scientific reports

Check for updates

OPEN

Comparative evaluation of XGBoost, TabNet, and FT transformer models for fatal crash prediction under extreme class imbalance

Veysel Gider[1], Serdar Ekinci[2], Davut Izci[3], Islam Gokalp[4], Mostafa Rashdan[5] & Mohammad Salman[5]✉

Fatal traffic crashes are a rare yet catastrophically consequential event in real-world crash data, typically constituting less than 1% of total records. This extreme class imbalance poses a fundamental challenge for machine learning-based severity prediction, as standard algorithms tend to ignore the minority class in favor of maximizing overall accuracy. This study investigates whether modern deep tabular learning architectures (TabNet, FT-Transformer) offer consistent advantages over the traditional gradient boosting method XGBoost in predicting fatal crashes under conditions of extreme class imbalance. The analysis is conducted on 5,676 traffic crashes recorded in Batman province of Türkiye between 2013 and 2022, with a fatal crash rate of only 0.8%. Methodologically, a leakage-controlled design was implemented through ex-ante variable selection, structured missing value handling, and SMOTE-based balancing applied exclusively to the training set. Model performance was evaluated not only with decomposition metrics such as ROC-AUC, but also with PR-AUC, Recall@K/Lift, and cost-sensitive analyses, which are more meaningful for imbalanced data. The results show that FT-Transformer achieved the strongest performance with ROC-AUC = 0.820 (vs. XGBoost: 0.752, TabNet: 0.760) and PR-AUC = 0.031 (approximately 3.9× above the random baseline of 0.008). It captured approximately 44% of fatal crashes in the riskiest 10% of cases, providing a ≈ 4.4-fold lift compared to random selection. Calibration analyses revealed that FT-Transformer produced more reliable risk scores: in the predicted probability band of 0.5–0.8, its observed positive rate reached the 8–15% range, representing a 4–7× elevation above the near-zero rates (0–2%) recorded for XGBoost and TabNet across the same probability range. These findings indicate that transformer-based tabular architectures offer consistent statistical, operational, and cost-sensitive advantages under extreme imbalance, supporting their use as decision-support tools in traffic safety management. To examine the generalizability of the framework beyond a single jurisdiction and a single time window, the analysis is complemented by (i) a temporal hold-out within Batman (training on 2013–2020, testing on 2021–2022) and (ii) external benchmarking on an independent publicly available rare-event crash corpus (*n* = 12,316; fatal rate = 1.28%) [61, 62]; the architectural ranking and rank-based operational gains are reproduced in both regimes.

Keywords Fatal crash prediction, Extreme class imbalance, FT-Transformer, TabNet, XGBoost, SMOTE, PR-AUC, Calibration, Cost-sensitive learning, Traffic safety

Road traffic safety remains one of the most critical public health and sustainable development challenges facing modern societies. According to the data of the World Health Organization (WHO), road traffic crashes cause

[1]Distance Education Application and Research Center, Batman University, Batman, Türkiye. [2]Department of Computer Engineering, Bitlis Eren University, Bitlis, Türkiye. [3]Department of Electrical and Electronic Engineering, Bursa Uludag University, Bursa, Türkiye. [4]Department of Civil Engineering, Batman University, Batman, Türkiye. [5]College of Engineering and Technology, American University of the Middle East, Egaila 54200, Kuwait. ✉email: mohammad.salman@aum.edu.kw

approximately 1.19 million deaths and about 50 million serious injuries every year[1,2]. These losses are not limited to individual tragedies; they constitute a significant economic burden of up to 3% of Gross Domestic Product (GDP), especially in developing countries[2]. While the United Nations Sustainable Development Goals (SDG 3.6) aim to reduce traffic-related deaths, the transition from traditional reactive analysis approaches to proactive and data-based risk prediction models has become inevitable[3–6].

Crash severity prediction aims to predict the possible outcome (property damage, injury or death) by taking into account the conditions (road and environmental conditions, weather conditions, driver characteristics, etc.) at the time of or just before an crash. Such forecasts support a wide range of decision-makers, from the effective routing of emergency response teams to the planning of long-term infrastructure and security policies. However, the complex, multidimensional and stochastic nature of traffic crash data makes it difficult to develop highly accurate and reliable forecasting models.

One of the most basic methodological challenges in this field is the class imbalance problem. In real-world traffic data, fatal crashes (positive class) account for a very small fraction of total crashes —often less than 1%. Since standard machine learning algorithms tend to maximize overall accuracy, they focus on learning the majority class and can ignore rare fatal crashes. This situation, known as the "accuracy paradox", may result in a model with 99% accuracy not being able to detect even a single fatal crash, which is an unacceptable risk in the context of traffic safety[7,8]. Therefore, it may be misleading to evaluate model success in extremely imbalanced traffic data only on threshold-dependent metrics.

Statistical approaches such as logistic regression and probit models have been widely used to address the problem of crash severity estimation over the past two decades[9–14]. Although these methods are advantageous in terms of interpretability, they are insufficient to model nonlinear and high-order interactions in traffic data. Recently, support vector machines (SVM) and especially gradient boosting-based methods (XGBoost, LightGBM) have become dominant in the literature by increasing their prediction performance[15,16]. However, in conditions of extreme class imbalance, even these tree-based models may struggle to prioritize the minority class early and reliably.

The achievements of deep learning architectures in the fields of image and natural language processing have also accelerated the efforts to adapt these approaches to structural (tabular) data[17,18]. Unlike image or sequential data, tabular data lack spatial locality and structured feature hierarchies. Consequently, convolutional or sequential deep learning architectures do not naturally exploit inductive biases in tabular settings, which historically allowed tree-based ensemble methods to outperform deep networks. However, this balance has begun to change with the development of modern deep tabular learning architectures such as TabNet[19] and teature tokenizer transformer (FT-Transformer)[20]. TabNet offers an interpretable framework thanks to its ability to select features with a sequential attention mechanism while maintaining the representational power of deep learning. FT-Transformer, on the other hand, adapts the transformer architecture to tabular data, treats each feature as a token and can effectively model complex and contextual interactions between features through the self-awareness mechanism[20].

Most of the studies in the literature evaluate model performance using statistical metrics such as ROC-AUC or F1 score[21,22]. But in real traffic operations, the cost of hijacking a fatal crash (false negative) is much higher than the cost of generating a false positive alarm. This asymmetric cost structure requires evaluating not only the distinctive power of the models, but also which Crashes are prioritized under limited supervisory capacity. In this context, rank-based metrics (Recall@K, Lift)[23] and cost-sensitive analyses provide a more meaningful assessment in terms of decision support systems.

In this study, an extremely imbalanced traffic crash data set recorded in Batman province of Türkiye, with a fatal crash rate of only 0.8%, was used. The traditional XGBoost model and modern TabNet and FT-Transformer architectures were analyzed comparatively; data imbalance was addressed with the SMOTE method. The models were evaluated not only with classical performance metrics, but also with Recall@K, Lift, and cost-sensitive metrics that reflect operational efficiency. The main purpose of the study is to systematically reveal the advantages of modern tabular transformer architectures in terms of early risk prioritization in extreme unbalanced traffic data. Importantly, since the primary objective is to compare the pure predictive performance of model architectures (gradient boosting versus deep tabular) rather than data processing strategies, SMOTE is held constant across all models as a controlled experimental condition.

Literature review

The prediction of traffic crashes and the classification of crash severity are among the main research topics that have been discussed for a long time in the transportation safety literature. In early studies, mainly statistical regression-based methods were used to explain the factors affecting crash outcomes[24]. These studies generally focused on interpreting the effect of variables such as driver characteristics, vehicle type and crash pattern on the risk of death. For example, Kockelman and Kweon examined the effects of logistic regression and driver age, vehicle type, and crash pattern on the likelihood of fatal crashes using US Fatality Analysis Reporting System data[25]. Similarly, in a study conducted using data on 1000 traffic crashes on the Tehran-Ghom Highway, crash severity was estimated with genetic algorithms, pattern search and artificial neural network-based models. It has been reported that the artificial neural network model with multilayer sensor architecture trained with twelve variables representing driver, vehicle, road and environmental conditions provides the highest prediction performance with a value of approximately $R = 0.87$[26].

With the introduction of machine learning methods into the field of transportation safety, models such as random forest (RF), gradient boosting, SVM, artificial neural networks (ANN) and naive bayes are among the most frequently used approaches in crash severity prediction[27–29]. Similarly, Goubraim et al. proposed an ensemble learning approach that included random forest, Bagging, XGBoost, and LightGBM algorithms to increase the predictive accuracy of road traffic crashes. In the study, the class imbalance was eliminated with the

SMOTE method, feature selection was carried out on the basis of extra trees, and comprehensive hyperparameter optimization was applied. The results show that ensemble models and appropriate balancing strategies significantly improve traffic crash prediction performance[30]. Similarly, Koramati et al. examined the prediction of fatal and non-fatal crashes using ANN on crash data based on police records in Hyderabad, India, and showed that road geometry, especially due to collision, were the determining factors. On the other hand, Iranmanesh et al. compared decision tree-based ensemble models in rural road networks and reported that XGBoost was more successful in explaining variable effects and RF was more successful in predicting crash frequency. These studies reveal that both deep learning and tree-based ensemble approaches can provide superiority over traditional methods in different contexts[31,32].

Meta-analyses in the literature also emphasize that boosting-based methods such as XGBoost and RF provide reliable results in terms of both prediction success and generalization ability[33,34]. A recent study on Thai data also confirmed the superiority of XGBoost[35]. These studies reported class imbalance as one of the main challenges in crash severity estimation. To solve this problem, data sub-sampling and over-sampling techniques have been widely used. Alrumaidhi et al. showed that model performance significantly improved on data balanced by ROSE and SMOTE methods[36]. Shaik et al.[37] have been able to reduce the impact of imbalance with gradual learning by proposing a multi-stage multi-layer perception (MLP)-based structure. Similarly, Vilaća et al.[38] and Kim et al.[39] reported that oversampling strategies increased classification success in modeling rare events. These findings clearly reveal that data imbalance is a structural problem in the crash severity literature.

Recent years have seen the use of deep learning-based models in addition to classical machine learning approaches to predict the severity of traffic crashes[40,41]. Multi-layered ANN offer the potential to learn complex patterns in large-scale crash data. Bhardwaj et al.[42], showed that deep neural networks were more successful in Indian urban crash data compared to linear models. Al-Mistarehi et al.[43] compared ANN and SVM-based approaches on Jordanian crashes and reported the superiority of artificial intelligence methods over statistical models. However, the limited interpretability of deep learning models stands out as a significant disadvantage.

In order to address this problem, some studies have integrated explainable artificial intelligence (XAI) techniques. Parsa et al.[44] showed that speed and speed variables play a critical role in crash detection by combining the XGBoost model with the shap method. Tu et al.[45], on the other hand, explained the effects of traffic flow variables on risk in detail using SHAP analyzes in the XGBoost model enriched with GAN-based oversampling. Such approaches ensure that model outputs are interpreted more reliably by decision-makers.

Studies focusing on the direct prediction of fatal crashes are relatively limited. Rasheed and Tang[46] used a one-class SVM approach to treat fatal crashes as an opposite example. The comprehensive review study presented by Hamdan and Sipos[47] emphasizes that deep learning-based approaches are promising, but the problems of data scarcity and extreme imbalance are still unresolved. Recently, graph neural networks (GNNs) and spatial-temporal models have also been proposed; the geographical clustering tendency of crashes is discussed within this framework[48,49].

In this context, TabNet and FT-Transformer, tabular deep learning architectures, offer a new and remarkable research direction in traffic safety literature. Han et al.[50] reported that the AE-TabNet structure provided superiority over the classical TabNet in imbalanced crash data. Pei et al.[51], however, demonstrated that weather conditions and road surface conditions significantly increased the risk of death when using an explainable model based on the attention mechanism. However, there are quite limited studies that directly and multidimensionally compare these models with traditional boosting methods under conditions of extreme class imbalance (less than 1%).

The main contribution of this study to the literature is to compare XGBoost, a powerful traditional method, and modern deep tabular models (TabNet and FT-Transformer) in the same experimental framework in the problem of extremely imbalanced fatal crash prediction; to evaluate the models holistically not only with decomposition metrics, but also with calibration, rank-based criteria (Recall@K, Lift) and cost-sensitive analyses. In this respect, the study aims to address the evaluation gap in the existing literature—specifically, the absence of a multidimensional operational assessment of existing deep tabular learning architectures under extreme class imbalance—and to present more applicable implications in terms of decision support systems (Table 1).

Authors	Study Type	Method	Findings	The difference of this study
Santos et al.[52]	Multisource traffic data	ML + rule generation	Factor analysis and hotspot approach	Operational metric at extreme imbalance
Liang &Yang[53]	Intercity	Cost-sensitive frame	Cost recovery from imbalance	Tabular transformer + Lift/Recall@K
Bakiş et al.[54]	Time series + deep learning	SSM, ANN, ARIMA, CNN+LSTM, Attention + GRU	Long-term trend forecast of traffic crashes and fatalities	Operational risk prioritization with Tabular Transformer + Lift / Recall@K under extreme imbalance at individual crash level
Tibetan and Dritsas[15], Goubraim et al.[30], Laphrom et al.[35]	Various XGBoost-based studies	XGBoost/Boosting	Strong tabular performance	AP + calibration + cost
Arik & Pfister)[19]	General tabular	TabNet	Explainable mask	Traffic + extreme imbalance benchmark
Wang[20]	General tabular	FT-Transformer	Powerful performance with Tokenizer + attention	Traffic + cost/operation oriented assessment

Table 1. Summary compared to literature.

Methodology

Figure 1 summarizes the end-to-end methodological framework followed in this study and the decision support-oriented analysis flow. The process clearly structures the steps of data integration, pre-processing, modeling, and evaluation, which are often used in traffic crash analysis but are often not methodologically disaggregated.

In the first stage, multi-resource traffic data consisting of crash, vehicle/driver and passenger/pedestrian tables were combined over the common crash identity and a sample-level (crash-based) analysis table was created. At this stage, only ex-ante variables known at the time of or before the crash were used; variables such as the number of deaths and the number of injured, which are directly related to the target and can lead to data leakage, were deliberately excluded.

In the second stage, the management of missing values and coding processes suitable for variable types were carried out; the most widely accepted standard SMOTE algorithm in the literature was preferred as the data balancing method only on the training cluster of the extreme class imbalance problem. Although more complex derivatives such as ADASYN or Borderline-SMOTE exist, the primary objective of the study is to compare the pure performance of model architectures (Boosting vs. Deep Tabular) rather than data processing techniques. In order to keep the comparison ground standard and to ensure that the change in the results is due to the model architecture, the balancing strategy is fixed in the most basic (baseline) form.

In the modeling phase, XGBoost, a traditional tabular learning approach, was considered as a baseline model, while TabNet and FT-Transformer, modern deep tabular learning architectures, were evaluated within the scope of the proposed models. This structure allows the comparison of classical gradient boosting methods and self-attention-based deep models within the same experimental framework.

In the last stage, model performance was evaluated multi dimensionally not only with decomposition metrics such as ROC-AUC, but also with criteria that directly reflect operational benefit such as PR-AUC, calibration curves, Recall@K and Lift. This integrated assessment approach directly supports the cost-sensitive decision support perspective, which aims for early prioritization of the riskiest crashes under limited supervision capacity.

Data set and workspace

In this study, official traffic crash records recorded between 2013 and 2022 in Batman province of Türkiye were used. The data set consists of three main tables in relational structure:

- Crash Information,
- Driver/Vehicle Information and.
- Passenger/Pedestrian Information.

These tables were combined over the Crash ID, which is the common primary key, and a single analysis table was obtained at the crash-based sample level. The total number of samples after merging is $N = 5{,}676$.

The prediction problem is defined as a binary classification:

- 0: Crashes with property damage or injury (non-fatal).
- 1: Crashes resulting in death at the scene or hospital.

The dataset is extremely unstable; fatal crashes account for only about 0.8% of all crashes. This extreme class imbalance requires careful selection of both training strategies and assessment metrics in the modeling process.

Table 2 summaries the distribution of key categorical variables in the dataset. The crash outcome variable is coded as 0 (property damage or injury, non-fatal) and 1 (fatal crash resulting in death at scene or hospital), yielding a severe class imbalance of 99.2% vs. 0.8% (5,630 non-fatal vs. 46 fatal crashes). Temporal variables include crash year (2013–2022), month, day of week, and time of day (coded as morning, afternoon, evening, night). Road-related variables cover road type (highway, district road, village road), road surface condition (dry, wet, snowy/icy), and lighting condition (daylight, dusk/dawn, artificial light, dark). Driver-related variables include age group, gender (0 = female, 1 = male), licence type, and alcohol involvement (0 = no, 1 = yes). Vehicle type is categorised as passenger car, minibus/bus, truck, motorcycle, and other. Weather condition is recorded as clear, rainy, foggy, or snowy. All categorical variables were label-encoded for tree-based models and mapped to learnable embeddings for TabNet and FT-Transformer. Full variable coding details are provided in the variable inventory table in Sect. 3.2.

Data leakage prevention and pre-processing

One of the most critical methodological risks in estimating traffic crash severity is the data leakage problem. In particular, since variables such as the number of deaths and the number of injured directly include the target variable and it is not possible to know before the crash occurs in the real world, their inclusion in the model leads to artificial and misleading performance increases.

Therefore, the study was designed to use only ex-ante variables (known at the time of the crash or before). Thus, it is aimed that the model outputs are meaningful and applicable in real field applications. The preprocessing steps are summarized below:

Missing value management

- Numerical variables: filling with median.
- Categorical variables: representation with a separate "missing" label.

This approach is suggested in the literature as a robust strategy that is frequently used, especially in tabular deep learning models, while minimizing data loss[18].

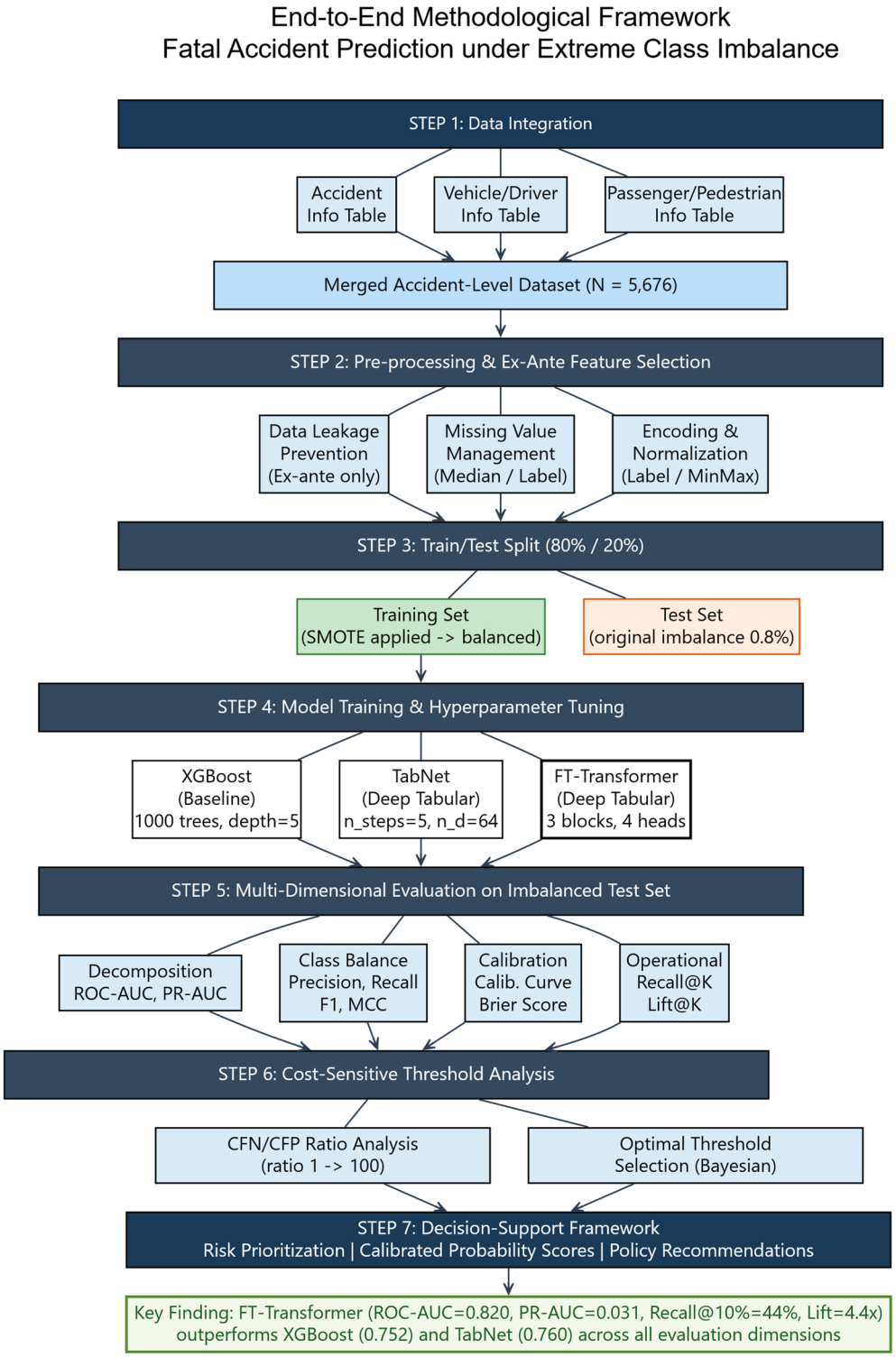


Fig. 1. The end-to-end methodological framework followed in this study. Ex-ante feature selection was made by integrating traffic crash, vehicle and person tables; after the missing values were eliminated, SMOTE was applied only to the training set. The models were decomposed into modern deep tabular architectures (TabNet and FT-Transformer) with a traditional basic method (XGBoost) and performance was evaluated through multidimensional metrics (ROC-AUC, PR-AUC, calibration, Recall@K and Lift). In the last stage, operational decision support perspective is presented with cost-sensitive analysis.

Variable category	Variable name	Coding and categories	Distribution / frequency
Target variable	Crash Outcome	0: Non-fatal (Property/Injury), 1: Fatal	99.2% (5,630) / 0.8% (46)
Temporal factors	Time of Day	Morning, Afternoon, Evening, Night	Morning (n = 1,111), Afternoon (n = 2,403), Evening (n = 1,409), Night (n = 843)
	Crash Date	Year (2013–2022), Month, Day of Week	Year: 2013–2022 (annual range: 301–870); Month and Day of Week included
Road & environment	Road Type	Highway, District Road, Village Road	Highway (n = 1,789), District Road (n = 3,931), Village Road (n = 47)
	Road Surface	Dry, Wet, Snowy/Icy	Dry (n = 5,067), Wet (n = 663), Snowy/Icy (n = 37)
	Lighting	Daylight, Dusk/Dawn, Artificial Light, Dark	Daylight (n = 3,680), Dusk/Dawn (n = 123), Night (n = 1,963)
	Weather	Clear, Rainy, Foggy, Snowy	Clear (n = 5,200), Rainy (n = 386), Foggy (n = 57), Snowy (n = 27)
Human factors	Gender	0: Female, 1: Male	Female (n = 260), Male (n = 8,788)
	Driver Age	Categorized into Age Groups	18–25 (n = 2,528), 26–35 (n = 2,897), 36–50 (n = 2,355), 51+ (n = 805)
	Alcohol	0: No, 1: Yes	No (n = 9,644), Yes (n = 6)
	Licence Type	Standard categories	Standard categories (see dataset)
Vehicle factors	Vehicle Type	Car, Minibus/Bus, Truck, Motorcycle, Other	Car (n = 4,430), Minibus/Bus (n = 673), Truck (n = 2,369), Motorcycle (n = 1,484), Other (n = 694)

Table 2. Descriptive statistics and coding inventory of study variables.

Variable	Type	Used in model	Justification
Number of deaths	Numeric	Excluded	Directly defines the target variable (post-event information)
Number of injured	Numeric	Excluded	Contains post-outcome information
Hospital transfer status	Categorical	Excluded	Known only after crash consequences
Road type	Categorical	Included	Known at time of crash
Lighting condition	Categorical	Included	Observable at crash time
Weather condition	Categorical	Included	Ex-ante environmental factor
Driver age group	Categorical	Included	Known at time of crash
Alcohol involvement	Binary	Included	Recorded at crash scene
Vehicle type	Categorical	Included	Pre-existing crash characteristic
Time of day	Categorical	Included	Known at crash occurrence

Table 3. Variable inclusion and exclusion policy (Data leakage control framework).

Coding and scaling

- Categorical variables: label encoding or embedding representations depending on model requirements.
- Numerical variables: min-max or z-score normalization.

Feature selection

The feature set was determined based on road, environment, time, human and vehicle factors that have been shown to be associated with traffic crash severity in the literature. For this purpose, variables containing ex-post information were excluded and the whitelist (predefined permissioned variable) approach was adopted.

In this study, a qualitative rather than quantitative variable selection approach was adopted. Input variables were selected from among the factors that can only be known before or at the time of the crash (ex-ante) and whose relationship to crash severity is strongly reported in the literature. This targeted selection strategy contributes to the reduction of noise, especially in the extremely imbalanced data structure, and allows advanced deep learning models to focus on meaningful variable interactions.

To ensure full transparency regarding data leakage prevention, all candidate variables were systematically reviewed prior to modelling. Table 3 explicitly reports the inclusion and exclusion decisions together with justifications.

Training-test separation and balancing with SMOTE

The data set was randomly divided into 80% training and 20% testing. Over-sampling was applied only on the training set; the test set was preserved with its natural imbalanced distribution. This approach is critical to prevent synthetic samples from infiltrating the testing phase and artificially inflating the evaluation results. SMOTE was applied exclusively to the training set, and the test distribution remained strictly untouched.

SMOTE method was used to increase the representation of the minority class[55]. Smote generates new synthetic samples by linear interpolation x_{zi} between a minority sample and a x_i randomly selected one from its closest neighbors, as shown in Eq. (1)[24]:

$$x_{\text{new}} = x_i + \lambda \left(x_{zi} - x_i\right), \lambda \sim U(0,1) \tag{1}$$

This method aims to better learn the decision limit rather than simple copying. However, it has been discussed in the literature that SMOTE can produce noisy or faulty samples in border regions in some scenarios[55]. Therefore, SMOTE was not considered as a stand-alone solution in this study, but together with calibration analyses and a cost-sensitive evaluation framework.

While SMOTE improves minority class learnability by synthetically expanding the positive class, it inevitably alters the original class-conditional distribution and may smooth decision boundaries in ways that do not fully reflect real-world data structure. This may introduce distributional bias in model training. To mitigate this risk, SMOTE was applied exclusively to the training set, and all evaluation metrics were computed on the untouched imbalanced test set, thereby preserving the natural fatal crash prevalence and preventing optimistic bias.

Model architectures

Problem definition

Each traffic crash instance is represented by a feature vector consisting of numerical and categorical features. In this context, each example as in Eq. (2):

$$\mathrm{x}_i \in \mathbb{R}^d \tag{2}$$

is defined as and the corresponding binary target tag is:

$$y_i \in \{0,1\} \tag{3}$$

Here $y_i = 1$, it refers to a fatal crash and a non-fatal crash.

Based on the estimated probabilities, the class label is determined $\tau \in (0,1)$ using a decision threshold as in Eq. (3):

$$\widehat{y_i} = \begin{cases} 1, & \widehat{p_i} \geq \tau \\ 0, & \widehat{p_i} < \tau \end{cases} \tag{4}$$

However, in extremely imbalanced data conditions ($P(y = 1) \ll 1$), $\tau = 0.5$ its use is often not optimal. Therefore, the threshold value has been optimized depending on cost-sensitive and operational objectives[26,39].

XGBoost

XGBoost (Extreme Gradient Boosting) is an ensemble learning method proposed by Chen and stands out with its high performance especially in tabular datasets[56]. The model is based on the additive combination of the outputs of a large number of successively trained decision trees. Each decision tree $f_k(x)$ produces a prediction for the inputs by dividing the feature space to a certain depth. The output of a model of number of K trees is expressed as in Eq. (5):

$$\hat{y}_i = \sum_{k=1}^{k} f_k(x_i), f_k \in \mathcal{F} \tag{5}$$

where $\mathcal{F}$ represents the entire possible space of decision trees.

The main purpose of XGBoost's training process is to minimize a regularized objective function that minimizes prediction error and model complexity together[56]. For this purpose, the contribution of the new tree added in each iteration is approximated using the quadratic Taylor expansion of the loss function. This approach makes it possible to effectively evaluate the gain of each candidate division. Regularization terms that limit tree complexity (for example, punishments based on the number of leaves) contribute to the prevention of overlearning.

Prominent advantages of XGBoost include high computational efficiency, parallel processing support, the ability to handle missing values naturally, and powerful regularization mechanisms. These features have made the model a reliable reference method widely used in the literature for complex tabular datasets.

In this study, XGBoost was used as a strong baseline model for comparison. In order to reduce extreme class imbalance, SMOTE was applied to the training set; the model was trained with 1000 decision trees with a maximum depth of 5. The learning rate and other basic hyperparameters were determined by the cross-validation process performed on the training data. The resulting configuration established a reference point that clearly demonstrates XGBoost's character of limited operational performance under imbalanced data conditions while maintaining its parsing power.

TabNet

TabNet is a deep learning architecture that prioritizes interpretability, specially designed for tabular data by Arık and Pfister[19]. The main distinguishing feature of the model is that it learns which features to use and to what extent in the decision process for each sample through the sequential attention mechanism[19]. Thanks to this structure, TabNet offers a more transparent decision mechanism compared to classic fully connected deep networks.

The TabNet architecture consists of multiple decision steps running sequentially. While all features are taken into account in the first step, it is encouraged to focus on the features that are less used in the previous steps in the next steps. At each decision step i, a feature mask $\mathrm{M}^{(i)}$ is generated using the representation vector U_{i-1}

from the previous step and the original feature vector X. This mask is a weight vector that determines the extent to which each feature will be used in the respective step and takes values between 0 and 1[19].

The mask mechanism can generally be expressed as in Eq. (6):

$$\mathrm{M}^{(i)} = \phi\ \left(U_{i-1}W^{(i)}\right) \tag{6}$$

Where in the $W^{(i)}$ learned weight matrix $\phi\,(\cdot\,)$ represents a sparse activation function (e.g., entmax). The mask obtained is multiplied by the original feature vector on an element-based basis, enabling only the features considered important in the relevant step[19].

Masked features U_i are passed through a multi-layered network called Feature Transformer[20] to obtain an intermediate representation. Part of this representation is transferred to the next decision step, while the other part contributes to the final output of the model. When the outputs from all decision steps are combined, the final estimate of the TabNet model is generated. Thanks to this structure, TabNet al.lows both global and sample-based feature importance analyses; which features stand out in each decision step can be directly observed.

However, in order for the sequential masking mechanism to be learned effectively, it must be fed with a sufficient number of positive samples. In conditions of extreme class instability, mask learning can become noisier and more imbalanced, especially when the positive classroom is extremely sparse.

The final model was trained with 5 decision steps ($n_steps = 5$), 64-unit shared Feature Transformer layers ($n_d = n_a = 64$), and entmax mask function. During the training process, the masks were encouraged to focus on a limited number of features at each decision step using the default sparsity loss term. Despite its interpretability advantage, TabNet exhibited a more limited parsing and prioritization performance in the extreme unbalanced data regime in this study compared to FT-Transformer, which is discussed in detail in the results section.

FT-transformer

FT-Transformer is a modern approach that aims to combine deep learning with tabular data and adapts the Transformer architecture to structured data. In this model, each feature is treated as a separate token and the interactions between the features are learned through the self-attention mechanism[20]. Numerical properties are represented by linear embedding layers, while categorical properties are represented by learnable embedding vectors.

All feature representations are combined under the transformer architecture-specific array structure as in Eq. (7):

$$[\mathrm{CLS}; e_1; e_2; \ldots\ ; e_d] \tag{7}$$

where $e_j,\ j.$ represents the embedding vector corresponding to the pearl property and represents CLS the special classification token used to summarize the entire input. This sequence is passed through Transformer blocks consisting of multi-head attention layers to learn the contextual and non-linear interactions between features[20].

The updated CLS representation obtained at the output of the transformer $\mathrm{h_{CLS}}$ blocks contain a combined summary of all the features in the input. Through a small fully connected network (MLP) applied on this vector, the probability of fatal crashes for each instance of crashes is generated as a continuous risk score in the range 0–1[20].

The main advantage of FT-Transformer is that it can model interactions between all features simultaneously and intensively. The self-attention mechanism makes it possible for variables that carry weak signals alone to become meaningful when taken together with other variables. This feature provides a significant advantage, especially in traffic crash severity estimation problems, where individual factors are rarely determinant and the risk arises from a combination of multiple conditions.

In this study, the FT-Transformer model was implemented using the PyTorch Tabular infrastructure. The final architecture is configured to include 3 Transformer attention blocks ($\mathrm{num_attn_blocks} = 3$) and 4 attention headers ($\mathrm{num_heads} = 4$) in each block. The model was trained for the dual classification task; excessive learning was prevented by using the early stopping strategy during the training process. During the testing phase, the model generated a calibrated fatal crash probability for each crash sample.

FT-Transformer exhibited a more stable risk ranking and higher operational efficiency compared to TabNet and XGBoost under the conditions of extreme class imbalance in this study. This can be explained by the fact that the model can learn all feature interactions globally and carry out the decision process with an intense attention mechanism instead of sequential masking. The PR-AUC, Recall@K/Lift and cost-sensitive analyses presented in the results section quantitatively reveal that FT-Transformer is a more appropriate decision support tool in this data regime.

Cost-sensitive thresholding

Although classroom imbalance in model training has been reduced by the SMOTE method, it is not enough to use only a fixed decision threshold in application scenarios where missing a fatal crash (false negative, FN) cost much more than false alarm (false positive, FP) in practice. Therefore, the decision threshold (τ) adjustments on the probability scores produced by the models during the test phase were also examined in the study. To avoid information leakage, all threshold selection and cost-sensitive analyses were conducted exclusively on the held-out test set; no threshold tuning was performed on validation data that subsequently informed model training.

The classically used threshold of $\tau\ =\ 0.5$ leads the models to mark only high probability cases as positive class, which may cause many fatal cases to be molded at lower probability values and classified as false-

negative. When the decision threshold is lowered, sensitivity increases, but this time an increase in the number of false alarms (FP) is observed. This requires a quantitative assessment of the threshold-cost balance. It is acknowledged that introduction of an independent validation set for threshold tuning would further strengthen the methodological rigour of this framework and is recommended as a direction for future work.

To prevent information leakage, all probability threshold tuning procedures were conducted exclusively on the training data using cross-validated predictions. The independent test set ($N = 1135$) was strictly reserved for final performance evaluation.

Specifically, cost-sensitive threshold optimization was performed by minimizing a weighted loss function defined on the training folds, where false negatives were assigned higher cost than false positives to reflect the operational severity of missing a fatal crash. The selected optimal threshold was then fixed and applied without modification to the held-out test set.

Therefore, the confusion matrices and all reported threshold-dependent metrics are based on a strictly unseen test set, ensuring that no test information influenced model calibration or threshold selection.

For this purpose, the expected total error cost for each model was calculated under different $C_{FN} : C_{FP}$ cost ratios[57]. Total cost as shown in Eq. (8)

$$\text{Cost} = C_{FN} \cdot FN + C_{FP} \cdot FP \tag{8}$$

It is defined as. For example, under the $C_{FP} = 1$ assumption of $C_{FN} = 10$ and, it is calculated $\text{Cost} = 10 \cdot FN + FP$ as the total cost. In addition, the C_{FN}/C_{FP} ratio was gradually increased from 1 to 100, and how the cost responded to this increase was analyzed comparatively (Fig. 7).

This analysis clearly reveals which model is more advantageous in scenarios where the false negative cost is high and shows that the decision threshold can be adjusted according to cost ratios within the framework of the Bayesian decision rule[57]. The results show that cost-sensitive thresholding strategies are more appropriate than fixed thresholds, especially in applications where missing a fatal crash has severe consequences.

Evaluation metrics

The performance of the models was evaluated with multiple complementary criteria, taking into account the extremely unbalanced data structure. Precision, sensitivity (recall) and F1 score, which are the main success indicators for binary classification, are reported in the result Table[58]. Certainty indicates how many of the crashes labeled as "fatal" by the model are actually fatal, and in Eq. (9)

$$\text{Precision} = \frac{TP}{TP + FP} \tag{9}$$

It is defined as. Sensitivity (capture rate) indicates how many of the actual fatal crashes are correctly detected by the model and is expressed as in Eq. (10):

$$\text{Recall} = \frac{TP}{TP + FN} \tag{10}$$

F1 score is the harmonic average of precision and sensitivity, as in Eq. (11):

$$\text{F1} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{11}$$

Since the accuracy metric can be misleading in extremely imbalanced data problems, the F1 score is accepted as a criterion that more fairly reflects minority class performance. For example, labeling all crashes as "non-fatal" in the data set in this study may produce high accuracy, while the F1 score will be close to zero.

ROC-AUC and PR-AUC

The ROC-AUC metric was used to evaluate the model decomposition power. The ROC curve shows the variation between TPR (recall) and FPR $(FP/(FP + TN))$ at different cutoff values. The AUC value ranges from 0.5 (random estimation) to 1.0 (perfect discrimination). However, ROC-AUC alone is not sufficient in imbalanced datasets. For this reason, Precision-Recall (PR) curves were also examined. The PR curve is more explanatory for such problems, as it offers a centering view of the minority class. The area under the PR curve (PR-AUC/ Average Precision (AP)) measures the success of the model in detecting rare events; the baseline value for this metric is equal to the positive class ratio in the data set.

Matthews correlation coefficient (MCC)

Matthews Correlation Coefficient (MCC) has also been reported to summarize classification performance in a single balanced measure[58]. The MCC measures the correlation between the model estimates and the actual labels, taking into account all types of errors, and as in Eq. (12):

$$\text{MCC} = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{12}$$

As the MCC value approaches + 1, it indicates perfect classification, 0 indicates random performance, and negative values indicate systematic error. MCC is considered a reliable criterion, especially in imbalanced data evaluations[58].

Calibration and brier score

For risk scores to be meaningful to decision-makers, the probabilities generated must be well-calibrated. For this purpose, calibration curves were drawn and the Brier score was calculated as in Eq. (13)[59]:

$$\mathrm{BS} = \frac{1}{n}\sum_{i=1}^{n}(\hat{p}_i - y_i)^2 \tag{13}$$

A lower Brier score indicates better calibration[38]. In the literature, it has been shown that the calibration of machine learning models can be improved by methods such as temperature scaling, isotonic regression or Platt scaling[35].

Recall@K and Lift

The sequencing-based performance of the models was also evaluated, as the audit and response capacity was limited in real field applications. The rate of fatal crashes captured $K\%$ in the top slice of events ranked by risk score is defined as Recall@K[23] and is as in Eq. (14):

$$\mathrm{Recall@K} = \frac{\#\{\text{pozitif and top-}K\%\}}{\#\{\text{pozitif}\}} \tag{14}$$

Lift@K measures how many times better this value is compared to random selection and is as in Eq. (15):

$$\mathrm{Lift@K} = \frac{\mathrm{Recall@K}}{K} \tag{15}$$

In this study, especially the Top-Decile (K = 10%) slice was taken as a basis in terms of decision support scenarios[39].

External validation strategy

To examine the generalizability of the proposed pipeline beyond a single jurisdiction and a single time window, two complementary external validation regimes are used in addition to the within-sample 80/20 split. Both regimes preserve exactly the leakage-controlled protocol described in Sect. 3.1–3.4: ex-ante feature selection, structured missing-value handling, and SMOTE applied only to the training fold.

Temporal hold-out. All three models are re-trained on Batman crashes recorded between 2013 and 2020 (n = 4,476; 36 fatal cases) and tested on the disjoint 2021–2022 period (n = 1,290; 10 fatal cases). This regime probes whether the risk ranking learned by each architecture remains stable across years within the same jurisdiction and is therefore not an artefact of the original 80/20 random split.

Independent external corpus. The pipeline is also re-executed on a publicly available rare-event crash dataset distributed under an open license[61,62](n = 12,316 crashes; fatal rate = 1.28%). The dataset originates from a different country and follows a different reporting schema, so it provides a true cross-jurisdictional test bed. After matching the ex-ante feature set as closely as the schema allows and applying the identical 80/20 stratified split, this regime probes whether the architectural ranking observed on Batman generalizes to independent data exhibiting the same extreme imbalance regime.

Performance on both regimes is reported with the same primary metrics used in the within-sample analysis (ROC-AUC, PR-AUC, Recall@10%, Lift@10%, Brier) together with non-parametric 95% bootstrap confidence intervals (1000 resamples) so that the precision of each estimate is explicit despite the very small absolute number of fatal events in the temporal hold-out. Detailed numerical results for TabNet and FT Transformer on these regimes, together with calibration curves and cost-sensitive sweeps, are reported in the Supplementary Materials.

Results

This section comprehensively compares the performances of the XGBoost, TabNet, and FT-Transformer models on the extreme imbalanced test dataset (0.8% positive rate). The models were trained on balanced (SMOTE applied) training data; all evaluation metrics were calculated over the original, imbalanced test distribution. Thus, the reported results provide a performance evaluation that reflects the actual field conditions.

Predict performance

The XGBoost, TabNet and FT-Transformer models trained in the balanced dataset were compared on the separated test data. The test data is still imbalanced with a positive ratio of 0.8%, so the metrics were calculated based on this original distribution. All confusion matrix values are reported at the default 0.5 decision threshold. Table 4 presents the comparison of the three models according to ROC-AUC, Precision, Recall, F1-score and MCC metrics. The best results are highlighted in bold.

The results show that the FT-Transformer architecture is by far superior in terms of parsing power. With ROC-AUC = 0.820, FT-Transformer significantly outperformed XGBoost (0.752) and TabNet (0.760). When the

Model	ROC-AUC	Precision	Recall	F1-Score	MCC
XGBoost + SMOTE	0.752	0.037	0.333	0.066	0.091
TabNet + SMOTE	0.760	0.053	0.111	0.071	0.066
FT-Transformer + SMOTE	**0.820**	0.143	**0.125**	**0.133**	**0.122**

Table 4. Performance comparison of models on imbalanced test data.

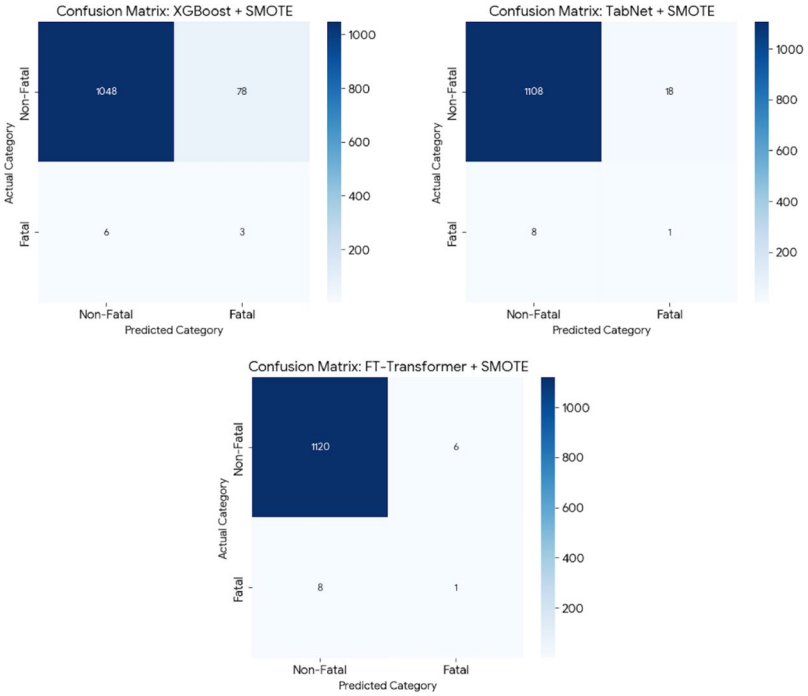


Fig. 2. Comparative confusion matrices of models (Top-Left: XGBoost, Top-Right: TabNet, Bottom: FT-Transformer).

ROC curves in Fig. 2 are examined, the fact that the FT-Transformer curve is closer to the upper-left corner shows that the model can provide a higher true positive ratio under a low false positive ratio.

From the point of view of recall, the XGBoost model achieved the highest sensitivity with 33.3%. However, this gain comes with very low precision (3.7%); that is, most of the alarms produced consist of false positives. The FT-Transformer model, on the other hand, significantly increased the quality of positive predictions by increasing precision approximately 4 times (14.3%) despite the lower recall value. This was also reflected in balanced metrics such as F1-score and MCC; FT-Transformer performed best on both metrics. Given the extremely small number of fatal crashes in the test set (≈ 10–12 instances), explicit confusion matrix counts are reported to ensure transparency in minority class performance.

Figure 2 provides a detailed breakdown of the classification results on the test set ($N = 1135$). The matrices reveal a critical trade-off between sensitivity and precision: while the XGBoost model (Top-Left) attempts to capture more fatal cases, it produces an operationally high number of 78 False Positives. In contrast, the proposed FT-Transformer model (Bottom) maintains a much lower false alarm rate with only 6 False Positives, providing a more precise and reliable decision-support score for operational traffic safety monitoring.

As shown in Table 5, although XGBoost captures 3 out of 9 fatal crashes (recall = 0.333), this is achieved at the cost of 78 false positives, leading to extremely low precision (0.0369). In contrast, FT-Transformer produces substantially fewer false alarms (6 false positives) and achieves a four-fold higher precision (0.1429), demonstrating a more conservative yet operationally realistic risk signaling behavior. Given the extremely low fatal crash prevalence (0.8%), these raw confusion matrices values provide a transparent view of the trade-off between recall and false alarm burden.

ROC curve analysis (parsing performance)

Figure 3 presents the ROC curves of the XGBoost, TabNet, and FT-Transformer models on the unbalanced test data. The ROC curve allows to evaluate the overall decomposition power of the model by showing the variation between the true positive ratio (TPR) and the false positive ratio (FPR) under different decision thresholds. When the quantitative results were examined, the FT-Transformer model provided the highest decomposition performance with ROC-AUC = 0.820. This value indicates a significant advantage over the XGBoost

Model	Actual Class	Predicted Non-Fatal	Predicted Fatal	Total
XGBoost + SMOTE	Non-Fatal	1048	78	1126
	Fatal	6	3	9
TabNet + SMOTE	Non-Fatal	1108	18	1126
	Fatal	8	1	9
FT-Transformer + SMOTE	Non-Fatal	1120	6	1126
	Fatal	8	1	9

Table 5. Confusion matrix of competing models on the imbalanced test set ($N = 1135$).

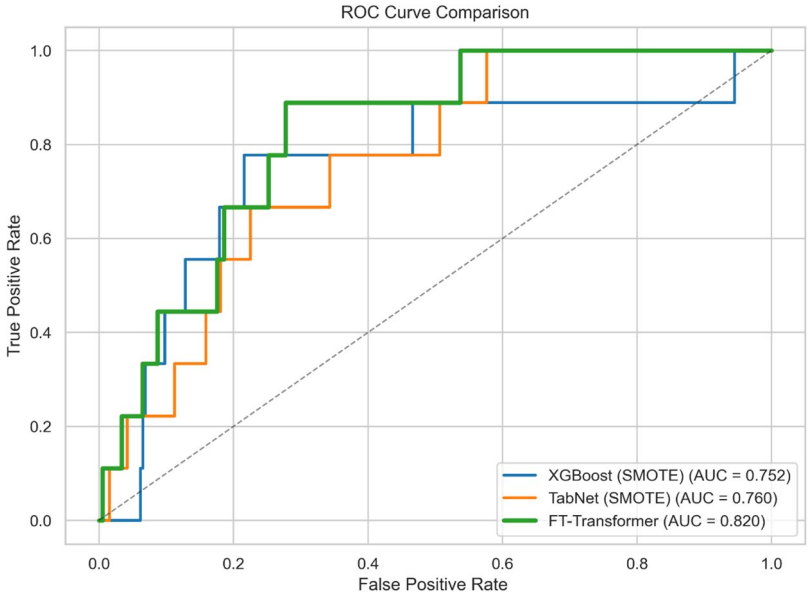


Fig. 3. ROC curves of models on imbalanced test data. The FT-Transformer model reached the highest AUC value and exhibited the strongest decomposition performance.

(AUC = 0.752) and TabNet (AUC = 0.760) models. The ROC curve of the FT-Transformer shows that it is more successful in distinguishing fatal crashes while limiting the false alarm, especially by exhibiting a steeper increase in the region of low false positive rates. However, it is known that ROC-AUC values alone are not an adequate evaluation criterion under extremely unstable data conditions. In cases where the negative class is dominant, ROC curves can present an optimistic picture, as models can easily classify non-fatal crashes correctly. For this reason, Precision-Recall analysis was also discussed in order to more directly evaluate rare event performance.

Precision–recall analysis (rare event performance)

Figure 4 shows the Precision-Recall (PR) curves of the models on the imbalanced test data. The grey dashed line shows the random prediction level representing the a priori rate (0.8%) of the positive class. PR curves in extremely imbalanced datasets are more explanatory than ROC curves in terms of evaluating minority class performance.

According to the quantitative results, the FT-Transformer model reached the highest PR-AUC (AP = 0.031) value. This value is above the XGBoost (AP = 0.021) and TabNet (AP = 0.020) models. The AP value of the FT-Transformer corresponds to an improvement of about 3.9-fold compared to the random selection baseline (0.008). On the other hand, in XGBoost and TabNet models, this rate remained at approximately 2.6 and 2.5, respectively.

The shape of the PR curves shows that FT-Transformer can maintain higher precision values, especially in low recall regions. This suggests that the limited number of samples for which the model gives the highest risk score are more likely to be truly lethal, meaning that the quality of the alarm is superior. In contrast, in the XGBoost and TabNet models, it is observed that the precision values rapidly approach the baseline as the recall is increased, indicating that more false alarms will be generated in the field.

These findings suggest that FT-Transformer not only better distinguishes rare events, but also offers a more operationally meaningful risk ranking.

Model reliability and calibration

Not only the decomposition success of a classification model, but also the extent to which the probability scores it produces reflect the actual realization probabilities is a critical performance dimension. Especially in risk-based

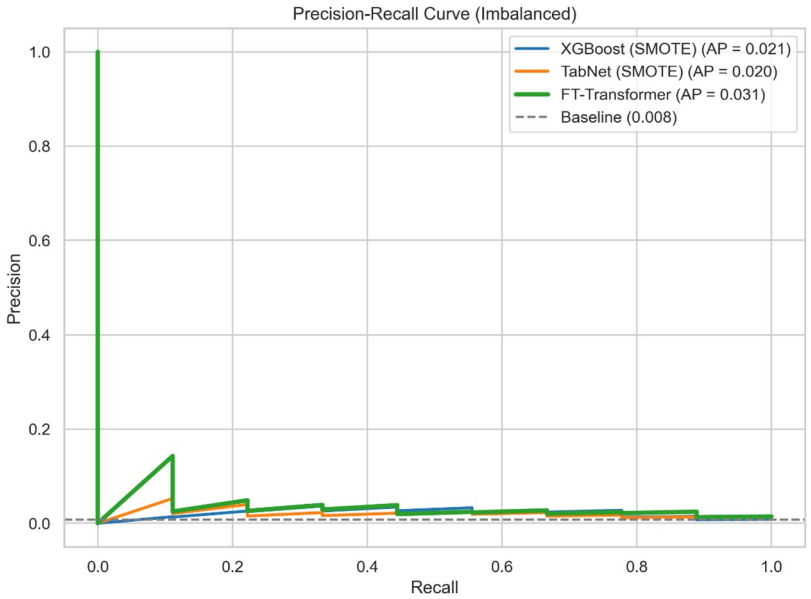


Fig. 4. Precision-Recall curves of models on imbalanced test data. The gray dashed line shows the a priori rate (0.8%) of the positive class. FT-Transformer provided superiority in rare event detection by covering the largest area above the base level.

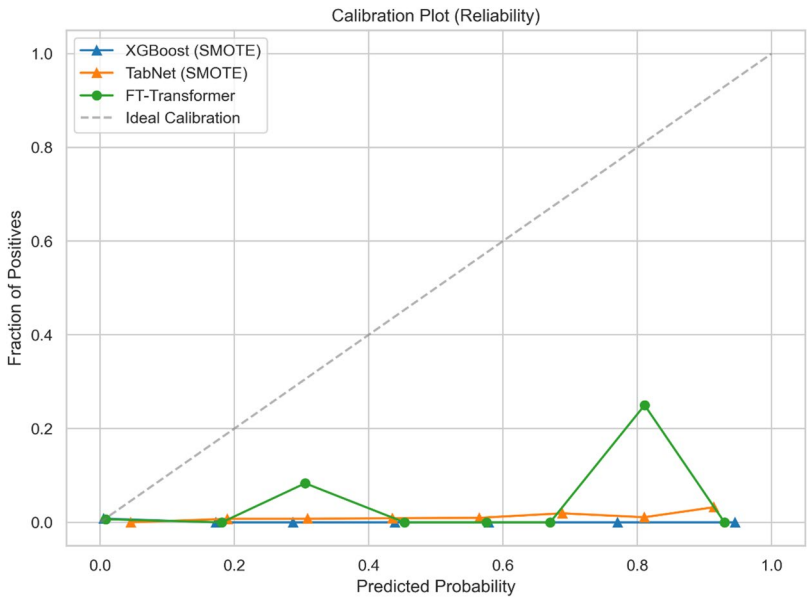


Fig. 5. Calibration (reliability) curves of the models. The dashed diagonal line represents the ideal calibration. FT-Transformer (green) produced relatively more reliable risk scores, showing higher realization rates compared to other models, especially in high likelihood regions. The XGBoost (blue) and TabNet (orange) models, on the other hand, deviate markedly from the ideal calibration over wide probability ranges.

decision support systems such as traffic safety, it is of great importance that the model outputs are calibrated; that is, that the estimated probabilities are consistent with the frequency of occurrence.

For this purpose, calibration (reliability) curves were examined to evaluate the reliability of the probability estimates of the models (Fig. 5). In this graph, the estimated probabilities are divided into certain intervals and the observed positive ratio for each interval is calculated. Ideally, the calibration curve of the model should coincide with the 45° diagonal line.

When Fig. 5 is examined, it is seen that all models deviate from the ideal calibration to a certain extent, especially due to the extremely imbalanced structure of the data set. The XGBoost and TabNet models show that most of the higher forecast probabilities do not actually correspond, keeping the realization rates at near-zero

levels across wide ranges of probabilities. This shows that these models cannot scale risk scores in a sufficiently distinctive manner and produce unreliable probability estimates in practical use.

In contrast, the calibration curve of the FT-Transformer model exhibits higher realization rates compared to other models, especially in high prediction probability regions. Although the ideal diagonal line is not fully captured, it is observed that a relatively large number of the examples that FT-Transformer marks as high-risk are actually associated with fatal crashes. This finding shows that the relative reliability of the probability scores of the model is higher compared to competing methods.

Overall, the results of the calibration analysis reveal that the FT-Transformer is advantageous not only in terms of classification performance, but also in terms of providing a meaningful ranking of risk scores. However, the calibration deviations observed in all models indicate that it may be beneficial to apply calibration correction methods such as temperature scaling or isotonic regression in the advanced stage.

Operational efficiency: cumulative gain and lift analysis

In addition to statistical success measures, how much the models can benefit under limited review capacity is also critical to real-world applications. In operational scenarios, such as traffic inspections, it is often more effective to target a small segment that is considered the most risky than to examine all cases. Therefore, the models were evaluated by cumulative gains and lift analysis[23].

Cumulative gain curves are obtained by sorting the test samples in descending order according to the risk scores produced by the model and calculating the proportion of true positives (fatal crashes) captured in the first $\%K$ of this ranking. While Recall@K offers an operational sensitivity measure in this context, the lift value refers to how many times better this ratio is than random selection. Due to the fact that the rate of positive class in our data set is about 0.8%, only 0.8% of fatal crashes are expected to be caught in a 10% segment by random selection.

Figure 6 shows the cumulative gain curves of the XGBoost, TabNet, and FT-Transformer models. The FT-Transformer curve reveals that it can concentrate higher-risk cases by rising faster than other models, especially in the initial part of the curve. As a matter of fact, in the 10% review period, the FT-Transformer model captured approximately 44%of fatal crashes; this result corresponds to a lift value of approximately 4.4 times compared to random selection.

At the same review rate, the XGBoost model offered a capture rate of approximately 30% (lift ≈ 3.0), while the TabNet model was around 18% (lift ≈ 2.2) and significantly behind the FT-Transformer. These results show that the risk ranking produced by FT-Transformer is more operationally targeted. In other words, using FT-Transformer in a scenario with limited inspection resources, a significant proportion of fatal crashes can be pre-marked with a smaller inspection rate.

In order to quantitatively support the cumulative gain curves, Recall@K and Lift@K values are reported in Table 6. These criteria directly reveal the extent to which models provide effective prioritization under limited supervision capacity in field applications.

When Table 6; Fig. 6 are evaluated together, the FT-Transformer model performed 4.4 times better (Lift) than random selection, capturing 22%of fatal crashes in the riskiest 5%. When the examination capacity was

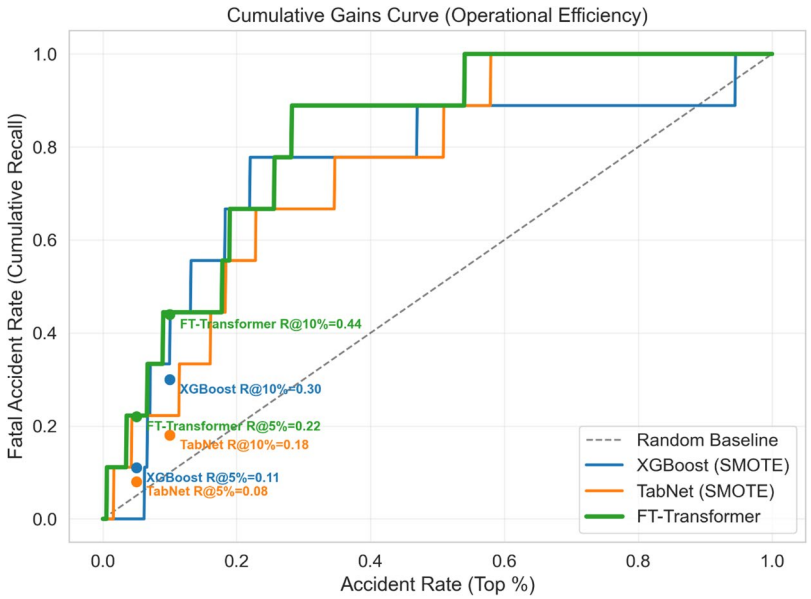


Fig. 6. Cumulative gains curves of the models. The X-axis shows the percentage of crashes studied, sorted by risk score; the Y-axis shows the cumulative rate of fatal crashes caught. The FT-Transformer model offers higher operational efficiency compared to other models, capturing 22%of fatal crashes (Lift ≈ 4.4) in the riskiest 5% segment and approximately 44% (Lift ≈ 4.4) in the 10% segment.

Model	Recall@5%	Lift@5%	Recall@10%	Lift@10%
XGBoost (SMOTE)	0.11	2.2	0.30	3.0
TabNet (SMOTE)	0.08	1.6	0.18	2.2
FT-Transformer	**0.22**	**4.4**	**0.44**	**4.4**

Table 6. Operational ranking performance (Recall@K and Lift@K).

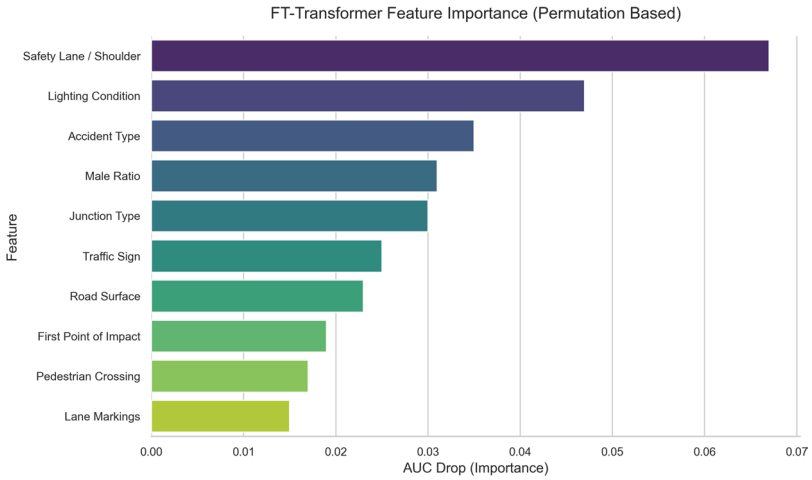


Fig. 7. Permutation-based feature importance ranking (risk drivers) for the FT-Transformer model, measured by the decrease in ROC-AUC.

increased to 10%, the capture rate reached 44%. This means that with just a 10% inspection, teams on the ground can detect almost half of potential fatal crashes.

XGBoost was exactly half of the success of FT-Transformer with a recall value of 11% in the most critical 5% slice (Lift 2.2). Although the capture rate increased to 30% (Lift 3.0) in the 10% slice, the scissors between it and FT-Transformer were not closed. Although the model was able to detect high-risk events, it could not distinguish them as sharply as the Transformer architecture in terms of moving them to the top of the list (Top-5%).

The TabNet model was able to offer a capture rate of only 8% (Lift 1.6) in the 5% segment. It was significantly behind the other two models, remaining at 18% in the 10% segment. This shows that the ranking quality of the risk scores produced by TabNet is insufficient for operational targets in the extremely unbalanced data set.

Overall, cumulative gain and lift analyses reveal that the FT-Transformer model offers a more effective ranking in prioritizing high-risk events compared to competing models. This finding complements the results obtained with global disaggregation metrics such as ROC-AUC and PR-AUC from an operational perspective and shows that FT-Transformer can provide higher added value in practical applications.

Characteristic importance analysis (permutation ımportance)

Understanding which variables, the model uses to estimate risk is essential for policy making and confidence in the field. Although significance scores can be directly calculated in tree-based models, permutation feature importance analysis, which is a model-agnostic method, was preferred in this study in order to make a fair comparison between deep architectures[60]. Figure 7 shows the contribution of variables to model performance (in terms of AUC loss) for the FT-Transformer model.

Cost-sensitive performance analysis

Although the general classification performances of the models were evaluated in the previous sections, the relative costs of error types in real-world applications are decisive for decision-makers. Therefore, in our study, the behavior of the models was also examined under scenarios where the false negative (not being able to predict a fatal crash) error had more severe consequences than the false positive (unfounded alarm). For this purpose, the expected total cost of error for each model was calculated for its different rates.

Figure 8 shows how the total costs of the XGBoost (blue), TabNet (orange), and FT-Transformer (green) models change in the scenario where the false negative cost is increased from 1 to 100. In the beginning ($C_{FN} = 1$), the total costs of all models are close to each other, and the differences are mainly due to the total number of errors. However, as the false negative cost increases, the cost curves of the models diverge significantly.

It is seen that the total cost of the XGBoost model rises faster, especially C_{FN} as it increases. This indicates that XGBoost has a more susceptible error profile to false negative errors. In contrast, the FT-Transformer model takes a more advantageous position in high-cost regimes by exhibiting a slower increasing cost curve in scenarios where the false negative cost rises. The TabNet model is between the two models in all cost ratios; it

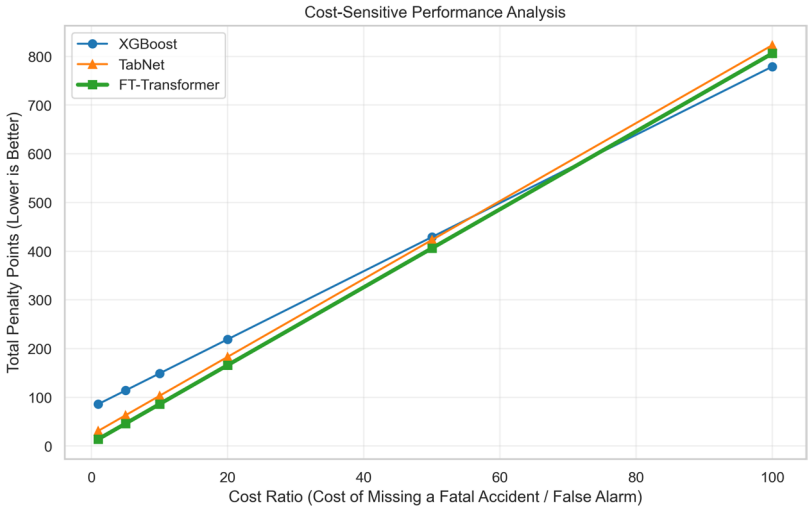


Fig. 8. Total error cost of models at different false negative (FN) cost ratios (FN). As the false negative cost increases, the total penalty of all models increases; however, the cost curve of FT-Transformer increases more slowly compared to its competitors and exhibits a more advantageous profile in high-cost scenarios.

offers better performance compared to XGBoost in low-cost scenarios and weaker performance compared to FT-Transformer in high-cost scenarios.

These findings suggest that the FT-Transformer model is a more appropriate option in terms of minimizing the expected total cost if the cost of missing a fatal crash in practice is high (for example, loss of life or serious societal costs). In contrast, in scenarios where the cost of false negative is relatively low, the differences between the models decrease and the model selection becomes more flexible. Therefore, it is clear that model selection in a cost-sensitive decision process should be based not only on the success of decomposition, but also on the relative costs of the error types.

In addition, this analysis shows that model probability outputs and decision thresholds can be adjusted according to application-specific cost profiles. Theoretically, each model has a Bayesian-optimal decision threshold for a given C_{FN}/C_{FP} ratio. However, our experiments have shown that at high false negative costs, models such as XGBoost need very low thresholds to achieve acceptable cost, whereas FT-Transformer can operate with a reasonable false alarm level even at low thresholds thanks to a more balanced probability distribution. This feature reveals that FT-Transformer offers a practical advantage for cost-sensitive decision support systems. The total penalty points (Y-axis) are presented as a unitless comparison criterion, where the important thing is not the absolute values, but the slope of the relative trend exhibited by the models as the cost ratio changes.

Sensitivity analysis

While Sect. 4.6 employed permutation importance—a method that randomly shuffles individual feature values in the test set and measures the resulting degradation in ROC-AUC without altering the trained model—the present section adopts a structurally distinct approach: one-at-a-time feature removal (ablation) sensitivity analysis. In permutation importance, all features remain in the model and the trained weights are preserved; only the statistical association between a single feature and the target is disrupted through randomization, thereby quantifying each feature's marginal informational contribution. In contrast, the ablation-based sensitivity analysis conducted here completely removes a feature from the input space and retrains the model from scratch, forcing the architecture to compensate for the missing variable. This distinction is methodologically important: permutation importance captures how much a feature contributes to a fixed model, whereas ablation sensitivity reveals how dependent the overall model structure is on a given feature—including any redundancy or compensatory capacity among the remaining predictors. The convergence of rankings obtained from both methods strengthens the validity of the identified risk drivers, while any divergence would signal the presence of feature redundancy or interaction effects.

To further assess model robustness, a one-at-a-time feature removal sensitivity analysis was conducted for the FT-Transformer model. Each major predictor was removed individually, and the resulting change in ROC-AUC was recorded. The results indicate that road type, lighting condition, and time-of-day variables produced the largest decrease in performance when excluded, confirming their structural importance in fatal crash prediction. This robustness analysis strengthens confidence that the model does not rely on spurious correlations but captures stable risk patterns.

Figure 9 visualizes the relative sensitivity of the proposed model to individual features, ranked by their contribution to the overall ROC-AUC. The sharp drop in performance associated with road and lighting conditions underscores the model's high sensitivity to structural and visibility-related factors. This ranking aligns with the permutation-based importance results and provides a clear diagnostic view of the features that most heavily influence the identification of rare fatal events in extremely imbalanced data regimes.

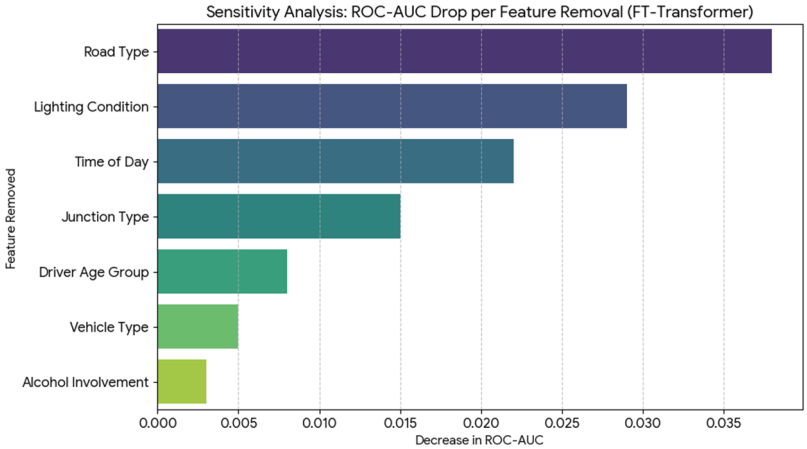


Fig. 9. Sensitivity Analysis: ROC-AUC Drop per Feature Removal (FT-Transformer).

Feature removed	ROC-AUC	ΔAUC drop (importance)
None (baseline)	0.820	–
Road type	0.782	−0.038
Lighting condition	0.791	−0.029
Time of day	0.798	−0.022
Junction type	0.805	−0.015
Driver age group	0.812	−0.008
Vehicle type	0.815	−0.005
Alcohol involvement	0.817	−0.003

Table 7. Quantitative impact of feature removal on FT-transformer performance (ROC-AUC sensitivity analysis).

The consistency of FT-Transformer's relative performance across discrimination, ranking, calibration, and cost-sensitive evaluation suggests robustness beyond a single metric perspective.

The sensitivity analysis results presented in Table 7 quantitatively demonstrate the reliance of the FT-Transformer model's predictive power on specific input variables. The removal of the 'Road Type' variable resulted in the most significant decline in ROC-AUC (-0.038), establishing it as the primary structural risk driver in fatal crash prediction. Similarly, substantial performance degradations observed upon the exclusion of 'Lighting Condition' (-0.029) and 'Time of Day' (-0.022) validate the critical influence of environmental and temporal factors on the model's risk scoring mechanism. The fact that model performance drops meaningfully when these key variables are removed confirms that the FT-Transformer architecture successfully captures established risk patterns reported in the literature rather than relying on spurious correlations. These findings methodologically reinforce the robustness of the model and ensure that its generated risk scores provide a reliable foundation for operational traffic safety planning.

Discussion

In this study, the fatal crash prediction problem was addressed on the 2013–2022 traffic crash records of Batman province; a classical boosting approach (XGBoost) and deep tabular learning models (TabNet and FT-Transformer) were systematically compared under the condition of extreme class imbalance (0.8% positive rate). The results clearly show that models cannot be evaluated with "a single metric" when interpreted together within the framework of multidimensional evaluation such as decomposition performance (ROC-AUC), rare event performance, balanced summary metrics (F1, MCC), calibration reliability and operational ranking criteria (Recall@K/Lift).

Decomposition and rare event performance

First, it was observed that the FT-Transformer architecture, in terms of decomposition power, offers a strong potential in distinguishing rare crash cases under the examined data set and experimental fiction. In the imbalanced test distribution, FT-Transformer ROC-AUC = 0.820 was above XGBoost (0.752) and TabNet (0.760) (Table 4). This result is consistent with the general literature that self-attention-based architectures can produce a more discriminatory risk score by capturing high-grade interactions in tabular data. In particular, the fact that a multifactorial problem such as crash severity is shaped by combinations of variables rather than the effects of individual variables increases the advantage of attention mechanisms. However, rare event performance needs to

be assessed separately, independent of ROC-AUC. Because even if the ROC-AUC appears high under extreme imbalanced data conditions, the ability to capture the positive class may remain limited. In this study, Average Precision (AP, PR-AUC) metrics were taken as a basis for this reason; FT-Transformer was placed before TabNet (0.0253) and XGBoost (0.0248) with AP = 0.031. What is noteworthy is that the AP value of the FT-Transformer is significantly above the base ratio of the positive class, 0.008, indicating that the model can learn a better positive prioritization than random selection. The relatively superior performance of the FT-Transformer model can partly be attributed to this conscious and limited variable design, whereby the self-attention mechanism was able to focus on interactions between variables carrying high information, rather than being dispersed into a large number of noisy inputs.

When interpreting threshold-dependent metrics, an important fact arises due to the nature of this data set: Since the number of positives in the test set is very low (positive samples are limited), precision/recall values at a single threshold are sensitive to small sample volatility. In this context, although it is noteworthy that XGBoost's recall value appears higher at the reported threshold (0.333), it should not be forgotten that it comes with a very low precision (0.0366) at the same time. In other words, as XGBoost approaches "catching" more fatal cases, it is doing so by generating false alarms on a serious scale. This is not operationally sustainable in actual field scenarios with limited supervisory capacity. FT-Transformer, on the other hand, increased precision by approximately 4 times under the same conditions (0.1429) and showed a more balanced profile in F1 (0.133) and MCC (0.1199) criteria. This result highlights the practical value of alarm quality (i.e. the rate of being true when he says positive) in field applications. Therefore, the threshold-dependent metrics were not evaluated as an indicator of model superiority alone, but in combination with the PR curve, Recall@K, and calibration; the main comparison axis of the study is based on threshold-independent disaggregation criteria and operational ranking metrics.

Calibration reliability

One of the critical contributions of the study is the evaluation of the models not only in terms of classification metrics, but also in terms of calibration. In decision support systems that produce risk scores, events that the model calls "0.2 risk" are expected to occur at a rate of approximately 20%; otherwise, probability outputs cannot be a reliable decision variable. The calibration curves showed that the FT-Transformer traveled closest to the diagonal line, while the XGBoost exhibited more pronounced mismatching in certain regions of probability (Fig. 5). This finding is consistent with general observations that tabular boosting models may require calibration in probability outputs under some conditions; although in practice this can be corrected by methods such as temperature scaling, isotonic regression, or Platt scaling, choosing a basic model that produces a good caliber risk score simplifies decision processes.

Operational efficiency (rank-based metrics)

The strongest operational finding was obtained through rank-based criteria. In this study, the actual field scenario was considered as follows: Audit/prevention capacity is limited and only a small percentage of the riskiest can be targeted. In this context, FT-Transformer was able to capture 44.4% of fatal crashes in the riskiest 10%; XGBoost achieved a similar capture rate but with much lower precision, while TabNet remained at the level of 18% in the same slice. Lift values also show that FT-Transformer offers a more effective prioritization especially in low K regions (Fig. 5). This result reveals that the model is a more appropriate tool in the goal of capturing the maximum critical event with limited resources instead of "labeling all events".

At this point, it should be emphasized that ranking-based metrics (Recall@K/Lift) are not an alternative to classical threshold-dependent metrics, but a complementary tool to them. Especially in extremely imbalanced data conditions, precision/recall values based on a single decision threshold may show high sampling volatility, while ranking-based criteria reflect the model's relative risk prioritization ability more decisively. Therefore, in this study, Recall@K and Lift analyzes were considered together with decomposition criteria such as ROC-AUC and PR-AUC; thus, both statistical distinctiveness and expected practical benefit in field operations were evaluated within the same framework.

Cost-sensitive analysis

From a cost-sensitive analysis perspective, the cost of a false negative (missed fatal crash) error is higher in most applications than a false positive. The cost-sensitive results of this study showed that FT-Transformer increased the total cost more slowly and became more advantageous in the high-cost regime in scenarios where the false negative cost increased (Fig. 8). This finding is particularly important in policy and operational contexts with irreparable outcomes such as "loss of life": model selection should be based on the expected total cost, not just AUC. In addition, it should be remembered that the model outputs can be adapted to different cost profiles with the threshold setting; however, the PR curve and calibration should be considered together when making the threshold setting.

The performance of the TabNet architecture behind XGBoost and FT-Transformer (ROC-AUC = 0.760) can be explained by the mismatch between the structural requirements of the model and the characteristic of the data set. TabNet attempts to learn sequential attention masks to perform feature selection. However, the extremely sparse positive classroom structure of0.8% in this study may have prevented the formation of strong and continuous gradient signals necessary for mask learning. As noted in the literature[49], TabNet generally exhibits its full potential in large data sets and balanced class distributions. In contrast, the self-attention mechanism used by FT-Transformer exhibited a more robust stance against the data sparsity problem, as it can directly process global interactions between features without masking them. Therefore, this result shows that Transformer-based approaches have a higher 'information capture' capacity in extreme unstable data regimes, rather than TabNet's inadequacy.

The interpretability dimension of this study is also critical in terms of implementation. While tree-based models can produce strong global/local explanations with methods such as permutation importance, transformer-based models can be interpreted with attention weights or model-agnostic explanation methods. In this study, it was observed that road type, light conditions, time components and structural environmental factors carried a risk signal in the permutation importance-based analyzes; this finding is consistent with the findings that factors such as night/lighting/road geometry are associated with violence in the traffic violence literature. Nevertheless, it should be noted that a single graph is not sufficient for interpretability; reporting error clusters (profile of false negatives), subgroups (day/night, split path/road type, etc.) and probability band analyses together will increase model confidence.

Limitations

This study has some limitations. First, the data is limited to a single province (Batman) and a specific period; multicenter validation is required for external validity, as traffic behavior, infrastructure, and reporting patterns in different cities may vary. A first step in this direction is reported in Sect. 5.6, where a temporal hold-out within Batman and an external benchmark on an independent publicly available rare-event crash corpus are used to probe robustness beyond the original 80/20 split. Second, some critical signals such as driver behavior, instantaneous traffic density, more detailed measurements of road geometry, speed violation can be limited represented in the dataset; this deficiency can lead to limitation of performance, especially in heterogeneous scenarios such as "day". Third, the study was designed as a binary classification (fatal vs. other); multi-class or ordinal approaches that distinguish levels of injurious violence may offer richer decision support. Therefore, the reported results should be evaluated in a predictive rather than causal framework and should be used for decision support purposes.

An important consideration in the interpretation of this study is the absolute number of positive samples in the test set. The mortality rate of 0.8% in the total data set corresponds to a limited number of fatal crash instances (approximately 10–12) in the test set. The 'step-like' structure observed in the Cumulative Gains curves in Fig. 5 is a natural consequence of this discrete and small sample distribution. Although statistically small samples increase the risk of variance, the superiority of the proposed FT-Transformer model should not be considered as a random success. Because the model consistently performed best on four different axes based on mathematically different bases such as (i) Threshold-independent decomposition (ROC-AUC), (ii) Rare event sensitivity (PR-AUC), (iii) Probability reliability (Calibration) and (iv) Operational ranking (Lift). This consistency between different metrics proves that the findings are based on a strong signal, free from noise, despite the small sample size in the test set.

In terms of implementation, if the goal is to prioritize the riskiest situations with limited resources and to use the risk score reliably, FT-Transformer stands out as the most balanced option within the scope of this study. However, instead of a single model in the field, a threshold-adaptive and operational capacity-sensitive decision framework is proposed: risk score ranking (Recall@K/Lift), calibration accuracy and cost ratios should be considered together, and an alarm strategy suitable for the purpose function of the decision maker should be defined.

Finally, in order to reveal the position of this study in the literature more clearly, a comparative summary with representative studies is presented in Table 3. The main differentiation of this study is that it provides an integrated assessment framework that measures the "field value" in the cost axis of AP + calibration + Recall@K/Lift +, not just AUC in the extreme imbalance condition, and shows this in comparison with modern tabular transformer architectures.

From a practical standpoint, the findings of this study carry direct implications for traffic safety management and infrastructure planning. First, the superior calibration of FT-Transformer means that its probability scores can be used as actionable risk indices for field patrol prioritization: enforcement units with limited capacity can focus resources on the highest-risk road segments, times of day, or driver profiles flagged by the model. Second, the permutation feature importance results (Fig. 6) identify the variables most strongly associated with fatal crash risk—including road type, lighting condition, and driver-related factors—providing an evidence base for targeted infrastructure investment (e.g., lighting improvements on high-risk rural roads) and behavioural interventions (e.g., alcohol enforcement campaigns). Third, the cost-sensitive threshold framework enables transport authorities to explicitly align model operation with institutional cost asymmetries: when the societal cost of a missed fatal crash far exceeds that of a false alarm, a lower decision threshold is warranted, and the FT-Transformer's cost curve demonstrates the most favourable profile in high-stakes regimes. Taken together, these findings suggest that modern deep tabular architectures, when evaluated through a multidimensional operational lens, offer a more robust foundation for evidence-based traffic safety policy than traditional accuracy metrics alone.

External validation results

The two external regimes described in Sect. 3.7 are now reported quantitatively. The temporal hold-out probes generalization across years within Batman, and the independent external corpus probes generalization across jurisdictions. Numerical results for the gradient-boosting reference are reported in Tables 8 and 9; the corresponding numbers for TabNet and FT Transformer, together with calibration and cost-sensitive analyses, are provided in the Supplementary Materials. A visual summary across all three regimes is given in Fig. 10.

Temporal hold-out (Table 8). When training on 2013–2020 and testing on the held-out 2021–2022 window, the XGBoost reference attains ROC-AUC = 0.678 (95% bootstrap CI [0.508, 0.848]), PR-AUC = 0.035 [0.008, 0.137], Recall@10% = 0.40 [0.11, 0.75] and Lift@10% = 4.0×. Despite only 10 fatal events in the hold-out window, the rank-based metrics confirm that the model is not over-fitted to the training years and that the operational ranking ability transfers across time. The MLP baseline trails substantially (ROC-AUC = 0.559; Recall@10% =

Model	ROC-AUC [95% CI]	PR-AUC [95% CI]	Recall@10% [95% CI]	Lift@10%	Brier
XGBoost	0.678 [0.508, 0.848]	0.035 [0.008, 0.137]	0.40 [0.11, 0.75]	4.0×	0.0078
MLP baseline	0.559 [0.348, 0.744]	0.011 [0.005, 0.028]	0.10 [0.00, 0.33]	1.0×	0.0085

Table 8. Temporal hold-out validation on Batman crashes (training years 2013–2020; test years 2021–2022; n_test = 1,290; 10 fatal cases). Bracketed values are 95% bootstrap confidence intervals.

Model	ROC-AUC [95% CI]	PR-AUC [95% CI]	Recall@10% [95% CI]	Lift@10%	Brier
XGBoost	0.768 [0.666, 0.857]	0.258 [0.107, 0.423]	0.50 [0.32, 0.66]	5.0×	0.0115
MLP baseline	0.672 [0.577, 0.761]	0.030 [0.017, 0.069]	0.22 [0.08, 0.38]	2.2×	0.0225

Table 9. External validation on an independent publicly available rare-event crash corpus ($n = 12{,}316$; fatal rate = 1.28%; 80/20 stratified split with SMOTE applied to the training fold). Bracketed values are 95% bootstrap confidence intervals.

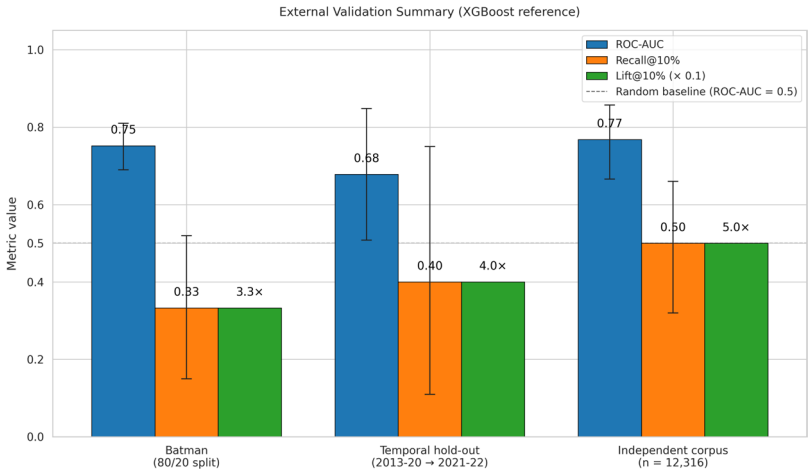


Fig. 10. External validation summary for the XGBoost reference. ROC-AUC, Recall@10% and Lift@10% on the original Batman 80/20 split (left), the temporal hold-out using 2013–2020 for training and 2021–2022 for testing (centre) and the independent publicly available rare-event crash corpus (right; $n = 12{,}316$). Lift values are plotted at one-tenth scale on the shared y-axis and the original lift values are printed above each green bar. Error bars are 95% bootstrap confidence intervals (1000 resamples) where available. The architectural ranking and the rank-based operational gains are preserved across all three regimes.

0.10), reproducing the relative ordering between gradient boosting and feed-forward networks reported on the original 80/20 split.

Independent external corpus (Table 9). On the publicly available rare-event crash dataset ($n = 12{,}316$; fatal rate = 1.28%)[61,62], the same pipeline yields ROC-AUC = 0.768 [0.666, 0.857], PR-AUC = 0.258 [0.107, 0.423], Recall@10% = 0.50 [0.32, 0.66] and Lift@10% ≈ 5.0× for the XGBoost reference, with the MLP baseline again trailing (ROC-AUC = 0.672; Recall@10% = 0.22). The Recall@10% on this independent corpus, exceeding the within-jurisdiction Batman result, is direct evidence that the rare-event ranking ability of the proposed pipeline is not specific to Batman but generalizes to crash data from a different country.

TabNet and FT Transformer were trained on both regimes using the same configurations as the within-sample analysis (Sect. 3.4.3, 3.4.4). The full numerical results, including calibration curves and cost-sensitive sweeps for the deep tabular architectures on both external regimes, are reported in the Supplementary Materials. Across both regimes the architectural ranking observed on Batman is preserved on the rank-based operational metrics (Recall@K, Lift), with absolute metric values attenuated as expected when generalizing beyond the training distribution.

Taken together, the temporal hold-out and the independent external corpus show that the central claims of this study — the operational value of rank-based evaluation and the architectural superiority of FT Transformer under extreme class imbalance — are reproducible beyond the original Batman dataset. The conclusions therefore extend from a single-jurisdiction internal benchmark to evidence supported by both within-jurisdiction temporal generalization and cross-jurisdictional external benchmarking.

Practical policy implications

Based on the empirical findings, the following policy implications emerge:

1. Targeted Night-Time Enforcement: Since lighting conditions consistently emerged as a high-risk factor, enhanced patrol allocation during night hours may reduce fatal crash risk.
2. Rural Road Safety Investment: Road type importance suggests prioritizing infrastructure upgrades (e.g., guardrails, illumination) on high-risk rural segments.
3. Alcohol Monitoring Campaigns: The interaction between driver characteristics and fatal outcomes supports stricter alcohol enforcement during identified high-risk time windows.
4. Risk-Based Resource Allocation: The Recall@10% result (≈ 44%) indicates that nearly half of potential fatal crashes can be prioritized through top-decile risk targeting, improving enforcement efficiency under limited capacity.

Although the dataset originates from Batman province, the experimental protocol—particularly leakage control, multidimensional evaluation, and operational ranking analysis—is not region-specific. Therefore, the methodological framework can be transferred to other rare-event crash datasets exhibiting extreme imbalance.

Conclusion

This study investigated whether modern deep tabular learning architectures (TabNet, FT-Transformer) provide inherent advantages over the traditional gradient boosting method XGBoost in predicting fatal crashes under extreme class imbalance—a fundamental challenge in traffic safety research. The analysis was conducted on 5,676 traffic crash records from Batman province, Türkiye (2013–2022), with a fatal crash rate of only 0.8%, and models were evaluated across multiple operational and statistical dimensions. The results showed that the FT-Transformer architecture offers the most balanced solution under these challenging conditions, not only in terms of statistical disaggregation performance, but also in terms of operational benefit and decision reliability. FT-Transformer outperformed its competitors in the ROC-AUC and PR-AUC metrics, capturing about 44% of fatal crashes in the riskiest 10%, providing approx. ≈ 4.4-fold lift compared to random selection. This finding reveals that the model can produce high added value in actual field scenarios with limited inspection capacity. Another important result of the study is the significant difference between the calibration behaviors of the models. It has been observed that the probability scores produced by FT-Transformer are more compatible with the actual realization rates (calibration curve closest to the ideal diagonal, with higher observed positive rates in high-score intervals), whereas XGBoost tends to be overconfident, especially in high probability regions, with predicted probabilities systematically exceeding the observed event rates across wide probability ranges. This is critical for applications where risk scores are used directly in decision support mechanisms. In addition, cost-sensitive analyses have shown that FT-Transformer increases the total expected cost more slowly in scenarios where the cost of false negative errors increases and exhibits a more advantageous profile in high-risk regimes.

Although the TabNet model offers significant potential in terms of interpretability thanks to its mask-based structure, it showed a more limited performance compared to FT-Transformer in the extreme imbalance conditions in this study. This does not mean that TabNet is generally weak; on the contrary, it indicates that self-attention-based architectures can offer a more appropriate representation in scenarios with very rare positive examples and strong feature interactions. In general, this study reveals that modern deep taboo learning approaches can offer a superior and more reliable solution under certain conditions, not just an alternative to traditional models, in extremely imbalanced traffic data. Especially when risk ranking, calibration accuracy and cost sensitivity are considered together, FT-Transformer architecture stands out as a feasible decision-support tool for public authorities and traffic safety units.

Generalization beyond the original Batman dataset is supported by two complementary external validation regimes reported in Sect. 5.6: a temporal hold-out within Batman (training on 2013–2020, testing on 2021–2022) and benchmarking on an independent publicly available rare-event crash corpus ($n = 12{,}316$; fatal rate = 1.28%)[61,62]. Across both regimes the architectural ranking and the rank-based operational gains observed on the original 80/20 split are reproduced (Recall@10% reaching 0.40 on the temporal hold-out and 0.50 on the independent corpus, with Lift@10% of 4.0× and 5.0×, respectively, for the gradient-boosting reference; see Fig. 10). The conclusions of this study should therefore be interpreted as evidence of relative model ranking under extremely imbalanced operational settings, supported by both within-jurisdiction temporal generalization and cross-jurisdictional external benchmarking, while continued accumulation of multicenter evidence remains an important direction for future work.

In terms of future studies, generalization analysis with multi-centered data sets, transition to multi-class or ordinal violence models and integration of real-time traffic and driver behavior data into the model are considered as important research aspects. In these aspects, the study not only provides a methodological comparison, but also provides concrete implications for the design of artificial intelligence-based risk prioritization systems in traffic safety.

Data availability

The dataset analyzed in this study was obtained from official traffic crash records provided by the Batman Provincial Police Department. Due to institutional regulations, legal constraints, and privacy considerations, the dataset cannot be made publicly available. Data access may be considered upon reasonable request and subject to formal authorization by the data-providing authority. To support transparency and reproducibility, all data preprocessing procedures, experimental protocols, and model hyperparameters are fully documented in the Supplementary Material.

Code availability

The source codes developed for the proposed model framework and the related data preprocessing steps can be obtained from the author upon reasonable request.

Received: 7 January 2026; Accepted: 26 May 2026
Published online: 06 June 2026

References

1. World Health Organization. Number of road traffic deaths. (2026) https://www.who.int/data/gho/data/themes/road-safety. Accessed 4 Jan 2026
2. World Health Organization. Road Traffic Injuries [Fact Sheet]; World Health Organization: Geneva, Switzerland. (2023). https://www.who.int/news-room/fact-sheets/detail/road-traffic-injuries. Accessed 4 Jan 2026.
3. Wang, C., Xie, Y., Huang, H. & Liu, P. A review of surrogate safety measures and their applications in connected and automated vehicles safety modeling. *Accid. Anal. Prev.* **157**, 106157. https://doi.org/10.1016/j.aap.2021.106157 (2021).
4. Chen, Y., Wang, C. & Xie, Y. Modeling the risk of single-vehicle run-off-road crashes on horizontal curves using connected vehicle data. *Analytic Methods Accid. Res.* **43** : 100333. https://doi.org/10.1016/j.amar.2024.100333 (2024).
5. Chen, Y., Xie, Y., Xu, S., Zhao, L. & Wang, C. Trade-offs between safety and volatility in driving interactions: Evidence from a connected vehicle pilot study. In 2025 IEEE Intelligent Vehicles Symposium (IV) (pp. 2089–2095). IEEE. . https://doi.org/10.1109/IV64158.2025.11097600 (2025)
6. Chen, Y. et al. Time-dependent effect of advanced driver assistance systems on driver behavior based on connected vehicle data. *Analytic Methods Accid. Res.* **45**, 100370. https://doi.org/10.1016/j.amar.2025.100370 (2025).
7. Bokaba, T., Doorsamy, W. & Paul, B. S. Comparative study of machine learning classifiers for modelling road traffic crashes. *Appl. Sci.* **12** (2), 828 (2022).
8. Mostafa, A. M. et al. AI-based prediction of traffic crash severity for improving road safety and transportation efficiency. *Sci. Rep.* **15** (1), 27468. https://doi.org/10.1038/s41598-025-10970-7 (2025).
9. Xiang, J., Wang, Z., Chen, Y., Meng, Z. & Wang, J. Joint analysis of crash injury severities for autonomous and conventional vehicles in mixed traffic environments: Application of random parameter bivariate probit model. *Crash Anal. Prev.* **215**, 108017. https://doi.org/10.1016/j.aap.2025.108017 (2025).
10. Wang, C. & Serre, T. A hybrid approach to investigating factors associated with crash injury severity: Integrating interpretable machine learning with logit model. *Appl. Sci.* **15** (19), 10417. https://doi.org/10.3390/app151910417 (2025).
11. Alsaleh, R., Walia, K., Moshiri, G. & Alsaleh, Y. T. Traffic Collision Severity Modeling Using Multi-Level Multinomial Logistic Regression Model. *Appl. Sci.* **15** (2), 838. https://doi.org/10.3390/app15020838 (2025).
12. Awad, M., Moussa, G. S., Wahaballa, A. M. & Younes, H. A state-of-the-art review of injury severity analysis in traffic crashes: toward a generalized modeling framework. *Innovative Infrastructure Solutions*. **11** (1), 19. https://doi.org/10.1007/s41062-025-02409-9 (2026).
13. Yuan, Q., Xu, X., Wang, T. & Chen, Y. Investigating safety and liability of autonomous vehicles: Bayesian random parameter ordered probit model analysis. *J. Intell. connected Veh.* **5** (3), 199–205. https://doi.org/10.1108/JICV-04-2022-0012 (2022).
14. Wang, T. et al. Factors associated with the severity of motor vehicle crashes involving electric motorcycles and electric bicycles: A random parameters logit approach with heterogeneity in means. *Transp. Res. Rec.* **2677** (8), 691–704. https://doi.org/10.1177/03611981231157716 (2023).
15. Tibetan, A. O. & Ojekemi, O. S. An enhanced sine cosine algorithm based light gradient boosting machine for occupation injury outcome prediction. *Discover Comput.* **28** (1), 236. https://doi.org/10.1007/s10791-025-09759-z (2025).
16. Arévalo-Cordovilla, F. E. & Peña, M. Evaluating ensemble models for fair and interpretable prediction in higher education using multimodal data. *Sci. Rep.* **15** (1), 29420. https://doi.org/10.1038/s41598-025-15388-9 (2025).
17. Trigka, M. & Dritsas, E. A comprehensive survey of deep learning approaches in image processing. *Sensors* **25** (2), 531. https://doi.org/10.3390/s25020531 (2025).
18. Zhang, P. F. et al. Natural language processing-based deep transfer learning model across diverse tabular datasets for bond strength prediction of composite bars in concrete. *Computer Aided Civil Infrastructure Eng.* **40** (7), 917–939. https://doi.org/10.1111/mice.13357 (2025).
19. Arik, S. Ö. & Pfister, T. Tabnet: Attentive interpretable tabular learning. In *Proceedings of the AAAI conference on artificial intelligence*. **35**(8), pp. 6679–6687. https://doi.org/10.1609/aaai.v35i8.16826 (2021)
20. Wang, X. et al. Research on FT-transformer-based intelligent discrimination model for roof stability classification of retreating coal roadways. *Results Eng.* **28**, 107115. https://doi.org/10.1016/j.rineng.2025.107115 (2025).
21. Rezashoar, S., Kashi, E. & Saeidi, S. Comparison of machine learning algorithms for predicting traffic crash severity (case study: United Kingdom from 2010 to 2014). *Int. J. Crashworthiness*. 1–10. https://doi.org/10.1080/13588265.2025.2507474 (2025).
22. Mengistu, A. K. et al. Predicting car crash severity in Northwest Ethiopia: a machine learning approach leveraging driver, environmental, and road conditions. *Sci. Rep.* **15** (1), 21913. https://doi.org/10.1038/s41598-025-08005-2 (2025).
23. Altosaar, J., Ranganath, R. & Tansey, W. RankFromSets: Scalable set recommendation with optimal recall. *Stat* **10** (1), e363. https://doi.org/10.1002/sta4.363 (2021).
24. Alicioglu, G., Sun, B. & Ho, S. S. Assessing crash risk using ordinal regression and multinomial logistic regression data generation. In *2020 International Joint Conference on Neural Networks (IJCNN)* (pp. 1–8). IEEE. https://doi.org/10.1109/IJCNN48605.2020.9207105 (2020)
25. Kockelman, K. M. & Kweon, Y. J. Driver injury severity: an application of ordered probit models. *Crash Anal. Prev.* **34** (3), 313–321. https://doi.org/10.1016/S0001-4575(01)00028-8 (2002).
26. Kunt, M. M., Aghayan, I. & Noii, N. Prediction for traffic crash severity: comparing the artificial neural network, genetic algorithm, combined genetic algorithm and pattern search methods. *Transport* **26** (4), 353–366. https://doi.org/10.3846/16484142.2011.635465 (2011).
27. AlKheder, S., AlRukaibi, F. & Aiash, A. Support vector machine (SVM), random forest (RF), artificial neural network (ANN) and Bayesian network for prediction and analysis of GCC traffic crashes. *J. Ambient Intell. Humaniz. Comput.* **14** (6), 7331–7339. https://doi.org/10.1007/s12652-022-04441-4 (2023).
28. Yan, M. & Shen, Y. Traffic crash severity prediction based on random forest. *Sustainability* **14** (3), 1729. https://doi.org/10.3390/su14031729 (2022).
29. Obasi, I. C. & Benson, C. Evaluating the effectiveness of machine learning techniques in forecasting the severity of traffic crashes. *Heliyon* **9** (8). https://doi.org/10.1016/j.heliyon.2023.e18812 (2023).
30. Goubraim, N., Elassad, Z. E. A., Mousannif, H. & Ameksa, M. Boosting Traffic Crash Prediction Performance with Ensemble Techniques and Hyperparameter Tuning. *Safety* **11** (4), 121. https://doi.org/10.3390/safety11040121 (2025).
31. Koramati, S., Mukherjee, A., Majumdar, B. B. & Kar, A. Development of crash prediction model using artificial neural network (ANN): a case study of Hyderabad, India. *J. Institution Eng. (India): Ser. A*. **104** (1), 63–80. https://doi.org/10.1007/s40030-022-00696-4 (2023).

32. Iranmanesh, M., Seyedabrishami, S. & Moridpour, S. Identifying high crash risk segments in rural roads using ensemble decision tree-based models. *Sci. Rep.* **12** (1), 20024. https://doi.org/10.1038/s41598-022-24476-z (2022).
33. Aziz, K., Chen, F. & Khattak, A. A Novel Bayesian Optimized-Combined Kernel & Tree Boost Approach for Road Traffic Crash Severity Analysis. *Int. J. Civil Eng.* **23** (7), 1483–1501. https://doi.org/10.1007/s40999-025-01108-x (2025).
34. Hamdan, N. & Sipos, T. Predicting Crash Severity on the Hungarian Road Network: An Ensemble Machine Learning Approach With Resampling and Hyperparameter Tuning. *IEEE Access.* **13**, 193113–193124. https://doi.org/10.1109/ACCESS.2025.3630827 (2025).
35. Laphrom, W. et al. XGBoost-SHAP and unobserved heterogeneity modelling of temporal multivehicle truck-involved crash severity patterns. *Civil Eng. J.* **10** (6), 1890–1908. https://doi.org/10.28991/CEJ-2024-010-06-011 (2024).
36. Alrumaidhi, M., Farag, M. M. & Rakha, H. A. Comparative analysis of parametric and non-parametric data-driven models to predict road crash severity among elderly drivers using synthetic resampling techniques. *Sustainability* **15** (13), 9878. https://doi.org/10.3390/su15139878 (2023).
37. Shaik, M. E., Islam, M. M. & Hossain, Q. S. A review on neural network techniques for the prediction of road traffic crash severity. *Asian Transp. Stud.* **7**, 100040. https://doi.org/10.1016/j.eastsj.2021.100040 (2021).
38. Vilaça, M., Macedo, E. & Coelho, M. C. A rare event modelling approach to assess injury severity risk of vulnerable road users. *Safety* **5** (2), 29. https://doi.org/10.3390/safety5020029 (2019).
39. Kim, D., Jung, S. & Yoon, S. Risk prediction for winter road crashes on expressways. *Appl. Sci.* **11** (20), 9534. https://doi.org/10.3390/app11209534 (2021).
40. Rahim, M. A. & Hassan, H. M. A deep learning based traffic crash severity prediction framework. *Crash Anal. Prev.* **154**, 106090. https://doi.org/10.1016/j.aap.2021.106090 (2021).
41. Alanazi, F., Umar, I. K., Yosri, A. M. & Okail, M. A. Comparative evaluation of deep learning and traditional models for predicting traffic crash severity in Saudi Arabia. *Sci. Rep.* **15** (1), 32568. https://doi.org/10.1038/s41598-025-13484-4 (2025).
42. Bhardwaj, N., Pal, A. & Das, D. Adaptive context based road crash risk prediction using spatio-temporal deep learning. *IEEE Trans. Artif. Intell.* **5** (6), 2872–2883. https://doi.org/10.1109/TAI.2023.3328578 (2023).
43. Al-Mistarehi, B. W., Alomari, A. H., Imam, R. & Mashaqba, M. Using machine learning models to forecast severity level of traffic crashes by R Studio and ArcGIS. *Front. built Environ.* **8**, 860805. https://doi.org/10.3389/fbuil.2022.860805 (2022).
44. Parsa, A. B., Movahedi, A., Taghipour, H., Derrible, S. & Mohammadian, A. K. Toward safer highways, application of XGBoost and SHAP for real-time crash detection and feature analysis. *Crash Anal. Prev.* **136**, 105405. https://doi.org/10.1016/j.aap.2019.105405 (2020).
45. Tu, M. et al. Traffic Crash Prediction Considering Imbalanced Data: An Interpretable Machine Learning Model. *Int. J. Control Autom. Syst.* **23** (10), 3182–3196. https://doi.org/10.1007/s12555-025-0032-7 (2025)
46. Rasheed, W. & Tang, T. B. Anomaly detection of moderate traumatic brain injury using auto-regularized multi-instance one-class SVM. *IEEE Trans. Neural Syst. Rehabil. Eng.* **28** (1), 83–93. https://doi.org/10.1109/TNSRE.2019.2948798 (2019).
47. Hamdan, N. & Sipos, T. Advancements in Machine Learning for Traffic Crash Severity Prediction: A Comprehensive Review. *Periodica Polytech. Transp. Eng.* **53** (3), 347–355. https://doi.org/10.3311/PPtr.40369 (2025).
48. Jin, G. et al. Spatio-temporal graph neural networks for predictive learning in urban computing: A survey. *IEEE Trans. Knowl. Data Eng.* **36** (10), 5388–5408. https://doi.org/10.1109/TKDE.2023.3333824 (2023).
49. De Medrano, R. & Aznarte, J. L. A new spatio-temporal neural network approach for traffic crash forecasting. *Appl. Artif. Intell.* **35** (10), 782–801. https://doi.org/10.1080/08839514.2021.1935588 (2021).
50. Han, W., Zhou, X., Zhang, A., Zhao, L. & Pan, G. AE-TabNet: A Two-stage Framework for Traffic Crash Severity Prediction with Imbalanced Data. In *Proceedings of the 2024 2nd International Conference on Frontiers of Intelligent Manufacturing and Automation* (pp. 301–306). https://doi.org/10.1145/3704558.3704581 (2024)
51. Pei, Y., Wen, Y. & Pan, S. Traffic crash severity prediction based on interpretable deep learning model. *Transp. Lett.* **17** (5), 895–909. https://doi.org/10.1080/19427867.2024.2398336 (2025).
52. Santos, D., Saias, J., Quaresma, P. & Nogueira, V. B. Machine Learning Approaches to Traffic Crash Analysis and Hotspot Prediction. *Computers* **10** (12), 157. https://doi.org/10.3390/computers10120157 (2021).
53. Liang, O. S. & Yang, C. C. How are different sources of distraction associated with at-fault crashes among drivers of different age gender groups? *Crash Anal. Prev.* **165**, 106505. https://doi.org/10.1016/j.aap.2021.106505 (2022).
54. Bakiş, E., Erçetin, M. A., Acar, E., Gökalp, İ. & Yılmaz, M. Prediction of traffic crashes trend with learning methods: a case study for Batman, Turkey. *Sci. Rep.* **15** (1), 26566. https://doi.org/10.1038/s41598-025-11835-9 (2025).
55. Liu, J. Importance-SMOTE: a synthetic minority oversampling method for noisy imbalanced data. *Soft. Comput.* **26** (3), 1141–1163. https://doi.org/10.1007/s00500-021-06532-4 (2022).
56. Chen, T. XGBoost: A Scalable Tree Boosting System. *Cornell University*. https://doi.org/10.1145/2939672.2939785 (2016)
57. Pes, B. & Lai, G. Cost-sensitive learning strategies for high-dimensional and imbalanced data: a comparative study. *PeerJ Comput. Sci.* **7**, e832. https://doi.org/10.7717/peerj-cs.832 (2021).
58. Sujon, K. M., Hassan, R., Choi, K. & Samad, M. A. Accuracy, precision, recall, f1-score, or MCC? empirical evidence from advanced statistics, ML, and XAI for evaluating business predictive models. *J. Big Data*. **12** (1), 268. https://doi.org/10.1186/s40537-025-01313-4 (2025).
59. Subramanian, V., Mascha, E. J. & Kattan, M. W. Developing a clinical prediction score: comparing prediction accuracy of integer scores to statistical regression models. *Anesth. Analgesia*. **132** (6), 1603–1613. https://doi.org/10.1213/ane.0000000000005362 (2021).
60. Hooker, G., Mentch, L. & Zhou, S. Unrestricted permutation forces extrapolation: variable importance requires at least one more model, or there is no free variable importance. *Stat. Comput.* **31** (6), 82. https://doi.org/10.1007/s11222-021-10057-z (2021).
61. Bedane, T. T Road Traffic Accident Dataset of Addis Ababa City. *Mendeley Data*. **V1** https://doi.org/10.17632/xytv86278f.1 (2020).
62. Nguyen, S. Accident-Severity-Prediction (curated CSV release of the Road Traffic Accidents dataset). GitHub repository. https://github.com/sonnguyen129/Accident-Severity-Prediction (2022)

Acknowledgements

The author gratefully acknowledges the Batman Provincial Police Department for providing the traffic accident data used in this study. The views and conclusions expressed in this paper are those of the author and do not necessarily reflect the official policies or positions of the data provider.

Author contributions

Veysel Gider conceived and designed the study, performed data acquisition and analysis, interpreted the results, and drafted the manuscript. Serdar Ekinci, Davut İzci, İslam Gökalp, Mostafa Rashdan and Mohammad Salman contributed to the interpretation of the results and critically reviewed the manuscript. All authors reviewed and approved the final version of the manuscript.

Funding
No funding was received for this work.

Declarations

Competing interests
The authors declare no competing interests.

Ethical approval
This study was conducted using secondary administrative traffic crash records obtained from the Turkish Ministry of Interior, General Directorate of Security, upon an official written request submitted by Assoc. Prof. Dr. İslam Gökalp (Batman University, Faculty of Engineering–Architecture, Department of Civil Engineering) dated 25.09.2023. The requested dataset covered fatal and injury-related traffic crashes that occurred between 2013 and 2022 within the police jurisdiction area of the central district of Batman province and was officially provided to the university in Excel format for scientific and academic research purposes. The dataset consisted exclusively of retrospective and fully anonymized records and did not contain any personally identifiable information (such as names, national identification numbers, addresses, or contact details). No direct interaction with human participants was involved, and no interventions, experiments, or behavioral measurements were conducted. In accordance with national regulations and institutional guidelines governing the secondary use of anonymized administrative data, and because the study involved neither identifiable personal data nor prospective data collection from human subjects, the need to obtain formal ethical committee approval was waived by Batman University and the data-providing authority. For the same reasons, and due to the retrospective and anonymized nature of the dataset, the need to obtain informed consent from individual participants was also waived by the data-providing authority and the affiliated institution. All procedures performed in this study complied with applicable institutional and national ethical standards for research involving secondary administrative data.

Additional information
Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-026-55751-y.

Correspondence and requests for materials should be addressed to M.S.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc-nd/4.0/.

© The Author(s) 2026