IJARST
International Journal For Advanced Research In Science & Technology
A peer reviewed international journal www.ijarst.in
ISSN: 2457-0362

DATAFITS A HETEROGENEOUS DATA FUSION FRAMEWORK FOR TRAFFIC AND INCIDENT PREDICTION

[1]DR.L.MALLIGA, [2]E.SOWMYA, [3]I.ASHRITHA, [4]K.SARIKA

[1]Assistant Professor, Department of Electronics and Communication Engineering,Malla Reddy Engineering College For Women, Maisammaguda, Dhulapally Kompally, Medchal Rd, M, Secunderabad, Telangana.

[2,3]Student, Department of Electronics and Communication Engineering,Malla Reddy Engineering College For Women, Maisammaguda, Dhulapally Kompally, Medchal Rd, M, Secunderabad, Telangana.

ABSTRACT— This paper introduces DataFITS (Data Fusion on Intelligent Transportation System), an open-source framework that collects and fuses traffic related data from various sources, creating a comprehensive dataset. We hypothesize that a heterogeneous data fusion framework can enhance information coverage and quality for traffic models, increasing the efficiency and reliability of Intelligent Transportation System (ITS) applications. Our hypothesis was verified through two applications that utilized traffic estimation and incident classification models. Data-fits collected four data types from seven sources over nine months and fused them in a spatiotemporal domain. Traffic estimation models used descriptive statistics and polynomial regression, while incident classification employed the k-nearest neighbors (k-NN) algorithm with Dynamic Time Warping (DTW) and Wasserstein metric as distance measures. Results indicate that DataFITS significantly increased road coverage by 137% and improved information quality for up to 40% of all roads through data fusion. Traffic estimation achieved an R2 score of 0.91 using a polynomial regression model, while incident classification achieved 90% accuracy on binary tasks (incident or non-incident) and around 80% on classifying three different types of incidents (accident, congestion, and non-incident).

Index Terms— Intelligent transportation systems, heterogeneous data fusion, traffic estimation, incident classification.

I. INTRODUCTION

Data availability is a critical aspect in the design of modern Intelligent Transportation Systems (ITSs), which implement models to understand better various patterns of the transportation system [1], thus improving mobility and safety for people and goods. With modern society depending heavily on efficient and reliable transportation, the importance of these systems has seen a rapid increase in significance over recent years. In Germany alone, both the number of registered cars and the number of carried passengers using public transportation, have shown a substantial increase, reaching their all-time Manuscript received 7 February 2023; revised 20 April 2023; accepted 25 May 2023. This work was suppported by the Bundeswehr through Federal Office of Bundeswehr Equipment, Information Technology, and In-Service Support (BAAINBw) and Bundeswehr Technical Center for Information Technology and Electronics (WTD81). The Associate Editor for this article was T. Tettamanti. *(Corresponding author: Paulo H. L.*

*Rettore.).*Philipp Zißner and Paulo H. L. Rettore are with the Communications Systems Department, Fraunhofer FKIE, 53177 Bonn, Germany(email:philipp.zissner@fkie.fraunhofer.de;paulo.rettore.lopes@fkie.fraunhofer.de). Bruno P. Santos is with the Department of Computer Science, Federal University of Bahia, Salvador 40170-110, Brazil (e-mail: bruno.ps@ufba.br). Johannes F. Loevenich is with the Communications Systems Department, Fraunhofer FKIE, 53177 Bonn, Germany, and also with the Department of Mathematics/Computer Science, University of Osnabrück, 49074 Osnabrück, Germany (e-mail: johannes.loevenich@fkie.fraunhofer.de). Roberto Rigolin F. Lopes is with the Secure Communications and Information (SIX), Thales Deutschland, 71254 Ditzingen, Germany (e-mail: roberto.rigolin@thalesgroup.com). Digital Object Identifier 10.1109/TITS.2023.3281752 highs of 48.5 million cars (2022) and 12.7 billion carried passengers (2019, before the pandemic) [2], [3]. As a result, urban areas experience an increasing number of traffic-related incidents (e.g., congestion and accidents), increasing time delays, emissions, and fuel consumption [4]. For this reason, academia and industry have driven efforts to create the next generation of transportation systems that are eco-friendly, cost-efficient, and powered by data analysis and communication technology. We hypothesize that a heterogeneous data fusion framework can enhance the coverage and quality of information serving as input for traffic models, thus increasing the efficiency and reliability of ITS applications.Therefore, we propose the Data Fusion on Intelligent Transportation System (DataFITS) framework, providing a spatiotemporal fusion of data used to train models for two ITS applications, traffic estimation, and incident classification.

DataFITS collects and combines real heterogeneous data (e.g., weather, traffic, incident) from various sources (e.g., open databases, map applications), preparing them by fixing errors, adapting the data structure, and finally fusing them in the exact

location and point in time. Our hypothesis is verified using data characterization to quantify the benefits of combining heterogeneous data sources and the proposal of two ITS applications. The performance of the two applications ratifies the benefits of larger data coverage/quality while estimating traffic and classifying incidents. Thus, the main contributions of this investigation are:

- An open-source framework DataFITS for heterogeneous spatiotemporal data fusion, covering the acquisition, processing and fusion of data, available in a public code repository.1
- The characterization of a heterogeneous dataset combining real traffic data from two cities in Germany, collected from seven sources over nine months and provided together with the repository.
- Two traffic estimation models, one using descriptive statistics and another using polynomial regression with different parameters such as time, road type, and weather, and a comparison between single and fused datasets.

II. RELATED WORK

This section reviews the literature on three main topics related to our proposed solution: (i) data collection and fusion, (ii) traffic estimation, and (iii) incident classification.

Finally, we summarize and compare the literature with our proposal.

A. Data Collection and Fusion

To develop ITS applications, significant data is required from real or virtual sensors [5]. Vitor et al. [4] present a platform to collect, process, and export heterogeneous data from smart city sensors, providing different statistics and visualizations. However, their platform concentrates on securing data. Similarly, [6] proposes a smart city data platform containing information from various cities. In contrast to our framework, we focus on improving the quantity and quality of the information by fusing data, and we assess the advantages of using fused data through two ITS applications. Data fusion combines data from multiple sources, enrichingspatiotemporal information [7], [8], [9], [10]. Several applications benefit from data fusion, such as emergency management [11] and path planning [12]. However, fusing heterogeneous data requires additional preprocessing to combine various data types and features [13], [14]. This investigation focuses on two applications supported through data fusion: traffic estimation and incident classification, and the methods to achieve their goals, such as data acquisition, fusion, machine learning, correlation, and different data types.

B. Traffic Estimation

Traffic estimation is a crucial smart city application for better transportation management. This review focuses on data fusion, spatiotemporal correlation, and machine learning techniques to achieve accurate and reliable traffic estimation using historical data. The increasing availability of open databases (kept by governmental authorities) and Application Programming Interfaces (APIs) to commercial applications (Bing, Google Maps, etc.) results in a vast collection of trafficrelated data, making big data an opportunity for heterogeneous data fusion [15]. The challenge is to combine stationary sensor data (e.g., traffic cameras or loop detectors) and probe vehicle information (e.g., cameras, GPS, cellular data, or vehicular sensors). Anand et al. [16] used a Kalman filter to fuse traffic flow values (from cameras) and travel time (from GPS), improving a traffic estimation approach. Many recent traffic estimation models use Machine Learning(ML) [17], [18], [19], [20], [21], [22], [23], [24], [25]. Reference [17] proposes an auto-regressive model that uses data from a traffic simulator and adapts to events like accidents.Their results showed that estimation up to 30 minutes ahead has an error of 12%. Meanwhile, [18] employs deep learning algorithms for traffic estimation, showing an improvement of accuracy and efficiency. These approaches discuss the usage of ML to create accurate models for traffic estimation, but do not consider further methods, such as data fusion, correlation, etc. Some ML approaches use spatiotemporal correlation to improve traffic estimation quality. In [19], a neural network(NN)-based estimation using Graph Convolutional Network (GCN) and Gated Recurrent Unit (GRU) models is proposed with full public access. The GCN captures spatial dependencies from the road network, and GRU detect dynamic changes in traffic data and captures temporal dependencies. Other NN-based approaches, such as [20] and [21], show similar improvements in accuracy using data correlation. Wang et al. [22] propose an open-source deep learning framework using GCN to estimate network-wide traffic multiple steps ahead in time. Zheng et al.

[23] introduce another opensource solution, the Graph Multi Attention Network (GMAN), using an encoder-decoder architecture to provide long-term traffic estimation up to one hour ahead. These approaches also include correlation to improve the discussed models and offer access to their data but do not propose a solution for collecting or fusing data. Limited literature combines data fusion, spatiotemporal correlation, and ML to estimate traffic, similar to our solution. In [26], the authors fuse traffic data from stationary and dynamic sensors, considering the spatiotemporal correlation between traffic levels of road segments.A Multiple Linear Regression (MLR) model processes the fused information to enhance traffic estimation accuracy. Unlike our solution, this approach relies solely on traffic data from sensors but does not consider different data types and sources. Zhao et al. [24] propose a general platform for spatiotemporal data fusion to enhance traffic estimation. The approach introduces a fusion method to improve accuracy by combining direct and indirect traffic-related data as input for two different ML models. The indirect traffic-related data features contain information about weather and points of interest and are used to improve the estimation quality. However, their model uses pre-existing datasets, offering no solution for data collection, and our study focuses on incident-related data, while the authors in [24] consider points of interest and weather conditions. In [27], the authors introduce a model to estimate traffic within a small urban area in Zurich, with data acquired as part of a video measurement campaign. Their solution fuses information from Loop detectors, traffic lights, and other sensors (e.g., video plus license plate recognition, thermal cameras) and trains different MLR models with this data. Finally, they evaluate the various sensors' accuracy and robustness. In contrast to our solution, they investigate the quality of a regression model using different sensor data fused to stationary data. Furthermore, their data is acquired using sensors that are not publicly available, covering only a small urban area.Finally, [25] proposes a traffic speed prediction by integrating heterogeneous data from various sensors, including exogenous data like weather, into a hybrid spatiotemporal features space. The main contributions are a hybrid model using Long short-term memory (LSTM) and GRU, comparing the model against other well-known classical deep learning models, showing the highest efficiency and lowest error metric. In contrast to our study, this investigation focuses on the prediction using only vehicle speed and has no open access to their solution and the data.

C. Incident Classification

Numerous ML and deep-learning models are also used for incident classification [28], [29], [30], [31]. These models improve road safety in urban areas by facilitating traffic management, warning systems, and emergency rescue operations. Other applications, such as incident detection, are proposed in [32] and [33], which provide additional traffic management enhancements, including the ability to control traffic lights from emergency vehicles. In [28], the authors introduce a Convolutional Neural Network (CNN) model to predict traffic accidents using a state matrix with influencing traffic features. Their solution achieves high prediction accuracy, but limited training data affects CNN model quality, which could be improved by using data fusion. Park et al. [29] propose a big data approach using the *Hadoop framework* to combine incident-related and other traffic data. The study classifies data into groups of traffic incidents. Data fusion benefits the approach, but incorporating spatiotemporal aspects could further increase model accuracy. In [30], the authors propose a recurrent neural network to predict traffic accident risk by combining incident data with a spatiotemporal traffic correlation. The model has high accuracy and can be used for accident prevention and integrated into traffic control systems. However, its main limitation is the consideration of only directly-related incident data. Other traffic-related features (e.g., traffic flow, weather, vehicular data, etc.) could be fused to improve accuracy. Shang et al. [31] propose a hybrid approach for automatic incident detection using random forest-recursive feature elimination and a LSTM network with Bayesian optimization. Their approach provides an accurate binary classification of incidents, outperforming other state-of-the-art solutions, but does not classify them into different types. Other approaches use location-based social media (LBSM) to improve the detection and classification of incidents. Rettore et al. [34] propose a framework containing two data services, one to detect traffic-related events. The framework collects data from social media platforms (e.g., Twitter), which is used in a road incident detection model based on heterogeneous data fusion to provide more descriptive transportation system data. The free access of user data through Twitter's API improves the availability of incident data, an essential aspect in developing ITS solutions. Also, in [35], the authors describe a real-time traffic event detection solution using Twitter posts. Their solution is based on a text classification algorithm to identify traffic-related tweets with their location and classify the information into different classes of events.

D. Comparison & Summary

Table I summarizes the reviewed literature, categorizing them into five applications: smart city, emergency, traffic estimation, incident classification, and our solution. The second and third columns list the key aspects and the corresponding references. The remaining columns denote the following labels: *Data Acquisition*, *Data Fusion*, *ML*, *Correlation*, *Stationary*, *Probe*, and *LBSM*. These labels indicate whether the approach collects data, uses data fusion techniques, utilizes ML and deep-learning models, incorporates data correlation, employs stationary sensor data, uses probe vehicle data, or utilizes georeferenced social media data. Moreover, we classify the availability of the source code and data of all solutions into three categories using different colors *no*, *limited*, or *yes* public access. A paper labeled with *no public access* does not offer access to their data or solution, unlike solutions that provide *full public access* to source code and data. *Limited public access* describes the usage of datasets that are not accessible anymore or solutions that plan to offer open access in theory but currently do not fulfill this aspect. The last row of Table I compares our investigation with the literature, highlighting the coverage and contributions of our proposed solution. Compared with most of the literature, we provide a methodology that covers four of five stages of the data cycle (acquisition, preparation, processing, use) [13], providing an open-source framework,1 and access to the collected datasets. Making the models and datasets available, or the means to acquire and process them, is crucial to enable a fair comparison between models/methodologies, which we did not find in most literature. Moreover, the DataFITS framework is designed to support multiple data types, including stationary and probe data, and can potentially incorporate additional types of information like LBSM. We perform spatiotemporal data fusion to provide enriched information used as an input for two, but not limited to, data applications showing the benefit of using fused heterogeneous data. In contrast, other approaches in the literature focus on specialized solutions that combine only a subset of the listed features in the context of ITS.

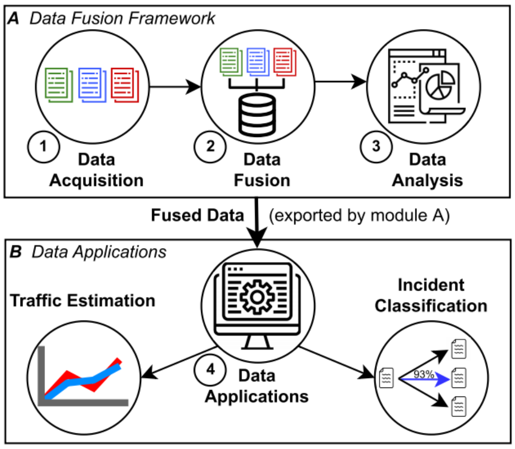

Fig :Workflow

III. THE DESIGN

This research proposes a solution including two different modules: A data fusion framework DataFITS and two data applications traffic estimation and incident classification. The DataFITS design follows a three-stage workflow, as presented in Fig. 1-A. It starts by gathering data from heterogeneous transportation-related data sources using APIs and web crawlers (1). In sequence, all acquired data are fused geographically by mapping them to road segments and aligned temporally (2). After fusing the data, we can perform data analysis to identify and visualize specific data characteristics (e.g., traffic and incident statistics) (3). DataFITS can export data which then can be used as input for different applications, depicted in Fig. 1-B. In this article, we use the fused data in two applications: traffic estimation and incident classification that can benefit from fused data (see Section III-B) providing a more comprehensive perspective of the results (4).

A. Data Fusion Framework

1) Data Acquisition: Within the data acquisition, Fig. 2 (1),DataFITS collects information from different predefined data sources according to a set of user-defined parameters (e.g., geographical area and time interval). Currently, DataFITS supports multiple methods to collect traffic, incident, vehicular, and weather data. In addition, the framework parses heterogeneous information and stores them in standardized CSV files. The acquisition follows a modular application design, ensuring easy expandability of the framework functionalities and allowing the specification of additional data sources.

2) Data Preparation: The compiled data undergo an additional preparation step as illustrated in Fig. 2 (2). The key component of the preparation stage is data standardization, converting different feature names and types into a uniform representation and a set of user-customizable data mappings to deliver consistent data types. In sequence, the data is prepared to be mapped onto geographical locations. Leveraging OpenStreetMap (OSM), a free map database, DataFITS gathers shapefiles according to the bounding box parameter specified in the data acquisition stage, using OSMNX [40]. A shapefile stores road network information identified by the primary key (*fid*) for each road segment, which is used in the map-matching procedure conducted within the data processing. DataFITS can also extract the road type and speed limit from shapefiles. Finally, the collected data is converted into "trip files", a representation of the input used by the mapmatching.

3) Data Processing:

a) Temporal fusion: Fig. 2 (3) displays the temporal data fusion. This process groups the complete data within an arbitrary time window aggregation (e.g., hourly, daily, or 10 minutes for the results in this paper), adapting the time interval from the collection process.

b) Spatial fusion: DataFITS leverages the map-matching technique, taking GPS points and aligning them to established coordinates under a predetermined degree of accuracy based on an underlying road network. This results in a balanced level of accuracy and associate all geo-located data with the same road network. Among different strategies of map-matching, DataFITS integrates Fast Map Matching (FMM), an open-source tool, which provides two different algorithms for achieving optimal performance based on the given road network size [41]. To this end, FMM uses the trip and shapefiles created in the prior stage and connects all input data points to a corresponding road network. Each data entry within the trip file contains a *Linestring* representing the GPS coordinates (path) of a road segment, except for incident data entries, which only contain coordinates of a start and end point. In addition to the matched points of each input entry, the algorithm returns two arrays, *opath* and *cpath*, that contain a set of road identifiers (*fids*) from the OSM. The first array, *opath*, stores the *fid* for each matched point, representing a list of road segments that got matched to the input data entry (data source coordinates with OSM road map). The *cpath*, second array, stores the *fid* values that create a path between all matched road segments.This process is performed on each record of the vehicular and incident data sources, while the geo-location of the traffic data sources is only matched once for each area, as those are static and do not change between data acquisitions. This strategy significantly reduces the execution time and computation required for map-matching. Instead of processing all data points within each acquisition, the main amount of data points, namely the traffic-related information, is only matched once. On our nine-month data time frame and a 10-minute acquisition time, this reflects a single matching procedure instead of 38,800 procedures, significantly reducing the runtime and required computation power.

4) Data Usage: The last stage, (4) in Fig. 2, describes different use cases of the fused dataset, e.g., as an input to various data applications or being characterized through different types of statistics and visualizations for spatiotemporal data analysis. For example, DataFITS provides heat maps and density plots separated by each source and different features, such as the number of observations, traffic levels, speed, and types of incidents. In the scope of temporal analysis, DataFITS provides time-series statistics for a specific time window (e.g., by the hour, day of the week, month, and season) and shows the correlation between different features. Moreover, the fused data is exported in different data structures, allowing to be used by various data applications, such as our proposed models or other third-party tools (e.g., *ArcGIS*).

B. Applications

1) Traffic Estimation: The proposed traffic estimation application is organized into two phases, as shown in Fig. 3. Phase (1) prepares the data, groups it by intersecting areas,identifies similar traffic regions based on correlating traffic patterns and performs a train-test-split. A traffic region is defined as the set of connected paths (road segments) reported from a data source, represented through unique road identifiers (*fids*). By intersecting areas, we obtain a list of unique traffic regions and are able to measure the similarities between them. In phase (2), the prepared data is used to create and evaluate two traffic estimation models using: i) descriptive statistics (naive); and ii) polynomial regression. Each model estimates traffic values for a single area within an arbitrarily defined time interval and can also utilize data from correlating regions with similar traffic behavior. Furthermore, the process considers optional input parameters like weekday, weather, and road type to createmore specific models for the given characteristics. This research mainly focuses on the regression-based model Fig. 3. Design of the traffic estimation application. but also discusses the model based on descriptive statistics. and gives a comparative evaluation of both approaches in section IV.

a) Preprocessing: The fused data from DataFITS is cleaned, removing all incident-related information, as it is not required by the model, and grouped into traffic areas containing one or multiple road segments. Using a data aggregation over the array of road identifiers (*cpath*), we create a list of areas contributing traffic information to the dataset. Due to the traffic area grouping, the data may contain overlapping areas due to the data fusion that merges traffic areas from different sources. Those intersecting areas describe the same spatial region but with minor differences in the covered road segments. Combining them removes potential duplicated areas, resulting in a final set of unique traffic areas. The underlying function iterates through all existing areas, calculates pairwise intersections, and combines them if the overlapping road segments exceed a predefined threshold *thoverlap*. Finally, the initial set of fused data is re-grouped according to the new set of combined traffic areas, resulting in an input dataset for the traffic estimation models that contains the combined information for each area.

V. CONCLUSION

In this paper, we introduce DataFITS, an open-source data fusion framework that integrates diverse data by collecting, analyzing, and fusing it. We hypothesize that heterogeneous data fusion increases data quantity and quality, thereby improving datasets for ITS applications. To verify this, we developed two ITS applications: one used polynomial regression to estimate traffic levels, while the other combined traffic and incident data to classify events into accident, congestion, or non-incidents. Using real heterogeneous data from two German cities, we quantified the advantages of DataFITS by compiling a fused dataset. Our results indicate that DataFITS integrated

data from multiple sources for 40% of all roads, thereby increasing the overall road coverage by 137%. In addition, the traffic estimation model, which uses polynomial regression, outperformed our previous approach based on descriptive statistics, achieving a high R2 score of 0.91, low error metrics of 0.05, and provides accurate traffic estimations using the fused dataset. Compared to using a single sources dataset, the fused dataset estimation showed minor accuracy improvements but drastically improved the spatiotemporal coverage of the estimated areas. Our incident classification model relies on the fusion of traffic and incident data, achieving a 90% binary classification accuracy rate within our evaluation. Preprocessing the data, such as removing unclear traffic patterns, improved accuracy by an average of 29%. The classification of incidents into different categories resulted in a slightly lower accuracy of 86%, with unequal performance among classes indicated by F1 scores. To mitigate this problem, we oversampled the training dataset to create a more uniform representation of the data, resulting in an 80% accuracy for each class. Collecting

more accident data can also solve this problem. We plan to expand the DataFITS framework by collecting and fusing more data types, improving its performance and data quality, and expanding its data analysis. We focus on data types such as social media and images, which require methods such as Natural Language Processing (NLP) and image processing. For ITS applications, we aim to use automated machine learning to explore different models and hyper-parameters and compare them with our current models. We also plan to analyze the correlation between traffic and incidents and incorporate it into the traffic estimation models. In addition, we intend to explore the use of big data in military scenarios, combining information from the civilian and military fields to support strategic operations in urban warfare. To this end, our framework can be enhanced to collect and combine different types of information (image, text) to create common operational pictures and verify/authenticate information, thereby avoiding misinformation that may influence political decisions.

REFERENCES

[1] L. Zhu, F. R. Yu, Y. Wang, B. Ning, and T. Tang, “Big data analytics in intelligent transportation systems: A survey,” *IEEE Trans. Intell. Transp. Syst.*, vol. 20, no. 1, pp. 383–398, Jan. 2019.

[2] Umweltbundesamt. (2022). *Verkehrsinfrastruktur und fahrzeugbestand*. Accessed: Dec. 12, 2022. [Online]. Available: https://www.umweltbundesamt.de/daten/verkehr/verkehrsinfrastruktur fahrzeugbestand

[3] German Federal Statistical Office (Destatis). (2022). *Passengers Carried in Germany*. Accessed: Jul. 12, 2022. [Online]. Available: https://www.destatis.de/EN/Themes/Economic-Sectors Enterprises/Transport/Passenger-Transport/Tables/passengerscarried.html

[4] G. Vítor, P. Rito, and S. Sargento, “Smart city data platform for real-time processing and data sharing,” in *Proc. IEEE Symp. Comput. Commun. (ISCC)*, Sep. 2021, pp. 1–7.

[5] A. B. Campolina, P. H. L. Rettore, M. Do Val Machado, and A. A. F. Loureiro, “On the design of vehicular virtual sensors,” in *Proc. 13th Int. Conf. Distrib. Comput. Sensor Syst. (DCOSS)*, Jun. 2017, pp. 134–141.

[6] S. Jeong, S. Kim, and J. Kim, “City data hub: Implementation of standard-based smart city data platform for interoperability,” *Sensors*, vol. 20, no. 23, p. 7000, Dec. 2020. [Online]. Available:

https://www.mdpi.com/1424-8220/20/23/7000

[7] L. Zhang, Y. Xie, L. Xidao, and X. Zhang, “Multi-source heterogeneous data fusion,” in *Proc. Int. Conf. Artif. Intell. Big Data (ICAIBD)*, May 2018, pp. 47–51.

[8] P. H. L. Rettore, B. P. Santos, A. B. Campolina, L. A. Villas, and A. A. F. Loureiro, “Towards intra-vehicular sensor data fusion,” in *Proc. IEEE 19th Int. Conf. Intell. Transp. Syst. (ITSC)*, Nov. 2016, pp. 126–131.

[9] P. H. L. Rettore, A. B. Campolina, L. A. Villas, and A. A. F. Loureiro, “A method of eco-driving based on intra-vehicular sensor data,” in *Proc. IEEE Symp. Comput. Commun. (ISCC)*, Jul. 2017, pp. 1122–1127.

[10] P. H. L. Rettore, A. B. Campolina, A. Souza, G. Maia, L. A. Villas, and A. A. F. Loureiro, “Driver authentication in VANETs based on intravehicular sensor data,” in *Proc. IEEE Symp. Comput. Commun. (ISCC)*, Jun. 2018, pp. 00078–00083.