Crash Severity Prediction Using ExcelFormer, a Transformer-based Model for Tabular Data

***Abstract*—Predicting crash severity is a crucial challenge in ensuring transportation safety. Although standard machine learning models have advanced this task, deep learning approaches have historically underperformed tree-based ensembles on tabular crash data due to their heterogeneity and complex feature interactions. This study evaluates *ExcelFormer*, a modern Transformer-based model designed to overcome the limitations of prior deep learning models on tabular data. By leveraging a Semi-Permeable Attention (SPA) mechanism and feature mixing, *ExcelFormer* effectively models the intricate patterns inherent in tabular data. We have evaluated the performance of *ExcelFormer* using a 10-year traffic crash dataset from Kansas, United States, and compared it against the performance of five tree-based models and the *TabNet* deep learning model. To address the significant class imbalance of the dataset, the models are evaluated using the weighted F1-score, precision, and recall. The results demonstrate that *ExcelFormer* achieves the highest weighted F1-score, outperforming the other models. The model was also able to identify the underrepresented class — fatal or injury crashes — with improved performance. In addition, the model's attention mechanism provides interpretability, highlighting speed limit, crash classification, and road surface type as the most influential predictors. The developed model will support transportation agencies in designing effective traffic safety countermeasures.**

***Index Terms*—Road safety, traffic crash analysis, crash severity prediction, predictive modeling, machine learning, deep learning, tabular data, class imbalance, model comparison, interpretability, tree-based models, transformer models, attention mechanism, *Random Forest*, *Gradient Boosting*, *XGBoost*, *LightGBM*, *CatBoost*, *TabNet*, *ExcelFormer*.**

I. Introduction

Traffic crashes remain a leading public safety concern, causing over 1.19 million deaths annually worldwide, along with millions of non-fatal injuries that carry long-term social and economic burdens [1]. Accurately predicting traffic crash severity remains a critical challenge in transportation safety research, particularly when working with structured crash datasets characterized by categorical, ordinal, and numerical attributes [2]. Given the highly imbalanced nature of crash data, where property-damage-only (PDO) cases vastly outnumber injury-related incidents—binary classification has emerged as a pragmatic strategy. By grouping all injury severities (fatal, serious, and minor) into a single class, this setup simplifies the modeling task while improving classifier sensitivity to injury-related outcomes [3].

To handle the complexity of crash data, researchers have increasingly turned to machine learning (ML) techniques that flexibly capture nonlinear interactions and complex patterns in high-dimensional data. Tree-based methods like *Random Forest* (RF) [4], *Adaptive Boosting (AdaBoost)* [5], *Gradient Boosting (GradientBoost)* [6], *eXtreme Gradient Boosting (XGBoost)* [7], *Light Gradient Boosting Machine (LightGBM)* [8], and *Categorical Boosting (CatBoost)* [9] provide flexibility in identifying nonlinear patterns and handling high-dimensional data without requiring strict assumptions about data distribution. However, these models [4]–[9] face challenges in capturing highly complex and nonlinear interactions or temporal dependencies inherent in crash datasets [10]. Recent studies including one by Zeng et al. [11], have emphasized the need for more expressive modeling techniques in the traffic safety domain.

Deep learning (DL) models are known to capture highly complex, non-linear dependencies by automatically learning complex feature hierarchies, leading to enhanced generalization and pattern recognition capabilities in many application domains, incuding traffic safety [12]. Architectures such as *TabNet* [13] have been developed specifically for tabular data, though empirical results show that they often underperform gradient boosting methods in structured settings [14], [15]. This gap has recently been addressed by *ExcelFormer* [16] a Transformer-based architecture [17] designed for tabular data, which demonstrates superior performance by incorporating semi-permeable attention and column-wise feature mixing.

Tabular datasets combine mixed data types, missing values, and complex feature dependencies—challenges that recent self-supervised, transformer-based models are beginning to address [18]. The ability of the transformer to dynamically weigh inputs shows promise for the prediction of crash severity, where the importance of the features can change by scenario. *ExcelFormer* [16] combines SPA with feature-mixing to dynamically weight inputs, offering a promising alternative to both tree ensembles and earlier deep nets.

The primary objective of this study is to evaluate the effectiveness of the Transformer-based *ExcelFormer* model in predicting crash severity (specifically, fatal/injury crashes versus non-injury crashes) using a comprehensive 10-year crash dataset. We have compared the *ExcelFormer* against established tree-based models, including *RF*, *Gradient Boosting*, *XGBoost*, *LightGBM*, and *CatBoost*, as well as the deep learning model *TabNet*. Specifically, we assess the performance of the models in terms of precision, recall, and F1-score. We hypothesize that *ExcelFormer* will outperform traditional tree-based models and *TabNet* due to its SPA attention and feature mixing techniques, which can effectively capture complex, nonlinear relationships within tabular data. This study aims to investigate the potential of such advanced Transformer-based models in enhancing crash severity prediction and to identify key predictors influencing severe crash outcomes. Through this

objective, we aim to answer the following questions:

- How does *ExcelFormer* perform in predicting crash severity compared to tree-based models and *TabNet*?
- What are the most significant predictors of crash severity identified by our best *ExcelFormer* model?

Our findings will provide valuable insights into the factors influencing crash severity and demonstrate the potential of advanced machine learning models in improving road safety strategies and resource allocation.

II. Related Work

In recent years, non-parametric ML methods have gained traction as an alternative to traditional statistical approaches in traffic safety applications. Non-parametric ML methods capture nonlinear relationships and complex feature interactions within crash data without presuming specific distributions. Studies have shown that ML classifiers such as *RF* and *kNN* outperform Multinomial Rogistic Regression (*MNL*) models in predicting severe crashes, offering improved accuracy on complex datasets [10], [19], [20]. Iranitalab and Khattak [21] demonstrated that k-means clustering and nearest-neighbor classification outperformed traditional statistical methods on heterogeneous crash data. By accommodating interacting variables-traffic volumes, weather conditions, and roadway geometries—these ML approaches offer a greater understanding of crash severity outcomes [10], [22].

Recent research has also explored deep learning (DL) models as a means of advancing crash severity prediction. For instance, Jeon et al. [23] fused a three-layer *CNN* for spatial and kinematic feature extraction with a *RF* classifier to combine deep representations and ensemble decision rules. Fan et al. [24] integrated a class-weighted loss into a multi-layer deep network to counter crash-data imbalance. Wu et al. [25] leveraged Large Short-Term Memory (*LSTM*) architectures to model temporal dependencies in sequential crash records.

Among recent works, Gyawali et al. [14] provide a notable example of ML applied to crash severity prediction. Using ensemble-based models such as *GradientBoost* and *CatBoost*, their study demonstrates the effectiveness of tree-based methods in handling high-dimensional data while maintaining interpretability through explainable AI tools like SHapley Additive exPlanations (SHAP) [26]. Their findings align with broader trends in the field, highlighting the importance of integrating flexible ML models with interpretable frameworks to address the challenges of high-stakes applications.

Transformers [17], which represent a state-of-the-art architecture in deep learning, have shown exceptional performance in many application domains, but traditionally lagged behind ensemble tree-based models when used on tabular data. Several transformer architectures have been proposed in recent years to address this limitation, including *TabNet* [13] and *ExcelFormer* [16]. *TabNet* uses sequential attention to identify salient features with the goal of achieving both high predictive power and inherent interpretability. The *ExcelFormer* model introduces Semi-Permeable Attention (SPA) and feature mixing techniques to dynamically prioritize relevant features. This architecture builds upon the strengths of earlier transformer models, offering a novel approach to addressing the complexities of tabular data prediction, which makes it an ideal candidate for crash severity prediction in the traffic safety domain. While *TabNet* and *ExcelFormer* emphasize leveraging the attention mechanisms for robust feature representation, they contrast with traditional tree-based models, which are inherently interpretable due to their ability to trace decisions along hierarchical splits. In comparison, transformer-based architectures such as *ExcelFormer* and *TabNet* provide no inherent transparency and depend on post hoc explainability methods (e.g., SHAP value analysis or attention-weight visualizations) to clarify their predictions.

Building on prior work on Transformer-based models for tabular data, we compare the *ExcelFormer* model against tree-based methods and *TabNet* on the crash severity prediction task using a 10-year Kansas dataset. By exploiting *ExcelFormer*'s SPA attention, we uncover complex environmental, roadway, and traffic factors, advancing both predictive accuracy and interpretability for targeted safety interventions.

III. Problem Statement

Given that a crash has already occurred, this study addresses the predictability of crash severity, specifically classifying crashes as resulting in fatal or injury crashes versus non-injury crashes using a robust Transformer-based model. Traditional models, including decision tree-based methods and earlier deep learning frameworks such as *TabNet*, face challenges in capturing the complex, nonlinear interactions and heterogeneity inherent in crash severity datasets. These limitations are further exacerbated by the class imbalance typical of such datasets, where severe crashes are rare relative to non-injury incidents. This research leverages a comprehensive 10-year Kansas crash dataset, encompassing diverse variables such as roadway characteristics, traffic characteristics, environmental conditions, traffic factors and temporal patterns, to develop a robust predictive framework. Using a Transformer-based model, *ExcelFormer*, designed to handle tabular data complexities, the study evaluates its performance against established decision tree-based models and *TabNet*. Additionally, it identifies key factors contributing to severe crash outcomes, using attention mechanisms to improve interpretability and predictive accuracy. By addressing these challenges, this study aims to provide actionable insights for data-driven decision-making, ultimately enhancing road safety outcomes and reducing the impact of severe crashes.

IV. Methods

We compare deep learning models and classical tree-based machine learning models to predict crash severity.

A. Deep Learning Models

TabNet [13] utilizes a sequential attention mechanism to dynamically select the most relevant features at each decision step, thereby enhancing interpretability and learning

efficiency. Additionally, *TabNet* supports self-supervised pre-training, which improves performance in scenarios where labeled data is limited.

ExcelFormer [16] was designed to address challenges inherent in tabular data. It incorporates innovations such as the Semi-Permeable Attention (SPA) module, data augmentation techniques (Feat-Mix and Hid-Mix), and an attentive feedforward network.

B. Tree-Based ML Models

Random Forest (RF) [4] is an ensemble method that constructs multiple decision trees using random subsets of data and features. By aggregating predictions from individual trees, *RF* reduces overfitting and improves generalization, making it particularly effective for high-dimensional data.

Adaptive Boosting (AdaBoost) [5] combines weak classifiers iteratively by assigning higher weights to misclassified instances in each iteration. This re-weighting strategy focuses the model on harder-to-classify samples, thereby enhancing overall accuracy.

Gradient Boosting [6] builds decision trees sequentially, with each tree fitting the negative gradient of the loss function to minimize residual errors. This gradient descent approach creates a robust model suitable for both classification and regression tasks, especially on noisy datasets.

eXtreme Gradient Boosting (XGBoost) [7] is an optimized version of *Gradient Boosting* that incorporates regularization (L1 and L2) and parallelization to enhance performance and reduce overfitting. Its sparsity-aware algorithm effectively handles missing data and scales efficiently to large datasets.

Light Gradient Boosting Machine (LightGBM) [8] is optimized for efficiency on large datasets through histogram-based techniques and methods like Gradient-based One-Side Sampling (GOSS) and Exclusive Feature Bundling (EFB). These optimizations accelerate training while maintaining high accuracy, especially for high-dimensional datasets.

Categorical Boosting (CatBoost) [9] specializes in handling categorical data using ordered boosting and a permutation-driven approach to prevent target leakage and reduce overfitting. It delivers robust performance with minimal preprocessing, making it ideal for datasets with mixed categorical and numerical features.

V. Experimental Setup

A. Dataset Description

We utilized a dataset obtained from the Kansas Department of Transportation [27], comprising 36,681 crash reports with 75 features. Several features in the dataset contained missing values. For these features, missing values were imputed using the mode. All other selected features had complete data. Only a subset of 25 features was selected to use in this study. The features selected together with the description and sample values are shown in Table I. The feature selection process and missing value handling adheres to the methodology outlined in [14]. Before running any models, all categorical features were one-hot encoded.

B. Target Variable

The target variable is `CRASH_SEVERITY`, which is binary, indicating *No Injury* or *Fatal/Injury* crashes. This categorization defines the problem as a binary classification task. The class distribution in the dataset used is shown in Figure 1. As can be seen, despite grouping *Fatal* and *Injury* crashes in one class, the distribution is still highly imbalanced towards the *No Injury* class, with 77.8% of the data corresponding to *No Injury*.

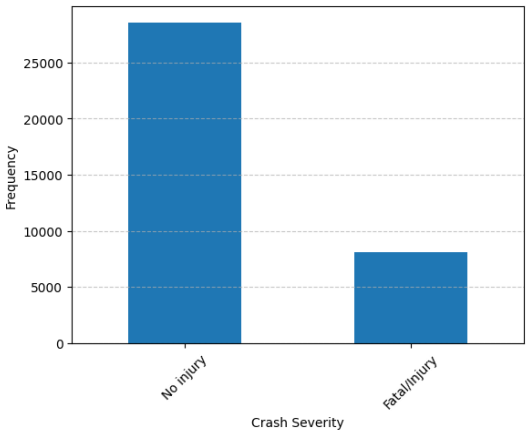


Fig. 1: Class Imbalance of Crash Severity

C. Evaluation Metrics

Given the high class imbalance presented by our dataset, accuracy was not employed as an evaluation metric, as it may be misleading. Instead, we report class-sensitive metrics such as precision, recall, and F1-score to provide a more informative evaluation.

We denote by TP, TN, FP, and FN the true positives, true negatives, false positives, and false negatives, respectively.

Precision is the proportion of true positive predictions among all positive predictions and is defined as:

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

Recall is the proportion of true positive predictions among all actual positive instances and is defined as:

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

The F1-Score is the harmonic mean of precision and recall, providing a balanced metric:

$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

To further account for the class imbalance, we look at the weighted versions of these metrics. This approach computes the metric (Precision, Recall, or F1-score) for each class (c) individually and then calculates a weighted average. The weight for each class (w_c) is its proportion of the total number of samples, defined as:

#	Feature Name	Description	Examples
1	ON_ROAD_KDOT_TYPE	Type of road where the situation began	AVE (Avenue), BLVD (Boulevard), etc.
2	AT_ROAD_KDOT_TYPE	Type of road where the crash occurred	AVE, BLVD, PKWY (Parkway), etc.
3	ON_ROAD_SPEED_LIMIT	Speed limit (in mph) of the road where the situation began	55 mph, 65 mph, 75 mph, etc.
4	ON_ROAD_SURFACE_TYPE	Surface type of the road where the situation began	Concrete, Gravel, Dirt, etc.
5	ON_ROAD_SURFACE_COND	Surface condition of the road	Dry, Wet, Snow, etc.
6	ON_ROAD_SURFACE_CHAR	Surface characteristics of the road	Straight & Level, Curved & Level
7	ON_ROAD_CONST_MAINT	Type of construction or maintenance on the roadway	None, Construction Zone, etc.
8	ON_ROAD_NBR_OF_LANES	Number of lanes on the road	1, 2, 3, etc.
9	CRASH_CLASS_FHE	Classification of the first harmful event (FHE)	Pedestrian, Railway train, etc.
10	CRASH_CLASS_MHE	Classification of the most harmful event (MHE)	Pedestrian, Pedal cyclist, etc.
11	COLLISION_W_OTHER_VEH_FHE	Type of collision with the first harmful event (FHE)	Head-on, Rear end, etc.
12	COLLISION_W_OTHER_VEH_MHE	Type of collision with the most harmful event (MHE)	Head-on, Backed into, etc.
13	CRASH_LOCATION	Location on the roadway where the crash occurred	Non-intersection, Toll Plaza, etc.
14	LIGHT_CONDITIONS	Lighting condition at the time of the crash	Daylight, Dawn, Dusk, etc.
15	WEATHER_CONDITIONS	Weather condition at the time of the crash	No adverse conditions, Fog, etc.
16	KDOT_WORK_ZONE_FLAG	Indicates if the crash occurred in a work zone	Yes/No
17	KDOT_PROP_DAMAGE_FLAG1	Property damage occurred (first object damaged)	Yes/No
18	KDOT_PROP_DAMAGE_FLAG2	Property damage occurred (second object damaged)	Yes/No
19	SPECIAL_JURISDICTION	Crash occurred in a special jurisdiction	Normal Jurisdiction, Military, etc.
20	DUI_FLAG	Driver under the influence of alcohol or drugs	Yes/No
21	ALCOHOL_INVOLVEMENT	Alcohol involvement in the crash	Yes/No
22	DRUG_INVOLVEMENT	Drug involvement in the crash	Yes/No
23	MONTH_OF_CRASH	Month when the crash occurred	1, 2, 10, etc.
24	DAY_OF_WEEK	Day of the week when the crash occurred	Sunday, Monday, etc.
25	HOUR_OF_CRASH	Hour of the day when the crash occurred	3, 5, 15, 23, etc.

TABLE I: Selected Feature Descriptions

$$w_c = \frac{\text{Number of samples in class } c}{\text{Total number of samples}}$$

The final weighted scores are calculated with the following formulas:

$$\text{Weighted Precision} = \sum_c w_c \times \text{Precision}_c$$

$$\text{Weighted Recall} = \sum_c w_c \times \text{Recall}_c$$

$$\text{Weighted F1-Score} = \sum_c w_c \times \text{F1-Score}_c$$

In addition to reporting performance in terms of weighted metrics, we also report the metrics independently for each class, *Fatal/Injury* and *No Injury*, as this allows us to understand how well the models predict the minority class *Fatal/Injury* versus the majority class *No Injury*.

D. Data Splitting

The dataset was divided into three subsets: 70% for training, 15% for validation, and 15% for testing. The validation split was utilized to tune hyperparameters, while the test split was reserved for the final evaluation of the models.

E. Hyperparameter Optimization

To optimize the performance of our models, we employed Optuna [28], a sophisticated hyperparameter optimization framework that leverages techniques such as Bayesian optimization to efficiently explore the hyperparameter space. For each model, we conducted a hyperparameter search consisting of 100 trials with the objective of maximizing the F1-Score on the validation split. The specific hyperparameters tuned, their respective search spaces, and the selected optimal values are detailed in Table II. The best hyperparameters identified were subsequently used to train the final model on the training split before evaluating its performance on the test split.

Fig. 2: Training and validation loss curves for the final *ExcelFormer* model.

VI. Results

The training and validation loss curves (Fig. 2) indicate that the *ExcelFormer* model converged well, with no signs of overfitting. The early stopping mechanism ensured that training halted when no further improvements were observed, enhancing generalization to the test split. We also chose the best epoch on the validation split to evaluate on the test split.

Tables III, IV, and V present the performance metrics of all models (with their best hyperparameters) evaluated on the test split. Tables III and IV show the metrics for the

TABLE II: Hyperparameter Search Space and Selected Optimal Values

Model	Hyperparameter	Search Space	Selected Value
RF	`n_estimators`	50 to 300 (step: 50)	200
	`max_depth`	5 to 50 (step: 5)	25
	`criterion`	{`gini`, `entropy`}	`gini`
AdaBoost	`n_estimators`	50 to 300 (step: 50)	100
	`learning_rate`	Log-uniform from 0.001 to 0.1	0.1
Gradient Boosting	`n_estimators`	50 to 300 (step: 50)	200
	`learning_rate`	Log-uniform from 0.001 to 0.1	0.1
	`max_depth`	3 to 20	5
XGBoost	`n_estimators`	50 to 300 (step: 50)	200
	`learning_rate`	Log-uniform from 0.001 to 0.1	0.0589
	`max_depth`	3 to 20	8
LightGBM	`n_estimators`	50 to 300 (step: 50)	300
	`learning_rate`	Log-uniform from 0.001 to 0.1	0.0848
	`max_depth`	5 to 50 (step: 5)	30
CatBoost	`n_estimators`	50 to 300 (step: 50)	300
	`learning_rate`	Log-uniform from 0.001 to 0.1	0.0804
	`max_depth`	3 to 10	5
TabNet	`batch_size`	{128, 256, 512, 1024}	1024
	`learning_rate`	Log-uniform from 10^{-4} to 10^{-2}	0.00482
	`n_d`	8 to 64 (step: 8)	40
	`n_a`	8 to 64 (step: 8)	56
	`n_steps`	3 to 10	5
	`n_independent`	1 to 5	1
	`n_shared`	1 to 5	2
	`gamma`	1.0 to 2.0	1.2042
	`momentum`	0.01 to 0.4	0.3085
	`lambda_sparse`	Log-uniform from 10^{-5} to 10^{-3}	0.0004187
	`step_size`	{10, 20, 30, 40, 50}	40
	`sched_gamma`	0.7 to 0.99	0.9065
	`mask_type`	{`sparsemax`, `entmax`}	`entmax`
ExcelFormer	`batch_size`	{128, 256, 512, 1024}	128
	`learning_rate`	Log-uniform from 10^{-5} to 10^{-2}	0.0021
	`n_layers`	{4, 6, 8}	6
	`n_heads`	{4, 8, 16}	8
	`d_token`	{64, 128, 256}	64
	`attention_dropout`	Uniform from 0.0 to 0.3	0.2362
	`ffn_dropout`	Uniform from 0.0 to 0.3	0.1485
	`residual_dropout`	Uniform from 0.0 to 0.3	0.0001

Fatal/Injury and *No Injury* classes, respectively, while Table V shows the weighted metrics over the two classes. Showing the metrics for each class and as a weighted average allows us to provide a comprehensive assessment of each model's global classification capabilities, as well as their ability of predicting the minority *Fatal/Injury* class.

Model	Precision	Recall	F1-Score
RF	0.6280	0.2151	0.3204
AdaBoost	0.0000	0.0000	0.0000
Gradient Boosting	0.6018	0.2735	0.3761
XGBoost	0.6197	0.2711	0.3772
LightGBM	0.6074	0.2914	0.3939
CatBoost	**0.6284**	0.2443	0.3518
TabNet	0.5508	0.2946	0.3839
ExcelFormer	0.5378	**0.3693**	**0.4379**

TABLE III: Performance on *Fatal/Injury* Class

As shown in the tables, model performance varies significantly by class for all models. Notably, *AdaBoost* completely failed to identify the minority *Fatal/Injury* class, rendering

Model	Precision	Recall	F1-Score
RF	0.8097	0.9632	0.8798
AdaBoost	0.7761	**1.0000**	0.8740
Gradient Boosting	0.8189	0.9478	0.8787
XGBoost	0.8191	0.9520	0.8806
LightGBM	0.8223	0.9457	0.8797
CatBoost	0.8147	0.9583	**0.8807**
TabNet	0.8206	0.9307	0.8722
ExcelFormer	**0.8332**	0.9085	0.8692

TABLE IV: Performance on *No Injury* Class

Model	Precision	Recall	F1-Score
RF	0.7690	0.7957	0.7546
AdaBoost	0.6024	0.7761	0.6783
Gradient Boosting	0.7703	0.7968	0.7662
XGBoost	**0.7744**	**0.7996**	0.7679
LightGBM	0.7742	0.7992	0.7709
CatBoost	0.7730	0.7985	0.7623
TabNet	0.7602	0.7883	0.7629
ExcelFormer	0.7670	0.7878	**0.7726**

TABLE V: Weighted Average Performance

it ineffective for this imbalanced classification task. For the critical *Fatal/Injury* class, *ExcelFormer* achieves the highest Recall and F1-Score. Considering the weighted average metrics, although *XGBoost* showed the highest weighted precision and recall, *ExcelFormer* achieved the best overall performance with the highest F1-Score. To better understand the predictions of the models, the normalized confusion matrices for all evaluated models are presented in Fig. 3.

Fig. 4 illustrates the top ten most important features as determined by the attention weights from the *ExcelFormer* model. To investigate the relationships among the most influential predictors, a correlation matrix of the top five most important features was computed. Fig. 5 presents this matrix, offering insights into potential multicollinearity. The matrix shows that the most notable relationship is a moderate positive correlation of 0.35 between `ON_ROAD_SURFACE_TYPE` and `ON_ROAD_SURFACE_COND`.

VII. Discussion

Historically, tree-based models like *RF* and *XGBoost* have outperformed deep learning models on tabular data. Despite the success of deep learning in domains such as image and text processing, tabular data presents unique challenges, including heterogeneous feature types, smaller sample sizes, and irregular patterns. Tree-based models excel due to their inherent inductive biases, such as robustness to uninformative features, ability to learn irregular functions, and non-reliance on rotation invariance. In contrast, neural networks often require extensive hyperparameter tuning and regularization to achieve competitive performance on tabular datasets.

This trend was observed in the crash severity prediction task, where *TabNet* shows considerably lower performance compared to tree-based models, particularly in F1-score and precision metrics [14]. However, *ExcelFormer* [16] is designed to address these challenges inherent in tabular data. It incorporates innovations such as the Semi-Permeable Attention (SPA) module, data augmentation techniques (Feat-Mix and Hid-Mix), and an attentive feedforward network. These enhancements enable *ExcelFormer* to effectively capture irregular feature-target relationships, and crucially for this study, its attention-based architecture allows for the analysis of attention weights to identify the most influential features contributing to high crash severity. Given this context, the results of this study demonstrate the advantages of applying *ExcelFormer*, to the crash severity prediction task, confirming its potential to surpass not only other deep learning approaches but traditional tree-based models as well.

First, *ExcelFormer* achieved the highest F1-Score among all tested models, outperforming tree-based models such as *RF*, *XGBoost*, *LightGBM*, and *CatBoost*, as well as the Transformer-based *TabNet*. This result underscores its ability to handle the complexities of crash severity data, particularly in distinguishing between the critical categories of *Fatal/Injury* and *No Injury*.

While tree-based models can be interpreted through SHAP values and *TabNet* through its own attention mechanisms, the superior performance of *ExcelFormer* suggests that the insights derived from its Semi-Permeable Attention (SPA) module may be more reliable and informative. The attention mechanism within this module identified features such as `ON_ROAD_SPEED_LIMIT`, `CRASH_CLASS_FHE`, and `ROAD_SURFACE_TYPE` as critical to crash severity prediction. These results align with domain knowledge, emphasizing the importance of speed limits, crash type, and road conditions in influencing crash outcomes. By dynamically weighing feature relevance, *ExcelFormer* not only excels in prediction accuracy but also provides actionable insights for targeted interventions.

A notable insight from the correlation matrix is that the road surface type (`ON_ROAD_SURFACE_TYPE`) and the road surface condition (`ON_ROAD_SURFACE_COND`) have a moderate correlation. This suggests that different road types are affected differently by factors like precipitation. Having precise, quantitative insights into this relationship can help with road planning.

The findings have significant implications for road safety. Policymakers can leverage the model's insights to prioritize safety measures, such as enforcing speed limits and improving road infrastructure. By focusing on the high-risk factors identified by *ExcelFormer*, resources can be effectively allocated to reduce the likelihood of severe crashes.

VIII. Conclusions and Future Work

In conclusion, *ExcelFormer* establishes itself as the most effective model in this study for crash severity prediction, outperforming both tree-based and other Transformer-based models. Its strong performance highlights the potential of advanced machine learning methods to inform road safety policies and reduce crash-related fatalities and injuries.

Future work can expand on these findings by performing further analyses of attention weights, particularly investigating interactions between feature pairs or groups to uncover more nuanced relationships. For instance, studying how combinations like speed limits and road surface conditions interact can provide deeper insights into crash dynamics. Additionally, exploring hybrid approaches that combine *ExcelFormer* with ensemble techniques could refine predictions and further boost performance.

References

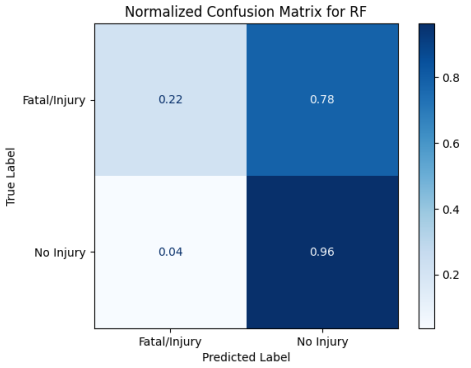
[1] World Health Organization, "Road traffic injuries: Fact sheets," 2023, accessed: 2025-01-09. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/road-traffic-injuries

[2] Y. Ali, F. Hussain, and M. M. Haquev, "Advances, challenges, and future research needs in machine learning-based crash prediction models: A systematic review," *Accident Analysis & Prevention*, vol. 194, no. 107378, 2024.

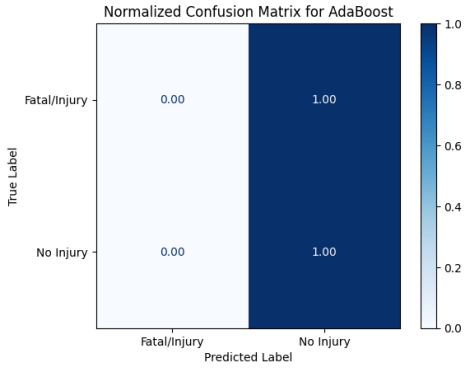
[3] S. Kim, Y. Lym, and K.-J. Kim, "Developing crash severity model handling class imbalance and implementing ordered nature: Focusing on elderly drivers," *Int. J. Environ. Res. Public. Health*, vol. 18, no. 4, 2021.

[4] L. Breiman, "Random forests," *Machine Learning*, vol. 45, pp. 5–32, 2001.

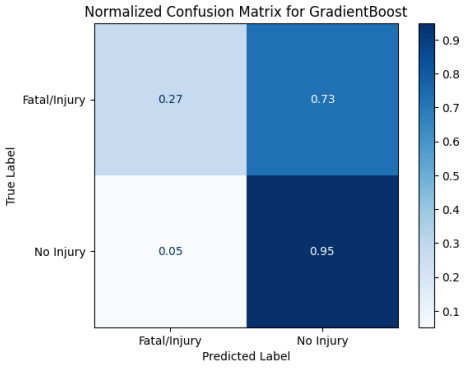
[5] Y. Freund and R. E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," *Journal of Computer and System Sciences*, pp. 119–139, 1997.



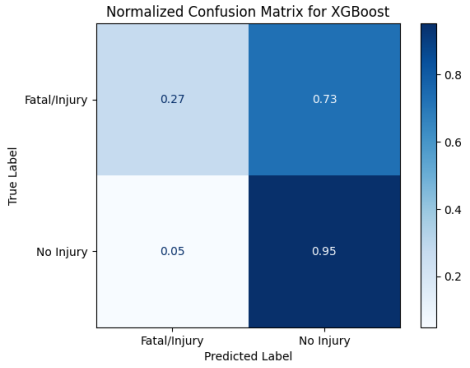
(a) *RF*



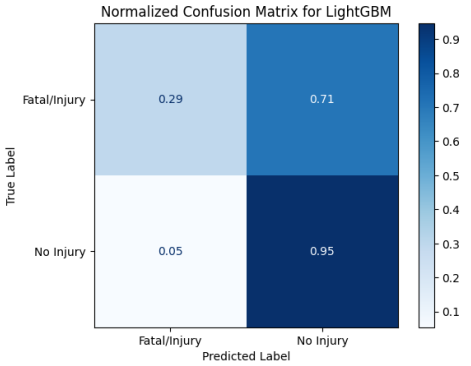
(b) *AdaBoost*



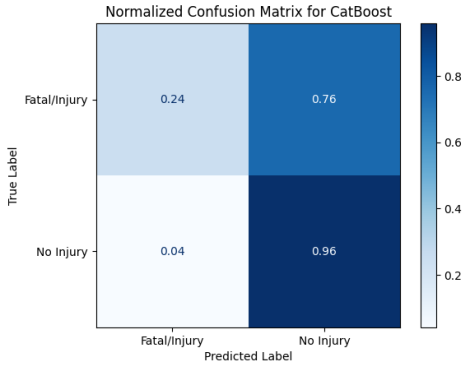
(c) *Gradient Boosting*



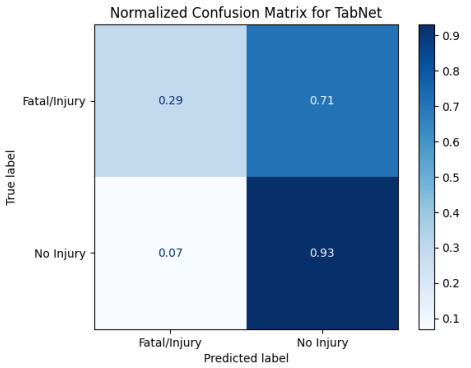
(d) *XGBoost*



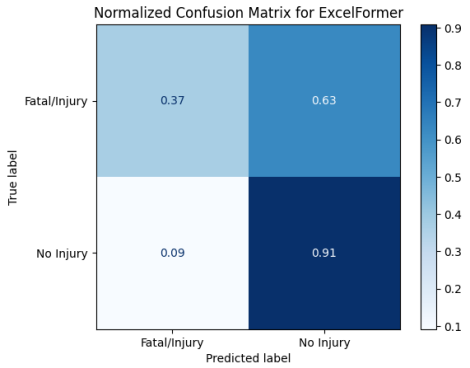
(e) *LightGBM*



(f) *CatBoost*



(g) *TabNet*



(h) *ExcelFormer*

Fig. 3: Normalized confusion matrices for all evaluated models on the test split.

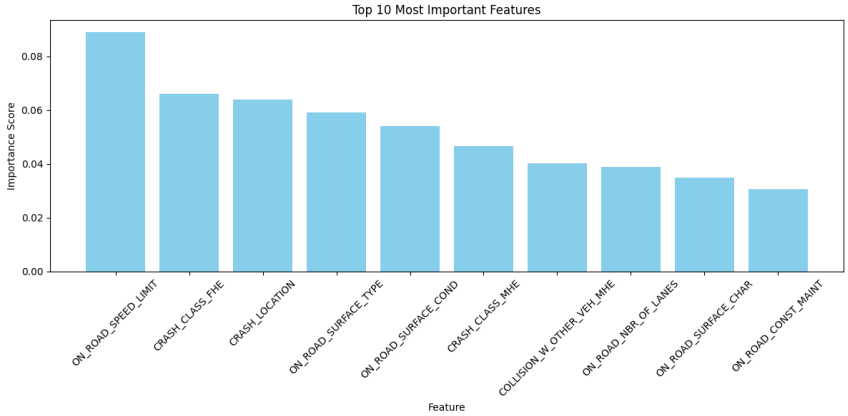


Fig. 4: Top ten most important features based on attention weights from the *ExcelFormer* model

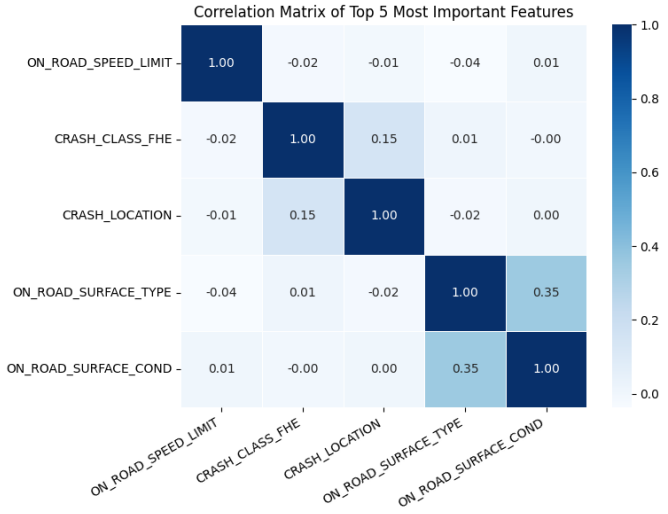


Fig. 5: Correlation matrix of the top five most important features.

[6] J. H. Friedman, "Greedy function approximation: A gradient boosting machine," *Annals of Statistics*, vol. 29, no. 5, pp. 1189–1232, 2001.

[7] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 785–794.

[8] G. Ke *et al.*, "Lightgbm: A highly efficient gradient boosting decision tree," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 2017, pp. 3146–3154.

[9] A. V. Dorogush, V. Ershov, and A. Gulin, "Catboost: gradient boosting with categorical features support," in *Proceedings of the Workshop on Handling Unstructured and Structured Data in the Same Database (co-located with KDD 2018)*, 2018.

[10] A. Iranitalab and A. Khattak, "Comparison of four statistical and machine learning methods for crash severity prediction," *Accident Analysis & Prevention*, vol. 108, pp. 27–36, 2017.

[11] Q. Zeng, H. Wen, and H. Huang, "Traffic crash injury severity analysis: A review and outlook," *Accident Analysis & Prevention*, vol. 98, pp. 29–38, 2017.

[12] M. Chakraborty, T. Gates, and S. Sinha, "Causal analysis and classification of traffic crash injury severity using machine learning algorithms," *Data Science for Transportation*, vol. 5, p. 3, 2023.

[13] S. Ö. Arik and T. Pfister, "Tabnet: Attentive interpretable tabular learning," *CoRR*, vol. abs/1908.07442, 2019. [Online]. Available: http://arxiv.org/abs/1908.07442

[14] N. Gyawali, S. Khanal, D. Caragea, H. M. A. Aziz, and E. J. Fitzsimmons, "Predicting commercial motor vehicle crash severity in kansas using explainable machine learning," in *The 103rd Transportation Research Board (TRB) Annual Meeting, Advances in Truck and Bus Safety Research*, Washington DC, 2024.

[15] L. Grinsztajn, E. Oyallon, and G. Varoquaux, "Why do tree-based models still outperform deep learning on tabular data?" *ArXiv Preprint*, 2022.

[16] J. Chen, J. Yan, Q. Chen, D. Z. Chen, J. Wu, and J. Sun, "Excelformer: A neural network surpassing gbdts on tabular data," in *Proceedings of the 30th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2024.

[17] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[18] W. Ren, T. Zhao, Y. Huang, and V. Honavar, "Deep learning within tabular data: Foundations, challenges, advances and future directions," 2025.

[19] J. Zhang, Z. Li, Z. Pu, and C. Xu, "Comparing prediction performance for crash injury severity among various machine learning and statistical methods," *IEEE Access*, vol. 6, pp. 60 079–60 087, 2018.

[20] T. G. Meghna Chakraborty, Shakir Mahmud, "Analysis of trends and correlation in child restraint use and seating position of child passengers in motor vehicles: Application of a bivariate probit model," 2021.

[21] A. Iranitalab and A. J. Khattak, "Comparison of four statistical and machine learning methods for crash severity prediction," *Accident Analysis & Prevention*, vol. 108, pp. 27–36, 2017.

[22] M. Chakraborty, H. Singh, P. T. Savolainen, and T. J. Gates, "Examining correlation and trends in seatbelt use among occupants of the same vehicle using a bivariate probit model," *Transportation Research Record*, vol. 2675, no. 7, pp. 288–298, 2021.

[23] D. Jeon, S. Ryu, E.-H. Choi, and S. Min, "Rfcnn: Random forest-based convolutional neural network for road accident severity prediction," in *2021 IEEE Intelligent Transportation Systems Conference (ITSC)*. Indianapolis, IN, USA: IEEE, 2021, pp. 1994–1999. [Online]. Available: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9536744

[24] Y. Fan, Y. Zhang, and Y. Liu, "Deep learning-based vehicle crash severity prediction: A comparative study on imbalanced datasets," *Journal of Transportation Safety & Security*, 2020.

[25] K. Wu, J.-E. Kang, and J. Zhu, "Deep learning approaches for traffic crash severity prediction: A comparative study," *Transportation Research Record*, vol. 2676, no. 7, pp. 640–654, 2022.

[26] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," *Advances in neural information processing systems*, vol. 30, 2017.

[27] Kansas Department of Transportation, "Crash statistics," https://www.ksdot.gov/burtransplan/prodinfo/accista15.asp, 2024.

[28] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, "Optuna: A next-generation hyperparameter optimization framework," *CoRR*, vol. abs/1907.10902, 2019. [Online]. Available: http://arxiv.org/abs/1907.10902