IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received January 16, 2022, accepted January 29, 2022, date of publication February 4, 2022, date of current version February 14, 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3149059

Entropy-Based Traffic Flow Labeling for CNN-Based Traffic Congestion Prediction From Meta-Parameters

MOUNA ZOUARI MEHDI[1], HABIB M. KAMMOUN[2], (Senior Member, IEEE), NORHENE GARGOURI BENAYED[3], DORRA SELLAMI[1], (Senior Member, IEEE), AND ALIMA DAMAK MASMOUDI[1]

[1]CEM Laboratory, National Engineering School of Sfax, Sfax University, Sfax 3038, Tunisia
[2]REsearch Groups in Intelligent Machines (ReGIM-Laboratory), National Engineering School of Sfax, Sfax University, Sfax 3038, Tunisia
[3]Research Digital Center of Sfax, Technopole of Sfax, Sfax 3021, Tunisia

Corresponding author: Mouna Zouari Mehdi (mouna.zouari.enis@gmail.com)

ABSTRACT Traffic congestion affects quality of life by inducing frustration and wasting time. The congestion is also critical to vehicles with high emergencies such as ambulances or police cars. This leads to additional CO2 emissions. Traffic management requires the accurate modeling of congestion levels. Two main observable parameters identify the congestion state of a city: vehicle speed and density. Congestion has an intuitive definition rather than a quantitative one, and is associated with the disorder and randomness occurring in traffic parameters. Therefore, statistical analysis offers an efficient and natural framework for modeling such disorders. In this study, a differential-entropy-based approach was proposed for labelling purposes. Subsequently, supervised congestion prediction from traffic meta-parameters based on a convolutional neural network was proposed. Traffic parameters includes node localization, date, day of the week, time of day, special road conditions, and holidays. The proposed model is validated on the CityPulse dataset, which is a set of vehicle traffic records, collected in Aarhus city in Denmark over a period of six months, for 449 observation nodes. Simulation results on the CityPulse dataset illustrate that the proposed approach yields accurate prediction rates for different nodes considered. The proposed system can prevent traffic congestion by reorienting the drivers to follow other itineraries.

INDEX TERMS Shannon entropy, deep CNN, smart city, traffic congestion state prediction.

I. INTRODUCTION

Recent years, population explosion, and the increase in vehicles as a transport mode have caused a tremendous increase in the number of vehicles, leading to problems in traffic such as congestion, environmental pollution, noise, and mobility latency. Congestion problems can also cause traffic non safety under certain circumstances and event occurrences. Such problems affect the quality of life of citizens and can be penalized for the economy. In Brazil, congestion causes economic losses, estimated at 80 $ billion (BRL) per year. In Europe, such loss is about 2 % of its domestic product (GDP), and in USA, it is about 160 billion $ [1], while the infrastructure issue for congestion mitigation is far from ensuring an optimal use of resources, it can be very expensive and requires time to be ready. Traffic management system (TMS)-based solutions can raise such challenges by optimal management of existing resources.

The explosion of IoT technologies makes traffic management systems a modern alternative that can assist in congested situations [2]–[4]. IoT technologies are used to collect data from the traffic nodes. They handle many types of sensors, to detect the number of vehicles and their speed. These parameters allow for the direct observation of the traffic state., A roadside unit (RSU) is often used for analyzing congestion through cloud computing and displaying it to users through mobile applications. Several technologies can be used for this purpose d. Radio-frequency identification detection n (RFID) can be used to detect vehicles using RFID tag, and the results can be made available on a network. Light detection and ranging (LIDAR) sensors can determine vehicle count and traffic state. The number of vehicles and their speed can be

The associate editor coordinating the review of this manuscript and approving it for publication was Michail Makridis.

estimated using infrared (IR) sensors. The DGT sensors can also sense the flow of vehicles.

It is worth noting that the development of a traffic flow prediction model can be applied to many scenarios. Apart from the centralized collection system, which is an inter-vehicle communication system, can be used to estimate the congestion level for traffic flow optimization in a transport system. This information on congestion is disseminated between the vehicles. The main goal of such a challenging distributed system is to improve traffic state visibility using a direct inter-vehicle network (Internet of Vehicles) and then reduce the travel time of vehicles by recommending alternative paths at lower congestion levels [1], [5]. Such systems have not yet been generalized and are not within our application framework. However, our proposed prediction model can be applied to such scenarios without a lack of generality.

Other new trends in traffic monitoring systems are based on visual information obtained from video cameras. These systems, also called incident detection systems (IDS), are based on video frame analysis to detect congestion or traffic jams in less than 5s [2], [6], [7]. Some commercial products are available but are proprietary and expensive [2]. In our study, we consider non-visual information for their availability at a lower cost and, thus, their ability to be easily generalized in a whole city at a reasonable cost. In this study, our major research concern was the development of reliable decision-making systems for traffic congestion time prediction. Several studies have considered the study of traffic congestion from different perspectives [8]. One research challenge in a congestion-state modeling system is the lack of an accurate definition of congestion out of the intuitive common definition. The latter can lose its accuracy within a quantitative interpretation, which is an essential step in modeling.

With a deeper understanding of traffic flow, we can note that when congestion occurs, the vehicle speed as well as the vehicle count decrease. However, prior to this step, crowding of vehicles occurs, leading to a relative increase in their density [9]. Then, a recurrent process takes place, where vehicle speed and density influence each other; the higher the crowd (vehicle number), the higher the decrease in the vehicle speed, causing in this way more crowding, where the speed may reach very low values, producing either severe congestion or traffic paralysis. Subsequently, a relative increase in the number of vehicles is observed.

Currently, while evacuating congested vehicles, the speed can increase until it reaches mean values and completely recovers from congestion. It is worth noting that in such mutual interaction between speed vehicles and their density, a small delay is observed. Thus, the series product of vehicle speed and vehicle count cannot be informative regarding congestion. A time series analysis allows an understanding of the intrinsic dynamics of the traffic model, while prediction systems prevent traffic congestion by identifying the congestion level based on constraints or meta parameters.

To tackle the problem of meta-parameters that interfere highly in the congestion state, an important step lies in their identification. The random aspects of some of these parameters also induce prediction errors. An obvious ascertainment is that given the dependency of such a process on human will, we cannot get rid of or even re-conciliate with the uncertainty aspect of traffic flow. Rather, we can look at predicting some influences and evolution or progress of the flow, with respect to given scenarios and parameters. Regarding all human activities, smart city transport is highly influenced by weather conditions. However, most smart city datasets do not include records of meta-parameters. For example, no previous work has addressed the correlation between traffic flow and weather conditions in a general framework. Nevertheless, weather conditions, such as date and month, are implicitly incorporated into meta-condition records. The question of traffic congestion prediction for prevention purposes remains open. Research challenges include a more accurate and adaptive definition of congestion and more efficient identification of meta-parameters.

In this paper, we propose a short-time congestion state labelling system based on vehicle speed and count, followed by CNN prediction based on meta-parameters. Our research context is a smart city urban project aiming at decision making based on different source modalities. First, congestion is modeled with a differential entropy-based approach with respect to traffic parameters which are vehicle speed and traffic density. Entropy-based approaches ensure prediction efficiency even at a highly random traffic parameters. Then, a deep neural network is devised for mean speed prediction from traffic meta-parameters. Our proposed approach achieves good performance over existing approaches and allows congestion flow prediction with higher accuracy.

The remainder of this paper is organized as follows: In Section 2, we present related work and preliminaries. In Section 3, we provide a general overview of the proposed method. In Section 4, we focus on the entropy-based bloc for labelling and cellular neural network for congestion prediction. The experimental results are illustrated in Section 5. Finally, in Section 6, we present our conclusions and discuss future work.

II. RELATED WORK

Researchers of smart city traffic modeling have explored earlier linear and statistical theories [9], [10] [11], [12]. With the maturity of recent advanced algorithms and artificial intelligence, and the emergence of big data owing to advances in data collection technology, a new background of traffic big data has been created, allowing researchers to propose more accurate approaches for smart city modeling and prediction, as well as more accurate investigations of urban traffic conditions and their interference in traffic flow [1], [8], [13], [14] [15], [16].

In [17], Kidando *et al.* highlighted that research on traffic flow prediction has transformed from linear and nonlinear prediction to intelligent prediction, exploiting the multimodal

and heterogeneous big data available in digital smart city platforms, thus making a revolution in prediction. Nevertheless, such a problem can be tackled by the disposal of the ground truth. Given the number of samples, automatic labelling is an important issue. In [16], a review of existing traffic prediction approaches was presented, and some possible future development trends were identified. Hawes *et al.* addressed the problem of traffic states estimation within segments of road using a particle filter and traffic measurements at the segment boundaries [18], [19]. Authors consider the scenario where traffic measurements are missing and propose an approximation based on the mean of the historical measurements from a suitable time period. Simulation results shows significant improvement of the traffic state estimation accuracy.

Other related work were developed based on statistical approaches for smart city congestion analysis. In [10], the authors considered the modeling of traffic flow on holidays. They estimated traffic patterns based on the assumption of a Gaussian mixture speed distribution using the expectation maximization algorithm (EM). This study claims that the EM can be used only under binomial traffic flow and for a proper node context, which is the actual framework of their work. However, for normal traffic scenarios, the binomial aspect of traffic is not straightforward in terms of the speed distribution. Because the EM algorithm does not apply any prior knowledge on practical values of vehicle speeds, it can then generate two different components, whose average speed values are very close to each other, identifying other Gaussian mixtures than the one corresponding to the congestion. Thus, the authors were constrained to fix the threshold value between the Gaussian centers in the EM algorithm. Another limitation of the Gaussian mixture model is that it depends on the choice of bin number in the histogram. Exceeding the bin number, a discontinuity in the histogram makes the EM algorithm identify false mixtures [10]. In addition, the authors illustrated the speed difference and mixing proportions in the computed mixture model in their results. An in-depth analysis of these results indicates that the higher the speed difference, the more severe the congestion state. In [20], the authors evaluated the influence of traffic density, in addition to vehicle speed, on traffic flow conditions, based on the Bayesian Dirichlet process mixtures of generalized linear models (DPM-GLM). Their study estimated the speed cut-off point values of the traffic state using a Bayesian change-point detection (BCD) technique.

Based on the BCD model, they computed the possible threshold speed values, separating traffic states into homogeneous groups, facilitating the classification of traffic conditions. In their results, the authors illustrated the change point detection. Nevertheless, they neither indicate the adopted time window widths for estimating the mixture model parameters nor illustrate change-point detection on real speed data transiting between different congestion states. In their experimental results, the authors illustrated all the parameters of the EM model and the different computed cutoff points. Nevertheless, the authors did not use any ground truth or experimental setup to confront their findings. Although statistical approaches have a significant ability to learn internal characteristics from traffic distributions, they cannot handle complex traffic nodes with more than two modalities at different locations.

Recently, the increasing demand for smart city application efficiency has leveraged data collection systems, making the application of advanced knowledge-based approaches possible. We describe the following recent approaches. In [13], the authors proposed a smart city management system based on artificial neural networks for traffic state classification and genetic algorithms for vehicle flow optimization. The system is devoted to intelligent vehicle network systems but can be used for any smart city application. A system integrating different intelligent systems, thus named by the author multi-intelligent system.A genetic-algorithm-based decision is proposed for more efficient infrastructure and optimal resource use. In [14], authors propose a data prediction system for dynamic planning algorithms in a smart city transportation context. Short-term traffic prediction refers to estimation of the basic parameters of traffic within 15 min s in the future. Accordingly, the authors applied a k-NN classifier to improve prediction accuracy by avoiding the influence of subjective classification in traffic state forecasting. This approach tackles the problem of the lack of an exact definition of congestion.

In [1], the k-NN classifier was s used in conjunction with two other classifiers: the fuzzy classifier and the ANN-MLP classifier. This aggregation involved a voting criterion that considered t the average performance of each classifier. Three levels of congestion were established: moderate, between the rate and congestion d, and yielding a faster convergence. Similarly, in [4], the authors introduced a multi-perspective framework for IoT data fusion approach classification, which can be devoted to smart city application, where: a common technique for data fusion is IoT data association, based on the similarity between data sources.

In [21], authors combine a random forest algorithm with density based spatial clustering for short-term traffic congestion prediction. The DBSCAN estimate the level of congestion, based on vehicle speed, by dividing the space into clusters with sufficient sample density, while the random forest is used to predict the vehicle speed and vehicle density. Moreover, authors do not integrate in their study any environmental parameters, but rather propose a historical prediction (as authors call it), i.e. a time-dependent model, of mean average speed vehicles. In [22], the problem of traffic congestion prediction in smart city transportation systems was n addressed. In this work, the authors applied two binary classifiers, namely, the support vector machine (SVM) and the Multinomial Naïve Bayes (MNB), within a conventional machine learning scheme, for traffic fluid classification of manually labeled samples. The SVM classifier yielded a better performance. It is worth noting g that there are two types of short-time prediction n methods: (i) a parametric prediction based on a certain mathematical model or (ii) non-parametric

knowledge-based intelligent prediction method. The latter is not explicit or predefined but has a strong ability to model nonlinearity. Moreover, smart cities can be approached from several aspects: time series method analysis of traffic flow parameters, non-parametric regression method, and artificial neural networks for traffic congestion prediction from meta-parameters.

The above related works have been validated on small non-exhaustive datasets, including either short-interval l traffic or limited recorded meta-parameters, neglecting some relevant influencing parameters. Indeed, although in [10], a deep understanding of traffic flow complexity is carried out, it lacks an experimental validation of the claimed hypothesis and the proposal. Similarly, in [11], the validation of the proposed approach was based on archived one-year traffic data collected on a single traffic way, Florida's interstate e freeway corridor. In [23], is composed of 1124 data points from different sections of the Shanghai traffic management information department, taking into consideration as model attributes the weather conditions, time peaks, holiday conditions, other special conditions, and road quality. Given a reduced validation dataset, the performance results cannot confirm the efficiency of such an approach. However, it is interesting to note that this approach can sort the relative importance of influencing factors.

Another important research issue is related to the influencing factors, also called meta-parameters, which interfere highly in the congestion state and externally influence the traffic. Nevertheless, in some contexts of traffic vulnerability, the control system service, designed for controlling the traffic, is no longer efficient, and the traffic flow prediction becomes more complex owing to such internal factors [16]. An accurate definition of a meta-parameter conditions the accuracy of prediction but is not faithfully considered in the literature. In [15], the authors explored the effects of weather conditions on a particular segment of user activity in a smart city, which is a cyclist in London. They applied a set of machine learning classifiers, such as Bayes networks, nearest neighbors, and J48, to correlate data to weather conditions. Although all the different classifiers yield good prediction performance, J48, based on data entropy-based analysis, offers the best performance.

However, it exhibited the lowest performance in terms of the considered time interval for building the model. In [9], the authors proposed a method for mining dependencies between smart city parameters (highway input vehicle count and exit vehicle count) by considering the time lag as an influencing parameter and by devising an expectation-maximization (EM) algorithm to learn the model parameters. A prediction model has been improved, considering two kinds of hierarchical characteristics: feature level and time level. Indeed, we can cope with the randomness caused by internal and external parameters by intuitively learning their influence through data by applying short-time non-parametric models. Indeed, some obvious characteristics can be identified through closer observation, such as the periodicity of traffic congestion, suggesting the existence of internal rules that can simplify the analysis.

Another important issue in smart city modeling based on supervised approaches is annotation. Many authors have generated labels for the used dataset without describing their annotation method, despite the impact of an improper label generation process on accuracy. In [14], although the proposed model reduces complexity, no quantitative assessment of its accuracy was undertaken in this study, and the authors did not indicate how they built the labeled samples. In [22], given that the most critical point in existing methods is improper label generation, the authors considered an intuitive solution for label generation based on the average speed of vehicles and their density. The authors generated a new feature, which is the normalized product of the average vehicle speed by the average vehicle count. They performed a semi-supervised process, starting with manual labeling of small sample rates, by thresholding this feature. The data instance is labeled congested for values less than a threshold of 0.5. In [22], the authors are based on a set of instances generated from the publicly available traffic dataset CityPulse relative to Aarhus city in Denmark [5], [24], [25]. In this study, the process of generating the ground truth is important. Although congestion is intuitively clear, there is no explicit definition in the literature. However, the rules applied for labelling are cursory and shallow.

III. GENERAL OVERVIEW OF THE PROPOSED SYSTEM

A first challenge in our work was to undertake accurate labelling of the dataset. Direct and accurate labelling can be based on a differential entropy measure for both traffic conditions: vehicle speed and vehicle density, as shown in Figure1.

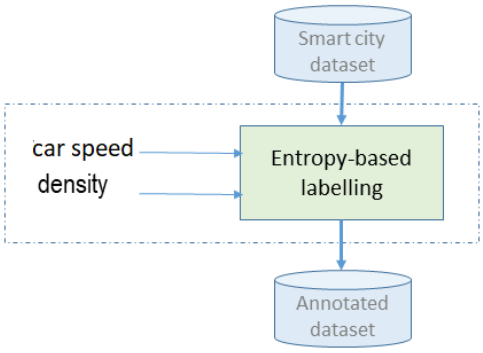


FIGURE 1. The differential entropy based labelling bloc.

In the following, we describe differential-entropy-based labelling module and CNN-based prediction models.

A. DIFFERENTIAL ENTROPY BASED LABELLING

For performance assessment, a ground truth should be established and compared with the experimental results of the proposed automatic congestion prediction algorithm. In the case of a small dataset, the ground truth is generally manually

TABLE 1. Comparative study of smart city approaches.

References	Year	Theme and proposed method	Results
Jungwook Jun [10]	2010	Traffic congestion detection, Expectation Maximization	-
Yingjie Xia et al. [12]	2016	Traffic flow estimation, Video based traffic flow estimation	-
Liu, Yunxiang and Wu, Hao [23]	2017	Traffic congestion prediction, Random Forest	87.5 %
Chin Jeannette et al. [15]	2017	Effect of weather data (raining), J48	100 %
Xiabing Zhou et al. [9]	2017	Intelligent Transportation System, Expectation Maximization	-
Kidando Emmanuel et al. [11]	2017	Traffic congestion detection, Bayesian model	-
Hawes, Matthew et al. [18]	2017	Traffic state estimation, Particle filter and Bayesian statistical inference	-
Chen, Xiangyang and Chen, Ruqing [16]	2019	Intelligent Transportation System, Review	-
Billy Pik Lik Lau et al. [4]	2019	Data fusion in smart city, Review	-
Izhar Aamish et al. [22]	2020	Label generation for congestion prediction, SVM and MNB	99 %
Daming Li et al. [13]	2020	Smart city managment, Genetic algorithm and Neural Networks	-
Geraldo P. Rocha Filho et al. [1]	2020	traffic congestion detection, Ensemble classifier	94%
Sun, Ning et al. [14]	2020	Dynamic planning based on data prediction, KNN	-
Shenghua, Huang et al. [21]	2020	Traffic congestion prediction, DBSCAN and Random Forest	94.36 %

performed by experts in a particular domain. In such situations, experts will make use of implicit rules for congestion estimation or the segmentation of data. Because we have big data, we attempt to identify the rules that are intuitively applied in a congestion flow estimation problem.

It is worth underlying that the congestion is a subjective definition. It cannot be associated with particular values of the speed or the count; indeed, while a mean speed value of 50 km/h in a town road at a limitation rule of 50 km/h, denotes a very high value for a case of fluid traffic, the same speed in a highway can be a sign of a heavy congestion state. Generally, congestion denotes a qualitative judgement reflecting the user perception of the flow state, associated with frustration caused by an extra delay over a normal flow situation. A closer understanding of this state yields semantic annotation of the traffic state.

Since the congestion state is associated with a decrease in speed and settling within a time interval corresponding to congestion degradation. When the speed decreases permanently or is subject to an overall decrease over a relatively short time interval, we can decide that the traffic is undergoing a congestion state. Now, the more the speed decreases, the heavier is the congestion. One can then fix an upper limit under which the traffic state is congested and a minimum time interval of the speed decrease to avoid false congestion detection, which is not followed by a vehicle crowd or by a recurrent influence between speed and vehicle account. Now, the more the speed decreases, the heavier is the congestion.

Similarly, the more lasting is the decrease, the more blocked is the traffic flow. Mathematically traducing such intuitive definition of the congestion is rather a semantic annotation of data.

In our case, the labelling was based on dual traffic conditions: vehicle speed and vehicle density. Although it has no interest in prediction because it is based on instantaneous direct measures of vehicle speed and vehicle density, this model can help prediction by offering a direct labelling method for different samples.

Therefore, to localize events within a signal frame, Shannon's entropy should be helpful in distinguishing random variations induced by traffic congestion. Figure 3 illustrates the signal randomness associated with the condition state.

Thus, our hypothesis is that the signal corresponding to congestion has potential information and reflects a specific event that differs from ordinary traffic flow conditions. Hence, the corresponding entropies were different. Thus, a differential-entropy-based analysis can be efficient for congestion detection.

In the following, we present the concept of applying the entropy for labelling.

Shannon entropy was introduced as a metric to modelling the amount of information in a given source [26]. It can consider the entire dataset of some particular sets.

Let o_n be an observed event with a corresponding probability p_n with $n = 1, \ldots, L$. The entropy H is expressed by Equation (1).

$$H = \Sigma_1^L p_n \log_2(p_n) \tag{1}$$

Hence, based on a statistical analysis of the vehicle speed and vehicle density signals, from our real-world dataset,

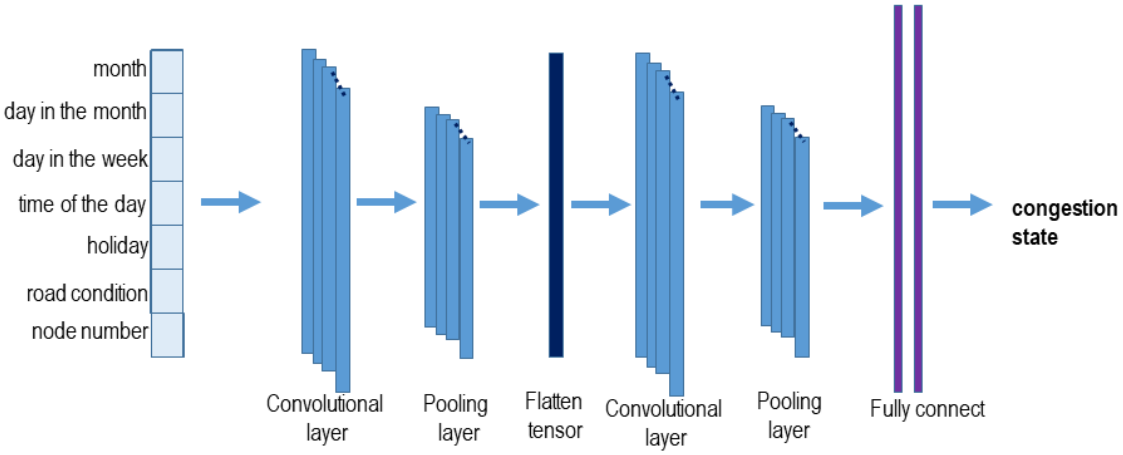


FIGURE 2. CNN architecture.

we found that the entropy of the differential signal is subject to raise when a congestion occurs. Therefore, we can deduce that for a given vehicle speed time-signal, we statistically have a differential entropy much higher in a congestion-related window, than in normal traffic flow. Accordingly, the vehicle speed and density time signals are partitioned into fixed-size overlapping sliding windows. The sliding window's size is empirically defined based on the congestion evolution and dynamics. As the overlapping sliding windows should keep the useful information in a continuous way, we take as a forward stride one sample. Subsequently, the Shannon's entropy of the differential signals was calculated for each window. Then, we localize the maxima occurring in the entropy, in a sliding window range and select only a one maximum per sliding window, to prevent false detection. Therefore, to maintain only relevant entropy peaks, a fixed threshold is adopted for assigning either the label fluid or traffic congestion label.

Based on the above description, we deduce a labeling algorithm of the traffic flow state. The corresponding algorithm is illustrated in the following:

In this algorithm, in addition to using the entropy of the differential speed and density signals as key conditions for assigning the congestion level, we assume additional hypothesis on the decrease in speed, the decrease in speed (through the speed derivative signal) follows: $\Delta Vh_{speed} \leq -\Delta S_{th}$, while the induced change in the density signal follows: $\Delta Vh_{density} \geq \Delta D_{th}$, where S_{th} and D_{th} are constants that are empirically adjusted. Following these hypothesis, when both the differential entropy of speed vehicle and density vehicle, exceeds D_{th}, the congestion state is labelled as high. Otherwise, if the corresponding entropy are lower that the same threshold D_{th}, the congestion state is labelled as Fluid. Through an adequate empirical adjustment of the different constants in this algorithm, we end by labelling the different samples in the dataset.

Algorithm 1 Detect and Estimate a Congestion State

$cong \leftarrow Fluid$, initialize cong
$H_{dV_{speed}}$, compute the histogram of dV_{speed}
$H_{dV_{density}}$, compute the histogram of $dV_{density}$
$P_{dV_{speed}}$, compute the probability distribution of dV_{speed}
$P_{dV_{density}}$, compute the probability distribution of $dV_{density}$
D_{ES}, compute the differential entropy of dV_{speed}
D_{ED}, compute the differential entropy of $dV_{density}$
Require: $dVh_{speed} \leq 0$ for $\Delta_{time} \geq \Delta_0$
Require: $dVh_{density} \geq 0$ for $\Delta_{time} \geq \Delta_0$
while $\Delta Vh_{speed} \leq -\Delta S_{th}$ **do**
  while $\Delta Vh_{density} \geq \Delta D_{th}$ **do**
    if $D_{ES} \geq D_{th}$ and $D_{ED} \geq D_{th}$ **then**
      $Cong \rightarrow High$
    else
      if $D_{ES} \leq D_{th}$ or $D_{ED} \leq D_{th}$ **then**
        $cong \rightarrow Medium$
      else
        if $D_{ES} \leq D_{th}$ and $D_{ED} \leq D_{th}$ **then**
          $cong \rightarrow Fluid$
        end if
      end if
    end if
  end while
end while

For further assessment, we manually annotated a randomly selected subset of our dataset. Next, we compared the labelling results with those of the manual target.

B. CNN DEEP NEURAL NETWORK BASED PREDICTION MODEL FROM META-PARAMETERS

Because of the multitude of meta-parameters, the traffic flow is subject to large randomness. Therefore, the neural network classifier framework offers the most reliable prediction tool,

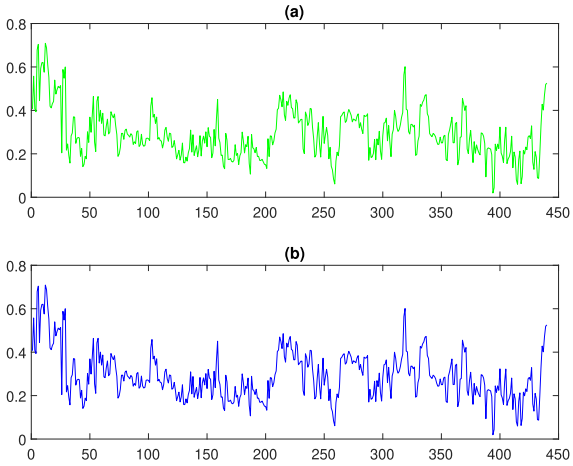


FIGURE 3. Normalized mean traffic conditions on the different major roads of the smart city used dataset: (a) mean vehicle speed, (b) mean vehicle count.

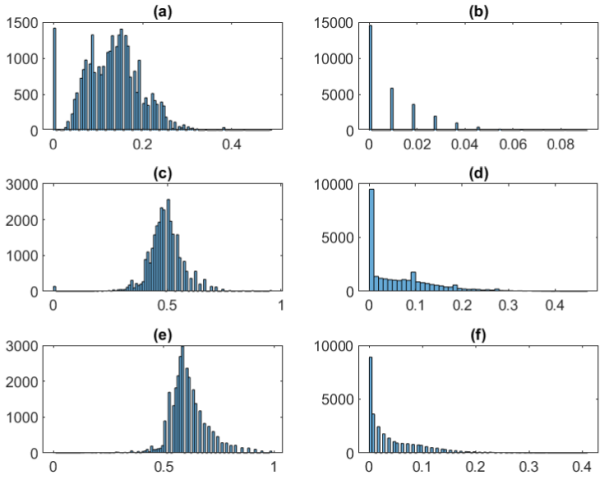


FIGURE 4. Normalized vehicle speed and count distribution for three major roads of different congestion levels: (a) and (b) Vehicle mean speed and count distribution for the major road of the dataset in the file trafficData190126.csv, respectively, (c) and (d) Vehicle mean speed and count distribution for the major road of the dataset in the file trafficData190501.csv, (e) and (f) Vehicle mean speed and count distribution for the major road of the dataset in the file trafficData197544.csv respectively.

enabling a built-in adaptive system to any potential amendment in meta-parameters, such as sudden weather change events or non-scheduled social events. The CNN is conventionally applied to a 2D image signal, for its ability to capture specific local patterns in the whole signal. A similar structure for 1D time-series can be used. The architecture of the deep CNN is illustrated in Figure 2. It is composed of an input layer, a first convolutional layer, a pooling layer, a flattened layer, a second convolutional layer, a second pooling layer, a fully connected layer, and output layer. The set of meta-parameters, including node localization, date, day in the week, time of day, special road conditions if there are any, and holidays, are fed into the convolutional layer. A set of kernels produce a feature map. Then, the pooling layer is intended for reducing the size of the feature map. In the flatten layer, we get a one dimensional signal that will be fed to the second convolutional layer, and applied to a pooling layer. Finally, the fully connected layer is fed into the output layer, producing the congestion state. In this study, the activation function is the Rectified Linear Unit (ReLU). Accordingly, all labelled datasets were used for training a 1-D convolutional Neural Network.

IV. EXPERIMENTAL RESULTS

A. DATASET

The authors classified data sources into four categories: (i) physical data sources collected from sensors on a physical platform, and (ii) cyber data sources commonly obtained from the Internet domain, such as social media information. (iii) Participatory data sources including crowd-sensing and crowdsourcing, sensor-devices fixed on vehicles, and (iv) hybrid data sources. All the above issues concern different applications in smart cities, where prediction is an important block intended for optimization, risk management, and efficient management.

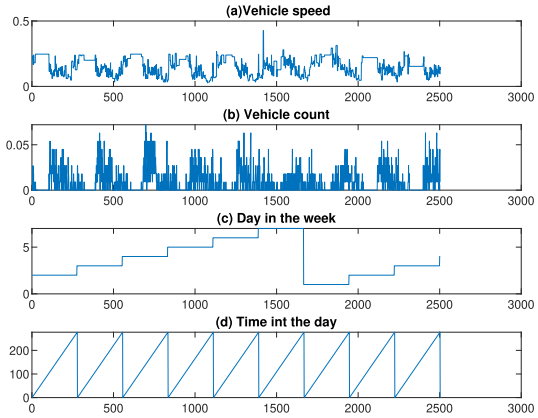


FIGURE 5. Meta-parameters extracted from the dataset attributes highlighting the periodicity in the traffic road for a low flow traffic node.

In our study, the vehicle traffic datasets were provided by the city of Aarhus, Denmark. The datasets are publicly available in comma-separated values (CSV) raw format and the semantically interpreted format provided in the framework of the CityPulse EU FP7 project. In addition, it is licensed under the Attribution of Creative Commons Attribution 4.0 International License. For further details, one can refer to several related studies [24], [25], [27] [5] [28]. In this study, we used different raw datasets relative to the 455 nodes, each available from February 2014 to June 2014. Each dataset is a collection of observations of traffic conditions between two different points in high-roads, for the above-mentioned time durations. This study include data records of 5-min intervals. Each dataset consists of nine attributes, namely, ''status'', ''avgMeasuredTime'',

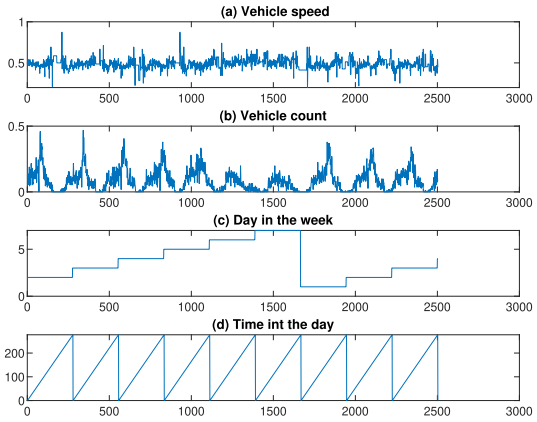


FIGURE 6. Meta-parameters extracted from the dataset attributes highlighting the periodicity in the traffic road for a high flow traffic node.

''avgSpeed'', ''exitID'', ''medianMeasuredTime'', ''TIMESTAMP'', ''vehicleCount'', ''Xid'' and ''REPORTID''.

Figure 3 illustrates the mean vehicle speed and count speed records for the different major roads of the considered smart city set in the CityPulse dataset. As can be seen, the mean values of the speed and count vehicles are sparse, suggesting different categories of traffic flows and potentially different internal rules.

Figure 4 illustrates the distribution of the mean vehicle speed and count speed on three typical major roads with different traffic aspects inherent from the road aspect. In Figure 4 (a) and (b), the major road considered has a lower mean speed and heavy traffic, resulting in a multimodal distribution of the mean speed. Indeed, the distribution has more than two modalities, and the lower modality corresponds to the congestion mode. Figure 4 (c) and (d), corresponding to a major road of medium mean vehicle speed and less heavy traffic, we can observe that the multimodal aspect of the mean speed distribution is less marked than in the first case. This is confirmed for the third major road considered in Figure 4 (d) and (e), corresponding to a major road with higher vehicle mean speed and a more fluid traffic. Besides, if the bin number in the above histograms is reduced, we can identify only one mode, corresponding to a fluid mixture.

B. EXPERIMENTAL RESULTS

A set of pre-processing steps were conducted for structuring the metadata and resorting more pertinent parameters to the congestion context. Indeed, out of the given nine traffic conditions, we have used in our work the conditions that interfere with the congestion state which are ''avgMeasuredTime'', ''avgSpeed'' (averagespeed) measured in km/h, ''medianMeasuredTime'' and ''vehicleCount''. In addition, in order to make more visible the natural periodicity carried in the attribute 'Time', we have generated other history attributes such as 'dayInTheWeek', 'day', 'TimeInTheDay', and 'month'. In the attribute 'dayInTheWeek', we aim to relieve the intuitive information weekend, which is generally governed by other traffic rules. For example, people on weekends can adopt other transport means and follow different mobility timings. For the attribute 'TimeInTheDay', we map the cyclic information during a day and the potential peaks relative to jam traffic incidents.

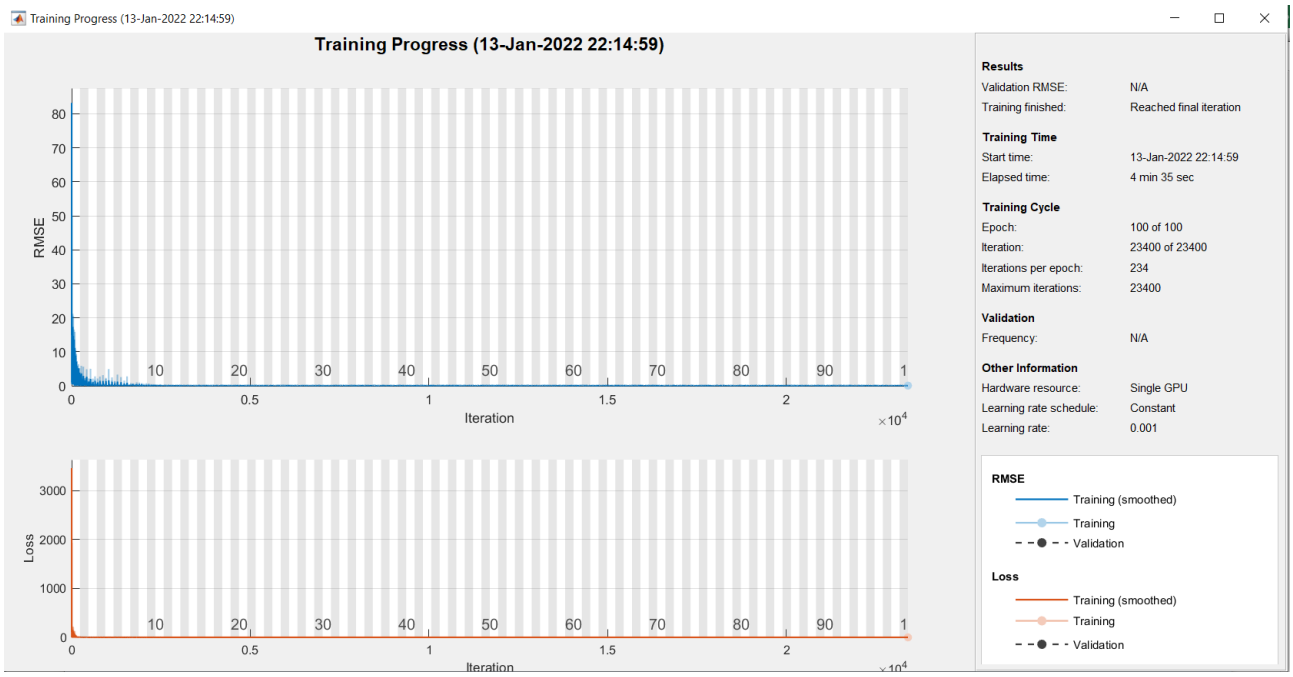


FIGURE 7. Learning progress of the deep CNN.

Each 'TimeInTheDay' cycle includes a record of 278 samples, each separated from the next, by 5 min time interval. In the attribute 'day' and 'month', we implicitly incorporate the dependency with respect to weather conditions and the duration of the light in a day. Indeed, the weather as well as sunrise and sunset times, which highly influence traffic hours, are variable during months and days in the month. Therefore, using these new attributes, we recover from the missing meta-conditions and gather a more complete set of traffic meta-conditions for an accurate prediction. The last attribute that we add is the node number, which is intended to relieve the spatio-localization dependency of traffic. Indeed, the difference in road width, state, commercial, residential, or other vocations can cause an internal difference in the traffic governing rules.

The deep CNN has been trained and tested at a sample number of ratios of data of 80% and 20% respectively for training and testing. The network parameters are listed in Table 2. The network training progress is illustrated in Figure 7. An acceptable training error is achieved. For illustration, Figure 8 depicts the accuracy prediction rate of different nodes (Node 1, Node 10, Node 50, Node 100, Node 200, Node 350 and node 450). As can be seen, the highspeed way node is the most unpredictable. Further investigation of the influence of the learning time-interval variation on the prediction accuracy is depicted in Figure 9 for a single node (Node 50). As can be seen, shortening the time interval increased the prediction quality. This effect can be attributed to the fact that less parameter variability is encountered in this case.

TABLE 2. The deep CNN model parameters.

convolution1	(3x3x16)
Activation function	Relu
Fully connected layer 1	384
Fully connected layer 2	100
Output	1

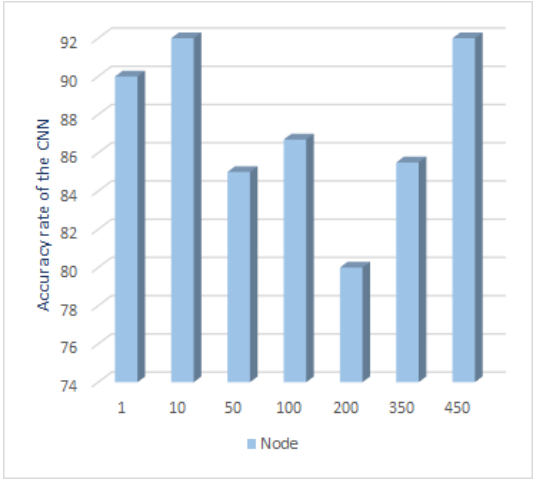


FIGURE 8. Prediction rate of some nodes in the city network (Node 1, Node 10, Node 50, Node 100, Node 200, Node 350 and node 450).

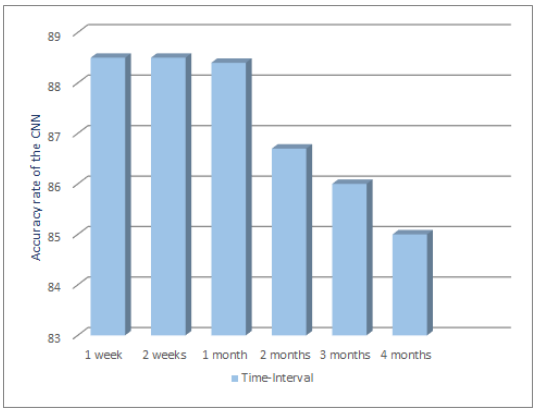


FIGURE 9. Prediction for a single node at different time intervals.

This suggests that reinforcement learning can significantly improve the results.

Figure 5 and 6 illustrate the different extracted meta-conditions with the traffic flow parameters for an example of a major road (extracted from the dataset in the file trafficData190126.csv, which correspond to a low and high-flow traffic on major roads. As can be observed from this figure, our proposed meta-conditions are highly correlated with the dynamics of the flow parameters. As can be observed in these figures, we can identify much easier in the high-flow major road the weekend and the different peaks during a day.

V. CONCLUSION AND PERSPECTIVES

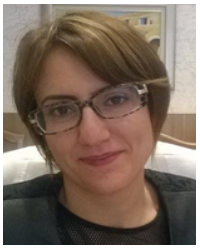
In this study, we have tackled the problem of traffic congestion prediction based on meta-parameters. These conditions include localization, time interval, month, day in the month, time in the day, holiday, road condition, and node number. Accordingly, a differential entropy-based approach is proposed for labelling purposes for two traffic conditions: vehicle speed and vehicle density. Subsequently, a supervised congestion prediction from traffic meta-parameters based on a convolutional neural network is proposed.

We validated our proposed model based on the City-Pulse dataset, which is a collection of vehicle traffic achieved in Aarhus city in Denmark over a period of six months for 449 observation nodes. The simulation results illustrate that the proposed approach yields accurate prediction rates for the different nodes considered. Moreover, shortening the prediction time interval may improve prediction accuracy. The proposed system can help prevent traffic congestion by reorienting users to other itineraries. As a perspective issue to our work, we can improve the prediction accuracy by applying a reinforcement learning paradigm based on recent updates of the prediction process, enabling the integration of the flow evolution data.

REFERENCES

[1] G. P. R. Filho, R. I. Meneguette, J. R. T. Neto, A. Valejo, L. Weigang, J. Ueyama, G. Pessin, and L. A. Villas, "Enhancing intelligence in traffic management systems to aid in vehicle traffic congestion problems in smart cities," *Ad Hoc Netw.*, vol. 107, Oct. 2020, Art. no. 102265.

[2] S. Felici-Castell, M. García-Pineda, J. Segura-Garcia, R. Fayos-Jordan, and J. Lopez-Ballester, "Adaptive live video streaming on low-cost wireless multihop networks for road traffic surveillance in smart cities," *Future Gener. Comput. Syst.*, vol. 115, pp. 741–755, Feb. 2021.
[3] K. Ramesh, A. Lakshna, and P. N. Renjith, "Smart traffic congestion model in IoT a review," in *Proc. 4th Int. Conf. Electron. Commun. Aerosp. Technol.*, 2020, pp. 651–658.
[4] B. P. L. Lau, S. H. Marakkalage, Y. Zhou, N. U. Hassan, C. Yuen, M. Zhang, and U.-X. Tan, "A survey of data fusion in smart city applications," *Inf. Fusion*, vol. 52, pp. 357–374, Dec. 2019.
[5] S. Kolozali, M. Bermudez-Edo, D. Puschmann, F. Ganz, and P. Barnaghi, "A knowledge-based approach for real-time IoT data stream annotation and processing," in *Proc. IEEE Int. Conf. Internet Things (iThings), IEEE Green Comput. Commun. (GreenCom), IEEE Cyber, Phys. Social Comput. (CPSCom)*, Sep. 2014, pp. 215–222.
[6] M. Staniek, "Road pavement condition diagnostics using smartphone-based data crowdsourcing in smart cities," *J. Traffic Transp. Eng., English Ed.*, vol. 8, no. 4, pp. 554–567, Aug. 2021.
[7] W. C. Tchuitcheu, C. Bobda, and M. J. H. Pantho, "Internet of smart-cameras for traffic lights optimization in smart cities," *Internet Things*, vol. 11, Sep. 2020, Art. no. 100207.
[8] S. B. Atitallah, M. Driss, W. Boulila, and H. B. Ghézala, "Leveraging deep learning and IoT big data analytics to support the smart cities development: Review and future directions," *Comput. Sci. Rev.*, vol. 38, Nov. 2020, Art. no. 100303.
[9] X. Zhou, H. Hong, X. Xing, K. Bian, K. Xie, and M. Xu, "Discovering spatio-temporal dependencies based on time-lag in intelligent transportation data," *Neurocomputing*, vol. 259, pp. 76–84, Oct. 2017.
[10] J. Jun, "Understanding the variability of speed distributions under mixed traffic conditions caused by holiday traffic," *Transp. Res. C, Emerg. Technol.*, vol. 18, no. 4, pp. 599–610, Aug. 2010.
[11] E. Kidando, R. Moses, E. E. Ozguven, and T. Sando, "Bayesian nonparametric model for estimating multistate travel time distribution," *J. Adv. Transp.*, vol. 2017, pp. 1–9, Feb. 2017.
[12] Y. Xia, X. Shi, G. Song, Q. Geng, and Y. Liu, "Towards improving quality of video-based vehicle counting method for traffic flow estimation," *Signal Process.*, vol. 120, pp. 672–681, Mar. 2016.
[13] D. Li, L. Deng, and Z. Cai, "Intelligent vehicle network system and smart city management based on genetic algorithms and image perception," *Mech. Syst. Signal Process.*, vol. 141, Jul. 2020, Art. no. 106623.
[14] N. Sun, H. Z. Shi, G. J. Han, B. Wang, and L. Shu, "Dynamic path planning algorithms with load balancing based on data prediction for smart transportation systems," *IEEE Access*, vol. 8, pp. 15907–15922, 2020.
[15] J. Chin, V. Callaghan, and I. Lam, "Understanding and personalising smart city services using machine learning, the Internet-of-Things and big data," in *Proc. IEEE 26th Int. Symp. Ind. Electron. (ISIE)*, Jun. 2017, pp. 2050–2055.
[16] X. Chen and R. Chen, "A review on traffic prediction methods for intelligent transportation system in smart cities," in *Proc. 12th Int. Congr. Image Signal Process., Biomed. Eng. Informat. (CISP-BMEI)*, Oct. 2019, pp. 1–5.
[17] E. Kidando, A. E. Kitali, B. Kutela, M. Ghorbanzadeh, A. Karaer, M. Koloushani, R. Moses, E. E. Ozguven, and T. Sando, "Prediction of vehicle occupants injury at signalized intersections using real-time traffic and signal data," *Accident Anal. Prevention*, vol. 149, no. 4, pp. 599–610, 2021.
[18] M. Hawes, H. M. Amer, and L. Mihaylova, "Traffic state estimation via a particle filter over a reduced measurement space," in *Proc. 20th Int. Conf. Inf. Fusion (Fusion)*, Jul. 2017, pp. 1–8.
[19] M. Hawes, H. M. Amer, and L. Mihaylova, "Traffic state estimation via a particle filter with compressive sensing and historical traffic data," in *Proc. 19th Int. Conf. Inf. Fusion (FUSION)*, 2016, pp. 735–742.
[20] E. Kidando, R. Moses, E. E. Ozguven, and T. Sando, "Evaluating traffic congestion using the traffic occupancy and speed distribution relationship: An application of Bayesian Dirichlet process mixtures of generalized linear model," *J. Transp. Technol.*, vol. 7, no. 3, pp. 318–335, 2017.
[21] H. Shenghua, N. Zhihua, and H. Jiaxin, "Road traffic congestion prediction based on random forest and DBSCAN combined model," in *Proc. 5th Int. Conf. Smart Grid Electr. Automat. (ICSGEA)*, Jun. 2020, pp. 323–326.
[22] A. Izhar, S. M. K. Quadri, and S. A. M. Rizvi, "Hybrid feature based label generation approach for prediction of traffic congestion in smart cities," in *Proc. 3rd Int. Conf. Intell. Sustain. Syst. (ICISS)*, Dec. 2020, pp. 991–997.
[23] Y. Liu and H. Wu, "Prediction of road traffic congestion based on random forest," in *Proc. 10th Int. Symp. Comput. Intell. Design (ISCID)*, Dec. 2017, pp. 361–364.
[24] M. I. Ali, F. Gao, and A. Mileo, "Citybench: A configurable benchmark to evaluate rsp engines using smart city datasets," in *Proc. 14th Int. Semantic Web Conf. (ISWC)*, Bethlehem, PA, USA, 2015, pp. 374–389.
[25] R. Tönjes, P. Barnaghi, A. M. M. Ali, F. G. M. Hauswirth, S. Ganea, B. Kjærgaard, D. Kuemper, S. Nechifor, D. Puiu, A. Sheth, V. Tsiatsis, and L. Vestergaard, "Real time IoT stream processing and large-scale data analytics for smart city applications," Measures Methods Reliable Inf. Process., Tech. Rep., Feb. 2015. [Online]. Available: https://cordis.europa.eu/project/id/609035/reporting
[26] C. E. Shannon, "A mathematical theory of communication," *Bell Syst. Tech. J.*, vol. 27, pp. 623–656, Jul./Oct. 1948.
[27] D. Puiu, P. Barnaghi, R. Tönjes, D. Kümper, M. I. Ali, A. Mileo, J. X. Parreira, M. Fischer, S. Kolozali, N. Farajidavar, F. Gao, T.-L. Pham, C.-S. Nechifor, D. Puschmann, and J. Fernandes, "CityPulse: Large scale data analytics framework for smart cities," *IEEE Access*, vol. 4, pp. 1086–1108, 2016.
[28] S. Bischof, A. Karapantelakis, C.-S. Nechifor, A. Sheth, A. Mileo, and P. Barnaghi, "Semantic modeling of smart city data," in *Proc. 14th Int. Semantic Web Conf. (ISWC)*. Berlin, Germany: W3C Workshop on the Web of Things: Enablers and services for an open Web of Devices, 2014, pp. 25–26.



MOUNA ZOUARI MEHDI was born in Sfax, Tunisia, in 1988. She received the bachelor's degree in math section, in June 2007, and the Ph.D. degree, in 2016. After that, she studied for two years with the Preparatory Institute for Engineering Studies of Sfax Section Math-Physics and succeed with the National Competition for Admission to Engineering Schools to continue her study with the Electrical Department, National Engineering School of Sfax (ENIS), in 2009. Since 2012, she has been qualified as an Electrical Engineer. Her research interests include pattern recognition, image and video processing, and feature extraction and classification.

HABIB M. KAMMOUN (Senior Member, IEEE) received the Ph.D. degree in computer science from the National Engineering School of Sfax (ENIS), University of Sfax, Tunisia. He is currently the Head of the Department of Computer Science, Faculty of Sciences, University of Sfax. He is a member of the Research Groups in Intelligent Machines (REGIM)-Laboratory, University of Sfax. He is a member of the Machine Intelligence Research Labs (MIR Labs), USA. His research interests include soft computing, computational intelligence, intelligent transportation systems, and smart cities. He is a program committee member within IEEE conferences and journals.

NORHENE GARGOURI BENAYED was born in Sfax, Tunisia, in 1985. She received the B.S. degree in electronic from the Faculty of Sciences of Sfax, Tunisia, in 2008, the M.Sc. degree in electrical engineering from the National Engineering School of Sfax (ENIS), University of Sfax, Tunisia, in 2010, and the Ph.D. degree, in 2013. She is currently an Assistant Professor with the Research Center Sfax. Her research interest includes medical image processing applications.

DORRA SELLAMI (Senior Member, IEEE) was born in Sfax, Tunisia, in 1969. She received the degree in engineering from the Sfax National Engineering School, in 1994, the Ph.D. degree in electronics system design, in 1998, and the H.D.R. degree, in 2006. She has been awarded by the President of the Republic of Tunisia. Subsequently, she joined the IMS Laboratory at Bordeaux. Since 1999, she has been an Assistant Professor at the Sfax National Engineering School. Since 2011, she has been a Full Professor. Her research interests include image processing, computer vision and pattern recognition, possibility theory, deep leaning, it covers a large spectrum of applications: hidden biometry and soft biometry, medical image processing, computer aided diagnosis, and fabric defect detection.

ALIMA DAMAK MASMOUDI received the Engineering degree in electrical engineering and the M.Sc. degree in automatic-computer-industrial engineering from the National Engineering School of Sfax, Tunisia, in 2005 and 2006, respectively, and the Ph.D. degree, in 2010. Subsequently, she joined the Control and Energy Management Laboratory (CEM-Laboratory), Computers Imaging and Electronics Systems (CIELS) Groups, to work towards her thesis. Her research interests include image processing applications and neural network implementations. She is currently a Professor with the Physical Department, Faculty of Sciences, University of Sfax.