Computers in Digital Forensics Using Machine Learning and Big Data

Balingan Sangameshwar[*1], Gamidelli Yedukondalu[2] & Pramod Kumar Amaravarapu[3]

[*1]Assistant Professor, Dept. of CSE-(CYS,DS) and AI&DS, VNR Vignana Jyothi Institute of Engineering and Technology, Hyderabad

[2]Assistant Professor, Dept. of CSE-(CYS,DS) and AI&DS, VNR Vignana Jyothi Institute of Engineering and Technology,
Hyderabad,

[3]Assistant Professor, Dept. of CSE-(CYS,DS) and AI&DS, VNR Vignana Jyothi Institute of Engineering and Technology, Hyderabad

Abstract:

The current growth of large data led to the development of AI and machine learning. The idea of strengthening the accuracy and usefulness of AI applications is also gaining popularity as big data and machine learning take off. In the realm of traffic applications, machine learning techniques improve guard safety in risky traffic situations. For vulnerable road users (VRUs), data privacy is the biggest problem with the current architectural designs. The primary cause of pedestrian traffic control failure is improper user privacy handling. The user data are vulnerable to several privacy and security flaws and are therefore at danger. If an intruder is able to break into the system, exposed data may be maliciously manipulated, manufactured, and misrepresented for illicit purposes. In this paper, a machine learning-based architecture is suggested for effectively analyzing and processing massive data in a secure setting. The suggested model takes user privacy into account when processing massive data. To achieve secure big data analytics, the suggested architecture is a layered framework with a parallel and distributed module that uses big data. The suggested architecture uses a machine learning classifier to create a unique unit for privacy management. The architecture also has a stream processing unit to process the data. Real-time datasets from a variety of sources are used to understand the proposed system, and experimental testing using credible datasets demonstrates the usefulness of the suggested architecture. Along with the training and validation outcomes, the results of the data import are also presented.

Keywords: Computers, Digital, Forensics, Machine Learning, Big Data, Vulnerable, Road , Architecture.

DOI: 10.24297/j.cims.2023.15

1. Introduction

[12][14] Data are accumulating quickly in the modern technological world, and humans largely rely on data. In addition to the rate at which the data increase, it is becoming impossible to store the data on any particular server. The amount of data on the earth today is immense and is growing rapidly at a very fast rate while also being unsecure [1,4,5,9]. With the development of the web, the entire world has also become online, and every action we take creates a digital map that is vulnerable [2,7,8,11]. The idea of increasing accuracy and expanding the usefulness of AI projects is becoming more important and widely acknowledged with the rise of big data and machine learning [3, 4,6,7,8]. The development of technology, social media, and the Internet of Things (IoT) are a few of the elements influencing data evolution. One of the newest ideas in the modern era, IoT is mostly used in applications for monitoring and regulating traffic. Secure IoT will transform the world's objects of today into intelligent and smart objects in the future [4,7,12,14,16].

[6,9, 14,17,19] IoT components including sensors and actuators, process input connectivity, and humans are all part of smart systems. All emerging systems are supported by sensors and actuators [7, 9, 12]. A new kind of smart application and service is produced by the interactions between all these elements. The concept of edge computing is also gaining popularity and is widely acknowledged with the rise of IoT devices. In the realm of traffic applications, machine learning solutions improve guard safety in risky traffic situations [5-7-8-7].

2. Literature Review

[1-2-3-4-5]Given the reliability of edge computing, a malicious attack on user services can be very expensive [21, 22]. In order to address the issues with data security in smart traffic applications, this study provides a secure architecture for data supervision [7-9-12-14]. Significant work has gone into the proposed architecture's data analytics and machine learning components for smart traffic data management. The usage of conventional MR cluster, insufficient data stacking, intangible structure, and use of just specific dataset are the major issues highlighted in the architecture [10, 23]. A plan was thoroughly discussed in relation to V2X connections [24].

[12-15-16] the conventional techniques to data analytics, which include Tiers responsible for specific procedures and activities, are also taken into account. Despite having a complete four-tier design with tiers for data collection, data analysis, and utilisation, the performance is slowed down by the use of a traditional map-reduce framework [25, 26]. Additionally, data aggregation prior to data loading is prioritised while data loading proficiency is disregarded. In this architecture, data loading before analysis is disregarded in favour of data aggregation of outcomes.

3. Proposed Architecture

[12-15-17] The suggested machine learning-based architecture links the various smart community departments, such as the department in charge of traffic monitoring and control. Big data for traffic monitoring and management makes up the data sources [4-7-9-14]. Figure 1 shows the process for the proposed parallel and distributed approach. The relevant components collect data from numerous sources for traffic control, such as sensors and cameras [6-9-12-17]. The data must be carefully examined prior to processing in order to design an efficient parallel and distributed architecture [2-6-9]. Various technologies, including environmental sensors, security monitoring sensors, traffic cameras, and transportation monitoring sensors, produce the data. The numerous departments, including the traffic-control authority, appropriately gather the data. Secure data collecting is the method in question [7-8-9]. Utilising machine learning, the data are categorised [5-8-12].

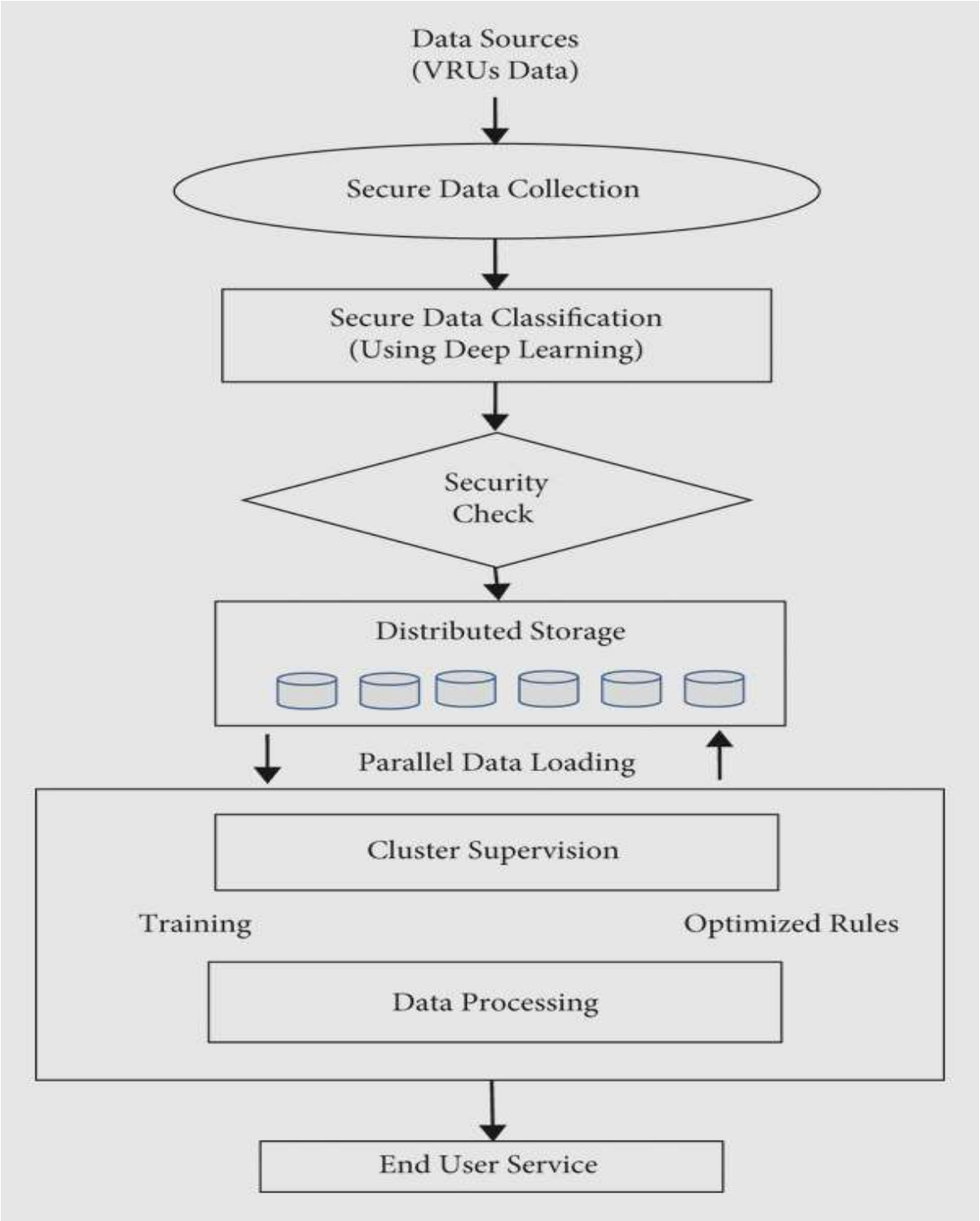

Fig.1:

The results are then sent to the relevant suppliers of smart society services for decision-making before being delivered to the users. Data stored in the distributed storage mechanism are filtered before being processed by the Hadoop processing unit [7-9-12-15-8]. The data are then used for community planning, which is the final step. Departments provide the information, and the decisions are then forwarded to the community development departments. Realising a

smart traffic system is the goal in order to process the data while maintaining privacy[9-15-18-22-23]. The aforementioned community departments serve as a liaison between the system and the user as well as data sources for the proposed system. Architecturally, the projected solution is composed of three modules: data organization, processing, and security which are shown in Figure 2.

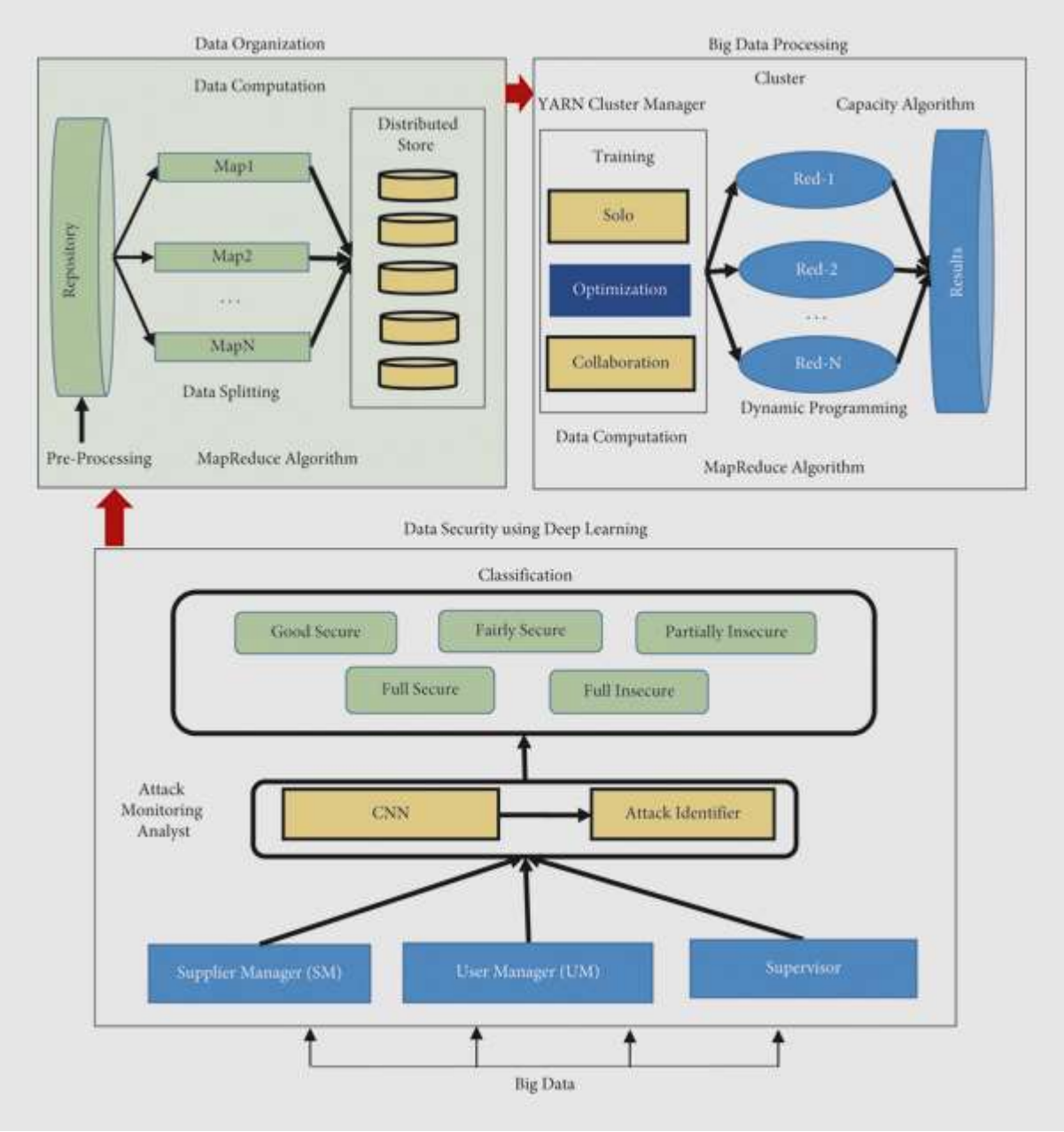

Fig.2: Proposed architecture.

4. Data Security Layer

The suggested structural layout includes a security layer to protect the data of the VRUs against intrusions. It's a component of intelligent traffic architecture. In response to the attacks, it suggests being adaptable. The security layer for the components consists of the supplier

manager (SM), user manager (UM), and supervisor. While the supervisor uses the machine learning algorithm, the SM and UM focus on the supplier and the user. It incorporates the CNN DL approach to categorise data as secure or insecure. Every supplier's profile must be kept up to date by the SM, and users' profiles must be kept up to date.

Big Data Organization

The big data organisation system includes data aggregation, gathering, and storage as well as overall data management. The data are split among different nodes so that computation can be loaded from a central server or cloud. Data from numerous local devices is obtained via the Internet to facilitate intelligent applications. The information of the environment is captured by a number of devices, such as sensors, cameras, and object-mounted devices. These data are subsequently analysed to gain knowledge and make wise decisions. It is the first layer's responsibility to compile the data from various local departments for use in managing the services for smart community development.

Big Data Processing

This unit is the primary processing component that initially preprocesses the raw data, integrating any missing values, numbers outside of the acceptable range, and irrational data combinations. If the data are not checked for these issues, decision-making may result in inaccurate findings. As a result, the transformation also involves scaling the data to a particular scale. A parallel processing unit, the core of the suggested architecture, then takes the data. The Map Reduce parallel computing model is the foundation of the suggested architecture. In order to implement big data analytics, Map Reduce is introduced.

5. Results and Discussion

Hadoop 3.0's parallel and distributed platform is used to implement the suggested approach. Apache Spark is available as a module for Hadoop. The trustworthy datasets are used. In Python 3.8, the pyspark package is used. Similar to this, a detailed configuration and a machine learning classification module are used to evaluate resilient agents. In Python 3.8, the machine learning library is also used and put into practice. With the existing suggestions, a comparative study of the suggested design is given.

The comparison and experimental results show how successful the suggested design is. This section contains a discussion of the findings. To evaluate the suggested architecture based on parallel and distributed paradigms utilizing planned algorithms, results are created using a variety of valid datasets. On top of our suggested architecture, we ran a noise and anomaly removal operation on the data. The Kilman algorithm and min-max normalizations are used to eliminate the anomalies. The map-only algorithm is used to accomplish the data ingestion.

Data Security Results

The needed training of the dataset using an ML classifier is included in the findings and experiments of the security layer. Secure and unsecure interactions with the suggested architecture are used to train the model. In a certain environment, the security layer is evaluated. The model was initially trained using matrices. Figure 3 depicts the Nave Bayes classifier's training procedure.

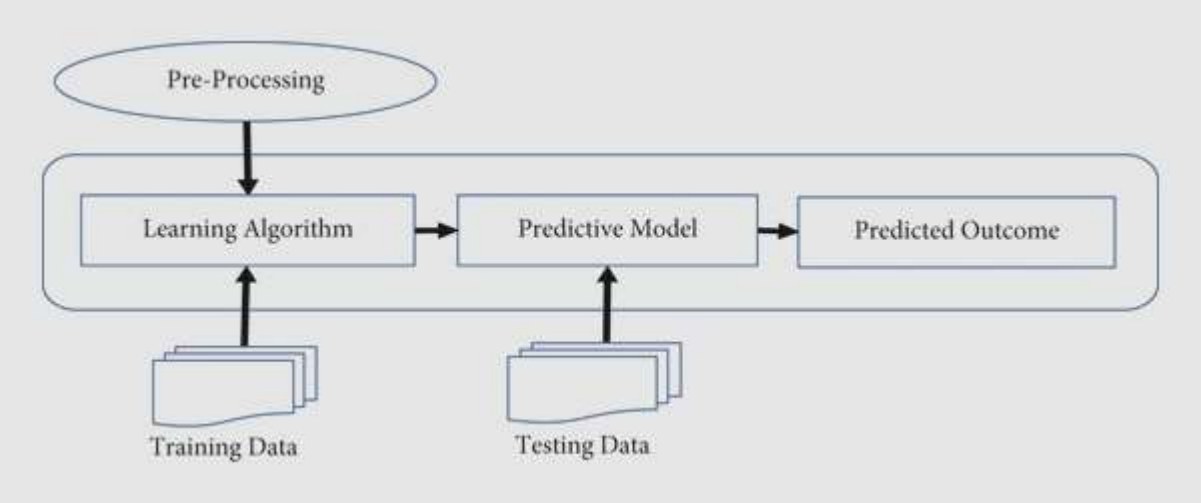

Fig.3: Model training.

Training and Validation Results

The higher accuracy of the validation and training is confirmed in Figure 4. The increased number of epochs (for example, 200 epochs) is responsible for the improved level of accuracy in training and validation. Figure 5 also shows the validation and training loss of the suggested model, which is a sign of little loss. Due to the increased number of epochs (for example, 200 epochs), the loss was reduced.

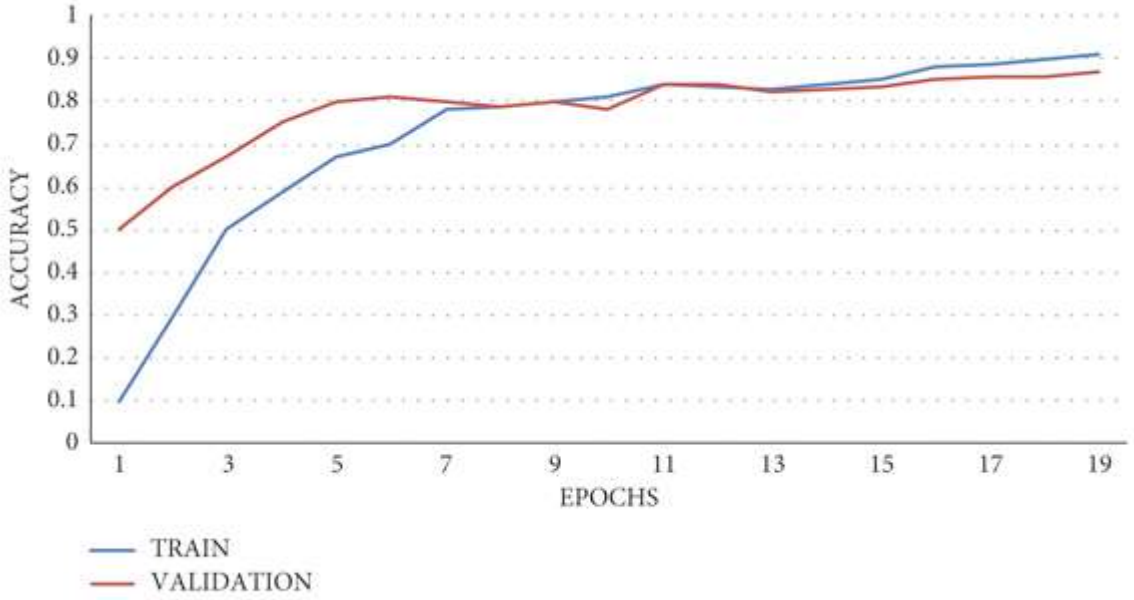

Fig.4: Accuracy.

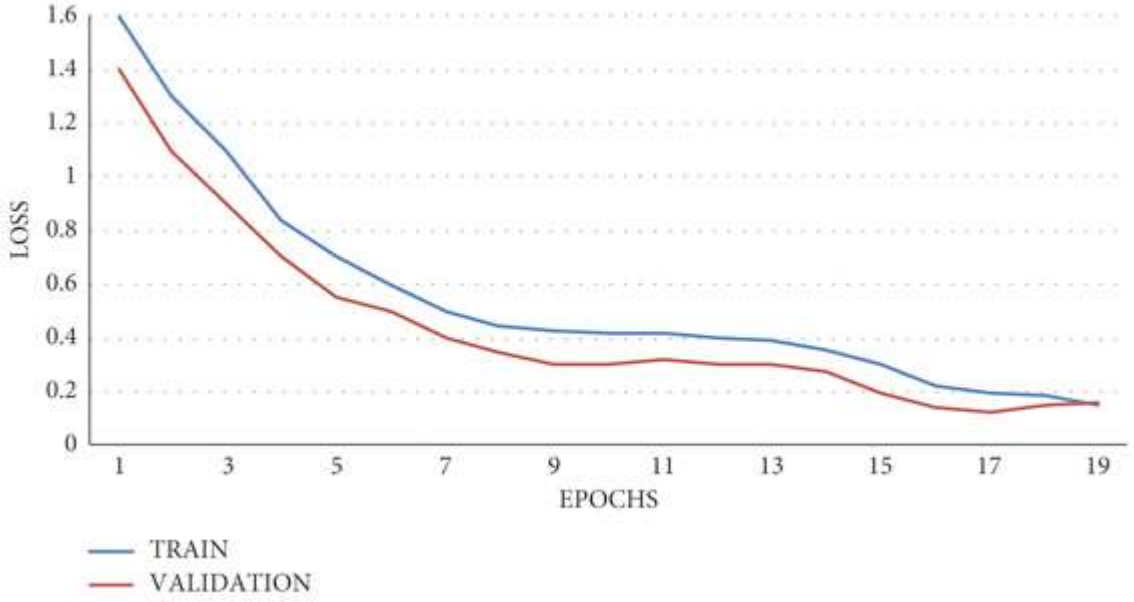

Fig.5: Loss.

Data Ingestion Results

When the dataset size is modest, it takes almost no time to load it into Hadoop. The loading of data, whether done manually or with a tool, takes about the same amount of time. Experimental evidence has shown that loading a small dataset into Hadoop takes almost no time at all. Figure 6 displays the suggested system's overall efficiency taking into account how all the settings affect how much data is loaded. The threshold for all parameter modifications of data loading

using the suggested approach is shown in Figure 7 in the same manner. The scary set value known as the threshold draws attention to the starting point where the differences between the existing and suggested methods begin. In terms of data ingestion, the proposed approach is manual, but in terms of classification and processing, it is automated.

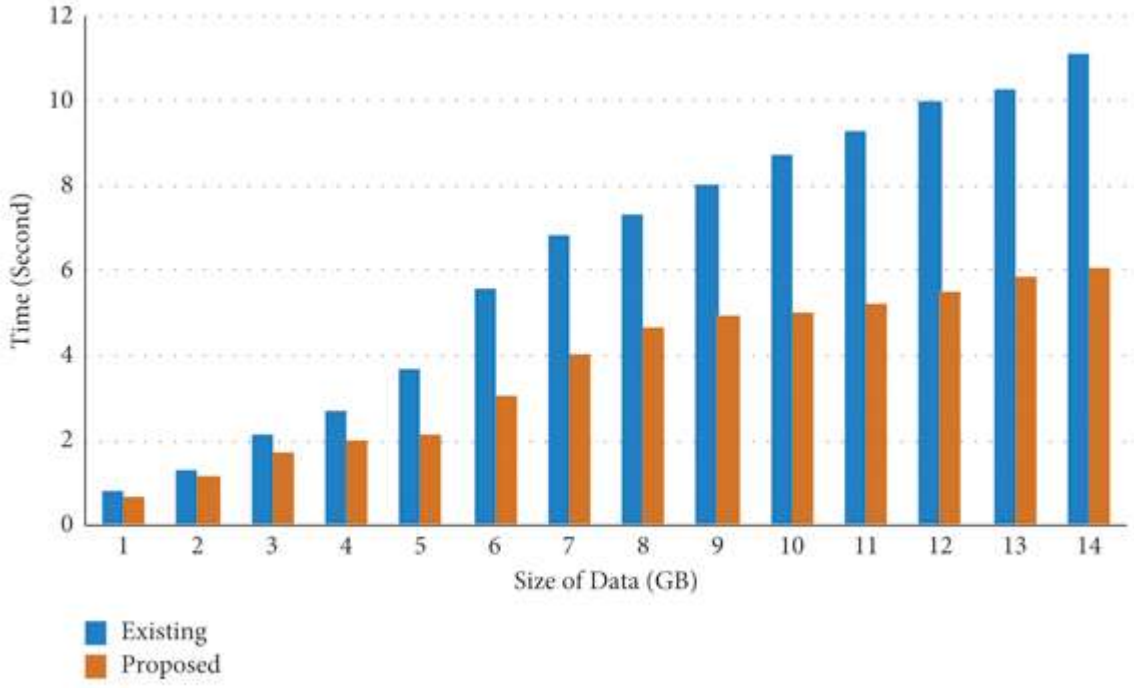

Fig.6: Overall data loading efficiency.

6. Conclusion

The widespread adoption of IoT-connected devices is taken into account as a smart traffic application, especially with the emergence of Big Data and machine learning. When compared to the effectiveness and accuracy of the underlying machine learning models, machine learning solutions deliver effective results. However, because smart traffic management and user surveillance generate a significant amount of big data that must be handled and analysed effectively, it becomes difficult to address user privacy issues. In this paper, a machine learning-based architecture is suggested for processing massive data effectively while protecting user privacy in a secure setting. To achieve secure big data analytics, the suggested architecture is a layered framework with a parallel and distributed module that uses big data.

References

1. M. I. Razzak, M. Imran, G. Xu, and G. Xu, "Big data analytics for preventive medicine," *Neural Computing & Applications*, vol. 32, no. 9, pp. 4417–4451, 2020.
2. N. Paltrinieri, L. Comfort, and G. Reniers, "Learning about risk: machine learning for risk assessment," *Safety Science*, vol. 118, pp. 475–486, 2019.

3. S. K. Maurya and A. Choudhary, "Deep learning based vulnerable road user detection and collision avoidance," in *Proceedings of the 2018 IEEE International Conference on Vehicular Electronics and Safety (ICVES)*, pp. 1–6, IEEE, Madrid, Spain, 2022 September.
4. M. Ahmad, T. Younis, M. A. Habib, R. Ashraf, and S. H. Ahmed, "A review of current security issues in internet of things," *Recent Trends and Advances in Wireless and IoT-enabled Networks*, Springer, Singapore, 2021.
5. M. Garcia-Venegas, D. A. Mercado-Ravell, and C. A. Carballo-Monsivais, "On the safety of vulnerable road users by cyclist orientation detection using Deep Learning," 2020, M. Goldhammer, S. Köhler, S. Zernetsch, K. Doll, B. Sick, and K. Dietmayer, "Intentions of vulnerable road users—detection and forecasting by means of machine learning," *IEEE Transactions on Intelligent Transportation Systems*, vol. 21, no. 7, pp. 3035–3045, 2019.
6. Z. Ahmed and R. Iniyavan, "Enhanced vulnerable pedestrian detection using deep learning," in *Proceedings of the 2019 International Conference on Communication and Signal Processing (ICCSP)*, pp. 0971–0974, IEEE, Chennai, India, April 2022.
7. M. Babar and F. Arif, "Real-time data processing scheme using big data analytics in internet of things based smart transportation environment," *Journal of Ambient Intelligence and Humanized Computing*, vol. 10, no. 10, pp. 4167–4177, 2019.
8. E. Carter, P. Adam, D. Tsakis, S. Shaw, R. Watson, and P. Ryan, "Enhancing pedestrian mobility in smart cities using big data," *Journal of Management Analytics*, vol. 7, pp. 1–16, 2020.
9. W. Tabone, J. de Winter, C. Ackermann et al., "Vulnerable road users and the coming wave of automated vehicles: expert perspectives," *Transportation Research Interdisciplinary Perspectives*, vol. 9, Article ID 100293, 2021.
10. Y. Lv, Y. Duan, W. Kang, Z. Li, and F.-Y. Wang, "Traffic flow prediction with big data: a deep learning approach," *IEEE Transactions on Intelligent Transportation Systems*, vol. 16, no. 2, pp. 865–873, 2014.
11. A. Sharma, G. Singh, and S. Rehman, "A review of big data challenges and preserving privacy in big data," in *Advances in Data and Information Sciences*, pp. 57–65, Springer, Singapore, 2020.
12. S. Boubiche, D. E. Boubiche, A. Bilami, and H. Toral-Cruz, "Big data challenges and data aggregation strategies in wireless sensor networks," *IEEE Access*, vol. 6, pp. 20558–20571, 2018.
13. D. Nallaperuma, R. Nawaratne, T. Bandaragoda et al., "Online incremental machine learning platform for big data-driven smart traffic management," *IEEE Transactions on Intelligent Transportation Systems*, vol. 20, no. 12, pp. 4679–4690, 2019.
14. A. Ahmad, M. Babar, S. Din et al., "Socio-cyber network: the potential of cyber-physical system to define human behaviors using big data analytics," *Future Generation Computer Systems*, vol. 92, pp. 868–878, 2019.
15. M. Elkhodr, B. Alsinglawi, and M. Alshehri, "Data provenance in the internet of things," in *Proceedings of the 2018 32nd international conference on advanced information networking and applications workshops (WAINA)*, pp. 727–731, IEEE, Krakow, Poland, May 2018.

16. S. G. Farrag, N. Sahli, Y. El-Hansali, E. M. Shakshuki, A. Yasar, and H. Malik, "STIMF: a smart traffic incident management framework," *Journal of Ambient Intelligence and Humanized Computing*, vol. 12, no. 1, pp. 85–101, 2021.
17. A. Ahmad, M. Khan, A. Paul et al., "Toward modeling and optimization of features selection in big data based social internet of things," *Future Generation Computer Systems*, vol. 82, pp. 715–726, 2018.
18. J. Yang, Y. Han, Y. Wang, B. Jiang, Z. Lv, and H. Song, "Optimization of real-time traffic network assignment based on IoT data using DBN and clustering model in smart city," *Future Generation Computer Systems*, vol. 108, pp. 976–986, 2020.
19. N. Cárdenas-Benítez, R. Aquino-Santos, P. Magaña-Espinoza, J. Aguilar-Velazco, A. Edwards-Block, and A. Medina Cass, "Traffic congestion detection system through connected vehicles and big data," *Sensors*, vol. 16, no. 5, p. 599, 2016.
20. J. Zhang, B. Chen, Y. Zhao, X. Cheng, and F. Hu, "Data security and privacy-preserving in edge computing paradigm: survey and open issues," *IEEE Access*, vol. 6, pp. 18209–18237, 2018.
21. C. Johnsson, A. Laureshyn, and T. De Ceunynck, "In search of surrogate safety indicators for vulnerable road users: a review of surrogate safety indicators," *Transport Reviews*, vol. 38, no. 6, pp. 765–785, 2018.
22. A. Siulagi, J. F. Antin, and L. J. MolnarS. Bai, S. Reynolds, O. carsten, and R. greene-roesel, "Vulnerable road users: how can automated vehicle systems help to keep them safe and mobile?" in *Road Vehicle Automation*, vol. 3, pp. 277–286, Springer, Cham, 2016.