A Machine Learning Approach to Analyzing Road Traffic Accident: Case Study of Rawalpindi, Pakistan

Samiul Basir Bhuiyan
2111006642
Dept. of Electrical and Computer Engineering
North South University
samiul.bhuiyan@northsouth.edu

Md Sazzad Hossain Adib
2132025642
Dept. of Electrical and Computer Engineering
North South University
sazzad.adib@northsouth.edu

Mohammed Aman Bhuiyan
2131864642
Dept. of Electrical and Computer Engineering
North South University
mohammed.aman@northsouth.edu

Abstract—In this study, machine learning approaches have been applied to predict critical outcomes of road traffic accidents, specifically *Injury Type* and *Patient Status*. This can assist emergency response teams and healthcare providers in optimizing resource allocation and improving patient care. We have used the "Road Traffic Accident Dataset, Rawalpindi-Punjab, Pakistan," containing real-world accident data. The study achieved an accuracy of 89% for predicting injury type and 99% for patient status using the Random Forest classifier. The novelty of this research lies in its effective application of machine learning techniques to analyze road traffic accidents and provide actionable insights for improving road safety and emergency response.

Keywords: Machine Learning, Traffic Data Analysis, Road traffic accidents, Injury prediction, Patient status, classification, Model Comparison, Predictive Systems, Transportation Safety.

I. INTRODUCTION

Road traffic accidents are a significant global concern, ranking among the leading causes of death and contributing to substantial socio-economic losses. According to the World Health Organization (WHO), road traffic injuries claim approximately 1.35 million lives annually, with millions more suffering non-fatal injuries (WHO, 2021) [1]. While many countries have made strides in improving road safety, developing nations, such as Pakistan, continue to face disproportionately high accident rates, resulting in devastating human and financial costs (Gershenson et al., 2019) [2]. In Pakistan, the situation is particularly dire in urban areas like Rawalpindi, where rapid motorization and urbanization have exacerbated the issue. In 2020 alone, nearly 8,000 fatalities were reported due to road traffic accidents, highlighting the urgent need for effective safety interventions (PBS, 2020) [3]. Addressing this challenge requires innovative approaches to predict and mitigate accidents, leveraging the potential of emerging technologies.

In recent years, researchers have paid increasing attention to determining factors that significantly affect severity of driver injuries caused by traffic accidents. Machine learning (ML) has shown great promise in the field of traffic safety, enabling the analysis of complex data to predict accident occurrences, identify high-risk zones, and assess accident severity (Zheng et al., 2019) [4]. Despite the global advancements in this area, research focused on Pakistan, and specifically on Rawalpindi, remains scarce. This study seeks to bridge this gap by developing ML-based models to predict the severity of traffic accidents and analyze contributing factors using a dataset from 2020 to 2023[5].

The remainder of this paper is structured as follows: Section II provides a review of related literature, Section III describes the dataset and research methodology, Section IV presents the model architecture & training, followed by section V which contains results and evaluation, and Section VI, VII, VIII and IX concludes with discussion, conclusion, acknowledgements and references.

II. LITERATURE REVIEW

Earlier studies, such as Abdel-Aty and Abdelwahab (2004) compared two distinct paradigms of Artificial Neural Networks (ANN) and the fuzzy Adaptive Resonance Theory (ART) for predicting the severity level of a crash [6]. Kabeer (2016) proposed the use of an ensemble technique to analyze road accidents in Leeds [7]. It is shown that ensemble techniques can increase the prediction accuracy to 78.03% while the accuracy of Naïve Bayes and DTs was 58.76% and 51.22%, respectively. A combination of clustering and classification is used by Iranitalab and Khattak (2017) to improve classification accuracy [8], while Mashfiq Rizvee et al. (2021) employed data mining for identifying accident-prone areas [9], offering valuable insights for policymakers. Recent studies have increasingly utilized machine learning for crash severity prediction. Ijaz et al. (2021) presented a comparative study of Decision Tree, and Random Forest for fatality prediction involving crashes of three-wheeled motorized rickshaws [10]. Dong et al. (2022) employed boosting-based ensemble learning models, highlighting significant variables such as month and driver age [11]. Madushani et al. (2023) investigated crash severity factors in Sri Lanka using explainable machine learning methods [12], emphasizing their superiority over traditional regression analysis. Eboli et al. (2020) utilized logistic regression to dissect crash-contributing factors, providing practical implications to formulate road safety measures [13]. Li et al. (2017) analyzed fatal accidents using data mining techniques [14], emphasizing the role of human factors like intoxication. Rocha et al. (2023) introduced a framework for identifying informative variables for distinguishing fatal from non-fatal accidents, offering valuable insights for accident prevention strategies [15]. Dealing with imbalanced data remains a challenge in crash severity classification. Building upon the foundation of previous studies, our research leverages classification models to address the critical challenge of accurately predicting injury type and patient status from road traffic accident (RTA) data in Rawalpindi, Pakistan. By employing advanced machine learning techniques and

addressing challenges such as imbalanced datasets, this study demonstrates the potential of achieving high predictive accuracy. Our approach not only builds on the efficacy of classification methods such as Decision Trees, Random Forests, and boosting-based ensembles but also emphasizes interpretability, ensuring the results can inform actionable road safety measures.

III. DATASET AND METHODOLOGY

A. *Dataset Description*

The dataset utilized in this study was collected from Harvard Dataverse, a publicly accessible data repository for researchers across disciplines to share, archive, and explore research data. The dataset encompasses road traffic accident data from 2020 to 2023, representing 46,189 recorded incidents. It includes 25 columns that detail various aspects of each accident, ranging from the timing and location to patient outcomes and injury types. Below is a detailed breakdown of the dataset's structure and contents:

1. **EcYear:** The year in which the accident occurred, capturing temporal trends in accident frequencies.
2. **EcNumber**: A unique identifier for each recorded accident, ensuring traceability.
3. **CallTime:** The timestamp indicating when the emergency call was received, enabling response time analysis.
4. **EmergencyArea:** The specific location of the accident, facilitating geographical analysis of incident hotspots.
5. **TotalPatientsInEmergency:** The number of individuals involved in the accident, ranging from single-patient cases to multi-casualty scenarios.
6. **Gender:** Gender of the patient(s), providing demographic insights into affected populations.
7. Age: The age of the patient(s), which is crucial for understanding risk patterns among different age groups.
8. **HospitalName:** The hospital to which the patient(s) were transported, indicating the healthcare facility's involvement.
9. **Reason:** The reason or cause of the accident, capturing both immediate triggers (e.g., "Bike Slip") and broader contextual factors.
10. **ResponseTime:** The time (in minutes) taken for emergency services to respond to the call, critical for evaluating system efficiency.
11. **EducationTitle:** The educational background of the patient(s), potentially linked to risk awareness or behavior.
12. **Cause:** A textual description of the underlying reason for the accident, offering granular contextual data.
13. **Vehicle Involvement:** Binary indicators for the types of vehicles involved, such as bikes, cars, trucks, and others, enabling multi-vehicle incident analysis.
14. **InjuryType:** The type of injury sustained by the patient(s), a key target variable for this study.
15. **PatientStatus:** The health status of the patient(s) post-accident, another critical target variable.

The dataset is structured with 23 features and 2 primary target variables:

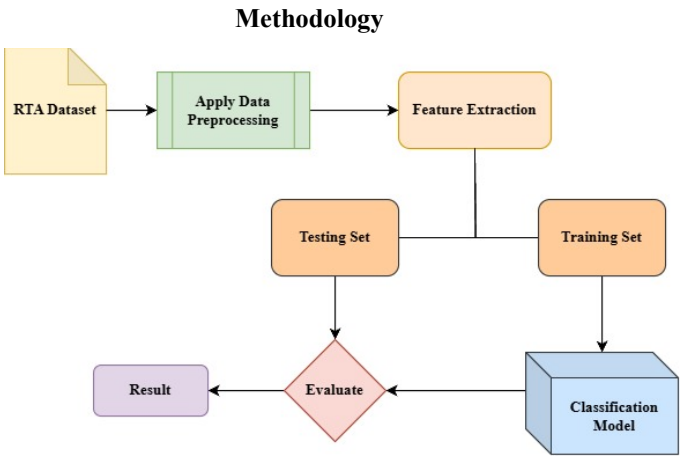


Fig. 1. Flowchart describing the methodology followed during the study

B. *Data Preprocessing*

Data preprocessing is an essential part of any machine learning project, transforming raw data into a clean and structured form suitable for analysis. For this dataset, preprocessing included handling missing data, feature scaling, feature engineering, outlier detection, and encoding categorical variables. Below is the detailed step-by-step explanation of the preprocessing pipeline.

1. **Checked missing values for each feature**: Evaluate the dataset for missing values across all features to identify data gaps. This step ensures data completeness and informs necessary preprocessing measures. Checked null values and visualized via barplots to assess the extent of missing data and determine if imputation or removal is needed.

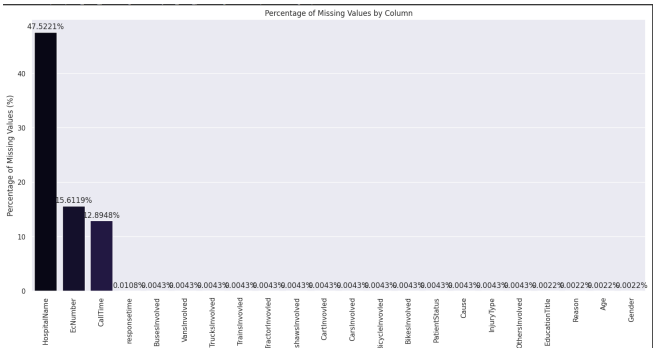


Fig. 2. Barplot for checking missing values

2. **Checked Correlation**: Analyze the correlation between features to identify relationships and potential multicollinearity. This helps in feature selection and understanding the dataset's structure.

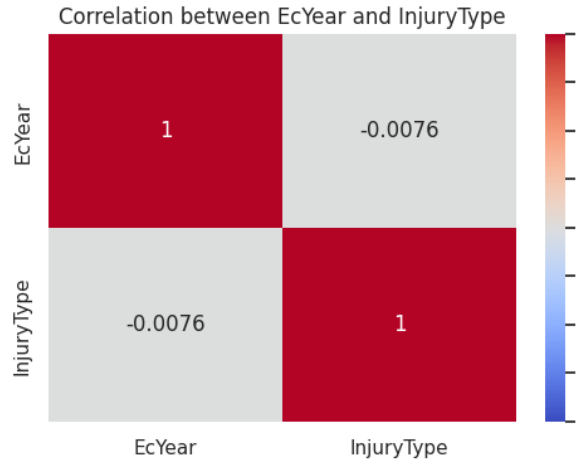


Fig. 3. Correlation between EcYear and InjuryType

3. **Feature Selection:** Features were selected based on their relevance to the target variables using correlation analysis and domain knowledge. Irrelevant or redundant columns were dropped to streamline the dataset for analysis. We dropped features checked correlation and other necessary condition, and those features are: EcYear, EcNumber.

   After dropping irrelevant features we left with:

```
Index: 40231 entries, 33190 to 16460
Data columns (total 23 columns):
 #   Column                    Non-Null Count  Dtype
---  ------                    --------------  -----
 0   EmergencyArea             40231 non-null  object
 1   TotalPatientsInEmergency  40231 non-null  object
 2   Gender                    40231 non-null  object
 3   Age                       40231 non-null  float64
 4   HospitalName              40231 non-null  object
 5   Reason                    40231 non-null  object
 6   responsetime              40231 non-null  float64
 7   EducationTitle            40231 non-null  object
 8   InjuryType                40231 non-null  object
 9   Cause                     40231 non-null  object
 10  PatientStatus             40231 non-null  object
 11  BicycleInvovled           40231 non-null  float64
 12  BikesInvolved             40231 non-null  float64
 13  BusesInvolved             40231 non-null  float64
 14  CarsInvolved              40231 non-null  float64
 15  CartInvovled              40231 non-null  float64
 16  RickshawsInvolved         40231 non-null  float64
 17  TractorInvovled           40231 non-null  float64
 18  TrainsInvovled            40231 non-null  float64
 19  TrucksInvolved            40231 non-null  float64
 20  VansInvolved              40231 non-null  float64
 21  OthersInvolved            40231 non-null  float64
 22  Hour                      40231 non-null  int32
```

Fig. 4. Remaining Features

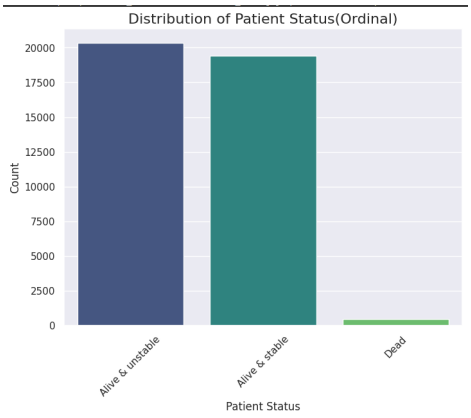
4. **Extracted hour from Call Time:** The 'hour' was extracted from the 'CallTime' feature using datetime parsing to analyze time-based patterns more effectively.

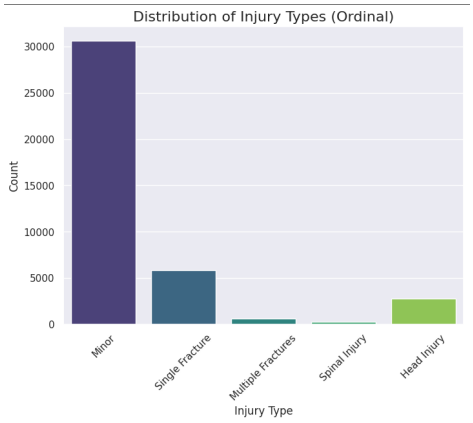
5. **Encoding Categorical features:**

   - Hospital name with Label Encoding
   - Emergency Area with Frequency Encoding
   - Gender encoded with Label Encoding
   - Education title with Ordinal Encoding
   - Injury Type with ordinal Encoding and mapped:
   - Reason encoded with Target Encoding with Injury Type
   - Cause with Label Encoding
   - Patient status with Ordinal Encoding and mapped

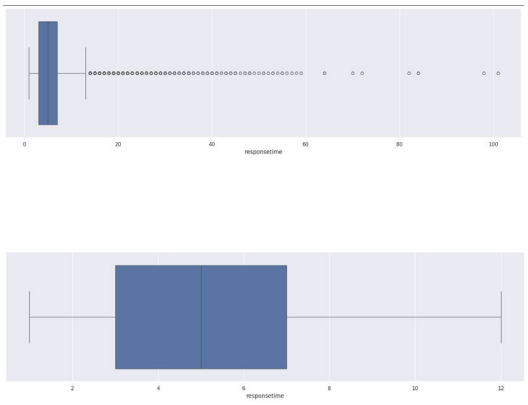
   After encoding all remaining features, they converted into numerical datatype for applying further computation

```
RangeIndex: 40231 entries, 0 to 40230
Data columns (total 23 columns):
 #   Column                    Non-Null Count  Dtype
---  ------                    --------------  -----
 0   EmergencyArea             40231 non-null  int64
 1   TotalPatientsInEmergency  40231 non-null  float64
 2   Gender                    40231 non-null  int64
 3   Age                       40231 non-null  float64
 4   HospitalName              40231 non-null  int64
 5   Reason                    40231 non-null  float64
 6   responsetime              40231 non-null  float64
 7   Cause                     40231 non-null  int64
 8   BicycleInvovled           40231 non-null  float64
 9   BikesInvolved             40231 non-null  float64
 10  BusesInvolved             40231 non-null  float64
 11  CarsInvolved              40231 non-null  float64
 12  CartInvovled              40231 non-null  float64
 13  RickshawsInvolved         40231 non-null  float64
 14  TractorInvovled           40231 non-null  float64
 15  TrainsInvovled            40231 non-null  float64
 16  TrucksInvolved            40231 non-null  float64
 17  VansInvolved              40231 non-null  float64
 18  OthersInvolved            40231 non-null  float64
 19  Hour                      40231 non-null  int32
 20  EducationTitle_Ordinal    40231 non-null  float64
 21  InjuryType_Ordinal        40231 non-null  float64
 22  Patient_Status_Ordinal    40231 non-null  float64
dtypes: float64(18), int32(1), int64(4)
```

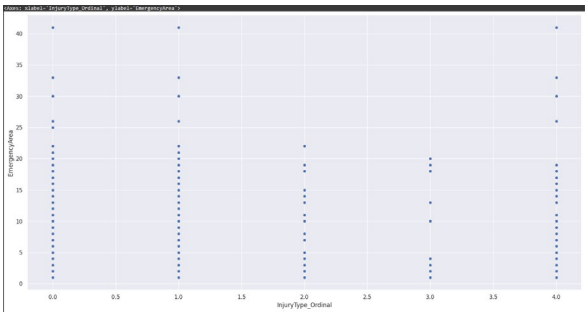




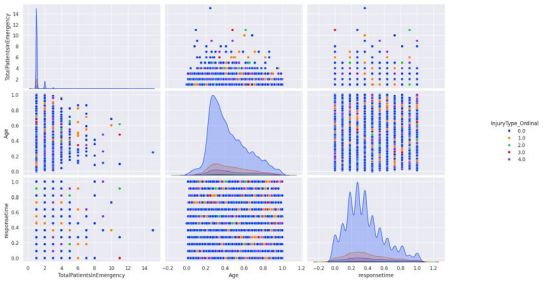
6. **Box plot and outlier remove:** Box plots were used to identify outliers in the dataset by visualizing feature distributions. Outliers were removed to prevent them from negatively impacting model performance and ensure data quality.



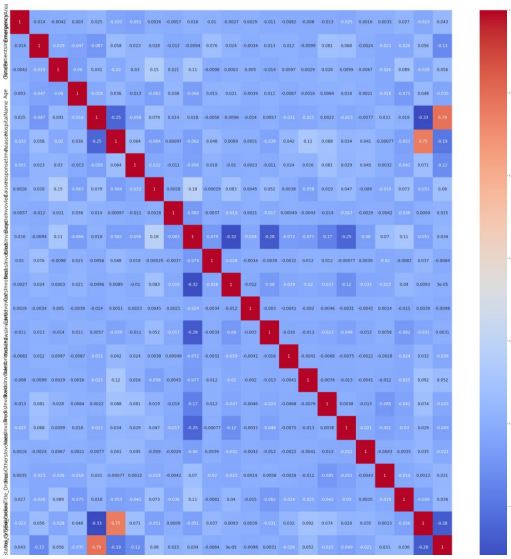
7. **Normalization for feature Scaling:** Normalization was applied to scale features to a uniform range, typically [0, 1], ensuring that all features contribute equally during model training and improving convergence performance.



8. **Pair plot for visualization:** A pair plot was created to visualize pairwise relationships between selected features. This provided insights into feature distributions and potential patterns or trends in the data.



9. **Heatmap for checking correlation:** A heatmap was generated to visualize the correlation matrix, highlighting the relationships between features. This helped identify highly correlated features for potential removal.



10. **Data portioning(train & test split)**: The dataset was partitioned into training and testing sets to evaluate model performance. A standard split ratio

of 80:20 was used to ensure sufficient data for both training and validation.

1. Target Variable InjuryType

Best Performing Model: XGBoost & Random Forest

- Accuracy: 89%
- Precision: 88%
- Recall: 89%
- F1 Score: 88%

Analysis:

The **XGBoost & Random Forest** model demonstrated the best performance for InjuryType predictions, achieving the highest accuracy (89%), precision(88%), recall (89%) and f1(88%) among all models. Thier ability to model complex patterns in the dataset is a key factor in its success.

Conclusion:

XGBoost & Random Forest is the best-performing model for InjuryType predictions due to its highest accuracy, precision, recall and f1 making it a reliable choice for this target variable.

Target	Best Model	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)
Injury Type	XGBoost	89	88	89	88
Injury Type	Random Forest	89	88	89	88
Patient Status	Random Forest	99	98	99	98

Table I. Final Comparison of Best Models

Final Insights

- InjuryType Target: **XGBoost & Random Forest** emerged as the best-performing model, achieving the highest accuracy and recall. Its ability to model complex relationships in the data gave it an edge over other models.

- PatientStatus Target: **Random Forest** also excelled in predicting PatientStatus, with consistent performance across all metrics, making it the most reliable model for this task.

2. Target Variable PatientStatus

Best Performing Model: Random Forest

- Accuracy: 99%
- Precision: 98%
- Recall: 99%
- F1 Score: 98%

Analysis:

The **Random Forest** model also delivered the best performance for PatientStatus predictions. Its accuracy (99%), precision (98%), recall (99%) and F1 score(98%) demonstrate consistent and reliable results across all metrics.

Conclusion:

Random Forest stands out as the best model for PatientStatus predictions due to its strong and consistent performance across all evaluation metrics almost close to near perfect.

IV. Model Architectures and Training

This section describes the models employed in the study, their architecture, hyperparameter tuning configurations, and training methodologies.

A. Models Used

Six machine learning models were trained and evaluated for predicting two critical targets: Patient Status and Injury Type. These models, along with their hyperparameters and methodologies, are summarized below:

1. Logistic Regression

- Description: Logistic Regression is a statistical method for binary and multiclass classification problems. It works by modeling the relationship between the independent variables and the dependent variable (target) using a sigmoid function, which maps predicted values to probabilities.
- Hyperparameters:
  - Solver: Algorithms like liblinear, lbfgs, newton_cg, newton_cholesky, sag and saga used for optimization.
  - Iterated though various combination for finding best result.

- Methodology:
    - Data scaled using standardization for better performance.
    - Hyperparameter tuning performed using grid search and cross-validation.

2. Decision Tree

- Description: Decision Tree models classify data by recursively splitting it into subsets based on feature thresholds, forming a tree structure.
- Hyperparameters:
    - Maximum Depth: Limits the depth of the tree to avoid overfitting. Typical values tested: 10 to 100.
    - Minimum Samples Split: Minimum number of samples required to split an internal node. Tested values: 2 to 25.
    - Criterion: Metrics like Gini Impurity or Entropy for deciding splits.
- Methodology:
    - Hyperparameter tuning performed using grid search and cross-validation.

3. **Support Vector Machine (SVM)**

- Description: SVM is a supervised learning model that separates classes using a hyperplane in a high-dimensional space. It is effective for linear and non-linear classification.
- Methodology:
    - Features scaled using standard scaling before train and test.
    - Default parameter gives similar result to hypertuning.

4. Random Forest

- Description: Random Forest is an ensemble learning method that constructs multiple decision trees and combines their outputs for more accurate predictions.
- Hyperparameters:
    - Number of Estimators: Number of trees in the forest. Typical values: 50 to 200.
    - Maximum Depth: Limits the depth of each tree between [10, 50, None].
    - Criterion: Gini Impurity or Entropy for splits.
    - Also, minimum sample split and minimum samples leaf were set into [2, 5, 4] and [1, 2, 4].
    - Max feature parameter was sqrt, log2 and none.
- The cross-validation process involves fitting 5 folds during the tuning.

5. XGBoost

- Description: XGBoost (Extreme Gradient Boosting) is an advanced gradient boosting framework that improves speed and performance.
- Hyperparameters:
    - Learning Rate (Eta): Step size shrinkage used to prevent overfitting (values: 0.01–0.2).
    - Max Depth: Maximum depth of trees (typical range: 3–9).
    - Number of Estimators: Number of boosting rounds (tested values: 100 to 250).
    - Subsample: Proportion of samples used for training each tree.
    - Also, parameter like min_child_weight and scale_pos_weight was used in tuning.
- Methodology:
    - RandomizedSearchCV was used for tuning with 3-fold cross validation.

6. AdaBoost

- Description: AdaBoost (Adaptive Boosting) combines multiple weak learners (typically decision stumps) to create a strong classifier by focusing more on misclassified instances in subsequent iterations.
- Hyperparameters:
    - Number of Estimators: Number of weak learners (tested value: 50).
    - Learning Rate: Shrinks the contribution of each weak learner (value: 1.0).
    - Base Estimator: Typically, a decision tree with max depth=1 (stump).
- Methodology:
    - Features not scaled, as decision trees are used as base estimators.
    - Evaluated for performance gains over unboosted models.

V. RESULTS AND EVALUATION

A. Model Comparison

The performance of each model was evaluated based on accuracy, precision, recall, F1 score, and confusion matrix. The table below summarizes the results for Injury Type predictions and Patient Status:

Model Performance Metrics

Model	Accuracy	Precision	Recall	F1 Score
Logistic Regression	**0.863973**	**0.826833**	**0.858082**	**0.839149**
Decision Tree	0.885890	0.882304	0.885890	0.875674
SVM	0.865753	0.845508	0.865753	0.848439
Random Forest(HyperTuned)	0.890000	**0.887221**	0.890000	0.880588
XG Boost(HyperTuned)	**0.890548**	0.887101	**0.890548**	**0.882189**
ADA Boost	0.887945	0.885581	0.887945	0.877906

Table II. Final Comparison of All Models(Injury Type)

Model	Accuracy	Precision	Recall	F1 Score
Logistic Regression	**0.930959**	**0.911985**	**0.913151**	**0.909767**
Decision Tree	0.982603	0.982766	0.982603	0.982683
SVM	0.955205	0.947931	0.955205	0.950577
Random Forest	**0.990274**	**0.988822**	**0.990274**	**0.988634**
XG Boost	0.989041	0.987605	0.989041	0.988096
ADA Boost	0.985068	0.986791	0.985068	0.985854

Table III. Final Comparison of All Models(Patient Status)

Table II & III. Analysis of Model Performance

- XGBoost and Random Forest demonstrated the best performance in predicting both **Injury Type** and **Patient Status** targets, achieving superior **accuracy, precision, recall** and **f1** compared to all other models tested. These two models consistently outperformed their counterparts, highlighting their robustness in handling the complexities of the **Road Traffic Accident Dataset**.
- Additionally, **AdaBoost** exhibited promising results that were very close to the performance of XGBoost and Random Forest for both target variables. This indicates that AdaBoost is a strong contender in terms of predictive accuracy and robustness, though slightly behind the top two models.
- Conversely, models such as **Logistic Regression**, **Support Vector Machine (SVM)**, and **Decision Tree** underperformed in comparison. These models failed to achieve similar levels of accuracy and recall, indicating their limitations when applied to the given dataset and the nature of the predictive tasks involved.
- These observations suggest that ensemble-based methods like **XGBoost, Random Forest, and AdaBoost** are better suited to capture the non-linear relationships and feature interactions present in the **Road Traffic Accident Dataset**.

B. Challenges in the Study

This study encountered several challenges during its execution, ranging from data quality issues to computational difficulties. These challenges are summarized below:

1. **Data Quality and Organization:**
   a. The dataset was messy and unorganized, which required significant preprocessing efforts.
   b. Mixed-language entries in columns like "Reason" (including Urdu) complicated the text-cleaning process, requiring manual intervention and specialized tools for translation and uniformity.
   c. Missing values in critical columns, such as "CallTime" and "HospitalName," posed additional difficulties, necessitating careful imputation strategies.
2. **Model-Specific Challenges:**
   a. Support Vector Machine (SVM): SVM took over two hours to complete training due to its high computational complexity, particularly with larger datasets.
   b. While tuning the ensemble model for checking different combination it took while to figure out best parameter for training for our dataset
3. **Computational Constraints:**
   a. Processing a large dataset with complex models like Random Forest, XGBoost, and ADABoost required substantial computational resources.
   b. Model hyperparameter tuning, cross-validation increased computation time, demanding optimization strategies such as early stopping and learning rate adjustment.
4. **Balancing Model Performance:**
   a. Achieving a balance between precision, recall, and F1 score, especially for imbalanced classes, required iterative fine-tuning of models and hyperparameters.
   b. Logistic Regression, while interpretable, struggled with non-linear relationships in the data, necessitating reliance on more complex models for better accuracy.
5. **Evaluation and Interpretation:**
   a. Comparing models across diverse metrics and ensuring a fair evaluation required extensive validation techniques, including confusion matrices and performance metric plots, to ensure the robustness of conclusions.

VI. DISCUSSION

1. **Model Performance**:
   The **Random Forest** and **XGBoost** models demonstrated the best performance in predicting **Patient Status** and **Injury Type**, thanks to their ability to handle complex data distributions and class imbalances effectively. Also, **ADABoost showed promising performance similar to those models.** These models outperformed **Logistic Regression, SVM** and **Decision Tree**, which showed limitations, particularly in predicting **Injury Type**, due to their struggles with capturing non-linear relationships.

2. **Impact of Features achieving high accuracy:**

Analysis of Hospital Name's Impact on Predicting Patient Status

Upon analyzing the features in the dataset, we observed a notable relationship between the *hospital name* and the predictive score for *patient status.* Initially, we noticed that the *hospital name* feature contained nearly 50% null values. Due to this high level of missing data, we first considered dropping this column altogether, as it did not seem to contribute effectively to predicting patient status.

However, we decided to replace the missing values with the label **"unknown"** instead. This change led to a significant improvement in the classification model's ability to predict patient status. The results of this adjustment are shown below:

Model	Accuracy	Precision	Recall	F1 Score
Logistic Regression	0.718630	**0.747900**	0.724521	0.706424
Decision Tree	0.635205	0.635287	0.635205	0.635220
SVM	**0.534932**	**0.523716**	**0.534932**	**0.484315**
Random Forest	0.721233	0.722682	0.721233	0.716825
XG Boost	**0.734932**	0.738681	**0.734932**	**0.730436**
ADA Boost	0.725205	0.732663	0.725205	0.719494

Table IV. Results while removing Hospital name

Model	Accuracy	Precision	Recall	F1 Score
Logistic Regression	**0.930959**	**0.911985**	**0.913151**	**0.909767**
Decision Tree	0.982603	0.982766	0.982603	0.982683
SVM	0.955205	0.947931	0.955205	0.950577
Random Forest	**0.990274**	**0.988822**	**0.990274**	**0.988634**
XG Boost	0.989041	0.987605	0.989041	0.988096
ADA Boost	0.985068	0.986791	0.985068	0.985854

Table V. Results while replacing null with 'Unknown' label

Analysis of Reason encoded with Target Encoding with InjuryType Impact on Predicting Injury Type

Initially, we applied label encoding to the reason feature and ordinal encoding to the injury type feature. These initial encoding methods yielded average predictive scores, indicating limited performance in capturing relationships.

Upon further analysis of the relationship between reason and injury type, we applied target encoding instead. This approach resulted in a significant improvement in the predictive performance of the model. The comparative results are shown below:

Model Performance Metrics

Model	Accuracy	Precision	Recall	F1 Score
Logistic Regression	**0.770411**	**0.597974**	0.773288	**0.674424**
Decision Tree	0.774658	0.683627	0.774658	0.693989
SVM	0.779041	**0.735676**	0.779041	0.688841
Random Forest	0.770685	0.701228	**0.770685**	0.713276
XG Boost	0.771644	0.706908	0.771644	**0.714951**
ADA Boost	**0.779589**	0.728793	**0.779589**	0.689745

Table VI. Result before Target Encoding

Model Performance Metrics

Model	Accuracy	Precision	Recall	F1 Score
Logistic Regression	**0.863973**	**0.826833**	**0.858082**	**0.839149**
Decision Tree	0.885890	0.882304	0.885890	0.875674
SVM	0.865753	0.845508	0.865753	0.848439
Random Forest(HyperTuned)	0.890000	**0.887221**	0.890000	0.880588
XG Boost(HyperTuned)	**0.890548**	0.887101	**0.890548**	**0.882189**
ADA Boost	0.887945	0.885581	0.887945	0.877906

Table VII. Results after applying Target Encoding

3. **Future Work:**
   a. Further hyperparameter tuning to optimize individual model performance.
   b. Exploring advanced ensemble methods like stacking or blending.
   c. Testing additional deep learning architectures, such as Convolutional Neural Networks (CNNs) or Recurrent Neural Networks (RNNs).
   d. Expanding the dataset to include more diverse accident scenarios and regions.

VII. CONCLUSION

In conclusion, this study establishes a comprehensive framework for predicting road traffic accident outcomes in Rawalpindi, Pakistan, utilizing machine learning models. Among the models evaluated, **Random Forest and XGBoost** proved to be the most effective due to their adaptability, robustness, and superior performance in handling complex patterns within the data. The findings highlight that implementing **class weighting and resampling techniques** significantly improves model performance, particularly when addressing challenges associated with imbalanced datasets. This demonstrates the potential of machine learning approaches to provide accurate and reliable predictions in real-world road traffic accident analysis. By systematically comparing models across key evaluation metrics, this research highlights the potential of machine learning in traffic safety analytics. These findings aim to support policymakers, urban planners, and road safety practitioners in designing data-driven interventions to reduce the socio-economic burden of road accidents. Future studies may explore additional models, datasets, and methodologies to further advance this critical area of research.

VIII. ACKNOWLEDGEMENTS

The authors extend their gratitude to their instructors and peers for their invaluable guidance and unwavering support throughout this project. Special appreciation is also due to the data providers and the machine learning community,

whose open-source contributions were instrumental in the successful completion of this research.

IX. REFERENCES

1. WHO. (2021). Road traffic injuries. World Health Organization. https://www.who.int/news-room/fact-sheets/detail/road-traffic-injuries.
2. Gershenson, P., Schmidt, L., & Zetterberg, S. (2019). Machine learning for road accident prediction: A systematic review. International Journal of Transportation Science and Technology, 8(1), 1-12.
3. PBS. (2020). Pakistan Bureau of Statistics Report on Road Accidents. Pakistan Bureau of Statistics.
4. Zheng, L., Li, Y., & Zhang, M. (2019). Application of machine learning methods for accident severity prediction in urban traffic. Accident Analysis & Prevention, 125, 103-110.
5. M Shujaat Abid, 2024, "Road Traffic Accident Dataset, Rawalpindi-Punjab, Pakistan", Harvard Dataverse, V1.
6. Mohamed A Abdel-Aty and Hassan T Abdelwahab. Predicting injury severity levels in traffic crashes: a modeling comparison. Journal of transportation engineering, 130(2):204–210, 2004.
7. Syed Ibrahim Kabeer. Analysis of Road accident in Leeds. PhD thesis, Dublin, National College of Ireland, 2016.
8. Amirfarrokh Iranitalab and Aemal Khattak. Comparison of four statistical and machine learning methods for crash severity prediction. Accident Analysis & Prevention, 108:27–36, 2017.
9. Md Mashfiq Rizvee, Md Amiruzzaman, and Md Rajibul Islam. Data mining and visualization to understand accidentprone areas. In Proceedings of International Joint Conference on Advances in Computational Intelligence: IJCACI 2020, pages 143–154. Springer, 2021.
10. Muhammad Ijaz, Muhammad Zahid, Arshad Jamal, et al. A comparative study of machine learning classifiers for injury severity prediction of crashes involving three-wheeled motorized rickshaw. Accident Analysis & Prevention, 154: 106094, 2021.
11. Sheng Dong, Afaq Khattak, Irfan Ullah, Jibiao Zhou, and Arshad Hussain. Predicting and analyzing road traffic injury severity using boosting-based ensemble learning models with shapley additive explanations. International journal of environmental research and public health, 19(5):2925, 2022.
12. JPS Shashiprabha Madushani, RM Kelum Sandamal, DPP Meddage, HR Pasindu, and PI Ayantha Gomes. Evaluating expressway traffic crash severity by using logistic regression and explainable & supervised machine learning classifiers. Transportation Engineering, 13:100190, 2023.
13. Laura Eboli, Carmen Forciniti, and Gabriella Mazzulla. Factors influencing accident severity: an analysis by road accident type. Transportation research procedia, 47:449–456, 2020.
14. Liling Li, Sharad Shrestha, and Gongzhu Hu. Analysis of road traffic fatal accidents using data mining techniques. In 2017 IEEE 15th International Conference on Software Engineering Research, Management and Applications (SERA), pages 363–370. IEEE, 2017.
15. Miriam Karla Rocha, Michel José Anzanello, Gabrielli Harumi Yamashita, Felipe Caleffi, and Helena Cybis. Identifying the most informative variables to discriminate between fatal and non-fatal road accidents. Case studies on transport policy, 14:101093, 2023.