Computational Framework for Analytical Operation in Intelligent Transportation System using Big Data

Mahendra G[1], Roopashree H. R[2]
Research Scholar, GSSSIETW, Mysuru, India[1],
Department of CSE, GEC, Kushalnagar, India[1],
Associate Professor, GSSS Institute of Engineering and Technology for Women,
Mysuru, India[2]

***Abstract*—Intelligent Transportation System (ITS) is the future of the current transport scheme. It is meant to incorporate an intelligent traffic management operation to offer vehicles more safety and valuable traffic-related information. A review of existing approaches showcases the implementation of varied scattered schemes where analytical operation is mainly emphasized. However, some significant shortcomings are witnessed in efficiently managing complex traffic data. Therefore, the proposed system introduces a novel computational framework with a joint operation toward analytical processing using big data targeting to manage raw and complex traffic data efficiently. As a novel feature, the model introduces a data manager who can manage the complex traffic stream, followed by decentralized traffic management, that can identify and eliminate artefacts using statistical correlation. Finally, predictive modelling is incorporated to offer knowledge discovery with the highest accuracy. The simulation outcome shows that Random Forest excels with 99% accuracy, which is the highest of all existing machine learning approaches, along with the accomplishment of 11.77% reduced overhead, 1.3% of reduced delay, and 67.47% reduced processing time compared to existing machine learning approaches.**

***Keywords*—Intelligent transportation system; traffic managemen; machine learning; artifacts; prediction**

I. INTRODUCTION

With the proliferation of advancements in network and communication technologies, transportation services have been visioned to incorporate smart features [1]. From this perspective, Intelligent Transportation System (ITS) has evolved in the form of innovative services associated with smart traffic management [2]. The prime agenda of ITS is to offer the user the most valuable information associated with traffic to make the driving experiences safer and well-synchronized [3]. Adoption of ITS is found to offer more reduction of cost owing to the adoption of preemptive preventive measures, minimizes the consumption of resources, and significantly emphasizes accident prevention [4]. Various application of ITS includes smart ports and maritime [5], smart airport [6], smart railways [7], fleet management [8], and smart road [9]. The prime idea of ITS is basically to capture all the essential parameters that connect to road navigation and safety and subject them to specific analytical processing. In contrast, the system's outcome of analytical processing is used to formulate a strategy for effective traffic management [10]. At present, various sets of research are being carried out towards improving the performance of ITS from various perspectives [11]-[15]. However, various scattered challenges are associated with the research-based studies and practical implementation of ITS. The research-based studies on ITS have reported various challenges, e.g., the need for cost-effective analytical processing, faster learning schemes with higher reliability, cost-effective computational modelling, and benchmarked schemes. At the same time, the practical implementation of ITS suffers from different challenges, e.g., constraints of budgets, integration with the conventional scheme, coping with updated standards and technology, rules and traffic regulation of different countries, etc. However, there is no denying that the success factor of an ITS majorly depends upon the accuracy and performance of its analytical capability. At present, the big data approach is the most suitable mechanism to deal with complex data [16]; however, they are in a very nascent stage of development. The various complexities associated with the demanded analytical processing of data in ITS are: i) ITS is characterized by the massive generation of traffic data with higher complexities within it, ii) the point of traffic data generation is multiple, and the rate of transmission is uneven and depends upon various network performance leading to error-prone data, iii) with increasing adoption of artificial intelligence and machine learning, the improvement in the predictive analytical model is significantly less owing to the inclusion of critical size of dynamic constraints. Therefore, the proposed scheme presents a novel computational framework of joint analytical operation using a big data approach to improve ITS knowledge discovery. Unlike any existing approaches reported in the literature, the study contributes toward a novel analytical model capable of extracting knowledge from raw and complex forms of traffic data. Another significant contribution of the proposed scheme is offering quality traffic data by performing a fusion of data and identifying and eliminating artefacts. The scheme implements a novel yet simplified machine learning scheme to optimize the predictive performance of traffic data further. The manuscript's organization is as follows: Section II discusses the existing scheme of analytical operations in ITS and highlights research problems in Section III. Section IV briefs about the research methodology, while an elaborated discussion of system design concerning the algorithm is carried out in Section V. Section VI presents the result and justification of the acquired result in the proposed analysis. At the same time, Section VII makes a

conclusive remark about the paper with highlights of novel contributions.

II. Review of Literature

Various work have been carried out to improve the analytical performance of traffic data in ITS. The work by Qi et al. [17] has presented a discussion about the influence of mobility patterns in investigating multiple constraints in traffic management. The study has implemented an ensemble clustering method along with a tensor-based factorization scheme for this purpose. Further adoption of clustering is also witnessed in the work of Huang et al. [18], which emphasizes evaluating the significance of traffic nodes in ITS considering geographic road networks. Liu et al. [19] discuss a unique study model to investigate variability in travel patterns in public transport systems. Analytical modelling is carried out towards quantifying intra-personal variability. Gu et al. [20] have used a deep learning-based approach to predict short-term traffic volume. The study model has used an enhanced Bayesian model integrated with multiple deep learning approaches where correlation analysis is carried out for traffic flow for current and prior time. Chen et al. [21] used the Gated Recurrent Unit (GRU) neural network model to forecast vehicle speed in urban traffic systems. This predictive model facilitates formulating strategies to control traffic and support navigation systems.

All the schemes mentioned above are mainly associated with a single flow of traffic, which are relatively less seen in practical ITS networks. This gap is addressed in the work of Wang et al. [22], where traffic flow estimation is carried out based on trajectories of license plate information embedded with GPS data. The adoption of big data is witnessed in Bakdi et al. [23] work has investigated the possibilities of multiple risk factors associated with ship trials. The study integrated an autonomous identification system with a digital map with a big data approach to realize the spatial and temporal dependencies of various connecting behaviour of vessels. Further adoption of the big data approach is seen in the work of Qin et al. [24], which is meant for tracking routes of vehicles in urban transport systems. The work identifies the vehicle from license plate recognition, while fuzzy logic tracks routes.

It has also been noted that Artificial Intelligence (AI) significantly contributes to the transport system. Besinovic et al. [25] have discussed current insights into railway transport applications using AI. Zhu et al. [26] have presented analytical modelling of traffic data considering transport systems using Hadoop, Spark, and multiple open-source-based software. Predictive analysis is carried out by Long Short-Term Memory (LSTM) and Support Vector Regression (SVR) to investigate various travel-related operations. The work carried out by Gunes et al. [27] has presented an analytical model deployed for the management of traffic lights and scheduling in ITS. The relationship between edge computing and ITS is discussed by Zhou et al. [28], where the study specifies significant challenges in sensing in ITS. The discussion presented by Lucic et al. [29] also stated the usage of Crowdsourcing in ITS. Mirboland and Smarsly [30] developed a novel and simplified information modelling for ITS using a semantic model. Asaithambi et al. [31] have presented a big-data architecture for micro-services for ITS facilitating the processing of both stream and batch of big data. Choosakun et al. [32] have presented insight into cooperative ITS. Yoo et al. [33] have presented an ITS scheme considering sensory data processing using a big data approach based on open-source software. A similar approach is also adopted in the work of Alexakis et al. [34]. Dudek and Kujawski [35] have presented a big data scheme for optimizing route planning and interval of the passage of vehicles considering image features. Table I summarizes the above-mentioned methodologies' effectiveness with respective advantages and limitations. The following section outlines research issues associated with the literature.

TABLE I. Summary of Methodologies Mentioned in Literature

Authors	Method	Advantage	Drawbacks
Qi et al. [17]	Ensemble Clustering, tensor	Simplified evaluation	Highly iterative scheme
Huang et al. [18]	Machine learning, clustering	Effectively identify critical nodes	Induces overhead in increased traffic
Liu et al. [19]	Analytical model	Variability evaluation for travel pattern	Narrowed constraint modelling
Gu et al. [20]	Bayesian, Deep Learning	Higher accuracy	Computationally complex
Chen et al. [21]	Bidirectional GRU	Minimize overfitting, benchmarked	It doesn't deal with user-based data
Wang et al. [22]	Data-driven approach	Supports Multi-fold validation	Dependent on specific data
Bakdi et al. [23]	Autonomous system identification, big data	A practical definition of manifold risk factors	Scope limited to specific forms of traffic
Qin et al. [24]	Big data, fuzzy logic	Satisfactory accuracy	Dependent on rule-sets
Besinovic et al. [25]	Review work	Emergence of AI	Study limited to rail transport
Zhu et al. [26]	LSTM, SVR, open-source software	User-friendly application	No benchmarking
Gunes et al. [27]	Traffic signal management	Offers adaptive control	Applicable only at the intersection
Zhou et al. [28], Lucic et al. [29], choosakun et al. [32]	Review on edge computing, Crowdsourcing, cooperation in ITS	Elaborated discussion	-N/A-
Mirboland & Smarsly [30]	Information model, semantic	A simplified and effective model	No benchmarking
Asaithambi et al. [31]	Big data, stream/batch processing	Improved predictability	No benchmarking
Yoo et al. [33], Alexakis et al. [34]	Open-source Software	Simplified implementation	Cannot reach scalability
Dudek & Kujawski [35]	Image, big data, thresholding	Optimized route planning	Applicable to image data only

III. Research Problem

Existing approaches towards improving analytical operations in ITS reviewed in prior sections offer some of the essential mechanisms; however, some of the shortcomings of the existing schemes are as noted below:

- Non-Inclusion of Complex Traffic Data: Most existing approaches [20]-[35] have considered highly structured and detailed datasets publicly available. However, the practical implementation of multiple devices in traffic monitoring of ITS results in complex traffic data, which are difficult to analyze and occupy a more significant segment of storage units. Existing studies do not address such issues, which are the preliminary steps of any analytical operation in ITS.
- Non-Inclusion of Big Data Characteristics: A larger-sized data with streamed data, including possible artefacts, and less value added is not emphasized for the existing big data-based approach. Some existing approaches have used open-source tools, e.g., Hadoop, to address this, but such software requires more reengineering and is also associated with various reported pitfalls [26].
- Lack of Storage Utilization Aspects: In most existing studies, the traffic data is first subjected to storage, and the claimed analytical algorithms are subjected to them in the same place to acquire knowledge [26][30]. The outcomes are stored in multiple storage containers in this process. Unfortunately, such a process over-utilizes the storage space, adversely affecting the query processing and leading to higher delay.
- Adoption of Learning Approaches: There is no denying the fact that machine learning and deep learning significantly contribute to predictive traffic data analysis [21][26]. However, most schemes adopt a complex learning strategy to acquire higher accuracy. Such accomplishment of accuracy is claimed at the cost of highly iterative training operation where higher accuracy is proportional to higher availability of trained error-free data. Confirming error-free data in distributed traffic networks with heterogeneous sensing devices is a computationally challenging task that is not reported to be addressed.
- Lack of Consideration of Model Cost: A model for ITS is required to respond faster, which entirely depends upon a more brilliant construction of its logical condition and lesser usage of traffic attributes with more meaningful insights. This demands a novel and benchmarked strategy implementation towards the knowledge discovery process, considering the decentralized environment of ITS. Very few models are yet reported to accomplish this objective.

All the research problems mentioned above are identified to have concrete solutions; hence, the proposed scheme deploys a novel strategy to address these open-end research problems in ITS. The following section summarizes the methodology adopted to evolve a novel ITS management solution.

IV. Research Methodology

The prime goal of the proposed research work is to introduce a unique knowledge discovery method using a novel big data approach in ITS. The core ideology constructed in the proposed scheme is to acquire a raw form of complex traffic data from multiple sources, followed by performing a decentralized scheme of analyzing the traffic data by uniquely transforming them. The implementation of the proposed scheme is carried out in analytical research methodology, and its architecture is presented in Fig. 1 as follows:

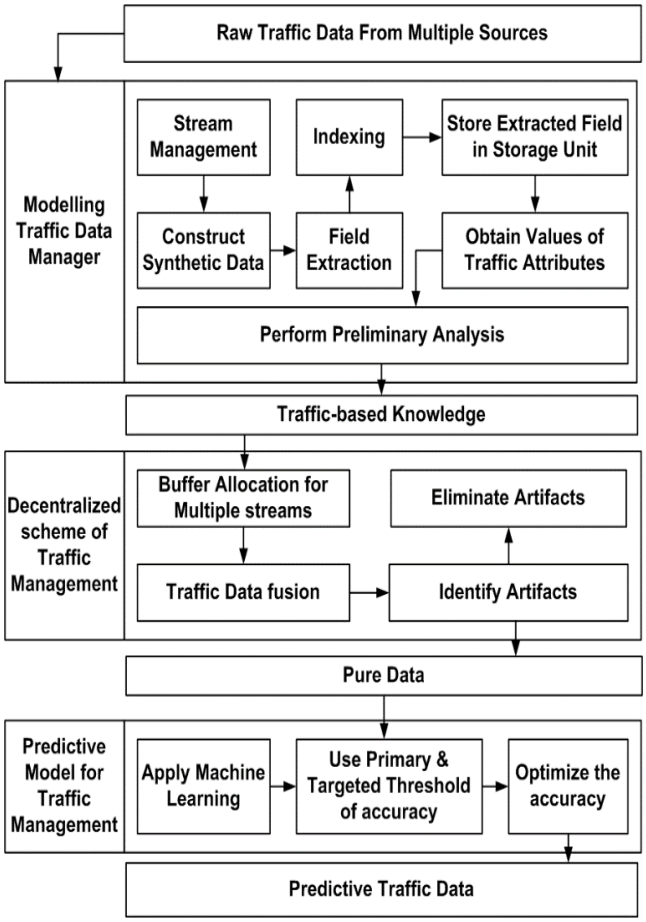

Fig. 1. Proposed architecture.

The reasons for adopting the above-mentioned architecture to address the identified problems are manifold.

- A closer look into the research problems showcases that analytical studies towards ITS don't offer emphasis to address the issues of essential big data characteristics and storage management. Both are considered as two different sets of issues. However, the proposed study introduces a mechanism where both issues are handled simultaneously, exhibiting the cost-effective operational features towards ITS.
- It was also noted that existing analytical approaches demands complex usage of mining or knowledge discovery process overlooking the implementation cost involved in it. However, the proposed architecture offers a single-window operation where indexing, transformation, and progressive knowledge discovery processes are carried out on the same platform. This reduces the dependency on including heavy-weighted analytical algorithms for processing and analyzing large-scale ITS data.

- From the perspective of model cost, it is observed that existing schemes take the input of the complete stream, process the queue, store the raw data in a cloud storage unit and then perform analytical operations. This consumes a lot of processing time and demands the involvement of various resources. At the same time, the proposed architecture can process the stream of ITS data while it stores a part of the data in the cloud and part of the data in a temporary buffer, which reduces the load of processing and storing significantly. Owing to this phenomenon, the filled-up queue can be emptied soon, and the network can process more incoming data. Hence, the proposed architecture offers a practical solution for processing and analyzing extensive ITS data even in peak traffic conditions.

The prime limitation of existing distributed data storage and analytical architecture is that it cannot jointly address the identified research problem. Apart from this, no existing architecture can optimize the spontaneous storage saturation issues in storage units or offer a faster, more progressive, and less iterative knowledge discovery process for evolving ITS data. The architecture exhibited in Fig. 1 offers three explicit blocks of operation toward knowledge discovery of traffic data in ITS. Considering the input of raw traffic data, the first block of operation performs stream management where a series of transformations is carried out over programmatically formulated synthetic data to acquire preliminary traffic-based knowledge finally. The outcome of the first operational block is further subjected to traffic data fusion over a buffer allocated for multiple streams. This operation assists in identifying possible forms of artefacts present in traffic data, eliminating them, and generating pure data. A statistical correlation-based mechanism is adapted to generate the possible value with the highest correlation, followed by substituting the artefact with statistically computed data. The third operational block is finally responsible for retaining the computed data's highest reliability and trustworthiness. A machine learning scheme is applied for this purpose which uses primary and targeted thresholds to assess the genuineness of the accuracy obtained in the outcome of prior block operation. The outcome of this model offers predictive traffic data, a knowledge discovery considered value-added analytical information. The block operation is elaborated and illustrated in the next section.

V. System Design

The proposed scheme's prime idea is to apply a novel analytical operation to offer a value-added ITS performance. An explicit system design using a big data approach has been developed, considering a typical use case for managing and analyzing ITS traffic information streams. The core motive is to develop a simplified and lightweight analytical framework for efficient knowledge discovery of complex forms of traffic data. The description of the proposed system design implemented for the proposed scheme follows.

A. Modelling Traffic Data Manager

This is the first module of implementation of the proposed system, which targets minimizing the complexities associated with streaming traffic-related information from various sensing devices installed on roads and vehicles. To carry out an investigation, there are two mechanisms to acquire the traffic-related information, viz., depending on publicly available datasets and acquiring realistic traffic data. There are currently various publicly available datasets for ITS, e.g. [36][37]; however, this dataset does not possess any characteristic of big data complexities of voluminous and various factors and is not featured with uncertainties. Hence, adopting already engineered and processed traffic-related information will not allow the model to be assessed for its capability to process a complex data stream. On the other hand, the acquisition of realistic data demands the involvement of an extensive experimental infrastructure with the inclusion of sensors. Even if small experimental prototyping is done, the outcome cannot be evaluated to prove its applicability for large-scale scenarios.

TABLE II. Sample Traffic Data

Item	Traffic Attribute	Values
1	Vehicle ID	001
2	Vehicle Type	Car
3	Sensor Type	T1
4	Date	20/02/2023
5	Location ID	NH48
6	Sensor Status	Active
7	Sensor Range	10m
8	Driver Name	John Thomas
9	Driving Experience	The lane is heavily congested

Hence, the proposed scheme develops synthetic traffic data characterized by the complexity attributes of big data in ITS. The complexity is added by streaming the data in massive volume and programmatically making the data highly unstructured, which is challenging for any processor to read and analyze. This scenario can be mapped with complex data acquisition events in a heavy incoming stream of sensed traffic data in ITS. Table II highlights the sample traffic data that are synthetically generated. To ensure better compliance with the significant data form of the repository, the proposed study has visited the existing format for publicly available big data [38] and ensured that synthetic data carry the exact form. The synthetic data is in plain text to ensure that such data could have possible artefacts apart from the autonomous data entry-based errors from the sensing devices. This is done intentionally to ensure the possibility of data with an artefact. Further, the data is programmatically converted to its most complex form as follows:

1VehicleID0012VehicleTypeCar3SensorType

T14Date20/02/20235LocationIDNH486Sensor StatusActive7SensorRange10m8DriverNameJohn

Thomas9DrivingExperienceThelaneisheavilycongested

A closer look into the above data exhibits that such information is highly challenging for a machine to read. Hence, a module of traffic data manager has been evolved to manage such highly unstructured data. The algorithm follows the traffic data manager's overall mechanism.

Algorithm for Traffic Data Manager

Input: s_d
Output: *k*
Start
1. **For** i=1:s_d
2. $U_d=f_1(i)$
3. $f_{ex} \leftarrow f_2(U_d)$
4. d=[f_{ex}, d_f, *val*]
5. $F_{ex} \rightarrow s_u(ind(f_{ex}))$
6. $sem_d=f_2(d(val))$
7. Apply $f_3(sem_d) \rightarrow d_{int}$
8. $k \leftarrow f_4(d_{int})$
9. $k \rightarrow s_u(ind(F_{ex}, k))$
10. **End**
End

The discussion of the algorithmic steps is as follows: The algorithm takes the input of s_d (stream of data) that generates a resultant of *k* (extracted knowledge) upon processing. The proposed scheme considers a stream of data s_d as an input; however, processing an infinite data stream is infeasible. Hence, sd is considered a sampled data stream based on the number of packets queued. Considering the specific queue capacity, the s_d is sampled and considered an input for the proposed scheme.

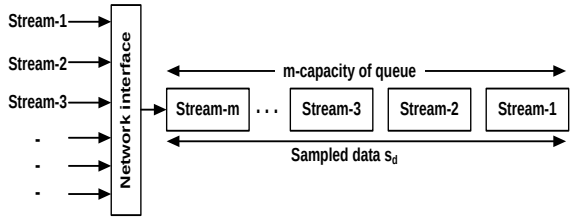

Fig. 2. Preparing for streamed input of traffic data.

Fig. 2 showcases that the proposed scheme considers its input s_d with a size equivalent to the size of the queue capacity *m* maintained by the service provider or application buffer. Hence, it is technically possible to process the streamed data s_d from different traffic zones (Line-1). The streamed data s_d is then programmatically transformed to unstructured data U_d using an explicit function $f_1(x)$. The prime task of this function is to obtain the stream data and eliminate all white spaces to form U_d, which is challenging for the machine to read (Line-2). The following line of execution of the algorithm is to extract significant fields f_{ex} using an explicit function $f_2(x)$ (Line-3). The prime tasks of this function are: i) it reads the complete string of the stream data sd and captures the significant fields f_{ex} (basically the traffic attributes in Table II). Finding and confirming the fields are simple as they will keep repeating in one individual data in s_d (however, their corresponding values *val* will differ), and ii) The function also identifies a differentiator d_f that exists between field (F_{ex}) and value (*val*). Further, a record d is constructed, which retains information about extracted field f_{ex}, differentiator d_f, and corresponding values *val* (Line-4).

The part of the implementation of the proposed algorithm associates it with the preliminary storage optimization. The algorithm stores extracted field f_{ex}, indexes them and stores them in storage unit s_u (Line-5). Hence, the variable f_{ex} and F_{ex} mean original and indexed features, respectively. However, the algorithm doesn't store the corresponding values *val*; instead, it keeps them in a temporary memory where it is further subjected to a transformation process (Line-6). An explicit function $f_2(x)$ is constructed, which uses document tagging for all the corresponding values (as well as indexed extracted features) to generate a semi-structured traffic data sem_d (Line-6). This operation leads to all the intermediate data d_{int}. The term intermediate data d_{int} will mean that each extracted value *val* is mapped with defined contextual factors constructed in the corpus while developing the synthetic data. The core reason behind this is that the last traffic attribute in Table II bears a longer string message which is computationally challenging for the machine to process and analyze. Hence, a small repository of contextual traffic factors is designed and subjected to string comparison with all the individual words in traffic attributes to extract the actual contextual meaning the machine can understand. This operation is carried out using function $f_4(x)$, which finally generates knowledge *k* (Line-8). The extracted knowledge *k* is indexed using the self-constructed *ind* method and stored alongside the priorly extracted field F_{ex} (Line-9).

The contribution of this algorithm is as follows:

- This algorithm can process voluminous amounts of traffic-related streamed data, unlike existing approaches applied to static data.
- The algorithm permanently stores a part of iterative message fields in storage and a part of the message for further processing to offer better storage utilization. Existing approaches consider storing entire data and processing where the processed data is kept in different locations while original data is considered meta-data. This increases query processing time in existing approaches.
- A unique indexing mechanism that reduces the processing time for streamed dataset is introduced.

B. Decentralized Scheme of Traffic Management

This second implementation module incorporates a decentralization scheme in proposed traffic management. According to the highlights of Fig. 3, the proposed study considers that multiple traffic environments collect traffic information to generate an individual data stream. Upon passing through the network interface using the prior algorithm, the data transform to a streamed data sd, which ultimately generates a knowledge *k*. However, all this knowledge-based information is further indexed and reposited to the storage unit su in the prior algorithm, which is again subjected to analytical data operation. To ensure the practicality of this implementation, it is necessary to introduce a data fusion technique in a decentralized manner over distributed cloud environment. This will further lead to the possibility of artefacts. Hence, this decentralization scheme further assists in identifying the location of artefacts, followed by adopting statistical operations to eliminate the artefacts. This process leads to the generation of pure traffic data.

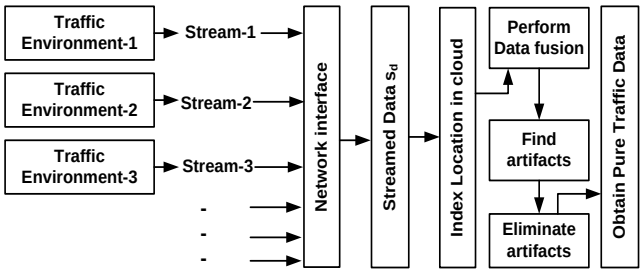

Fig. 3. Decentralized data fusion and quality improvement process.

For this purpose, an algorithm is designed to perform this decentralization operation toward further traffic data management. It is to be noted that this algorithm is explicitly executed in a storage unit on top of the final data (i.e., *k*) collected by the prior algorithm. To optimize computational speed and utilization of storage units, a buffer space matrix is constructed by extracting the available empty spaces of storage units in sharing form from multiple storage units. It will eventually mean that algorithm must be executed in a decentralized manner. Following are the steps of the proposed algorithm:

Algorithm for Decentralized Traffic Management

Input: *k*, s_u
Output: d_p
Start
1. **For** *j*=1: *k*
2. alloc *j*→buf(s_u)
3. d_{art}←$f_5(s_u, j)$
4. c_{sol}=f_6(cell(d_{rt}), $s_u(k)$)
5. d_p←$arg_{max}(c_{sol})$
6. **End**

End

The discussion of the algorithmic steps is as follows: The algorithm takes the input of k (knowledge) and s_u (index values in a storage unit) that, after processing, yields an outcome of d_p (pure data). A counter variable *j* represents the number of incoming data streams mainly focusing on newly analyzed data of *k* from the prior algorithm (Line-1). Each *j* stream of k data is then allocated to a matrix formulated by a shared buffer from all the available decentralized storage units s_u using method *buf* (Line-2). The prime reason to perform this operation is to ensure that no operation towards fused stream data is carried out directly over storage units, unlike the majority of the existing schemes. Fig. 4 offers a pictorial representation of the allocation of streams over unused memory over existing storage units. The advantage of this mechanism is that it utilizes unused memory from different cloud storage units to process the incoming streams of analyzed data. The following line of operation is associated with identifying errors or artefacts in the data (d_{art}). An explicit function $f_5(x)$ is constructed to identify an artefact in the *j* stream concerning distributed index storage units s_u (Line-3). The outcome will lead to the generation of the location address of the cell in shared memory, which is witnessed with artefacts (d_{art}). The term *artefact* will represent illegitimate or illogical data that offers no significance towards either simplified understanding or could ever be processed.

Such data could be caused due to network or sensor-based errors while propagating the data. Upon identifying the artefacts (dart), the next action step will be eliminating them. The proposed scheme implements a unique artefact elimination process: i) an explicit function $f_6(x)$ is constructed to obtain candidate solution c_{sol}. The variable c_{sol} will represent an alternate solution from different traffic data considering the specific traffic attribute where the artefact has been witnessed, ii) the function $f_6(x)$ performs a correlation assessment between the contextual value of other fields with the field where the artefact is present (Line-4). At the same time, this operation will lead to the generation of multiple correlation values; iii) the correlation value found to be maximum is substituted in the place of artefact dart, leading to the generation of pure data d_p (Line-5).

To offer a clear idea about the proposed identification and elimination of artefacts, a pictorial representation of Fig. 5 further illustrates this process. Fig. 5 showcases that there are series of *m* number of streamed data concerning analyzed data, i.e., $sd_1(k_1)$, $sd_2(k_2)$, …, $sd_m(k_m)$. Although the proposed scheme uses nine explicit traffic attributes (as shown in Table II), this explanation considers five explicit traffic attributes (T1, T2, T3, T4, and T5) to fit the information in pictorial representation. Also, assume that user (node) information present in each stream of data is U_1, U_2, ….., etc. The corresponding values of each user concerning traffic attributes are represented as $v_{1,1}$, $v_{1,2}$, …. Consider that the proposed algorithm, till Line-3 using function $f_5(x)$, found the presence of artefact as d_{art}=$v_{2,3}$ that belongs to second user U_2 in streamed data. In such case, the proposed algorithm obtains candidate correlation for all the user rows to obtain c_{sol1}, c_{sol2}, ….c_{sol99}, c_{sol100}. The algorithm further applies function $f_6(x)$, which upon execution, found multiple possibilities based on correlation analysis. The outcome shows two higher values say $v_{7,3}$ and $v_{97,3}$ belonging to 7th and 97th user respectively. It will mean that the possibility of a better proximal solution in the cell for artefact $v_{2,3}$ resides for either $v_{7,3}$ or $v_{97,3}$. The algorithm compares $v_{7,3}$ and $v_{97,3}$ to find that the higher value is $v_{97,3,}$ which is then substituted in the specific matrix cell for $s_{d1}(k_1)$. That means the older value of artefact $v_{2,3}$ is now eliminated and substituted with the new value of $v_{97,3}$.

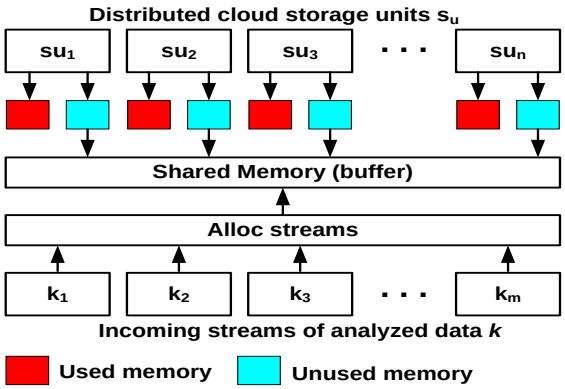

Fig. 4. Mechanism of stream allocation.

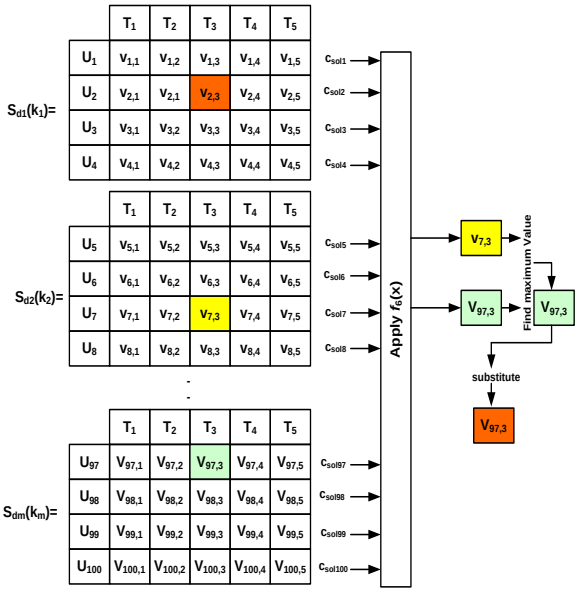

Fig. 5. Mechanism of finding and eliminating artifacts.

The novelty of this algorithm is as follows:

- The algorithm can identify the artefacts in more extensive indexed data in simplified steps.
- A non-iterative correlation-based assessment is carried out to obtain the proximal candidate solution, which replaces the artefact value in a specific cell only within the matrix.
- The algorithm is designed to work on artefact identification on massive traffic data.

C. Predictive Model for Traffic Management

The prior module of implementation assists in offering the highest possible data purity using correlation-based analysis statistically. P*ure data* (d_p) refers to a matrix formation of complete information. Each matrix cell is filled with the legitimate or authorized format of individual data complying with the respective traffic attribute. However, on the verge of seamless incoming of stream traffic data, there is still a possibility of accuracy in substituted value in the prior module. Hence, a predictive scheme is introduced in a proposed scheme to assess the degree of accuracy in the obtained data. There are two sets of operations performed in the proposed predictive model viz. i) to identify the degree of accuracy of the newly substituted value based on the learning process, and ii) in case of identification of faulty accuracy score, the predictive model assists in obtaining the actual value with higher accuracy. The operation of the predictive algorithm is as follows:

Algorithm for Predictive Traffic Management

Input: d_p, Th
Output: i_{sol1}
Start
1. **For** i=1:d_p
2. $i_{sol}=f_7(i)$
3. **If** $i_{sol}\geq$Th
4. flag i_{sol} as true
5. **Else**
6. $i_{sol1}=f_7(i_{sol}, Th_{tar})$
7. **For** $i_{sol1}\geq Th_{tar}$
8. flag i_{sol1} as true
9. $d_p(d_{art})\leftarrow i_{sol1}$
10. **End**
11. **End**

End

The discussion of those mentioned above predictive algorithmic steps are as follows: The algorithm takes the input of d_p (pure data) and Th (threshold) that, after processing, yields an outcome of i_{sol1} (predicted value). The algorithm considers all the outcomes of the prior module of implementation, i.e., pure data d_p (Line-1), where it subjects all the data to various machine learning approaches. A function $f_7(x)$ is constructed where multiple machine learning approaches are applied to the input data i to obtain an intermediate solution i_{sol} (Line-2). The prime reason behind applying multiple machine learning approaches is to assess the best-fit model toward optimal accuracy. The algorithm then compares the obtained value of i_{sol} with a primary accuracy threshold *Th* (Line-3). The optimal anticipated accuracy of i_{sol} should be more than Th, while the system flags the assessed value of i_{sol}>Th as true accuracy (Line-4). Otherwise (Line-5), the algorithm reform the implication of function $f_7(x)$ to the obtained value of i_{sol} as the second iteration (Line-6). A new target threshold Th_{tar} is set, which is more than the primary threshold Th. The iteration is continued until the objective function towards converging to a new threshold Th_{tar} is met. For optimal solution, the algorithm re-checks the newly obtained value, i.e., isol1 is more than Thtar while the truth condition (Line-7) confirms higher accuracy (Line-8). The obtained value of i_{sol1} is now substituted in the cell recorded with artefact data in the d_p matrix (Line-9). It should be noted that the primary threshold *Th* is fixed by a user based on application/service demand. At the same time, the user can set the value of Th_{tar} by adjusting the prior *Th* value to a slightly higher level. A closer look into the pictorial representation in Fig. 6 of this algorithm will further illustrate the proposed predictive traffic management mechanism.

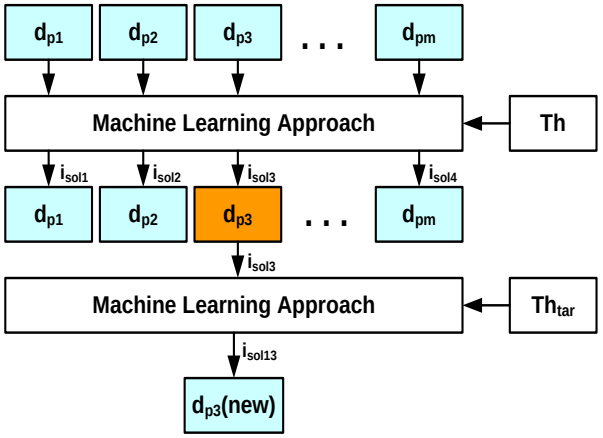

Fig. 6. Mechanism of obtaining predictive value.

According to Fig. 6, the algorithm considers m number of pure data values, i.e., d_{p1}, d_{p2}, ….$d_{pm,}$ subjected to the machine learning model. The respective obtained solution i_{sol1}, i_{sol2}, ….i_{solm} are compared concerning primary threshold Th to find out that one of the data, i.e., d_{p3} is found to under-performed (which means its accuracy is lower than cut-off), while the rest other values (i.e., d_{p1}, d_{p2}, … are found to be optimal, i.e., more than *Th*). Hence, the algorithm is explicitly iterated for d_{p3} with a fine-tuned value of target threshold Th_{tar} (Th_{tar}>Th) that finally leads to the new predictive outcome of $d_{p3.}$ This completes the overall algorithm implementation. The outcome of this model from the viewpoint of the application or service of ITS can be stated as a value-added analytical result, which could assist in more profound insights into the traffic scenario with higher reliability and accuracy.

The novelty of the proposed predictive algorithm is as follows:

- A simplified predictive model doesn't demand any form of the reengineering process towards the ITS framework.
- The accuracy outcome of the model can eventually be fine-tuned by the severity of the application or services demanded in ITS.
- The proposed machine learning algorithm can perform more progressive operations and requires less iteration to obtain optimal accuracy.

A closer look into the entire algorithm implementation shows that the proposed scheme offers a novel and sophisticated learning scheme for traffic management. The following section discusses the results accomplished in the study.

VI. Result Analysis

This section discusses the results obtained by implementing the algorithms discussed in prior sections. The discussion is carried out concerning the assessment environment, and the result is accomplished with more highlights on the result discussion and learning outcome.

A. Strategies for Result Analysis

Before initiating the proposed scheme, 25 records were chosen for a pilot study from the primary dataset as a smaller sample size. The idea was to assess the appropriateness of the proposed algorithm towards data analysis concerning accuracy. With the 25 records, 93.1-98.2% of accuracy has been obtained for almost all the learning approaches. This accomplishment offers a concrete proof-of-concept which is then subjected to the original size of the dataset for further assessment. From the data elicitation process perspective, the proposed scheme develops multiple data nodes responsible for streaming the data to the distributed cloud interface. The proposed scheme develops five distributed ITS data nodes (which can be mapped with a gateway node that keeps track of all traffic data to be disseminated). This information is forwarded as a stream to a standard cloud interface, where further data aggregation, processing, and analysis are carried out. The proposed scheme uses a supervised learning approach where the data points with significant predictive errors can be identified and solved. The implementation environment consists of multiple ITS traffic system data nodes that generate the traffic data and forward it to the core cloud interface. This standard interface is used to aggregate the data, identify with the elimination of the artefacts, perform normalization, and apply semantic and syntactical approaches to discover knowledge. Further multiple supervised learning approach is used for the predictive analysis of ITS traffic data. From the perspective of the balanced dataset, the proposed scheme splits the complete data into ten sets, where each set bears 100 records, and one set is allocated to a single stream. This facilitates effective monitoring of network performance parameters like overhead and delays while attempting to perform data transmission. Further, from the perspective of the learning approach, 70% of the data has been considered for training, while 30% of residual data is considered for testing. Further discussion of the assessment environment follows next.

B. Assessment Environment

A computational framework is constructed in MATLAB to assess the proposed scheme, considering a regular 64-bit Windows machine with i5 processing capability. The raw data from traffic is generated as a stream from multiple sources of traffic environment, which are further subjected to the first algorithm towards accomplishing preliminary knowledge. This output is further subjected to a second algorithm where the artefacts are identified, followed by a statistical correlation-based approach to substitute the specific cell of an artefact with new values to achieve data purity. Further, a set of machine learning algorithms are applied to assess the correctness of obtained outcome of the second algorithm using dual thresholding methods. For simplification in evaluation, the primary threshold (*Th*) for accuracy is maintained at 0.5, while the secondary threshold (Th_{tar}) is maintained at 0.7. The thresholding values can constantly be amended based on service or application severity towards analytical operations. Following is further information on the assessment environment:

1) Dataset: The implementation of the complete work is carried out by constructing a synthetic dataset with nine traffic attributes (e.g., Vehicle ID, Vehicle Type, Sensor Type, Date, Location ID, Sensor Status, Sensor Range, Driver Name, Driving Experience) and their corresponding values. The construction of the dataset is carried out by adhering to the standard format of big data [38]. At the same time, the inclusion of traffic attributes is formulated based on an existing publicly available dataset of ITS [36][37]. A total of 1000 datasets is acquired; each dataset further consists of 100 records of an individual traffic environment. The datasets are maintained in plain-text form, which is further given as input to MATLAB script for further algorithmic operation.

2) Performance Parameters: The performance parameters considered in the proposed assessment are accuracy, communication overhead, delay, and processing time.

3) Comparison with Existing Scheme: The proposed scheme has used multiple machine learning schemes that are

reported to be frequently adopted in existing schemes, e.g., Random Forest, Artificial Neural Network, Support Vector Machine, Logistic Regression, and Naïve Bayes. The comparative analysis is carried out with each other adopted machine learning model to investigate the best-fit model towards performing an analytical operation in ITS.

C. Result Accomplished

The numerical outcome of the proposed scheme concerning various learning-based analytical approaches is showcased in Table III with respect to various performance metrics adopted for assessment.

TABLE III. NUMERICAL OUTCOME OF STUDY

Techniques	Accuracy
Random Forest (RF)	99
Artificial Neural Network (ANN)	91
Support Vector Machine (SVM)	86
Logistic Regression (LR)	67
Naïve Bayes	70
Techniques	**Overhead (ms)**
Random Forest (RF)	1.11
Artificial Neural Network (ANN)	2.21
Support Vector Machine (SVM)	2.31
Logistic Regression (LR)	2.37
Naïve Bayes	2.22
Techniques	**Delay (s)**
Random Forest (RF)	0.05
Artificial Neural Network (ANN)	0.31
Support Vector Machine (SVM)	0.12
Logistic Regression (LR)	0.1
Naïve Bayes	0.19
Techniques	**Processing Time (s)**
Random Forest (RF)	3.67
Artificial Neural Network (ANN)	10.56
Support Vector Machine (SVM)	11.21
Logistic Regression (LR)	10.03
Naïve Bayes	9.87

The discussion of the numerical outcomes is further illustrated in graphical form from Fig. 7 to 10 for better inference to the accomplished outcome.

D. Analysis of Accuracy

The proposed scheme computes accuracy by evaluating several correct predictions made by different sets of machine learning approaches toward the analyzed data. The outcome of accuracy is shown in Fig. 7.

The inference of the result exhibited in Fig. 7 is as follows:

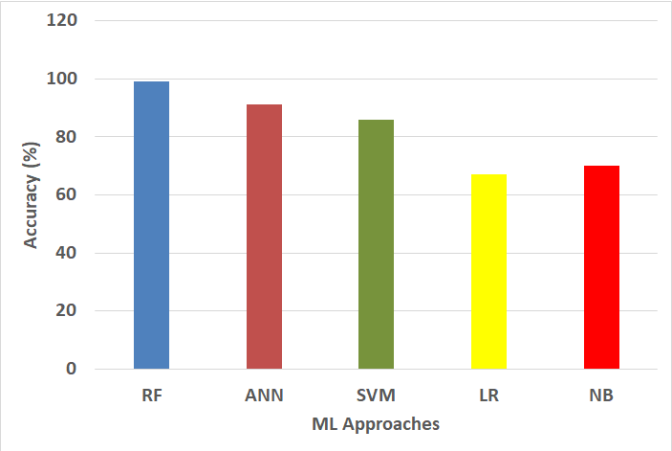

Fig. 7. Comparative analysis of accuracy.

1) Discussion of Results: The outcome showcases that the RF model performs better than others on the accuracy scale. The next exhibit of better performance is from ANN and SVM models, although the SVM model has slightly less accuracy than ANN. LR and NB's accuracy trend is nearly the same, with no significant difference. The prime justification is that the LR method exhibits less reliability when exposed to a continuous data stream and hence witnesses degraded accuracy. At the same time, the assumption of independent predictors in the NB model doesn't map well with the proposed study. Although SVM offers better ranges of classification for heterogeneous traffic data, its accuracy starts to decline when the size of the data increases. In this perspective, ANN performs better than SVM, which performs iterative operations to acquire a higher accuracy state. On the other hand, the RF model can execute both regression and classification, potentially contributing to higher accuracy.

2) Learning Outcome: Based on the result, it can be stated that the RF model is highly suited when exposed to significant streams of ITS applications/services that demand higher accuracy in the prioritized traffic environment. The applicability of ANN and SVM is suited for low-medium scale traffic, which also demands a medium-high range of accuracy. On the other hand, the applicability of the LR and NB model is best suited for performing an analytical operation that doesn't demand instantaneous response delivery or doesn't need higher accuracy.

E. Analysis of Communication Overhead

Communication overhead is computed by evaluating cumulative packets destined to be forwarded from one vehicle to another vehicular node in ITS. A sample of 2500 test packet bytes is used to assess this performance metric. An ITS with an optimal design plan should always keep the communication overhead as low as possible to resist a communication bottleneck situation. The outcome of communication overhead is shown in Fig. 8.

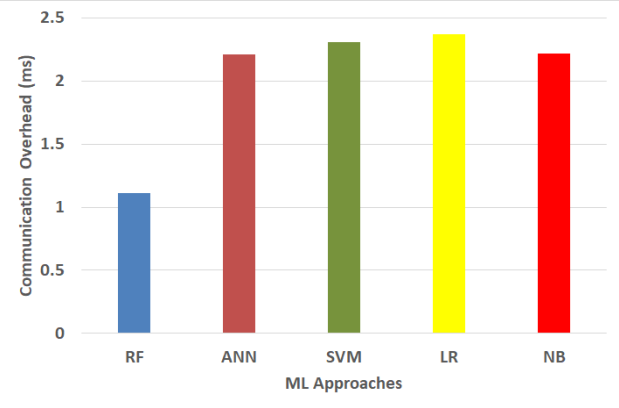

Fig. 8. Comparative analysis of communication overhead.

The inference of the result exhibited in Fig. 8 is as follows:

1) Discussion of results: The outcome in Fig. 8 shows two distinct trends viz. i) lower communication overhead shown by the RF model and ii) higher communication shown by the ANN, SVM, LR, and NB models. Although there is a numerical difference in the outcome among ANN, SVM, LR, and NB models (as shown in Table III), the difference is less significant. The prime justification behind this is as follows: A closer look into ANN, SVM, LR, and NB models shows that they perform quite an iterative analytical process while performing both training and validation operations (especially the training). This fact leads to a more significant communication overhead when exposed to many data packets. It is to be noted that data packets and their sizes are derived from the existing synthetic dataset itself. However, RF models offer a comparatively less iterative process and a more streamlined operation for continuous incoming packets, resulting in less communication overhead.

2) Learning outcome: Based on the outcome, the adoption of RF is more suitable in ITS when it calls for performing analytical operations in dense traffic conditions, whereas other machine learning models are applicable only in lesser dense traffic environments. The reliability of RF processing and analyzing data is relatively higher than others.

F. Analysis of Delay

Owing to the decentralized scheme in the proposed system, it can be assumed that all the algorithmic operations are carried out on different terminals in ITS with higher synchronous operation. Hence, delay is a suitable parameter to justify any form of lag of interval in receiving and analyzing data. The proposed system computes delay by evaluating the duration interval for one module's data packet to reach another.

As the proposed architecture jointly implements three different algorithms, it is anticipated to exhibit a delayed trend as lower as possible. The inference of the result exhibited in Fig. 9 is as follows:

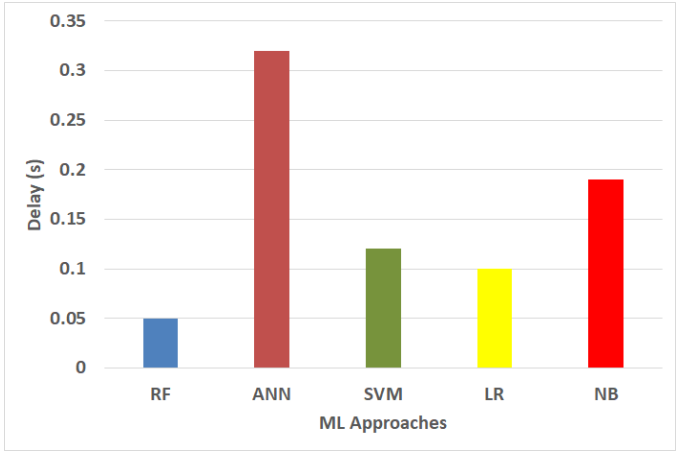

Fig. 9. Comparative analysis of delay.

1) Discussion of results: A unique observation is noted in Fig. 9 concerning delay, which is entirely different from a prior comparison of accuracy and communication overhead. Fig. 9 showcases that RF is the performing model, while the next performing model is SVM and LR. On the other hand, the performance of ANN and NB is shown to consume more delay. The specific reason behind this outcome is mainly associated with increasing training operations in ANN to meet the anticipated accuracy (Th_{tar}). Although NB performs slightly better than ANN, it cannot feature learning from the associated relationship of traffic attributes. From this context, LR and SVM perform better as both can deal with high-dimensional spaces between the data; however, SVM doesn't excel well compared to LR as it demands increasing time for training.

2) Learning outcome: From the outcome perspective, it can be stated that prioritized applications/services in ITS are well-suited when executed with the RF model. In contrast, the application/services of ITS only work well with ANN and SVM if the sample size is reduced. However, it is not practically possible in the genuine environment of ITS.

G. Analysis of Processing Time

Processing time is computed as the time required for the complete algorithm to execute jointly. An effective architecture deployment always demands lower processing time. The outcome of processing time is shown in Fig. 10.

The inference of the result exhibited in Fig. 10 is as follows:

1) Discussion of results: The outcome in Fig. 10 showcases RF to offer reduced processing time compared to other machine learning approaches. A similar justification discussed for another performance metric can be attributed to stating the reason behind this outcome. Lower processing time also refers to lower time complexity stating that the RF model offers a computationally cost-effective analytical process in ITS compared to others.

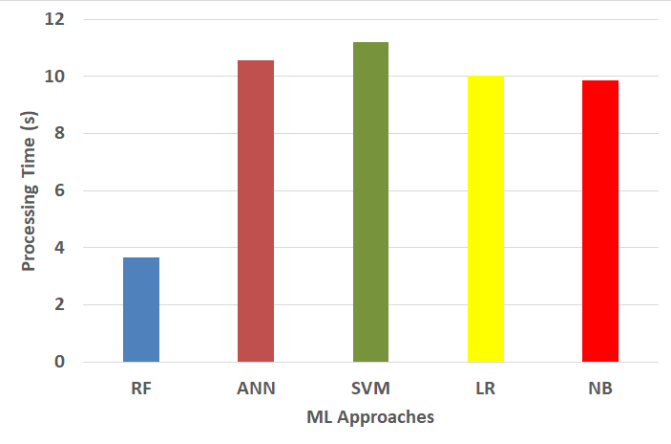

Fig. 10. Comparative analysis of processing time.

2) Learning outcome: As the ITS environment included the usage of multiple resource-constraint devices, it is anticipated that algorithmic operation should not be computationally complex. Hence, the RF model offers better predictive operation on low resource-based devices in ITS, while others are witnessed with higher computational complexity.

Therefore, from the perspective of accomplished outcome, it can be stated that the proposed analytical model is well suited with RF to exhibit a best-fit machine learning model in ITS. However, to understand the efficiency of the proposed scheme apart from standard learning approaches, the proposed scheme also analyzes comparison with existing state-of-art methods, as exhibited in Table IV.

TABLE IV. COMPARISON WITH STATE-OF-ART

Method	A_1	A_2	A_3	A_4
Ensemble Clustering, tensor [17]	No	Low	Medium	High
Machine learning, clustering [18]	No	Low	Low	Medium
Analytical model [19]	no	Low	Medium	High
Bayesian, Deep Learning [20]	No	Low	Faster	Medium
Bidirectional GRU [21]	No	Medium	Medium	High
Data-driven approach [22]	No	Low	Medium	Medium
Autonomous system identification, big data [23]	No	Medium	Low	Medium
Big data, fuzzy logic [24]	No	Low	Low	High
LSTM, SVR, and open-source software [26]	No	High	Medium	Medium
Traffic signal management [27]	No	Low	Low	Medium
Information model, semantic [30]	No	Medium	Medium	High
Big data, stream/batch processing [31]	No	Low	Higher	High
Open-source Software [33]	No	Low	Slower	Medium
Image, big data, and thresholding [35]	No	Medium	Slower	High
Proposed	Yes	High	Faster	Very low

Table IV highlights the comparison with the state-of-the-art methods presented in Section II. The prime headers used in Table IV are A_1: Decentralization A_2: Accuracy, A_3: Timeliness, A_4: Complexity. From the outcome shown in Table IV, the following inference of novelty is drawn:

- The entire modelling of the proposed scheme is carried out to support the scheme's decentralization applicability, which is not exhibited by either of the existing methods, irrespective of their accomplished outcomes. This is a significant novelty of the proposed scheme, which leads to the practical ground of implementation in urban traffic systems.
- Without using any sophisticated or iterative learning principle, the proposed scheme offers higher accuracy than existing methods without compromising the other performance attributes shown in Table IV.
- The proposed algorithm has reduced dependencies of iterative computation while it is more progressive, providing a speedier algorithm processing time. This is a significant accomplishment of novelty attributes of the proposed scheme compared to existing schemes where complex approaches are used.
- The proposed approach also offers very low computational complexity, contributing to other novel features. The prime reason behind this is the formulation of algorithms emphasising optimal data quality in predictive operation. Further, the adoption of shared memory makes it a more lightweight operation.

VII. CONCLUSION

A closer look into the proposed experimental analysis showcases that the proposed scheme can consider the input of multiple streams of ITS traffic data via various data nodes. This is a practical scenario where a gateway node can collect the traffic data and subject it to analysis. Further, the information is subjected to various rounds of processing, right from transformation to the final stage of knowledge discovery. All these individual blocks of operation can be carried out in distributed form. Yet, owing to the inclusion of a unique indexing mechanism, all the data are closely synced with each other. Further, the benchmarking over the accomplished outcome clearly states the distinguishable performance for each supervised machine learning scheme. This shows that the proposed scheme can act as a robust and highly flexible evaluation platform for analyzing the ITS traffic data with cost-effective measures without affecting data transmission performance. The proposed scheme presents a novel architecture capable of performing an efficient analytical operation over complex traffic data in ITS. The contribution of the proposed study is as follows:

1) The proposed model can transform the raw and complex stream of traffic data into highly structured data rendering it with higher suitability for analytical processing,

2) The proposed model offers the implementation of the highly decentralized scheme of the analytical process

considering streams of multiple traffic data from different origins followed by a unique data fusion,

3) The proposed model can identify the position of artefacts in massive data sizes followed by eliminating the artefacts using cost-effective statistical correlation-based analysis.

4) The proposed scheme also introduces a joint test-bed benchmarked with multiple machine learning models to show the best-fit model towards predictive analytical operations in the ITS environment.

5) Quantification of outcome states that the proposed model, when executed with the RF model, shows improvement in accuracy by 20%, reduction in communication overhead by 14%, minimization of delay by 13%, and diminished processing time by 7% in comparison to other machine learning approaches (e.g., ANN, SVM, LR, and NB model).

Future work will be carried out toward further optimizing the learning model by including network-based parameters involved in the assessment. An assessment model can be constructed to investigate the impact of different network types, storage types, and communication standards on analytical operations. Further, work can be extended to understand the adoption of different data formats toward predictive accuracy.

REFERENCES

[1] M. Cenite, "Google Books," in The SAGE Guide to Key Issues in Mass Media Ethics and Law, 2455 Teller Road, Thousand Oaks California 91320: SAGE Publications, Inc., 2015, pp. 847–858.

[2] G. Dimitrakopoulos, L. Uden, and I. Varlamis, The future of Intelligent Transport Systems, 1st ed. Elsevier, 2020

[3] X. (joyce) Liang, S. I. Guler, and V. V. Gayah, "Decentralized arterial traffic signal optimization with connected vehicle information," J. Intell. Transp. Syst., vol. 27, no. 2, pp. 145–160, 2023.

[4] Y.-H. Chen, Y. Cheng, and G.-L. Chang, "Incorporating bus delay minimization in design of signal progression for arterials accommodating heavy mixed-traffic flows," J. Intell. Transp. Syst., vol. 27, no. 2, pp. 187–216, 2023.

[5] N. Kapkaeva, A. Gurzhiy, S. Maydanova, and A. Levina, "Digital platform for maritime port ecosystem: Port of hamburg case," Transp. Res. Procedia, vol. 54, pp. 909–917, 2021.

[6] K. Kováčiková, A. Novák, M. Kováčiková, and A. N. Sedláčková, "Smart parking as a part of Smart airport concept," Transp. Res. Procedia, vol. 65, pp. 70–77, 2022.

[7] A. Balboa, O. Abreu, J. González-Villa, and D. Alvear, "Intelligent emergency management system for railway transport," Transp. Res. Procedia, vol. 58, pp. 193–200, 2021.

[8] B. Rojas, C. Bolaños, R. Salazar-Cabrera, G. Ramírez-González, Á. Pachón de la Cruz, and J. M. Madrid Molina, "Fleet Management and control system for medium-sized cities based in Intelligent Transportation Systems: From review to proposal in a city," Electronics (Basel), vol. 9, no. 9, p. 1383, 2020.

[9] A. Sumalee and H. W. Ho, "Smarter and more connected: Future intelligent transportation system," IATSS Res., vol. 42, no. 2, pp. 67–71, 2018.

[10] C. Creß, Z. Bing, and A. C. Knoll, "Intelligent transportation systems using external infrastructure: A literature survey," arXiv [cs.RO], 2021.

[11] D. Mans et al., "Recommendations for actions concerning supporting ITS developments for VRUs," Eur. Transp. Res. Rev., vol. 9, no. 2, 2017.

[12] J. Wang, X. Yu, Q. Liu, and Z. Yang, "Research on key technologies of intelligent transportation based on image recognition and anti-fatigue driving," EURASIP J. Image Video Process., vol. 2019, no. 1, 2019.

[13] X. Shi, "More than smart pavements: connected infrastructure paves the way for enhanced winter safety and mobility on highways," J. Infrastruct. Preserv. Resil., vol. 1, no. 1, 2020.

[14] I. Damaj, S. K. Al-Khatib, T. Naous, W. Lawand, Z. Z. Abdelrazzak, and H. T. Mouftah, "Intelligent transportation systems: A survey on modern hardware devices for the era of machine learning," J. King Saud Univ. - Comput. Inf. Sci., vol. 34, no. 8, pp. 5921–5942, 2022.

[15] T. Yuan, W. Rocha Neto, C. E. Rothenberg, K. Obraczka, C. Barakat, and T. Turletti, "Machine learning for next-generation intelligent transportation systems: A survey," Trans. Emerg. Telecommun. Technol., vol. 33, no. 4, 2022.

[16] J. R. Montoya-Torres, S. Moreno, W. J. Guerrero, and G. Mejía, "Big data analytics and intelligent transportation systems," IFAC-PapersOnLine, vol. 54, no. 2, pp. 216–220, 2021.

[17] G. Qi, A. Ceder, A. Huang and W. Guan, "A Methodology to Attain Public Transit Origin–Destination Mobility Patterns Using Multi-Layered Mesoscopic Analysis," in *IEEE Transactions on Intelligent Transportation Systems*, vol. 22, no. 10, pp. 6256-6274, Oct. 2021, doi: 10.1109/TITS.2020.2990719.

[18] X. Huang, J. Chen, M. Cai, W. Wang, and X. Hu, "Traffic Node Importance Evaluation Based on Clustering in Represented Transportation Networks," in *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 9, pp. 16622-16631, Sept. 2022, doi: 10.1109/TITS.2022.3163756.

[19] S. Liu, T. Yamamoto, E. Yao, and T. Nakamura, "Exploring Travel Pattern Variability of Public Transport Users Through Smart Card Data: Role of Gender and Age," in *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 5, pp. 4247-4256, May 2022, doi: 10.1109/TITS.2020.3043021.

[20] Y. Gu, W. Lu, X. Xu, L. Qin, Z. Shao, and H. Zhang, "An Improved Bayesian Combination Model for Short-Term Traffic Prediction With Deep Learning," in *IEEE Transactions on Intelligent Transportation Systems*, vol. 21, no. 3, pp. 1332-1342, March 2020, doi: 10.1109/TITS.2019.2939290.

[21] D. Chen, X. Yan, X. Liu, S. Li, L. Wang, and X. Tian, "A Multiscale-Grid-Based Stacked Bidirectional GRU Neural Network Model for Predicting Traffic Speeds of Urban Expressways," in *IEEE Access*, vol. 9, pp. 1321-1337, 2021, doi: 10.1109/ACCESS.2020.3034551

[22] P. Wang, J. Lai, Z. Huang, Q. Tan, and T. Lin, "Estimating Traffic Flow in Large Road Networks Based on Multi-Source Traffic Data," in *IEEE Transactions on Intelligent Transportation Systems*, vol. 22, no. 9, pp. 5672-5683, Sept. 2021, doi: 10.1109/TITS.2020.2988801

[23] A. Bakdi, I. K. Glad and E. Vanem, "Test-bed Scenario Design Exploiting Traffic Big Data for Autonomous Ship Trials Under Multiple Conflicts With Collision/Grounding Risks and Spatio-Temporal Dependencies," in *IEEE Transactions on Intelligent Transportation Systems*, vol. 22, no. 12, pp. 7914-7930, Dec. 2021, doi: 10.1109/TITS.2021.3095547.

[24] G. Qin, S. Yang, and S. Li, "A Vehicle Path Tracking System With Cooperative Recognition of License Plates and Traffic Network Big Data," in *IEEE Transactions on Network Science and Engineering*, vol. 9, no. 3, pp. 1033-1043, 1 May-June 2022, doi: 10.1109/TNSE.2020.3048167.

[25] N. Bešinović *et al*., "Artificial Intelligence in Railway Transport: Taxonomy, Regulations, and Applications," in *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 9, pp. 14011-14024, Sept. 2022, doi 10.1109/TITS.2021.3131637.

[26] Y. Zhu, C. Huang, Y. Wang, and J. Wang, "Application of bionic algorithm based on CS-SVR and BA-SVR in short-term traffic state prediction modeling of urban road," Int. J. Automot. Technol., vol. 23, no. 4, pp. 1141–1151, 2022.

[27] F. Gunes, S. Bayrakli, and A. H. Zaim, "Smart cities and data analytics for intelligent transportation systems: An analytical model for scheduling phases and traffic lights at signalized intersections," Appl. Sci. (Basel), vol. 11, no. 15, p. 6816, 2021.

[28] X. Zhou, R. Ke, H. Yang, and C. Liu, "When intelligent transportation systems sensing meets edge computing: Vision and challenges," Appl. Sci. (Basel), vol. 11, no. 20, p. 9680, 2021

[29] M. C. Lucic, X. Wan, H. Ghazzai, and Y. Massoud, "Leveraging Intelligent Transportation Systems and smart vehicles using Crowdsourcing: An Overview," Smart Cities, vol. 3, no. 2, pp. 341–361, 2020.

[30] M. Mirboland and K. Smarsly, "BIM-based description of intelligent transportation systems for roads," Infrastructures, vol. 6, no. 4, p. 51, 2021.

[31] S. P. R. Asaithambi, R. Venkatraman, and S. Venkatraman, "MOBDA: Microservice-Oriented Big Data Architecture for smart city transport systems," Big Data Cogn. Comput., vol. 4, no. 3, p. 17, 2020.

[32] A. Choosakun, Y. Chaiittipornwong, and C. Yeom, "Development of the cooperative intelligent transport system in Thailand: A prospective approach," Infrastructures, vol. 6, no. 3, p. 36, 2021.

[33] A. Yoo, S. Shin, J. Lee, and C. Moon, "Implementation of a sensor big data processing system for autonomous vehicles in the C-ITS environment," Appl. Sci. (Basel), vol. 10, no. 21, p. 7858, 2020.

[34] T. Alexakis, N. Peppes, K. Demestichas, and E. Adamopoulou, "A distributed big data analytics architecture for vehicle sensor data," Sensors (Basel), vol. 23, no. 1, p. 357, 2022

[35] T. Dudek and A. Kujawski, "The concept of big data management with various transportation systems sources as a key role in smart cities development," Energies, vol. 15, no. 24, p. 9506, 2022.

[36] "Department of transportation - open data portal," *Dot.gov*. https://www.its.dot.gov/data/ (accessed Jul. 18, 2023).

[37] "Data.World," *data.world*. https://data.world/datasets/transportation (accessed Jul. 18, 2023).

[38] "Publicly available big data sets :: Hadoop illuminated," *Hadoopilluminated.com*. https://hadoopilluminated.com/hadoop_illuminated/Public_Bigdata_Sets.html (accessed Jul. 18, 2023).