A Hybrid Digital Twin Framework for Road Accident Prediction in Butembo (DR Congo) Using Machine Learning, Graph Neural Networks, and Spatio-Temporal Simulation

Nsenge Mpia Héritier[*,1], Kasoki Luhala Christelle[1], Mbusa Murumba Kombi[1], Katembo Kasolene Moïse[1], Kyamakya Kyandoghere[2], Wojciech Kustra[3], Wido Hamel[4]

[1] Université de l'Assomption au Congo (UAC), Democratic Republic of the Congo

[2] Alpen-Adria Kagenfurt University (AAU), Austria

[3] Gdansk University of Technology (GUT), Poland

[4] Bauhaus Universitaet Weimar (BUW), Germany

***Corresponding author:** Nsenge Mpia Héritier, nsengempia@uaconline.edu.cd

Keywords: Hybrid Digital Twin for Road Safety, Road Accident Prediction in Data-Constrained Cities, Machine Learning and Ensemble Modelling, Graph Neural Networks and Spatio-Temporal Analysis, AfroTrans project.

Background

Road traffic accidents remain a persistent public safety challenge in many African cities undergoing rapid urban growth. These environments are often characterised by informal transport systems, heterogeneous road infrastructure and uneven enforcement of traffic regulations. Although accident data are commonly collected by local authorities, their use is often limited to descriptive reporting, limiting their value for preventive and strategic road safety planning. This situation is particularly evident in medium-sized cities such as Butembo, located in eastern Democratic Republic of the Congo, where motorcycle taxis dominate urban mobility and where traffic conditions vary markedly across space and time. In recent years, digital twin concepts have been promoted as powerful tools for integrating data, predictive models and simulation to support decision-making in transport systems. However, most existing applications focus on well-instrumented cities with access to dense sensor networks and high-quality real-time data. In Sub-Saharan African contexts, such conditions are rarely met due to data sparsity, inconsistent reporting practices and limited technical resources. This gap calls for approaches that can adapt the digital twin paradigm to data-constrained urban settings while retaining practical relevance for policy analysis.

Aim

The aim of this study is to develop a hybrid digital twin framework for road accident risk analysis in the city of Butembo. The proposed framework seeks to combine machine learning, spatio-temporal modelling, and simulation to predict accident severity and explore the potential

effects of safety interventions. A central objective is to demonstrate that an operational digital twin can be constructed using routinely collected accident records, without reliance on advanced sensing infrastructure, and that such an approach can support evidence-based road safety planning in similar urban contexts.

Method

The study is based on 10,168 road accident reports recorded by municipal authorities between 2014 and 2025. Each report includes information on the time of occurrence, road segment, vehicle characteristics, driver condition, road surface state and accident severity. Personal identifiers were removed prior to analysis. Due to limitations in reporting detail, accident severity was modelled as a binary outcome distinguishing fatal from non-fatal accidents. Significant preprocessing was required to ensure data consistency and usability. Time information was standardised to an hourly scale from 0 to 23, categorical variables were harmonised, and missing values were handled using conservative imputation strategies. All preprocessing steps were implemented programmatically to ensure transparency and reproducibility. The modelling framework integrates three complementary components. Gradient boosting models (XGBoost, LightGBM and CatBoost) are employed to capture nonlinear relationships in the structured accident data. Temporal dynamics are represented using a Long Short-Term Memory (LSTM) network trained on aggregated hourly accident frequencies, allowing the identification of recurrent intra-day risk patterns. Spatial dependencies are modelled using a Graph Neural Network (GNN), where road segments are treated as nodes connected by geographical adjacency. Predictions from the three components are combined using a weighted ensemble approach, producing spatio-temporal probabilities of fatal accidents for each road segment and hour of the day. These probabilities feed a simulation layer that updates risk states over a 24-hour cycle and enables the evaluation of counterfactual scenarios, such as targeted infrastructure improvements, speed management measures or temporary road segment closures.

Results

Model performance is assessed using stratified train–validation–test splits to preserve class proportions, complemented by a temporal validation strategy in which earlier records are used for training and more recent records for testing. Evaluation metrics include the F1-score, the area under the ROC curve and probability calibration measures. Results indicate that each modelling component contributes distinct information: gradient boosting models capture contextual and categorical effects, the LSTM identifies temporal concentration of risk, and the GNN accounts for spatial correlation between neighbouring road segments. The ensemble consistently delivers more stable and better-calibrated predictions than individual models.

Conclusions

The study demonstrates that a functional digital twin for road accident analysis can be developed using existing administrative data in resource-constrained urban environments. By combining machine learning, spatio-temporal modelling and simulation, the proposed framework supports both predictive risk assessment and exploratory evaluation of safety interventions. The approach provides a practical and transferable pathway for strengthening data-driven road safety planning in medium-sized African cities.