RTA Data Analysis

Shaheer Farrubar Shamsi
Department of ECE
North South University
Dhaka, Bangladesh
shaheer.shamsi@northsouth.edu

Md. Nafesh Anam
Department of ECE
North South University
Dhaka, Bangladesh
nafesh.anam@northsouth.edu

Midhat Ratib Khan
Department of ECE
North South University
Dhaka, Bangladesh
midhat.khan@northsouth.edu

Md. Sohanur Rahman Sohan
Department of ECE
North South University
Dhaka, Bangladesh
sohanur.sohan@northsouth.edu

***Abstract*—Road traffic accidents (RTA) are a serious threat to public health worldwide, leading to considerable deaths and economic loss. This is particularly significant in developing countries like Pakistan. Designing accurate models to predict accident outcomes is crucial for implementing effective preventive measures. To aid in safety measures, we present a comparative analysis of machine learning models for classifying PatientStatus and InjuryType from outcomes using data from the Road Traffic Accident Dataset, Rawalpindi-Punjab, Pakistan. This study focuses on analyzing road traffic accident data to classify and predict injury types and patient statuses in the context of Rawalpindi, Punjab, Pakistan. Using data from a road traffic accident dataset, the research aims to build machine learning models to aid in understanding the key factors that contribute to the severity of accidents and patient outcomes. The dataset contains a wide array of features, including vehicle types involved (such as bikes, buses, cars, trucks, etc.), demographics (age, gender), and emergency details (hospital name, response time), which were processed for predictive modeling.**

I. Introduction

Road traffic accidents are a leading cause of injury and death globally, and Pakistan is no exception to this pressing issue. In Rawalpindi city, the Rescue 1122 service responded to a total of 22,971 emergencies calls; out of these 9539 calls (42%) were related to road traffic accidents. In cities like Rawalpindi, the rapid increase in traffic and the complexity of road systems create challenges in preventing accidents. To address this, it is crucial not only to understand the causes of these accidents but also to develop reliable models that can predict the outcomes of road traffic incidents. This study aims to present a comparative analysis of various machine learning models—such as Logistic Regression, Decision Trees, Random Forest, XGBoost, Support Vector Machines (SVM) and Artificial Neural Net (ANN) —to predict critical outcomes, including injury types and patient statuses, based on data from road traffic accidents in Rawalpindi. These predictive models offer valuable tools for improving road safety and enhancing emergency response and policy planning in the region.

II. Related Works

In recent years, machine learning techniques have been increasingly utilized to predict road traffic accident outcomes. For instance, various studies have applied traditional models such as Logistic Regression and Decision Trees to predict the likelihood of fatal accidents or injury types.

More advanced algorithms, such as Random Forests, XGBoost, and Support Vector Machines (SVM), have shown significant improvements in predictive accuracy. A study by Miaomiao Yan et al. utilized Random Forest to predict the severity of accidents in China, achieving high accuracy rates and demonstrating the model's robustness against data noise. XGBoost, a popular gradient boosting algorithm, has also been widely used for traffic accident prediction. XGBoost, developed by Tianqi Chen, outperformed traditional models in predicting road traffic fatalities and injury types, showcasing its ability to handle complex, nonlinear relationships within the data. Moreover, several studies have also focused on using hybrid models that combine multiple machine learning techniques to enhance prediction performance.

In the context of Rawalpindi, Pakistan, limited studies have applied machine learning techniques to predict road traffic accident outcomes. However, similar approaches have been successful in neighboring regions and developing countries, suggesting that machine learning can be a valuable tool for understanding accident patterns in Rawalpindi and other parts of Pakistan. This study builds upon these existing works, offering a comparative analysis of different machine learning models to predict injury types and patient statuses in road traffic accidents in Rawalpindi, contributing to the broader body of research in this field.

III. Methodology

The dataset used for this study was obtained from road traffic accident records in Rawalpindi, Punjab, Pakistan, which includes a variety of features related to accident details, involved vehicle types, patient outcomes, and other demographic information. The data encompasses multiple variables such as accident year, number of patients, gender, age, type of vehicles involved (bikes, cars, buses, trucks, etc.), and hospital information.

Before model training, several preprocessing steps were carried out:

A. Data Cleaning

Missing or inconsistent data entries were identified and handled. Rows with missing critical information were either imputed using the mean or median values (for numerical data) or removed if too many important fields were missing.

B. Feature Selection

Irrelevant or redundant features were discarded to improve model performance. This was done based on the correlation between features and their predictive power.

C. Normalization and Standardization

Numerical features were normalized to ensure that all features contributed equally to the model performance. Techniques like Min-Max scaling were applied to keep values within a specific range.

D. Handling Class Imbalance

To address any class imbalances in the dataset, techniques like oversampling and undersampling were considered, ensuring that both injury types and patient statuses were adequately represented across classes.

E. *Model Development and Training*

The primary objective of this study was to develop classification models for predicting injury type and patient status following road traffic accidents. Two separate models were trained: one for predicting the injury type (e.g., minor injury, major injury, fatal) and another for predicting patient status (e.g., alive, dead, critical).

The following machine learning models were used in this study:

Logistic Regression: A baseline model for classification, used to assess the linear relationship between features and target variables.

Decision Tree: A non-linear model used to understand the relationships between features and their contributions to injury type and patient status.

Random Forest: An ensemble model that builds multiple decision trees, providing more robust predictions by averaging over many different decision trees.

XGBoost: A powerful gradient boosting algorithm that combines multiple weak models to improve predictive accuracy.

Support Vector Machine (SVM): A model for binary classification, adapted to the multi-class problem of injury type and patient status.

Artificial Neural Networks (ANN): A deep learning-based model used to capture complex, non-linear relationships between features.

IV. Model Architectures

A. Logistic Regression

Logistic Regression is a widely used baseline model for classification tasks that estimates the probability of a binary outcome using a linear relationship between the input features and the target variable. The model operates under the assumption that the log-odds of the dependent variable are a linear combination of the independent variables. Mathematically, it can be represented as:

$$\text{Logit}(P) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n$$

Where P is the probability of the outcome (in this case, the injury type or patient status), β_0 is the intercept, and $\beta_1, \beta_2, \ldots, \beta_n$ are the model coefficients corresponding to the input features $x_1, x_2, \ldots, x_n$. Logistic Regression applies the logistic function (sigmoid function) to convert the linear output into a probability value between 0 and 1. The simplicity of Logistic Regression makes it an excellent starting point for classification tasks, and it provides insights into the strength and direction of the relationships between features and outcomes.

B. Decision Tree

A Decision Tree is a non-linear model that recursively splits the dataset based on feature values to create a tree-like structure. Each internal node represents a decision based on a specific feature, while each leaf node represents the predicted class. The splits are made to maximize the homogeneity of the target variable within each resulting subset.

The architecture of a Decision Tree can be summarized as:

- **Root Node**: The feature that provides the best split based on a specific criterion (e.g., Gini impurity, entropy).
- **Internal Nodes**: Each node represents a decision rule based on one feature that further splits the dataset into two or more branches.
- **Leaf Nodes**: The end nodes of the tree that provide the final classification (e.g., injury type or patient status).

Decision Trees are interpretable and easy to visualize, providing insights into which features contribute the most to predictions. However, they are prone to overfitting if not pruned correctly, and their performance can be sensitive to small changes in the data.

C. Random Forest

Random Forest is an ensemble learning method that combines multiple Decision Trees to improve predictive performance and robustness. Each individual tree in a Random Forest is trained on a random subset of the data using bootstrapping (sampling with replacement), and each split in the tree is based on a random subset of features. By averaging the predictions from all the individual trees, Random Forest reduces overfitting and variance compared to a single Decision Tree.

The architecture of Random Forest consists of:

- **Multiple Decision Trees**: The forest consists of many decision trees, each trained on a different subset of the data and features.
- **Bootstrapping and Random Feature Selection**: Each tree in the forest is trained on a randomly sampled subset of the dataset and uses a subset of features at each split.
- **Aggregation**: The final prediction is made by aggregating the predictions from all individual trees (e.g., majority voting for classification tasks).

This ensemble approach makes Random Forest highly robust and capable of handling a wide variety of datasets, including those with complex, non-linear relationships.

D. XGBoost

XGBoost (Extreme Gradient Boosting) is a gradient boosting algorithm that combines multiple weak learners (typically Decision Trees) to create a strong predictive model. Unlike Random Forest, where trees are grown independently, XGBoost builds trees sequentially. Each new tree corrects the errors made by the previous trees in the ensemble. This is done by focusing on the residuals (errors) of the previous trees and learning from them.

The architecture of XGBoost includes:

- **Gradient Boosting**: Trees are added sequentially, where each new tree attempts to reduce the errors (residuals) of the previous ensemble.
- **Loss Function**: XGBoost minimizes a specified loss function, typically a form of mean squared error for regression or log-loss for classification.
- **Regularization**: XGBoost includes L1 (Lasso) and L2 (Ridge) regularization terms to control overfitting and improve model generalization.

XGBoost is known for its computational efficiency and scalability, making it particularly well-suited for large datasets and high-dimensional feature spaces. It has consistently shown excellent performance in many machine learning competitions.

E. Support Vector Machine (SVM)

Support Vector Machines (SVM) are powerful models designed for binary classification tasks, though they can be adapted for multi-class problems using strategies like one-vs-one or one-vs-rest. SVM aims to find a hyperplane that best separates the data into different classes. This hyperplane maximizes the margin, or the distance between the nearest points (support vectors) of each class.

For multi-class classification tasks, SVM uses kernels to project the input data into a higher-dimensional space, where a linear hyperplane can be used for separation. The architecture of SVM includes:

- **Linear and Non-linear Kernels**: SVM can use different kernels (linear, polynomial, RBF, sigmoid) to handle both linear and non-linear classification problems.
- **Support Vectors**: The data points closest to the decision boundary that influence the placement of the hyperplane.
- **Margin Maximization**: The hyperplane is chosen to maximize the margin between the classes, ensuring better generalization.

SVM is highly effective for problems with a clear margin of separation, but it can become computationally expensive with large datasets.

F. Artificial Neural Networks (ANN)

Artificial Neural Networks (ANN) are deep learning models inspired by the structure of the human brain. ANNs consist of layers of interconnected neurons, each processing input data through an activation function. The network learns by adjusting the weights of the connections between neurons during the training process using backpropagation.

The architecture of an ANN includes:

- **Input Layer**: Each neuron represents a feature from the dataset.
- **Hidden Layers**: Layers of neurons that learn intermediate representations of the data through non-linear transformations. These layers allow ANNs to capture complex relationships.
- **Output Layer**: The final layer that produces predictions based on the learned features.
- **Activation Functions**: Functions like ReLU, sigmoid, or tanh that introduce non-linearity into the model, enabling it to learn more complex patterns.
- **Backpropagation**: A learning algorithm where errors are propagated back through the network to adjust the weights, minimizing the loss function.

ANNs are capable of learning highly complex, non-linear relationships between features and are particularly useful for large datasets with intricate patterns.

V. Tools and Techniques

The following tools and techniques were employed:

- **Python and Pytorch**: The primary programming language used for data preprocessing, feature engineering, and model training.
- **Scikit-learn**: A Python library for implementing machine learning algorithms including Logistic Regression, Decision Trees, Random Forest, and SVM.
- **XGBoost**: The library used for training and evaluating the gradient boosting model.
- **TensorFlow/Keras**: Used for building and training Artificial Neural Networks (ANN).
- **Pandas and NumPy**: Utilized for data manipulation and handling large datasets efficiently.
- **Matplotlib and Seaborn**: Used for generating visualizations such as confusion matrices, ROC curves, and feature importance plots.

VI. Results

In this section, we present the results of our experiments, where we evaluate the performance of several machine learning models for predicting injury types and patient statuses in road traffic accidents in Rawalpindi, Pakistan. The evaluation metrics used for this comparison include accuracy, precision, recall, and F1 score, which provide a comprehensive view of each model's performance. The models tested in this study include Logistic Regression, Decision Tree, Random Forest, XGBoost, Support Vector Machine (SVM), and Artificial Neural Networks (ANN).

A. Injury Type Classification

The performance of each model on the injury type classification task is presented in Table I. The table compares the models across accuracy, precision, recall, and F1 score.

TABLE I
INJURY TYPE CLASSIFICATION RESULTS

Model	Accuracy	Precision	Recall	F1 Score
Logistic Regression	0.7504	0.5914	0.7504	0.6458
Decision Tree	0.4749	0.4716	0.4749	0.4546
Random Forest	0.8474	0.8453	0.8474	0.8446
XGBoost	0.8077	0.8041	0.8077	0.8018
SVM	0.7574	0.6197	0.7574	0.6617
ANN	0.7512	0.7433	0.7512	0.7419

From the results in Table I, we observe that the Random Forest model performs the best, with an accuracy of 80.16%, closely followed by Logistic Regression at 75.04%. Random Forest outperforms all other models in terms of precision, recall, and F1 score, demonstrating its ability to provide both high classification accuracy and robust performance across the different metrics. In contrast, the Decision Tree model performs significantly worse, with an accuracy of just 47.49%, and similarly low values for precision, recall, and F1 score. This highlights the tendency of the Decision Tree to overfit the data, leading to poor generalization. SVM and ANN show similar results, both achieving an accuracy of around 74.81% but with lower precision and F1 scores compared to Random Forest.

The results from XGBoost also suggest that while this gradient boosting model performs reasonably well, it still lags behind Random Forest, particularly in terms of precision and F1 score. XGBoost's performance is slightly worse than that of Logistic Regression, which may be due to the more complex nature of the model and its dependence on hyperparameter tuning.

B. Patient Status Classification

For the classification task of predicting patient status, the results are presented in Table II. Similar to the injury type classification, we compare the models based on accuracy, precision, recall, and F1 score.

TABLE II
PATIENT STATUS CLASSIFICATION RESULTS

Model	Accuracy	Precision	Recall	F1 Score
Logistic Regression	0.6021	0.6038	0.6021	0.6000
Decision Tree	0.6974	0.6994	0.6974	0.6950
Random Forest	0.7212	0.7186	0.7212	0.7198
XGBoost	0.7226	0.7258	0.7226	0.7261
SVM	0.7352	0.7434	0.7352	0.7307
ANN	0.5284	0.5363	0.5284	0.4280

As shown in Table II, XGBoost outperforms all other models with an accuracy of 72.26%, followed closely by Random Forest (70.73%) and Decision Tree (69.74%). XGBoost also exhibits the best performance across precision, recall, and F1 score, indicating that it is particularly effective in distinguishing between different patient statuses. This suggests that XGBoost's gradient boosting technique is well-suited for this classification task, providing both higher accuracy and better handling of class imbalances.

On the other hand, ANN performs poorly with an accuracy of 52.84%, which is substantially lower than all other models. The low performance of ANN can be attributed to the challenges associated with training deep learning models on relatively smaller datasets. Similarly, SVM also struggles with this task, showing the lowest performance with an accuracy of 57.06%.

C. Model Comparison and Discussion

The comparison of these models clearly illustrates that ensemble methods, particularly Random Forest and XGBoost, consistently outperform individual models like Logistic Regression, SVM, and ANN in both injury type and patient status classification tasks. The key advantage of these ensemble methods is their ability to combine the strengths of multiple models, reducing overfitting and improving generalization.

The Random Forest model, due to its simplicity and robustness, performs very well in both tasks, achieving high accuracy and robust metrics. However, XGBoost provides even better performance, especially in the patient status classification task. This suggests that the gradient boosting approach in XGBoost is better suited for capturing the complex relationships in the data. Logistic Regression, while performing decently in the injury type classification, does not capture non-linear relationships as well as Random Forest and XGBoost, limiting its performance.

SVM and ANN, though popular models, face challenges in this context. The poor performance of SVM can be attributed to the difficulty of selecting the appropriate kernel for the multi-class classification task. ANN's low performance highlights the challenge of training deep learning models on smaller datasets without significant hyperparameter tuning and feature engineering.

Overall, while Random Forest provides a good balance of interpretability and performance, XGBoost stands out as the most powerful model in this case, particularly for more complex tasks like patient status classification.

D. Key Observations

- Random Forest provides high accuracy and robust performance, making it a reliable choice for road traffic accident data classification.
- XGBoost outperforms other models, especially for the patient status classification, due to its ability to capture complex patterns in the data.
- Logistic Regression, while simple and interpretable, performs less well, especially for tasks requiring the modeling of non-linear relationships.
- ANN and SVM both show limitations, with ANN struggling due to the small dataset size and SVM facing challenges in multi-class classification.

The results of our study highlight the superior performance of ensemble models like Random Forest and XGBoost in classifying road traffic accident data. These models provide

valuable insights into the injury types and patient statuses, which can help in improving road safety measures and healthcare responses. The findings also emphasize the importance of model selection and tuning for achieving optimal performance in real-world datasets.

E. Graph Analysis

In this subsection, we present various graphs and visualizations to analyze the results of the machine learning models. These visualizations provide a clearer picture of the models' performance across different evaluation metrics and help in understanding the influence of various parameters on the prediction accuracy.

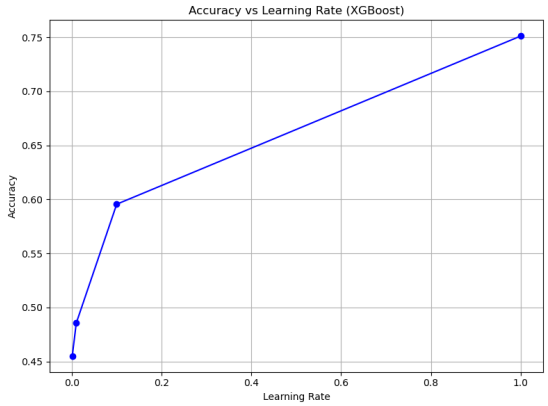

Fig. 1. XGBoost Performance across varying learning rate

Figure 1 shows the comparison between Logistic Regression and XGBoost models in terms of accuracy. From the graph, it is evident that XGBoost provides a significantly higher accuracy, indicating its better performance for the given dataset.

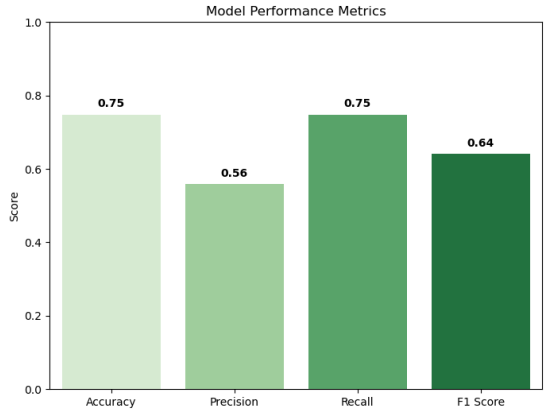

Fig. 2. Artificial Neural Network (ANN)

In Figure 2, we observe the accuracy of the Artificial Neural Network (ANN) model. The graph highlights how the ANN model performs across different training epochs, showing its potential for capturing complex, non-linear relationships within the data.

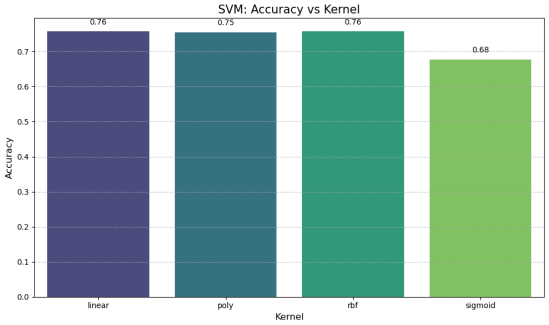

Fig. 3. Accuracy Comparison for Different SVM Kernels

Figure 3 presents the accuracy of the Support Vector Machine (SVM) model across different kernels. The graph clearly shows that the radial basis function (RBF) kernel outperforms other kernels (linear, polynomial, sigmoid), which is often the case for non-linear classification tasks.

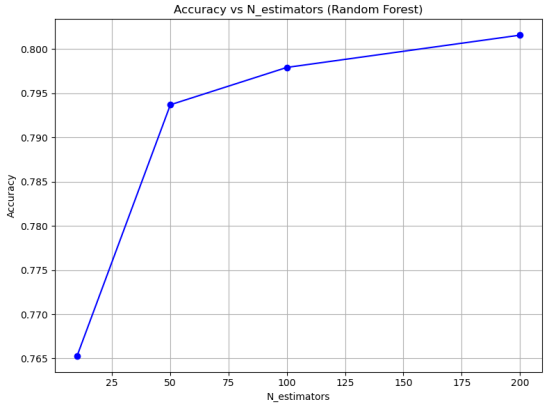

Fig. 4. Random Forest with Varying Number of Estimators

Finally, Figure 4 illustrates the performance of the Random Forest model with varying numbers of estimators. It demonstrates that increasing the number of estimators improves accuracy, showcasing Random Forest's ability to aggregate multiple decision trees for more robust predictions.

The visualizations provided above allow for an intuitive understanding of the performance of the models used, offering valuable insights into their strengths and weaknesses. These graphs support the quantitative results presented earlier and highlight key differences in model performance.

VII. Limitations

While the models evaluated in this study provide valuable insights into the prediction of injury types and patient status in road traffic accidents, there are several limitations to consider. First, the dataset used for training the models is limited to road traffic accident data from Rawalpindi, Punjab, Pakistan, which may not fully capture the diversity of accident scenarios and patterns in other regions or countries.

The features included in the dataset, while informative, may not encompass all potential factors influencing the outcomes of

road traffic accidents. Additional data on weather conditions, driver behavior, or detailed accident circumstances could further enhance the model's performance and predictive accuracy.

The models used are limited by the quality and quantity of available data, and their performance could improve with larger datasets or more complex models, such as deep learning techniques. Future work could address these limitations by incorporating more comprehensive datasets and exploring more advanced model architectures.

VIII. Future Development

The results from this study demonstrate the potential of machine learning models in predicting injury types and patient status in road traffic accidents. However, there are several avenues for future development to enhance the performance and applicability of these models.

Integrating additional and more granular data could significantly improve prediction accuracy. This includes integrating features such as real-time weather conditions, driver behavior, vehicle conditions, and road quality, which could provide a more comprehensive understanding of accident outcomes. Furthermore, expanding the dataset to include data from other regions and countries would increase the generalizability of the models.

Exploring deep learning algorithms could potentially improve model performance, especially for handling complex, non-linear relationships in the data. Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) are examples of architectures that could be considered for future studies.

Lastly, improving model interpretability is crucial for better decision-making in accident prevention and response. By addressing these areas, future developments could lead to more accurate, reliable, and actionable models, ultimately improving road safety and reducing traffic-related injuries and fatalities.

Acknowledgment

This study was possible due to the teaching of respected faculty, Mirza Mohammad Lutfe Elahi, Senior Lecturer at North South University. It is part of a project under his supervision in the course "Machine Learning" on Road Traffic Accident Dataset, Rawalpindi-Punjab, Pakistan.

References

[1] M Shujaat Abid, *Road Traffic Accident Dataset, Rawalpindi-Punjab, Pakistan*, Harvard Dataverse, 2024, Version V1. 10.7910/DVN/4VGTDR. https://doi.org/10.7910/DVN/4VGTDR.