A Comparative Analysis of Machine Learning Algorithms For Healthcare Device Data of Social IoT

D. Bhavya
Research Scholar
JSS Science and Technology University (Formerly SJCE)
Mysuru, Karnataka
India
bhavyadb@gmail.com

D.S. Vinod , S.P. Shiva Prakash
JSS Science and Technology University (Formerly SJCE)
Mysuru, Karnataka
India
(dsvinod, shivasp)@sjce.ac.in

***Abstract*—In recent times, Social IoT(SIoT) plays a important role in everyday life. These devices communicate between each other and generate various data, in which, few healthcare devices helps in monitoring the activities of human and keep track about the health. Data accumulated from these devices are analyzed based on the sensitivity of device that adds a weightage in understanding the criticality of the person's health. Many different machine learning models exists for discerning the imminence of health. Hence, in this paper, a comparative study of several machine learning algorithms that are used for classification of data are contrived. The experiment have been conducted using R Programming platform and healthcare dataset. The result shows Naive bayes model performs better with both time and space complexities for enormous healthcare data.**

I. Introduction

Social IoT(SIoT) is a next leap in advancement of IoT technology. It involves device interaction with sensors and actuators to receive and execute commands while attaining an predefined objective. SIoT makes device interaction as argumentation for readability and tractability of decisions making it reliable.

Healthcare is one of the several domains where IoT applications are used to monitor and track patients records. Nowadays data generated from each of these devices are huge as the data is recorded at small intervals of time and each recorded data is critical in patients diagnosis. Fig. 1 shows the sensors connected on the host system that generate data. This data is analyzed over various techniques with their parameters and a result is obtained. Finally, the result is transmitted over the internet.

Categorizing and analyzing of such vast amount of data in real time is critical, as these data lead to congestion in network leading to errors in generated data and delays in transmission. Furthermore, compromising the system performance and efficiency. Ultimately, making device unreliable at emergency situations.

Many methods are proposed to address the issues in data collection and analyzation in IoT devices, one such method is machine learning technique. The machine learning technique in IoT uses minimal storage, minimal network and minimal time complexity to collect and analyze the data. Hence, leading to overall performance improvement in the devices.

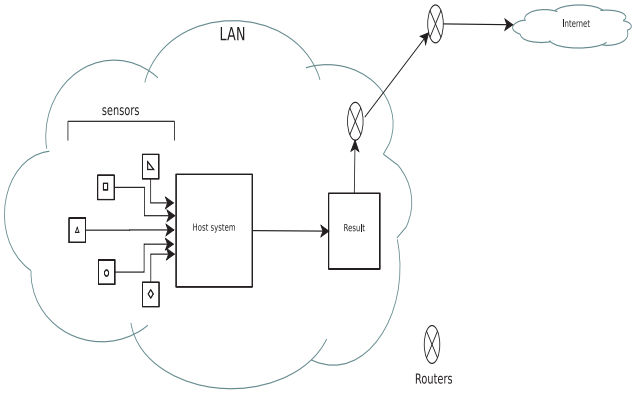

Fig. 1. Device communication network

The rest of the paper is structured as comply. The overview of the related works are discussed in section II. Section III defines problem statement. The comparative analysis is shown in IV. The results obtained are discussed in Section V and Section VI presents conclusion.

II. Related works

In this section, work related to IoT, machine learning techniques are shown.

Authors Arjun Chandra et.al. [1] proposes framework in addition to presenting detailed empirical results and comparisons with a wide range of algorithms in the machine learning literature. This framework shows its effectiveness by allowing accuracy and diversity as evolutionary pressures exerted at multiple levels of abstraction. Authors Wenting Tu et.al. [3], proposes a framework to combine domain-merge objectives and class-separate objectives to achieve cross-domain representation learning. Authors Y. Bengio et.al. [6] surveyed recent works on the area of deep learning and unsupervised feature learning, gives insights on the advancement in auto encoders, probabilistic models, deep networks, and manifold learning. Authors Qiu et.al. [14] reviewed the machine learning tech-

niques and highlight some promising learning methods, such as representation learning, deep learning, distributed and parallel learning, transfer learning, active learning, and kernel-based learning. Authors Jayavardhana Gubbi et.al. [5] proposed the key enabling technologies and application domains that drives the future IoT researches. A Cloud implementation using Aneka, which is based on coordination of private and public clouds is also discussed in the model. Authors C. Tsai et.al. [11] discussed on the IoT and gave review of the features of "data from IoT" and "data mining for IoT". Authors Gil et.al. [15] reviewed the IoT related surveys in order to provide well incorporated and context aware intelligent services for IoT. Authors C. Perera et.al. [8] presented the survey that address a broad range of techniques, methods, models, functionalities, systems, applications, and middle ware solutions related to context awareness and IoT. Authors Islam et.al. [12] reviewed advances in IoT-based healthcare technologies and surveyed the state-of-the-art network architectures/platforms, applications, and industrial trends in IoT-based healthcare solutions. Authors S. Athmaja et.al. [17] presented a literature survey of different machine learning techniques. Authors Raouf Boutaba et.al. [20] presented the application of diverse machine learning techniques in routing, traffic prediction and classification, fault management, congestion control, network security, and QoS and QoE management, limitations, insights, research challenges and future opportunities to advance machine learning in networking. Authors Kumar Donta et.al. [21] presented various machine learning algorithms for wireless sensor networks with their advantages, drawbacks, and parameters effecting the network lifetime. Furthermore, discussed on machine learning algorithms for congestion control, synchronization, energy harvesting and mobile sink scheduling. Authors M. G. Kibria et.al. [22] discussed on data sources and the role of machine learning, artificial intelligence in making the system intelligent regarding being self-adaptive, self-aware, dedicated and authoritative.

Authors Atzori et.al. [4] proposes the following: i) identify appropriate policies for the establishment and the management of social relationships among devices in order to make navigable social network, ii) describe a IoT architecture with functionalities that are required to integrate things into a social network. iii) analyze the characteristics of the SIoT network structure by means of simulations. Authors M. Lippi et.al [18] illustrate how argumentation naturally enables a form of conversational coordination between the devices through practical examples and a case study scenario. Authors Atzori et.al. [2] propose to build a social network and a framework that can be applied for the implementation of the SIoT. Authors H. Z. Asl et.al. [8] identified the objects that are likely to play a important role in the interactions among smart objects in the IoT. Authors L. Atzori et.al [9] investigated the opportunities from the integration of social networking concepts into the Internet of Things, presented the critical technical challenges of ongoing research activities.

Authors Badraddin Alturki et.al. [16] explored a hybrid approach in cloud level and network level, to build effective data analytics in IoT to overcome their specific strengths and respective weaknesses. Authors Yoshua Bengio et.al. [7] gives better representation of the model that can produce Markov chains that mixing faster between modes. Mixing between modes that are more efficient at higher representation levels. The higher level samples uniformly fills the occupied space and high density unfolds manifold when represented at higher levels. Authors J. L. Berral-García et.al. [13] proposed platforms and tools that are being served and developed in order to help researchers, students to learn from their data automatically, most of those platforms are coming from big companies like Microsoft or Google, or from incubators at the Apache Foundation.

III. Problem Statement

Increase in usage of SIoT devices in healthcare domain escalates data generation, leading to affliction in analyzing and interpreting the data. To overcome this situation, many models such as supervised and unsupervised algorithms under machine learning are proposed. But, no algorithm is preferred for healthcare data originated from heterogeneous SIoT. Hence, there is a need to identify a better machine learning algorithm to determine the criticality of data.

IV. Dataset

Healthcare data is taken from [23] and analyzed on R programming platform. This dataset consists of 296792 cases of 8 attributes as shown in Table I and 8 sensors as represented in Table II. Table I describes the features present in the dataset, here $type$ feature has values "symbolic" stating the values are discreet and "continuous" for the values that are uniformly increasing. The "measure" stores a value generated at every 10 ms time interval from each device. Table II consists of different sensors that are participating in the data generation and their respective type id's.

TABLE I. Features of Dataset

No	Attribute	Type	Description
1	device_id	symbolic	it shows different number of devices in dataset
2	device_type	symbolic	it shows various types of devices in dataset
3	id_user	symbolic	it is the patient's id
4	timestamp_start	continuous	it represents the start time of the device
5	timestamp_stop	continuous	it represents the stop time of the device
6	x	symbolic	defines the position of the device
7	y	symbolic	defines the position of the device
8	measure	symbolic	it records the measurement in device at each timestamp_stop

TABLE II. Features of Sensors

Sensor	Device type
Electrocardiogram	8
Blood Volume Pulse	9
Galvanic Skin Response	10
Respiration	11
Skin Temperature	12
Surface Electromyography (on the Zygomaticus major muscles)	13
Surface Electromyography (on the Corrugator supercilli muscles)	14
Surface Electromyography (on the Trapezius muscles)	15

V. Evaluation of Machine Learning Techniques

Data from the device is collected at fixed intervals of time. After collection of data, pre-processing of dataset is made by removing null values and data standardization. The data is dimensionally reduced by removing x and y as they are the physical positions of the sensors. Then, the dataset is

segregated into training dataset and testing dataset in the ratio of [7 3]. Later training dataset is used to train machine learning models as shown in Fig. 2. Finally, testing dataset is analyzed with comparison to training by each machine learning model to assess their performance.

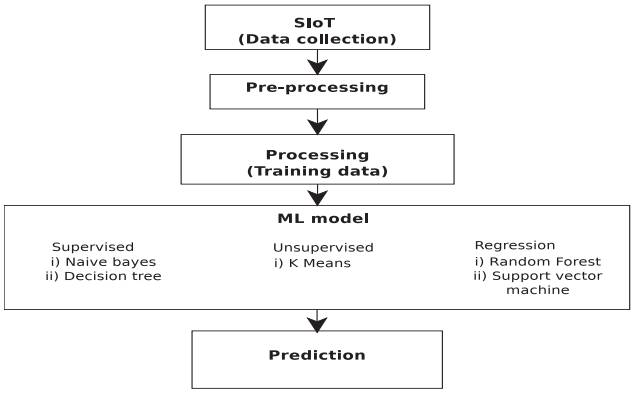

Fig. 2. Factors affecting solutions of machine learning

TABLE III. PROBABILITY DISTRIBUTIONS OF MACHINE LEARNING ALGORITHMS

No	Algorithm	Distribution
1	Naive bayes	Gaussian
2	Decision tree	Non parametric
3	K means	Normal
4	Random forest	Normal
5	Support vector machine	Gaussian

Each of the machine learning algorithm participates as shown in Fig. 3, where training of the dataset depends upon the probability distribution as shown in Table III. Decision tree is non-parametric in nature as it won't presume the distribution of the data. Whereas, other algorithms have gaussian and normal distributions.

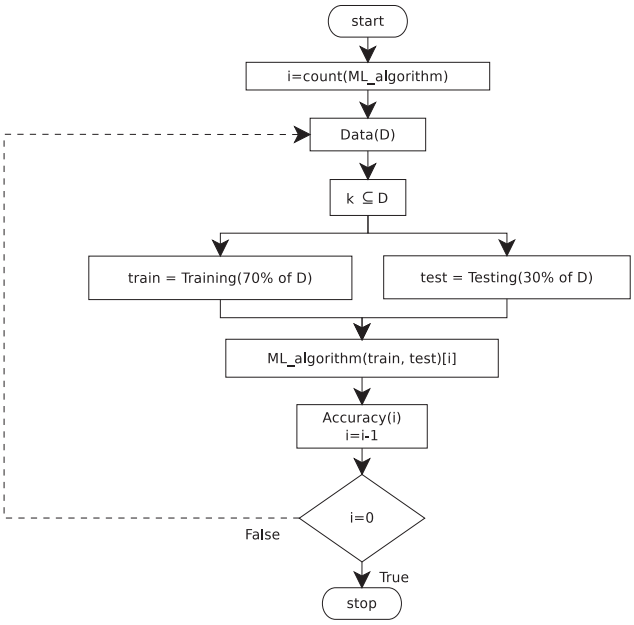

Fig. 3. Flow chart represents the working mechanism to obtain accuracy of i number of machine learning algorithms

Fig. 4 reveals the method to calculate K clusters in K means algorithm, where the sharp slope in the graph obtained from the within the sum of squares(wss) with respect to increasing number of clusters. Wss is obtained from the data points close to centroid of each cluster. The performance of K means is shown in Fig. 5, where *total_ss* is a feature that encompasses complete data points with a initial centroid, *within_ss* features a average data points close to each cluster's centroid. *between_ss* gives the data points between *total_ss* and *within_ss*.

Figures Fig. 6 to Fig. 12 shows the performance evaluation of algorithms from there respective confusion matrix.

Naive bayes, Decision tree, and Random forest excels in performance because of the following reasons, i) the classification of data is categorized in such a way that all new occurrences of data fall under any one of the classes. ii) all data is considered important and the algorithms would not prune least occurrences. iii) data considered is in huge volume.

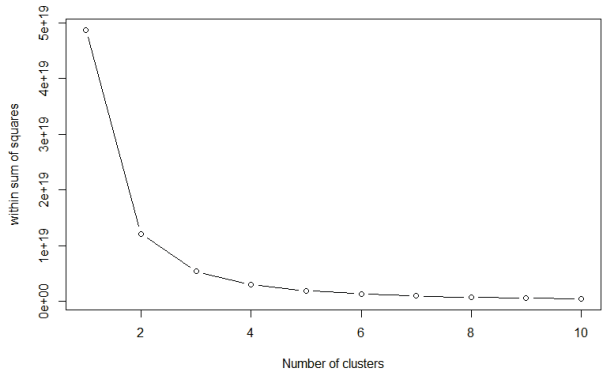

Fig. 4. Elbow graph to calculate k clusters value in k means, x-axis raising with number of clusters and y-axis elevating within sum of squares(wss); data points within a common cluster center. The start to have diminishing value of wss returns by increasing number of clusters represents elbow

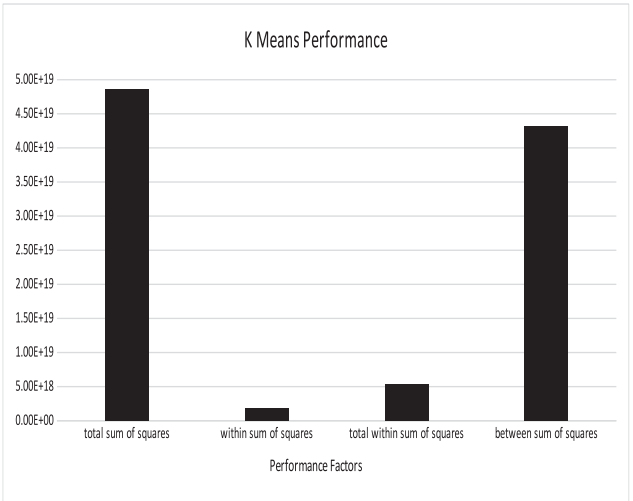

Fig. 5. Manifestation of K Means performance

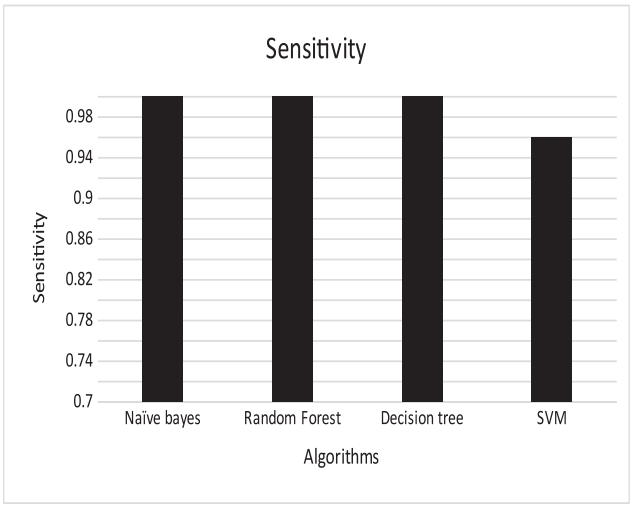

Fig. 6. Illustrates the sensitivity that predicts the actual critical occurrences

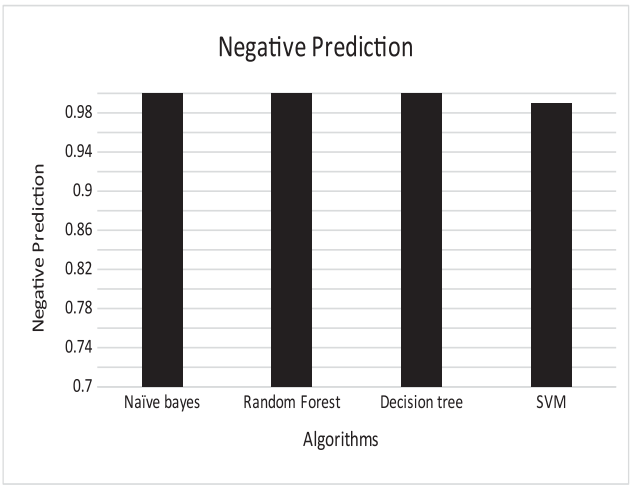

Fig. 9. Portrays the negative predictions that exhibits the correct negative prediction

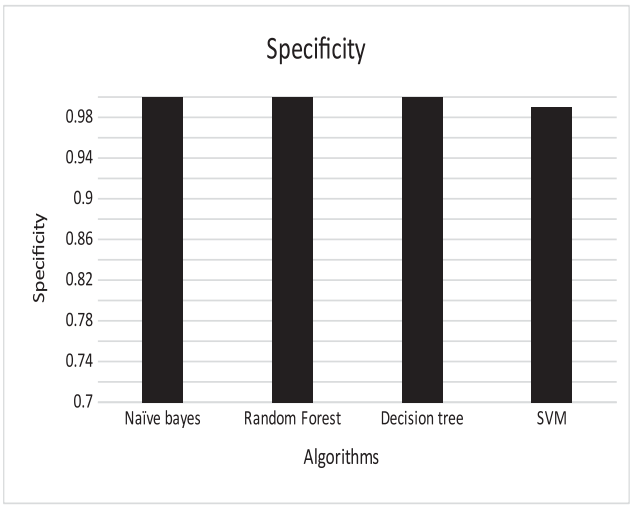

Fig. 7. Shows the specificity in predicting non critical data

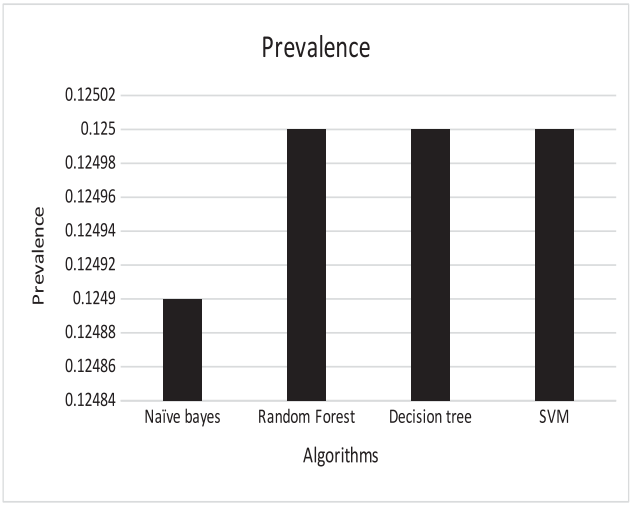

Fig. 10. Interprets the prevalence; the frequency of true critical data occurrences.

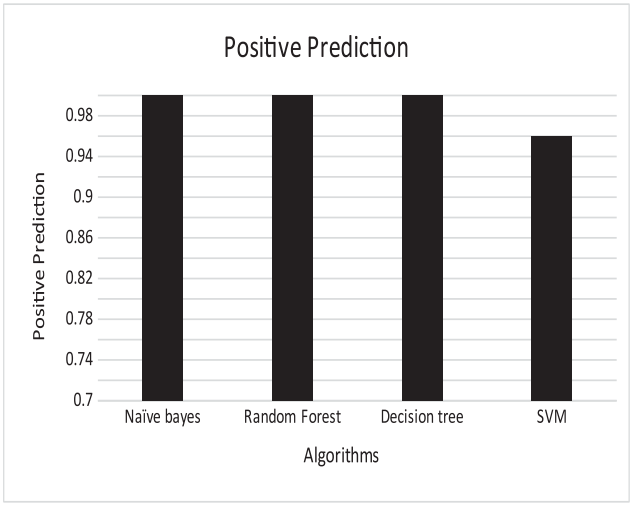

Fig. 8. Represents the positive predictions or precision that indicates the rate of correct prediction in the given data

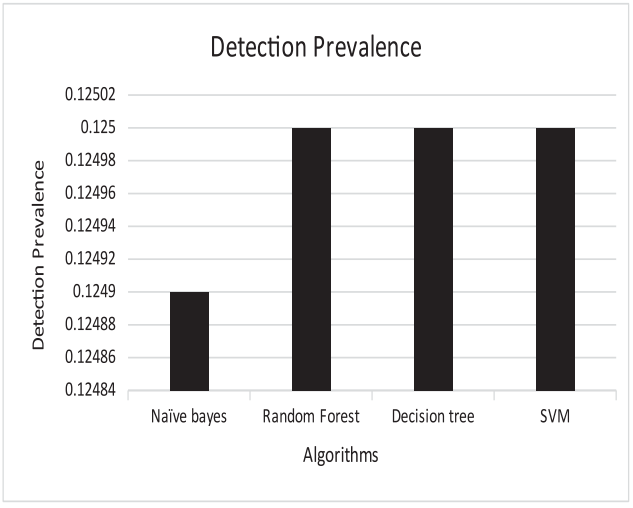

Fig. 11. Depicts the true critical occurrences on overall data

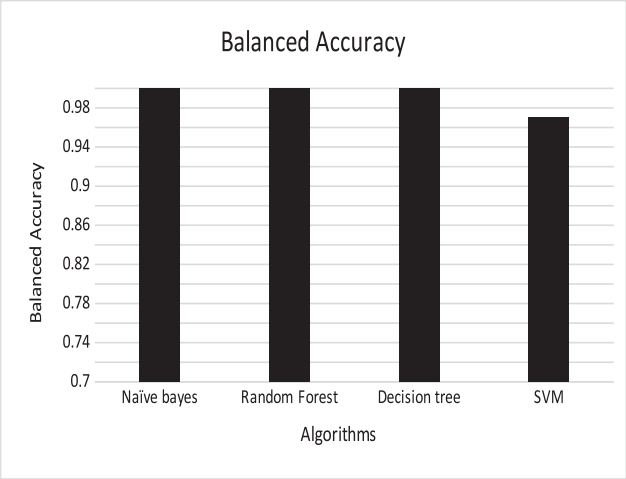

Fig. 12. Interprets balanced accuracy of all the algorithms

VI. Comparative Analysis

The experimental analysis on the given healthcare data indicate that Naive bayes, Decision tree, and Random forest methods are the best algorithms with maximum accuracy of 1, in comparison to Support vector machine and K means clustering that produced relatively less accuracy's, respectively.

Table IV illustrates the classification accuracy of Naive bayes, Decision tree, regression accuracy of Random Forest, Support vector machine and clustering accuracy of K means.

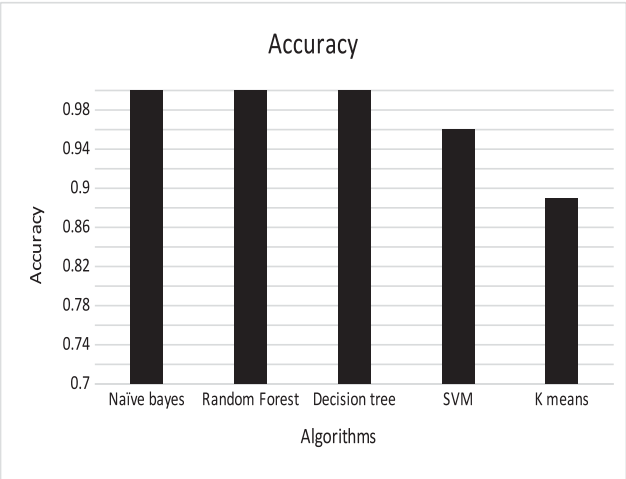

Fig. 13. Shows accuracy comparison between applied machine learning algorithms. Vertical axis shows increase in accuracy values and horizontal axis lists the algorithms.

TABLE IV. Accuracy in various machine learning algorithms

Algorithm	Accuracy
Naive bayes	1.00
Decision tree	1.00
K means	0.89
Random forest	1.00
Support vector machine	0.96

VII. Conclusion

Determining the critical condition of a person from gathered heterogeneous SIoT data based on the weitage given to each SIoT device. The comparison of Naive bayes, Decision tree, K means, Random forest, and Support vector machine algorithms are analyzed. The experimental analysis shows that Naive bayes, Decision tree and Random forest produce same maximum accuracy. Although these three algorithms results accuracy of 1, Naive bayes is chosen over Decision tree and Random forest. This is because Decision tree will prune small occurrences of data where each data in healthcare monitoring and tracking systems is considered important. Further, it is hard to implement pruning procedures for healthcare data. Whereas, Naive bayes have a parallel map reduce implementation for devices that record huge data in small intervals of time. Therefore, Naive bayes provide a better accuracy considering all occurrences of data with both time and space complexities for very huge data.

As healthcare technologies are getting portable options such as ECG devices, BP devices, smartphones, smart watches etc., more research and development is required to monitor and track the person's health in real time and take necessary actions like, to inform guardian, nearest healthcare unit as the condition of the person's health get critical.

References

[1] Arjun Chandra, Xin Yao, "Evolving hybrid ensembles of learning machines for better generalization", *Neurocomputing.*, vol. 69, 2006, pp. 686-700.

[2] Atzori, Luigi and Iera, Antonio and Morabito, Giacomo, "Making things socialize in the Internet — Does it help our lives?", *Proceedings of ITU Kaleidoscope*, 2011, pp. 1-8.

[3] Wenting Tu and Shiliang Sun, "Cross-domain representation-learning framework with combination of class-separate and domain-merge objectives", *(CDKD '12) ACM.*, 2012, pp. 18-25.

[4] Atzori, Luigi and Iera, Antonio and Morabito, Giacomo and Nitti, Michele, "Internet of Things (IoT): A vision, architectural elements, and future directions", *Computer Networks.*, vol. 56, 2012.

[5] Jayavardhana Gubbi, Rajkumar Buyya, Slaven Marusic, Marimuthu Palaniswami, "Cross-domain representation-learning framework with combination of class-separate and domain-merge objectives", *Future Generation Computer Systems.*, vol. 29, 2013, pp. 1645-1660.

[6] Y. Bengio, A. Courville and P. Vincent, "Representation Learning: A Review and New Perspectives", *IEEE Transactions on Pattern Analysis and Machine Intelligence.*, vol. 35, 2013, pp. 1798-1828.

[7] Yoshua Bengio, Grégoire Mesnil, Yann Dauphin, and Salah Rifai, "Better mixing via deep representations.", *ICML'13.*, vol. 28, 2013, pp. 552-560.

[8] H. Z. Asl, A. Iera, L. Atzori and G. Morabito, "How often social objects meet each other? Analysis of the properties of a social network of IoT devices based on real data", *IEEE Global Communications Conference (GLOBECOM)*, 2013, pp. 2804-2809.

[9] L. Atzori, A. Iera and G. Morabito, "From "smart objects" to "social objects": The next evolutionary step of the internet of things", *IEEE Communications Magazine*, vol. 52, no. 1, 2014, pp. 97-105. '

[10] C. Perera, A. Zaslavsky, P. Christen and D. Georgakopoulos, "Context Aware Computing for The Internet of Things: A Survey", *IEEE Communications Surveys and Tutorials.*, vol. 16, 2014, pp. 414-454.

[11] C. Tsai, C. Lai, M. Chiang and L. T. Yang, "Data Mining for Internet of Things: A Survey", *IEEE Communications Surveys and Tutorials.*, vol. 16, 2014, pp. 77-97.

[12] Islam, S.M. Riazul & Kwak, Daehan & Kabir, Md. Humaun & Hossain, Mahmud & Kwak, Kyung, "The Internet of Things for Health Care: A Comprehensive Survey", *IEEE Access.*, vol. 3, 2015, pp. 678-708.

[13] J. L. Berral-García, "A quick view on current techniques and machine learning algorithms for big data analytics", *18th (ICTON).*, 2016, pp. 1-4.

[14] Qiu J, Wu Q, Ding G, "A survey of machine learning for big data processing", *EURASIP Journal on Advances in Signal Processing.*, vol. 2016, 2016, pp. 67.

[15] Gil, David Ferrández, Antonio Mora-Mora, Higinio Peral, Jesús Journal, "Internet of Things: A Review of Surveys Based on Context Aware Intelligent Services", *Sensors.*, vol. 16, 2016, pp. 1069.

[16] Badraddin Alturki, Stephan Reiff-Marganiec, and Charith Perera, "A hybrid approach for data analytics for internet of things", *(IoT'17), ACM.*, 2017, pp. 71-78.

[17] S. Athmaja, M. Hanumanthappa and V. Kavitha, "A survey of machine learning algorithms for big data analytics", *ICIIECS.*, 2017, pp. 1-4.

[18] M. Lippi, M. Mamei, S. Mariani and F. Zambonelli, "An Argumentation-Based Perspective Over the Social IoT", *IEEE Internet of Things Journal.*, vol. 5, 2018, pp. 2537-2547.

[19] C. Perera, A. Zaslavsky, P. Christen and D. Georgakopoulos, "Context Aware Computing for The Internet of Things: A Survey", *IEEE Communications Surveys and Tutorials.*, vol. 16, 2014, pp. 414-454.

[20] Raouf Boutaba, Mohammad A. Salahuddin, Noura Limam, Sara Ayoubi, Nashid Shahriar, Felipe Estrada-Solano, Oscar M. Caicedo, "A comprehensive survey on machine learning for networking: evolution, applications and research opportunities", *Springer.*, vol. 9, 2018.

[21] Kumar Donta, Praveen & Tarachand, Amgoth & Sekhar, Chandra, "Machine learning algorithms for wireless sensor networks: A survey", *Information Fusion.*, vol. 49, 2018, pp. 1-25.

[22] M. G. Kibria, K. Nguyen, G. P. Villardi, O. Zhao, K. Ishizu and F. Kojima, "Big Data Analytics, Machine Learning, and Artificial Intelligence in Next-Generation Wireless Networks", *IEEE Access.*, vol. 6, 2018, pp. 32328-32338.

[23] Sharma Karan, Castellini Claudio, van den Broek Egon L, Albu-Schäeffer Alin, Schwenker Friedhelm, *A dataset of continuous affect annotations and physiological signals for emotion analysis*. 2018.