IJIRSET
International Journal of Innovative Research in
SCIENCE | ENGINEERING | TECHNOLOGY

e-ISSN: 2319-8753 | p-ISSN: 2347-6710

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

Volume 13, Issue 1, January 2024

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.423

9940 572 462 | 6381 907 438 | ijirset@gmail.com | www.ijirset.com

International Journal of Innovative Research in Science, Engineering and Technology (IJIRSET)

|e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 13, Issue 1, January 2024 ||

| DOI:10.15680/IJIRSET.2024.1301020 |

Optimization Accuracy of Diagnosis Disease Prediction for Smart Healthcare using Machine Learning Algorithm

[1]Shambhu Kumar, [2]Prof. Santosh Nagar, [3]Prof. Anurag Shrivastava

M. Tech. Scholar, Department of Computer Science and Engineering, NIIST, Bhopal, India[1]

Assistant Professor, Department of Computer Science and Engineering, NIIST, Bhopal, India [2]

Head of Dept., Department of Computer Science and Engineering, NIIST, Bhopal, India [3]

ABSTRACT: - Data has become an integral part of the digital world with the advancement in computing technologies. The collection of data is very crucial with regards to data analytics. Every industry makes use of data analytics ranging from financial to other commercial applications but it becomes even more important in healthcare domain for the analysis of healthcare data. The present research work is mainly focused on classification/prediction problems of healthcare data based on machine learning (supervised) approaches using data mining techniques. Machine learning provides a reliable and excellent support for prediction of a Diabetes and Heart disease with correct case of training and testing. In this paper studied of Diagnosis of diabetes and heart disease desires great support of different classifiers algorithm to detect diabetes and heart disease in early stage, since it cannot be cured which brings great complication to our health system.

KEYWORDS: -Machine Learning, Diabetes Disease, Heart Disease

I. INTRODUCTION

Many renovate studies are taking place in the digital world with the advancement in computing technologies. Data has become an integral part of this digital world. The collection of data is very crucial with regards to data analytics. The study of data with an intension to derive a conclusion (valuable insights), based on information is known as data analytics. Every industry makes use of data analytics. In comparison to the financial and other commercial businesses, data analytics is more important in the healthcare industry [2, 3]. Hence, effective utilization of data analytics in healthcare domain can improve the quality of healthcare and, more importantly, facilitate preventive measures. Notably, a recent report on big data suggests that the entire potential of healthcare data is estimated to be approximately $300 billion [4]. In addition, Sun et al. [5], in 2012 have discussed that the computerized version of medical data from all across the globe was assessed at 500 petabytes, which could reach at 25,000 petabytes by 2020. The main role of data analysts in healthcare is to understand and develop knowledge from healthcare data in order to extract useful and noteworthy (actionable) information from it. In recent times, several data mining applications are being developed for healthcare sectors to support decision making [3, 6]. As reported in [1], the author discussed the analysis, visualization and mining of healthcare data and determines the way in which data can be efficiently managed which could further enhance the organization ability to produce revenue, control risk, and cost. Raja et al. [2] have described and summarized how medical services are changing through clinical innovation, electronic health records (EHRs), medical devices, and other factors related to healthcare. In particular, healthcare data analytics has proven to play a vital role for handling various issues in healthcare disciplines [1, 2].

II. MACHINE LEARNING ALGORITHM

Uproarious information is available in the heap of substance that will be identified through the anomaly strategies. The information can be spatial or can be a transient method spatial connected with the geological conditions and worldly connected with the time perspectives [14, 15]. The principle point of exception identification is to deal with the loud

|e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 13, Issue 1, January 2024 ||

| DOI:10.15680/IJIRSET.2024.1301020 |

information that is introduced in the heap of text. Different methods for recognizing abnormalities in Text are specified in below:

Learning

The main property of an ML is its capability to learn. Learning or preparing is a procedure by methods for which a neural system adjusts to a boost by making legitimate parameter modifications, bringing about the generation of wanted reaction. Learning in an ML is chiefly ordered into two classes as [16].

- Supervised learning
- Unsupervised learning

Supervised Learning

Regulated learning is two stage forms, in the initial step: a model is fabricated depicting a foreordained arrangement of information classes or ideas. The model developed by investigating database tuples portrayed by traits. Each tuple is expected to have a place with a predefined class, as dictated by one of the qualities, called to have a place with a reclassified class, as controlled by one of the traits called the class name characteristic. The information tuple are dissected to fabricate the model all things considered from the preparation dataset [17].

Unsupervised learning

It is the kind of learning in which the class mark of each preparation test isn't knows, and the number or set of classes to be scholarly may not be known ahead of time. The prerequisite for having a named reaction variable in preparing information from the administered learning system may not be fulfilled in a few circumstances.

Data mining field is a highly efficient techniques like association rule learning. Data mining performs the interesting machine-learning algorithms like inductive-rule learning with the construction of decision trees to development of large databases process. Data mining techniques are employed in large interesting organizations and data investigations. Many data mining approaches use classification related methods for identification of useful information from continuous data streams.

Nearest Neighbors Algorithm

The Nearest Neighbor (NN) rule differentiates the classification of unknown data point because of closest neighbor whose class is known. The nearest neighbor is calculated based on estimation of k that represents how many nearest neighbors are taken to characterize the data point class. It utilizes more than one closest neighbor to find out the class where the given data point belong termed as KNN. The data samples are required in memory at run time called as memory-based technique. The training points are allocated weights based on their distances from the sample data point. However, the computational complexity and memory requirements remained key issue. For addressing the memory utilization problem, size of data gets minimized. The repeated patterns without additional data are removed from the training data set [18].

Naive Bayes Classifier

Naive Bayes Classifier technique is functioned based on Bayesian theorem. The designed technique is used when dimensionality of input is high. Bayesian Classifier is used for computing the possible output depending on the input. It is feasible to add new raw data at runtime. A Naive Bayes classifier represents presence (or absence) of a feature (attribute) of class that is unrelated to presence (or absence) of any other feature when class variable is known. Naïve Bayesian Classification Algorithm was introduced by Shinde S.B and Amrit Priyadarshi (2015) that denotes statistical method and supervised learning method for classification. Naive Bayesian Algorithm is used to predict the heart disease. Raw hospital dataset is employed. After that, the data gets preprocessed and transformed. Finally by using the designed data mining algorithm, heart disease was predicted and accuracy was computed.

Support Vector Machine

SVM are used in many applications like medical, military for classification purpose. SVM are employed for classification, regression or ranking function. SVM depends on statistical learning theory and structural risk minimization principal. SVM determines the location of decision boundaries called hyper plane for optimal separation of classes as described in figure 3. Margin maximization through creating largest distance between separating hyper plane and instances on either side are employed to minimize upper bound on expected generalization error. Classification accuracy of SVM not depends on dimension of classified entities. The data analysis in SVM is based on

|e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 13, Issue 1, January 2024 ||

| DOI:10.15680/IJIRSET.2024.1301020 |

convex quadratic programming. It is expensive as quadratic programming methods need large matrix operations and time consuming numerical computations.

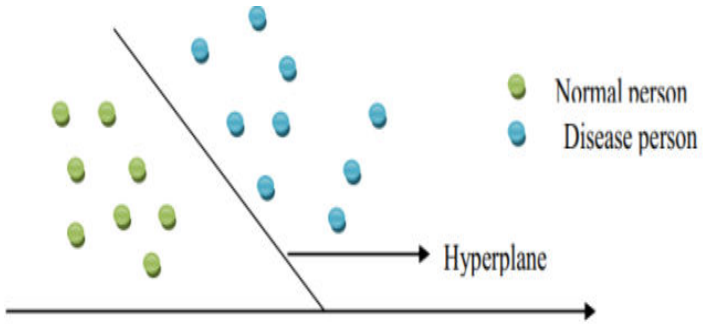


Fig. 1: Support Vector Classification

III. PROPOSED METHODOLOGY

In today's digital environment, healthcare has turned into a data-driven industry. A common use of data in healthcare is to develop an electronic Medical Disease Diagnosis System (MDDS)- based on clinical data, domain knowledge & machine learning in order to assist doctors and medical professionals to detect any critical situation or errors during the treatment of patients [2, 3, 6, 7] as well as to improve healthcare decision. For this purpose, collection of data is very crucial with regards to healthcare data analytics.

- Descriptive Analytics Analysis of healthcare data using descriptive analytics enables medical specialists to know what is going on in a specific situation.
- Predictive Analytics Various statistical tools and machine learning approaches are used in predictive analytics to assess and derive important information from healthcare data in order to predict the future outcomes.
- Prescriptive Analytics Prescriptive analytics is comparatively a new analytical process that incorporates both descriptive and predictive analytics.

Healthcare professionals can use prescriptive analytics to recommend the right treatments for the patients. There is huge wealth of healthcare data available in the healthcare industries. It would be almost impossible to process such large data for human minds. Hence machine learning techniques are required to solve such complex information efficiently, effectively and quickly. For analysis of health-related data, several standard machine learning algorithms are being used with data mining methods so as to better anticipate the outcomes of healthcare data. Supervised, unsupervised and reinforcement are the three main types of machine learning techniques that are primarily utilized for health data analysis. Most of the healthcare applications are based on supervised learning techniques, in which data are classified into predetermined categories. The current research focuses on using machine learning methodologies to address the challenges of classification problems in healthcare domain.
Arthur Samuel coined the term 'Machine Learning' in 1959 to play strategic games like chess. Machine Learning can be defined as a mechanism that enables computer to selflearn without having to be explicitly programmed. The main aim of machine learning is to develop a program that can access and use data for learning purposes. It is primarily based on statistics and probability. However, it is more effective than traditional statistical approaches while making computational decision. In fact, machine learning is also known as the kernel to the success of AI. It is widely applicable in various areas such as medical science, engineering, and society. Machine learning relies on data. Relevant data will lead to more accurate clinical decisions in healthcare [17]. It is the foundation of any model. The algorithms of machine learning are used to mine actionable insights from healthcare data in order to make excellent clinical diagnostic decision. In this regard, till date several standard learning methods like Naïve Bayes Support Vector Machine, K-Nearest Neighbour, Neural Network, Decision Trees etc. have been developed.

IV. SIMULATION PARAMETER

Dataset was separated into two datasets (70%/30%, preparing/testing) to keep away from any predisposition in preparing and testing. Of the information, 70% was utilized to prepare the ML model, and the excess 30% was utilized

|e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 13, Issue 1, January 2024 ||

| DOI:10.15680/IJIRSET.2024.1301020 |

for testing the presentation of the proposed movement arrangement framework. The articulations to compute accuracy and review are given in Equations (2) and (3).

Accuracy gives a proportion of how precise your model is in anticipating the real up-sides out of the absolute up-sides anticipated by your framework. Review gives the quantity of real up-sides caught by our model by grouping these as obvious positive. F-measure can give a harmony among accuracy and review, and it is liked over precision where information is uneven.

Accordingly, F-measure was used in this review as a presentation metric to give a decent and fair measure utilizing the equation.

$$\text{Precision} = TP/(TP + FP) \times 100$$
$$\text{Recall} = TP/(TP + FN) \times 100$$

$$F - \text{measure} = 2 * \text{Precision} * \text{Recall}/(\text{Precision} + \text{Recall}) \times 100$$

Where,

TP—True Positive, FP—False Positive, FN—False Negative

V. SIMULATION RESULTS

Results as follow

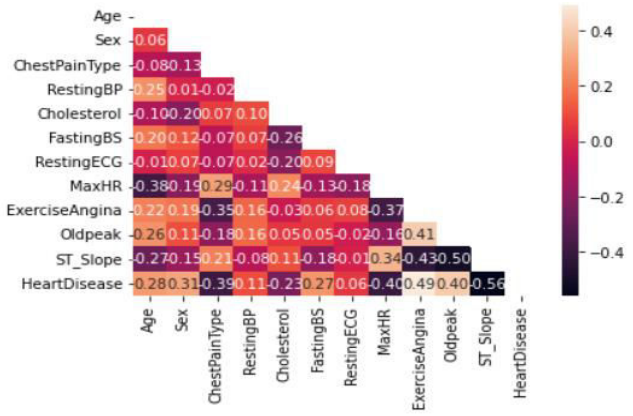
1. Load data and convert it into a pandas data frame

	Age	Sex	ChestPainType	RestingBP	Cholesterol	FastingBS	RestingECG	MaxHR	ExerciseAngina	Oldpeak	ST_Slope	HeartDisease
0	40	M	ATA	140	289	0	Normal	172	N	0.0	Up	0
1	49	F	NAP	160	180	0	Normal	156	N	1.0	Flat	1
2	37	M	ATA	130	283	0	ST	98	N	0.0	Up	0
3	48	F	ASY	138	214	0	Normal	108	Y	1.5	Flat	1
4	54	M	NAP	150	195	0	Normal	122	N	0.0	Up	0

2. Label encoding for changing string value to Numeric

	Age	Sex	ChestPainType	RestingBP	Cholesterol	FastingBS	RestingECG	MaxHR	ExerciseAngina	Oldpeak	ST_Slope	HeartDisease
0	40	M	ATA	140	289	0	Normal	172	N	0.0	Up	0
1	49	F	NAP	160	180	0	Normal	156	N	1.0	Flat	1

3. Find Correlation among data

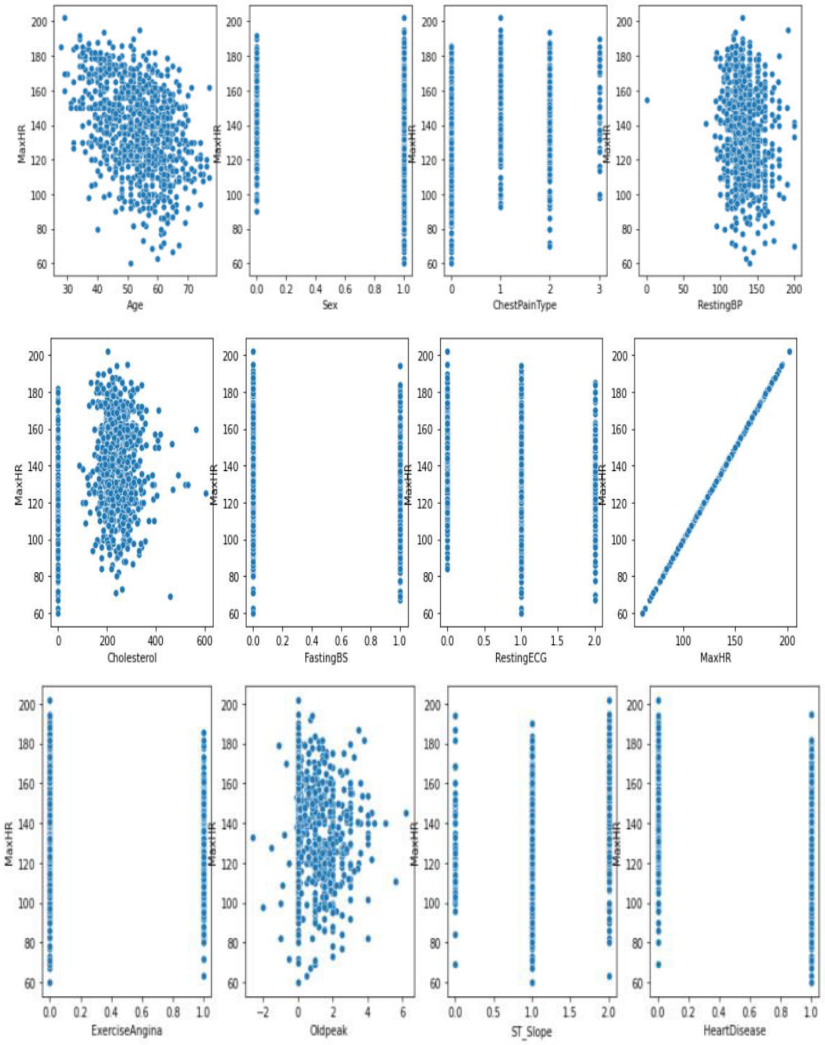


|e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

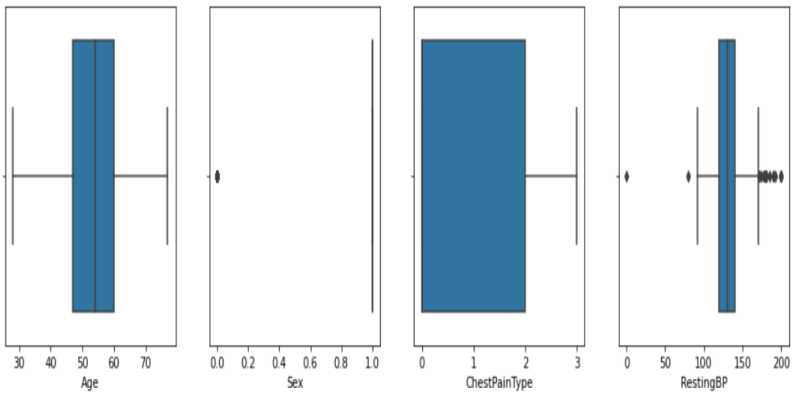
|| Volume 13, Issue 1, January 2024 ||

| DOI:10.15680/IJIRSET.2024.1301020 |

4. Visualization



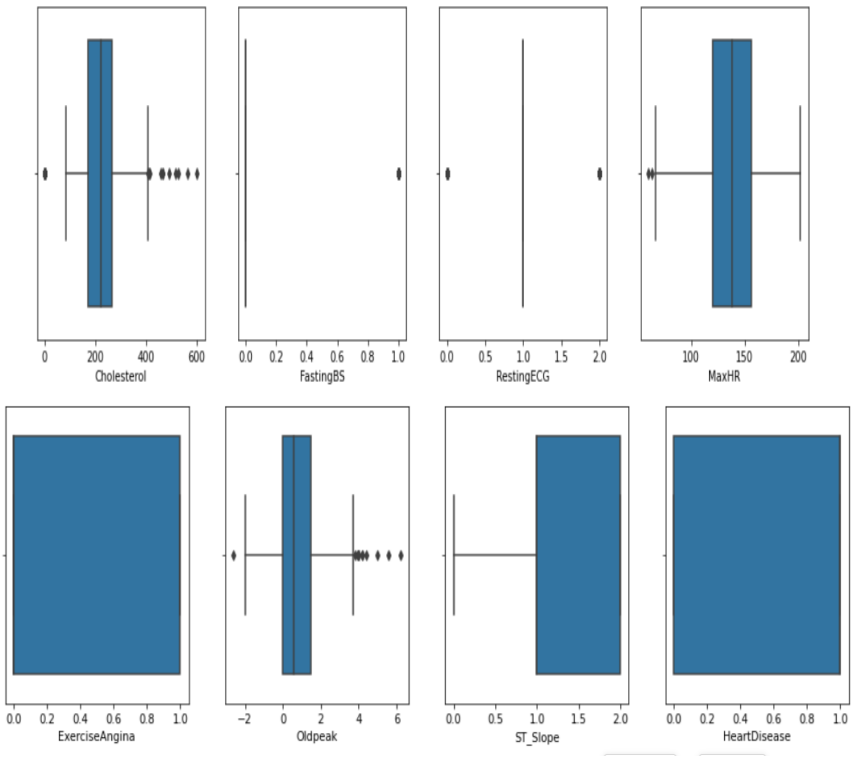
5. Outlier detection in data



|e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 13, Issue 1, January 2024 ||

| DOI:10.15680/IJIRSET.2024.1301020 |



6. Remove Outliers
7. Feature selection and data splitting into train and test data in the ratio of 80:20
8. Choosing model as random forest for primary data matrices check
9. Perform standard scaling
10. Use synthetic minority oversampling technique (SMOTE) for oversampling to solve imbalance data
11. Perform cross-validation Hyperparameters

N_estimators	9, 16, 25, 36, 49, 64, 81
n_jobs	-1
random_state	42, none

Training SVM, Decision Tree, and Gradient Boost classifiers on Heart dataset

Algorithms	**Accuracy**	**Precision**	**Recall**	**F score**
Romany et al. [1]	88.80	92.04	87.04	-
SVM	90.6	93.6	88.90	89.58
Decision Tree	92	94.86	91.42	91.22
Gradient Boost	94.6	96.78	93.52	93.42

Training SVM, Decision Tree, and Gradient Boost classifier on Diabetic dataset

Algorithms	**Accuracy**	**Precision**	**Recall**	**F score**
Romany et al. [1]	89.00	92.10	87.70	-
SVM	91.12	93.02	88.90	88.94
Decision Tree	91.88	93.98	89.12	90.74
Gradient Boost	92.94	94.64	91.78	92.34

|e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 13, Issue 1, January 2024 ||

| DOI:10.15680/IJIRSET.2024.1301020 |

VI. CONCLUSION

In recent time, cardiovascular disorders or cardiovascular diseases (CVDs) are considered one of the leading causes of death among men and women across the world. CVD is very dangerous (deadly) diseases that affect blood vessels and heart. According to the recent survey by WHO, over 17 million deaths have been estimated annually due to cardiovascular diseases in the past few years, which represents about 31% of all deaths globally, including USA and out of these, 85% are caused by heart attacks and strokes. This number is further expected to be increased to 23.6 million by 2030. CVDs are the major global health problem in the modern medicine. Nowadays, many researchers and clinicians have been using several data mining techniques by employing machine learning approaches in the diagnosis of cardiovascular disorder.

REFERENCES

[1] Romany Fouad Mansour, Adnen El Amraoui, Issam Nouaouri, Vicente García Díaz , Deepak Gupta, and Sachin Kumar, “Artificial Intelligence and Internet of Things Enabled Disease Diagnosis Model for Smart Healthcare Systems”, IEEE Access 2021.

[2] G. Muhammad, M. S. Hossain, and N. Kumar, ``EEG-based pathology detection for home health monitoring," IEEE J. Sel. Areas Commun., vol. 39, no. 2, pp. 603610, Feb. 2021,

[3] A. A. Mutlag, M. K. A. Ghani, M. A. Mohammed, M. S. Maashi, O. Mohd, S. A. Mostafa, K. H. Abdulkareem, G. Marques, and I. de la Torre Díez, ``MAFC: Multi-agent fog computing model for healthcare critical tasks management," *Sensors*, vol. 20, no. 7, p. 1853, Mar. 2020.

[4] M. S. Hossain and G. Muhammad, ``Deep learning based pathology detection for smart connected healthcare," *IEEE Netw.*, vol. 34, no. 6, pp. 120125, Nov. 2020.

[5] M. S. Hossain G. Muhammad and A. Alamri "Smart Healthcare Monitoring: A Voice Pathology Detection Paradigm for Smart Cities" Multimedia Systems vol. 25 no. 5 pp. 565-75 Oct. 2019

[6] V Krishnapraseeda, M S Geetha Devasena, V Venkatesh and A Kousalya, “Predictive Analytics on Diabetes Data using Machine Learning Techniques”, 7th International Conference on Advanced Computing and Communication Systems (ICACCS), pp. 458-463, IEEE 2021

[7] Valasapalli Mounika, Devi Sree Neeli, Gorla Suma Sree, Parimi Mourya and Modala Aravind Babu, “Prediction of Type-2 Diabetes using Machine Learning Algorithms”, International Conference on Artificial Intelligence and Smart Systems (ICAIS), pp. 167-173, IEEE 2021

[8] I.K. Mujawar, B.T. Jadhav, V.B. Waghmare and R.Y. Patil, “Development of Diabetes Diagnosis System with Artificial Neural Network and Open Source Environment”, International Conference on Emerging Smart Computing and Informatics (ESCI), pp. 778-784, IEEE 2021

[9] Cecilia Saint-Pierre;Florencia Prieto;Valeria Herskovic;Marcos Sepúlveda, “Team Collaboration Networks and Multidisciplinarity in Diabetes Care: Implications for Patient Outcomes”, IEEE Journal of Biomedical and Health Informatics, Vol. 14(1), pp. 319-329, 2020.

[10] M. S. Hossain and G. Muhammad, ``Emotion-aware connected healthcare big data towards 5G," IEEE Internet Things J., vol. 5, no. 4, pp. 23992406, Aug. 2018.

[11] M. Pham, Y. Mengistu, H. Do, andW. Sheng, ``Delivering home healthcare through a cloud-based smart home environment (CoSHE)," Future Gener. Comput. Syst., vol. 81, pp. 129140, Apr. 2018.

[12] A. Kaur and A. Jasuja, ``Health monitoring based on IoT using raspberry PI," in Proc. Int. Conf. Comput., Commun. Autom. (ICCCA), Greater Noida, India, May 2017, pp. 13351340.

[13] U. Satija, B. Ramkumar, and M. Sabarimalai Manikandan, ``Realtime signal quality-aware ECG telemetry system for IoT-based health care monitoring," IEEE Internet Things J., vol. 4, no. 3, pp. 815823, Jun. 2017.

[14] O. S. Alwan and K. Prahald Rao, ``Dedicated real-time monitoring system for health care using ZigBee," Healthcare Technol. Lett., vol. 4, no. 4, pp. 142144, Aug. 2017.

[15] P. Kakria N.K. Tripathi and P. Kitipawang "A Real-Time Health Monitoring System for Remote Cardiac Patients Using Smartphone and Wearable Sensors" t'l. J. Telemedicine and Applications vol. 2015 2015.

|e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 13, Issue 1, January 2024 ||

| DOI:10.15680/IJIRSET.2024.1301020 |

[16] G. Villarrubia, J. Bajo, J. De Paz, and J. Corchado, ``Monitoring and detection platform to prevent anomalous situations in home care," Sensors, vol. 14, no. 6, pp. 99009921, Jun. 2014.

[17] M.A. Alsheikh et al. "Machine Learning in Wireless Sensor Networks: Algorithms Strategies and Applications" IEEE Commun. Surveys Tutorials vol. 16 no. 4 pp. 1996-2018 Fourth Quarter 2014.

[18] F. F. Gong, X. Z. Sun, J. Lin, and X. D. Gu, ``Primary exploration in establishment of China's intelligent medical treatment," (in Chinese), *Mod. Hos Manag.*, vol. 11, no. 2, pp. 28_29, 2013.

[19] A. Krizhevsky I. Sutskever and G. E. Hinton "ImageNet Classification with Deep Convolutional Neural Networks" Proc. 25th Int'l. Conf. Neural Info. Processing Systems pp. 1097-1105 2012.

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi cross ref

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

www.ijirset.com

Scan to save the contact details