Journal of Theoretical and Applied Information Technology
15th December 2025. Vol.103. No.23
© Little Lion Scientific

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

AN ANALYTICAL MACHINE LEARNING MODEL FOR PREDICTING DEFECTS IN IoT BASED APPLICATIONS

MANUJAKSHI B C[1], SHASHIDHAR T M[2], LATHARANI T R[3], RENUKADEVI S[4], N SHIVAKUMAR[5]

[1]Associate Professor, School of Computer Science and Engineering, Faculty of Engineering and Technology, JAIN(Deemed-to-be-University), Bengaluru, Karnataka, India
[2] Professor, Department of Computer Science and Engineering (Artificial Intelligence and Machine Learning), Harsha Institute of Technology, Bengaluru, Karnataka, India
[3]Associate Professor, Department of Computer Science & Engineering, Ramaiah University of Applied Sciences, Bengaluru, Karnataka, India
[4] Assistant professor, School of CS & IT, JAIN (Deemed- to-be University), Bengaluru, Karnataka, India.
[5]Associate Professor, Department of Computer Engineering, Marwadi University, Rajkot, Gujarat, India

manujakshibc@gmail.com, shashilara@gmail.com, lathaquick@gmail.com, selvamrenukadevi@gmail.com, drsivakumar.nadarajan@gmail.com

ABSTRACT

In past half decade, there has been considerable progressive advancement towards performing big data analysis over large scale complex network of Internet-of-Things (IoT) using machine learning approach for performing predictive analysis. The proposed study presents a simplified approaches to perform defect analysis using machine learning in large scale IoT based applications like healthcare services. In this study various distinct sensory data are obtained from multiple sources in healthcare services which are further subjected to data aggregation and preprocessing which not only improves the data quality but also reduces the computational burden of training operation. The proposed system is evaluated with multiple machine learning model like random forest which can perform identification as well as classification of defect with respect to its logical classes defined in this study model. The study outcome shows that random forest offers 45% of improved accuracy and 79% of faster processing in comparison to other frequently used machine learning models.

Keywords: *Internet-of-Things, Machine Learning, Prediction, Accuracy, Aggregation*

1. INTRODUCTION

With an increasing adoption of Internet-of-Things (IoT), sensory-based applications or devices are frequently being used to capture data that are further subjected for sophisticated analytical operations [1][2]. In the perspective of complex data analysis, the role of big data is highly essential that assists in extracting intellectual information from the problem scenario [3]. At present, there are various approaches towards big data analytics which is capable of processing sensory data in large scale applications [4]-[6]. However, there are various challenges associated with this approach. The primary challenge is the adoption of artificial intelligence and learning-based approach in big data in order to extract knowledge [7]. Adoption of machine learning calls for dependencies towards training data as well as selection of an appropriate learning model itself is a biggest challenge in varying use cases of IoT big data analysis [8]. Further, adoption of deep learning-based approaches too is shown to offer highest accuracy in its predictive performance; however, computational challenges associated with it cannot be denied [9]. A closer look into the various study models shows that although there are various analytical approaches in big data but there are few approaches which can be deemed as reliable. The specific reason behind this is its incapability to carry out analysis of defect in data management or analysis while performing predictive operation. At present, there are multiple approaches towards defect analysis of software modelling, however, they are studied from the perspective of software engineering and not from the perspective of machine learning approach. The methodology of solving problem in both are quite different with varying perspective. Hence, it is initially necessary to assess the existing big data approaches in order to understand its effectiveness prior to study its defect. Existing methodology of machine learning and big data has been already proving itself quite beneficial towards solving the complex problems of real-world data management. Hence, the prime motivation of the proposed study is to harness the complex problem-solving capability of big data in order to identify the defects associated with more practical world data. Hence, the proposed scheme considers a use-case of IoT where various smart appliances are connected to generated a massive set of physical world data, which in due course of time takes the shape of complex data.

Therefore, the proposed study model introduces a simplified and cost-effective baseline architecture where practical world defect can not only be identified but also can be subjected to an effective classification. The contribution of the proposed study model are as follows:

i) an analytical model is presented which can aggregate streams of data from multiple sensory devices in large scale IoT environment,
ii) the model is capable of performing preprocessing to improve the quality of the data prior to perform an analytical operation,
iii) the proposed scheme is assessed with simplified multiple machine learning approach towards performing defect forecasting considering real-time use cases of IoT,
iv) a unique baseline architecture is designed which is applicable for any form of dynamic streams of incoming traffic using big data approach which is not only lightweight but also offer better balance between computational accuracy and faster response time.

The organization of the paper is as follows: Section 2 presents highlight of existing literatures associated with analytical approaches followed by briefing of identified problems in Section 3. Section 4 presents briefing of adopted research methodology while discussion of the system implementation is carried out in Section 5. Accomplished result of the study model and its discussion is carried out in Section 6 while Section 7 provides summary of study contribution and future direction of the research work.

2. REVIEW OF LITERATURE

At present, there are various study models being developed towards analyzing big data-based approach on sensory data. The work carried out by Hashidzume et al. [10] have implemented a model towards integrating big data and device physics for investigating various form of parametric failures. The study has developed an experimental prototype towards identification of failures. Similar form of integrated model is also witnessed in work of Yu et al. [11] where big data is integrated with IoT for effectively monitoring the state of health. The study model has developed a filter for feature of high noise in order to perform an autonomous selection of sensors for better data quality. March et al. [12] have developed an approach for improving the management of telemetry data. Study towards identifying the security threats in IoT distributed scenario is carried out by Li et al. [13] where the model can capture an essential insight towards time-series data on the basis of temporal and spatial characteristic of data in order to model the malicious behavior in IoT using big data approach. Yang et al. [14] have presented a model which is capable of recovering the artifacts in cloud-enabled big data. The model carries out approximation of data sources in order to substitute the identified data with presence of artifacts in cloud environment. Adoption of big data is also reported in work of Huang et al. [15] where communication channel is analyzed for 5G based services. A hybrid machine learning model using radial basis and feed-forward neural network is designed to investigate the statistical properties of channel.

Adoption of fuzzy-rule based approach for enhancing analytical operation in big data is carried out by Manogaran et al. [16] where the idea is to perform a prediction towards sensory information of image data using semantic analysis. Yin et al. [17] have carried out a scheme where a mining approach is applied over a sensory data in IoT. The idea of this work is to capture the historical trend of sensor data where the correlation is obtained among different physical attributes in industrial big data. Adoption of reinforcement learning towards performing analytical operation over big data is carried out by Xu et al. [18] where neural network is combined with reinforcement learning for developing a task scheduling approach. The study model further constructs an unidirectional bridge network as well as random pool sampling on the basis of historical data in order to control the operational cost of datacenters. Subramaniam et al. [19] have developed a computational model towards improving management and storage of big data in IoT. Kumar et al. [20] have used MapReduce and Hadoop, an existing software framework, where machine learning is applied over big data.

Nearly similar adoption of Hadoop was also witnessed in work of Mjhool et al. [21] considering educational big data. Study towards retaining higher degree of data quality is carried out by Pratibha et al. [22] using a simplified analytical modelling. Guo et al. [23] have presented a stud which can control the degree of data redundant while perform energy harvesting. The idea of this work is to present a unique clustering model towards facilitating transmission of non-redundant data. Zhu et al. [24] have developed a mechanism towards monitoring big data attributes considering a use case of high-end manufacturing. The technique uses deep learning approach over heterogeneous big data considering various machining dynamics. Li et al. [25] have developed an ontology-based framework towards improving monitoring mechanism in healthcare system. The study model integrates ontology and bridge structure in order to accomplish higher accuracy over big data platform. Zhu et al. [26] have developed a stochastic model towards analyzing industrial big data with an idea towards determining the relevance on the basis of feature weights. The study also uses bagging mechanism for parallel computing in order to maintain better stability of system.

Kulkarni et al. [27] have used big data based analytical operation in order to perform analysis of road traffic using Bluetooth. The idea is towards performing data aggregation and analysis on the basis of Bluetooth-based communication channel. Qiu et al. [28] have used big data approach on IoT in order to monitor the fitness factors using particle swarm optimization. Further a dual-direction chaotic search approach is implemented in order to overcome the problems of conventional swarm optimization. A reverse learning approach is implemented for resisting the particle towards converging in local optima. Zhao et al. [29] have developed a model which can perform monitoring of health using sensory data that furnishes information about plantar stress and electromyography using machine learning. Yang and Ge [30] have used regression approach to analyzing sensory data of industry. The model has developed for parallel computation while Bayesian factor of non-linear variant is used for regression on the basis of probability modelling.

Puttinaovarat and Horkaew [31] have used machine learning towards large crowdsourced data in order to forecast natural calamities. Sergiyenko and Tyrsa [32] have developed an optimized communication scheme to be used in robotics for improving group navigation. The idea of work is mainly towards data fusion and path planning using big data approach. Wang et al. [33] have developed a unique scheme towards improving the training operation for industrial big data in IoT using tensor train, which is basically a distributed processing technique for big data. The idea of this work is to carry out decomposition of tensor train data in order to improve the computational performance when the system model carry out analysis of big data. The study outcome has exhibited better verification model. The work carried out by Khudhur and Jeiad [34] have developed a big data based streaming approach using multiple learning approach e.g., random forest, K-Nearest Neighbor, Decision Tree, etc. Based on the above-mentioned detailed survey, the conclusion is drawn to have a efficient model to identify the defect present in any of the IoT applications.

3. RESEARCH PROBLEM

Apart from the above-mentioned studies, the motivation of the proposed study is drawn from current work towards defect forecasting using big data [35]-[37]. These studies have been carried out towards identifying a defect present in software modelling where random forest has been exhibited to show better results from the perspective of machine learning approach. The identified research problems in existing schemes are: i) there are few reporting of any study model which offers a cost effective learning scheme towards defect analysis ii) majority of the work in big data has been focused on classification with few emphasis towards identifying defects in large scale monitoring environment, and iii) there are less studies being carried out in practical context of use case of IoT towards defect analysis, they are either highly iterative or dependent on predefined trained data and not reported of any applicability towards analyzing streamed data.

4. PROPOSED METHODOLOGY

This method is an extension of our prior work [38][39] towards evolving up in novel data management in bigdata. This part of the solution mainly focuses on rectifying the possibilities of any artifacts present in data which could significantly affect in prediction[42] performance. The proposed scheme takes the input sensory data that is further subjected to machine learning approach in order to identify and classify the defects. The architecture of proposed scheme is as shown in Fig.1

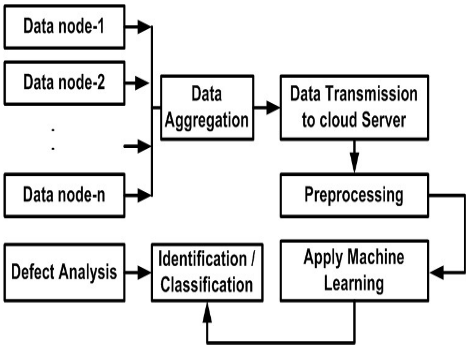

Figure 1 Architecture of Proposed Method

According to Fig.1, the proposed scheme initially performs data aggregation from various streams of *n* sensory data nodes. The aggregated information is then forwarded to the cloud server where the prime operation of defect analysis is carried out. The acquired data is initially preprocessed followed by classification on the basis of multiple machine learning approach. The machine learning models are developed using a use-case of multiple sensory devices to state the condition of proper functionalities and error-prone operational defect. The training operations are carried out in order to yield the final outcome of identification of defect, forecasting defect, and also offering the solution towards mitigating the forecasted defects. The novelty of the adopted methodology from the existing system are that it considers a specific use-cases of multiple and heterogeneous data nodes which generates an explicit sensory information subjected to further analysis. Adopting the framework of our prior model [38][39], the scheme can easily address the unstructured issues of incoming stream by transforming them to highly semi-structured data. Another significant novelty of this scheme is that although it considers input from incoming stream of raw data, but it identifies the prime categories of data types of each data nodes and stores them in permanent cloud storage unit. This process thereby resists reading all the redundant information from incoming sensory stream leading to less saturation state in cloud storage units. Further, adoption of machine learning approach assists not only to acquire an optimal state of accuracy towards predictive analysis but also attain less computational effort while performing the task of detection and classification. The next section further elaborates the adopted methodology of proposed scheme.

5. PROPOSED METHODOLOGY

The proposed system design targets towards developing a robust computational model that can carry out monitoring of real-world data as well as it is capable of forecasting possible defect. The idea is to positively determine the defect in the emergency services/devices followed by alerting the user about the defect causes and origin. The proposed scheme is design considering sensors and a processing unit mainly while module of machine learning has been newly constructed. The study considers adopting random forest in order to carry out classification. Fig.2 highlights the usage of multiple forms of sensors that are connected with the control units of design associated with healthcare sector. There are possibilities of various forms of sensors; however, the sensors adopted for implementation are fire sensor, air quality sensor, ultrasonic sensor, humidity sensor, temperature sensor, and current sensor. The control unit captures the target environmental data from the sensors where the data are restored in memory of these devices while some of the surplus data are also allowed to be migrated to the cloud server. The concept of data processing is used for processing the practical world data by the cloud server while the problem related to classification [43] and defect forecasting is carried out by the machine learning approach. The proposed scheme considers an extensive amount of data for the purpose of training using machine learning. Therefore, the proposed scheme demands the data to be processed for defect forecasting along with classification that performs updating of the data with corresponding outcomes. Once the fault forecasting is successfully accomplished, the computed data is retained in cloud server that can be accessed by the end-users via any form of handheld device or application. The score of defects associated with the prediction is carried out using classification of random forest for all the incoming stream of data. The front end of the cloud server is used for collected all the historical data associated with each different types of sensors.

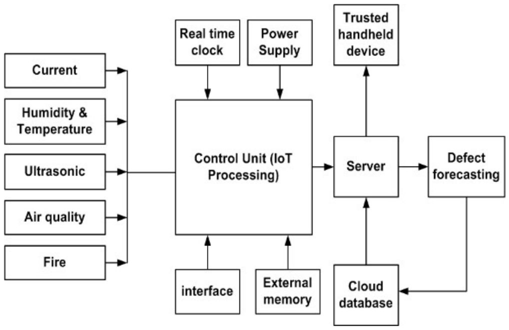

Figure 2 Adopted System Design for Defect forecasting in healthcare Sector

It should be noted that proposed scheme adopts Fig.2 as the system design which is responsible for generating a voluminous structure of data which are not only vast but also streamed. Table 1 showcase the sample of test data being generated by the sensory module which are fused in one node followed by forwarding them to the cloud server. The back-end process of cloud server receives this generated data by the individual sensor node for the purpose of defect forecasting while application of machine learning is carried out for similar purpose. The data is rigorously processed by the proposed model that is followed by the computation of defect outcome residing in the database system of the cloud server. The study model also applies an embedded scheme in order to consider the capacity towards reading and writing data over the database on cloud servers. Table 1 highlights the finally processed data where a sub-document is created in order to embedded various individual sensory data.

Table 1 Example of processing of sensory data in proposed scheme

Initial Sensory Data	Processed Outcome

TimeDate: “1900-01-23-2023”
fire: “0”
humidity: “0”
temperature: “1”
motion: “1”
qualityAir: “1”
currentSwitch: “1”
light: “1.002677”
television: “3.68891”
airCondition: “9.56711”
	ProcID: “766HGF900011”
	sensorData:{
	TimeDate: “1900-01-23-2023”
	fire: “0”
	humidity: “0”
	temperature: “1”
	motion: “1”
	qualityAir: “1”
	Currentswitch: “1”
	forecastOutcome: {
	television: “OFF”
	airCondition: “REPAIR”
	light: “REPLACE”}

In the proposed scheme, the mechanism of the defect forecasting is carried out in order to perform surveillance of the data from multiple computing devices and sensory applications. Irrespective of any form of structure of the room or indoor location, the proposed model is highly extensible of its design applicability.

5.1 Defect Forecasting

The prime agenda of the proposed scheme by implementing machine learning is to ensure that the system model used in healthcare is working without any error or defect. Assuming various form of big data-based application is running within a healthcare sector, it is quite inevitable that there are fair possibilities of defect when it comes to perform an analytical assessment to some of the clinical data which may look redundant but difficult to confirm this fact. Fig.3 highlights the flow of the implementation where the proposed scheme initially carries out aggregation of data followed by performing normalization and feature extraction in its preprocessing operation over the considered dataset. The proposed scheme also implements multiple machine learning approach in order to carry out classification of data e.g., DT (Decision Tree), Random Forest (RF), Gaussian Naïve Bayesian (GNB), and K-Nearest Neighborhood (KNN) algorithm. The process of identification as well as classification of classes of defects in the clinical data is carried out by random forest algorithm. The aggregation of the data is carried out from various devices and sensors for performing processing and assessing its predictive outcomes. Multiple number of instances are considered in order to categorize into 4 classes of defect i.e., properly functional, supply of power to turn in off state, requirement of replacement of device, and requirement of repair/maintenance of device during the entire predictive process.

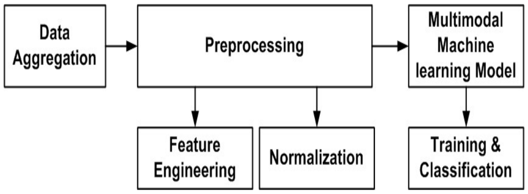

Figure 3 Proposed Implementation Flow

Fig.4 highlights the mechanism of implementation carried out in proposed scheme. Initially, the value of current sensor associated with television, refrigerator, light, printer, surveillance camera, and air-conditioner are chosen. These sensory data are then forwarded to IoT based processing unit that further uses available network (IEEE 802.11) to transmit the data to the cloud server. The obtained data is then subjected to the machine learning approach for performing defect forecasting. In case there is no defect being observed, the proposed scheme saves the data to the cloud database otherwise it recommends for solution as stated in Table 1. The database saved in cloud served can be further accessed by the user from their trusted handheld device.

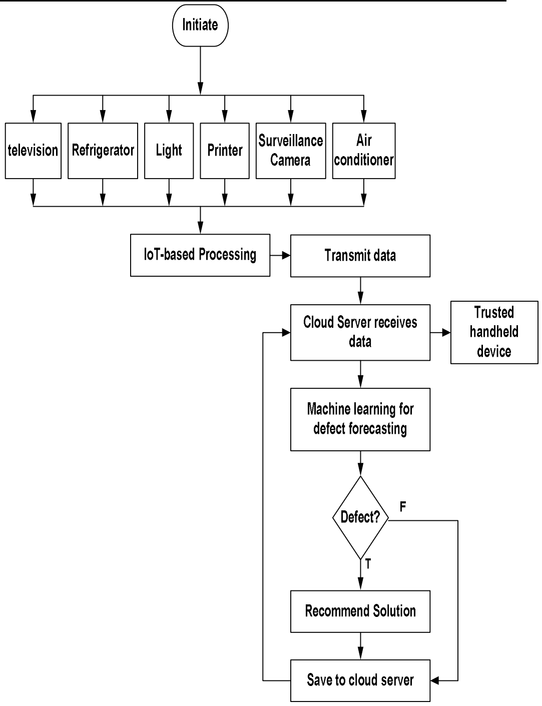

Figure 4 Adopted Mechanism of internal operation

Therefore, it can be seen that a very simplified form of architectural implementation is carried out in proposed scheme in order to identify and classify the form of defect present in clinical data using simplified machine learning. The next section outlines the discussion of the outcome being accomplished after implementing the proposed scheme.

6. RESULTS AND ANALYSIS

The proposed scheme is analytically implemented in python scripts where an IoT environment is considered considering television, refrigerator, light, printer, surveillance camera, and air-conditioner. A synthetic dataset is constructed which is complying of standard big data [40]. The scripting of the python is carried out in highly flexible form which is also capable of processing real-time values captured from experimental IoT devices (Raspberry Pi). Hence, irrespective of the source of data being collected, the emphasis is actually given to ensure that dataset is suitable for analysis. For this purpose, the dataset is first subjected to our prior model [38][39] which transforms the data into highly structured suitable for applying further analytics. Further, statistical information of mean, standard deviation, and variance are estimated. Table 2 highlights the numerical values of the synthetic data used:

Table 2 Synthetic Data Adopted For Experiment

Classes of Defect	param	air-conditioner	refrigerator	television	printer	surveillance camera	light
properly functional	Mean	11.775	7.787	4.833	3.988	3.706	2.624
	Standard deviation	2.484	2.478	2.455	2.37	2.469	1.104
	Variance	2.997	2.981	2.9174	2.665	2.9667	1.9218
requirement of replacement of device	Mean	11.543	7.5472	4.5513	3.5476	3.5497	1.1001
	Standard deviation	1.5711	1.451	1.4892	1.4891	1.4887	1.3998
	Variance	1.3227	1.2696	1.2541	1.2547	1.2537	1.194
requirement of repair/maintenance of device	Mean	11.5515	7.5462	4.573	3.5533	3.5542	2.114
	Standard deviation	1.485	1.489	1.4938	1.4963	1.49166	1.3918
	Variance	1.2499	1.2546	1.2575	1.2519	1.255	1.1956
supply of power to turn in off state	Mean	1.1120	1.1110	1.11146	1.1112	1.1113	1.1119
	Standard deviation	1.25243	1.1878	1.1454	1.1332	1.1362	1.1228
	Variance	1.1200	1.1160	1.1122	1.1115	1.1117	1.1112

According to above Table 2, the four different classes of defect is being analyzed on the considered synthetic data which this data is subjected to the machine learning approach. In this case, the dataset is subjected to data processing followed by performing an exploratory analysis of data while classification is carried out using machine learning algorithm. Finally, the classification of the instances is carried out considering normal or defective score. The overall outcome has been assessed with multiple accuracy-based performance parameters.

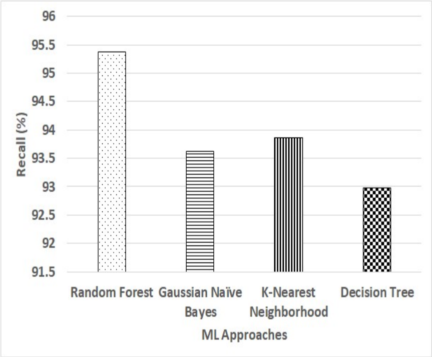

Figure 5 Comparative Analysis Of Recall

Fig.5 highlights the outcome of the comparative analysis where Random Forest has exhibited higher recall rate in comparison to other existing machine learning approaches. A closer look into Fig.6 showcases that precision of certain approaches (e.g., DT) is higher on the contrary to the lower recall rate observed in Fig.5. This can be justified by the learning-based approaches which is more inclined towards capturing majority of positive instances causing slight involvement of false positives or vice-versa.

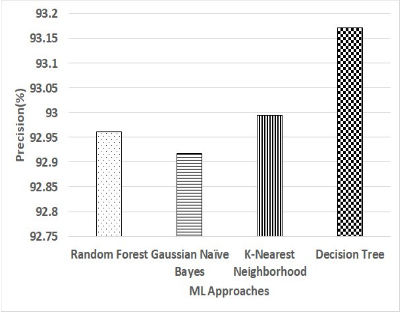

Figure 6 Comparative Analysis Of Precision

Similar trend is also observed for precision (Fig.6), F1-Score (Fig.7), and Accuracy (Fig.8). Based on all these outcomes, it can be stated that proposed scheme excels better predictive result of defect analysis and classification using random forest.

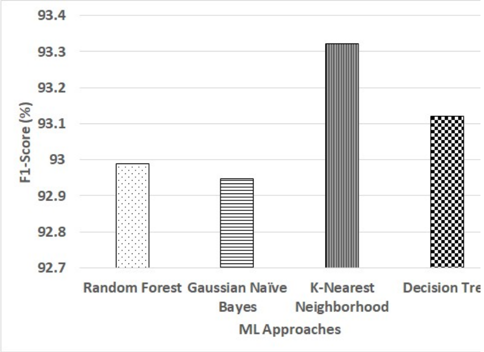

Figure 7 Comparative Analysis Of F1-Score

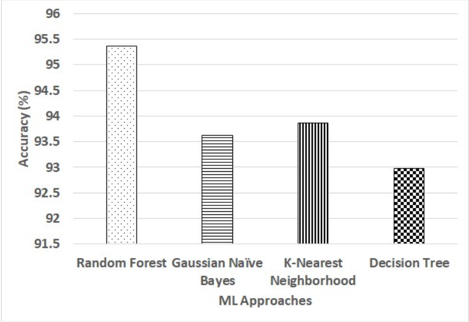

Figure 8 Comparative Analysis Of Accuracy

A closer look into each of the graphical analysis shows that accuracy of random forest is approximately 45% better compared to other machine learning approach. Although, KNN approach has too showed slightly increased accuracy (Fig.8), still it is highly sensitive to increasing stream of data. Gaussian Naïve Bayes assumptions towards independent features are bit unpractical and hence not suggested for critical healthcare applications. Further, a higher degree of instability is objected for decision tree when exposed to large data and hence its accuracy is potentially affected. It should be noted that accuracy outcomes obtained in this model is actually highly optimized value. It can be justified by the higher accuracy score which is already obtained by deploying our prior model [38][39] and while the high-quality preprocessed data when subjected to proposed scheme results in additional score of accuracy. Hence, optimal accuracy scores can be witnessed in proposed scheme.

The proposed scheme also performs evaluation of the processing time which represents the overall duration to execute the entire model. One of the prime impending rationales behind the adoption of this performance parameter is that its majority of the existing approaches lacks evaluation of final outcome with respect to computational effort. Lack of evaluation of actual processing time will lead to incomplete study of effectiveness in perspective of model utilization on practical environment in distributed environment.

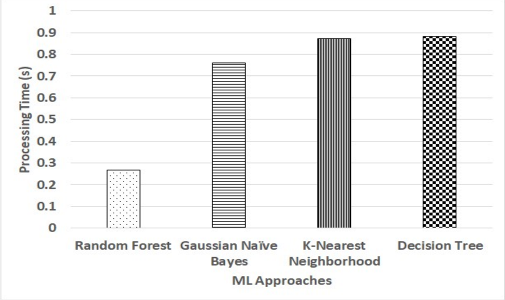

Figure 9 Comparative Analysis Of Processing Time

Fig.9 exhibits the comparative analysis of processing time where it shows that proposed system using random forest offers approximately 79% of reduced processing time in contrast with other machine learning schemes.

The discussion of the outcomes are as follows:

- *Decision Tree*: When decision tree is used for predictive analysis, it is noted that a minor alteration in the stream of data causes significant instability. Apart from this, it also suffers from overfitting problem. Further analysis showcase that if a static data is used than accuracy is slightly increased; however, when exposed to stream of large data in healthcare, the accuracy is adversely affected.
  - *Inference*: Decision Tree is suitable for small set of operation with smaller dataset and not to be used for continuous values of streamed data.
- *K-Nearest Neighborhood*: Irrespective of being a simplified predictive operation with a capability to carry out both regression and classification, this approach is found to have higher dependencies towards data quality. Apart from this, it heavily occupies an extensive memory while performing training operation. Its accuracy is affected when irrelevant features are encountered.
  - *Inference*: This machine learning approach is suitable for simplified classification especially of binary form. In presence of larger set of classification rules based on continuous data, the processing time is quite higher as well as accuracy parameters are less.
- *Gaussian Naïve Bayes*: One of the best parts of this approach is its lesser degree of dependency towards training data. Apart from this, both discrete and continuous data can be managed by this algorithm; however, it assigns a zero probability many times when it a categorical value is present its test data that is found to be absent in training data.
  - *Inference*: This approach is not meant for handling dynamic stream of dataset irrespective of its capabilities.
- *Random Forest*: Although the interpretability of this approach is quite less compared to decision tree, but it is capable of maintaining a higher accuracy which is not affected by any artifacts much. Hence, defect forecasting with higher accuracy is possible much cost effectively in this approach.
  - *Inference*: This approach is found suitable to perform predictive analysis of presence as well as classification of defects of data that are maintained in dynamic streams of incoming flow of traffic. It can maintain a good balance between higher accuracy and lower computational processing time.

Therefore, from the perspective of the outcome obtained in proposed scheme, it can be stated that proposed scheme offers a simplified baseline architecture towards performing defect analysis for any set of given data. Apart from this, the model is computationally found to be less complex when applied with random forest approach for prediction.

7. CONCLUSION

The proposed study model introduces a simplified analytical model which can perform defect analysis based on sensory information from IoT as well as cloud and applied with machine learning approach. The proposed scheme can identify and classify the predictive defects. The baseline architecture of proposed scheme also exhibits that joint operation of cloud with sensory model in IoT is capable of processing large incoming streams of data. The proposed study exhibits its applicability towards any critical as well as non-critical a IoT based application with a significant control over its intermittent occurring defect. The random forest offers the higher accuracy in less processing time.

The future work of this idea will be further investigated with inclusion of more complex use cases where a dynamic mobility can be introduced to the network data. Further work could be also carried out to investigate the impact of artifacts over various heterogeneous use-cases in IoT by considering prioritized services.

REFERENCES

[1] . D. N. Jha, P. Michalák, Z. Wen, R. Ranjan and P. Watson, "Multiobjective Deployment of Data Analysis Operations in Heterogeneous IoT Infrastructure," in IEEE Transactions on Industrial Informatics, vol. 16, no. 11, pp. 7014-7024, Nov. 2020, doi: 10.1109/TII.2019.2961676.

[2]. M. Taneja, N. Jalodia and A. Davy, "Distributed Decomposed Data Analytics in Fog Enabled IoT Deployments," in IEEE Access, vol. 7, pp. 40969-40981, 2019, doi: 10.1109/ACCESS.2019.2907808.

[3]. M. S. Mahmud, J. Z. Huang, S. Salloum, T. Z. Emara and K. Sadatdiynov, "A survey of data partitioning and sampling methods to support big data analysis," in Big Data Mining and Analytics, vol. 3, no. 2, pp. 85-101, June 2020, doi: 10.26599/BDMA.2019.9020015.

[4]. D. Syed, A. Zainab, A. Ghrayeb, S. S. Refaat, H. Abu-Rub and O. Bouhali, "Smart Grid Big Data Analytics: Survey of Technologies, Techniques, and Applications," in IEEE Access, vol. 9, pp. 59564-59585, 2021, doi: 10.1109/ACCESS.2020.3041178.

[5]. X. Cao, L. Liu, Y. Cheng and X. Shen, "Towards Energy-Efficient Wireless Networking in the Big Data Era: A Survey," in IEEE Communications Surveys & Tutorials, vol. 20, no. 1, pp. 303-332, Firstquarter 2018, doi: 10.1109/COMST.2017.2771534

[6]. M. Jahanbakht, W. Xiang, L. Hanzo and M. Rahimi Azghadi, "Internet of Underwater Things and Big Marine Data Analytics—A Comprehensive Survey," in IEEE Communications Surveys & Tutorials, vol. 23, no. 2, pp. 904-956, Secondquarter 2021, doi: 10.1109/COMST.2021.3053118

[7]. A. Cravero, S. Pardo, S. Sepúlveda, and L. Muñoz, “Challenges to Use Machine Learning in Agricultural Big Data: A Systematic Literature Review,” Agronomy, vol. 12, no. 3, p. 748, Mar. 2022, doi: 10.3390/agronomy12030748

[8]. P. Borrellas and I. Unceta, “The Challenges of Machine Learning and Their Economic Implications,” Entropy, vol. 23, no. 3, p. 275, Feb. 2021, doi: 10.3390/e23030275

[9]. A. H. Gandomi, F. Chen, and L. Abualigah, “Machine Learning Technologies for Big Data Analytics,” Electronics, vol. 11, no. 3, p. 421, Jan. 2022, doi: 10.3390/electronics11030421

[10]. T. Hashidzume, T. Yasui and T. Tanaka, "Rapid Resolution of Parametric Failures in the Process Development Period by Integrating Device Physics and Big Data," in IEEE Transactions on Semiconductor Manufacturing, vol. 34, no. 3, pp. 352-356, Aug. 2021, doi: 10.1109/TSM.2021.3072367

[11]. W. Yu, Y. Liu, T. Dillon, W. Rahayu and F. Mostafa, "An Integrated Framework for Health State Monitoring in a Smart Factory Employing IoT and Big Data Techniques," in IEEE Internet of Things Journal, vol. 9, no. 3, pp. 2443-2454, 1 Feb.1, 2022, doi: 10.1109/JIOT.2021.3096637.

[12]. R. De March, C. Leuzzi, M. Deffacis, F. Caronte, A. F. Mulone and R. Messineo, "Innovative Approach for PMM Data Processing and Analytics," in IEEE Transactions on Big Data, vol. 6, no. 3, pp. 452-459, 1 Sept. 2020, doi: 10.1109/TBDATA.2020.2995242.

[13]. F. Li et al., "Online Distributed IoT Security Monitoring With Multidimensional Streaming Big Data," in IEEE Internet of Things Journal, vol. 7, no. 5, pp. 4387-4394, May 2020, doi: 10.1109/JIOT.2019.2962788.

[14]. C. Yang, X. Xu, K. Ramamohanarao and J. Chen, "A Scalable Multi-Data Sources Based Recursive Approximation Approach for Fast Error Recovery in Big Sensing Data on Cloud," in IEEE Transactions on Knowledge and Data Engineering, vol. 32, no. 5, pp. 841-854, 1 May 2020, doi: 10.1109/TKDE.2019.2895612.

[15]. J. Huang et al., "A Big Data Enabled Channel Model for 5G Wireless Communication Systems," in IEEE Transactions on Big Data, vol. 6, no. 2, pp. 211-222, 1 June 2020, doi: 10.1109/TBDATA.2018.2884489.

[16]. G. Manogaran et al., "FDM: Fuzzy-Optimized Data Management Technique for Improving Big Data Analytics," in IEEE Transactions on Fuzzy Systems, vol. 29, no. 1, pp. 177-185, Jan. 2021, doi: 10.1109/TFUZZ.2020.3016346.

[17]. Y. Yin, L. Long and X. Deng, "Dynamic Data Mining of Sensor Data," in IEEE Access, vol. 8, pp. 41637-41648, 2020, doi: 10.1109/ACCESS.2020.2976699.

[18]. C. Xu, K. Wang, P. Li, R. Xia, S. Guo and M. Guo, "Renewable Energy-Aware Big Data Analytics in Geo-Distributed Data Centers with Reinforcement Learning," in IEEE Transactions on Network Science and Engineering, vol. 7, no. 1, pp. 205-215, 1 Jan.-March 2020, doi: 10.1109/TNSE.2018.2813333.

[19]. A. Subramaniam, N.A. Ibrahim, S.N. Jabar, S.A. Rahman, "Driving cycle tracking device big data storing and management", International Journal of Electrical and Computer Engineering, vol.12, No.2, pp.1402-1410, 2022. DOI: 10.11591/ijece.v12i2.pp1402-1410

[20]. S. Kumar, J., B.K. Raghavendra, S. Raghavendra, Meenakshi, "Performance evaluation of Map-reduce jar pig hive and spark with machine learning using big data" International Journal of Electrical and Computer Engineering, vol.10, No.4, pp.3811-3818, 2020. DOI: 10.11591/ijece.v10i4.pp3811-3818

[21]. A.Y. Mjhool, A.H. Alhilali, S Al-augby, "A proposed architecture of big educational data using hadoop at the University of Kufa", International Journal of Electrical and Computer Engineering, vol.9, No.6, pp.4970-4978, 2019. DOI: 10.11591/ijece.v9i6.pp4970-4978

[22]. B. Pratibha, K.J. Sankar, V. Sumalatha, "Novel framework of retaining maximum data quality and energy efficiency in reconfigurable wireless sensor network", International Journal of Electrical and Computer Engineering, vol.9, No.4, pp.2893-2901, 2019. DOI: 10.11591/ijece.v9i4.pp2893-2901

[23]. Z. Guo et al., "Minimizing Redundant Sensing Data Transmissions in Energy-Harvesting Sensor Networks via Exploring Spatial Data Correlations," in IEEE Internet of Things Journal, vol. 8, no. 1, pp. 512-527, 1 Jan.1, 2021, doi: 10.1109/JIOT.2020.3004554.

[24]. K. Zhu, G. Li and Y. Zhang, "Big Data Oriented Smart Tool Condition Monitoring System," in IEEE Transactions on Industrial Informatics, vol. 16, no. 6, pp. 4007-4016, June 2020, doi: 10.1109/TII.2019.2957107.

[25]. R. Li, T. Mo, J. Yang, S. Jiang, T. Li and Y. Liu, "Ontologies-Based Domain Knowledge Modeling and Heterogeneous Sensor Data Integration for Bridge Health Monitoring Systems," in IEEE Transactions on Industrial Informatics, vol. 17, no. 1, pp. 321-332, Jan. 2021, doi: 10.1109/TII.2020.2967561.

[26]. J. Zhu, M. Jiang, G. Peng, L. Yao and Z. Ge, "Scalable Soft Sensor for Nonlinear Industrial Big Data via Bagging Stochastic Variational Gaussian Processes," in IEEE Transactions on Industrial Electronics, vol. 68, no. 8, pp. 7594-7602, Aug. 2021, doi: 10.1109/TIE.2020.3003583.

[27]. A. R. Kulkarni, N. Kumar and K. R. Rao, "Efficacy of Bluetooth-Based Data Collection for Road Traffic Analysis and Visualization Using Big Data Analytics," in Big Data Mining and Analytics, vol. 6, no. 2, pp. 139-153, June 2023, doi: 10.26599/BDMA.2022.9020039.

[28]. A. R. Kulkarni, N. Kumar and K. R. Rao, "Efficacy of Bluetooth-Based Data Collection for Road Traffic Analysis and Visualization Using Big Data Analytics," in Big Data Mining and Analytics, vol. 6, no. 2, pp. 139-153, June 2023, doi: 10.26599/BDMA.2022.9020039.

[29]. Y. Zhao et al., "Flexible and Wearable EMG and PSD Sensors Enabled Locomotion Mode Recognition for IoHT-Based In-Home Rehabilitation," in IEEE Sensors Journal, vol. 21, no. 23, pp. 26311-26319, 1 Dec.1, 2021, doi: 10.1109/JSEN.2021.3058429

[30]. Z. Yang and Z. Ge, "Industrial Virtual Sensing for Big Process Data Based on Parallelized Nonlinear Variational Bayesian Factor Regression," in IEEE Transactions on Instrumentation and Measurement, vol. 69, no. 10, pp. 8128-8136, Oct. 2020, doi: 10.1109/TIM.2020.2993980.

[31]. S. Puttinaovarat and P. Horkaew, "Flood Forecasting System Based on Integrated Big and Crowdsource Data by Using Machine Learning Techniques," in IEEE Access, vol. 8, pp. 5885-5905, 2020, doi: 10.1109/ACCESS.2019.2963819

[32]. O. Y. Sergiyenko and V. V. Tyrsa, "3D Optical Machine Vision Sensors With Intelligent Data Management for Robotic Swarm Navigation Improvement," in IEEE Sensors Journal, vol. 21, no. 10, pp. 11262-11274, 15 May15, 2021, doi: 10.1109/JSEN.2020.3007856.

[33]. X. Wang, L. T. Yang, Y. Wang, L. Ren and M. J. Deen, "ADTT: A Highly Efficient Distributed Tensor-Train Decomposition Method for IIoT Big Data," in IEEE Transactions on Industrial Informatics, vol. 17, no. 3, pp. 1573-1582, March 2021, doi: 10.1109/TII.2020.2967768.

[34]. S. D. Khudhur and H. A. Jeiad, "LSDStrategy: A Lightweight Software-Driven Strategy for Addressing Big Data Variety of Multimedia Streaming," in IEEE Access, vol. 10, pp. 111794-111810, 2022, doi: 10.1109/ACCESS.2022.3215531.

[35]. K. Zhao, Z. Xu, T. Zhang, Y. Tang and M. Yan, "Simplified Deep Forest Model Based Just-in-Time Defect Prediction for Android Mobile Apps," in IEEE Transactions on Reliability, vol. 70, no. 2, pp. 848-859, June 2021, doi: 10.1109/TR.2021.3060937.

[36]. R. Duan, H. Xu, Y. Fan and M. Yan, "The Impact of Duplicate Changes on Just-in-Time Defect Prediction," in IEEE Transactions on Reliability, vol. 71, no. 3, pp. 1294-1308, Sept. 2022, doi: 10.1109/TR.2021.3061618.

[37]. F. Alghanim, M. Azzeh, A. El-Hassan and H. Qattous, "Software Defect Density Prediction Using Deep Learning," in IEEE Access, vol. 10, pp. 114629-114641, 2022, doi: 10.1109/ACCESS.2022.3217480.

[38]. B C Manjujakshi, K.B. Ramesh, "Novel holistic architecture for analytical operation on sensory data relayed as cloud services", International Journal of Electrical and Computer Engineering, vol.10, No.4, pp.4322-4330, 2020. DOI: 10.11591/ijece.v10i4.pp4322-4330

[39]. B C Manjujakshi, K.B. Ramesh, "Framework for cost-effective analytical modelling for sensory data over cloud environment", International Journal of Electrical and Computer Engineering, vol.9, No.5, pp.3822-3832, 2019. DOI: 10.11591/ijece.v9i5.pp3822-3832

[40]. https://hadoopilluminated.com/hadoop_illuminated/Public_Bigdata_Sets.html

[41]. Sivakumar, Mariyappan, & Prakash, P. G. Om. (2022). "Machine Learning-Based Algorithmic Approach for Enhanced Anomaly Detection in Financial Transactions." In Lecture Notes on Data Engineering and Communications Technologies (pp. 59). DOI: 10.1007/978-981-16-6605-6_59

[42]. Prakash, O. P. G., Abdul Rahman, A., Nagaraj, J., & Sivakumar, N. (2022). "Forecasting the User Prediction from Weblogs Using Improved IncSpan Algorithm." In Lecture Notes on Data Engineering and Communications Technologies (pp. 58). DOI: 10.1007/978-981-16-6605-6_58

[43]. Srivinay, C., M. B., Kabadi, M. G., Naik, N. & Chandrasekhara, S. P. R. (2023). Stock Price Classification Based on Hybrid Feature Selection Method. *Journal of Computer Science*, *19*(2), 274-285. https://doi.org/10.3844/jcssp.2023.274.285