successive model that unites a series of the basic test efficiently and cohesively where a numeric feature is compared to a threshold value in each test. The conceptual rules are much easier to construct than the numerical weights in the neural network of connections between nodes. Mainly for grouping purposes, DT is used. Moreover, DT is a usually utilized classification model in Data Mining. The nodes and branches are composed of each tree[63]. Each node represents features in a category to be classified and each subset defines a value that can be taken by the node. Because of their simple analysis and their precision on multiple data forms, decision trees have found many implementation fields.

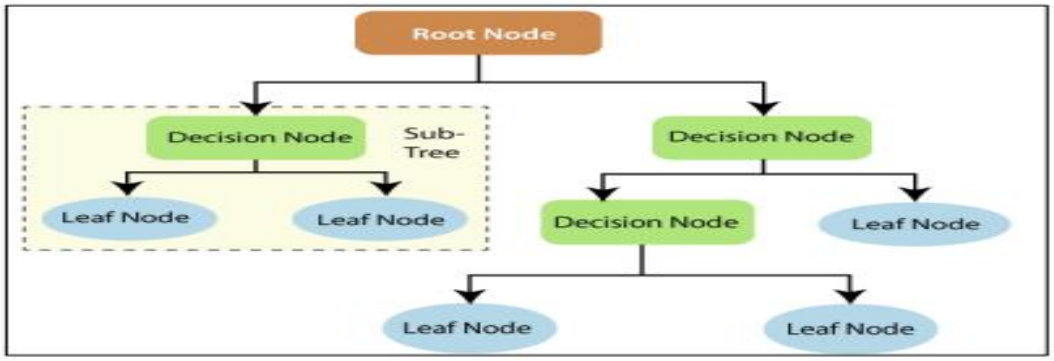

Figure 2.15. Decision Tree Flow Chart[63].

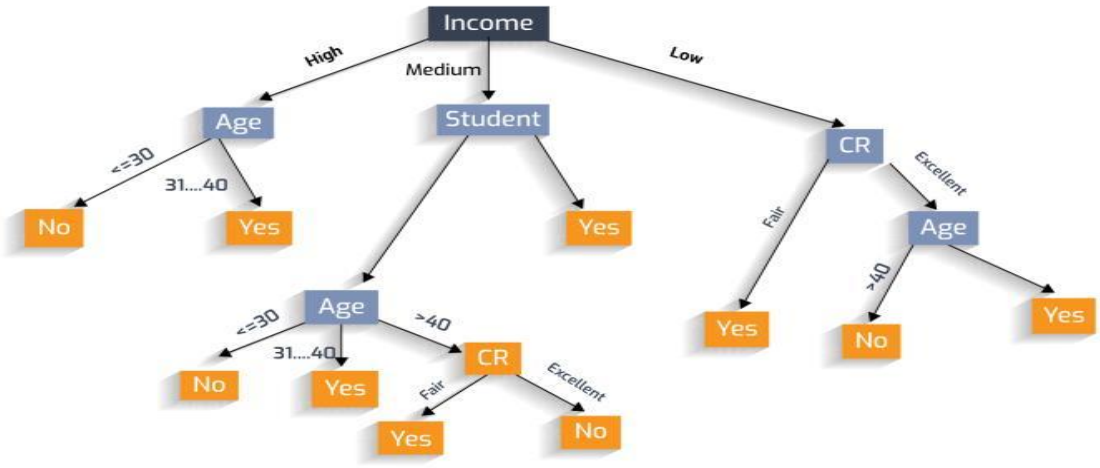

Figure 2.16. Decision Tree[63].

2.2.2.1 Types of Decision Tree Algorithms :

There are several Types of DT algorithms such as: Iterative Dichotomies 3 (ID3), Successor of ID3 (C4.5), Classification And Regression Tree(CART), CHi-squared Automatic Interaction Detector(CHAID), Multivariate Adaptive Regression Splines (MARS), Generalized, Unbiased, Interaction Detection and Estimation (GUIDE), Conditional Inference Trees (CTREE), Classification Rule with Unbiased Interaction Selection and Estimation (CRUISE), Quick, Unbiased and Efficient Statistical Tree (QUEST)[64].

In decision algorithms, entropy and information gain entropy is employed to measure a dataset's impurity or randomness. The value of entropy always lies between 0 and 1. Its value is better when it is equal to 0 while it is worse when it is equal to 0, i.e. the closer its value to 0 the better. As shown in "Figure 2.11". If the target is

$$Entropy\ (S) = \sum_{i=1}^{c} \quad P_i\ log\ 2^{p_i} \tag{30}$$

Where P_i is the ratio of the sample number of the subset and i-th attribute value.

Benefits of Decision Tree

The DT algorithm is part of the supervised learning algorithm family, and its main objective is to construct a training model that can be used to predict the class or value of target variables through learning decision rules inferred from the training data. The DT algorithm can be used to

i. solve regression and classification problems
ii. Simple to comprehend
iii. Quickly translated to a set of principle for production

iv. Can classify both categorical and numerical outcomes, but the attribute generated must be categorical

However, DT has some draw backs which include;

i. The optimal decision making mechanism can be deterred and incorrect decisions can follow

ii. There are lots of layers in the decision tree, which makes it interesting

iii. For more training samples, the decision tree's calculation complexity may increase

2.3 Review of Related Works

2.3.1 Classification Based on Decision Tree Algorithm for Machine Learning.

A decision tree was used in several machine learning and data mining tasks as a classifier.

In a study that utilized a decision tree (j48), Random Forest (RF), and neural network algorithms for diabetes mellitus prediction. The dataset is physical research data for hospitals in Luzhou, China. There are 14 characteristics involved.

Training array randomly extracts data from 68994 stable human and diabetic patients, respectively. They used the full significance of minimum Redundancy Maximum Relevance (mRMR) and Principal Component Analysis (PCA) to minimize dimensionality. In some ways, the effects of RF, as opposed to each other, seemed to be higher than the other classifiers. Also, 0.8084 is the best outcome in the Luzhou data collection[65].

In another study that utilized the DT classification process to classify the handwritten digits of the standard data set of kaggle digits and estimate the accuracy of the model for each digit from 0 to 9. The kaggle features include 42,000 rows and 720 columns

used for machine training, vector features are used for pixels of digital images. They used a highly efficient language named "python programming" for the application of machine learning algorithms to map the classifier's success rate graph in the realization of handwritten digits. The findings suggested that the 83.4% accuracy and decision tree classifier had an impact on handwritten number recognition[66].

In a study that suggested a decision tree algorithm to recognize known and novel clinical indications before treatment for survival in Locally Advanced Rectal Cancer (LARC). The analytics showed that even non-experts in the field, in particular classification trees, can easily interpret the tree-based machine learning process. Validation errors need to be managed to even achieve their statistical capacity. Around 2007 and 2014, patients with histologically confirmed LARC had their data checked. The Kaplan-Meier approach has been used to determine overall survival (OS). It involved a total of 100 patients. 76.4 % and 71.3% were the 5-year and 7-year OS points. Age, comorbidity, tumor size, Clinical Tumor classification (CT), and clinical node classification are important predictive variables for tree composition (CN). The results showed that the highest survival rates were in elderly patients with a tumor size of less than 5 cm and patients under the age of 65 years who had cT3. A decision tree is a way of getting better clinical practice decision-making, based on broad data sets[67].

Another work presented a Behavioral Decision Tree named "BehavDT" context-aware structure that takes into account consumer behavior-oriented generalization according to the degree of personal choice.

In exceptional cases of association, the BehavDT model provided comprehensive decisions as well as context-specific decisions. Experiments were carried out on real

smartphone datasets of individual users through the efficiency of the BehavDT model. The results indicated that the Behav DT context-aware model, whose accuracy is up to 90%, is the model that is most energetic compared to other conventional machine learning models[68].

In a work that illustrated the first practical algorithm to optimize decision trees for binary variables. The algorithm is a co-design of analytical limits involving a dedicated bit vector library and data structures that minimize the search area and current application technologies. They used the Binary Optimal Classification Trees (BinOCT) method, which is the current publicly available method, to assess the accuracy and compare it with the Optimal Sparse Decision Trees (OSDT). As well as they utilized text datasets from the University of California, Irvine (UCI) Machine Learning Repository and numeric datasets from the other ProPublica COMPAS datasets. The findings showed that when a COMPAS dataset, the optimal decision tree produced by OSDT, its accuracy 66.90 %. Besides, when BinOCT and OSDT generated the UCI dataset, decision trees, their accuracy is 76.722 %, 82.881 %, respectivel[y69].

Another study introduced a Distributed Spark Tree (DST) to better execute the DT algorithm in terms of model construction time without losing accuracy. Besides, they suggested using them in Spark's climate. Data in Spark's shared architecture does not perform horizontal parallel execution. Spark functions well and coherently in-memory computations, RDD, and map reduction. The dataset that was used from the UCI ML repository and four classes were chosen.

Wide data files are utilized to test performance regarding model build time for DST, PySpark (PT), and MLLib (MLT). The findings showed that in terms of accuracy,

DST performed better than both PT and MLT, as its lowest value was 81.445 % and the highest according to the scale of the dataset was 99.9 %[70].

Another work offered a modern approach, namely a Pixel Label Decision Tree (PLDT), and checks whether it can achieve better detailed femur segmentation efficiency in DXA imaging. PLDT includes extraction and selection of the trait. PLDT was used to uncover secret patterns found in DXA pictures in contrast to photographic images. To decide the best feature set for the model, PLDT generates seven new feature sets and uses Global Threshold (GT), Region Growing Threshold (RGT), and Artificial Neural Networks (ANN). The results revealed that in segmenting DXA images, PLDT exceeds other conventional partition techniques. For each algorithm such as this PLDT, the accuracy is 91.4%, GT is 68.4%, RGT is 76%, and ANN is 84.4%[71].

In a study that proposed a new approach that affects the amplitude of signals from the Global Navigation Satellite System (GNSS) and was used to detect ionic scintillation events that are concerned with accuracy, reliability, and readiness. A broad collection of 50 Hz post-correlation data was supplied by the GNSS recipient.

The outcomes showed that this method, in terms of accuracy and F-score, exceeds state-of-the-art techniques and can achieve a human-driven standard, which is the level of manual annotation. It improves greatly as it gains 98 % of identification, very similar to handdriven human-driven classification[72].

In a work that proposed a structure based on a decision tree named Screen Content Coding (SCC) to make a fast decision in situations by testing their different features in the training sets. Moreover, to prevent the thorough search process, a sequential arrangement of decision trees was illustrated. In addition, SCBs were used as datasets

to balance the SCC with the Intra Block Copy (IBC) and PaLeTte (PLT) modes. The results indicated that the SCC system offers a 47.62 % decrease in computational complexity on average, with a small 1.42 % in Bjøntegaard delta bitrate (BDBR)[73].

Another study demonstrated a comparative analysis of accuracy and process length for each algorithm performed using the K-Nearest Neighbor (KNN) and Decision Tree (DT) algorithms for the detection of DDoS attacks. Moreover, they used the CICIDS2017 dataset that consists of the latest attacks and global packages, is standard and applicable to real-world data in a PCAP format. The findings showed that the accuracy of DT to detect DDoS attacks was higher than the KNN value, the accuracy of DT was 99.91 %, and the accuracy of KNN was 98.94 %[74].

In a study that presented a system to identify up to 10 irregular red blood cells and to know the accuracy rate for all abnormal red blood cells. Additionally, To detect irregular red blood cells, they employed a DT algorithm in image processing and used frames of former patients for the scheme in hospitals. Also, the camera was used to insert them into the software to capture the slides. The results showed that the accuracy rate averaged 89.31 % and the error rate averaged 10.69 %.

Furthermore, the central irregularity of the Codocyte pallor was found to be a cause for the mistake in the classification of abnormal red blood cells[75].

Another study proposed a model based on the decision tree machine learning algorithm named Extreme Gradient Boosting (XGBoost) for the prediction of regular smoking time. Furthermore, to create a simulated data set for smoking time data, the Chinese Center for Disease Control and Prevention collected people's information from smokers. Also, they used a module for extracting feature information. To see its output in the feature extraction module, they used the decision tree (XGBoost)

module and Random Forest machine learning algorithms. The results showed that DT efficiency is higher than RF, achieving 84.11 % with DT accuracy, while 58.11 % with RF accuracy[76].

In a study that discussed the effective methods of developing a machine learning model using some of the common algorithms that can distinguish whether mail is spam or ham. UCI's Machine Learning store was used as a dataset for Spambase. Besides, they evaluated the output of Logistic Regression (LR), DT, Naïve Bayes (NB), KNN, and Support Vector Machine (SVM) to construct an efficient machine learning model for spam. Using the Weka tool to train and evaluate the data collection. The results indicated that DT performance is comparable to and better than KNN performance, and the accuracy for both of them is as follows: DT is 99.93 percent, KNN is 99.93 %, LR is 93.13 %, SVM is 90.76 % and NB is 79.52 %[77].

In a study that developed a new machine learning approach for the hybrid decision tree and a genetic algorithm known as GADT for spam detection. The most significant algorithm for enhancing decision tree efficiency is the genetic algorithm. Also, it is efficient and reliable for text classification. A genetic algorithm has used the element of trust that governs decision tree pruning to optimize and detect its optimum value. They used the UCI Machine Learning Store spam dataset. Besides, they used the mechanism of main Principle Component Analysis (PCA) to delete features that are inappropriate for email message content and process them less frequently. The findings showed that after using PCA, the mixed GADT approach has an accuracy of 93.4 % before using PCA and an accuracy of 95.5 %. This implies that the extraction of inappropriate characteristics has a great impact on the PCA[78].

In a work that implemented a Principle Component Analysis (PCA) feature extraction algorithm to decrease the dimensions and demonstrate the high dimensions analyze evidence on gene expression. The KNN classification and DT algorithm were utilized to detection various biological structures and to Offer better value resolution as well as to detect new malaria genes and prediction tests. Ribonucleic acid (RNA-seq) sequencing is also used as a data collection. The results indicated that the performance of the KNN classification is better than the DT classification in the PCA feature extraction. The accuracy of KNN reaches 86.7% while the DT reaches 83.3%[79].

In a proposed work on a new technique that recognizes and removes the blood artery for correct segmentation of the Optic Disc (OD). This is done in two ways. First, the directional filter is used to build an efficient blood vessel identification and exclusion algorithm. In the second step, to detect the contour of the optic disc, the decision tree classifier is utilized to achieve an adaptive threshold. As well as, two separate databases were used, including 300 fundus images obtained from Kasturba Medical College (KMC) Manipal and also the RIM-ONE database that is publicly accessible. The results showed that a fully automatic OD segmentation technique that uses a decision tree classifier to achieve the segmentation threshold improves the robustness of the algorithm even for images containing exudate, vesicle atrophy, and reversals, Hence, resulting in an appropriate fractionation of OD.

The mathematical study demonstrates the effect of pretreatment. Therefore, the average values of accuracy obtained for KMC images are 99.61% and for the RIM-ONE database, the obtained average values of accuracy are 99.15%[80].

Another study introduced the Self-Inertia Weight Adaptive Particle Swarm Optimization with Gradient Base Local Search (SIW-APSO-LS) feature selection

approach was modified to conduct feature selection and the C4.5 decision tree method was used as a classifier to determine the sub-sets of features given. When comparing algorithms in feature selection problems, 16 datasets from the UCI Machine Learning Repository were used for the experiments. The experimental outcomes demonstrate that SIW-APSO-LS simplifies the collection of features by effectively decreasing the number of features picked, thus maintaining the best precision compared to other literature selection approaches for the same test functions. In the field of attribute collection, the experimental findings showed that the proposed approach is useful and the highest accuracy obtained from a total of 16 datasets is 99.88%[81].

In a study that proposed a new Intrusion Detection System (IDS) that incorporates diverse classification systems that are DT-based and rule-based concepts, namely the REP tree, JRip algorithm, and Forest PA. In specific, the first and second approaches take data set features as inputs and categorize the network traffic as Attack/Benign. In comparison to the results both the first and second classifiers for reference, the third classifier uses the attributes of the original data collection. The research findings achieved by using the CICIDS2017 dataset to analyze the IDS testify to their dominance in terms of accuracy, identification rate, false alarm rate, and time overhead relative to current state-of-the-art schemes. In thorough, with 94.457%, our model has the highest DR, the highest precision with 96.665%, and the lowest FAR with 1.145%, although its low computing time makes it quickly implemented into a soft real-time system[82].

A work provided an evidentiary decision tree to classify the fuzzy data set and the ding entropy has been used as an indicator of the partition rules for its construction. Moreover, the Basic Belief Assignment (BBAs) of Iris and wine Datasets are utilized

to calculate the optimal splitting feature. The lower the entropy of Deng, the more effective the feature will be to characterize the samples. In contrast to the standard mixture rules employed for the combination of BBAs, the proof DT can be extended specifically to the classification. The findings showed that the implementation of the proof DT based on conviction entropy effectively decreases the complexity of the fuzzy data classification whether the patient is either affected by the cancer type of Malign or Benign. The Wisconsin Breast Cancer dataset, containing 32 attributes and 569 data, was used. They were using a 10 fold cross-validation test to identify and analyze the algorithms. The accuracy is 95% when using Wine datasets, but the accuracy obtained by the Iris datasets is 98%[83].

In a work that used the DT algorithm under the supervised learning mechanism to reveal breast cancer. Breast cancer identification is conducted here and it focused on data, which separates the data for the preparation and testing process. The result obtained is thus contrasted between the algorithms KNN and DT. The findings reveal that the accuracy obtained by KNN is 97%, while DT reaches the maximum accuracy of 99%. Therefore, a decision tree algorithm that comes under supervised learning methods predicts the type of cancer[84].

2.3.2 Classification Based on Random Forest Algorithm for Machine Learning

The Random Forest Algorithm was used by several authors in several machine learning tasks as a classifier

In a research that proposed a model proposed to apply the Random Forest algorithm With an F1 Score of 0.866, improved by the AdaBoost algorithm On the patient dataset for COVID19. Also, pointed out that the Boosted Random Forest algorithm gives detailed forecasts on imbalanced datasets too. The knowledge reviewed in this

analysis has It indicated that among the Wuhan natives, death rates were higher. Non-natives as opposed. Male patients have had a higher percentage of Compared with female patients, the mortality risk. The largest of those impacted patients are aged 20 and 70 years of age[86].

In a work that used seven variable rating methods focused on Random Forests were tested in this research. To choose the best classification form, feature exclusion techniques were implemented using both CART and CIT models. CPVIM has been proven to be more reliable in providing stable and reasonable feature rankings from correlated remotely sensed data. The optimal model was found through the NRFE process based on the CART tree using CPVIM. It achieved an overall accuracy of 89.03% with ten features only, i.e., Green, NIR, SWIR1 and SWIR2, Greenness, MSAVI, NDII, ED, SVVI, and DEM[86].

Moreover, in a work that presented a time-domain characteristics derived from the single-lead ECG was critically chosen by their data quality, and the efficiency of the heartbeat classification using RF was reasonably assessed by adopting the (AAMI) and the inter-patient paradigm principles. The most discriminative features for the classification task were considered to be normalized features relative to R-R intervals and to the width of the main wave of the QRS complex. With the top six most insightful features and a 40-tree RF classifier, the best results were produced. The MIT-BIH Arrhythmia Database measurement culminated in an average precision of 96.14 percent for the NB, SVEB, and VEB groups, with individual F1 ratings of 97.97 percent, 73.06 percent, and 90.85 percent, respectively. Results are one of the best performances recorded to date in accordance with state-of-the-art methods tested in comparable conditions. The findings not only indicate that RF is an outstanding

heartbeat classification method, but also that relatively few features are necessary to achieve state-of-the-art efficiency[87].

In a study that discussed the influence of sample size on habitats was explored in this review. Plots for suitability utilizing RF .The outcome revealed that the predictive one was the efficiency of the approximate RF models is positive, the sample sizes were associated. Next, it was determined that the Plots for habitat suitability are often impacted by the sample size and Output as prediction. In the case of minimal accessible sample evidence, to find a realistic approach for delineating habitat suitability plots, the "average plot of habitat suitability" was suggested. This demonstrates that the typical habitat suitability plot can theoretically be Improves also in a habitat suitability plot calculation in a Small quantity of samples[88].

Additionally, a work proposed a framework for soil mapping by the integration of a methodology focused on similarities and Random Forests. To check its electiveness, the approach proposed was extended to the Heshan study field. The following conclusions can be taken from this study: (1) The SB-RF system achieved better accuracy output than either the RF or SB alone, demonstrating the integrated method's e-efficacy and superiority; (2) The similarity covariates provided by the similarity-based approach embedded useful data that can eeffectively boost the accuracy of the mapping. The precision of the SB-RF system is influenced by the sampling technique[89].

Another study used computer vision methods and machine learning To explain how they can be used to detect peaks and to conduct Random Forests binary classification. Grayscale imaging, RIP reduction, hat top Filtering and thresholding is used to minimize background and threshold noise. The picture is transformed to a binary

image. Segmentation in the Watershed helped diagnose each peak compound ion is labeled and classified as a separate compound. A Full Alignment a peak table summarizing the compounds identified was created by an algorithm using compensation voltage and retention time limits. Random Forests, a model of machine learning, demonstrated strong precision, suggesting that Random Forests are a stable model for predicting binary classification on GC/DMS samples[90].

2.3.3 Classification of Vehicle Accidents Using Machine Learning

In a study carried out based on the heavy vehicle crash data of 2014, extracted from the MIROS Road Accident and Analysis and Database System (M-ROADS). The main objective of this study is to identify significant variables associated with categories of injury severity as well as classify and predict heavy vehicle drivers' injury severity in Malaysia using the classification and regression tree (CART) and random forest (RF) methods. Both CART and RF found that types of collision, driver errors, number of vehicles involved, driver's age, lighting condition and types of heavy vehicle are significant factors in predicting the severity of heavy vehicle drivers' injuries. Both models are comparable, but the RF classifier achieved slightly better accuracy[91].

In another study that aims to identify rules induced by Decision Tree algorithms (DT) for detecting traffic accidents with victims in a road stretch from accidents records, as well as probable causes of the occurrence and type of accident. Data are from a road stretch of the *Régis Bittencourt* (highway BR-116) between km 509 to km 519 in the period 2012–2014, located in São Paulo, Brazil.

Through the main results obtained, it can be concluded that the CART algorithm of the Decision Tree is a useful tool in identifying potential sites of accidents with

victims. In this case, the two most important variables to identify the severity of accidents were the accident type and the accident cause[92].

In a similar paper that presents a methodology to establish incident duration estimation models by utilizing decision tree models of CHAID, CART, C4.5 and LMT. For this study, the data contained traffic incidents that occurred on the Istanbul Trans European Motorway were obtained and separated into three groups according to duration by utilizing some studies about classification of traffic incidents. By using classified data, decision tree models of CHAID, CART, C4.5 and LMT were established and validated to estimate the incident duration. According to the results, although the models used different variables, the decision tree models of CHAID, CART and C4.5 have nearly the same prediction accuracy which is approximately 74%. On the other hand, the prediction accuracy of decision tree model of LMT is 75.4% which is somewhat better than the others. However, C4.5 model required less number of parameters than the others, while its accuracy is the same with others[93].

Another study sought to predict crash frequency according to five severity levels of PDO, fatality, severe injury, other visible injuries, and complaint of pain. The multinomial logistic regression (MLR) model and data mining approaches, including artificial neural network-multilayer perceptron (ANN-MLP) and two decision tree techniques, (i.e., Chi-square automatic interaction detector (CHAID) and C5.0) are utilized based on traffic crash records for State Highways in California, USA. The comparison of the findings of the relative importance of ten qualitative and ten quantitative independent variables incorporated in CHAID and C5.0 indicated that the cause of the crash (X1) and the number of vehicles (X5) were known as the most influential variables involved in the crash. However, the cause of the crash (X1) and

weather (X2) were identified as the most contributing variables by the ANN-MLP model. In addition, the MLR model showed that the driver's age (X11) accounts for a larger proportion of traffic crash severity. Therefore, the sensitivity analysis demonstrated that C5.0 had the best performance for predicting road crash severity. Not only did C5.0 take a shorter time (0.05 s) compared to CHAID, MLP, and MLR, it also represented the highest accuracy rate for the training set. The overall prediction accuracy based on the training data was approximately 88.09% compared to 77.21% and 70.21% for CHAID and MLP models. In general, the findings of this study revealed that C5.0 can be a promising tool for predicting road crash severity[94].

In a study that proposed a hybrid model that integrates random forest (RF) and Bayesian optimization (BO). In the proposed model, BO-RF, RF is adopted as a basic predictive model and BO is used to tune the parameters of RF. Experimental results show that BO-RF achieves higher accuracy than conventional algorithms. Moreover, BO-RF provides interpretable results by relative importance and a partial dependence plot[95].

In a study on predicting crash injury severity with machine learning algorithm synergized with clustering technique. The authors developed machine learning (ML) models to predict crash injury severity using 15 crash-related parameters. Separate ML models for each cluster were obtained using fuzzy c-means, which enhanced the predicting capability. Finally, four ML models were developed: feed-forward neural networks (FNN), support vector machine (SVM), fuzzy C-means clustering based feed-forward neural network (FNN-FCM), and fuzzy c-means based support vector machine (SVM-FCM). Features that were easily identified with little investigation on crash sites were used as an input so that the trauma center can predict the crash

severity level based on the initial information provided from the crash site and prepare accordingly for the treatment of the victims.

The input parameters mainly include vehicle attributes and road condition attributes. This study used the crash database of Great Britain for the years 2011–2016. A random sample of crashes representing each year was used considering the same share of severe and non-severe crashes. The models were compared based on injury severity prediction accuracy, sensitivity, precision, and harmonic mean of sensitivity and precision (i.e., F1 score). The SVM-FCM model outperformed the other developed models in terms of accuracy and F1 score in predicting the injury severity level of severe and non-severe crashes. This study concluded that the FCM clustering algorithm enhanced the prediction power of FNN and SVM models[96].

In a related research to assess drivers' behavior and traffic accident analysis using Decision Tree method. The study's goal is to identify the major contributing factors to traffic accidents in connection to driver behavior and socioeconomic characteristics. In order to find the most probable causes in accordance with the major target variable, which is the level of severity of the crash, the study set out to identify the main attributes induced by the decision tree method (DT). The local people received a semi-structured questionnaire interview with closed-ended questions. The survey asked questions about drivers' attitude and behavior, as well as other contributing factors such as time of accidents and road type. The attributes were analyzed using the machine-learning method using DT with Python programming language. This method was able to determine the relationship between severe and non-severe crashes and other significant influencing elements. The Duhok city people participated in the survey, which was conducted in the Kurdistan area of northern Iraq. The results of the

study demonstrate that the number of lanes, time of the accident, and human attitudes, represented by their adherence to the speed limit, are the primary causes of accidents with victims[97].

In a study on Machine Learning Technique for Predicting Road Accidents, this study suggests a system for predicting accidents that can assist in analyzing potential safety concerns and foretelling whether an accident will occur or not. The most accurate model for predicting accidents was determined by contrasting several machine learning algorithms. The government records of accidents that took place in the UK served as the dataset for this study. A combination of the Xboost method with regression is used for prediction the of accidents in the earlier stage[98].

In a study that proposes a data-driven machine learning solution for black spot screening using features of road network and facilities. The accident neighborhood is a concept introduced in the paper that represents the nearby locations associated with the happening of accidents. The concept has been realized as graph embeddings of road network, which, together with a deep neural network classifier, are the two major components of the solution. An evaluation of the solution using data from a Hong Kong district indicates that recognition of both the surrounding road network structure and the local features near accident sites can yield accurate models for black spot prediction[99].

In another study that employed deep learning for prediction of traffic accident injury severity based on contributing factors. This paper proposes a comprehensive analytic framework that employs a deep learning model referred to as the stacked sparse autoencoder (SSAE) to predict the injury severity of traffic accidents based on contributing factors. The essential idea of the method is to integrate various analyses

into an analytical framework that performs corresponding data processing and analysis by different machine learning approaches. In the proposed method, the authors utilize a machine learning approach (i.e., Catboost) to analyze the importance and dependence of the contributing factors to injury severity and remove low correlation factors; second, according to the geographical information, we classify the data into different classes by utilizing a machine learning approach (i.e., *k*-means clustering); third, by employing high correlation factors, we employ an SSAE-based deep learning model to perform injury severity prediction in each data class. By experiments with a real-world traffic accident dataset, they demonstrated the effectiveness and applicability of the framework. Specifically, (1) the importance and dependence of contributing factors were obtained by CatBoost and the Shapley value, and (2) the SSAE-based deep learning model achieved the best performance compared to other baseline models[100].

In another study that aims to investigate the injury severity prediction (ISP) capability in machine-learning analytics based on five-different regional Level 1 trauma center enrolled patients in Korea. We study car crash-related injury data of 1417 patients enrolled in the Korea In-Depth Accident Study database from January 2011 to April 2021. Severe injury classification was defined using an Injury Severity Score of 15 or greater. A planar crash was considered by excluding rollovers to compromise an accurate prediction. Furthermore, dissimilarities of the collision partner component based on vehicle segmentation were assumed for crash incompatibility. To handle class-imbalanced clinical datasets, we used four data-sampling techniques (i.e., class-weighting, resampling, synthetic minority oversampling, and adaptive synthetic sampling). Machine-learning analytics based on logistic regression, extreme gradient

boosting (XGBoost), and a multilayer perceptron model were used for the evaluations. Each model was executed using five-fold cross-validation to solve overfitting consistent with the hyperparameters tuned to improve model performance. The area under the receiver operating characteristic curve of 0.896. Additionally, the present ISP model showed an under-triage rate of 6.1%. The Delta-V, age, and Principal ~ were significant predictors. The results demonstrated that the data-balanced XGBoost model achieved a reliable performance on injury severity classification of emergency department patients[101].

Using comparative analysis of machine learning algorithms for road accident severity prediction. The authors develop a prediction framework and implemented six different machine learning algorithms, namely: Naïve Bayes, Logistic Regression, Decision Tree, Random Forest, Bagging, and AdaBoost to predict the severity of the crash. Experimental results procured for the crash dataset published by the UK shows that Random Forest, Decision Tree, and Bagging significantly outperformed other algorithms in terms of all performance metrics. Furthermore, we analyze the huge; traffic data and extract insightful crash patterns to figure out the significant factors that have a clear effect on road accidents and provide beneficial suggestions regarding this issue[102].

In another research that used gradient boosted and random forest trees for road crashes analysis and prediction. The proposed work studied and analyzed several factors of road accidents to create an accurate and interpretable model that predicts the occurrence and severity of car accidents by investigating crash causal factors and crash severity factors. In the proposed work, we employed three machine learning

algorithms to vis-à-vis Decision Tree, Random Forest, and Gradient Boosted tree on Statewide Vehicle Crashes Dataset provided by Maryland State Police. The gradient boosted-based model reported the highest prediction accuracy and provided the most influencing factors in the predictive model. The findings showed that disregarding traffic signals and stop signs, road design problems, poor visibility, and bad weather conditions are the most important variables in the predictive road traffic crash model. Using the identified risk factors is crucial in establishing actions that may reduce the risks related to those factors[103].

In another similar research that used data mining techniques for road accidents investigation and forecasting. In this paper, the authors used used data mining techniques and geometric analysis on a dataset of road accidents to find the impact of attributes like road surface, weather conditions, lighting conditions, and casualty severity on a road accident. The Frequent Pattern (FP) Growth technique was used to discover the association rules. Classification models were made by some decision trees like J48 and Decision Tree (DT), Random Tree, and Hoeffding tree. The results showed that Random Tree Classifier performed well with 90.6% accuracy, followed by Hoeffding Tree with 85.58% accuracy and J48 with 84% accuracy[104].

In a work on classification of road traffic accident data using machine learning algorithms, the authors applied different machine learning classification algorithms and discussed here the six algorithms with high accuracy and best classification performances such as Fuzzy-FARCHD, Random Forest, Hierarchal LVQ, RBF Network (Radial Basis Function Network), Multilayer Perceptron, and Naïve Bayes on road traffic accident data set obtained from UK road traffic accident of the year 2016. The data set contains information on all road accident casualties across

Calderdale. The results from our analysis show that Fuzzy-FARCHD algorithm is effective to classify the dataset and achieves an accuracy of 85.94%. In this work, we have revealed that Lighting Conditions, 1st Road Class & No., Number of vehicles are the key features in selecting the attributes[105].

In a research that aimed to assess prediction model designs for RTAs to assist transport authorities and policymakers. It considered classifiers such as naïve Bayes, logistic regression, k-nearest neighbour, AdaBoost, support vector machine, random forest, and five missing data methods. These classifiers were evaluated using five evaluation metrics: accuracy, rootmean-square error, precision, recall, and receiver operating characteristic curves. Furthermore, the assessment involved parameter adjustment and incorporated dimensionality reduction techniques. The empirical results and analyses show that the RF classifier, combined with multiple imputations by chained equations, yielded the best performance when compared with the other combinations[106].

In another similar work titled "The prediction of road-accident risk through data mining". This work proposes a tool to predict the risk of road accidents. The developed system consists of three steps: data selection and collection, preprocessing, and the use of mining algorithms. The data were imported from the Portuguese National Guard database, and they related to accidents that occurred from 2019 to 2021. The results allowed us to conclude that the highest concentration of accidents occurs during the time interval from 17:00 to 20:00, and that rain is the meteorological factor with the greatest effect on the probability of an accident occurring[107].

In a study on clustering of road traffic accidents as a gestalt problem, this paper introduces and illustrates an approach to automatically detecting and selecting "critical" road segments, intended for application in circumstances of limited human or technical resources for traffic monitoring and management. The reported study makes novel contributions at three levels. At the specification level, it conceptualizes "critical segments" as road segments of spatially prolonged and high traffic accident risk. At the methodological level, it proposes a two-stage approach to traffic accident clustering and selection. The first stage is devoted to spatial clustering of traffic accidents. The second stage is devoted to selection of clusters that are dominant in terms of number of accidents. At the implementation level, the paper reports on a prototype system and illustrates its functionality using publicly available real-life data. The presented approach is psychologically inspired to the extent that it introduces a clustering criterion based on the Gestalt principle of proximity. Thus, the proposed algorithm is not density-based, as are most other state-of-the-art clustering algorithms applied in the context of traffic accident analysis, but still keeps their main advantages: it allows for clusters of arbitrary shapes, does not require an a priori given number of clusters, and excludes "noisy" observations[108].

Using data-driven analysis for fatal urban traffic accident characteristics. This work uses the statistical data of fatal road traffic accidents in Shenzhen from 2018 to 2022 as the basis to determine the characteristic patterns and the main influencing factors of the occurrence of fatal road traffic accidents. The accident description data are also analyzed using the analysis method based on Term Frequency-Inverse Document Frequency (TF-IDF) data mining to obtain the characteristics of accident fields, objects, and types. Furthermore, this work conducts a kernel density analysis

combined with spatial autocorrelation to determine the hotspot areas of accident occurrence and analyze their spatial aggregation effects. A principal component analysis is performed to calculate the factors related to the accident subjects. Results showed that weak safety awareness of motorists and irregular driving operations are the main factors for the occurrence of accidents. Finally, targeted safety management strategies are proposed based on the analysis results[109].

In a work on hybrid traffic accident classification models. This paper proposes a CCTV frame-based hybrid traffic accident classification model that enables the identification of whether a frame includes accidents by generating object trajectories. The proposed model utilizes a Vision Transformer (ViT) and a Convolutional Neural Network (CNN) to extract latent representations from each frame and corresponding trajectories. The fusion of frame and trajectory features was performed to improve the traffic accident classification ability of the proposed hybrid method. In the experiments, the Car Accident Detection and Prediction (CADP) dataset was used to train the hybrid model, and the accuracy of the model was approximately 97%. The experimental results indicate that the proposed hybrid method demonstrates an improved classification performance compared to traditional models[110].

In a similar study to assess highway accident severity prediction for optimal resource allocation of emergency vehicles and personnel. This paper uses real-life traffic and accident data for a Florida highway to build prediction models to predict traffic accident severity. Accurate severity prediction is beneficial for both the responders and the drivers. First responders are in high demand, and the pandemic has made the situation worse. When an accident occurs, the emergency center dispatches a random number of emergency vehicles. Unfortunately, this number exceeds the number of

vehicles needed most of the time, leaving fewer resources to respond to simultaneous accidents in different locations. Also, an increased number of emergency vehicles could introduce secondary accidents[111].

In a study on traffic accident severity prediction based on decision level fusion of machine and deep learning model. In this study, significant factors that are strongly correlated with the accident severity on highways are identified by Random Forest. Top features affecting accidental severity include distance, temperature, wind_Chill, humidity, visibility, and wind direction. This study presents an ensemble of machine learning and deep learning models by combining Random Forest and Convolutional Neural Network called RFCNN for the prediction of road accident severity. The performance of the proposed approach is compared with several base learner classifiers. The data used in the analysis include accident records of the USA from February 2016 to June 2020. Obtained results demonstrate that the RFCNN enhanced the decision-making process and outperformed other models with 0.991 accuracy, 0.974 precision, 0.986 recall, and 0.980 F-score using the 20 most significant features in predicting the severity of accidents[112].

Using Artificial Neural Network for road accident severity prediction, this research work focuses on the prediction of accident severity in the island of Mauritius. The authors experimented with different configurations of the multi-layer perceptron (MLP). The optimum values for hyperparameters were determined through systematic manual tuning over several experiments. They are as follows: 50 and 25 neurons in the hidden layers respectively, with a learning rate of 0.1, the Rectified Linear Unit (RELU) as an activation function and a maximum of 10,000 iterations. To test and avoid overfitting/underfitting, stratified 10-fold cross-validation was used. The

comparative analysis shows that the MLP outperformed all the other models previously implemented using the same dataset, with an accuracy of around 84.1%[113].

In a study on road accident prediction and classification using Machine Learning, the authors discussed about one such model where they make use of machine learning and data mining concepts to predict and analyze such accidents all over the country. They used regression and clustering types of Machine Learning algorithms to predict and analyze the accident rate for the year 2022 for all the States and UTs. They also used Linear Regression algorithm for the prediction of the accident rate for the year 2022. Furthermore, they classified all the states and UTs into two clusters where one cluster consists of states with High accident rate and another cluster consists of states with Low accident rate. The finally used K-Means clustering algorithm for the classification of the states and UTs into clusters[114].

2.4 Chapter Summary and Gap in Literature Reviewed

This chapter was organised into four sub-headings - conceptual review, theoretical review/framework, review of empirical studies related to the research topic and conceptual model. The conceptual review explained in depth the concepts of the study. These concepts are - road traffic accidents and machine learning. It also richly gave insights into sub-concepts such as classification of road traffic accident using machine learning which includes various classification algorithms like GBoost, Support Vector Regression, Extreme Learning Machines (SVM), Gaussian Naive Bayes (GNB), Multilayer Perceptron, Esemble Technique and K-Nearest Neighbour (KNN). Other sub-concept which includes analysis and interpretation of ML model, model optimization and data balancing were also explained.

In the methodological review, the main classification algorithms used in this study were fully explained. They are Random Forest (RF) Algorithms which involves two steps, one is Random Forest formation, and the other is to make a guess from the first step of the Random Forest classifier. The second is Decision Tree, where each node represents features in a category to be classified and each subset defines a value that can be taken by the node

In the review of empirical studies, several studies on classifications using Decision Tree and Random Forest algorithm were presented. Also, closely related studies on accident predictions using various algorithms were also presented.

The studies show that many empirical research works similar to the topic under study have been carried out. However, past empirical studies using other algorithms showed lower accuracy, lower f1 scores and precision. Also, these solutions did not address the severity of vehicle accidents based on factors such as vehicle speed, impact angle, and vehicle type. Previous studies are also scarce on using both Decision Tree and Random Forest in predicting road traffic accident. Therefore, this work tends to predict the severity of any road traffic accident and evaluation of multiple traffic accident attributes for the purpose of achieving a better prediction performance using two algorithms (the Random Forest model and the Decision Tree Classifier model). The empirical studies reviewed therefore shows that studies are lacking in the subject matter which identifies a gap in literature that needs to be filled.

Endnotes

[1]N.F Yaacob, N Rusli & S.N Bohari. *A Review Analysis of Accident Factor on Road Accident Cases Using Haddon Matrix Approach*. **In Proceedings of the Second International Conference on the Future of ASEAN (ICoFA) Springer Singapore 2017–Volume 2: Science and Technology** 2018, pp. 55-65.

[2]E.A Yihun. *The Causal Attribution of Road Traffic Accident among Victims and Drivers in Case of Gondar Town, North West Ethiopia*. **IJASSH**. Dec 12 2019.

[3]J.A Andeta. *Road-Traffic Accident Prediction Model : Predicting the Number of Casualties 2021*. Available from: http://urn.kb.se/resolve?urn=urn:nbn:se:his:diva-20146

[4]B. Kumeda, F Zhang, F Zhou, S Hussain, A Almasri & M Assef. *Classification of Road Traffic Accident Data Using Machine Learning Algorithms*. **In 2019 IEEE 11th international conference on communication software and networks (ICCSN) IEEE,** Jun 12 2019, pp. 682-687.

[5]P.A Nandurge & N.V Dharwadkar. *Analyzing Road Accident Data Using Machine Learning Paradigms*. **In 2017 International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud)(I-SMAC) IEEE** Feb 10 2017 pp. 604-610.

[6]Z.E Abou Elassad, H Mousannif & H Al Moatassime. *A Real-Time Crash Prediction Fusion Framework: An Imbalance-Aware Strategy for Collision Avoidance Systems*. **Transportation Research Part C: Emerging Technologies.** Sep 1 2020; 118:102708.

[7]H Park & A Haghani. *Real-Time Prediction of Secondary Incident Occurrences Using Vehicle Probe Data.* T**ransportation Research Part C: Emerging Technologies.** Sep 1 2016; 70:69-85.

[8]M Zahid, Y Chen, S Khan, A Jamal, M Ijaz & T Ahmed. *Predicting Risky and Aggressive Driving Behavior Among Taxi Drivers: Do Spatio-Temporal Attributes Matter?* **International Journal of Environmental Research and Public Health**. (11):3937, Jun 2020.

[9]D.H Kim DH, Ramjan LM & Mak KK. *Prediction of Vehicle Crashes by Drivers' Characteristics and Past Traffic Violations In Korea Using A Zero-Inflated Negative Binomial Model.* **Traffic Injury Prevention** 17(1). Jan 2 2016 86-90.

[10]C Wang, L Liu, C Xu & W Lv. *Predicting Future Driving Risk of Crash-Involved Drivers Based on a Systematic Machine Learning Framework*. **International Journal of Environmental Research And Public Health** 16(3): Feb 2019; 334.

[11]N O Onyemaechi & U R Ofoma. *The Public Health Threat of Road Traffic Accidents in Nigeria: A Call to Action.* **Annals of Medical and Health Sciences Research.** 6(4) 2016; 199-204.

[12]B. B Apeagee & S A Haaor. *A Logistic Regression Model of Road Traffic Fatalities in Benue State: Imp*lication *to Public Health***. Nigerian Annals of Pure and Applied Sciences.** 3(3a), Nov 15, 2020: 46-52.

[13]https://www.statista.com/statistics/1296025/total-number-of-road-casualties-in-nigeria/

[14]S. K Chinnamgari. *R Machine Learning Projects: Implement Supervised, Unsupervised, and Reinforcement Learning Techniques using R 3.5*. **Packt Publishing Ltd;** Jan 14, 2019.

[15]H. Hihn & D.A Braun. *Specialization in Hierarchical Learning Systems: A Unified Information-theoretic Approach for Supervised, Unsupervised and Reinforcement Learning. Neural Processing Letters*. 2020 Dec;52(3):2319-52.

[16]J. A Sidey-Gibbons & C.J Sidey-Gibbons. *Machine Learning in Medicine: A Practical Introduction*. **BMC MedicalR Methodology**. 19 Dec 2019: 1-8.

[17]I. H Sarker. *Machine Learning: Algorithms, Real-World Applications and Research Direction*s. **SN computer science**. 2(3), May 2021; 160.

[18]Y. Bengio, A Lodi & A Prouvost. *Machine Learning for Combinatorial Optimization: A Methodological Tour d'horizon*. **European Journal of Operational Research**. 290(2) Apr 16, 2021; 405-21.

[19]T Bokaba, W Doorsamy & B.S Paul. *Comparative Study of Machine Learning Classifiers for Modelling Road Traffic Accidents*. **Applied Sciences.** 12(2): Jan 2022; 828.

[20]H Saigo, S Nowozin, T Kadowaki, T Kudo & K Tsuda. *gBoost: A Mathematical Programming Approach to Graph Classification and Regression*. Machine Learning. Apr 2009; 75:69-89.

[21]A.V Konstantinov & L.V Utkin. *Interpretable Machine Learning with an Ensemble of Gradient Boosting Machine*s. **Knowledge-Based Systems**. Jun 21 2021; 222:106993.

[22]M Awad, & R Khanna. *Efficient Learning Machines: Theories, Concepts, and Applications for Engineers and System Designers* **Springer Nature**, 2015 (p. 268).

[23]Y Guo, TvHastie & R Tibshirani. *Regularized Linear Discriminant Analysis and its Application in Microarrays*. **Biostatistics** 8(1): Jan 1 2007; 86-100.

[24]S Ghosh, A Dasgupta & A Swetapadma. *A Study on Support Vector Machine Based Linear and Non-Linear Pattern Classification*. **In 2019 International Conference on Intelligent Sustainable Systems (ICISS) IEEE** Feb 21 2019, pp. 24-28.

[25]I Ahmad, M Basheri, M.J Iqbal & A Rahim. *Performance Comparison of Support Vector Machine, Random Forest, and Extreme Learning Machine for Intrusion Detection*. **IEEE** access. May 30 2018; 6:33789-95.

[26]X Zhai, M Chen & W Lu. *Fuel Ratio Optimization of Blast Furnace Based on Data Mining.* **ISIJ International**. 60(11) Nov 15 2020: 2471-6.

[27]M Sabzekar & S.M Hasheminejad. *Robust Regression Using Support Vector Regressions. Chaos, Solitons & Fractals*. Mar 1 2021; 144:110738.

[28]P. Tsirikoglou, S Abraham, F Contino, C Lacor & G Ghorbaniasl. *A Hyperparameters Selection Technique for Support Vector Regression Models.* **Applied Soft Computing**. 1 Dec 2017; 61:139-48.

[29]D.A Pisner & D.M *Schnyer. Support Vector Machine. In Machine learning* **Academic Press.** Jan 1 2020 (pp. 101-121).

[30]M.S Adnan, S Zaidi & P Bhargava. *A Novel Support Vector Regression (SVR) Model for the Prediction of Splice Strength of the Unconfined Beam Specimens. Construction and Building Materials.* Jul 10 2020; 248:118475.

[31]M Wang, S Jia, E Chen, S Yang, P Liu & Z Qi. *A Derived Least Square Fast Learning Network Model.* **Applied Intelligence**. Dec 2020; 50:4176-94.

[32]S Mangalathu, S.H Hwang, E Choi & J.S Jeon. *Rapid Seismic Damage Evaluation of Bridge Portfolios Using Machine Learning Techniques.* **Engineering Structures.** Dec 15 2019; 201:109785.

[33]S Afraei, K Shahriar & S.H Madani. *Developing Intelligent Classification Models for Rock Burst Prediction after Recognizing Significant Predictor Variables, Section 2: Designing classifiers. Tunnelling and Underground Space Technology*. Feb 1 2019; 84:522-37.

[34]L Ali, S.U Khan, N.A Golilarz, I Yakubu, I Qasim, A Noor & R Nour. *A Feature-Driven Decision Support System for Heart Failure Prediction Based on Statistical*

Model And Gaussian Naive Bayes. **Computational and Mathematical Methods in Medicine**. Nov 2019.

[35]D Bassi & H Singh. *Optimizing Hyperparameters for Improvement in Software Vulnerability Prediction Models.* **In Advances in Distributed Computing and Machine Learning: Proceedings of ICADCML Singapore: Springer Nature Singapore.** Jul 28 2022, pp. 533-544.

[36]V Mishra, S.M Agarwal & N Puri. *Comprehensive and Comparative Analysis of Neural Network.* **International Journal of Computer Application**. 2(8), 2018;:126-37.

[37]J.M Johnson & T.M Khoshgoftaar. *Survey on Deep Learning with Class Imbalance*. **Journal of Big Data**; 6(1). Dec 2019: 1-54.

[38]A Darwish, D Ezzat & A E Hassanien. *An Optimized Model Based on Convolutional Neural Networks and Orthogonal Learning Particle Swarm Optimization Algorithm for Plant Diseases Diagnosis*. **Swarm and evolutionary computation.** Feb 1 2020; 52:100616.

[39]Y Ren, L Zhang & P.N Suganthan. *Ensemble Classification and Regression-Recent Developments, Applications and Future Directions.* **IEEE Computational Intelligence Magazine** 11(1). Jan 12 2016; 41-53.

[40]A Mosavi, F Sajedi Hosseini, B Choubin, M Goodarzi, A.A Dineva & E R Sardooi. *Ensemble Boosting and Bagging Based Machine Learning Models for Groundwater Potential Prediction*. **Water Resources Management**. Jan 2021, 35:23-37.

[41]N. A Mashudi, N Ahmad & N.M Noor. *Classification of Adult Autistic Spectrum Disorder using Machine Learning Approach*. **IAES International Journal of Artificial Intelligence** 10(3)**.** Sep 1 2021: 743.

[42]A. Taherkhani, G Cosma & T.M McGinnity. *AdaBoost-CNN: An Adaptive Boosting Algorithm for Convolutional Neural Networks to Classify Multi-Class Imbalanced Datasets Using Transfer Learning*. **Neurocomputing**. Sep 3 2020; 404:351-66.

[43]E. Suganya & CRajan. *An Adaboost-Modified Classifier using Particle Swarm Optimization and Stochastic Diffusion Search in Wireless Iot Networks*. **Wireless Networks.** May 2021; 27:2287-99.

[44]R. Sibindi, R.W Mwangi & A.G Waititu. *A Boosting Ensemble Learning Based Hybrid Light Gradient Boosting Machine and Extreme Gradient Boosting Model for Predicting House Prices*. **Engineering Reports**. 2022:e12599.

[45]X. Yang, Y. Wang, R. Byrne, G Schneider & S Yang. *Concepts of Artificial Intelligence for Computer-Assisted Drug Discovery*. **Chemical reviews**. Jul 11 2019; 119(18):10520-94.

[46]C. Zhang, X Lei & L. Liu. *Predicting Metabolite–Disease Associations Based on LightGBM Model*. **Frontiers in Genetics**. Apr 13 2021;12:660275.

[47]O.I Abiodun, A Jantan, K.V Dada, N.A Mohamed & H Arshad. *State-of-theart in Artificial Neural Network Applications: A survey*. Heliyon 2018. 4 e00938. doi: 10.1016/j.heliyon. e00938.

[48]W Qi, H Su, C Yang, G Ferrigno, E De Momi & A Aliverti. *A Fast and Robust Deep Convolutional Neural Networks for Complex Human Activity Recognition using Smartphone*. Sensors. Aug 29 2019; 19(17):3731.

[49]A Golab, E Gooya, A Falou & M Cabon. *A Multilayer Feed-Forward Neural Network (MLFNN) For the Resource-Constrained Project Scheduling Problem (RCPSP)*. **Decision Science Letters.** 11(4): 2022; 407-18..

[50]A. B Shaik & S Srinivasan. *A Brief Survey on Random Forest Ensembles in Classification Model*. **In International Conference on Innovative Computing and Communications: Proceedings of ICICC 2018, Springer Singapore** Volume 2 2019, pp. 253-260.

[51]I Reis, D Baron & S Shahaf. *Probabilistic Random Forest: A Machine Learning Algorithm for Noisy Data Sets*. **The Astronomical Journal.** 2018 Dec 20; 157(1):16.

[52]B.O Yigin, O Algin & G Saygili. *Comparison of Morphometric Parameters in Prediction of Hydrocephalus Using Random Forests*. **Computers in Biology and Medicine.** Jan 1 2020;116:103547.

[53]N.M Abdulkareem & A.M Abdulazeez. *Machine Learning Classification Based on Radom Forest Algorithm: A review*. **International Journal of Science and Business**. 2021; 5(2):128-42.

[54]D Denisko & M.M Hoffman. *Classification and Interaction in Random Forests. Proceedings of the National Academy of Sciences*. Feb 20 2018;115(8):1690-2.

[55]L.V Utkin, M.S Kovalev & F.P Coolen. *Imprecise Weighted Extensions of Random Forests for Classification and Regression*. **Applied Soft Computing.** Jul 1 2020;92:106324.

[56]L Demidova & M Ivkina. *Defining the Ranges Boundaries of the Optimal Parameters Values for the Random Forest Classifier.* **In 2019 1st International**

Conference on Control Systems, Mathematical Modelling, Automation and Energy Efficiency (SUMMA) IEEE Nov 20 2019 pp. 518-522.

[57]M. A Sulaiman. *Evaluating data mining classification methods performance in Internet of things applications*. **Journal of Soft Computing and Data Mining**. Dec 6 2020;1(2):11-25.

[58]M.L Kolhe, S Tiwari, M.C Trivedi & K.K Mishra (Eds.). *Advances in Data and Information Sciences:* **Proceedings of ICDIS Vol. 94. Springer Singapore** 2019**.** https://doi.org/10.1007/978- 981-15-0694-9

[59]K Gajowniczek, I Grzegorczyk, T Ząbkowski, C Bajaj. *Weighted Random Forests to Improve Arrhythmia Classification*. **Electronics.** Jan 3 2020;9(1):99.

[60]Y Sun, Y Li, Q Zeng & Y Bian. *Application Research of Text Classification Based on Random Forest Algorithm.* **2020 3rd International Conference on Advanced Electronic Materials, Computers and Software Engineering (AEMCSE),** 2020 370–374. https://doi.org/10.1109/AEMCSE50948.2020.00086

[61]S Koley, A.K Sadhu, P Mitra, B Chakraborty & C Chakraborty. *Delineation and Diagnosis of Brain Tumors from Post Contrast T1-weighted MR Images Using Rough Granular Computing and Random Forest*. **Applied Soft Computing.** Apr 1 2016;41:453-65.

[62]B Charbuty & A Abdulazeez. *Classification Based on Decision Tree Algorithm for Machine Learning*. **Journal of Applied Science and Technology Trends**. 2021 Mar 24;2(01):20-8.

[63]Q. Zou, K. Qu, Y. Luo, D. Yin, Y. Ju, & H. Tang, "*Predicting Diabetes Mellitus with Machine Learning Techniques*," **Frontiers in Genetics,** vol. 9, 2018 p. 515

[64]T. A. Assegie & P. S. Nair, "*Handwritten Digits Recognition with Decision Tree Classification: A Machine Learning Approach*," **International Journal of Electrical and Computer Engineering**, vol. 9, no. 5, 2019, p. 4446,

[65]F. De Felice, D Crocetti, M Parisi, V Maiuri, E Moscarelli, R Caiazzo, N Bulzonetti, D Musio & V Tombolini."*Decision Tree Algorithm in Locally Advanced Rectal Cancer: An Example of Over-Interpretation and Misuse of a Machine Learning Approach,*" **Journal of Cancer Research and Clinical Oncology**, vol. 146, no. 3, 2020 pp. 761–765

[66]I. H. Sarker, A. Colman, J. Han, A. I. Khan, Y. B. Abushark, & K. Salah, "*Behavdt: A Behavioral Decision Tree Learning to Build User-centric Context-Aware Predictive Model*," **Mobile Networks and Applications**, vol. 25, no. 3, 2020, pp. 1151–1161

[67]X. Hu, C. Rudin, & M. Seltzer, "*Optimal Sparse Decision Trees,*" **In Advances in Neural Information Processing Systems**, 2019, pp. 7267– 7275

[68]S. Patil & U. Kulkarni, "*Accuracy Prediction for Distributed Decision Tree using Machine Learning approach*," I**n 2019 3rd International Conference on Trends in Electronics and Informatics (ICOEI)**, Apr. 2019, pp. 1365–1371, doi: 10.1109/ICOEI.2019.8862580

[69]D. Hussain, M. A. Al-Antari, M. A. Al-Masni, S.-M. Han & T.-S. Kim, "*Femur Segmentation in DXA Imaging using a Machine Learning Decision Tree,*" **Journal of X-ray Science and Technology**, vol. 26, no. 5, 2018, pp. 727–746.

[70]N. Linty, A. Farasin, A. Favenza, & F. Dovis, "*Detection of GNSS Ionospheric Scintillations Based on Machine Learning Decision Tree*," **IEEE Transactions on Aerospace and Electronic Systems,** vol. 55, no. 1, Feb. 2019 pp. 303–317, doi: 10.1109/TAES.2018.2850385.

[71]W. Kuang, Y. Chan, S. Tsang, & W. Siu, "*Machine Learning-Based Fast Intra Mode Decision for HEVC Screen Content Coding via Decision Trees,*" **IEEE Transactions on Circuits and Systems for Video Technology**, vol. 30, no. 5, May 2020, pp. 1481–1496, doi: 10.1109/TCSVT.2019.2903547.

[72]I. Ramadhan, P. Sukarno, & M. A. Nugroho, "*Comparative Analysis of K-Nearest Neighbor and Decision Tree in Detecting Distributed Denial of Service,*" I**n 2020 8th International Conference on Information and Communication Technology (ICoICT)**, Yogyakarta, Indonesia, Jun. 2020, pp. 1–4, doi: 10.1109/ICoICT49345.2020.9166380.

[73]V. M. E. Batitis, M. J. G. Caballes, A. A. Ciudad, M. D. Diaz, R. D. Flores, & E. R. E. Tolentin, "*Image Classification of Abnormal Red Blood Cells Using Decision Tree Algorithm,*" I**n 2020 Fourth International Conference on Computing Methodologies and Communication (ICCMC),** Mar. 2020, pp. 498–504, doi: 10.1109/ICCMC48092.2020.ICCMC-00093.

[74]Y. Zhang, J. Liu, Z. Zhang, & J. Huang, "*Prediction of Daily Smoking Behavior Based on Decision Tree Machine Learning Algorithm*," **in 2019 IEEE 9th International Conference on Electronics Information and Emergency Communication (ICEIEC),** Jul. 2019, pp. 330–333, doi: 10.1109/ICEIEC.2019.8784698

[75]S. Nandhini & J. M. K.S, "*Performance Evaluation of Machine Learning Algorithms for Email Spam Detection*," I**n 2020 International Conference on Emerging Trends in Information Technology and Engineering (ic-ETITE)**, Feb. 2020, pp. 1–4, doi: 10.1109/ic-ETITE47903.2020.312.

[76]A. I. Taloba & S. I. Ismail, "*An Intelligent Hybrid Technique of Decision Tree and Genetic Algorithm for E-Mail Spam Detection*," I**n 2019 Ninth International Conference on Intelligent Computing and Information Systems (ICICIS),** Dec. 2019, pp. 99–104, doi: 10.1109/ICICIS46948.2019.9014756

[77]M. O. Arowolo, M. Adebiyi, A. Adebiyi, & O. Okesola, "*PCA Model for RNA-Seq Malaria Vector Data Classification using KNN and Decision Tree Algorithm,*" I**n 2020 International Conference in Mathematics, Computer Engineering and Computer Science (ICMCECS)**, Mar. 2020, pp. 1–8, doi: 10.1109/ICMCECS47690.2020.240881

[78]S. Pathan, P. Kumar, R. Pai, & S. V. Bhandary, "*Automated Detection of Optic Disc Contours in Fundus Images using Decision Tree Classifier*," **Biocybernetics and Biomedical Engineering**, vol. 40, no. 1, 2020, pp. 52–64

[79]A. A. Nagra, F Han, Q H Ling, M Abubaker, F Ahmad, S Mehta, A T Apasiba. "*Hybrid Self-Inertia Weight Adaptive Particle Swarm Optimisation with Local Search Using C4. 5 Decision Tree Classifier for Feature Selection Problems,*" **Connection Science,** vol. 32, no. 1, 2020, pp. 16–36.

[80]A. Ahmim, L. Maglaras, M. A. Ferrag, M. Derdour, & H. Janicke, *"A Novel Hierarchical Intrusion Detection System Based On Decision Tree And Rules-Based Models*," **in 2019 15th International Conference on Distributed Computing in Sensor Systems (DCOSS)**, 2019, pp. 228–233.

[81]M. Li, H. Xu, & Y. Deng, "*Evidential Decision Tree Based on Belief Entropy," Entropy*, vol. 21, no. 9, 2019, p. 897.

[82]P. Sathiyanarayanan, S. Pavithra, M. S. SARANYA, & M. Makeswari, "*Identification of Breast Cancer using the Decision Tree Algorithm*," I**n 2019 IEEE International Conference on System, Computation, Automation and Networking (ICSCAN),** 2019, pp. 1–6.

[83]C Iwendi, A.K Bashir, A Peshkar, R Sujatha, J.M Chatterjee, S Pasupuleti, R Mishra, S Pillai & O Jo. *COVID-19 Patient Health Prediction using Boosted Random Forest Algorithm*. **Frontiers in Public Health**. Jul 3 2020; 8:357.

[84]F Zhang, X Yang. *Improving Land Cover Classification in an Urbanized Coastal Area by Random Forests: The Role of Variable Selection*. **Remote Sensing of Environment.** Dec 15 2020;251:112105.

[85]J.F Saenz-Cogollo & M Agelli. *Investigating Feature Selection and Random Forests for Inter-Patient Heartbeat Classification*. **Algorithms.** Mar 25 2020; 13(4):75.

[86]R Shiroyama, M Wang & C Yoshimura. *Effect of Sample Size on Habitat Suitability Estimation using Random Forests: A Case Of Bluegill, Lepomis Macrochirus*. **In Annales de Limnologie-International Journal of Limnology Vol. 56, p. 13. EDP Sciences**, 2020.

[87]D Yeap, M.M McCartney, M.Y Rajapakse, A.G Fung, N.J Kenyon & C.E Davis. *Peak Detection and Random Forests Classification Software for Gas Chromatography/Differential Mobility Spectrometry (GC/DMS) Data.* **Chemometrics and Intelligent Laboratory Systems**. Aug 15 2020;203:104085.

[88]A Azhar, N.M Ariff, M.A Bakar & A Roslan. *Classification of Driver Injury Severity for Accidents Involving Heavy Vehicles with Decision Tree and Random Forest*. **Sustainability**. Mar 30 2022;14(7):4101.

[89]A da Cruz Figueira, C.S Pitombo & A.P Larocca. *Identification of Rules Induced Through Decision Tree Algorithm for Detection of Traffic Accidents with Victims: A Study Case from Brazi*l. **Case Studies on Transport Policy**. Jun 1 2017;5(2):200-7.

[90]A Saracoglu & H Ozen. *Estimation of Traffic Incident Duration: A Comparative Study of Decision Tree Models*. **Arabian Journal for Science and Engineering**. Oct 2020; 45(10):8099-110.

[91]G Shiran, R Imaninasab & R Khayamim. *Crash Severity Analysis of Highways Based on Multinomial Logistic Regression Model, Decision Tree Techniques, and Artificial Neural Network: A Modeling Comparison*. **Sustainability**. May 18 2021;13(10):5670.

[92]M Yan & Y Shen. *Traffic Accident Severity Prediction Based on Random Forest.* **Sustainability**. Feb 2 2022; 14(3):1729.

[93]K Assi, S.M Rahman, U Mansoor & N Ratrout. *Predicting Crash Injury Severity with Machine Learning Algorithm Synergized with Clustering Technique: A Promising Protocol.* **International journal of environmental research and public healt**h. Aug 17 2020 (15):5497.

[94]P Abdullah & T Sipos. *Drivers' Behavior and Traffic Accident Analysis Using Decision Tree Method*. **Sustainability**. Sep 9 2022; 14(18):11339.

[95]S Alagarsamy, P Nagaraj, B Srikanth, C.V Krishna, G Bharath & S.S Kalyan. *A Novel Machine Learning Technique for Predicting Road Accidents*. **In 2023 Third International Conference on Artificial Intelligence and Smart Energy (ICAIS)** Feb 2 2023, pp. 1547-1551.

[96]A Kwok-Fai Lui, Y.H Chan, K.H Lo, W.T Cheng & H.T Cheung. *Predictive Screening of Accident Black Spots based on Deep Neural Models of Road Networks*

and Facilities: A Case Study based on a District in Hong Kong. **In 2021 5th International Conference on Computer Science and Artificial Intelligence,** Dec 4 2021, pp. 422-428.

[97]Z Ma, G Mei & S Cuomo. *An Analytic Framework using Deep Learning for Prediction of Traffic Accident Injury Severity Based on Contributing Factors*. **Accident Analysis & Prevention**. Sep 1 2021;160:106322.

[98]J.S Kong, K.H Lee, O.H Kim, H.Y Lee, C.Y Kang, D Choi, S.C Kim, H Jeong, D.R Kang & T.E Sung. *Machine Learning-Based Injury Severity Prediction of Level 1 Trauma Center Enrolled Patients Associated with Car-To-Car Crashes in Korea.* **Computers in Biology and Medicine**. Feb 1 2023;153:106393.

[99]S Malik, H El Sayed, M.A Khan & M.J Khan. *Road Accident Severity Prediction—A Comparative Analysis of Machine Learning Algorithms.* **In 2021 IEEE Global Conference on Artificial Intelligence and Internet of Things IEEE. (GCAIoT)** Dec 12 2021, pp. 69-74.

[100]S Elyassami, Y Hamid & T Habuza. *Road Crashes Analysis and Prediction using Gradient Boosted and Random Forest Trees*. **In 2020 6th IEEE Congress on Information Science and Technology (CiSt) IEEE** Jun 5 2021, pp. 520-525.

[101]U Fareed, U Khadam, M.M Iqbal & M.J Iqbal. *Road Accidents Investigation and Forecasting using Data Mining Techniques.* **KIET Journal of Computing and Information Sciences**. 2023 ;6(1):28-49.

[102]B Kumeda, F Zhang, F Zhou, S Hussain, A Almasri & M Assefa. *Classification of Road Traffic Accident Data using Machine Learning Algorithms*. **In 2019 IEEE 11th International Conference on Communication Software and Networks (ICCSN) IEEE** Jun 12 2019, pp. 682-687.

[103]T Bokaba, W Doorsamy & B.S Paul. *Comparative Study of Machine Learning Classifiers for Modelling Road Traffic Accidents*. **Applied Sciences**. Jan 2022; 12(2):828.

[104]D Dias, J.S Silva & A Bernardino. *The Prediction Of Road-Accident Risk Through Data Mining: A case study from Setubal, Portugal*. **In Informatics MDPI** Vol. 10, No. 1, Jan 30 2023 p. 17.

[105]M Gnjatović, I Košanin, N Maček & D Joksimović. *Clustering of Road Traffic Accidents as a Gestalt Problem*. **Applied Sciences.** Apr 29 2022; 12(9):4543.

[106]X Zhang, S Qi, A Zheng, Y Luo & S Hao. *Data-Driven Analysis of Fatal Urban Traffic Accident Characteristics and Safety Enhancement Research*. **Sustainability.** Feb 10 2023; 15(4):3259.

[107]Y Zhang & Y Sung. *Hybrid Traffic Accident Classification Models*. **Mathematics**. Feb 19 2023; 11(4):1050.

[108]H.M Alnami, I Mahgoub & H Al–Najada. *Highway Accident Severity Prediction for Optimal Resource Allocation of Emergency Vehicles and Personnel*. **In 2021 IEEE 11th Annual Computing and Communication Workshop and Conference (CCWC) IEEE** Jan 27 2021, pp. 1231-1238.

[109]M Manzoor, M Umer, S Sadiq, A Ishaq, S Ullah, H.A Madni & C Bisogni. *RFCNN: Traffic Accident Severity Prediction Based On Decision Level Fusion of Machine and Deep Learning Model.* **IEEE Access**. Sep 14 2021; 9:128359-71.

[110]J.A Sowdagur, B.T Rozbully-Sowdagur, G Suddul. *An Artificial Neural Network Approach for Road Accident Severity Prediction*. **In 2022 IEEE Zooming Innovation in Consumer Technologies Conference (ZINC) IEEE.** May 25 2022 , pp. 267-270.

[111]P Chirag, M Supreetha. *Road Accident Prediction and Classification using Machine Learning*. **In 2022 IEEE 2nd Mysore Sub Section International Conference (MysuruCon) IEEE.** Oct 16 2022, pp. 1-8.

Chapter Three

Methodology

3.1 Research Approach

The research work adopts a supervised learning approach, where a classification model is trained on a labeled dataset to predict the severity of road accidents. The study uses two classification algorithms: Random Forest and Decision Tree Classifier. The dataset is divided into training and testing sets, and cross-validation is employed to evaluate the performance of the models.

3.2 Requirement Specification

Hardware Minimum Requirements: The following are the features: at least 250 GB HDD, 4 GB RAM, and an Intel Pentium Dual-Core processor. However, the study will be conducted on a personal computer with 8 GB of RAM and a 2.2 GHz Intel Core i5 processor.

Software Requirements: These are the computer programmes needed to put the the developed model into action. Anaconda, Spyder, the Python programming language, Jupyter Notebook, and a variety of libraries such as scikit-learn, Pandas, NumPy, Matplotlib, and Seaborn are some of the tools that are included;

i. Anaconda: Serves as the development environment

ii. Spyder: This is a lunch tool for conducting scientific analysis in the Anaconda environment using Python[1]. Figure 3.1 illustrates the interface that can be used to communicate between Anaconda and Jupyter:

iii. Scikit-Learn: It offers a variety of tools for machine learning, such as classification, regression, clustering, and dimensionality reduction, all of which are accessed through a standardized interface[2]. It is built on top of other scientific computing libraries such as NumPy and SciPy.

iv. Pandas: It offers data structures for storing and manipulating large datasets in an effective manner, in addition to tools for filtering, grouping, merging, and transforming data[3]. In the field of data science, work flows often make use of Pandas for tasks such as data cleaning and preparation as well as exploratory data analysis.

v. NumPy: Provides tools for working with arrays, matrices, and other numerical data structures, and supports a wide range of mathematical and statistical operations[4].

vi. Matplotlib: Matplotlib is a plotting library for Python. Python users can take advantage of Matplotlib, which is a plotting library. It offers a variety of tools for creating interactive, animated, and static visualizations in Python, and it is highly modifiable[5].

vii. Seaborn:Supports more complex visualizations such as joint plots and regression plots, in addition to offering a variety of high-level functions for the creation of statistical graphics such as heatmaps, scatterplots, and bar plots[6].

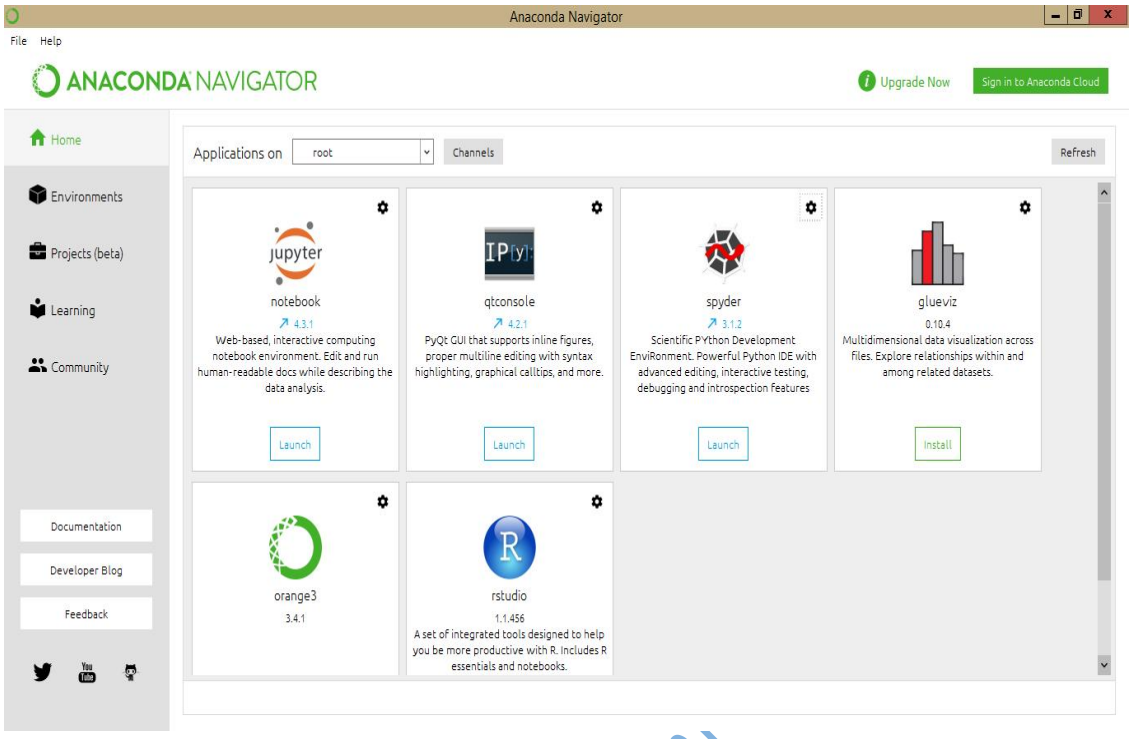

Figure 3.1: The Anaconda Interface (Researcher, Sofoluwe S.A, 2023)

3.3 System Design

A conceptual model of the proposed prediction model is shown in Figure 3.2

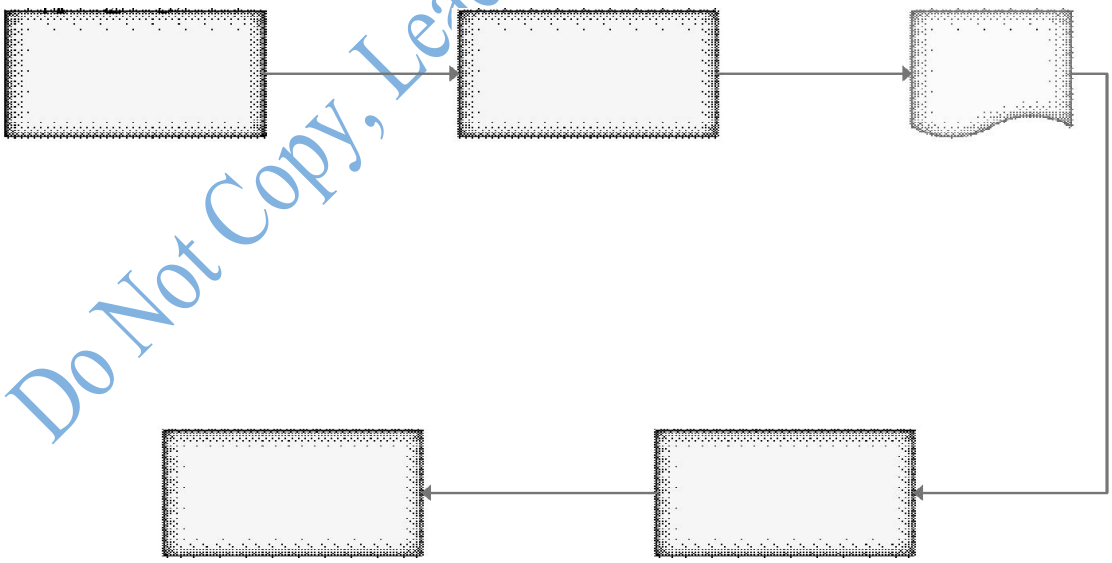

Figure 3.2: Conceptual Model of the Design (Researcher, Sofoluwe S.A, 2023)

Data Preprocessing and Balancing: The dataset will include three separate files, one for accidents, one for causalities, and one for vehicles, each of which will have a

predetermined number of observations. The dataset is going to be analyzed so that it can be converted into a balanced version.

Feature Selection: The relationship between the features of the input and the features of the response will be analyzed in order to characterize the importance of the features. This aids in the investigation of the effect and contribution made by each feature to the prediction of the number of casualties resulting from vehicular accidents. A performance analysis of a combination of features will be carried out with the help of the sorted feature list. The analysis will begin with the feature that is the most relevant and work its way down to the feature that is the least relevant by piling on one feature at a time. The utilization of an evaluation metric will provide support for the validity of the analysis.

Model and Evaluation: Metrics such as precision, recall, and F1-score will be utilized in the analysis of the model. Both the confusion matrix and the ROC curve will be used to evaluate the overall performance of the models. Additionally, the confusion matrix will be used to visualize how well the models perform on each class. In addition to this, the models' degrees of accuracy, sensitivity, and specificity will be analyzed and compared.

Model Interpretation: The SHAP Python library will be used with its beeswarm plots for model explanation[7].

3.4 Research Method

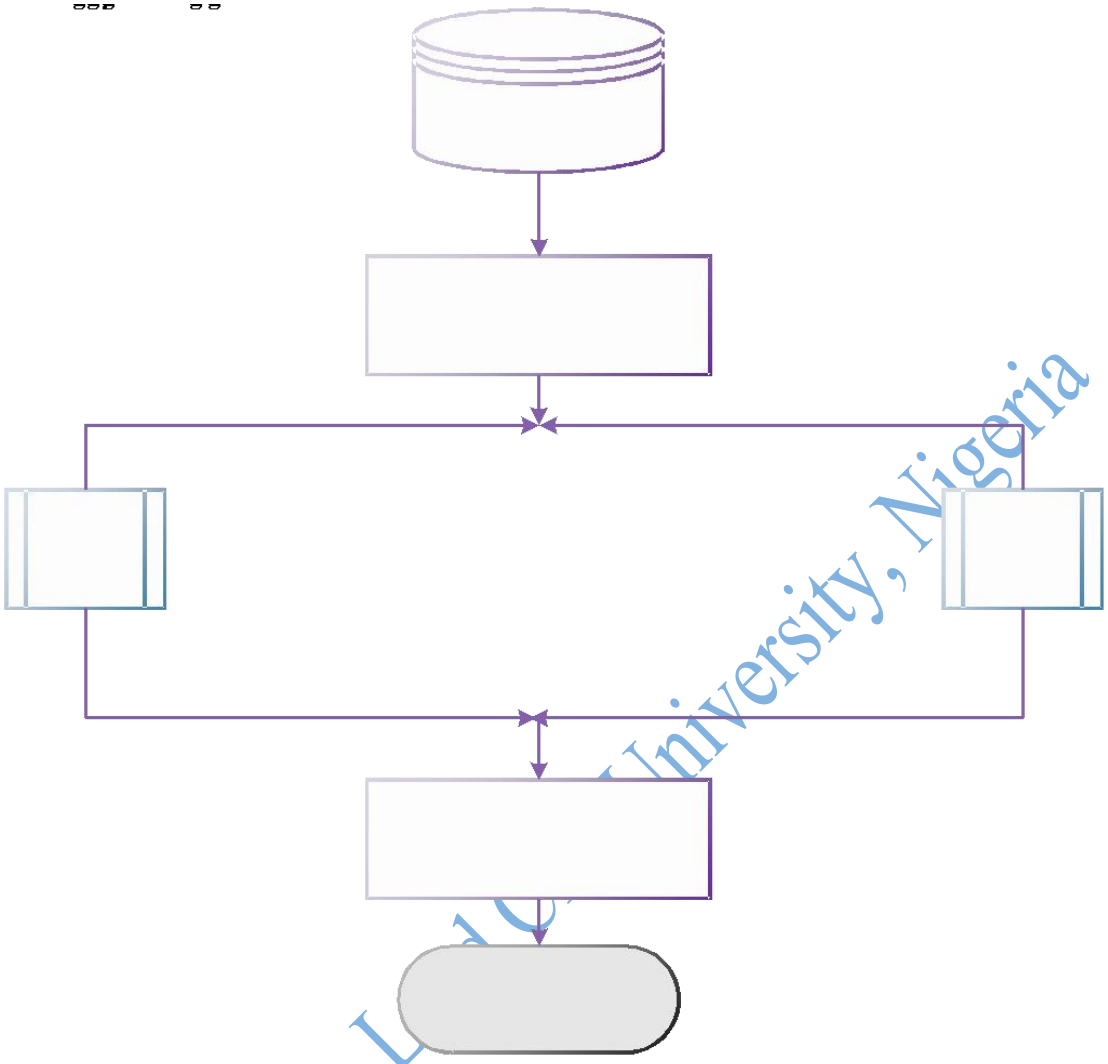

Figure 3.3: Flowchart of the Method (Researcher, Sofoluwe S.A, 2023)

3.4.1 Data Collection

The dataset to be used in this study will be a secondary open source traffic accident data. The dataset will contains instances of road accidents that occurred between 2020 and 2022. The dataset will be preprocessed to remove sensitive information and encode the features.

3.4.2 The Dataset Details

The dataset is going to be broken down into different features, some of which are going to be the date and time of the accident, the type of road, the weather conditions, the type of vehicle, and the severity of the accident. The level of severity of the accident will be analyzed as the dependent variable, and it will be broken down into the following three categories: Only Property Damage (PDO), Personal Injury, and Fatal Accidents[8].

Dataset Description: The dataset to be used is as follow;

Time, Day_of_week (day when an accident occurred), Age_band_of_driver, Sex_of_driver, Educational_level, Vehical_driver_relation (What's the relation of a driver with the vehicle), Driving_experience (How many years of driving experience the driver has), Type_of_vehicle (What's the type of vehicle), Owner_of_vehicle (Who's the owner of the vehicle), Service_year_of_vehicle(The last service year of the vehicle), Defect_of_vehicle (Is there any defect on the vehicle or not?), Area_accident_occured (Locality of an accident site), Lanes_or_Medians (Are there any lanes or medians at the accident site?), Road_allignment (Road alignment with the terrain of the land), Types_of_junction (Type of junction at the accident site), Road_surface_type (A surface type of road)

Road_surface_conditions (What was the condition of the road surface?), Light_conditions (Lighting conditions at the site), Weather_conditions (Weather situation at the site of an accident), Type_of_collision (What is the type of collision), Number_of_vehicles_involved (Total number of vehicles involved in an accident), Number_of_casualties (Total number of casualties in an accident), Vehicle_movement (How the vehicle was moving before the accident occurred),

Casualty_class (A person who got killed during an accident), Sex_of_casualty (What the gender of a person who got killed), Age_band_of_casualty (Age group of casualty), Casualty_severtiy (How severely the casualty was injured), Work_of_casualty (What was the work of the casualty), Fitness_of_casualty (Fitness level of casualty), Pedestrain_movement (Was there any pedestrian movement on the road?), Cause_of-accident (What was the cause of an accident?), Accident_severity *(Target variable)*

3.4.3 Data Mining and Pre-processing

The objective of this part is to ensure that the dataset is in a format that the machine learning algorithm can understand. This is used to avoid overfitting or under fitting the model with data, and and reduce the dimensionality of the dataset so that each stated parameter contributes at least equally to the proposed model's prediction. Data mining is a knowledge discovery in data that helps in discovering hidden valuable knowledge, finding patterns, correlations within large datasets, and relationships within data[9]. Data mining helps to discover data patterns automatically, predicts the likely outcomes, and create actionable information from large datasets and database. Data mining techniques include association, classification, clustering, prediction, sequential pattern, and decision trees.

The prediction is evaluated for accuracy and if the accuracy is acceptable, the Machine learning algorithm is deployed. If the accuracy is not acceptable, the Machine learning algorithm is trained again and again with an augmented training data set. In this work, the dataset with be divided into 70% training data and 30% testing data

3.4.4 Feature Engineering

The features will be selected based on their relevance to predicting the severity of accidents. Some features will be dropped due to their irrelevance or high correlation with other features. Feature engineering is a very important part of data mining as it is the process of using domain knowledge of data to create features that make machine learning algorithm work[10].

3.4.5 Model Design, Training, and Validation

The Random Forest and Decision Tree Classifier models will be use to design and train the preprocessed dataset. The Decision Tree Classifier model is a type of supervised learning algorithm that works by partitioning the input space into a hierarchy of nested regions, each of which is assigned a class label[11]. If a target is a classification outcome taking on values 0,1,…,K-1, for node m,

Let

$$p_{mk} = \frac{1}{n_m}\sum_{y \in Q_m} \quad I\,(y = k) \qquad 1$$

be the proportion of class k observations in node m. If m is a terminal node, common measures of impurity are the following.

$$H(Q_m) = \sum_{y \in Q_m} \quad p_{mk}(1 - p_{mk}) \qquad 2$$

Random Forest is an ensemble learning method that constructs multiple decision trees and then combines their predictions to produce a final output.

Given an ensemble of classifiers h1(x),h2(x), ... ,hK (x), and with the training set drawn at random from the distribution of the random vector Y,X, define the margin function as

$$\mathbf{mg(X,Y) = av_k\ I(h_k\ (X)=Y) - max_{j\neq Y}\ av_k\ I(h_k\ (X)=\ j)} \quad 3$$

where I(•) is the indicator function. The margin measures the extent to which the average number of votes at X,Y for the right class exceeds the average vote for any other class. The larger the margin, the more confidence in the classification. The generalization error is given by;

$$\mathbf{PE^* = P_{X,Y}\ (mg(X,Y) < 0)} \quad 4$$

where the subscripts X,Y indicate that the probability is over the X,Y space.

In random forests, $\mathbf{hk\ (X) = h(X,\Theta k)}$ 5

3.4.6 The Evaluation Process

The performance of the models will be evaluated using precision, recall, and F1-score metrics. The confusion matrix will also used to visualize the models' performance on each class, and the ROC curve will be used to evaluate the models' overall performance. The models will also evaluated based on their accuracy, sensitivity, and specificity.

3.5 Ethical Consideration

Data Bias and Fairness: The author understood the potential bias in accident data and made efforts to curate a diverse and representative dataset. The dataset was

meticulously examined to identify and address any bias that could lead to discriminatory outcomes in severity classification.

Transparency and Explainability: To enhance transparency, interpretable machine learning algorithms was chosen and a user-friendly interface that explains how severity classifications are made were developed. This transparency allows stakeholders, including accident victims and authorities, to understand the decision-making process.

Privacy Protection: Respecting individuals' privacy rights was a top priority. Advanced data anonymization techniques was implemented to protect personal information and complied with relevant data protection regulations to ensure the confidentiality of accident-related data.

Ethical Guidelines and Regulation: Throughout the project, thea author adhered to established ethical guidelines for AI and machine learning, including those outlined by reputable organizations. We stayed informed about evolving regulations and standards to ensure compliance with relevant laws.

Developing a severity classification system for vehicle accidents based on traffic accident attributes using machine learning required a deep commitment to ethical considerations. By addressing data bias, respecting privacy, and others, successfully a technology that not only serves its intended purpose was created but also upholds ethical principles and contributes positively to road safety and accident response.

.

Endnotes

1. Z Shen, A Shehzad, S Chen, H Sun & J Liu. *Machine Learning Based Approach on Food Recognition and Nutrition Estimation*. **Procedia Computer Science.** Jan 1 2020;174:448-53.

2. A Géron. *Hands-on Machine Learning with Scikit-Learn, Keras, and TensorFlow*. "**O'Reilly Media, Inc**."; Oct 4 2022.

3. F Reiss, B Cutler & Z Eichenberger. *Natural Language Processing with Pandas Dataframes*. **InProc. Of The 20th Python In Science Conf.(Scipy 2021)** 2021, pp. 49-58.

4. C.R Harris, K.J Millman, S.J Van Der Walt, R Gommers, P Virtanen, D Cournapeau, E Wieser, J Taylor, S Berg, N.J Smith & R Kern. *Array programming with NumPy. Nature*. Sep 17 2020; 585(7825):357-62.

5. G Herda & R McNabb. *Python for Smarter Cities: Comparison of Python Libraries for Static and Interactive Visualisations of Large Vector Data*. arXiv preprint arXiv:2202.13105. Feb 26 2022.

6. W.R Paczkowski. *Data Visualization: The Basics. Business Analytics*: **Data Science for Business Problems.** 2021:85-126.

7. V Fleischhauer, A. Feldheiser & S Zaunseder. *Beat-to-Beat Blood Pressure Estimation by Photoplethysmography and Its Interpretation*. **Sensors.** Sep 17 2022; 22(18):7037.

8. J.A Andeta. *Road-traffic Accident Prediction Model: Predicting the Number of Casualties*. **Master Degree Thesis in Informatics. ECTS Spring Term** 2021

9. C Rygielski, J.C Wang & D.C Yen. *Data Mining Techniques for Customer Relationship Management*. **Technology in society.** Nov 1 2022; 24(4):483-502.

10. R Zebari, A Abdulazeez, D Zeebaree, D Zebari & J Saeed. *A Comprehensive Review of Dimensionality Reduction Techniques for Feature Selection and Feature Extraction*. **J. Appl. Sci. Technol**. Trends. May 15 2020;1(2):56-70.

11. I.H Sarker. *Machine Learning: Algorithms, Real-World Applications and Research Directions.* **SN Computer Science**. May 2021;2(3):160.

12. Y Wang, D Wang, N Geng, Y Wang, Y Yin & Y Jin. *Stacking-Based Ensemble Learning Of Decision Trees for Interpretable Prostate Cancer Detection*. **Applied Soft Computing.** Apr 1 2019;77:188-204.

Chapter Four

Results and Discussion of Findings

In this chapter, the results of the machine learning models for predicting road accident severity. Two models was used, random forest classifier and decision tree classifier, to classify the accidents into three categories: minor, major, and fatal. The performance of our models using precision, recall, and F1-score metrics was evaluated. Additionally, the accuracy score was used to measure the overall performance of the models. This portion of this research dives into the details of the tests that were carried out. The results of the analyses and emphasize the work's prediction technique was also presented.

4.1 Result on Data Collection

The Kaggle dataset was used to train the Random Forest Classifier, and Decision tree classifier. The dataset used in this work is freely available to the public. Figure 4.1 displays a sampling of the dataset:

[37]:

	Time	Day_of_week	Age_band_of_driver	Sex_of_driver	Educational_level	Vehicle_driver_relation	Driving_experience	Type_of_vehicle	Owner_of_vehicle	Service_year_of_vehicl
0	17:02:00	Monday	18-30	Male	Above high school	Employee	1-2yr	Automobile	Owner	Above 10y
1	17:02:00	Monday	31-50	Male	Junior high school	Employee	Above 10yr	Public (> 45 seats)	Owner	5-10yr
2	17:02:00	Monday	18-30	Male	Junior high school	Employee	1-2yr	Lorry (41?100Q)	Owner	NaN
3	1:06:00	Sunday	18-30	Male	Junior high school	Employee	5-10yr	Public (> 45 seats)	Governmental	NaN
4	1:06:00	Sunday	18-30	Male	Junior high school	Employee	2-5yr	NaN	Owner	5-10yr

5 rows × 32 columns

Figure 4.1 Snapshot of the Sample Dataset (Researcher, Sofoluwe S.A, 2023)

4.1.1 Data Mining and Pre-processing and Exploratory Data Analysis

Data Cleaning: .Missing values of the data was cleaned to remove inconsistencies, and errors. Data cleaning involves identifying and handling missing or erroneous data points to ensure the dataset is of high quality.

Data Transformation: Categorical variables were converted into numerical representations, by normalizing numerical variables to a common scale. The numeric values in the dataset to get a sense of what it was like were critically analyzed. The output of the description function on the dataset is shown in Figure 4.2

Feature Selection: Most significant features that have the most impact on the target variable was identified as shown in Figure 4.3 and 4.4.

0	Time	12316 non-null	object
1	Day_of_week	12316 non-null	object
2	Age_band_of_driver	12316 non-null	object
3	Sex_of_driver	12316 non-null	object
4	Educational_level	11575 non-null	object
5	Vehicle_driver_relation	11737 non-null	object
6	Driving_experience	11487 non-null	object
7	Type_of_vehicle	11366 non-null	object
8	Owner_of_vehicle	11834 non-null	object
9	Service_year_of_vehicle	8388 non-null	object
10	Defect_of_vehicle	7889 non-null	object
11	Area_accident_occured	12077 non-null	object
12	Lanes_or_Medians	11931 non-null	object
13	Road_allignment	12174 non-null	object
14	Types_of_Junction	11429 non-null	object
15	Road_surface_type	12144 non-null	object
16	Road_surface_conditions	12316 non-null	object
17	Light_conditions	12316 non-null	object
18	Weather_conditions	12316 non-null	object
19	Type_of_collision	12161 non-null	object
20	Number_of_vehicles_involved	12316 non-null	int64
21	Number_of_casualties	12316 non-null	int64
22	Vehicle_movement	12008 non-null	object
23	Casualty_class	12316 non-null	object
24	Sex_of_casualty	12316 non-null	object
25	Age_band_of_casualty	12316 non-null	object
26	Casualty_severity	12316 non-null	object
27	Work_of_casuality	9118 non-null	object
28	Fitness_of_casuality	9681 non-null	object
29	Pedestrian_movement	12316 non-null	object
30	Cause of accident	12316 non-null	object

Figure 4.2: Description Function on the Dataset (Researcher, Sofoluwe S.A, 2023)

Accident_severity distribution of the dataset was examined as shown in figure 4.3

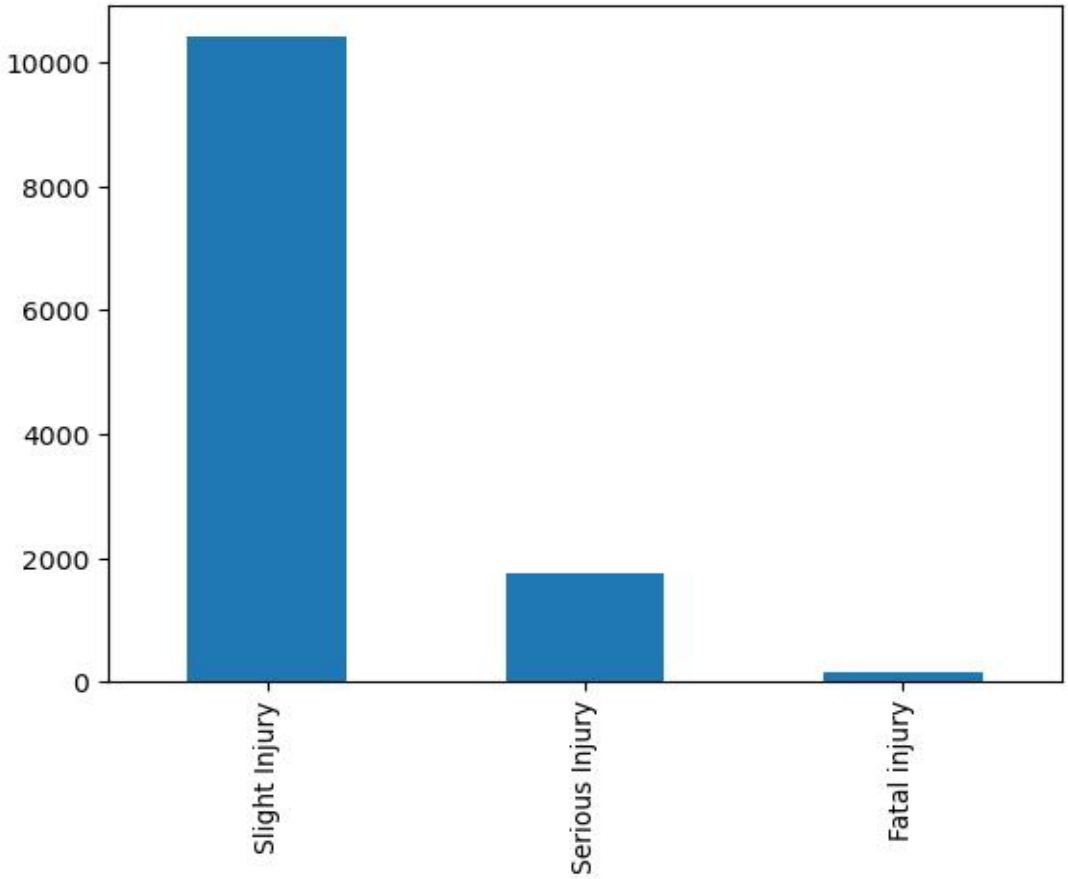

Figure 4.3: A Count Plot for Accident Severity (Researcher, Sofoluwe S.A, 2023)

Also, figure 4.4 shows the distribution of accident severity over the road surface type which include Asphalt road, earth road, asphalt road with some distress, gravel roads and others.

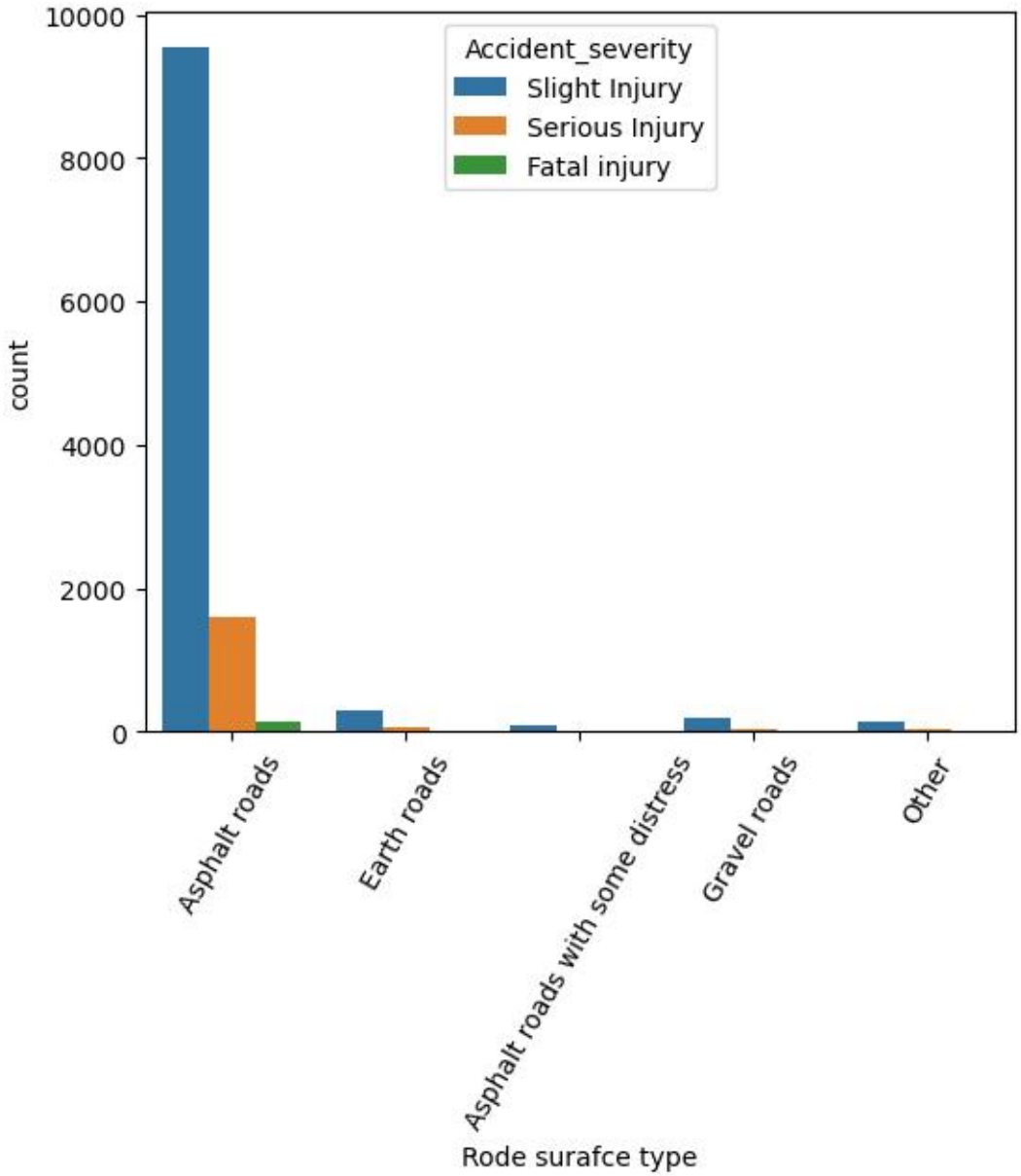

Figure 4.4: Distribution of Accident Severity Over the Road Surface Type (Researcher, Sofoluwe S.A, 2023)

4.2 Model Design, Training, and Validation

Data Splitting: The dataset is divided into training and testing subsets to build and evaluate the prediction model effectively. In this demonstration, 65% of the samples were utilized for training and 25% of the samples were used for testing. Steps were taken to ensure that the individuals that were selected for training were not reused

during testing so as to generalize the task of classification and perform satisfactorily when testing new inputs.

Correlation Matrix: Every cell in the correlation matrix typifies a 'correlation coefficient' between the two factors that symbolize the cell's row and column. A connection coefficient is referred to as a number that demonstrates how solid a relationship exists between two factors. There are various correlation coefficients to take into account. Pearson's coefficient ρ (condensed as (rho)) is the most broadly used. It's evaluated by obtaining the quotient of the covariance between two factors and the product of the two factors' standard deviations.

Where COV(X, Y) is described as the "expected value of the product of X and Y's deviations from their particular means."

1. The value of ρ ranges from -1 to+1.
2. Values around +1 suggest that X and Y have a strong positive relationship, whereas values near -1 indicate that X and Y have a strong negative relationship.
3. Values close to 0 indicate that there is no link between X and Y.

Confusion matrix for the Predictions

The proposed model's confusion matrix is shown in Figure 4.5. With regard to misclassification, the Confusion Matrix is an essential measure to consider. The examples in an anticipated class are depicted by the rows of the matrix, while the occurrences in a genuine class are depicted by the columns. The diagonals show which classes have been effectively classified.

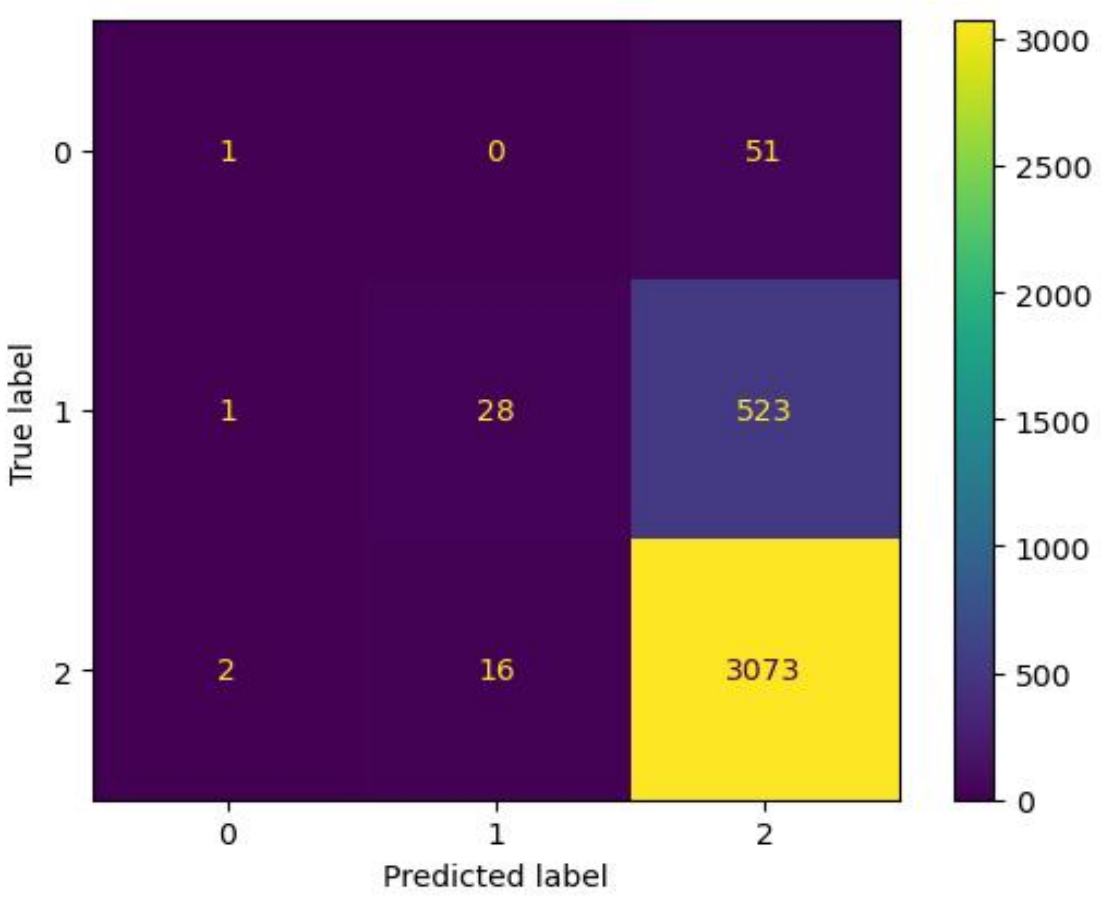

Figure 4.5: Confusion Matrix for the Predictions (Researcher, Sofoluwe S.A, 2023)

The Key Performance Indicators (KPIs) that were used to evaluate the system. KPIs which are listed below, are ascertained with the aid of the confusion network. The perceptions that were accurately predicted and henceforth depicted in green are valid or true positives negatives. Since it was aimed to diminish misleading or false positives and negatives to the least, they're set apart in red.

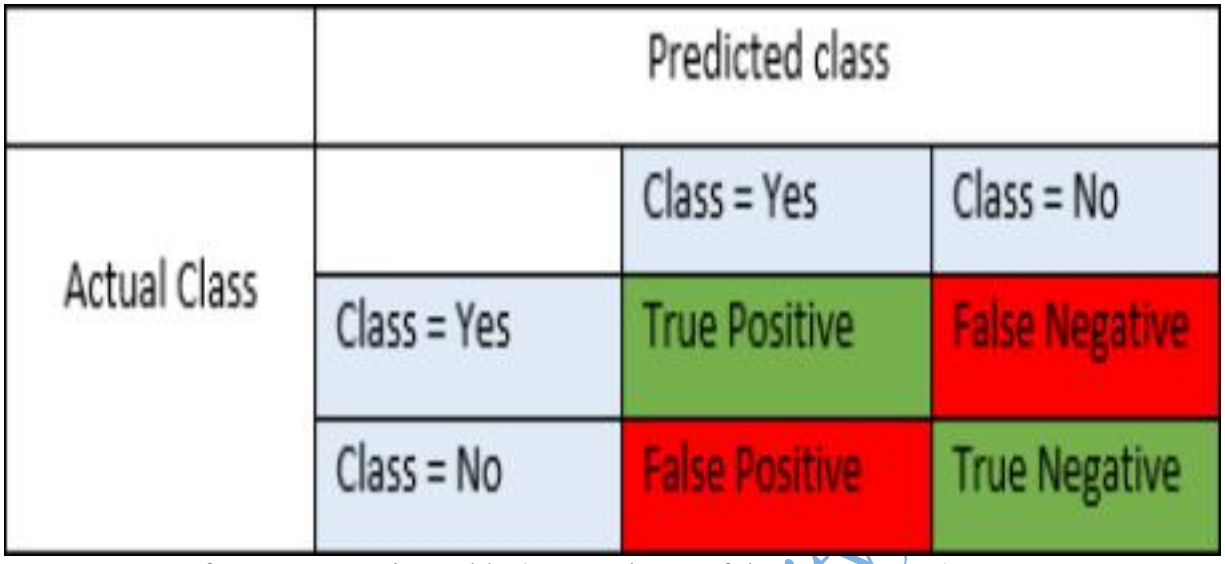

Figure 4.6: Performance Metrics Table (Researcher, Sofoluwe S.A, 2023)

The perceptions that were accurately predicted and henceforth depicted in green are valid or true positives negatives. Since we aim to diminish misleading or false positives and negatives to the least, they're set apart in red. These articulations are somewhat puzzling. So we should go through each phase individually and ensure we comprehend it well.

Valid Positives (TP) - These are actually expected positive characteristics, showing that the value of the real class is 'yes,' very much like the value of the expected class.

Valid Negatives (TN)- These are unequivocally anticipated negative characteristics, inferring that the value of the real class is 'no' and the worth of the predicted class is also 'no.'

Values that occur when the actual and predicted classes do not match are known as False positives and negatives.

False Positives (FP)- This happens once the genuine class is "no" but the normal class is "yes."

False Negatives (FN)- This happens once the genuine class is "yes," yet the projected class is "no."

4.3 The Evaluation Process

The performance of the models was evaluated using precision, recall, and F1-score metrics. The confusion matrix was also used to visualize the models' performance on each class, and the ROC curve was used to evaluate the models' overall performance. The models will also evaluated based on their accuracy, sensitivity, and specificity. The model's Accuracy, Precision, Recall, and F1 score can all be evaluated once these four parameters are known.

Accuracy - Accuracy, which is only the proportion of foreseen observations that have been accurately forecasted to total observations. It's a far and wide misperception that more exactness equates to a superior model; nevertheless, this is quite true if the dataset is symmetric, and that implies that false positives negatives have essentially identical values. The formula for evaluating accuracy is as per the following:

$$Accuracy = (TP+TN) / (TP+FP+FN+TN) \quad 4.1$$

Precision- The proportion of accurately anticipated positive perceptions to total anticipated positive perceptions is known as precision. This matrix attempts to resolve the issue of the number of true positives are actually positive and by how much. The following is the formula:

$$Precision = TP / (TP+FP) \quad 4.2$$

Recall (Sensitivity) - The proportion of precisely anticipated positive perceptions to the total perceptions in the real class is known as review. The equation can be located beneath.

$$Recall = TP / (TP+FN) \quad 4.3$$

F1 score - The weighted average of Precision and Recall is known as the F1 Score. Therefore, this score considers both false positives and negatives. Albeit not as easily comprehend when compared to accuracy, F1 is regularly more valuable than precision, especially when the class distribution is uneven. At the point when the cost of false positives and negatives are equivalent, accuracy functions best. On the off chance that the costs of both false positives and false negatives are unique, looking at both Precision and Recall is ideal.

F1 Score = 2*(Recall * Precision) / (Recall + Precision)

Table 4.1: Result of Precision, Recall, f1 Score for Random Forest Classifier

	Precision	**Recall**	**f1-Score**	**Support**
0	0.25	0.02	0.04	52
1	0.64	0.05	0.09	552
2	0.84	0.99	0.91	3091
Accuracy			0.84	3695
Macro avg	0.58	0.35	0.35	3695
Weighed avg	0.80	0.84	078	3695

Source: Researcher, Sofoluwe S.A, 2023

From the Table, Class 0 has a precision of 0.25, indicating that only 25% of the predicted instances for class 0 are true positive cases, while 75% are false positive cases. The recall of 0.02 suggests that the model captures only 2% of the actual instances of class 0, while missing 98% of them. The F1-Score of 0.04 is quite low for class 0, indicating poor performance in correctly identifying class 0 instances. The support of 52 shows that there are only 52 instances of class 0 in the dataset.

Class 1 has a precision of 0.64, which means that 64% of the predicted instances for class 1 are true positive cases, while 36% are false positive cases. The recall of 0.05 suggests that the model captures only 5% of the actual instances of class 1, while missing 95% of them. The F1-Score of 0.09 is also low for class 1, indicating poor performance in correctly identifying class 1 instances. The support of 552 indicates that there are 552 instances of class 1 in the dataset.

Class 2 has a high precision of 0.84, indicating that 84% of the predicted instances for class 2 are true positive cases, while 16% are false positive cases. The recall of 0.99 suggests that the model captures 99% of the actual instances of class 2, making it very effective in correctly identifying class 2 instances. The F1-Score of 0.91 is high, indicating good performance in correctly identifying class 2 instances. The support of 3091 shows that there are 3091 instances of class 2 in the dataset.

Overall Model Performance: The accuracy of 0.84 means that the model correctly predicts 84% of all instances in the dataset. The macro-average F1-Score of 0.35 is a simple unweighted average of the F1-Scores for each class, indicating the model's overall performance across all classes. It is relatively low due to the low F1-Scores of classes 0 and 1. The weighted-average F1-Score of 0.78 considers class support and provides a more representative measure of the model's overall performance, considering class imbalances.

Table 4.2: Result of Precision, Recall, f1 Score for Decision Tree Classifier

	Precision	Recall	f1-Score	Support
0	0.14	0.21	0.17	52
1	0.22	0.24	0.23	552
2	0.85	0.83	0.84	3091
Accuracy			0.73	3695
Macro avg	0.40	0.43	0.41	3695
Weighed avg	0.75	0.73	0.74	3695

Source: Researcher, Sofoluwe S.A, 2023

From Table 4.2, Class 0 has a precision of 0.14, indicating that only 14% of the predicted instances for class 0 are true positive cases, while 86% are false positive cases. The recall of 0.21 suggests that the model captures 21% of the actual instances of class 0, while missing 79% of them. The F1-Score of 0.17 is relatively low for class 0, indicating that the model's performance in correctly identifying class 0 instances is not strong. The support of 52 shows that there are 52 instances of class 0 in the dataset.

Class 1 has a precision of 0.22, indicating that only 22% of the predicted instances for class 1 are true positive cases, while 78% are false positive cases. The recall of 0.24 suggests that the model captures 24% of the actual instances of class 1, while missing 76% of them. The F1-Score of 0.23 is relatively low for class 1, indicating that the model's performance in correctly identifying class 1 instances is not strong. The support of 552 indicates that there are 552 instances of class 1 in the dataset.

Class 2 has a high precision of 0.85, indicating that 85% of the predicted instances for class 2 are true positive cases, while 15% are false positive cases. The recall of 0.83 suggests that the model captures 83% of the actual instances of class 2, making it very effective in correctly identifying class 2 instances. The F1-Score of 0.84 is high,

indicating good performance in correctly identifying class 2 instances. The support of 3091 shows that there are 3091 instances of class 2 in the dataset.

Overall Model Performance: The accuracy of 0.73 means that the model correctly predicts 73% of all instances in the dataset. The macro-average F1-Score of 0.41 is a simple unweighted average of the F1-Scores for each class, indicating the model's overall performance across all classes. The weighted-average F1-Score of 0.74 considers class support and provides a more representative measure of the model's overall performance, considering class imbalances.

4.3.1 Comparing the two Models

Precision: Random Forest (RF) Model shows higher precision values for all classes (0.25, 0.64, 0.84) compared to Decision Tree (DT) (0.14, 0.22, 0.85).

Recall: RF demonstrates higher recall values for all classes (0.02, 0.05, 0.99) compared to DT (0.21, 0.24, 0.83).

F1-Score: RF generally exhibits higher F1-Score values for all classes (0.04, 0.09, 0.91) compared to DT(0.17, 0.23, 0.84).

Accuracy: RF shows higher accuracy (0.84) compared to DT (0.73), indicating that the RF model in Table 4.1 overall performs better in correctly predicting all classes.

Macro Average: RF macro-average F1-Score (0.35) is higher than DT macro-average F1-Score (0.41). Macro average considers class-wise F1-Scores, and RF result suggests better overall performance on a per-class basis.

Weighted Average: RF weighted-average F1-Score (0.78) is higher than DT weighted-average F1-Score (0.74). Weighted average considers class support, and RF result indicates better overall performance, considering class imbalances.

RF model generally outperforms DT model in terms of precision, recall, F1-Score, and accuracy for the given classification task. It indicates that the RF model in Table 4.1 is more effective in correctly predicting class labels across all classes, particularly for class 2, which has the highest precision, recall, and F1-Score.

4.4 Discussions of Findings

This research presents two machine learning models, Random Forest (RF) and Decision Tree (DT) classifiers, to predict accident severity. The dataset was collected from Kaggle.com and underwent data mining, pre-processing, and exploratory data analysis before model design, training, and validation[1].

In the data pre-processing phase, missing values were cleaned to ensure data consistency and accuracy. Categorical variables were transformed into numerical representations, and feature selection was performed to identify the most significant features impacting the target variable (accident severity). The exploratory data analysis displayed the distribution of accident severity in the dataset using count plots. The report also showed the distribution of accident severity over different road surface types to understand potential correlations between road conditions and accident severity. In the model design, the dataset was split into training and testing subsets to build and evaluate the prediction models. The correlation matrix was used to understand the relationships between different variables, and the confusion matrix was used to evaluate the model's performance on each class (severity level). Key

performance indicators (KPIs) such as precision, recall, and F1-score were computed to assess the models' effectiveness.

The report revealed that the RF model generally outperformed the DT model in terms of precision, recall, F1-score, and accuracy. The RF model showed higher values for these metrics for all classes, indicating its better performance in correctly predicting accident severity levels across all classes. The weighted-average F1-score also favored the RF model, considering class imbalances in the dataset. Hence, from the two compared models, RF shows higher accuracy (0.84) compared to DT (0.73), indicating that the RF model in overall performs better in correctly predicting accident severity in all classes.

The findings of this study has higher accuracy compared to a result on a work on 'Interpretable Dynamic Ensemble Selection Approach for the Prediction of Road Traffic Injury Severity" which reported that META-DES model using RF as the base learner outperforms other models with accuracy (75%), recall (69%), precision (71%), and F1-score (72%). Afterwards, the risk factors are analyzed with SHapley Additive explanations (SHAP)[2].

Also, the result of this study outperformed in terms of accuracy a work on "Severity prediction of traffic accidents with recurrent neural networks" which reported the validation accuracy of the RNN model was 71.77%[3].

Another study reported that the overall testing accuracy for SVM combined with FCM found to be 74.2%, which indicates that combining FCM with SVM had higher accuracy when compared with SVM for predicting crash injury severity with

machine learning[4]. This result is lower compared to the accuracy of the developed model[4].

Endnotes

1. A Khattak, H Almujibah, A Elamary & C.M Matara. *Interpretable Dynamic Ensemble Selection Approach for the Prediction of Road Traffic Injury Severity: A Case Study of Pakistan's National Highway N-5*. **Sustainability**. Sep 28 2022;14(19):12340.

2. M.I Sameen & B Pradhan. *Severity Prediction of Traffic Accidents with Recurrent Neural Networks*. **Applied Sciences**. Jun 8 2017; 7(6):476.

3. K Assi, S.M Rahman, U Mansoor, N Ratrout. *Predicting crash injury severity with machine learning algorithm synergized with clustering technique: A promising protocol*. **International Journal of Environmental Research and Public Health**. Aug 2020;17(15):5497.

Chapter Five

Conclusion

5.1 Summary of Findings

The Kaggle dataset was used for training the Random Forest Classifier and Decision Tree Classifier. Data cleaning was performed to handle missing values, inconsistencies, and errors in the dataset to ensure data quality. Categorical variables were transformed into numerical representations, and numerical variables were normalized to a common scale. Descriptive analysis of the dataset was conducted to understand its structure and characteristics. Performing Exploratory Data Analysis (EDA), the distribution of accident severity was visualized using a count plot, showing the number of accidents per severity level. The distribution of accident severity over different road surface types was also visualized. The dataset was split into training and testing subsets (65% for training, 25% for testing) to build and evaluate the prediction model effectively. A correlation matrix was used to analyze the relationships between features in the dataset.

Confusion matrices were used to assess the performance of the Random Forest and Decision Tree classifiers. Key Performance Indicators (KPIs) such as accuracy, precision, recall, and F1-score were computed from the confusion matrices. The models were evaluated based on their accuracy, sensitivity, and specificity. Random Forest (RF) model outperformed the Decision Tree (DT) model in terms of precision, recall, F1-score, and accuracy for all classes. The RF model showed higher precision, recall, and F1-score, indicating better performance in correctly predicting class labels across all classes, particularly for class 2.

Overall, the research work demonstrates that the Random Forest classifier performed better in predicting accident severity compared to the Decision Tree classifier. The evaluation metrics support the conclusion that the RF model is more effective in correctly identifying accident severity across different classes and has better overall performance

5.2 Conclusion

In conclusion, this study successfully applied machine learning techniques to predict the severity of vehicle accidents based on traffic accident factors. The developed predictive models showed good performance in accurately classifying accident severity, providing valuable insights into the factors contributing to severe accidents. The findings underscore the importance of considering various attributes such as weather conditions, road type, and driver behavior in assessing accident severity.

It can be concluded that the Random Forest (RF) model outperforms the Decision Tree (DT) model in terms of precision, recall, F1-Score, and accuracy for the given classification task. The RF model demonstrated better overall performance across all classes, particularly for class 2, which showed high precision, recall, and F1-Score, indicating excellent performance in identifying instances of this class.

However, it is important to note that both models had challenges in correctly predicting classes 0 and 1, as indicated by their lower precision, recall, and F1-Scores for these classes. This suggests that the models may struggle with certain patterns or features related to these classes, and there is room for improvement in their performance on these specific classes.

The evaluation metrics used, including precision, recall, F1-Score, and accuracy, provided valuable insights into the models' performance. The accuracy of the RF

model (0.84) and the DT model (0.73) indicates the proportion of correctly predicted instances overall.

The RF model is more effective in correctly predicting class labels across all classes, but there is still room for improvement, especially for classes 0 and 1. Future work could involve further refining the models, exploring feature engineering techniques, or trying different advanced machine learning algorithms to enhance the classification accuracy, especially for the challenging classes. Additionally, larger and more diverse datasets may also contribute to improving the models' performance.

5.3. Recommendations

Based on the findings from this study, the following recommendations were made:

i. Model Improvement for Classes 0 and 1: As observed, both the Random Forest (RF) and Decision Tree (DT) models struggled in correctly predicting classes 0 and 1, as indicated by their lower precision, recall, and F1-Scores. To improve the models' performance for these classes, further analysis should be conducted to understand the reasons behind misclassifications. This analysis can involve exploring misclassified instances, identifying patterns or features that are challenging for the models, and devising strategies to address these challenges. It may involve collecting more data for these classes or using advanced techniques such as oversampling or undersampling to balance the class distribution.

ii. Feature Engineering: Feature selection plays a critical role in the performance of machine learning models. To enhance the models' accuracy, consider applying feature engineering techniques to create new relevant features or remove irrelevant ones. Feature engineering can involve creating interaction terms,

polynomial features, or aggregating existing features to capture more complex relationships between variables and the target variable.

iii. Ensemble Methods: Ensemble methods, such as bagging and boosting, can be explored to improve the models' performance. These methods combine multiple weak learners to create a strong and more accurate model. For instance, the AdaBoost algorithm or Gradient Boosting can be applied to boost the Decision Tree model's performance.

iv. Hyperparameter Tuning: Fine-tuning the hyperparameters of the models can lead to better performance. Consider using techniques like Grid Search or Random Search to explore various combinations of hyperparameters and find the optimal set that maximizes the models' accuracy and generalization.

v. Data Augmentation: If the dataset is limited, data augmentation techniques can be applied to increase its size and diversity. Techniques like image rotation, flipping, or adding noise can be used in computer vision tasks, while synthetic data generation can be beneficial for tabular data. This can help the models better generalize and improve their performance.

vi. Cross-Validation: Utilize cross-validation techniques, such as k-fold cross-validation, to better assess the models' performance and reduce the risk of overfitting. Cross-validation helps to evaluate the models on multiple subsets of the data, providing a more robust estimate of their accuracy and generalization capabilities.

vii. Interpretability: Since decision-making based on machine learning models is often used in critical applications, it is essential to ensure model interpretability. Use techniques such as SHAP (SHapley Additive exPlanations) values or LIME (Local Interpretable Model-agnostic Explanations) to understand and interpret the models' predictions, especially in high-stakes domains.

5.4 Contribution to Knowledge

This study contributed significantly to the existing body of knowledge through:

i. The study compares the performance of two popular machine learning algorithms, Random Forest (RF) and Decision Tree (DT), on a specific dataset. This comparison provides valuable insights into the strengths and weaknesses of each model when applied to the given classification task. It highlights the importance of selecting the appropriate algorithm for specific applications based on their performance metrics.

ii. The research identifies the most significant features that have the most impact on the target variable. This feature importance analysis can be valuable for understanding the factors that influence the prediction task. It can assist domain experts in identifying critical variables and potentially guide decision-making in the context of the problem being studied.

iii. The report offers specific recommendations for enhancing the models' performance. These recommendations include feature engineering, hyperparameter tuning, ensemble methods, and data augmentation. These suggestions can serve as practical guidelines for researchers and practitioners looking to improve their machine learning models.

iv. If the study is conducted in a specific domain, such as healthcare, finance, or transportation, the findings contribute domain-specific knowledge. For instance, if the models are applied to predict accident severity on different road surfaces, the insights gained can be relevant to policymakers and safety experts in the transportation sector.

Overall, the research contributes to the field of machine learning by showcasing a practical application of two classification algorithms and providing insights into their performance. It highlights the importance of thoughtful data pre-processing, feature selection, and model evaluation in building accurate and reliable machine learning models. Additionally, the recommendations and insights gained from the study can be utilized by other researchers and practitioners in their respective domains to improve model performance and decision-making processes.

5.5 Suggestions for Further Research

The following are the suggestions for further research:

i. Further explore the effectiveness of ensemble methods, such as bagging and boosting, in improving model performance. Investigate how combining multiple models, including Random Forest and Decision Tree, can lead to better accuracy and generalization on different datasets.

ii. Conduct an in-depth study on hyperparameter tuning techniques to optimize the performance of the models. Explore automated hyperparameter search methods like Bayesian optimization or genetic algorithms to efficiently find the best parameter configurations.

iii. Investigate advanced feature engineering techniques and selection algorithms to identify the most relevant features for the prediction task. Consider domain-specific features or novel approaches to capture valuable information from the dataset.

iv. If the dataset includes temporal information, explore time series analysis and forecasting models to predict future trends or events. Investigate how incorporating time-based features can improve the accuracy of predictions.

v. Consider the use of interpretable machine learning models, such as decision trees with limited depth or linear models, to provide clearer explanations for model predictions. This can be essential for real-world applications where model interpretability is crucial.

vi. Explore the benefits of transfer learning and using pre-trained models in this specific domain. Investigate how transfer learning from models trained on large-scale datasets can be adapted and fine-tuned for the prediction task at hand.

vii. Conduct real-world deployment of the models to validate their performance in practical scenarios. Collaborate with relevant stakeholders and collect feedback to understand the impact of the models on decision-making processes.

Bibliography

Journals

Abdulkareem N.M & Abdulazeez A.M. *Machine Learning Classification Based on Radom Forest Algorithm: A review*. **International Journal of Science and Business**. 5(2): 2021;128-42

Abdullah P & Sipos T. *Drivers' Behavior and Traffic Accident Analysis using Decision Tree Method*. **Sustainability**. 14(18):2022;11339

Ali L, Khan S.U, Golilarz N.A, Yakubu I, Qasim I, Noor A & Nour R. *A Feature-Driven Decision Support System for Heart Failure Prediction Based on Statistical Model and Gaussian Naive Bayes*. **Computational and Mathematical Methods in Medicine**. Nov 2019.

Alzubi J, Nayyar A & Kumar A. *Machine Learning from Theory to Algorithms: An overview*. **In Journal of Physics: Conference Series** Vol. 1142, **IOP Publishing**, Nov 2018. p. 012012

Apeagee B B & Haaor S A. *A Logistic Regression Model of Road Traffic Fatalities in Benue State: Imp*lication *to Public Health***. Nigerian Annals of Pure and Applied Sciences.** 3(3a), Nov 15, 2020: 46-52.

Assegie T. A. & Nair P. S., "*Handwritten Digits Recognition with Decision Tree Classification: A Machine Learning Approach*," **International Journal of Electrical and Computer Engineering**, vol. 9, no. 5, 2019, p. 4446,

Assi K, Rahman S.M, Mansoor U & Ratrout N. *Predicting Crash Injury Severity with Machine Learning Algorithm Synergized with Clustering Technique: A Promising Protocol.* **International journal of environmental research and public healt**h. Aug 17 2020 (15):5497.

Audu A.A, Iyiola O.F, Popoola A.A, Adeleye B.M, Medayese S, Mosima C, & Blamah N. *The Application of Geographic Information System as an Intelligent System Towards Emergency Responses in Road Traffic Accident in Ibadan*. **Journal of Transport and Supply Chain Management**. 2021 Mar 4;15:17

Awad M, & Khanna R. *Efficient Learning Machines: Theories, Concepts, and Applications for Engineers and System Designers* **Springer Nature**, 2015, p. 268.

Azhar A, Ariff N.M, Bakar M.A & Roslan A. *Classification of Driver Injury Severity for Accidents Involving Heavy Vehicles with Decision Tree and Random Forest*. **Sustainability**. 14(7). 2022; 4101.

Bengio Y, Lodi A & Prouvost A. *Machine Learning for Combinatorial Optimization: A Methodological Tour d'horizon*. **European Journal of Operational Research**. 290(2) Apr 16, 2021; 405-21.

Bokaba T, Doorsamy W & Paul B.S. *Comparative Study of Machine Learning Classifiers for Modelling Road Traffic Accidents*. **Applied Sciences.** 12(2): Jan 2022; 828.

Chinnamgari S. K. *R Machine Learning Projects: Implement Supervised, Unsupervised, and Reinforcement Learning Techniques using R 3.5*. **Packt Publishing Ltd;** Jan 14, 2019.

Darwish A, Ezzat D & Hassanien A E. *An Optimized Model Based on Convolutional Neural Networks and Orthogonal Learning Particle Swarm Optimization Algorithm for Plant Diseases Diagnosis*. **Swarm and evolutionary computation.** Feb 1 2020; 52:100616.

De Felice F., Crocetti D, Parisi M, Maiuri V, Moscarelli E, Caiazzo R, Bulzonetti N, Musio D & Tombolini V. "*Decision Tree Algorithm in Locally Advanced Rectal Cancer: An Example of Over-Interpretation and Misuse of a Machine Learning Approach,*" **Journal of Cancer Research and Clinical Oncology**, vol. 146, no. 3, 2020 pp. 761–765

Dias D, Silva J.S & Bernardino A. *The Prediction of Road-Accident Risk through Data Mining: A Case Study from Setubal, Portugal*. **In Informatics MDPI** Vol. 10, No. 1, Jan 30 2023 p. 17.

Fareed U, Khadam U, Iqbal M.M & Iqbal M.J. *Road Accidents Investigation and Forecasting using Data Mining Techniques.* **KIET Journal of Computing and Information Sciences**. 6(1), 2023:28-49.

Figueira da Cruz A, Pitombo C.S & Larocca A.P. *Identification of Rules Induced through Decision Tree Algorithm for Detection of Traffic Accidents with Victims: A Study Case from Brazi*l. **Case Studies on Transport Policy**. 5(2)2017:200-7.

Fleischhauer V, Feldheiser A. & Zaunseder S. *Beat-to-Beat Blood Pressure Estimation by Photoplethysmography and Its Interpretation*. **Sensors.** 22(18),Sep 17 2022; :7037.

Gajowniczek K, Grzegorczyk I, Ząbkowski T, C Bajaj. *Weighted Random Forests to Improve Arrhythmia Classification*. **Electronics.** Jan 3 2020;9(1):99.

Gnjatović M, Košanin I, Maček N & Joksimović D. *Clustering of Road Traffic Accidents as a Gestalt Problem*. **Applied Sciences.** Apr 29 2022; 12(9):4543.

Golab A, Gooya E, Falou A & Cabon M. *A Multilayer Feed-Forward Neural Network (MLFNN) For the Resource-Constrained Project Scheduling Problem (RCPSP)*. **Decision Science Letters.** 11(4): 2022; 407-18..

Guo Y, TvHastie & Tibshirani R. *Regularized Linear Discriminant Analysis And Its Application In Microarrays*. **Biostatistics** 8(1): Jan 1 2007; 86-100.

Hagenauer J & Helbich M. *A Comparative Study of Machine Learning Classifiers for Modeling Travel Mode Choice*. **Expert Systems with Applications**, 78:273-82. Jul 15 2017.

Heinsfeld A.S, Franco A.R, Craddock R.C, Buchweitz A, & Meneguzzi F. *Identification of Autism Spectrum Disorder Using Deep Learning and the ABIDE Dataset*. **NeuroImage: Clinical**; 17:16-23 Jan 1 2018.

Hu X., Rudin C., & Seltzer M., "*Optimal Sparse Decision Trees,*" **in Advances in Neural Information Processing Systems**, 2019, pp. 7267– 7275

Hussain D., Al-Antari M. A., Al-Masni M. A., S.-M. Han & T.-S. Kim, "*Femur Segmentation in DXA Imaging using a Machine Learning Decision Tree,*" **Journal of X-ray Science and Technology**, vol. 26, no. 5, 2018, pp. 727–746.

Iwendi C, Bashir A.K, Peshkar A, Sujatha R, Chatterjee J.M, Pasupuleti S, Mishra R, Pillai S & Jo O. *COVID-19 Patient Health Prediction Using Boosted Random Forest Algorithm*. **Frontiers in Public Health**. Jul 3 2020; 8:357.

Jamali A.A, Ferdousi R, Razzaghi S, Li J, Safdari R & Ebrahimie E. *DrugMiner: Comparative Analysis of Machine Learning Algorithms for Prediction of Potential Druggable Proteins*. **Drug Discovery Today**. 21(5)2016:718-24

Johnson J.M & Khoshgoftaar T.M. *Survey on Deep Learning With Class Imbalance*. **Journal of Big Data**; 6(1). Dec 2019: 1-54.

Khattak A, Almujibah H, Elamary A & Matara C.M. *Interpretable Dynamic Ensemble Selection Approach for the Prediction of Road Traffic Injury Severity: A Case Study of Pakistan's National Highway N-5*. **Sustainability**. Sep 28 2022;14(19):12340.

Koley S, Sadhu A.K, Mitra P, Chakraborty B & Chakraborty C. *Delineation and Diagnosis of Brain Tumors from Post Contrast T1-weighted MR Images using Rough Granular Computing and Random Forest*. **Applied Soft Computing.** Apr 1 2016;41:453-65.

Kong J.S, Lee K.H, Kim O.H, Lee H.Y, Kang C.Y, Choi D, Kim S.C, Jeong H, Kang D.R & Sung T.E. *Machine Learning-Based Injury Severity Prediction of Level 1 Trauma Center Enrolled Patients Associated with Car-To-Car Crashes in Korea.* **Computers in Biology and Medicine**. Feb 1 2023;153:106393.

Konstantinov A.V & Utkin L.V. *Interpretable Machine Learning with an Ensemble of Gradient Boosting Machines*. **Knowledge-Based Systems**. Jun 21 2021; 222:106993.

Ma Z, Mei G & Cuomo S. *An Analytic Framework using Deep Learning for Prediction of Traffic Accident Injury Severity Based on Contributing Factors*. **Accident Analysis & Prevention**. Sep 1 2021;160:106322.

Mangalathu S, Hwang S.H, Choi E & Jeon J.S. *Rapid Seismic Damage Evaluation of Bridge Portfolios using Machine Learning Techniques.* **Engineering Structures.** Dec 15 2019; 201:109785.

Mashudi N.A, Ahmad N & Noor N.M. *Classification of Adult Autistic Spectrum Disorder using Machine Learning Approach*. **IAES International Journal of Artificial Intelligence** 10(3)**.** Sep 1 2021: 743

Mishra V, Agarwal S.M & N Puri. *Comprehensive and Comparative Analysis of Neural Network*. **International Journal of Computer Application**. 2(8), 2018;:126-37.

Mosavi A, Sajedi Hosseini F, Choubin B, Goodarzi M, Dineva A.A & Sardooi E R. *Ensemble Boosting And Bagging Based Machine Learning Models For Groundwater Potential Prediction*. **Water Resources Management**. Jan 2021, 35:23-37.

Nagra A. A., Han F, Ling Q H, Abubaker M, Ahmad F, Mehta S, Apasiba A T. "*Hybrid Self-Inertia Weight Adaptive Particle Swarm Optimisation with Local Search Using C4. 5 Decision Tree Classifier for Feature Selection Problems,*" **Connection Science,** vol. 32, no. 1, 2020, pp. 16–36.

Onyemaechi N O & Ofoma U R. *The Public Health Threat of Road Traffic Accidents in Nigeria: A Call to Action.* **Annals of Medical and Health Sciences Research.** 6(4) 2016; 199-204.

Paczkowski W.R. *Data Visualization: The Basics. Business Analytics*: **Data Science for Business Problems.** 2021:85-126.

Qi W, Su H, Yang C, Ferrigno G, De Momi E & Aliverti A. *A Fast and Robust Deep Convolutional Neural Networks for Complex Human Activity Recognition using Smartphone*. **Sensors**. 19(17)2019:3731.

Reis I, Baron D & Shahaf S. *Probabilistic Random Forest: A Machine Learning Algorithm for Noisy Data Sets*. **The Astronomical Journal.** 157(1)2018:16.

Ren Y, L Zhang & Suganthan P.N. *Ensemble Classification and Regression-Recent Developments, Applications and Future Directions.* **IEEE Computational Intelligence Magazine** 11(1). Jan 12 2016; 41-53.

Tangirala S . *Evaluating the Impact of GINI Index and Information Gain on Classification using Decision Tree Classifier Algorithm*. **International Journal of Advanced Computer Science and Applications**. 11(2):612-9. 2020;

Saenz-Cogollo J.F & Agelli M. *Investigating Feature Selection and Random Forests for Inter-Patient Heartbeat Classification*. **Algorithms.** Mar 25 2020; 13(4):75.

Sameen M.I & Pradhan B. *Severity Prediction of Traffic Accidents with Recurrent Neural Networks*. **Applied Sciences**. Jun 8 2017; 7(6):476.

Saracoglu A & Ozen H. *Estimation of Traffic Incident Duration: A Comparative Study of Decision Tree Models*. **Arabian Journal for Science and Engineering**. Oct 2020; 45(10):8099-110.

Sarker I. H., Colman A., Han J., Khan A. I., Abushark Y. B., & Salah K., "*Behavdt: A Behavioral Decision Tree Learning To Build User-centric Context-Aware Predictive Model*," **Mobile Networks and Applications**, vol. 25, no. 3, 2020, pp. 1151–1161

Sarker I.H. *Machine Learning: Algorithms, Real-World Applications and Research Directions*. **SN computer science.** (3):160, May 2 2021.

Shiran G, Imaninasab R & Khayamim R. *Crash Severity Analysis of Highways Based on Multinomial Logistic Regression Model, Decision Tree Techniques, and Artificial Neural Network: A Modeling Comparison*. **Sustainability**. May 18 2021;13(10):5670.

Shiroyama R, Wang M & Yoshimura C. *Effect of Sample Size on Habitat Suitability Estimation using Random Forests: A Case of Bluegill, Lepomis Macrochirus*. **In Annales de Limnologie-International Journal of Limnology** Vol. 56, p. 13. EDP Sciences, 2020.

Sibindi R, Mwangi R.W & Waititu A.G. *A Boosting Ensemble Learning Based Hybrid Light Gradient Boosting Machine and Extreme Gradient Boosting Model for Predicting House Prices*. **Engineering Reports**. 2022:e12599.

Sidey-Gibbons J.A & Sidey-Gibbons C.J. *Machine Learning in Medicine: A Practical Introduction*. **BMC MedicalR Methodology**. 19 Dec 2019: 1-8.

Suganya E & Rajan C. *An Adaboost-Modified Classifier Using Particle Swarm Optimization and Stochastic Diffusion Search in Wireless Iot Networks*. **Wireless Networks.** May 2021; 27:2287-99.

Sulaiman M.A. *Evaluating data mining classification methods performance in Internet of things applications*. **Journal of Soft Computing and Data Mining**. Dec 6 2020;1(2):11-25.

Taherkhani A, Cosma G & McGinnity T.M. *AdaBoost-CNN: An Adaptive Boosting Algorithm for Convolutional Neural Networks to Classify Multi-Class Imbalanced Datasets using Transfer Learning*. **Neurocomputing**. Sep 3 2020; 404:351-66.

Tsirikoglou P, Abraham S, Contino F, Lacor C & Ghorbaniasl G. *A Hyperparameters Selection Technique for Support Vector Regression Models.* **Applied Soft Computing**. 1 Dec 2017; 61:139-48.

Utkin L.V, Kovalev M.S & Coolen F.P. *Imprecise Weighted Extensions of Random Forests for Classification and Regression*. **Applied Soft Computing.** Jul 1 2020;92:106324.

Varshney K.R. *Trustworthy Machine Learning and Artificial Intelligence*. *XRDS: Crossroads*, **the ACM Magazine for Students**. 25(3):26-9, Apr 10, 2019.

Wang D & Zhu A X. *Soil Mapping Based on the Integration of the Similarity-Based Approach and Random Forests*. **Land**. May 29 2020;9(6):174.

Wang Y, Wang D, Geng N, Wang Y, Yin Y & Jin Y. *Stacking-Based Ensemble Learning of Decision Trees for Interpretable Prostate Cancer Detection*. **Applied Soft Computing.** Apr 1 2019;77:188-204.

Wen H, Zhang X & Zeng Q. *Predicting Future Driving Risk of Crash-Involved Drivers Based on a Systematic Machine Learning Framework*. **International Journal of Environmental Research and Public Health**. 16(3), 2019;334-52.

Yan M & Shen Y. *Traffic Accident Severity Prediction Based on Random Forest.* **Sustainability**. Feb 2 2022; 14(3):1729.

Yang X, Wang Y, Byrne R, Schneider G & Yang S. *Concepts of Artificial Intelligence for Computer-Assisted Drug Discovery*. **Chemical reviews**. Jul 11 2019; 119(18):10520-94.

Yeap D, McCartney M.M, Rajapakse M.Y, Fung A.G, Kenyon N.J & Davis C.E. *Peak Detection and Random Forests Classification Software for Gas Chromatography/Differential Mobility Spectrometry (GC/DMS) Data.* **Chemometrics and Intelligent Laboratory Systems**. Aug 15 2020;203:104085.

Yigin B.O, Algin O & Saygili G. *Comparison of Morphometric Parameters in Prediction of Hydrocephalus using Random Forests*. **Computers in Biology and Medicine.** Jan 1 2020;116:103547.

Zahid M, Chen Y, Khan S, Jamal A., Ijaz M & Ahmed T. *Predicting Risky and Aggressive Driving Behavior among Taxi Drivers: do Spatio-Temporal Attributes Matter?* **International Journal of Environmental Research and Public Health**. 17(11), Jun 2020: 3937.

Zebari R, Abdulazeez A, Zeebaree D, Zebari D & Saeed J. *A Comprehensive Review of Dimensionality Reduction Techniques for Feature Selection and Feature Extraction*. **J. Appl. Sci. Technol**. Trends. May 15 2020;1(2):56-70.

Zhai X, Chen M & Lu W. *Fuel Ratio Optimization of Blast Furnace Based on Data Mining.* **ISIJ International**. 60(11) Nov 15 2020: 2471-6.

Zhang C, Lei X & Liu L. *Predicting Metabolite–Disease Associations Based on LightGBM Model*. **Frontiers in Genetics**. Apr 13 2021;12:660275.

Zhang F, Yang X. *Improving Land Cover Classification in an Urbanized Coastal Area by Random Forests: The Role of Variable Selection*. **Remote Sensing of Environment.** Dec 15 2020;251:112105.

Zhang X, Qi S, Zheng A, Luo Y & Hao S. *Data-Driven Analysis of Fatal Urban Traffic Accident Characteristics and Safety Enhancement Research*. **Sustainability.** Feb 10 2023; 15(4):3259.

Zou Q., Qu K., Luo Y., Yin D., Ju Y., & Tang H., "*Predicting Diabetes Mellitus with Machine Learning Techniques*," **Frontiers in Genetics,** vol. 9, 2018 p. 515

Conference Proceedings

Ahmim A., Maglaras L., Ferrag M. A., Derdour M., & Janicke H., "*A Novel Hierarchical Intrusion Detection System Based on Decision Tree and Rules-Based Models*," I**n 2019 15th International Conference on Distributed Computing in Sensor Systems (DCOSS)**, 2019, pp. 228–233.

Alagarsamy S, Nagaraj P, Srikanth B, Krishna C.V, Bharath G & Kalyan S.S. *A Novel Machine Learning Technique for Predicting Road Accidents*. **In 2023 Third International Conference on Artificial Intelligence and Smart Energy (ICAIS)** Feb 2 2023 (pp. 1547-1551). IEEE.

Alnami H.M, Mahgoub I & Al–Najada H. *Highway Accident Severity Prediction for Optimal Resource Allocation of Emergency Vehicles and Personnel*. **In 2021 IEEE 11th Annual Computing and Communication Workshop and Conference (CCWC) IEEE** Jan 27 2021 (pp. 1231-1238).

Arowolo M. O., Adebiyi M., Adebiyi A., & Okesola O., "*PCA Model for RNA-Seq Malaria Vector Data Classification using KNN and Decision Tree Algorithm,*" **in 2020 International Conference in Mathematics, Computer Engineering and Computer Science (ICMCECS)**, Mar. 2020, pp. 1–8, doi: 10.1109/ICMCECS47690.2020.240881

Batitis V. M. E., Caballes M. J. G., A. A. Ciudad, M. D. Diaz, R. D. Flores, & E. R. E. Tolentin, "*Image Classification of Abnormal Red Blood Cells using Decision Tree Algorithm,*" **in 2020 Fourth International Conference on Computing Methodologies and Communication (ICCMC),** Mar. 2020, pp. 498–504, doi: 10.1109/ICCMC48092.2020.ICCMC-00093.

Chirag P, Supreetha M. *Road Accident Prediction and Classification using Machine Learning*. **In 2022 IEEE 2nd Mysore Sub Section International Conference (MysuruCon) IEEE.** Oct 16 2022 (pp. 1-8).

Demidova L & Ivkina M. *Defining the Ranges Boundaries of the Optimal Parameters Values for the Random Forest Classifier.* **In 2019 1st International Conference on Control Systems, Mathematical Modelling, Automation and Energy Efficiency (SUMMA) IEEE** Nov 20 2019 (pp. 518-522).

Denisko D & Hoffman M.M. *Classification and Interaction in Random Forests.* **Proceedings of the National Academy of Sciences.** Feb 20 2018;115(8):1690-2.

Elassad Z.E Abou, Mousannif H & Moatassime H Al. *A Real-Time Crash Prediction Fusion Framework: An Imbalance-Aware Strategy for Collision Avoidance Systems*. **Transportation Research Part C: Emerging Technologies.** Sep 1 2020; 118:102708.

Elyassami S, Hamid Y & Habuza T. *Road Crashes Analysis and Prediction using Gradient Boosted and Random Forest Trees*. **In 2020 6th IEEE Congress on Information Science and Technology (CiSt) IEEE** Jun 5 2021 (pp. 520-525).

Finogeev A, Deev M & Kolesnikoff I. *Proactive Big Data Analysis for Traffic Accident Prediction*. **In 2020 5th International Conference on Innovative Technologies in Intelligent Systems and Industrial Applications (CITISIA), IEEE,** Nov 25 2020, pp. 1-9.

Géron A. *Hands-on Machine Learning with Scikit-Learn, Keras, and TensorFlow*. "**O'Reilly Media, Inc**."; Oct 4 2022.

Ghosh S, Dasgupta A & Swetapadma A. *A Study On Support Vector Machine Based Linear And Non-Linear Pattern Classification*. **In 2019 International Conference on Intelligent Sustainable Systems (ICISS) IEEE** Feb 21 2019 (pp. 24-28).

Haynes S, Estin P.C, Lazarevski S, Soosay M & Kor A.L. *Data analytics: Factors of Traffic Accidents in the UK,* **In 2019 10th International Conference on Dependable Systems, Services and Technologies (DESSERT) IEEE**, Jun 5 2019, pp. 120-126.

Kim D.H DH, Ramjan LM & Mak KK. *Prediction of Vehicle Crashes by Drivers' Characteristics and Past Traffic Violations In Korea Using A Zero-Inflated Negative Binomial Model.* **Traffic Injury Prevention** 17(1). Jan 2 2016 86-90.

Kuang W., Chan Y., Tsang S., & Siu W., "*Machine Learning-Based Fast Intra Mode Decision for HEVC Screen Content Coding via Decision Trees,*" **IEEE Transactions on Circuits and Systems for Video Technology**, vol. 30, no. 5, May 2020, pp. 1481–1496, doi: 10.1109/TCSVT.2019.2903547.

Kumeda B, Zhang F, Zhou F, Hussain S, Almasri A & Assefa M. *Classification of Road Traffic Accident Data using Machine Learning Algorithms*. **In 2019 IEEE 11th International Conference on Communication Software and Networks (ICCSN) IEEE** Jun 12 2019 (pp. 682-687).

Kwok-Fai Lui A, Chan Y.H, Lo K.H, Cheng W.T & Cheung H.T. *Predictive Screening of Accident Black Spots based on Deep Neural Models of Road Networks and Facilities: A Case Study based on a District in Hong Kong.* **In 2021 5th International Conference on Computer Science and Artificial Intelligence,** Dec 4 2021 (pp. 422-428).

Linty N., Farasin A., Favenza A., & Dovis F., "*Detection of GNSS Ionospheric Scintillations Based on Machine Learning Decision Tree*," **IEEE Transactions on Aerospace and Electronic Systems,** vol. 55, no. 1, Feb. 2019 pp. 303–317, doi: 10.1109/TAES.2018.2850385.

Lubbe N, Jeppsson H, Ranjbar A, Fredriksson J, Bärgman J, & Östling M. *Predicted Road Traffic Fatalities in Germany: The Potential and Limitations of Vehicle Safety Technologies from Passive Safety to Highly Automated Driving*. **In Proceedings of IRCOBI conference. Athena, Greece,** Sep 2018

Malik S, Sayed H El, Khan M.A & Khan M.J. *Road Accident Severity Prediction—A Comparative Analysis of Machine Learning Algorithms.* **In 2021 IEEE Global Conference on Artificial Intelligence and Internet of Things IEEE. (GCAIoT)** Dec 12 2021 (pp. 69-74).

Manzoor M, Umer M, Sadiq S, Ishaq A, Ullah S, Madni H.A & Bisogni C. *RFCNN: Traffic Accident Severity Prediction Based On Decision Level Fusion of Machine and Deep Learning Model.* **IEEE Access**. Sep 14 2021; 9:128359-71.

Nandhini S. & J. M. K.S, "*Performance Evaluation of Machine Learning Algorithms for Email Spam Detection*," **in 2020 International Conference on Emerging Trends in Information Technology and Engineering (ic-ETITE)**, Feb. 2020, pp. 1–4, doi: 10.1109/ic-ETITE47903.2020.312.

Nandurge P.A & Dharwadkar N.V. *Analyzing Road Accident Data using Machine Learning Paradigms*. **In 2017 International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC) IEEE**, Feb 10, 2017 (pp. 604-610).

Park H & Haghani A. *Real-Time Prediction of Secondary Incident Occurrences using Vehicle Probe Data.* Tr**ansportation Research Part C: Emerging Technologies.** Sep 1 2016; 70:69-85.

Pathan S., Kumar P., Pai R., & Bhandary S. V., "*Automated Detection of Optic Disc Contours in Fundus Images using Decision Tree Classifier*," **Biocybernetics and Biomedical Engineering**, vol. 40, no. 1, 2020, pp. 52–64

Patil S. & Kulkarni U., "*Accuracy Prediction for Distributed Decision Tree using Machine Learning approach*," **in 2019 3rd International Conference on Trends in Electronics and Informatics (ICOEI)**, Apr. 2019, pp. 1365–1371, doi: 10.1109/ICOEI.2019.8862580

Ramadhan I., Sukarno P., & Nugroho M. A., "*Comparative Analysis of K-Nearest Neighbor and Decision Tree in Detecting Distributed Denial of Service,*" **in 2020 8th International Conference on Information and Communication Technology (ICoICT)**, Yogyakarta, Indonesia, Jun. 2020, pp. 1–4, doi: 10.1109/ICoICT49345.2020.9166380.

Reiss F, Cutler B & Eichenberger Z. *Natural Language Processing with Pandas Dataframes*. **InProc. Of The 20th Python In Science Conf.(Scipy 2021)** 2021, pp. 49-58.

Sathiyanarayanan P., Pavithra S., Saranya M. S., & Makeswari M., "*Identification of Breast Cancer using the Decision Tree Algorithm*," **In 2019 IEEE International Conference on System, Computation, Automation and Networking (ICSCAN),** 2019, pp. 1–6.

Shaik A.B & Srinivasan S. *A Brief Survey on Random Forest Ensembles in Classification Model*. **In International Conference on Innovative Computing and Communications: Proceedings of ICICC 2018, Springer Singapore** Volume 2 2019 (pp. 253-260).

Shen Z, Shehzad A, Chen S, Sun H & Liu J. *Machine Learning Based Approach on Food Recognition and Nutrition Estimation*. **Procedia Computer Science.** Jan 1 2020;174:448-53.

Sowdagur J.A, Rozbully-Sowdagur B.T, Suddul G. *An Artificial Neural Network Approach for Road Accident Severity Prediction*. **In 2022 IEEE Zooming Innovation in Consumer Technologies Conference (ZINC) IEEE.** May 25 2022 (pp. 267-270).

Sun Y, Li Y, Zeng Q & Bian Y. *Application Research of Text Classification Based on Random Forest Algorithm.* **2020 3rd International Conference on Advanced Electronic Materials, Computers and Software Engineering (AEMCSE),** 2020 370–374. https://doi.org/10.1109/AEMCSE50948.2020.00086

Taloba A. I. & Ismail S. I., "*An Intelligent Hybrid Technique of Decision Tree and Genetic Algorithm for E-Mail Spam Detection*," **in 2019 Ninth International Conference on Intelligent Computing and Information Systems (ICICIS),** Dec. 2019, pp. 99–104, doi: 10.1109/ICICIS46948.2019.9014756

Wang C, Liu L, Xu C & Lv W. *Predicting Future Driving Risk of Crash-Involved Drivers Based on a Systematic Machine Learning Framework*. **International Journal of Environmental Research And Public Health** 16(3): Feb 2019; 334.

Xia X L, Nan B. & Xu C. *Real-time traffic Accident Severity Prediction using Data Mining Technologies*. **In 2017 International Conference on Network and Information Systems for Computers (ICNISC)** Apr 14, 2017, pp. 242-245. IEEE.

Yaacob N.F, Rusli N & Bohari S.N. *A Review Analysis of Accident Factor on Road Accident Cases using Haddon Matrix Approach*. **In Proceedings of the Second International Conference on the Future of ASEAN (ICoFA) Springer Singapore 2017–Volume 2: Science and Technology** 2018, (pp. 55-65).

Zhang Y., Liu J., Zhang Z., & Huang J., "*Prediction of Daily Smoking Behavior Based on Decision Tree Machine Learning Algorithm*," **in 2019 IEEE 9th International Conference on Electronics Information and Emergency Communication (ICEIEC),** Jul. 2019, pp. 330–333, doi: 10.1109/ICEIEC.2019.8784698

Dissertations

Abiodun O.I, Jantan A, Dada K.V, Mohamed N.A & Arshad H. *State-of-theart in Artificial Neural Network Applications: A survey*. Heliyon 2018. 4 e00938. doi: 10.1016/j.heliyon. e00938.

Adnan M.S, Zaidi S & Bhargava P. *A Novel Support Vector Regression (SVR) Model for the Prediction of Splice Strength of the Unconfined Beam Specimens. Construction and Building Materials.* Jul 10 2020; 248:118475.

Afraei S, Shahriar K & Madani S.H. *Developing Intelligent Classification Models for Rock Burst Prediction after Recognizing Significant Predictor Variables, Section 2: Designing classifiers. Tunnelling and Underground Space Technology*. Feb 1 2019; 84:522-37.

Andeta J.A. *Road-traffic Accident Prediction Model: Predicting the Number of Casualties*. Master Degree Thesis in Informatics. ECTS Spring Term 2021

.Herda G & McNabb R. *Python for Smarter Cities: Comparison of Python Libraries for Static and Interactive Visualisations of Large Vector Data*. arXiv preprint arXiv:2202.13105. Feb 26 2022.

Klein L.A. "S*ensor and Data Fusion for Intelligent Transportation Systems." Society of Photo-Optical Instrumentation Engineers*, 2019.

Li M., Xu H., & Deng Y., "*Evidential Decision Tree Based on Belief Entropy,*" *Entropy*, vol. 21, no. 9, 2019, p. 897.

Sabzekar M & Hasheminejad S.M. *Robust Regression Using Support Vector Regressions. Chaos, Solitons & Fractals*. Mar 1 2021; 144:110738.

Saigo H, Nowozin S, Kadowaki T, Kudo T & Tsuda K. *gBoost: A Mathematical Programming Approach To Graph Classification And Regression*. Machine Learning. Apr 2009; 75:69-89.

Salaudeen A.G. *Risk Factors and Safety Measures for Road Traffic Crashes Among Inter-City Commercial Drivers in Kwara State, Nigeria*, Doctoral Dissertation, University Of Ilorin. 2018

Uzondu C, Jamson S & Lai F. *Exploratory study involving observation of traffic behaviour and conflicts in Nigeria using the Traffic Conflict Technique*. Safety science. 1;110, Dec 2018:273-84

Wang M, Jia S, Chen E, Yang S, Liu P & Qi Z. *A Derived Least Square Fast Learning Network Model. Applied Intelligence*. Dec 2020; 50:4176-94.

Zhang Y & Sung Y. *Hybrid Traffic Accident Classification Models*. Mathematics. Feb 19 2023; 11(4):1050.

Appendix I

```
import pandas as pd

import numpy as np

import matplotlib.pyplot as plt

import seaborn as sns

# %matplotlib inline


df = pd.read_csv('RTA Dataset.csv')

df.head()

df.shape

# print the dataset information

df.info()


df.isnull().sum()/100


df['Accident_severity'].value_counts().plot(kind='bar')


import matplotlib.pyplot as plt


# Plot the bar chart

ax = df['Accident_severity'].value_counts().plot(kind='bar', color=['#FF6347', '#4169E1', '#32CD32'])
```

```
# Customize the plot
ax.set_xlabel('Accident Severity')
ax.set_ylabel('Count')
ax.set_title('Distribution of Accident Severity')
ax.legend(['Minor', 'Moderate', 'Severe'])
ax.grid(axis='y', linestyle='--')

# Save the plot
plt.savefig('accident_severity_plot.png')
plt.show()

"""This shows imbalance multiclass label on the dataset"""

# plot the bar plot of road_surface_type and accident severity feature
plt.figure(figsize=(6,5))
sns.countplot(x='Road_surface_type', hue='Accident_severity', data=df)
plt.xlabel('Rode surafce type')
plt.xticks(rotation=60)
plt.savefig('accident_severity_plot.png')
plt.show()

# convert object type column into datetime datatype column
df['Time'] = pd.to_datetime(df['Time'])

# Extrating 'Hour_of_Day' feature from the Time column
new_df = df.copy()
new_df['Hour_of_Day'] = new_df['Time'].dt.hour
```

```
df_new = new_df.drop('Time', axis=1)
df_new.head()


def fill_missing_values(df):
    # Loop over each column in the dataframe
    for col in df.columns:
        if df[col].dtype == 'float64' or df[col].dtype == 'int64': # Check if column is numeric
            # Fill missing values with mean
            df[col].fillna(df[col].mean(), inplace=True)
        else:
            # Fill missing values with mode
            df[col].fillna(df[col].mode()[0], inplace=True)
    return df

# Fill missing values using the function
df_new = fill_missing_values(df_new)

df_new.isnull().sum()

from sklearn.preprocessing import LabelEncoder

def label_encode_features(df):
    le = LabelEncoder() # create a label encoder object

    for col in df.columns:
        if df[col].dtype == 'object': # check if column is of type 'object'
            df[col] = le.fit_transform(df[col].astype(str)) # label encode the column
```

```
    return df

# Label encode the object-type features using the function
new_df = label_encode_features(new_df)

new_df.head()

df_new.columns

#handling imbalance multiclass
X = new_df.drop(['Accident_severity', 'Time'], axis=1)
y = new_df['Accident_severity']

X

!pip install imblearn

from sklearn.model_selection import train_test_split, cross_val_score
from sklearn.preprocessing import LabelEncoder, StandardScaler
from sklearn.pipeline import Pipeline
from imblearn.pipeline import make_pipeline
from imblearn.over_sampling import SMOTE
from sklearn.ensemble import RandomForestClassifier, GradientBoostingClassifier,VotingClassifier
from sklearn.metrics import classification_report, confusion_matrix, accuracy_score

le = LabelEncoder()
```

```
y = le.fit_transform(y)

sc = StandardScaler()

X = sc.fit_transform(X)

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=42)

smote = SMOTE(random_state=42)

X_train_res, y_train_res = smote.fit_resample(X_train, y_train)


# modelling using random forest baseline

rf = RandomForestClassifier(n_estimators=800, max_depth=20, random_state=42)


rf.fit(X_train_res, y_train_res)


# predicting on test data

predics = rf.predict(X_test)


cm = confusion_matrix(y_test, predics)

ConfusionMatrixDisplay(cm).plot()


# classification report on test dataset

classif_re = classification_report(y_test,predics)

print(classif_re)


from sklearn import tree

from sklearn.tree import DecisionTreeClassifier

from sklearn import metrics

from sklearn.metrics import confusion_matrix

from sklearn.metrics import classification_report
```

```
from sklearn.metrics import ConfusionMatrixDisplay

decisionTree = DecisionTreeClassifier(criterion='entropy')
print(decisionTree)

dtc_model = decisionTree.fit(X_train_res, y_train_res)

from matplotlib import pyplot

# feature importance

importance = dtc_model.feature_importances_
for i,v in enumerate(importance):
 print('Feature: %0d, Score: %.5f' % (i,v))

# Barchat for feature importance

pyplot.bar([x for x in range(len(importance))], importance)
pyplot.show()

prediction = dtc_model.predict(X_test)

cm = confusion_matrix(y_test, prediction)
ConfusionMatrixDisplay(cm).plot()
print(classification_report(y_test, prediction))
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.ensemble import RandomForestClassifier
```

```
from sklearn.metrics import accuracy_score, classification_report, confusion_matrix

# Load the dataset (replace 'your_dataset.csv' with your actual dataset file)
data = pd.read_csv('your_dataset.csv')

# Feature columns (replace 'feature1', 'feature2', etc. with the actual feature column names)
features = data[['feature1', 'feature2', 'feature3', ...]]

# Target column (replace 'target' with the actual column containing the severity codes)
target = data['target']

# Split the data into training and testing sets (adjust the test_size as needed)
X_train, X_test, y_train, y_test = train_test_split(features, target, test_size=0.25, random_state=42)

# Initialize the Random Forest Classifier
rf_classifier = RandomForestClassifier()

# Train the model
rf_classifier.fit(X_train, y_train)

# Make predictions on the test set
y_pred = rf_classifier.predict(X_test)

# Calculate accuracy
accuracy = accuracy_score(y_test, y_pred)
print(f"Accuracy: {accuracy}")
# Generate classification report and confusion matrix
```

```
print("\nClassification Report:")
print(classification_report(y_test, y_pred))
print("\nConfusion Matrix:")
print(confusion_matrix(y_test, y_pred))

import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
# %matplotlib inline

df = pd.read_csv('RTA Dataset.csv')
df.head()
df.shape
# print the dataset information
df.info()

df.isnull().sum()/100

df['Accident_severity'].value_counts().plot(kind='bar')

import matplotlib.pyplot as plt

# Plot the bar chart
ax = df['Accident_severity'].value_counts().plot(kind='bar', color=['#FF6347', '#4169E1', '#32CD32'])

# Customize the plot
```

```
ax.set_xlabel('Accident Severity')
ax.set_ylabel('Count')
ax.set_title('Distribution of Accident Severity')
ax.legend(['Minor', 'Moderate', 'Severe'])
ax.grid(axis='y', linestyle='--')

# Save the plot
plt.savefig('accident_severity_plot.png')
plt.show()

"""This shows imbalance multiclass label on the dataset"""

# plot the bar plot of road_surface_type and accident severity feature
plt.figure(figsize=(6,5))
sns.countplot(x='Road_surface_type', hue='Accident_severity', data=df)
plt.xlabel('Rode surafce type')
plt.xticks(rotation=60)
plt.savefig('accident_severity_plot.png')
plt.show()

# convert object type column into datetime datatype column
df['Time'] = pd.to_datetime(df['Time'])

# Extrating 'Hour_of_Day' feature from the Time column
new_df = df.copy()
new_df['Hour_of_Day'] = new_df['Time'].dt.hour
df_new = new_df.drop('Time', axis=1)
df_new.head()
```

```
def fill_missing_values(df):
    # Loop over each column in the dataframe
    for col in df.columns:
        if df[col].dtype == 'float64' or df[col].dtype == 'int64': # Check if column is numeric
            # Fill missing values with mean
            df[col].fillna(df[col].mean(), inplace=True)
        else:
            # Fill missing values with mode
            df[col].fillna(df[col].mode()[0], inplace=True)
    return df

# Fill missing values using the function
df_new = fill_missing_values(df_new)

df_new.isnull().sum()

from sklearn.preprocessing import LabelEncoder

def label_encode_features(df):
    le = LabelEncoder() # create a label encoder object

    for col in df.columns:
        if df[col].dtype == 'object': # check if column is of type 'object'
            df[col] = le.fit_transform(df[col].astype(str)) # label encode the column

    return df
```

```
# Label encode the object-type features using the function
new_df = label_encode_features(new_df)

new_df.head()

df_new.columns

#handling imbalance multiclass
X = new_df.drop(['Accident_severity', 'Time'], axis=1)
y = new_df['Accident_severity']

X

!pip install imblearn

from sklearn.model_selection import train_test_split, cross_val_score
from sklearn.preprocessing import LabelEncoder, StandardScaler
from sklearn.pipeline import Pipeline
from imblearn.pipeline import make_pipeline
from imblearn.over_sampling import SMOTE
from sklearn.ensemble import RandomForestClassifier, GradientBoostingClassifier,VotingClassifier
from sklearn.metrics import classification_report, confusion_matrix, accuracy_score

le = LabelEncoder()
y = le.fit_transform(y)
sc = StandardScaler()
```

```
X = sc.fit_transform(X)

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=42)

smote = SMOTE(random_state=42)

X_train_res, y_train_res = smote.fit_resample(X_train, y_train)


# modelling using random forest baseline

rf = RandomForestClassifier(n_estimators=800, max_depth=20, random_state=42)


rf.fit(X_train_res, y_train_res)


# predicting on test data

predics = rf.predict(X_test)


cm = confusion_matrix(y_test, predics)

ConfusionMatrixDisplay(cm).plot()


# classification report on test dataset

classif_re = classification_report(y_test,predics)

print(classif_re)


from sklearn import tree

from sklearn.tree import DecisionTreeClassifier

from sklearn import metrics

from sklearn.metrics import confusion_matrix

from sklearn.metrics import classification_report

from sklearn.metrics import ConfusionMatrixDisplay
```

```
decisionTree = DecisionTreeClassifier(criterion='entropy')
print(decisionTree)

dtc_model = decisionTree.fit(X_train_res, y_train_res)

from matplotlib import pyplot

# feature importance

importance = dtc_model.feature_importances_
for i,v in enumerate(importance):
 print('Feature: %0d, Score: %.5f' % (i,v))

# Barchat for feature importance

pyplot.bar([x for x in range(len(importance))], importance)
pyplot.show()

prediction = dtc_model.predict(X_test)

cm = confusion_matrix(y_test, prediction)
ConfusionMatrixDisplay(cm).plot()
print(classification_report(y_test, prediction))
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import accuracy_score, classification_report, confusion_matrix
```

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
# %matplotlib inline

df = pd.read_csv('RTA Dataset.csv')
df.head()
df.shape
# print the dataset information
df.info()

df.isnull().sum()/100

df['Accident_severity'].value_counts().plot(kind='bar')

import matplotlib.pyplot as plt

# Plot the bar chart
ax = df['Accident_severity'].value_counts().plot(kind='bar', color=['#FF6347', '#4169E1', '#32CD32'])

# Customize the plot
ax.set_xlabel('Accident Severity')
ax.set_ylabel('Count')
ax.set_title('Distribution of Accident Severity')
ax.legend(['Minor', 'Moderate', 'Severe'])
ax.grid(axis='y', linestyle='--')
```

```
# Save the plot
plt.savefig('accident_severity_plot.png')
plt.show()

"""This shows imbalance multiclass label on the dataset"""

# plot the bar plot of road_surface_type and accident severity feature
plt.figure(figsize=(6,5))
sns.countplot(x='Road_surface_type', hue='Accident_severity', data=df)
plt.xlabel('Rode surafce type')
plt.xticks(rotation=60)
plt.savefig('accident_severity_plot.png')
plt.show()

# convert object type column into datetime datatype column
df['Time'] = pd.to_datetime(df['Time'])

# Extrating 'Hour_of_Day' feature from the Time column
new_df = df.copy()
new_df['Hour_of_Day'] = new_df['Time'].dt.hour
df_new = new_df.drop('Time', axis=1)
df_new.head()

def fill_missing_values(df):
    # Loop over each column in the dataframe
    for col in df.columns:
```

```
        if df[col].dtype == 'float64' or df[col].dtype == 'int64': # Check if column is numeric
            # Fill missing values with mean
            df[col].fillna(df[col].mean(), inplace=True)
        else:
            # Fill missing values with mode
            df[col].fillna(df[col].mode()[0], inplace=True)
    return df

# Fill missing values using the function
df_new = fill_missing_values(df_new)

df_new.isnull().sum()

from sklearn.preprocessing import LabelEncoder

def label_encode_features(df):
    le = LabelEncoder() # create a label encoder object

    for col in df.columns:
        if df[col].dtype == 'object': # check if column is of type 'object'
            df[col] = le.fit_transform(df[col].astype(str)) # label encode the column

    return df

# Label encode the object-type features using the function
new_df = label_encode_features(new_df)
```

```
new_df.head()


df_new.columns


#handling imbalance multiclass
X = new_df.drop(['Accident_severity', 'Time'], axis=1)
y = new_df['Accident_severity']


X


!pip install imblearn


from sklearn.model_selection import train_test_split, cross_val_score
from sklearn.preprocessing import LabelEncoder, StandardScaler
from sklearn.pipeline import Pipeline
from imblearn.pipeline import make_pipeline
from imblearn.over_sampling import SMOTE
from sklearn.ensemble import RandomForestClassifier, GradientBoostingClassifier,VotingClassifier
from sklearn.metrics import classification_report, confusion_matrix, accuracy_score


le = LabelEncoder()
y = le.fit_transform(y)
sc = StandardScaler()
X = sc.fit_transform(X)
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=42)
smote = SMOTE(random_state=42)
X_train_res, y_train_res = smote.fit_resample(X_train, y_train)
```

```
# modelling using random forest baseline
rf = RandomForestClassifier(n_estimators=800, max_depth=20, random_state=42)

rf.fit(X_train_res, y_train_res)

# predicting on test data
predics = rf.predict(X_test)

cm = confusion_matrix(y_test, predics)
ConfusionMatrixDisplay(cm).plot()

# classification report on test dataset
classif_re = classification_report(y_test,predics)
print(classif_re)

from sklearn import tree
from sklearn.tree import DecisionTreeClassifier
from sklearn import metrics
from sklearn.metrics import confusion_matrix
from sklearn.metrics import classification_report
from sklearn.metrics import ConfusionMatrixDisplay

decisionTree = DecisionTreeClassifier(criterion='entropy')
print(decisionTree)

dtc_model = decisionTree.fit(X_train_res, y_train_res)
```

```
from matplotlib import pyplot

# feature importance

importance = dtc_model.feature_importances_
for i,v in enumerate(importance):
 print('Feature: %0d, Score: %.5f' % (i,v))

# Barchat for feature importance

pyplot.bar([x for x in range(len(importance))], importance)
pyplot.show()

prediction = dtc_model.predict(X_test)

cm = confusion_matrix(y_test, prediction)
ConfusionMatrixDisplay(cm).plot()
print(classification_report(y_test, prediction))
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import accuracy_score, classification_report, confusion_matrix

# Load the dataset (replace 'your_dataset.csv' with your actual dataset file)
data = pd.read_csv('your_dataset.csv')

# Feature columns (replace 'feature1', 'feature2', etc. with the actual feature column
names)
```

```
features = data[['feature1', 'feature2', 'feature3', ...]]

# Target column (replace 'target' with the actual column containing the severity codes)
target = data['target']

# Split the data into training and testing sets (adjust the test_size as needed)
X_train, X_test, y_train, y_test = train_test_split(features, target, test_size=0.25,
random_state=42)

# Initialize the Random Forest Classifier
rf_classifier = RandomForestClassifier()

# Train the model
rf_classifier.fit(X_train, y_train)

# Make predictions on the test set
y_pred = rf_classifier.predict(X_test)

# Calculate accuracy
accuracy = accuracy_score(y_test, y_pred)
print(f"Accuracy: {accuracy}")
# Generate classification report and confusion matrix
print("\nClassification Report:")
print(classification_report(y_test, y_pred))
print("\nConfusion Matrix:")
print(confusion_matrix(y_test, y_pred))

import pandas as pd
```

```
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
# %matplotlib inline

df = pd.read_csv('RTA Dataset.csv')
df.head()
df.shape
# print the dataset information
df.info()

df.isnull().sum()/100

df['Accident_severity'].value_counts().plot(kind='bar')

import matplotlib.pyplot as plt

# Plot the bar chart
ax = df['Accident_severity'].value_counts().plot(kind='bar', color=['#FF6347', '#4169E1', '#32CD32'])

# Customize the plot
ax.set_xlabel('Accident Severity')
ax.set_ylabel('Count')
ax.set_title('Distribution of Accident Severity')
ax.legend(['Minor', 'Moderate', 'Severe'])
ax.grid(axis='y', linestyle='--')
```

```
# Save the plot
plt.savefig('accident_severity_plot.png')
plt.show()

"""This shows imbalance multiclass label on the dataset"""

# plot the bar plot of road_surface_type and accident severity feature
plt.figure(figsize=(6,5))
sns.countplot(x='Road_surface_type', hue='Accident_severity', data=df)
plt.xlabel('Rode surafce type')
plt.xticks(rotation=60)
plt.savefig('accident_severity_plot.png')
plt.show()

# convert object type column into datetime datatype column
df['Time'] = pd.to_datetime(df['Time'])

# Extrating 'Hour_of_Day' feature from the Time column
new_df = df.copy()
new_df['Hour_of_Day'] = new_df['Time'].dt.hour
df_new = new_df.drop('Time', axis=1)
df_new.head()

def fill_missing_values(df):
    # Loop over each column in the dataframe
    for col in df.columns:
        if df[col].dtype == 'float64' or df[col].dtype == 'int64': # Check if column is numeric
```

```
            # Fill missing values with mean
            df[col].fillna(df[col].mean(), inplace=True)
        else:
            # Fill missing values with mode
            df[col].fillna(df[col].mode()[0], inplace=True)
    return df

# Fill missing values using the function
df_new = fill_missing_values(df_new)

df_new.isnull().sum()

from sklearn.preprocessing import LabelEncoder

def label_encode_features(df):
    le = LabelEncoder() # create a label encoder object

    for col in df.columns:
        if df[col].dtype == 'object': # check if column is of type 'object'
            df[col] = le.fit_transform(df[col].astype(str)) # label encode the column

    return df

# Label encode the object-type features using the function
new_df = label_encode_features(new_df)

new_df.head()
```

```
df_new.columns

#handling imbalance multiclass
X = new_df.drop(['Accident_severity', 'Time'], axis=1)
y = new_df['Accident_severity']

X

!pip install imblearn

from sklearn.model_selection import train_test_split, cross_val_score
from sklearn.preprocessing import LabelEncoder, StandardScaler
from sklearn.pipeline import Pipeline
from imblearn.pipeline import make_pipeline
from imblearn.over_sampling import SMOTE
from sklearn.ensemble import RandomForestClassifier, GradientBoostingClassifier,VotingClassifier
from sklearn.metrics import classification_report, confusion_matrix, accuracy_score

le = LabelEncoder()
y = le.fit_transform(y)
sc = StandardScaler()
X = sc.fit_transform(X)
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=42)
smote = SMOTE(random_state=42)
X_train_res, y_train_res = smote.fit_resample(X_train, y_train)

# modelling using random forest baseline
```

```
rf = RandomForestClassifier(n_estimators=800, max_depth=20, random_state=42)

rf.fit(X_train_res, y_train_res)

# predicting on test data
predics = rf.predict(X_test)

cm = confusion_matrix(y_test, predics)
ConfusionMatrixDisplay(cm).plot()

# classification report on test dataset
classif_re = classification_report(y_test,predics)
print(classif_re)

from sklearn import tree
from sklearn.tree import DecisionTreeClassifier
from sklearn import metrics
from sklearn.metrics import confusion_matrix
from sklearn.metrics import classification_report
from sklearn.metrics import ConfusionMatrixDisplay

decisionTree = DecisionTreeClassifier(criterion='entropy')
print(decisionTree)

dtc_model = decisionTree.fit(X_train_res, y_train_res)

from matplotlib import pyplot
```

```
# feature importance

importance = dtc_model.feature_importances_
for i,v in enumerate(importance):
 print('Feature: %0d, Score: %.5f' % (i,v))

# Barchat for feature importance

pyplot.bar([x for x in range(len(importance))], importance)
pyplot.show()

prediction = dtc_model.predict(X_test)

cm = confusion_matrix(y_test, prediction)
ConfusionMatrixDisplay(cm).plot()
print(classification_report(y_test, prediction))
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import accuracy_score, classification_report, confusion_matrix

# Extrating 'Hour_of_Day' feature from the Time column
new_df = df.copy()
new_df['Hour_of_Day'] = new_df['Time'].dt.hour
df_new = new_df.drop('Time', axis=1)
df_new.head()

def fill_missing_values(df):
```

```
    # Loop over each column in the dataframe
    for col in df.columns:
        if df[col].dtype == 'float64' or df[col].dtype == 'int64': # Check if column is numeric
            # Fill missing values with mean
            df[col].fillna(df[col].mean(), inplace=True)
        else:
            # Fill missing values with mode
            df[col].fillna(df[col].mode()[0], inplace=True)
    return df

# Fill missing values using the function
df_new = fill_missing_values(df_new)

df_new.isnull().sum()

from sklearn.preprocessing import LabelEncoder

def label_encode_features(df):
    le = LabelEncoder() # create a label encoder object

    for col in df.columns:
        if df[col].dtype == 'object': # check if column is of type 'object'
            df[col] = le.fit_transform(df[col].astype(str)) # label encode the column

    return df

# Label encode the object-type features using the function
```

```
new_df = label_encode_features(new_df)

new_df.head()

df_new.columns

#handling imbalance multiclass
X = new_df.drop(['Accident_severity', 'Time'], axis=1)
y = new_df['Accident_severity']

X

!pip install imblearn

from sklearn.model_selection import train_test_split, cross_val_score
from sklearn.preprocessing import LabelEncoder, StandardScaler
from sklearn.pipeline import Pipeline
from imblearn.pipeline import make_pipeline
from imblearn.over_sampling import SMOTE
from sklearn.ensemble import RandomForestClassifier, GradientBoostingClassifier,VotingClassifier
from sklearn.metrics import classification_report, confusion_matrix, accuracy_score

le = LabelEncoder()
y = le.fit_transform(y)
sc = StandardScaler()
X = sc.fit_transform(X)
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=42)
```

```
smote = SMOTE(random_state=42)
X_train_res, y_train_res = smote.fit_resample(X_train, y_train)

# modelling using random forest baseline
rf = RandomForestClassifier(n_estimators=800, max_depth=20, random_state=42)

rf.fit(X_train_res, y_train_res)

# predicting on test data
predics = rf.predict(X_test)

cm = confusion_matrix(y_test, predics)
ConfusionMatrixDisplay(cm).plot()

# classification report on test dataset
classif_re = classification_report(y_test,predics)
print(classif_re)

from sklearn import tree
from sklearn.tree import DecisionTreeClassifier
from sklearn import metrics
from sklearn.metrics import confusion_matrix
from sklearn.metrics import classification_report
from sklearn.metrics import ConfusionMatrixDisplay

decisionTree = DecisionTreeClassifier(criterion='entropy')
print(decisionTree)
```

```
dtc_model = decisionTree.fit(X_train_res, y_train_res)

from matplotlib import pyplot

# feature importance

importance = dtc_model.feature_importances_
for i,v in enumerate(importance):
 print('Feature: %0d, Score: %.5f' % (i,v))

# Barchat for feature importance

pyplot.bar([x for x in range(len(importance))], importance)
pyplot.show()

prediction = dtc_model.predict(X_test)

cm = confusion_matrix(y_test, prediction)
ConfusionMatrixDisplay(cm).plot()
print(classification_report(y_test, prediction))
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import accuracy_score, classification_report, confusion_matrix

import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
```

```
import seaborn as sns
# %matplotlib inline

df = pd.read_csv('RTA Dataset.csv')
df.head()
df.shape
# print the dataset information
df.info()

df.isnull().sum()/100

df['Accident_severity'].value_counts().plot(kind='bar')

import matplotlib.pyplot as plt

# Plot the bar chart
ax = df['Accident_severity'].value_counts().plot(kind='bar', color=['#FF6347', '#4169E1', '#32CD32'])
```

Biodata

A. Personal Data

1. **Full Name:** Sofoluwe Segun Abayomi
2. **Date and Place of Birth:** 15th September 1998.
3. **Nationality:** Nigerian
4. **Marital Status:** Single
5. **No. of Children & their ages:** Nil
6. **Name and Address of Spouse:** Nil
7. **Name and Address of Next of Kin:** Bola Sofoluwe
8. **Faculty:** Natural and Applied Sciences
9. **Department:** Computer Science

B. Educational Background

Educational Institutions Attended with Dates and Qualification:

i.	Gloryville Primary School	Primary School Leaving Certificate	2010
ii.	Kolmor Metropolitan College	WAEC Certificate	2015
iii.	Elizade University	BSc. Computer Science	2019

C. Work Experience: With Dates

Elizade Motors

2017

Elizade University ICT center 2018

Oyo State Basic Educational Board 2020

Sesako Investment, Lagos State

2022

D. Awards and Fellowship Nil

E. Membership of Academic/Professional Bodies Nil

F. Publications - Predicting the Severity of Vehicle Accidents Based on Traffic Accident Attributes Using Machine Learning**.**

G.

H. Attended with Dates - Google Digital Skills for Africa

2020

I. Names and Addresses of References

Dr. A.A. Waheed
Senior Lecturer
Lead City University, Ibadan
Department of Mathematics
Waheed.azeez@lcu.edu.ng

Dr. Kehinde Agbele
HOD Computer Science
Elizade University
Kehinde.agbele@elizadeuniversity.edu.ng

Dr. R. O Kolapo
Lecturer I
Computer Science Department
Lead City University, Ibadan
Kolapo.ridwan@lcu.edu.ng

J. Date & Signature:

…… ……..

Signature Date

The University Compliance Certification

This is to certify that this thesis by Segun Abayomi SOFOLUWE with Matriculation Number LCU/PG/002221in the Department of Computer Science, Faculty of Natural and Applied Sciences, Lead City University, Ibadan is in full compliance with the approval of the University's format and style.

…… ……..

Signature Date