Predictive Analysis of Patient Outcomes in Road Traffic Accidents using Ensemble Machine Learning Methods

Motasim Abid
Department of Electrical and Computer Engineering (ECE)
North South University(NSU)
Dhaka, Bangladesh
motasim.abid@northsouth.edu

Naima Zaman Roshni
Department of Electrical and Computer Engineering (ECE)
North South University(NSU)
Dhaka, Bangladesh
Naima.roshni@northsouth.edu

Muhammad Ruhul Amin
Department of Electrical and Computer Engineering (ECE)
North South University(NSU)
Dhaka, Bangladesh
ruhul.amin15@northsouth.edu

Ekra Bintay Ismail
Department of Electrical and Computer Engineering(ECE)
North South University(NSU)
Dhaka, Bangladesh
ekra.bintay@northsouth.edu

Ms. Silvia Ahmed
Department of Electrical and Computer Engineering(ECE)
North South University(NSU)
Dhaka, Bangladesh
silvia.ahmed@northsouth.edu

Abstract**-- Timely prediction of patient conditions following road traffic accidents is critical for effective emergency response and healthcare resource management. This study presents a machine learning approach to classify patient outcomes—categorized as stable, critical, or deceased—based on real-world accident and hospital data collected between 2020 and 2023. In the first phase, baseline models such as Logistic Regression and Decision Tree were used to establish performance benchmarks. The second phase extended this work by implementing ensemble methods including Gradient Boosting, XGBoost, Bagging, AdaBoost, Stacking, and Voting. These models were evaluated using standard classification metrics, with hyperparameter tuning applied to enhance generalization. Gradient Boosting achieved the highest performance with 99.36% accuracy. Feature importance analysis identified Cause_Reason, injury type, and patient count as key predictors. The results demonstrate that ensemble learning significantly improves predictive performance, supporting their application in real-time triage and emergency decision-making systems.**

Keywords—Road Traffic Accidents, Machine Learning, Ensemble Methods, Patient Outcome Prediction, Gradient Boosting, Emergency Triage

I. INTRODUCTION

Road traffic accidents remain a significant public health issue, especially in rapidly urbanizing areas where vehicle congestion and human error frequently lead to emergencies. In such critical scenarios, predicting the condition of accident victims at the earliest possible stage is essential for effective triage and resource allocation. Traditional methods often rely heavily on human judgment, which, under pressure, may lack consistency and speed. The growing availability of structured accident and hospital data presents an opportunity to introduce machine learning-based decision support systems that can assist in making these assessments faster and more reliably.

This project aims to predict patient status—classified as stable, critical, or deceased—based on features such as the nature of the accident, patient demographics, emergency response time, and types of vehicles involved. The work was conducted in two phases: initial evaluation using baseline models, followed by implementation of more advanced ensemble techniques to improve accuracy and robustness.

Our main contributions in this project include:

- Preprocessing and cleaning a real-world dataset consisting of over 46,000 road accident records.
- Designing and comparing baseline models, including Logistic Regression and Decision Tree classifiers, to establish performance benchmarks.
- Applying ensemble learning methods—such as Gradient Boosting, XGBoost, Voting, Bagging, AdaBoost, and Stacking—for improved predictive accuracy.
- Conducting thorough hyperparameter tuning and cross-validation to optimize each model
- Performing feature importance analysis to identify the most critical factors influencing patient outcomes.
- Achieving over 99% accuracy with ensemble models, particularly Gradient Boosting, showing significant improvement over individual classifiers.
- Presenting a comparative analysis of all models with insights into their performance, strengths, and practical applications in emergency response.

Through this structured approach, our study not only demonstrates the benefits of ensemble learning in healthcare prediction tasks but also provides a clear foundation for developing automated triage tools for emergency medical services.

II. Literature Review

Previous research on road traffic accident prediction has shown promising results using machine learning techniques to classify accident severity and improve traffic safety planning. Labib et al. (2019) [1] conducted a study focused on accident severity analysis in Bangladesh, using Decision Tree, KNN, Naïve Bayes, and AdaBoost to categorize accidents into multiple levels, such as fatal and grievous. Their findings indicated that AdaBoost performed best among the models, emphasizing the potential of ensemble methods in improving predictive accuracy. The paper also highlighted the importance of feature selection and data preprocessing—two steps we followed closely in our own project.

Similarly, Ballamudi (2019) [2] reviewed various machine learning approaches applied to road accident severity classification and suggested that algorithms like Decision Tree and AdaBoost can identify complex patterns in traffic data when traditional statistical methods fall short. The study also emphasized the growing role of ensemble learning in capturing hidden relationships within accident datasets, aligning with the motivation behind our final-phase work.

Inspired by these papers, our project extended the use of ensemble techniques—specifically Gradient Boosting, XGBoost, Voting, and Stacking—on a real-world dataset, aiming to improve prediction of patient outcomes after traffic accidents. The consistent success of AdaBoost across both papers provided a strong foundation for selecting boosting methods in our final implementation.

III. Methodology

A. Dataset Overview

The dataset used in this project was collected from emergency response and hospital records, covering road traffic accident cases reported between 2020 and 2023. It contains detailed information about the nature of each accident, the patients involved, emergency response metrics, and various contextual factors.

The dataset comprises a total of 46,190 records with 25 features, including both numerical and categorical variables. Each row corresponds to a unique accident case, and the primary target variable is PatientStatus, which indicates whether the patient was stable, critical, or deceased following the accident.

Key features include:

- Demographic Information: Age, Gender, EducationTitle
- Accident Context: EmergencyArea, Cause, Reason, ResponseTime
- Injury-related Variables: InjuryType,

  TotalPatientEmergency.
- Vehicle Involvement: Count of different vehicle types like Cars, Buses, Trucks, Rickshaws, Trains etc.

The dataset presented a diverse set of predictors that could influence post-accident outcomes. The richness and structure of the dataset made it suitable for classification tasks, especially for models focused on handling a mix of categorical and continuous inputs.

In both phases of this project, the same dataset was used. Phase 1 established baseline performance using conventional classifiers, while Phase 2 applied advanced ensemble techniques on the same data to improve classification accuracy and generalization.

B. Data Preprocessing

In both Phase 1 and the Final Phase of our project, we followed a consistent approach for handling missing values and duplicates. Numerical columns with missing entries were filled using the median to reduce the influence of outliers, while categorical columns were filled using the most frequent value (mode). Duplicate records were removed to ensure data integrity.

However, several important improvements were made in the Final Phase that significantly impacted the model's predictive capability. One major area of refinement was the encoding of categorical variables. While Phase 1 relied on straightforward encoding techniques such as ordinal encoding for ordered features and one-hot encoding for binary categories like gender, the Final Phase incorporated more nuanced strategies. For example, target encoding was applied more selectively, particularly to the InjuryType column, which showed a strong relationship with the target variable. In addition, the handling of features such as EmergencyArea, Reason, and Cause was revised in the Final Phase to facilitate the construction of a new composite feature.

This newly engineered feature, named Cause_Reason, was created by combining the Cause, Reason, and EmergencyArea columns into a single categorical variable. The rationale behind this combination was to capture a more holistic context of the accident, integrating environmental, behavioral, and geographical information into one feature. This feature proved to be highly valuable—emerging as the most important predictor across all models. Its introduction led to a dramatic performance improvement, raising model accuracy from around 77% in Phase 1 to over 99% in the Final Phase.

Despite the high accuracy, our results do not indicate overfitting. The dataset's large size—over 46,000 records—provided ample diversity for the models to generalize effectively. The models performed consistently well across training, validation, and test sets, and similar results were obtained using different ensemble algorithms. Furthermore, cross-validation and hyperparameter tuning were carried out carefully to ensure stability and prevent the model from memorizing patterns in the training data. These observations suggest that the increase in accuracy was a result of more informative feature construction and refined encoding strategies, rather than overfitting.

C. Models

In Phase 1, we used Logistic Regression and Decision Tree as baseline models to establish initial performance benchmarks. These models were chosen for their simplicity and interpretability. While Logistic Regression performed slightly better overall, both models had difficulty predicting the minority class accurately.

In the Final Phase, we shifted our focus to ensemble learning techniques to improve predictive performance and generalization. We implemented Bagging, AdaBoost, Gradient Boosting, XGBoost, Stacking, and Voting classifiers. These models combine multiple learners to reduce variance and bias, capturing more complex relationships in the data. Among them, Gradient Boosting and Stacking delivered the most notable improvements, with test accuracies approaching 99%.

All models were tuned using appropriate hyperparameter search techniques to ensure fairness and optimal performance. A detailed comparison of these models, along with evaluation results, will be discussed in the Results and Discussion section.

Model	Type	Purpose
Logistic Regression	Base Model	Benchmark, interpretable results
Decision Tree	Base Model	Benchmark, tree-based classification
Bagging	Ensemble (Bagging)	Reduces variance, improves stability
AdaBoost	Ensemble (Boosting)	Focuses on difficult examples
Gradient Boosting	Ensemble (Boosting)	Strong learner, best performance
XGBoost	Ensemble (Boosting)	Optimized gradient boosting
Voting Classifier(SOFT)	Ensemble (Voting)	Combines predictions from models
Stacking	Ensemble (Stacking)	Meta-learning, leverages multiple models

Fig. 1. Summary of Models Used in Phase 1 and Final Phase

IV. EXPERIMENTS

To evaluate model performance, we split the dataset into training, validation, and test sets. The training set was used to fit the models, while the validation set helped in tuning hyperparameters and avoiding overfitting. The final evaluation was conducted on the test set to ensure unbiased performance comparison.

In Phase 1, we trained Logistic Regression and Decision Tree classifiers using default settings and later refined them with Grid Search. These models served as performance baselines.

For the Final Phase, we implemented a range of ensemble models—Bagging, AdaBoost, Gradient Boosting, XGBoost, Voting, and Stacking. Each model underwent hyperparameter tuning to identify the best-performing configuration. We used standard classification metrics such as accuracy, precision, recall, and F1-score to assess performance across all models.

In addition, we monitored consistency between training, validation, and test results to ensure model generalization. Ensemble models, particularly Gradient Boosting and Stacking, consistently outperformed the baselines, which we explore in more detail in the next section.

V. RESULTS AND DESCUSSION

In this section, we present the evaluation outcomes of both baseline and ensemble models, along with the impact of hyperparameter tuning. Metrics such as accuracy, precision, recall, and F1-score were used to assess the models, and confusion matrices and learning curves were analyzed to understand their generalization performance.

Baseline Model Results:

Among the base models, Logistic Regression achieved the highest test accuracy of 0.9945, closely followed by Decision Tree (0.9935). SVM showed comparatively lower performance with 0.9819 accuracy and reduced precision, especially for the minority class. Confusion matrices for Logistic Regression and Decision Tree (Figures X and Y) indicate that most misclassifications occurred in class 2, which represents the most critical cases. However, both models still maintained high recall for all classes.

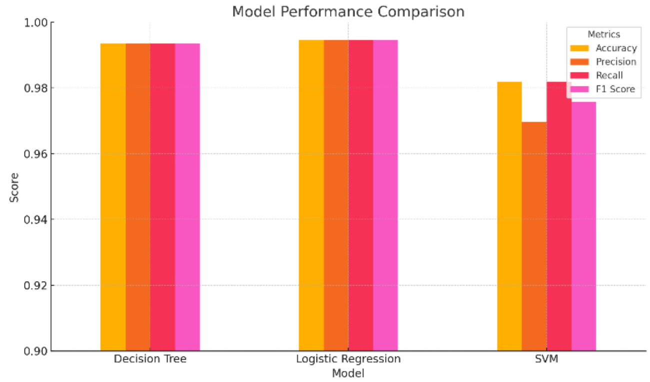

Fig. 2. Performance comparison of baseline models (Decision Tree, Logistic Regression, and SVM) using accuracy, precision, recall, and F1-score.

Learning Curve Analysis:

The learning curves for both Logistic Regression and Decision Tree (Figures A and B) show strong generalization with minimal gap between training and validation scores, indicating that the models are not overfitting. Logistic Regression in particular showed consistently high performance across all training sizes, further confirming its stability.

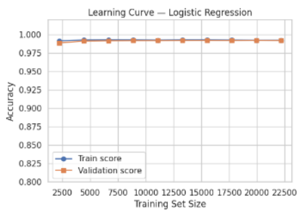

Fig. 3. Learning curve of Logistic Regression showing strong overlap between training and validation scores.

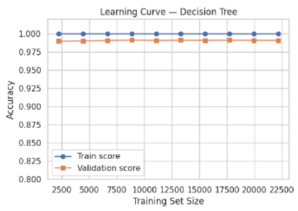

Fig. 4. Learning curve of Decision Tree demonstrating high training accuracy and stable validation performance

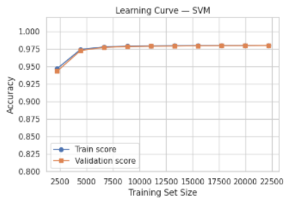

Fig. 5. Learning curve of SVM indicating a slightly wider gap between training and validation.

Feature Importance:

One of the most influential changes in the Final Phase was the creation of the Cause_Reason feature, which was engineered by combining Cause, Reason, and EmergencyArea. This single feature emerged as the top contributor across all models, capturing a richer context of the accident scenario. It played a significant role in boosting model performance, helping ensemble models exceed 99% accuracy without signs of overfitting.

Hyperparameter Tuning Impact:

Hyperparameter tuning led to marginal but meaningful improvements in model scores. As shown in the tuning plots for Logistic Regression (C vs Accuracy) and Decision Tree (max_depth vs Accuracy), the models were optimized for best cross-validation performance. Evaluation charts for all models (Figures C–F) display both pre- and post-tuning comparisons across accuracy, precision, recall, and F1-score.

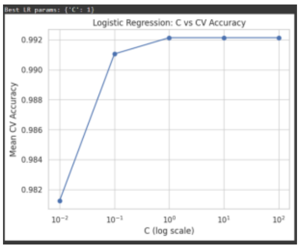

Fig. 6. Logistic Regression hyperparameter tuning showing the effect of regularization strength (C) on cross-validation accuracy.

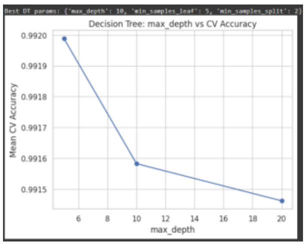

Fig. 7. Decision Tree tuning curve showing how varying max_depth impacts model performance.

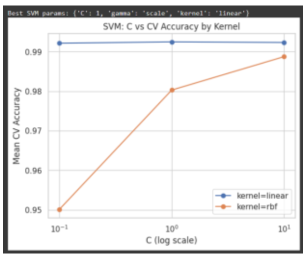

Fig. 8. SVM cross-validation accuracy comparison across different kernel types and regularization strengths.

Ensemble Model Results:

Ensemble methods outperformed base models across the board. Gradient Boosting emerged as the top-performing model with a post-tuning accuracy of 0.9936, followed by Stacking (0.9935) and Voting (0.9934). A detailed ranking is shown below:

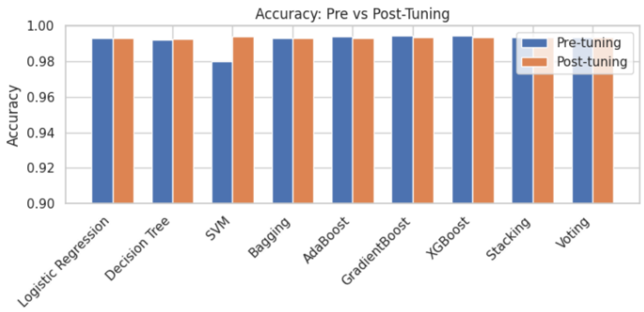

Fig. 9. Accuracy comparison of all models before and after hyperparameter tuning.

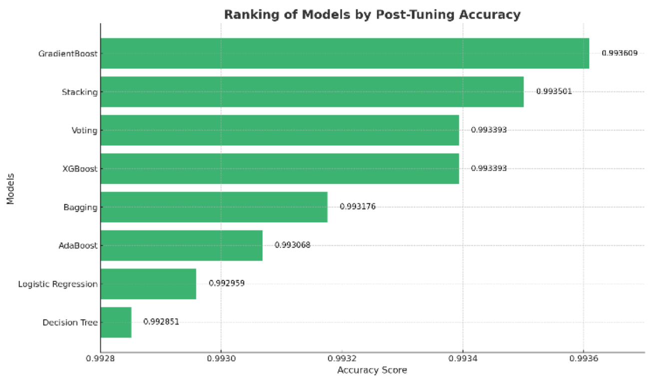

Fig. 10. Accuracy comparison of all models before and after hyperparameter tuning.

Although the improvements may appear small, in large-scale and high-stakes applications such as emergency triage prediction, even fractional gains in recall and precision can translate into significantly better patient outcomes. Additionally, consistent results across models, cross-validation, and confusion matrix evaluations affirm that these models generalize well.

CONCLUSION

This project explored the use of machine learning techniques to predict patient outcomes—categorized as stable, critical, or deceased—following road traffic accidents. By working with a real-world dataset consisting of over 46,000 records, we built and evaluated both baseline models and advanced ensemble methods to assess their effectiveness in this classification task.

In Phase 1, Logistic Regression and Decision Tree models served as strong baselines but struggled with class imbalance, particularly for the deceased category. In the Final Phase, we focused on ensemble techniques such as Gradient Boosting, XGBoost, Stacking, and Voting, all of which significantly improved prediction accuracy and robustness. The creation of the Cause_Reason feature played a key role in enhancing performance across models.

Gradient Boosting emerged as the top performer, achieving over 99% accuracy, with consistent recall and F1-scores across all classes. Despite such high scores, the models showed no signs of overfitting, supported by learning curves and validation metrics. Hyperparameter tuning further refined performance, demonstrating that small adjustments can lead to more stable and generalizable models.

Overall, the study confirms that ensemble learning methods, when combined with thoughtful feature engineering and careful tuning, can serve as powerful tools for predictive modeling in healthcare-related emergency systems. These findings open opportunities for real-time triage support tools that could assist medical responders in prioritizing care more efficiently.

REFERENCES

[1] (Koteswara, 2019) Road Accident Analysis and Prediction using Machine Learning Algorithmic Approaches.

Available: https://pdfs.semanticscholar.org/09a6/19f4f742bdd8b775a6950787f0f6ab41754a.pdf

[2] (Labib, Rifat, Hossain, Das, & Nawrin, 2019) Road Accident Analysis and Prediction of Accident Severity by Using Machine Learning in Bangladesh.

Available: https://ieeexplore.ieee.org/abstract/document/8843640