Analysis of the SUV Involved Pedestrian Crashes in Pennsylvania

Youssra Aaiad

Submitted to the
Institute of Graduate Studies and Research
in partial fulfillment of the requirements for the degree of

Master of Science
in
Civil Engineering

Eastern Mediterranean University
February 2024
Gazimağusa, North Cyprus

Approval of the Institute of Graduate Studies and Research

Prof. Dr. Ali Hakan Ulusoy
Director

I certify that this thesis satisfies all the requirements as a thesis for the degree of Master of Science in Civil Engineering.

Assoc. Prof. Dr. Eriş Uygar
Chair, Department of Civil Engineering

We certify that we have read this thesis and that in our opinion it is fully adequate in scope and quality as a thesis for the degree of Master of Science in Civil Engineering.

Assoc. Prof. Dr. Mehmet Metin Kunt
Supervisor

Examining Committee

1. Prof. Dr. Mustafa Ergil

2. Assoc. Prof. Dr. Mehmet Metin Kunt

3. Asst. Prof. Dr. Hüseyin Sevay

ABSTRACT

This thesis addresses the pivotal challenge of predicting the SUV involved pedestrian crash severity and proposes improvements to existing methodologies, underscoring the substantial threat posed by such incidents. Utilizing a comprehensive dataset spanning five years from the state of Pennsylvania, USA, the study acknowledges and addresses the challenge of class imbalance through the application of the Synthetic Minority Oversampling Technique (SMOTE) for data augmentation. Methodologically, diverse artificial neural network (ANN) architectures are explored, with meticulous evaluation through K-fold cross-validation to ensure the robustness of the model.

Descriptive statistics and correlation analyses are employed to investigate crash characteristics and inter-variable relationships. The outcomes underscore the efficacy of SMOTE in improving predictive accuracy. Beyond its primary predictive contributions, this research offers nuanced insights into factors impacting model efficacy. By addressing prevailing limitations and introducing an innovative approach to handling class imbalances, our research informs the development of interventions to enhance road safety. The findings carry crucial implications for policy and practice, with the ultimate goal of reducing pedestrian accidents and mitigating their severity.

Keywords: pedestrian crash severity, road safety, artificial neural networks, SMOTE.

ÖZ

Bu tez, SUV kaynaklı yaya kazası ciddiyetini tahmin etme konusundaki temel zorluğu ele almakta ve bu tür olayların oluşturduğu önemli tehdidin altını çizerek mevcut metodolojilerde iyileştirmeler önermektedir. USA’nin Pennsylvania eyaletinden beş yılı kapsayan kapsamlı bir veri setini kullanan çalışma, veri artırma için Sentetik Azınlık Aşırı Örnekleme Tekniğinin (SMOTE) uygulanması yoluyla sınıf dengesizliği sorununu kabul ediyor ve ele alıyor. Metodolojik olarak, modelin sağlamlığını sağlamak için K-fold çapraz doğrulaması yoluyla titiz bir değerlendirme yapılarak çeşitli yapay sinir ağı (YSA) mimarileri araştırılmaktadır.

Tanımlayıcı istatistikler ve korelasyon analizleri, çarpışma özelliklerini ve değişkenler arası ilişkileri araştırmak için kullanılır. Sonuçlar, SMOTE'un tahmin doğruluğunu artırmadaki etkinliğini vurgulamaktadır. Bu araştırma, birincil öngörü katkılarının ötesinde, model etkinliğini etkileyen faktörlere ilişkin incelikli bilgiler sunmaktadır. Araştırmamız, mevcut kısıtlamaları ele alarak ve sınıf dengesizliğini ortadan kaldırmak için yenilikçi bir yaklaşım sunarak, karayolu güvenliğini artırmaya yönelik müdahalelerin geliştirilmesine bilgi sağlamaktadır. Bulgular, yaya kazalarının azaltılması ve ciddiyetinin hafifletilmesi nihai hedefiyle politika ve uygulama açısından önemli çıkarımlar taşıyor.

Anahtar Kelimeler: yaya çarpma şiddeti, yol güvenliği, yapay sinir ağları, SMOTE.

DEDICATION

To my parents, whom unwavering support and love have been my foundation throughout this academic journey.

To my family, for their constant encouragement and understanding, enabling me to pursue and achieve my goals.

To my classmates, for sharing the challenges and triumphs of this educational voyage and creating lasting memories and friendships.

To my professors, your guidance and expertise have shaped my intellectual growth, and I am grateful for your dedication to education.

ACKNOWLEDGEMENT

I want to express my sincere gratitude to my professors for their steadfast guidance and scholarly insights, which have played a pivotal role in shaping my academic journey. Their commitment to education has been a driving force in my intellectual growth.

A special thanks goes to my supervisor, Assoc. Prof. Dr. Metin M. Kunt, for their invaluable mentorship and thoughtful contributions, which were essential in shaping the direction of my research. I appreciate their dedication to fostering academic excellence and maintaining high research standards.

I am also thankful for the support and encouragement from my peers and colleagues. Their collaboration has added depth to this academic experience, and I am grateful for the shared challenges and successes during this educational adventure.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

INTRODUCTION

1.1 Background and significance of the study

According to the Fatality Analysis Reporting System (FARS) of U.S. Department of Transportation, traffic crashes are the leading cause of death for people aged 4 to 7 and 16 to 20. In 2021, motor vehicle traffic collisions were the biggest cause of death for ages 4 through 8, 17 through 19, and 21 [1]. There are numerous factors causing this high rate of accident fatalities. Many traffic crashes are caused by reckless driving behaviors such as speeding, distracted driving, and driving under the influence of drugs or alcohol. Globally, vehicle ownership has increased substantially throughout the years. This increase contributes to traffic congestion, particularly in densely populated areas or during peak commute hours resulting in an increase in traffic crashes. According to the National Highway Traffic Safety Administration (NHTSA), there were 6516 pedestrians killed in traffic crashes in 2020, highest since 1990 and a 3.9% rise over 2019. In 2020, a pedestrian died every 81 minutes, accounting for 17% of all traffic fatalities on America's roadways.

Pedestrians, considered the most vulnerable road users (VRUs), face this vulnerability due to their lack of protection compared to motorized vehicles. The kinetic forces produced by various types of vehicles can have a considerable impact on the severity of pedestrian injuries when traffic crashes occur. It is important to understand that the severity of a car accident is determined by factors other than the physical forces involved. Other factors that influence the outcome of a collision include the location of the impact, the use of safety measures, and the design of the vehicles. The type of vehicle is a key component that will be explored throughout this research. Larger and heavier vehicles, such as trucks or SUVs, generate larger kinetic forces due to their mass, which significantly influences their momentum. The design of the vehicle's front end also influences the severity of pedestrian injuries. Higher bumpers or hoods are more likely to contact pedestrians at a higher position on their bodies, resulting in more severe head or upper body injuries [2]. As a result, identifying the characteristics of such accidents and working towards reducing casualties are deemed critical. Studies on collisions involving Vulnerable Road Users (VRUs) play a significant role in traffic crash research, addressing important aspects of road safety.

Previous studies have addressed various aspects of this matter. Different analytical methods were employed for investigating the factors influencing the pedestrian crashes, which include environmental variables, behavioral characteristics, and road design. Lee and Abdel-Aty (2005) studied vehicle-pedestrian collisions in Florida, and found that pedestrian and driver characteristics, vehicle size, and environmental conditions all contribute to severe injuries. Kim et al. (2008) investigated single-vehicle single-pedestrian collisions in North Carolina, finding that parameters such as the age of the pedestrians, male drivers, two-way roads, overspeed, dark-lighted conditions, and commercial areas increase the probability of fatal pedestrian-involved crash. Abdul-Aziz et al. (2013) studied pedestrian-vehicle crashes in New York, finding that roadway features, traffic attributes, and land-use features contribute to severe injuries. Research has shown that factors such as pedestrians over 65 years old, not wearing contrasting clothing, adult drivers, the summer season, time of day, multilane highways, darkness, and collisions with pickup trucks increase pedestrian injury severity. Vehicle type, drivers under the influence of alcohol and elderly pedestrians, dark lighting conditions, the presence of intersections without traffic lights or the absence of pedestrian crossings, and sport utility vehicles (SUVs) and vans contribute significantly to injury severity in urban areas. These studies highlight the unique factors that contribute to the severity of pedestrian injuries.

Crash prediction tools are critical techniques in transportation safety because they allow us to analyze historical crash data and identify factors associated with increased crash risk. Crash prediction models can range from simple regression models to more complex models using machine learning. These models usually include a variety of variables, such as driver characteristics, roadway features, and environmental factors. By examining these variables in relation to crash outcomes, researchers can find patterns and correlations that might not be immediately clear. Transportation planners and politicians may develop strategies and measures to lower the chance of crashes and improve overall safety on our roads by recognizing these risk variables.

Most traffic safety programs have aimed at reducing the frequency of pedestrian-vehicle collisions, but only a few of these programs have focused specifically on reducing the risk posed to pedestrians by large and heavier vehicles. In this research, the focus is on investigating the impact of sport utility vehicles (SUVs) on pedestrians. The study aims to analyze both environmental factors and human behaviors to understand their influence on the severity of pedestrian crashes. The research will employ a machine-learning model to analyze relevant data, allowing for a comprehensive examination of the topic. The study intends to provide insights into the factors influencing SUV-related pedestrian crashes and contribute to the development of effective strategies for prevention and mitigation.

With new vehicle designs and characteristics, it is necessary to investigate their potential impact on pedestrian crash severity. The effectiveness and limitations of these new designs in preventing severe pedestrian crashes can be assessed through research, as well as potential challenges and opportunities for their implementation. Besides that, using complex modeling and simulation techniques, researchers can better understand the dynamics of pedestrian crashes and predict the severity of such crashes. The development of advanced models that incorporate various variables, such as the interaction between vehicles and pedestrians, human behaviour, and environmental factors, can provide an understanding of crash severity patterns and potentially facilitate effective safety interventions.

1.2 Research objectives

This study aims to investigate pedestrians' crash severity with SUVs. The analysis was conducted in a certain geographical area and over a specified period. Crash data from three counties in Pennsylvania provided by the PennDOT Portal was used [3]. An artificial neural network model is employed to investigate the factors influencing the severity of pedestrian injuries. Crash severity levels are separated in order to determine which one is most frequently caused by SUVs. A prediction model was also developed to predict the severity of pedestrian injuries.

This knowledge is essential to providing evidence-based recommendations to politicians, including both politicians and technical experts, by investigating the factors that contribute to these crashes and their consequences and implementing stricter safety rules, enforcement measures, and public awareness campaigns. This research has the potential to influence vehicle design, driver training, pedestrian infrastructure, and resource allocation policies to address pedestrian safety problems.

1.3 Research methodology

The fundamental goal of this research was to create an ANN-based model for traffic crash analysis with an emphasis on accurate prediction of crash injury severity as well as the identification of significant contributing elements. A complete dataset containing numerous criteria, such as road conditions, road characteristics, driver behaviors, and vehicle types, was collected from the PennDOT portal. This dataset was fed to the ANN model, which was subjected to training in order to learn the complicated patterns and correlations between these variables and crash occurrences. The program would precisely estimate the possibility of crashes in various settings by employing machine learning techniques. Furthermore, the ANN model would identify the important contributing elements that have a substantial impact on the occurrence of crashes.

1.3.1 Data collection

Many sources offer valuable details on crash locations, categories, causes, and repercussions, assisting in the development of evidence-based policies and initiatives. One of the most important sources is data from government organizations. The information is gained through police department reports, insurance company databases, and hospital records. These resources include detailed information about injuries, treatment, and healthcare costs. Details regarding the severity, treatment, and public health consequences are recorded in hospital records. Insurance claims contain information about crash causes, car damage, and injury complaints, but they may not cover all occurrences because of private settlements or the failure of policyholders to file claims. Another source is a fully accessible database; most of them are developed with government funding and are freely accessible to the public. Access to these databases is possible via specialized websites or web services that provide search and discovery interfaces.

Many countries have established centralized databases to collect and store road crash data. These databases contain information from police reports, insurance claims, and other sources and provide data on crash characteristics, weather, and road hazards. Moreover, national or regional databases specifically dedicated to recording road crash fatalities are vital for understanding traffic crash mortality rates, identifying high-risk groups, and evaluating road safety interventions.

Intelligent Transportation Systems (ITS) are currently one of the most essential sources. Massive volumes of data are collected by GPS information via on-board electronic equipment (OBEs). Closed-circuit television cameras (CCTV) are another widespread technology that captures real-time information on traffic conditions and incidents.

1.3.2 Data preparation

Historical crash data was collected from the Pennsylvania department of transportation that covered the period 2017-2021. Preprocessing is the initial stage of preparing data for analysis. The goal of preprocessing is to remove any inconsistencies, missing values, or outliers from the data. This also includes changing the data into an analysis-ready format, such as normalizing or scaling the data. The next critical step in preparing the data for analysis is data normalization. It contributes to the elimination of any biases that may occur as a result of the various scales of distinct features. We ensure that the selected features contribute equally to the model's training process by scaling them, resulting in more accurate and dependable predictions.

The final stage in data preparation is to divide the preprocessed data into training, validation, and testing sets so that the performance of the ANN model developed can be evaluated. This is an important phase in machine learning since it allows us to determine how effectively the model generalizes to new data. The training set is used to train the model, the validation set is used to tune hyperparameters and pick the best model, and the testing set is used for evaluating the model's final performance.

1.3.3 Model development

Data-driven methodologies have gained popularity in crash analysis due to their advantages over traditional models. These methods do not necessitate apriori parametric presumptions, which may or may not be required in real-world settings. They can discover relationships and trends based on data without making predetermined assumptions, and they can handle complex and non-linear interactions between factors that standard statistical models may have difficulty with. For smaller datasets, statistical techniques are preferred. Traditional statistical methods do well at predicting parameters and evaluating hypotheses, allowing for a higher level of trust in drawing connections between variables[4].

One complaint levelled at machine learning techniques is the difficulty in determining causality. Machine learning models frequently deal with a huge number of characteristics or variables, and predicted accuracy is prioritized over interpretability. While these models can find predicted associations due to the complexity of some machine learning algorithms, determining the specific causal elements leading to the predictions can be complicated [5].

Machine Learning (ML) is a subfield of Artificial Intelligence (AI) that focuses on the creation of algorithms and models that can learn from data and make predictions or perform actions without being manually programmed. The goal of ML is to allow computers to learn and develop from experience in the same way that humans learn and improve over time. These algorithms evaluate and identify patterns and relationships in massive volumes of data. ML algorithms can generate predictions, spot patterns, classify or cluster data, and uncover insights that are not immediately evident to people by processing and learning from this data.

Artificial Neural Networks (ANN):

Artificial neural networks (ANNs) are advanced systems that can process information in an accelerated and decentralized fashion [6]. They can calculate the relationship between dependent and independent variables and estimate nonlinear models. These networks operate using error backpropagation with the purpose of minimizing the variance between the output of the network and the desired output. The goal of continuously altering the model's parameters is to get a low error and increase the model's accuracy. An artificial neural network is a computer framework that is roughly based on biological neural networks [7]. It is composed of neurons that communicate with each other through weighted connections. Each neuron is represented by a real variable, and the connections between neurons are quantified by a parameter called weight. There are three layers in the network: the input layer, the hidden layer, and the output layer, as shown in Figure 1.

The input layer is given crash-related features or variables, such as weather, roadway type, vehicle characteristics, and so on. Each characteristic is represented by its own input node. The feature values are transmitted from the input nodes to the hidden layers, where the actual calculation takes place. Hidden layers, which comprise several nodes, conduct computations on the input data, using weights and biases to generate final results. These final outcomes are then processed by an activation function, which adds nonlinearity to the network and assists in the capture of complicated connections. The output layer represents the final crash severity predictions or classifications. Each node in the output layer represents a different level of severity, such as minor, moderate, or major. The type of problem at hand determines the activation function employed in the output layer. The output values reflect the model's estimated probabilities according to each severity level.

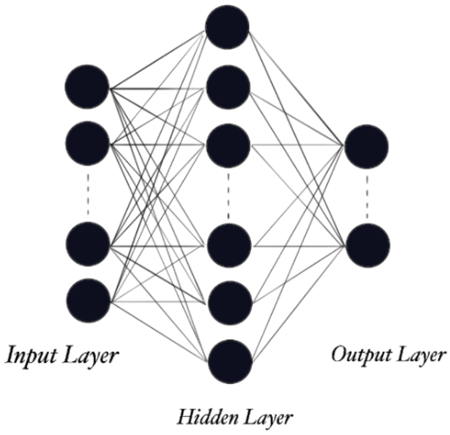

Figure 1: Structure of a typical ANN

The following two operations: weighted sum and activation function, are used in calculating a layer's output (Fig. 2).

[1] Each neuron in a layer receives input from the preceding layer, which is then multiplied by weights. The weighted sum is computed by adding the products of the input values and their weights. Mathematically, for a neuron in layer k, the weighted sum (y_k) is calculated in equation (1) as:

$$y_k = (w_{1k} * x_1) + (w_{2k} * x_2) + \ldots + (w_{nk} * x_n) + \theta \quad (1)$$

$$\text{Or,} \quad y_k = \sum_j w_{jk} x_j + \theta_k \text{ (for layer k)} \quad (2)$$

The weights define the degree of significance of each input in the overall algorithm. In order to reduce the error between expected and actual results, they are learned and modified throughout the training phase.

[2] Following the calculation of the weighted sum, an activation function is used to implement non-linearity and decide the neuron's output. The weighted total is fed into the activation function, which provides the final output value (Eq. (3)). This process increases the neural network's adaptability, allowing it to simulate complicated correlations and detect non-linear tendencies in data.

$$z_{\ k} = f_k(\sum w_{jk}\, x + \theta_j) \tag{3}$$

Depending on the analysis objectives, various activation methods can be used. The sigmoid function, ReLU, and Softmax are examples of frequently utilized activation functions [8].

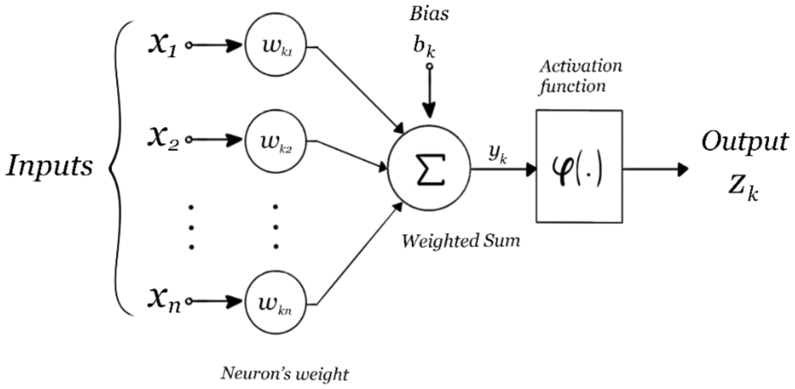

Figure 2: Structure of a neuron cell

Multilayer perceptron (MLP) neural networks:

Multilayer Perceptron networks are a prominent type of neural network that is used for a variety of applications, such as classification, regression, and recognition of patterns. MLP models are based on the mathematical development of Rosenblatt's perceptron theory from the 1950s [9]. They fall under the category of feedforward algorithms because the inputs are merged with the initial weights in a summation function, and applied to the activation function. Each layer feeds the results of its computation to the next, and this continues through the hidden layers into the output layer.

Back propagation is a method of learning that enables the MLP to systematically alter the network's weights in order to minimize the cost function. Back propagation must meet one strict criterion in order to function effectively. The function that mixes inputs and weights, the weighted sum, and the threshold function must be differentiable [10]. The following are typical steps in the training process: forward propagation, cost function calculation, backward propagation, and weights and biases update. Once the weighted sums are sent down all layers in every loop, the gradient of the Mean Squared Error (MSE) is calculated for all input and output units. The weights of the initial hidden layer are then modified with the gradient value in order to transmit it back. That represents the way the weights are transported back to the neural network's beginning. This process is repeated until a convergence criterion for each unit is reached. The number of hidden layers, the learning rate, the activation function, and the regularization parameters are all hyperparameters that must be adjusted before training. Making the right choice of these hyperparameters can have a considerable impact on network performance.

Review on SMOTE with ANN model:

Several studies have showed the efficiency of integrating SMOTE into ANN models. These studies have consistently demonstrated improvements in model performance in terms of accuracy, memory, F1-score, and AUC-ROC, which will be demonstrated in this research. Islam, Z. et al. (2021) used and compared three techniques: variational autoencoder (VAE), SMOTE, and ADASYN. Crash prediction models based on logical regression, support vector machines, and artificial neural networks were used to compare the generated data of different oversampling techniques.

The Synthetic Minority Oversampling TEchnique was introduced by Chawla et al. (2002) as a method for oversampling minority classes by producing synthetic data along the boundary portions linking related examples. This method alleviates class imbalance by increasing the diversity of the training data and balancing the class distribution. Class imbalance is a prevalent problem in many machine learning applications, in which one class vastly outweighs the others. This imbalance may cause erroneous predictions and a poor model performance. SMOTE, a popular resampling technique, generates computer-generated data that equalizes the class distribution and enhances the model's capacity to spot minority class tendencies.

Many studies have investigated the use of SMOTE and ANN models in conjunction to address class imbalances [11]. The structure of the ANN, extent of complexity of the dataset, and evaluation measures employed in these experiments differ. Common findings include increased minority class recognition, improved generalization performance, and reduced overfitting. Elamrani Abou Elassad, Z., Mousannif, H., & Al Moatassime, H. (2020) have created two refined models for crash prediction using the well-known machine learning techniques of Support Vector Machine (SVM) and neural network Multilayer Perceptron (MLP). To manage the imbalanced datasets, SMOTE was used to balance the training sets. Abou Elassad, Z. E., Mousannif, H., & Al Moatassime, H. (2020) employed resampling-based programs, including Bayesian learners (BL), k-nearest neighbors (kNN), support vector machines (SVM), and multilayer perceptron (MLP) to introduce diversity among models. To ensure that the proposed framework provides accurate and stable decisions, an imbalanced learning method using synthetic minority oversampling techniques (SMOTE) was adopted to address the problem of class imbalance, as collisions usually occur in rare cases.

Proper selection of SMOTE parameters (e.g., synthetic sample numbers and balance ratios) is essential to avoid overfitting or underfitting. Synthetic samples introduced by SMOTE may include noise that may affect model performance. Careful data pre-processing and adjustment are necessary. Synthetic sample generation can increase computational complexity, especially for larger datasets.

1.3.4 Activation functions

Activation functions are critical in defining the output of a neuron because they evaluate whether the neuron needs to be active depending on the weighted sum of its inputs. They introduce crucial non-linearity into the model to capture complex patterns. As a result, we must use an activation function that will render the network dynamic and give it the capacity to extract sophisticated and complex insights from data. Here are some of the commonly used AFs in ANN:

- *Sigmoid*

The sigmoid function converts the [-∞ ; +∞] input range to an S-shaped curve between 0 and 1. Sigmoid functions are commonly employed in binary classifications because they compress each unit's output into a range of 0 to 1. However, they do face the vanishing gradient problem, which makes the network extremely difficult to optimize. The Sigmoid activation function (Eq. (4)) is non-linear by nature with a smooth derivative, as shown in figure 3.

$$\sigma(x) = \frac{1}{1+e^{-x}} \tag{4}$$

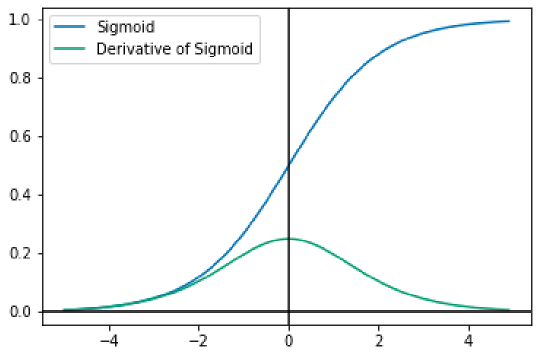

Figure 3: Plot of the Sigmoid activation function

- *Hyperbolic Tangent (TanH)*

The Hyperbolic Tangent function has a similar structure to the Sigmoid function, except it transforms the input into a curve between -1 and 1. This function is often used in hidden ANN layers. It solves the sigmoid function's zero-centered issue, but it remains afflicted by the vanishing gradient problem. This function is defined in Eq. (5):

$$\tanh(x) = \frac{2}{1+e^{-2x}} - 1 \tag{5}$$

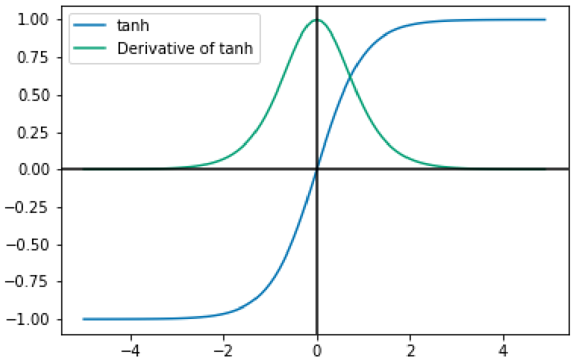

Figure 4: Plot of the TanH activation function

- *Rectified Linear Unit (ReLU)*

In machine learning, ReLU is one of the most widely utilized activation functions [12]. ReLU resets all negative values to zero while leaving positive values unaffected. ReLU's popularity is based on its higher training performance when compared to other activation functions such as the logistic and the hyperbolic tangent. This activation function is highly computationally effective and contributes to resolving the vanishing gradient problem [13]. The challenge of the vanishing gradient emerges when the gradients in the backpropagation process become extremely small, causing the weights to stop updating effectively. ReLU helps by maintaining a derivative of 1 for positive inputs, preventing gradient from vanishing. However, for negative inputs, the gradient is 0, leading to dead neurons. To address this issue, variants like Leaky ReLU and Parametric ReLU have been introduced [14]. The following is the definition of this function (Eq. (6)):

$$ReLU(x) = \max(0, x) \tag{6}$$

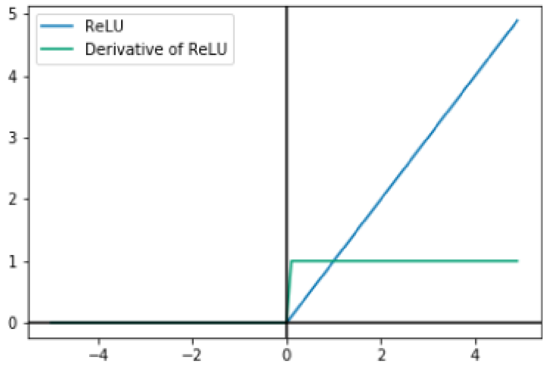

Figure 5: Plot of the ReLU activation function

- *Softmax*

The Softmax function is often employed in the output layer of a multi-class classification ANN. It consists of several sigmoid functions. It takes a vector of real values as input and turns it into a probability distribution with the total of all probabilities equal to one. The function returns the probability for each data point that can be expressed as follows (Eq. (7)):

$$\sigma(\vec{z})_i = \frac{e^{z_i}}{\sum_{j=1}^{k} e^{z_j}} \qquad (7)$$

Where,

$\vec{z}$ = input vector

e^{z_i} = standard exponential function for input vector

k = number of subclasses in the multi-class predictor

e^{z_j} = standard exponential function for output vector

1.3.5 Model Evaluation and Validation

The purpose of building a prediction model is to accurately predict data that has never been observed before. The approach to model training in machine learning entails exposing the model to a training dataset in order to learn the basic patterns and correlations between the independent features and the target variable. Once properly trained, the model should be able to make accurate predictions on new data that has an identical distribution as the training dataset. To evaluate a model's performance, several crucial evaluation measures are frequently employed. One of the key metrics is accuracy, which evaluates the proportion of accurately identified cases compared to the total number of cases. However, accuracy alone is not enough to evaluate the model's performance, particularly in unbalanced data sets where one class dominates the others. Other metrics, like precision, recall, and F1 score are used to overcome the constraints of accuracy. Precision expresses the model's ability to prevent false positives by measuring the fraction of real positive predictions over the total projected positives. Recall computes the proportion of true positives among all positive cases, emphasizing the model's capacity to recognize all positive situations. The F1 score includes precision and recall, resulting in a comprehensive metric that takes both false positives and false negatives into account. Evaluation criteria such as macro-average precision, recall, F1 score, cross-entropy loss, and confusion matrix are utilized for multi-class classification. The macro-average computes metrics separately for every class before calculating the average. The confusion matrix, on the other hand, is a tabular summary of what the model predicts in comparison to the actual classes.

- *Confusion matrix*

A confusion matrix is frequently used to assess the effectiveness of classification models, which are designed to predict a categorical class for each input. It is a two-dimensional matrix in which the rows represent the actual values and the columns represent the predicted values, along with the sum of the predictions. As illustrated in Figure 6, the matrix depicts the number of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) that resulted from the model.

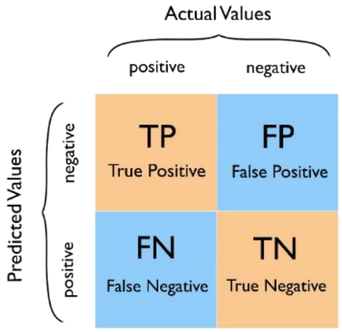

Figure 6: Representation of the confusion matrix

- *Accuracy*

Accuracy is a regularly used metric for assessing classification models, and it reflects the proportion of successfully categorized cases (including TP and TN) compared to the total number of cases in the data. It is presented as a percentage, and higher accuracy often represents improved model performance. It is computed as follows:

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

Note: TP: True Positive; TN: True Negative; FP: False Positive; FN: False Negative.

- *Precision*

Precision is the fraction of true positive predictions divided by the entire number of positive predictions made by the model (which includes both TP and FP predictions). It evaluates a model's ability to correctly detect positive occurrences. It is calculated as follows:

$$Precision = \frac{TP}{TP + FP}$$

- *Recall*

Recall, also known as sensitivity, is an important evaluation statistic used in both binary and multi-class classification problems. The percentage of true positive predictions divided by the total number of actual cases in the class (which includes both TP and FN predictions) is calculated as recall. A high recall score shows that the model has a small percentage of false negatives, implying that it can recognize positive examples effectively.

$$Recall = \frac{TP}{TP + FN}$$

- *F1 score*

The F1 score is beneficial in situations where both precision and recall are important and must be addressed simultaneously. This is especially critical in imbalanced data, in which the model can reach high accuracy by picking the majority class while doing poorly on the minority class. The F1 score goes from 0 to 1, with 1 indicating exact precision and recall and 0 indicating unsatisfactory in precision, recall, or both. It is measured as follows:

$$F1\,score = 2 \times \frac{Recall\ \times Precision}{Recall + Precision}$$

1.4 Contributions

The central research question that this study aims to address is: How do SUVs influence the severity of pedestrian crashes, and what are the significant contributing factors to such crash severity?

This study provides insights into the patterns and correlations between SUV-involved pedestrian crashes and their severity by analyzing crash data from Pennsylvania over five years. Understanding these factors may assist in the development of targeted interventions and policies aimed at lowering the frequency and severity of crashes. The creation of a prediction model based on an artificial neural network is an advanced approach for predicting crash occurrences. Machine learning techniques allow for a more accurate estimation of the likelihood of crashes under various conditions, thereby assisting in proactive crash prevention strategies.

This research has focused on identifying important factors that influence crash occurrence and have a practical impact. By identifying the most influential factors, transportation authorities can prioritize interventions, better allocate resources to address these elements, and mitigate their negative effects. The establishment of ANN-based crash analysis models not only provides valuable tools for this specific research but also prepares the foundations for future research. This approach can apply to different regions and periods and contribute to a broader understanding of the dynamics and factors of crashes in different contexts.

Chapter 2

LITERATURE REVIEW

2.1 Overview of traffic crash analysis

In recent years, the analysis and forecasting of traffic crashes have become major concerns for researchers. As a result, it is increasingly interested in developing accurate methods for predicting road crashes, with the goal that authorities and policymakers can undertake specific strategies and preventive actions to reduce the occurrence and severity of crashes. In order to address this issue, academics have been investigating numerous methodologies and techniques for predicting traffic crashes. Spatial analysis, machine learning, hybrid models, and big data analytics are all important tools used to improve crash forecasting. Spatial analysis uses Geographic Information Systems (GIS) and spatial statistical techniques to identify high-risk areas or crash hotspots. In machine learning algorithms, techniques such as use decision trees, random forests, support vector machines, and artificial neural networks are used to analyze crash data and develop predictive models. Big data analytics techniques such as data mining, machine learning, and natural language processing can uncover valuable insights and improve the accuracy of crash predictions.

Traffic crash analysis is the process of examining and studying traffic crashes to gain insights into their causes, contributing factors, and consequences. The analysis includes collecting and analyzing various types of crash data, including information about vehicles, road conditions, environmental factors, and driver behavior. The results of traffic crash analysis often include determining the underlying causes of crashes, identifying contributing factors, assessing countermeasures, and developing preventive measures. Data are collected from various sources, including police reports, witness reports, crash reconstruction studies, medical records, and traffic surveillance systems. To properly analyze and interpret data, statistical methods, data visualization techniques, and machine modeling are used. The ultimate objective is to reduce crashes, injuries, and deaths and to develop safer transportation infrastructure for everyone.

2.2 History and evolution of traffic crash analysis

In the early days of transportation, there was not much formal knowledge about traffic crashes and their causes. Crash investigations relied heavily on informal proof, witness statements, and police reports. Rather than addressing broader trends or cumulative concerns, the emphasis was on determining the immediate cause of an event.

In the 1980s, road safety audits were launched as a proactive approach to crash analysis. The purpose of these audits was to identify possible dangers and make the required adjustments. Crash simulation approaches improved as technology advanced. To reenact incidents, crash investigators began employing computer simulations, mathematical models, and forensic investigations. This enabled investigators to examine the order of events, vehicle dynamics, and factors influencing crash severity.

Traffic crash analysis has evolved as our understanding of transportation safety and data collection methods has improved. In the early 20th century, statistical analysis played an important role in identifying traffic crashes. Researchers began collecting data on crashes, injuries, and fatalities in order to discover trends and patterns. Basic statistical methods were used to examine crash data and calculate crash rates. The availability of large-scale data and progress in data analytics have changed crash analysis in recent years. Many countries have created extensive databases to collect detailed information on traffic crashes. Furthermore, with the development of technology, crash analysis techniques have become more sophisticated. For example, machine learning and data mining use data to detect trends and risk factors and develop predictive models for crash prevention.

2.3 Studies related to traffic crash analysis

2.3.1 Key factors in traffic crash analysis

- Human behaviors

Understanding how human behavior interacts with the driving environment could be beneficial in determining crash causes and developing effective preventative strategies. Driver distraction is one of the most crucial variables to consider. Drivers are more likely to be distracted while driving by their phones or other devices. This can result in reduced reaction times and a higher risk of collisions. Fatigue and drowsiness can also contribute to crashes involving tired drivers who struggle to stay attentive and focused on the road. Another significant cause is impaired driving, which includes driving while under the influence of drugs or alcohol. This can seriously impair a driver's ability to make informed decisions and respond promptly in an emergency.

Another aspect to be considered is the driver characteristics. Age is an important factor because older drivers may have decreased visual acuity, reaction times, and cognitive processing speed. They may also encounter difficulties with divided attention and navigating unfamiliar roads. Younger, inexperienced drivers, on the other hand, are more likely to engage in risky behaviors, such as speeding or distracted driving. Another critical factor is driver experience, as new drivers, particularly teenagers, have little on-road experience and are more likely to make mistakes. However, as drivers gain experience, they improve their ability to recognize and manage potential risks. It should be noted that passivity and overconfidence in experienced drivers could also lead to risky behavior.

In a study conducted by Tay, R., et al. (2011), the findings revealed male drivers were more likely than female drivers to be involved in catastrophic and fatal crashes. When compared to younger drivers, older drivers were less likely to be involved in major and fatal crashes. Alcohol-impaired drivers were more likely to be involved in deadly crashes. The characteristics of pedestrians, such as age, sex, and location, are important factors in the severity of pedestrian-vehicle crashes. Pedestrians over the age of 65 were more likely to be seriously injured and fatally harmed. Rashid, H. M. S., & Ismail, K. H. (2022) have identified human error as a significant factor contributing to the severity of injuries in these crashes. Single people and children were found to be more involved in car crashes, potentially because of distractions and risk-taking behaviors. Using cell phones while driving has been identified as a prominent cause of crashes among young drivers.

- Vehicle characteristics

Vehicle characteristics have a significant impact on the severity of crashes and the likelihood of injuries or fatalities. The size and weight of the vehicle, its speed and acceleration capabilities, braking performance, and overall structural integrity are some of the key vehicle characteristics that are frequently considered in crash analysis. Other factors to consider include safety features such as airbags, seat belts, and antilock brakes, as well as the vehicle's age and maintenance history. It is found that higher speeds are associated with more severe crashes, longer stopping distances, and reduced control, ultimately amplifying the severity of injuries sustained by vehicle occupants [15]. The literature also emphasizes the crucial role of adequate visibility in safe driving. It explains that the design and condition of headlights, taillights, turn signals, and reflective materials on vehicles play a key role in visibility, especially in low-light conditions or inclement weather.

Understanding the characteristics of different types of vehicles is also vital for crash analysis. Each vehicle type possesses unique traits that affect its maneuverability, stability, visibility, and vulnerability in crashes.

Passenger cars are the most common type of vehicle on the road. They are deemed stable and easy to maneuver. However, they are prone to rollovers in specific circumstances, particularly when taking sharp turns at high speeds. This characteristic can make them more susceptible to crashes in certain situations. Trucks and buses, on the other hand, are larger and heavier than passenger cars. This size and weight make maneuvering and stopping quickly more challenging for their drivers compared to passenger cars. Furthermore, trucks and buses have wider blind areas, making it more difficult for drivers to see other vehicles or people in their path. By considering these factors, crash investigators can gain a more complete understanding of what caused a specific crash and how it could have been avoided. Ulfarsson, G. F., & Mannering, F. L. (2004) found that in single-vehicle collisions, pickup and SUV drivers had a higher percentage of severe injuries than passenger car drivers, according to the data. In a study conducted by Oikawa, S., et al. (2016), the findings revealed that the risks of serious injury and death rose with increasing vehicle travel speeds for both sedans and light passenger cars. The study also evaluated the chances of serious pedestrian injuries with sedans and discovered that the risks of serious injury were higher in the middle-aged group compared to the early-aged group.

- Road conditions

Road conditions have a substantial effect on the frequency and severity of crashes. Weather conditions such as rain, mist, or ice can make roadways dangerous and disable visibility, expanding the danger of an crash. These conditions might lead to higher stopping distances and make it troublesome for drivers to preserve viable control of their cars. Furthermore, the road surface's condition might impact vehicle handling. Potholes, uneven surfaces, loose gravel, or debris on the road can cause vehicles to lose traction or exhibit unexpected changes in behavior, potentially resulting in an crash.

Road signs and indicators that are clear and visible are critical for guiding drivers and alerting them to possible dangers. Faded or absent signs, poor lane lines, or insufficient signaling can lead to confusion or inaccurate lane changes, all of which can lead to an crash. Another component that can increase the chance of an crash is heavy traffic congestion. In congested traffic, impatience and dangerous actions such as tailgating, rapid lane changes, or running red lights can occur, leading to crashes.

This study by Cheng, W., et al. (2023) found that collisions on wet-skid surfaces are more likely to result in severe injuries compared to collisions on dry surfaces. Switching from a dry road surface to a wet and slippery one increases the probability of a fatal injury by 3.38%. They also discussed the impact of light conditions on the severity of crashes. The study found that crashes that occurred at night without sufficient lighting were more likely to result in severe injuries compared to crashes that happened during the day. The article suggests several measures to address these issues. Lee, D., et al. (2023) found that ice-covered streets increase the probability of severe crashes, likely due to decreased maneuverability and compromised visibility. Bad weather, including rain and snow, can worsen these conditions and contribute to crashes that are more serious.

- Road geometry

Road geometry can have a significant impact on the behavior of drivers, car handling, and overall safety. The physical aspects of the road, such as its layout, alignment, and curvature, are referred to. Poor alignment, such as unexpected changes in direction or sudden curves, can catch drivers off guard and increase the risk of an crash. The design of intersections is critical to traffic safety. Crashes can be affected by factors like the type of intersection and the availability of designated turning lanes. Lane size influences driver actions and vehicle interactions. Tight lanes can make it difficult for drivers, especially heavy cars, to manage, increasing the possibility of an crash.

Koramati, S., et al. (2023) stated that road geometry had been found to be a crucial determinant influencing the severity of the crash. Among the twelve different types of road geometric features, straight roads were shown to cause the highest number of collisions, followed by curved roads and four-arm junctions. Previous research has also found that traveling over a long and straight stretch with a few activity paths within the same course increases the potential of a run-off collision, especially when the paths are small. Lee, D., et al. (2023) looked at how road type, intersection type, and road conditions influence the possibility of serious injury in crashes. The research includes reference situations such as four-way intersections, roundabouts, ramps, driveways, and diverse road conditions. Four-way crossroads and driveways, according to the research, are less likely to result in severe injury crashes. This is because automobiles must halt or slow down at these crossroads, resulting in enhanced care. Roundabouts and ramps, on the other hand, are more likely to be related to severe-injury crashes. It was explained that roundabouts have low visibility, making it difficult for drivers to identify vehicles approaching the roundabout.

2.3.2 Methods for crash severity modelling

To determine and comprehend the relationships between the attributes in connection to the severity of the crash, methods such as statistical models and machine learning models have been applied. However, ML models are gaining popularity because they can discover interactions between variables that would be challenging to predict directly using statistical models and can handle and interpret massive datasets.

- Statistical models

Binary logit, binary probit, Bayesian ordered probit, Bayesian hierarchical binomial logit, generalized ordered logit, mixed generalized ordered logit, multinomial logit, multivariate probit, ordered logit, and ordered probit are some of the statistical models that have been used in the literature to conduct traffic injury severities. Many researchers have noted that statistical modeling has limitations because it generates hypotheses about data distribution and predetermines the relationship between the target and the independent variables. Statistical models can be classified into three categories:

[1] Models with binary results: Binary injury-severity outcomes, such as injury vs. non-injury collisions or fatal vs. non-fatal collisions, have been studied using conventional discrete outcome models such as the binary logit and binary probit models.

[2] Models with ordered discrete results: In crash severity modeling, it is critical to account for the ordinal distribution of injury data (such as starting with no damage

to minor injury to serious injury to fatal injury). Traditional ordered probability models have been frequently used to account for the categorical structure of the data.

[3] Models with non-ordered multinomial discrete results: In the analysis of crash injury severity data, models that neglect to account for the ordinal character of injury data have also proven prevalent. While such models do not account for the ordering of injury-severity outcomes, they do not suffer from some constraints imposed by classic ordered probit and logit models.

Table 1: Previous studies using different statistical models

Model used		**Previous studies**
Binary	Bayesian hierarchical binomial logit	Wang, X., & Abdel-Aty, M. (2008). Huang, H., Chin, H. C., & Haque, M. M. (2008).
	Bayesian ordered probit	Xie, Y., Zhang, Y., & Liang, F. (2009). Karabulut, N. C., & Ozen, M. (2023).
	Binary logit / probit	Moudon, A. V., Lin, L., Jiao, J., Hurvitz, P., & Reeves, P. (2011). Kononen, D. W., Flannagan, C. A., & Wang, S. C. (2011). Gong, Y., Lu, P., & Yang, X. T. (2023). Lidbe, A., Adanu, E. K., Tedla, E., & Jones, S. (2022). Sobhani, A., Young, W., & Logan, D. (2011, September).
	Bivariate binary probit	Russo, B. J., Yu, F., & Smaglik, E. J. (2023). Lee, J., Abdel-Aty, M., & Choi, K. (2014). Li, L., Hasnine, M. S., Nurul Habib, K. M., Persaud, B., & Shalaby, A. (2017).
Ordered discrete	Bivariate ordered probit	Yamamoto, T., & Shankar, V. N. (2004). Chiou, Y. C., Fu, C., & Ke, C. Y. (2020). Russo, B. J., Savolainen, P. & Anastasopoulos, P. C. (2014).

	A copula-based multivariate	Ahmad, N., Gayah, V. V., & Donnell, E. T. (2023). Huang, H., Ding, X., Yuan, C., Liu, X., & Tang, J. (2023). Bhowmik, T., Rahman, M., Yasmin, S., & Eluru, N. (2021).
	Generalized ordered logit	Song, D., Yang, X., Anastasopoulos, P. C., Zu, X., Yue, X., & Yang, Y. (2023). Zhao, L., Wang, C., Yang, H., Wu, X., Zhu, T., & Wang, J. (2023). Mphekgwana, P. M. (2022).
Non-ordered multinomial discrete	Multinomial logit models	Adanu, E. K., Dzinyela, R., & Agyemang, W. (2023). Islam, S. M., Washington, S., Kim, J., & Haque, M. M. (2023).
	Sequential logit and probit models	Xu, C., Tarko, A. P., Wang, W., & Liu, P. (2013). Jung, S., Qin, X., & Noyce, D. A. (2010).
	Mixed logit models	Obaid, I., Alnedawi, A., Aboud, G. M., Tamakloe, R., Zuabidi, H., & Das, S. (2022). Hasan, A. S., Orvin, M. M., Jalayer, M., Heitmann, E., & Weiss, J. (2022). Rezapour, M., & Ksaibati, K. (2022).

- Machine-learning models

Machine learning (ML) is a branch of artificial intelligence that enables real-time data analysis, decision-making, and record preparation, as well as self-learning for computers with limited, sophisticated coding. ML algorithms can detect significant risk variables and provide helpful guidance for crash prevention measures by processing massive volumes of data. Through integrating new data, machine-learning algorithms can constantly learn and increase their accuracy.

Machine-learning approaches that can undertake crash severity analysis include random forest (RF), support vector machines (SVM), Bayesian networks (BN), genetic algorithms (GAs), and artificial neural networks (ANN), as illustrated in Table 2. These models can reveal hidden correlations and nonlinear associations among features and crash severity that would not be obvious using typical statistical methods.

Table 2: Previous studies using machine learning models

Models	Previous studies
Decision Trees	Lee and Li (2015); Toran Pour et al. (2017); Mafi et al. (2018); Montella et al. (2020)
Artificial Neural Networks	Kunt et al. (2011); Sameen and Pradhan (2017); Das et al. (2018); Zheng et al. (2019)
Random Forests	Mafi et al. (2018); Tang et al. (2019); Jiang et al. (2020)
Support Vector Machines	Abdel-Aty (2014); Iranitalab and Khattak (2017); Gu et al. (2018); Hadjidimitriou et al. (2020); Xi et al. (2019)
K-nearest Neighbours	Beshah and Hill (2010); Gu et al. (2018); Montella et al. (2020)
Naïve Bayes	Kwon et al. (2015); Jeong et al. (2018); Yahaya et al. (2019)

2.4 Current state of research on traffic crash analysis

The present status of road crash analysis includes a mix of disciplines combining data collection, sophisticated analytics, behavioral research, and developing technology.

Researchers have been attempting to improve data collection methods, such as combining data from multiple sources to get full crash data. However, traffic crash investigations have progressed beyond conventional police reports. Researchers and traffic safety authorities have integrated data from many sources, such as road cameras, GPS devices, car sensors, and feeds from social networks. Analysts can acquire a broader understanding of collision scenarios and the variables contributing to them by merging these various data sources. The use of machine learning and data mining approaches to examine large-scale crash datasets has been assessed. These strategies are useful in the detection of unnoticed trends and patterns, the prediction of crash hotspots, and the classification of crash types. Furthermore, new visualization methods and computer simulations have proven to be beneficial for traffic crash investigation. To successfully examine crash data, researchers are creating enhanced visualization and simulation software. These tools can produce 3D representations of crashes, simulate situations, and evaluate the efficacy of various safety measures. There have been advancements in Intelligent Transportation Systems (ITS), which provide prospects for improving traffic crash analysis. To improve safety, researchers have been investigating the implementation of vehicle-to-vehicle and vehicle-to-infrastructure communication networks. To prevent crashes while improving the circulation of traffic, these technologies provide real-time data transmission, collision mitigation, and traffic control. The development of linked vehicle technology has created new opportunities for traffic crash investigation. Sensors and communication features in vehicles can offer real-time data on speed, location, and probable crash risks. Many industries consider the use of artificial intelligence (AI) to improve their productivity and consumer interactions. Transportation engineering has been no exception. AI techniques such as image recognition and natural language processing have been employed to advance traffic crash investigations. AI-powered systems, for example, can scan surveillance camera data automatically to discover possible risks or identify driving habits connected with crashes.

Chapter 3

METHODOLOGY

3.1 Data collection methods

Researchers have frequently used established databases to address specific research questions. These databases include data compiled by government agencies, research institutions, or other entities. Additional analyses of existing datasets can help save valuable time and money while drawing insightful conclusions. This study used the publicly available database from the PennDOT portal in Pennsylvania.

Because this study included numerical data and required statistical analysis to reach accurate and generalizable findings, a quantitative approach was used. A quantitative method is suitable for determining the prevalence of a particular occurrence or investigating the links between various factors.

Pennsylvania Department of Transportation (PennDOT):

The Pennsylvania Department of Transportation facilitated the secure and effective conveyance of individuals and commodities. It is responsible for overseeing and enhancing Pennsylvania's enormous network of highways, bridges, trains, and airports. PennDOT oversees roads, public transportation in cities and rural areas, terminals, railways, and ports. It was founded in 1911 and is directly responsible for approximately 40,000 miles of roads and approximately 25,400 bridges. The state highway system is maintained, restored, and expanded by 11,579 PennDOT personnel. The personnel operate from the Harrisburg headquarters and ten additional districts

(Fig. 7) and they have facilities in all 67 counties. PennDOT also manages the state's 12.1 million vehicle registrations and 10.1 million driver licenses and identification cards, as well as safety and emission inspection programs.

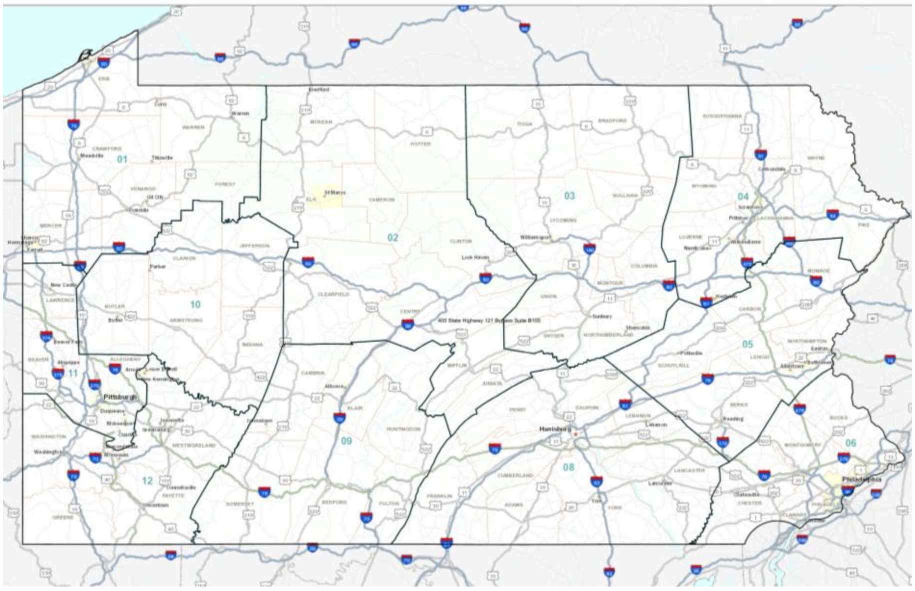

Figure 7: PennDOT interactive map [16]

The Pennsylvania Department of Transportation's (PennDOT) crash database is a comprehensive source of information on crashes involving vehicles that occur in Pennsylvania. This database is a useful tool for road safety assessments, research, and legislation. It includes an array of specific data on each crash, such as the time and location of the incident, the types of cars involved, and the number of people injured or killed. Furthermore, the database records pertinent aspects that might have led to the collision, such as weather, road features, driver behavior, and any probable alcohol or drug use. Furthermore, the collision database combines information from various sources, such as police reports, hospital records, and crash reports provided by individuals involved in crashes. This multi-source strategy ensures that crash data are comprehensive and accurate, allowing transport officials, legislators, and researchers to detect patterns, high-risk regions, and underlying causes of traffic collisions. This information is then gathered, processed, and stored in the PennDOT crash database. The database is updated regularly to guarantee that the information is recent and reliable, making it a significant resource for road safety analysis and decision-making in Pennsylvania.

The Pennsylvania Department of Transportation (PennDOT) utilizes various data-collection instruments to gather information for its crash database. Common devices include crash report forms, computerized crash reporting systems, speedometers and sensors, weather observatories, road assessment forms, surveys, geographic information systems (GIS), automatic traffic recorders (ATRs), and collision investigation teams. These tools aid in the preservation of precise and up-to-date data on traffic, crashes, and other transportation-related variables, thereby guaranteeing the safety and efficiency of PennDOT.

PennDOT data framework:

Upon accessing the PennDOT database portal, an extensive dataset can be obtained, structured into eight distinct CSV files, each catering to a specific category of information. These files, denoted as COMMVEH, CRASH, CYCLE, FLAG, PERSON, ROADWAY, TRAILVEH, and VEHICLE, collectively offer a multifaceted insight into vehicular incidents.

The COMMVEH file contains pertinent details concerning commercial vehicles, such as carrier information, cargo body types, and official registration numbers.

The CRASH file encapsulates essential crash-related data, including geographical dimensions such as county and municipality; temporal aspects such as time, day of the week, and month of the year; and quantifiable counts of elements such as individuals, vehicles, unbelted occupants, and fatalities.

The CYCLE file contains information related to motorcycle and pedal cycle involvement, emphasizing variables such as helmet use, suitable gear, and supplemental items such as side packs.

The FLAG file is a collection of binary indicators (0=No, 1=Yes) that serves to filter inquiries by denoting crash-related elements, such as the existence of a drunk driver, mobile phone usage, running on red light, motorbike involvement, and a variety of other determinants.

The PERSON file provides information on all individuals involved in the crash, including age, sex, drug and alcohol test results, and vehicle seating positions.

The ROADWAY file contains information regarding the highways involved, including route identifiers, sections, roadway types, and other defining characteristics.

The TRAILVEH file contains information regarding towed trailers, including several categories and types associated with the vehicles involved.

Finally, the VEHICLE file contains information about all vehicles involved in the crash, including body type, movement characteristics, spatial positioning, and other vehicle-related parameters. Collectively, these files comprehensively contribute to a robust and intricate representation of vehicular incidents within the PennDOT database.

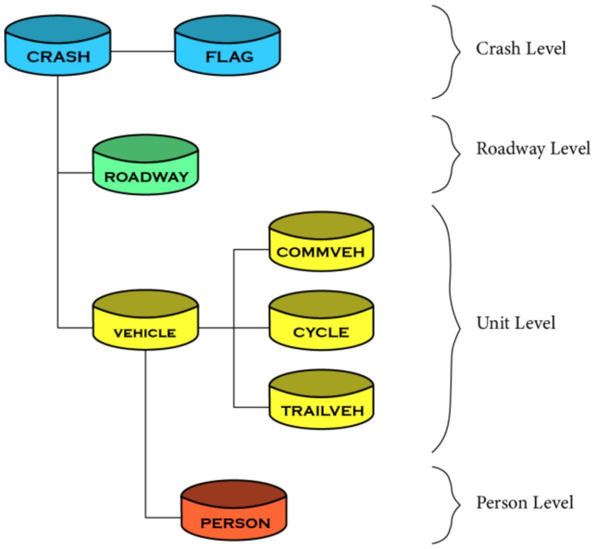

Figure 8: PennDOT database framework [17]

3.2 Overview of collected data

Data filtration was conducted by focusing on three prominent counties, namely Allegheny, Montgomery, and Philadelphia, out of the 67 counties within Pennsylvania. The rationale behind this selection is based on the substantial volume of data points represented by these specific counties compared to the remaining regions as seen in Fig. 9. To streamline the analytical process, many columns that did not contribute significantly to the analysis were removed, resulting in a refined dataset that contained 33 pertinent parameters. Table 3 presents a categorized display of the parameters involved within the model training process.

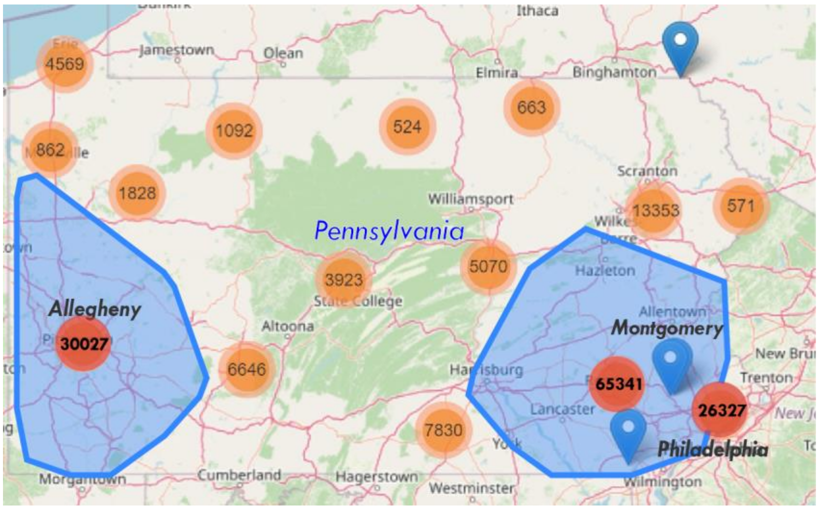

Figure 9: Top 3 counties of crash occurrences in Pennsylvania

Table 3: Final parameters included in model training

Crash characteristics	**Driver condition**	**Pedestrian characteristics**
		AGE_GROUP_[0,10]
TIME_OF_DAY_Afternoon	MATURE_DRIVER	AGE_GROUP_[10,20]
TIME_OF_DAY_Evening	YOUNG_DRIVER	AGE_GROUP_[20,30]
TIME_OF_DAY_Morning	AGGRESSIVE_DRIVING	AGE_GROUP_[30,40]
TIME_OF_DAY_Night	DRINKING_DRIVER	AGE_GROUP_[40,50]
SPEEDING_RELATED	DRUGGED_DRIVER	AGE_GROUP_[50,60]
RUNNING_RED_LT	DISTRACTED	AGE_GROUP_[60,60+]
INJ_SEVERITY	UNBELTED	SEX

Road condition	**Vehicle characteristics**
RDWY_ALIGNMENT_Curve_Left	TRAVEL_SPD
RDWY_ALIGNMENT_Curve_Right	
RDWY_ALIGNMENT_Straight	
INTERSECT_TYPE_Four-way_intersection	
INTERSECT_TYPE_Mid_Block	
INTERSECT_TYPE_T_intersection	
ILLUMINATION	
ROAD_CONDITION	
SIGNALIZED_INT	
INTERSECTION	

In the context of traffic-crash data analysis, imbalances often arise because of the small percentage of crashes on particular routes, typically resulting in sparse crash data. Addressing this issue requires careful consideration, as traditional approaches of undersampling non-collision data to align with crash data quantities may mistakenly ignore several valuable non-crash incidents. To ensure the integrity of the model training, it is necessary to correct this disparity by employing an oversampling technique specifically customized to crash data. Failure to do so might lead to an unfair bias towards non-collision occurrences within the model's performance.

To address this challenge effectively, data augmentation was performed in this study. The Synthetic Minority Oversampling Technique (SMOTE), a well-known method particularly suited for Artificial Neural Network (ANN) models, was employed. SMOTE contributes to the enhancement of crash data representation, thereby yielding a more balanced and representative dataset that is essential for robust model training and accurate predictions.

3.2.1 Preprocessing of PennDOT data

To gather reliable insights into my research project, a stratified sampling method was embarked upon. The rationale behind stratified sampling was to ensure the representation of different subgroups in the dataset. By dividing the dataset into specific strata, the sample was more likely to capture the diversity present in a large dataset.

For this thesis, a comprehensive dataset was created by consolidating data from 2017 to 2021. This involved merging information from the Crash, Flag, Vehicle, and Person sub-datasets into a unified dataset. The merging process was accomplished by connecting data points through the Crash Recorder Number (CRN), a unique identifier present in every sub-dataset file to identify each crash.

Next, focus was placed on specific groups that aligned with the research objective, which examined crashes involving SUVs and pedestrians across various age ranges.

To achieve this, the dataset was filtered using relevant criteria. Specifically, instances were selected where PERSON_TYPE from the PERSON file equaled 2, signifying pedestrians, and VEH_TYPE from the VEHICLE file equaled 6, representing SUVs. As a result, the dataset only contained information about crashes involving SUVs and pedestrians.

By honing these key variables, the aim is to intricate relationship between pedestrian activity and SUV presence, seeking valuable insights that might shape road safety measures and urban planning strategies.

To further refine the analysis, parameters from each sub-dataset were carefully selected. This process entailed considerable preprocessing and data manipulation, as parameter choices were iteratively adjusted based on considerations such as data availability and relevance to the analysis.

In the context of preprocessing, new columns were incorporated into the existing dataset. The AGE_GROUP variable was introduced based on the AGE parameter sourced from the PERSON file. This variable was discretized into bins representing 10-year intervals, spanning from 0 to 10 years, 10 to 20 years, and so forth, culminating in the final bin denoting individuals aged 60 years and above. Additionally, a TIME_OF_DAY feature was derived from the HOUR_OF_DAY attribute. The categorization was established as follows: 6 a.m. to 12 p.m. was designated as Morning, 12 p.m. to 6 p.m. as Afternoon, 6 p.m. to 12 a.m. as Evening, and 12 a.m. to 6 a.m. as Night.

The INJ_SEVERITY parameter within the PERSON file originally employed the following numerical values: 0 for no injury, 1 for fatal injury, 2 for suspected serious injuries, 3 for suspected minor injuries; and 4 for possible injuries. To align with the KABCO classification system, these labels were subsequently transformed: 0 was mapped to 'O’, 1 to 'K’, 2 to 'A’, 3 to 'B’, and 4 to 'C’. This parameter serves as the target variable within the forthcoming predictive model, a topic that will be discussed in subsequent sections.

In its initial iteration, the dataset encompassed 41 distinct parameters, collectively covering a range of collision-related parameters ranging from spatial and temporal attributes to pedestrian attributes and driver conditions. Ultimately, the initial compilation of the dataset consists of approximately 169,000 unique crash instances.

3.3 Data analysis techniques

The preparation and preprocessing of data is an essential part of any analytical process. The following sections describe the steps taken in data preprocessing, which provide the basis for a robust and informed analysis. After data preprocessing, a complete process of detailed refinement and organization was undertaken to produce the final iteration of the dataset, allowing integration into the model. First, the missing values and labels of unknown were eliminated in all columns. In addition, some columns, which were excessive for the requirements of the analysis, were removed. Following these procedures, an investigation of the interrelationships between existing features was conducted. A filter mechanism was established based on specific fixed attributes to reduce the dimensions of an extensive set of data. These attributes included exclusive consideration of urban areas, road conditions classified as dry or wet, and the categorical classification of intersections, especially mid-block, four-way intersections, and T-intersections, among others.

After the data preprocessing phase, the next step was data processing. However, a prerequisite is ensuring the appropriate scaling of the data. This process guarantees uniformity in the scope and distribution of characteristics, avoiding unnecessary domination of a particular attribute at different scales. One-hot encoding was applied to facilitate the representation of categorical variables. This method converts categorical attributes into a binary matrix and enables the incorporation of such attributes into the analysis [18]. In the case of the target variable, INJ_SEVERITY, which shows an internal ordinal structure, an alternative encoding strategy called ordinal encoding was used. Ordinary encoding preserves the hierarchical relationship between the different levels of the categorical variable, thus capturing the inherent order of the severity levels. This coding method ensures that the model can effectively distinguish and use severity levels during the learning process.

Table 4: Final categorical parameters before scaling

Categorical Parameters	Features
COUNTY	02 - Allegheny 46 - Montgomery 67 - Philadelphia
CRASH_YEAR	2017 2018 2019 2020 2021
CRASH_MONTH	January February March April May June July August September October November December

Categorical Parameters	**Features**
DAY_OF_WEEK	Sunday Monday Tuesday Wednesday Thursday Friday Saturday
TIME_OF_DAY	Morning Afternoon Evening Night
AGE_GROUP	[0,10] [10,20] [20,30] [30,40] [40,50] [50,60] [60,60+]
RDWY_ALIGNMENT	Curve Left Curve Right Straight
INTERSECT_TYPE	Four-way intersection Mid-Block T-intersection
INJ_SEVERITY	0 (C – Minor Injury) 1(B – Moderate Injury) 2 (A – Major Injury) 3 (K – Killed)

Table 5: Final binary parameters

Binary Parameters	Features
SEX	1 (Female) 0 (Male)
INTERSECTION	0 (No), 1 (Yes)
SIGNALIZED_INT	0 (No), 1 (Yes)
ROAD_CONDITION	1 (Dry) 0 (Wet)
ILLUMINATION	1 (Daylight) 0 (Dark_streetlights)
AGGRESSIVE_DRIVING	0 (No), 1 (Yes)

Binary Parameters	**Features**
DRINKING_DRIVER	0 (No), 1 (Yes)
DRUGGED_DRIVER	0 (No), 1 (Yes)
SPEEDING_RELATED	0 (No), 1 (Yes)
RUNNING_RED_LT	0 (No), 1 (Yes)
DISTRACTED	0 (No), 1 (Yes)
UNBELTED	0 (No), 1 (Yes)
MATURE_DRIVER	0 (No), 1 (Yes)
YOUNG_DRIVER	0 (No), 1 (Yes)

3.3.1 Model architecture

The architecture of the model involves several steps for constructing, training, and evaluating an artificial neural network (ANN) for predicting pedestrian crash severity. The model architecture can be broken down into the following components (Fig. 10): The training and testing data were divided into features and labels, respectively. Features are selected criteria related to pedestrian crashes, whereas labels represent the severity of the injuries. Both the training and testing datasets were transformed into 'NumPy' arrays and shuffled to ensure randomness. In the context of machine learning and data analysis, NumPy is used for handling numerical data and performing computations such as matrix operations, statistical calculations, and data manipulation. The K-fold cross-validation technique was employed with a specified number of folds (in this case, 10) to partition the training data into subsets for training and validation purposes. This technique helps assess the generalization performance of the model across different subsets of data.

A sequential neural network model is constructed within the loop for each fold. The model was sequentially composed of layers, starting with a hidden layer of 120 neurons. The layer uses the Parametric Rectified Linear Unit (PReLU) activation function, L2 regularization for both the bias and kernel, and the He uniform initialization method. This is followed by a batch normalization layer to improve convergence and speed up the training. The output layer consists of four units (for the four classes of injury severity) and uses a Softmax activation function for the probability distribution. The model was compiled using the Adam optimizer at a specified learning rate. The chosen loss function is sparse categorical cross-entropy, suitable for multi-class classification tasks, and 'accuracy' is chosen as the metric for evaluation. For each fold, training and validation data were extracted based on the current fold indices. The model was trained on the training data using a fitting method. The training process involved iterating through epochs (200 epochs are used) with a batch size of 16. After each epoch, the training accuracy and model loss were recorded. After training, the model was evaluated on the validation data to obtain validation loss and accuracy. The model's predictions on the validation data were used to compute a classification report, including precision, recall, and F1-score for each class of injury severity. This report provides detailed insights into the model's performance for each fold. For each fold, metrics, such as validation accuracy and loss, are saved in order to analyze the model's performance across different subsets of data. The classification reports are also saved, providing a comprehensive view of the model's performance in each class. This ANN architecture should train and evaluate the model iteratively using cross-validation to ensure that the model's performance is robust and consistent across different data divisions. The selection of activation functions, normalization, batch normalization, and assessment metrics all help to capture complex patterns and relationships in the data, ultimately improving pedestrian crash prediction accuracy.

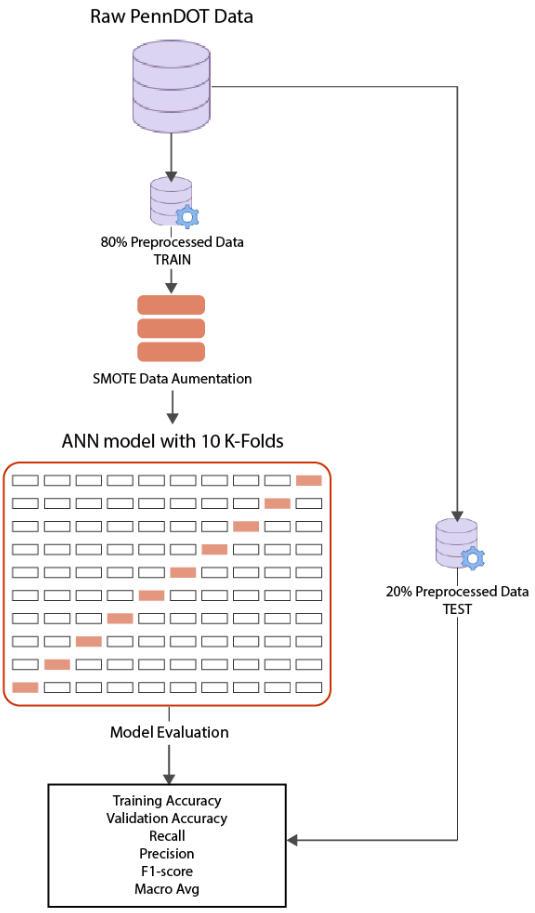

Figure 10: Flowchart of the model architecture

3.4 Software and tools

To analyze the data in this study, several instrumental software and tools were used to facilitate examination and model development. Python 3.9, TensorFlow 2.12, and Keras 2.12 are the key components in implementing artificial neural network models (ANNs). Python is a widely used and versatile programming language used to perform data manipulation, preprocessing, and analysis tasks. Python is known for its readability and extensive libraries, and provides an effective environment for orchestrating complex data operations [19]. TensorFlow, a well-known open-source machine-learning framework, plays a central role in the construction and training of ANN models. Its computational graph architecture allows for efficient numerical calculations, which makes it particularly suitable for large-scale data analysis and model optimization tasks [20]. Keras is a high-level network API integrated into TensorFlow that provides a simplified interface for building and training complex network architectures. Its user-friendly design has accelerated the process of defining model layers, loss functions, and optimization algorithms, thereby improving the development and experimental phases of ANN models [21]. The combination of Python, TensorFlow, and Keras contributed to the methodological rigor of the data analysis efforts. This integrated software toolbox facilitates the exploration of complex patterns and relationships in the dataset while simultaneously enabling the construction of complex ANN models for predictive purposes. Furthermore, it is worth noting that this integration of software tools exemplifies a comprehensive approach to utilizing modern technologies in the pursuit of insightful data analysis, thereby enhancing the academic integrity and practical implications of the study's findings.

The development of an artificial neural network (ANN) model in TensorFlow was an essential part of the study's data analysis methodology. The architecture of the ANN model is characterized by its complex arrangement of interconnected layers, which contribute to the model's ability to extract meaningful data patterns and make accurate predictions.

Chapter 4

RESULTS AND DISCUSSION

4.1 Descriptive statistics and discussion

4.1.1 Descriptive statistics

Summary statistics provide information about the distribution and characteristics of the various attributes of the dataset. Within the reported collisions, the data show the frequency of various behaviors, circumstances, and demographic characteristics. These observations can help comprehend the dataset patterns and potential links. The Count column indicates the number of instances for each feature, providing a measure of sample size. The Mean column shows the average value of the feature across all instances, while the Std column denotes the standard deviation, indicating the dispersion of values around the mean.

Table 6: Descriptive statistics of model features (one-hot encoded data version)

	count	mean	std	min	25%	50%	75%	max
TRAVEL_SPD	1619	0.295	0.163	0	0.220	0.266	0.358	0.908
TIME_OF_DAY_Afternoon	1619	0.394	0.489	0	0	0	1	1
TIME_OF_DAY_Evening	1619	0.296	0.457	0	0	0	1	1
TIME_OF_DAY_Morning	1619	0.241	0.428	0	0	0	0	1
TIME_OF_DAY_Night	1619	0.069	0.254	0	0	0	0	1
AGE_GROUP_[0,10]	1619	0.166	0.372	0	0	0	0	1
AGE_GROUP_[10,20]	1619	0.193	0.395	0	0	0	0	1
AGE_GROUP_[20,30]	1619	0.179	0.384	0	0	0	0	1
AGE_GROUP_[30,40]	1619	0.116	0.320	0	0	0	0	1
AGE_GROUP_[40,50]	1619	0.092	0.289	0	0	0	0	1
AGE_GROUP_[50,60]	1619	0.095	0.293	0	0	0	0	1
AGE_GROUP_[60,60+]	1619	0.159	0.366	0	0	0	0	1
RDWY_ALIGNMENT_Curve_Left	1619	0.027	0.163	0	0	0	0	1

RDWY_ALIGNMENT_Curve_Right	1619	0.038	0.190	0	0	0	0	1
RDWY_ALIGNMENT_Straight	1619	0.933	0.251	0	1	1	1	1
INTERSECT_TYPE_Four-way_intersection	1619	0.400	0.490	0	0	0	1	1
INTERSECT_TYPE_Mid_Block	1619	0.482	0.500	0	0	0	1	1
INTERSECT_TYPE_T_intersection	1619	0.118	0.323	0	0	0	0	1
AGGRESSIVE_DRIVING	1619	0.636	0.481	0	0	1	1	1
DRINKING_DRIVER	1619	0.051	0.221	0	0	0	0	1
DRUGGED_DRIVER	1619	0.033	0.178	0	0	0	0	1
SPEEDING_RELATED	1619	0.250	0.433	0	0	0	0.5	1
RUNNING_RED_LT	1619	0.096	0.294	0	0	0	0	1
SIGNALIZED_INT	1619	0.348	0.476	0	0	0	1	1
INTERSECTION	1619	0.518	0.500	0	0	1	1	1
DISTRACTED	1619	0.130	0.336	0	0	0	0	1
UNBELTED	1619	0.200	0.400	0	0	0	0	1
MATURE_DRIVER	1619	0.184	0.388	0	0	0	0	1
YOUNG_DRIVER	1619	0.180	0.385	0	0	0	0	1
SEX	1619	0.636	0.481	0	0	1	1	1
ILLUMINATION	1619	0.684	0.465	0	0	1	1	1
ROAD_CONDITION	1619	0.847	0.360	0	1	1	1	1
INJ_SEVERITY	1619	0.689	0.584	0	0	1	1	3

The presented summary statistics table (Table 6) offers valuable insights into the characteristics of one-hot encoded features within a dataset. The time of day influences incident occurrence, with afternoon accounting for the majority at 39.4%, followed by evening (29.6%), morning (24.1%), and night (6.9%). This indicates a pattern where incidents are more frequent during daylight hours, possibly due to higher traffic volumes and visibility. Age group distributions exhibited varying proportions, with the 60 and above category being the most prevalent at 15.9%. Most incidents occurred on straight road alignments (93.3%), while mid-block intersections made up 48.2% of the incidents. Aggressive driving behavior is notable, present in approximately 63.6% of cases, whereas occurrences of drinking and drug-related driving are less common (5.1% and 3.3%, respectively). The high prevalence of aggressive driving behavior is noteworthy, suggesting that it may be a contributing factor to many incidents. In contrast, while drinking and drug-related driving are less common, they do occur. Environmental factors indicated that incidents primarily occurred under illuminated conditions (68.4%) and on roads with favorable conditions (84.7%). Injury severity levels showed variability, with an average severity level of approximately 0.689 and a range of 0 to 3. A standard deviation of 0.584 indicates notable variability.

4.1.2 Frequency distribution

This study used a dataset that contains information on numerous factors that contribute to injury severity levels from a variety of traffic crashes. The frequency distribution table shown below depicts the proportions of the various severity levels for each key feature (Table 7).

Table 7: Frequency distribution of model features in injury severity classes

		Injury Severity Levels			
Features	**Binary input (0,1)**	**Minor**	**Moderate**	**Major**	**Fatal**
AGE_GROUP_[0,10]	1 = YES	20%	15%	6%	8%
AGE_GROUP_[10,20]	1 = YES	15%	21%	32%	34%
AGE_GROUP_[20,30]	1 = YES	17%	18%	15%	25%
AGE_GROUP_[30,40]	1 = YES	11%	12%	18%	17%
AGE_GROUP_[40,50]	1 = YES	9%	10%	3%	0%
AGE_GROUP_[50,60]	1 = YES	10%	9%	13%	8%
AGE_GROUP_[60,60+]	1 = YES	18%	15%	13%	8%
SEX	0 = Male	35%	36%	45%	83%
	1 = Female	65%	64%	55%	17%
MATURE_DRIVER	1 = YES	18%	18%	19%	25%
YOUNG_DRIVER	1 = YES	21%	15%	34%	33%

Features	Binary input (0,1)	Minor	Moderate	Major	Fatal
INTERSECTION	1 = YES	49%	54%	55%	33%
SIGNALIZED_INT	1 = YES	31%	37%	37%	33%
INTERSECT_TYPE_Four-way_intersection	1 = YES	36%	43%	40%	25%
INTERSECT_TYPE_Mid_Block	1 = YES	51%	46%	45%	67%
INTERSECT_TYPE_T_intersection	1 = YES	13%	11%	15%	8%
RDWY_ALIGNMENT_Curve_Left	1 = YES	2%	2%	7%	25%
RDWY_ALIGNMENT_Curve_Right	1 = YES	3%	4%	9%	25%
RDWY_ALIGNMENT_Straight	1 = YES	94%	94%	84%	50%
ROAD_CONDITION	0 = Wet	16%	15%	19%	0%
	1 = Dry	84%	85%	81%	100%
ILLUMINATION	0 = Dark_ streetlights	27%	32%	58%	75%
	1 = Daylight	73%	68%	42%	25%
TIME_OF_DAY_Afternoon	1 = YES	41%	40%	27%	8%
TIME_OF_DAY_Evening	1 = YES	27%	30%	48%	42%
TIME_OF_DAY_Morning	1 = YES	28%	22%	6%	17%
TIME_OF_DAY_Night	1 = YES	4%	8%	19%	33%
UNBELTED	1 = YES	16%	21%	40%	33%
RUNNING_RED_LT	1 = YES	7%	11%	12%	8%
SPEEDING_RELATED	1 = YES	33%	20%	28%	50%
AGGRESSIVE_DRIVING	1 = YES	69%	61%	55%	75%
DRINKING_DRIVER	1 = YES	3%	5%	16%	17%
DRUGGED_DRIVER	1 = YES	4%	3%	1%	0%

The frequency distribution table reveals the following major observations and implications:

- Incidents involving female and male drivers show distinct patterns in terms of injury outcomes. Notably, crashes involving female drivers tend to result in fewer fatal injuries than those involving male drivers. This gender-based disparity in injury severity implies the existence of potential differences in driving behaviors or physiological responses to crashes between the two sexes.

- An analysis of incidents involving mature drivers reveals intriguing trends in the distribution of injury severity. These incidents tend to exhibit a balanced distribution of injury severity levels, encompassing minor-to-major injuries. However, incidents involving young drivers display a high proportion of major and fatal injuries, indicating a pressing need for targeted interventions to enhance the safety of young drivers on the road.

- The type of road alignment is closely linked to injury outcomes in crashes. Specifically, incidents occurring on left curve road alignments have been associated with a higher proportion of major and fatal injuries. This trend may be attributed to the challenges posed by navigating curves, potentially leading to a higher likelihood of severe crashes. Similarly, right-curve road alignments also exhibit an elevated proportion of major and fatal injuries, which could indicate driver behavior on curved roads.

- Incidents occurring on wet roads revealed distinctive injury patterns. Such crashes display a slightly higher proportion of major injuries; however, the absence of fatal injuries raises intriguing questions. The absence of fatal injuries on wet roads might suggest the adoption of lower speeds by drivers during adverse weather conditions, thereby contributing to a reduction in the severity of crashes.

- The role of illumination in crashes cannot be understood, particularly under low-visibility conditions. Incidents that transpire under darker conditions with streetlights have demonstrated a significant correlation with major and fatal injuries. This underscores the critical importance of proper lighting infrastructure for road safety, particularly during times of reduced visibility. Moreover, the timing of crashes during the day has implications on injury severity. Incidents occurring at night and in the evening consistently exhibit higher proportions of major and fatal injuries. These findings underscore the heightened risks associated with reduced visibility and potentially increased fatigue during these times of day.

- Speeding-related incidents present distinct injury severity patterns. Crashes linked to speeding tend to have a notable proportion of major injuries. This observation can be attributed to the amplified impact forces resulting from higher speeds, leading to more severe injury outcomes. Aggressive driving behaviors are associated with a high percentage of injuries across all severity levels, including fatal injuries. The pronounced proportion of fatal injuries in incidents involving aggressive driving underscores the inherent dangers associated with such behaviors.

4.1.3 Discussion

- Demographic Factors

Age and gender play pivotal roles in shaping the outcomes of traffic incidents. For the age groups ranging from 0 to 60+, a discernible trend emerged where the proportion of major and fatal injuries increased with age until the 10-20 age group, at which point the likelihood of severe injury peaked. Notably, the absence of fatal injuries among individuals aged 40-50 raises intriguing questions about the risk profile of this age group. Gender disparities are evident, with a lower proportion of fatal injuries observed among female drivers than among their male counterparts. This gender-based variance underscores the potential influence of physiological differences, driving behavior, and risk-taking tendencies.

- Environmental Factors

The environment in which traffic incidents occur also significantly influences injury severity. Incidents transpiring at intersections, especially signalized ones, display a heightened prevalence of major and fatal injuries. This suggests the complexities and potential conflicts inherent in intersections that warrant targeted safety measures. Road conditions have a substantial impact as incidents transpiring on wet roads are linked to a relatively higher proportion of major injuries. Notably, crashes occurring on straight road alignments resulted in fewer severe injuries, highlighting the intrinsic safety of such segments. Similarly, different times of the day and illumination levels demonstrated distinct distributions of injury severity. Nighttime and evening incidents consistently exhibited higher proportions of major and fatal injuries, emphasizing the importance of adequate lighting and heightened caution during these periods.

- Behavioral and Human Factors

Human behavior and driver attributes further contribute to injury severity outcomes. The analysis revealed a disconcerting pattern among young drivers, where a substantial proportion of incidents lead to major and fatal injuries. Conversely, mature drivers appeared to have a more balanced distribution of injury severity levels. Aggressive driving behaviors, speeding, and running red lights correlated with higher proportions of moderate and major injuries, exposing the inherent risks associated with such practices. The presence of moderate injuries in incidents involving distracted driving underscores the multifaceted nature of the impact of distractions on road safety. Notably, impaired driving due to alcohol consumption and drug use consistently yields a higher proportion of major and fatal injuries, indicating the grave consequences of such behaviors.

4.2 Correlation analysis

Correlation matrices are important analytical tools for determining relationships between different factors in a dataset. It provides an organized image of how variables interact with each other and provides useful insights into possible patterns and relationships.

The principal diagonal of the matrix comprises unitary values, reflective of attributes correlating perfectly with themselves, yielding a correlation coefficient of one. The remaining entries within the matrix denote the correlations between the pairs of attributes. Positive values denote a positive correlation, whereas negative values indicate a negative one. The absolute values of these coefficients show the robustness of the correlation between the attributes.

It is critical to emphasize that this correlation does not imply causation. The observed correlations provide valuable insights and avenues for future investigation; however, they do not imply immediate cause-and-effect relationships between the variables. Furthermore, although the correlation matrix offers useful insights, careful consideration of domain expertise combined with advanced statistical analyses is required if the goal is to establish more robust relationships or make predictions based on these variables.

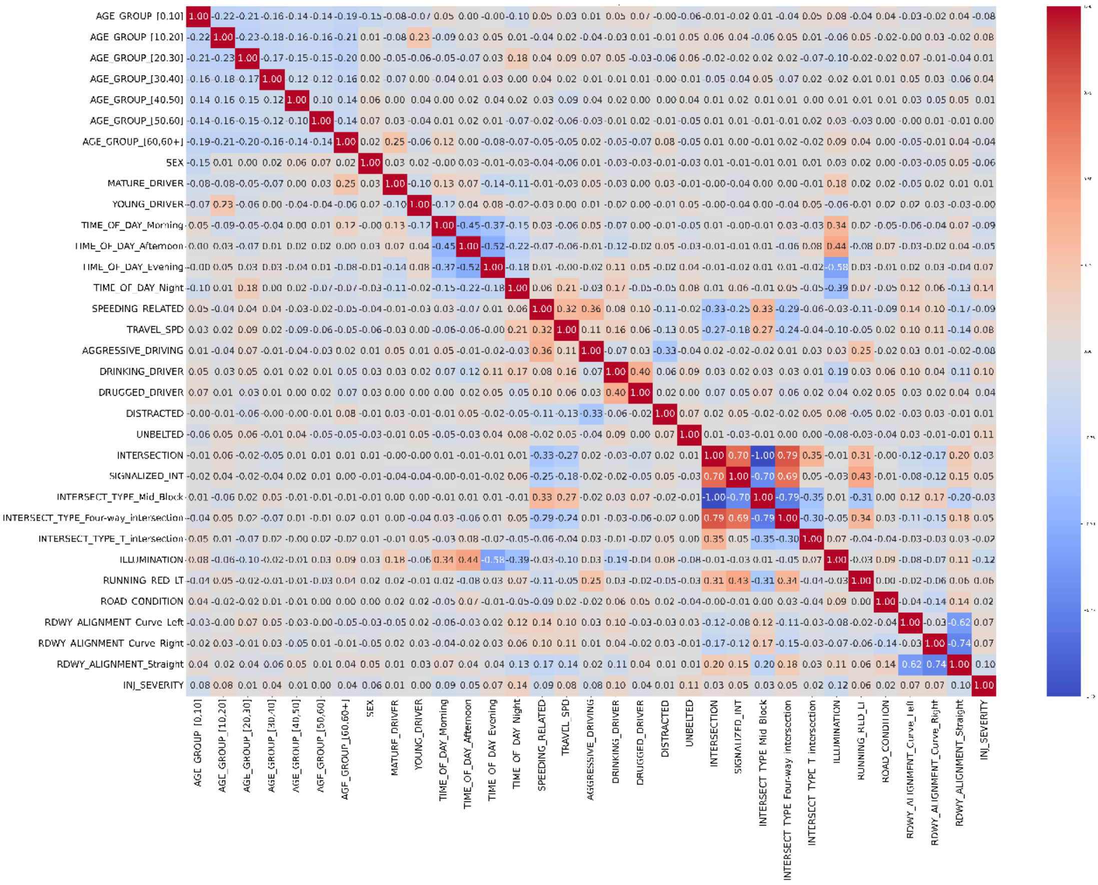

Figure 11: Correlation matrix of the SMOTE augmented data

The correlation matrix includes the correlation coefficients that define the relationships between various variables. The degree and direction of the linear correlations between variable pairs are represented by these coefficients. An inspection of the correlation matrix yields the following observations and insights:

Examination of the presented variables yields noteworthy insights into the associations between specific driving behaviors, intersection characteristics, road conditions, and injury severity in vehicular crashes. An evident relationship emerged between travel speed, denoted as TRAVEL_SPD, and crashes linked to speeding, as indicated by the positive coefficient. This relationship aligns with common intuition, as heightened travel speeds amplify the inherent crash risk. Furthermore, the coefficients pertaining to the AGGRESSIVE_DRIVING and DISTRACTED variables reveal their respective contributions to crash occurrence. Positive coefficients underscore the heightened likelihood of crashes when such behaviors are present.

The investigation extends to intersection-related factors, wherein the coefficients for the INTERSECTION, SIGNALIZED_INT, and different INTERSECT_TYPE variables provide insights into the relationships between intersections and crashes. Notably, the affirmative coefficients for SIGNALIZED_INT underscore an augmented propensity for crashes at signalized intersections. The presence of the variable RUNNING_RED_LT was significant, indicating a heightened crash likelihood associated with red light violations.

Considering injury severity, discernible patterns surface through coefficients related to diverse variables concerning INJ_SEVERITY. In particular, positive coefficients connected to DRINKING_DRIVER and DRUGGED_DRIVER accentuate the augmented severity of injuries stemming from crashes involving intoxicated drivers.

4.3 Model results

Table 8 provides a comprehensive overview of the different architectural configurations and augmentation methods utilized within the scope of the study. The aim of this analysis is to investigate the performance of distinct artificial neural network (ANN) models in predicting pedestrian crash severity. Each model is characterized by its augmentation method, number of hidden layers, nodes, activation function, K-fold parameter, and epoch count. Additionally, the table presents key performance metrics, including training accuracy, training loss, validation accuracy, and validation loss, for each respective configuration.

Table 8: Evaluation of different ANN architectures

Model	Augmentation method	Hidden layers	Nodes	Activation Function	K-fold	Epochs	Training accuracy	Training loss	Validation accuracy	Validation loss
ANN-1	None	2	100 - 70	ReLU	10	300	0.9446	0.2597	0.6181	1.0667
ANN-2	None	1	100	ReLU	10	300	0.9777	0.1496	0.6158	1.0790
ANN-3	None	1	100	ReLU	10	200	0.9537	0.2654	0.6111	0.9987
ANN-4	Oversampled	1	80	PReLU	7	200	0.9079	0.2616	0.7368	0.6082
ANN-5	Oversampled	1	100	PReLU	10	200	0.9327	0.2010	0.7427	0.5880
ANN-6	SMOTE	1	100	PReLU	8	200	0.9876	0.0685	0.8968	0.3008
ANN-7	**SMOTE**	**1**	**100**	**PReLU**	**10**	**200**	**0.9820**	**0.0685**	**0.9006**	**0.3008**

The models are denoted as ANN-1 through ANN-7, with each corresponding to a specific set of architectural attributes. ANN-1 to ANN-3 are trained without augmentation, while ANN-4 to ANN-7 are trained with oversampling or Synthetic Minority Over-sampling Technique (SMOTE) augmentation methods. The utilization of augmented data is aimed at addressing class imbalance concerns, potentially enhancing the model's capacity to discern patterns accurately (Fig. 12).

Across the spectrum of models, variations in hidden layer counts, neuron quantities, activation functions, and K-fold parameters are evident. Epoch counts also vary among the models, with training epochs ranging from 200 to 300. Training and validation accuracy metrics exhibit a range of values, reflecting the models' abilities to learn from the training data and generalize to the validation dataset. Similarly, training and validation loss values provide insights into the models' convergence rates and generalization performances.

In conclusion, this table encapsulates the nuanced interplay between architectural parameters and augmentation methods within the context of ANN models for predicting pedestrian crash severity. The tabulated results serve as a valuable reference point for evaluating and contrasting the efficacy of different model configurations, aiding in the identification of optimal approaches to achieve accurate predictions in the field of traffic safety analysis.

Several key observations can be drawn from the presented table, shedding light on the performance and characteristics of the different artificial neural network (ANN) models in predicting pedestrian crash severity:

- Augmentation Impact: Models employing data augmentation techniques (ANN-4 to ANN-7) exhibit improved validation accuracy compared to those without augmentation (ANN-1 to ANN-3). This suggests that oversampling and SMOTE contribute positively to addressing class imbalance issues, enhancing the models' ability to generalize and predict accurately.

- Activation Function Influence: The choice of activation functions, primarily ReLU and PReLU, appears to have varying effects on model performance. While both activation functions are employed across different models, their impact on accuracy and loss differs, highlighting the significance of activation functions in influencing the network's learning dynamics.

- Hidden Layer Variability: The number of hidden layers and neurons within these layers varies among the models. Models with fewer hidden layers (e.g., ANN-2 and ANN-3) demonstrate competitive validation accuracy compared to those with additional hidden layers (e.g., ANN-1). This suggests that increasing model complexity through more layers does not necessarily guarantee improved predictive performance.

- Epochs and Convergence: The number of training epochs spans a range from 200 to 300. While models with higher epoch counts may exhibit improved training accuracy, this does not necessarily translate to better validation accuracy. It underscores the importance of monitoring validation performance to avoid overfitting and optimize model generalization.

- Performance Discrepancies: Noticeable variations exist between training and validation accuracy, as well as training and validation loss, across the models. While some models achieve high training accuracy, their validation accuracy remains lower, indicating potential overfitting. Such disparities emphasize the necessity of monitoring model performance on validation data to ensure robust predictions.

- SMOTE's Efficacy: Models employing SMOTE augmentation (ANN-6 and ANN-7) consistently demonstrate high validation accuracy, suggesting that SMOTE's synthetic oversampling helps in generating informative samples that enhance the model's ability to recognize complex patterns in the data.

The observed trade-offs between training accuracy, validation accuracy, training loss, and validation loss underscore the importance of finding the right balance during model training. Optimizing these metrics collectively is crucial to ensuring the model's ability to generalize effectively. These insights contribute to a deeper understanding of the factors influencing model efficacy and assist in refining model design for accurate predictions in the realm of traffic safety analysis.

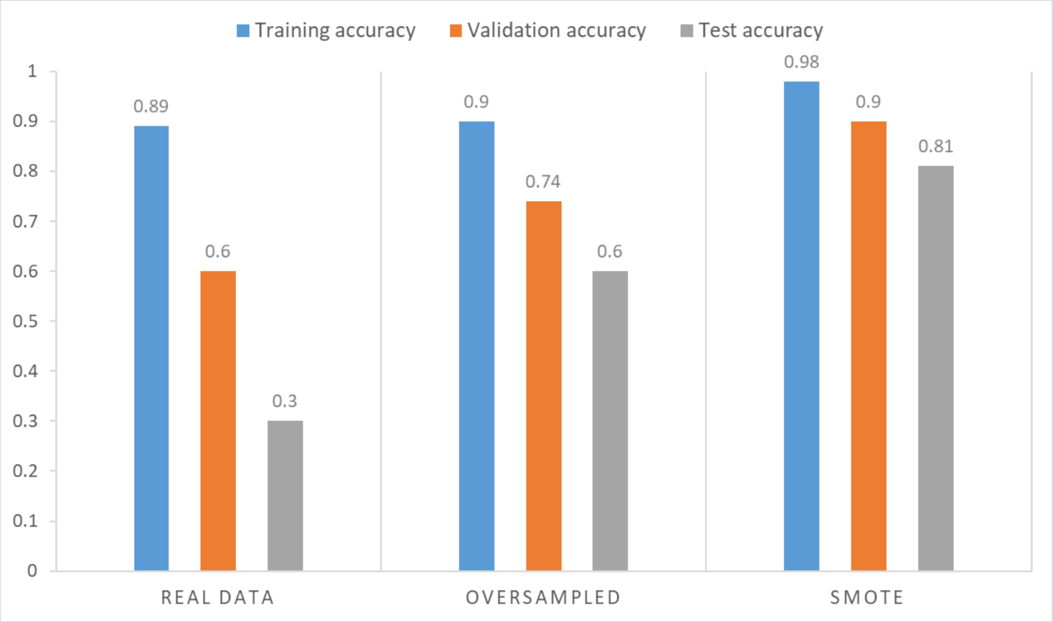

Figure 12: Comparison of the evaluation metrics across the model with data augmentation techniques

The provided confusion matrix in Fig. 13 portrays the performance of a multi-class classification model for injury severity prediction. The model effectively predicts Moderate Injury instances with 769 accurate classifications, while Minor Injury predictions are reasonably accurate in 473 instances. Challenges arise in classifying the rarer Major Injury and Fatal classes, with only 65 and 12 instances correctly predicted, respectively. Confusion between Minor Injury and Moderate Injury classes, as well as Minor Injury and Major Injury, indicates potential difficulty distinguishing between similar injury levels.

While the matrix offers valuable insights, a comprehensive evaluation of precision, recall, and F1-score metrics is essential for a nuanced assessment. Imbalanced class distributions can influence model performance, suggesting a need for techniques such as resampling and hyperparameter tuning.

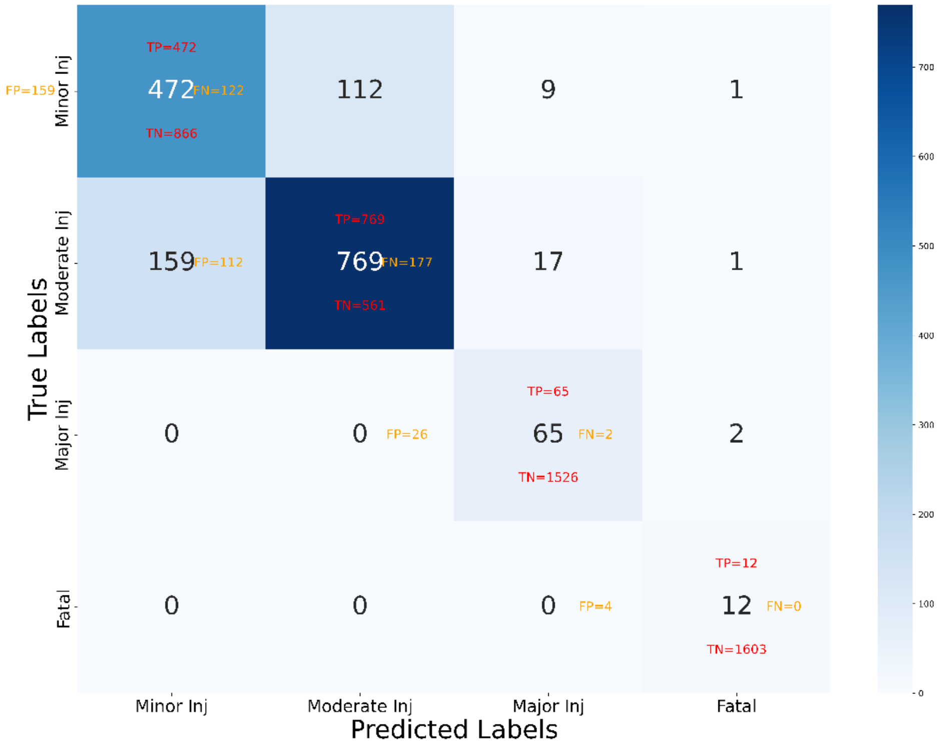

Figure 13: Confusion matrix of the classification model

The models exhibit a consistent pattern of accuracy, macro-average metrics, and weighted-average metrics across all folds, as seen in Table 9. This stability suggests that the models are generalizing well to different subsets of the data and are not excessively overfitting to any fold. The overall accuracy of 90% suggests that the models are proficient at correctly classifying injury severity. The prediction accuracy of 83.6% offers a comprehensive indication of the model's overall performance in making accurate predictions across the different classes. It is noteworthy to mention that, during the training process, the model achieved a notably high training accuracy of 99.34%, reflecting its capacity to learn from the training data and capture intricate patterns. The corresponding training loss of 0.0374 further highlights the model's efficiency in minimizing errors during training. Moreover, the balanced metrics across different classes underscore the models' ability to handle the multi-class nature of the prediction task effectively.

Across all folds, there is remarkable consistency in the precision, recall, and F1-score metrics for different classes of injury severity (Minor Injury, Moderate Injury, Major Injury, and Fatal). This suggests that the models are maintaining stable performance across various subsets of the data. The consistency observed in both macro and weighted average metrics further highlights a balanced distribution of performance across different injury severity classes, enhancing the models' reliability in capturing the intricacies of each class. The lower precision and F1 scores observed in the categories of Minor and Moderate Injuries, despite a higher volume of data points, can potentially be explained by the utilization of Synthetic Minority Over-sampling Technique (SMOTE) during the training phase. SMOTE is a well-recognized strategy for addressing class imbalance, involving the generation of synthetic instances for the minority class to achieve a balanced dataset. Although this technique effectively augments the representation of under-represented classes, it may introduce complexities into the decision boundaries, potentially affecting the model's capacity to accurately differentiate between various severity classes [22].

Table 9: Model evaluation metrics across the K-folds

Model Evaluation	**Prediction accuracy**			**83.6%**
	Avg. Precision	**Avg. Recall**	**Avg. F1-score**	**Cross-validation results**
Minor Injury	0.838	0.848	0.837	Mean Accuracy: **0.9051** (±0.0085) Mean Loss: **0.2836** (±0.0269)
Moderate Injury	0.848	0.801	0.822	
Major Injury	0.97	0.973	0.972	
Fatal	0.972	1	0.985	

Chapter 5

CONCLUSION

5.1 Conclusion of the study

This study focuses on predicting the SUV involved pedestrian crash severity using ANN models. Different model architectures are explored, with varied parameters and augmentation techniques. Augmenting data with SMOTE consistently improves validation accuracy, addressing class imbalance. Activation functions, hidden layers, and epochs influence model performance, while SMOTE proves effective in enhancing predictive accuracy.

The results of the analysis indicate a clear improvement over existing approaches, particularly in addressing class imbalance concerns. Models employing data augmentation techniques, such as oversampling and SMOTE, consistently demonstrated higher validation accuracy compared to those without augmentation. The ANN-6 model and the ANN-7 model, utilizing SMOTE augmentation, achieved validation accuracies of 89.68% and 90.06% respectively, surpassing the validation accuracies of non-augmented models, which were approximately 62%.

The exploration of activation functions, hidden layers, and epochs revealed nuanced impacts on model performance. Activation functions like PReLU exhibited superior performance in certain model architectures, leading to improved validation accuracy and lower validation loss. The ANN-6 model and the ANN-7 model, both utilizing

PReLU activation, achieved validation losses of 0.3008, significantly lower than the losses observed in models employing ReLU activation, which ranged from 0.5880 to 1.0790.

The models show stable performance across K-folds, and their evaluation metrics show high accuracy, precision, recall, and F1-score. The confusion matrix highlights accurate classifications, particularly for minor and moderate injuries. The models maintain consistent performance across different injury severity classes, ensuring reliable predictions.

In conclusion, the study integrates SMOTE into ANN models for crash prediction, addressing class imbalance and improving predictive accuracy. Descriptive statistics and correlation analysis provide insights into crash characteristics and relationships among variables. The improvements demonstrated by the model highlight the efficacy of this approach in accurately predicting pedestrian crash severity involving SUVs. Model results underscore the importance of architecture parameters and augmentation techniques, demonstrating the models' ability to accurately predict pedestrian crash severity and guiding the development of targeted interventions for safer roads.

5.2 Future work

In future work, further enhancing the predictive capabilities of the models could involve integrating additional datasets pertaining to pedestrian behavior, road infrastructure, weather conditions, and vehicle characteristics. By utilizing these diverse sources of information, the models can better capture factors influencing pedestrian crash severity, leading to more comprehensive and accurate predictions. Additionally, investigating methods to enhance interpretability of artificial neural networks, such as feature importance analysis and model visualization techniques, can empower stakeholders and decision-makers to better understand the underlying factors driving the model predictions. This increased transparency and understanding of the model's decision-making process can foster trust and facilitate the implementation of targeted interventions aimed at improving pedestrian safety on roadways.

REFERENCES

[1] Fatality Analysis Reporting System (FARS) visualization. *U.S. Department of Transportation*. https://www.transportation.gov/office-policy/transportation-policy/fatality-analysis-reporting-system-fars-visualization.

[2] Monfort, S. S., & Mueller, B. C. (2020). Pedestrian injuries from cars and SUVs: Updated crash outcomes from the vulnerable road user injury prevention alliance (VIPA). *Traffic Injury Prevention, 21*(sup1), S165-S167.

[3] *ArcGIS web application*. (n.d.). https://pennshare.maps.arcgis.com/apps/webappviewer/index.html?id=8fdbf046e36e41649bbfd9d7dd7c7e7e.

[4] Mannering, F. (2020). Big Data, traditional data and the tradeoffs between prediction and causality in highway-safety analysis. Analytic Methods in Accident Research, 25. https://doi.org/10.1016/j.amar.2020.100113.

[5] Riccardi, R. & Maria. (2022). Parametric and non-parametric analyses for pedestrian crash severity prediction in Great Britain. Sustainability, 14(6), 3188,. https://doi.org/10.3390/su14063188.

[6] Haghshenas, S. S., Guido, G., Vitale, A., & Astarita, V. (n.d.). Assessment of the level of road crash severity: Comparison of intelligence studies. *Expert Systems With Applications, 234, 121118.* https://doi.org/10.1016/j.eswa.2023.121118.

[7] Bishop, C. M. (2007). Pattern recognition and machine learning. Journal of Electronic Imaging, 16(4), 049901. https://doi.org/10.1117/1.2819119.

[8] Heaton, J. (2017). Ian Goodfellow, Yoshua Bengio, and Aaron Courville: Deep learning. *Genetic Programming and Evolvable Machines, 19(1–2), 305–307.* https://doi.org/10.1007/s10710-017-9314-z.

[9] Pinkus, A. (1999). Approximation theory of the MLP model in Neural Networks. *Acta Numerica*, *8,* 143–195. https://doi.org/10.1017/s0962492900002919.

[10] Khalid, F. (2008). Measure-based Learning Algorithms: An Analysis of Back-propagated Neural Networks.

[11] Shrinidhi, M., Kaushik Jegannathan, T. K., & Jeya, R. (2023). Classification of Imbalanced Datasets Using Various Techniques along with Variants of SMOTE Oversampling and ANN. *Advances in Science and Technology*, *124,* 504-511.

[12] Nair, V. S., & Hinton, G. E. (2010). Rectified linear units improve restricted Boltzmann machines. *International Conference on Machine Learning, 807–814.* https://icml.cc/Conferences/2010/papers/432.pdf

[13] Mahima, R., Maheswari, M., Roshana, S., Priyanka, E., Mohanan, N., & Nandhini, N. (2023). A comparative analysis of the most commonly used activation functions in deep neural network. *2023 4th International Conference*

on Electronics and Sustainable Communication Systems (ICESC. https://doi.org/10.1109/icesc57686.2023.10193390.

[14] He, K., Zhang, X., Ren, S., & Sun, J. (2015). Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification. *Proceedings of the IEEE International Conference on Computer Vision (ICCV).* https://doi.org/10.1109/iccv.2015.123.

[15] Kloeden, C. N., McLean, A. J., & Ponte, G. (2000). Travelling speed and the risk of crash involvement on rural roads. *Australian Transport Safety Bureau.*

[16] *PennDot interactive Map*. https://gis.penndot.gov/onemap/

[17] *ArcGIS web application*. (n.d.). https://pennshare.maps.arcgis.com/apps/webappviewer/index.html?id=8fdbf046e36e41649bbfd9d7dd7c7e7e.

[18] Han, D., Kwon, S., & Son, H. (2020, November). Productivity Prediction Integrating Data-Driven Method, Deep Neural Network and Exploratory Data Analysis in Montney Shale Plays. In *First EAGE Digitalization Conference and Exhibition* (Vol. 2020, No. 1, pp. 1-5). *European Association of Geoscientists & Engineers.*

[19] What is python? Executive summary. *Python.Org*. https://www.python.org/doc/essays/blurb/

[20] TensorFlow, W. https://www.tensorflow.org/about

[21] Team, K. *Keras documentation: About keras 3*. https://keras.io/about/

[22] Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). Smote: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research, 16*, 321–357. https://doi.org/10.1613/jair.953.

[23] Aziz, H. M. A. (2013). Exploring the determinants of pedestrian–vehicle crash severity in New York City. In *Accident Analysis &amp; Prevention (Vol. 50, pp. 1298-1309,)*. https://doi.org/10.1016/j.aap.2012.09.034.

[24] Cheng, W., Ye, F., Wang, C., & Bai, J. (2023). Identifying the factors contributing to freeway crash severity based on discrete choice models. *Sustainability*, *15*(3*), 1805*. https://doi.org/10.3390/su15031805.

[25] Chiou, Y.-C. (2020). Modelling two-vehicle crash severity by generalized estimating equations. In *Accident Analysis &amp; Prevention (Vol. 148).* https://doi.org/10.1016/j.aap.2020.105841.

[26] Elassad, E. A. & Zouhair. (2020a). A real-time crash prediction fusion framework: An imbalance-aware strategy for collision avoidance systems. *Transportation Research Part C: Emerging Technologies*, *118*. https://doi.org/10.1016/j.trc.2020.102708.

[27] Elassad, E. A. & Zouhair. (2020b). Class-imbalanced crash prediction based on real-time traffic and weather data: A driving simulator study. *Traffic Injury Prevention, 21*(3), *201-208,*. https://doi.org/10.1080/15389588.2020.1723794.

[28] Huang, H. (2008). Severity of driver injury and vehicle damage in traffic crashes at intersections: A bayesian hierarchical analysis. *Accident Analysis &amp; Prevention, 40*(1), *45-54,*. https://doi.org/10.1016/j.aap.2007.04.002.

[29] Islam, Z. (2021). Crash data augmentation using variational autoencoder. In *Accident Analysis &amp; Prevention* (Vol. 151, p. 105950,). https://doi.org/10.1016/j.aap.2020.105950.

[30] Karabulut, N. C., & Ozen, M. (2023). Exploring driver injury severity using latent class ordered probit model: A case study of turkey. *KSCE Journal of Civil Engineering*, *27*(3), 1312-1322,. https://doi.org/10.1007/s12205-023-0473-6.

[31] Kim, J.-K., Ulfarsson, G. F., Shankar, V. N., & Kim, S. (2008). Age and pedestrian injury severity in motor-vehicle crashes: A heteroskedastic logit analysis. *Accident Analysis &amp; Prevention*, *40*(5), 1695–1702. https://doi.org/10.1016/j.aap.2008.06.005.

[32] Kononen, D. W. (2011). Identification and validation of a logistic regression model for predicting serious injuries associated with motor vehicle crashes. *Accident Analysis &amp; Prevention*, *43*(1), 112-122,. https://doi.org/10.1016/j.aap.2010.07.018.

[33] Koramati, S., Mukherjee, A., Majumdar, B. B., & Kar, A. (2022). Development of crash prediction model using artificial neural network (ANN): A case study of hyderabad, India. *Journal of The Institution of Engineers (India): Series A*, *104*(1), 63–80. https://doi.org/10.1007/s40030-022-00696-4.

[34] Lee, C., & Abdel-Aty, M. (2005). Comprehensive analysis of vehicle–pedestrian crashes at intersections in Florida. *Accident Analysis &amp; Prevention*, *37*(4), 775–786. https://doi.org/10.1016/j.aap.2005.03.019.

[35] Lee, D., Guldmann, J.-M., & Rabenau, B. (2023). Impact of driver's age and gender, built environment, and road conditions on crash severity: A logit modeling approach. *International Journal of Environmental Research and Public Health*, *20*(3), 2338. https://doi.org/10.3390/ijerph20032338.

[36] Lee, J. (2014). Analysis of residence characteristics of at-fault drivers in traffic crashes. *Safety Science*, *68*, 6-13,. https://doi.org/10.1016/j.ssci.2014.02.019.

[37] Li, L. (2016). Investigating the interplay between the attributes of at-fault and not-at-fault drivers and the associated impacts on crash injury occurrence and severity level. *Journal of Transportation Safety &amp; Security*, *9*(4), 439-456,. https://doi.org/10.1080/19439962.2016.1237602.

[38] Moudon, A. V. (2011). The risk of pedestrian injury and fatality in collisions with motor vehicles, a social ecological study of state routes and city streets in King County, Washington. *Accident Analysis &amp; Prevention*, *43*(1), 11-24,. https://doi.org/10.1016/j.aap.2009.12.008.

[39] Russo, B. J., & Savolainen, P. T. (2014). Comparison of factors affecting injury severity in angle collisions by fault status using a random parameters bivariate ordered Probit Model. *Analytic Methods in Accident Research, 2, Apr*, 21-29,. https://doi.org/10.1016/j.amar.2014.03.001.

[40] Russo, B. J., & Yu, F. (2023). Examination of factors associated with fault status and injury severity in intersection-related rear-end crashes: Application of binary and Bivariate ordered Probit Models. *Safety Science*, *164*, 106187,. https://doi.org/10.1016/j.ssci.2023.106187.

[41] Saeed, R. H. M., & Kameran, I. H. (2022). Road Traffic Accidents and Associated Risk Factors in Erbil, Iraq: Retrospective (2017-2019) Households Based Study. *Bahrain Medical Bulletin, 44*.

[42] Smith, A. (2017). Perceptions of risk factors for road traffic accidents. *Advances in Social Sciences Research Journal*, *4*(1). https://doi.org/10.14738/assrj.41.2616.

[43] Tay, R., Choi, J., Kattan, L., & Khan, A. (2011). A multinomial logit model of pedestrian–vehicle crash severity. *International Journal of Sustainable Transportation*, *5*(4), 233–249. https://doi.org/10.1080/15568318.2010.497547.

[44] Ulfarsson, G. F., & Mannering, F. L. (2004). Differences in male and female injury severities in sport-utility vehicle, minivan, pickup and passenger car

accidents. *Accident Analysis &amp; Prevention*, *36*(2), 135-147,. https://doi.org/10.1016/s0001-4575(02)00135-5.

[45] Wang, X., & Abdel-Aty, M. (2008). Modeling left-turn crash occurrence at signalized intersections by conflicting patterns. *Accident Analysis &amp; Prevention*, *40*(1), 76-88,. https://doi.org/10.1016/j.aap.2007.04.006.

[46] Xie, Y. (2009). Crash injury severity analysis using bayesian ordered probit models. *Journal of Transportation Engineering*, *135*(1), 18-25,. https://doi.org/10.1061/(asce)0733-947x(2009)135:1(18).

[47] Yamamoto, T., & Shankar, V. N. (2004). Bivariate ordered-response Probit model of driver's and passenger's injury severities in collisions with fixed objects. *Accident Analysis &amp; Prevention*, *36*(5), 869-876,. https://doi.org/10.1016/j.aap.2003.09.002.

[48] Rashid, H. M. S., & Ismail, K. H. (2022). Road Traffic Accidents and Associated Risk Factors in Erbil, Iraq: Retrospective (2017-2019) *Households Based Study. Bahrain Medical Bulletin, 44(2).*

[49] Oikawa, S., Matsui, Y., Doi, T., & Sakurai, T. (2016). Relation between vehicle travel velocity and pedestrian injury risk in different age groups for the design of a pedestrian detection system. *Safety Science, 82, 361–367.* https://doi.org/10.1016/j.ssci.2015.10.003.