European Journal of Molecular & Clinical Medicine
ISSN 2515-8260 Volume 7, Issue 4, 2020

Diabetes Data Prediction in healthcare Using Hadoop over Big Data

Gajanand Sharma[1], *Ashutosh Kumar[2], Himanshu Sharma[3], Ashok Kumar Saini[4], Priyanka[5], S.R. Dogiwal[6]

[1,2,3,4,5] *Department of Computer Science & Engineering*
[6]*Department of Information Technlogy*
[1,2,3,4] *JECRC University, Jaipur, India*
[5,6] *Swami Keshvanand Institute of Technology, Management & Gramothan, Jaipur, India*
[1]*gajanan.sharma@gmail.com*, *[*2]*ashucse007@gmail.com*, [3]*himanshu.manu.sharma@gmail.com*, [4]*shksaini@gmail*, [5]*trikhapriyanka@gmail.com*, [6]*dogiwal@gmail.com*

Abstract: Diabetes mellitus is one of the major non-communicable diseases which have great impact on human life today. A huge amount of data is generated including a wide variety of the Electronic Medical Record (EMR), pharmacy reports, and laboratory reports, among other data related to patients. Big data analytics can be applied to this data to generate useful patterns and relation between different factors which affects diabetes. The results obtained from this analysis shows relation between different attributes which can be used to improve healthcare system. In this paper the analysis of the diabetes dataset is done using Hadoop framework, which is a distributive framework and can be used to analysis large amount of data. The dataset is taken from PIMA Indian Database, which includes different factors that affect diabetes like age, blood pressure, BMI (Body-Mass Index), skin thickness etc. Results produced by the analysis of data are projects on Power BI.

Keyword: Diabetes, Body-Mass Index, Electronic Medical Record, Hadoop, Prediction

1. Introduction

Countless individuals are experiencing various illnesses today. An enormous measure of diabetes has been produced by human services industry. The medicinal services information takes Electronic Health Reports (EHR) of patient's information, clinical reports, specialist's remedy, analytic reports, therapeutic pictures, drug store data and medical coverage related information. Diabetic Mellitus (DM) is one of the Non-Communicable Diseases (NCD) creating in various nations. Diabetes Mellitus or basically called Diabetes is the affliction wherein the human body doesn't produce appropriate measure of insulin. There are number of variables which influence diabetes which incorporates age, circulatory strain, insulin, skin thickness and so on. A lot of diabetes information is now accessible which can be broke down to discover the relations between various components to improve the social insurance framework and to give better treatment office.

1.1 Big Data

Big Data is a term which is utilized for tremendous measure of information. This data establishes both organized and unstructured data that is developing at a quick rate step by step. Associations are confronting difficulties to oversee and investigation this data to create some important outcomes, on the grounds that conventional database frameworks can't oversee huge data collections.

1.2 Big Data Characteristics (5 V's of Big Data)

Characteristics of Big data defined by these 5 V's of Big Data –

- Volume - Volume is the amount of data generated
- Velocity – Velocity describes the speed at which the data is generated.
- Variety – Variety defines the type of data as data can be categorized in structured, semi-structured or unstructured.
- Veracity - Veracity is the trustworthiness of the data. It shows accuracy of the data.
- Value – Value means how much meaningful or useful data we can extract.

1.3 Hadoop Distributed File System (HDFS)

Hapdoop Distributed File System is the most significant piece of Hadoop's Ecosystem. HDFS is the essential stockpiling arrangement of Hadoop. Hadoop circulated record framework (HDFS) is a java based document framework that gives adaptable, adaptation to non-critical failure, dependable and cost effective information stockpiling for Big data. (HDFS) is a dispersed record framework that utilizations and runs on ware equipment.

1.4 HDFS Nodes

HDFS consists of two nodes –

- Name Node: - It is called as Master Node. Real document information isn't put away by Name Node. It is capable of putting away the metadata for example number of squares made, area of the square, subtleties of the Data node on which the information is put away and other data of records.
- Data Node - It is otherwise called Slave Node. HDFS Data node stores the real document information as squares in HDFS. Data node performs information peruse and compose tasks. Data nodes additionally perform diverse square creation, cancellation and replication undertakings on the guidance of name node.

1.5 HDFS Architecture

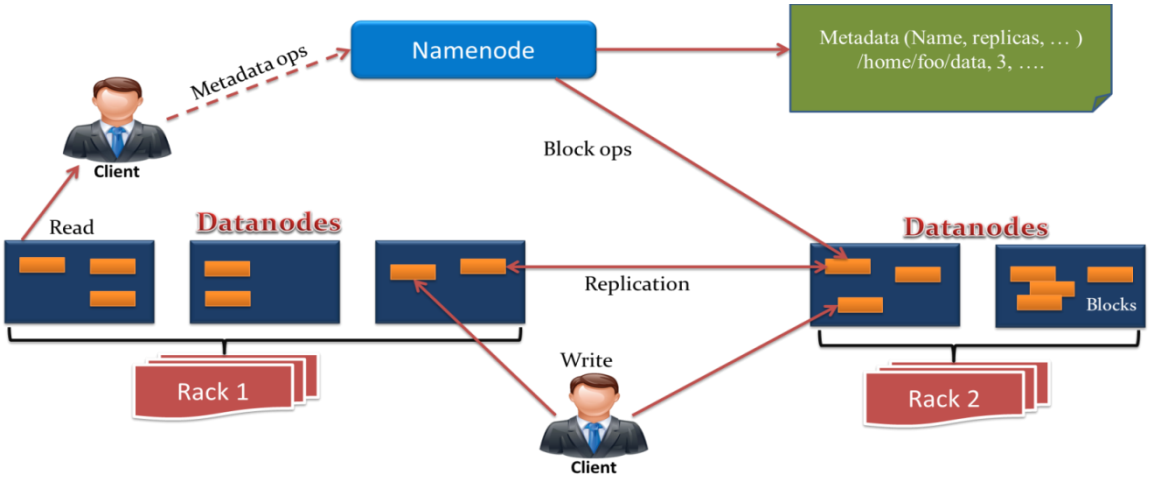

Figure 1. HDFS Architecture

1.6 MapReduce

Hadoop MapReduce is the handling segment of Hadoop biological system, which gives parallel preparing of information. Every one of the information which is put away in the Hadoop Distributed File framework

is prepared by MapReduce programs. Because of this parallel preparing of information huge datasets can be handled at a quicker rate.

1.7 MapReduce

MapReduce component of Hadoop which works in two phases, which are -

- Map phase
- Reduce phase

Each stage takes contribution to the type of key-esteem sets and furthermore creates yield as key-esteem structure. There are two capacities by which the handling is practiced – map work and decrease work. Guide work accepts a lot of information as info and produces a lot of key-esteem sets. Yield created by Map work is called as middle of the road yield. Yield from the Map work is provided as a contribution to the Reduce work and the conclusive outcome is produced by the Reduce work.

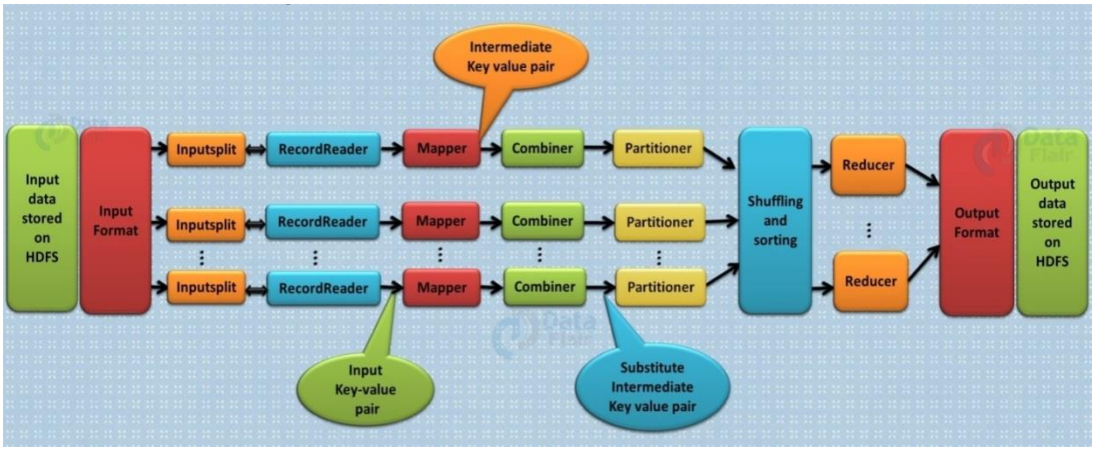

Figure 2. Working of MapReduce

1.8 Yet Another Resource Negotiator (YARN)

Apache Yarn :– "One more Resource Negotiator" , is the significant asset the executives layer of Apache Hadoop . The Yarn was presented in Hadoop 2.x. Various information handling motors like chart preparing, intuitive preparing, stream preparing just as clump handling are upheld by YARN which can run and process information put away in HDFS (Hadoop Distributed File System).Job Scheduling is additionally done by YARN.

1.9Apache Hive

Apache Hive is an open source information distribution center, which based over Hadoop. It is utilized for questioning and examining enormous measure of information (organized or semi-organized) put away HDFS. Composing MapReduce occupations for an assignment is a mind-boggling task however with hive information can be investigated utilizing SQL inquiries. Hive utilizes a SQL comparative language which is known as Hive Query Language (HQL). SQL like inquiries composed by the software engineer are naturally converted into the MapReduce employments.

1.10 Apache Pig

Pig is a deliberation over MapReduce calculation. With regards to huge informational collections just as to speak to them as information streams, we for the most part use Apache Pig, we use it with Hadoop. Pig offers a significant level language to compose programs which is called as Pig Latin.It is a procedural

programming language and fits normally in the pipeline worldview. At the point when questions are perplexing with a large portion of joins and channels then it is firmly prescribed to utilize it.

1.11 Power BI

Power BI is a Data Visualization and Business Intelligence apparatus gave by microsoft that permits the investigation and perception of information by giving diverse visuals. Dataset from various information sources can be imported to Power BI, which can be changed over to intuitive dashboards and BI reports.

2. LITERATURE REVIEW

This section comprises of summary and comparative study of the research papers that are related to our problem statement.

In this paper [1] author describe the Diabetes Data Prediction Using Spark and Analysis in Hue Over Big Data." This paper utilizes Big information Analytics for investigation of Diabetes information. The creator has gathered dataset from the Pima Indian database. The strategy utilized for this object is Hadoop MapReducee and Hue. Flash was utilized for quick preparing. At that point the code is written in Java in Hadoop in which the prescient examination calculation is utilized to break down the clinical dataset. The inquiries were written in Hive and indexes were made in Hue.

In this paper [2] author describe the Predictive Methodology for Diabetic Data Analysis in Big Data." The principle motivation behind this examination was to build up a framework which will utilize the prescient investigation calculation in Hadoop/Map Reduce condition to anticipate the diabetes types common, confusions related with it and the kind of treatment to be given. By changing different wellbeing records of diabetic patients to a helpful broke down outcome, these outcomes will make the patient ready to know the entanglements of the sickness and they can make expected move to decrease its complications.

In this paper [3] the creator has discussed the difficulties that are related with diabetes data. This look into paper depicts all factors that can influence Type 2 diabetes in various age gatherings and sexual orientation This paper likewise shows that how Big information examination is superior to anything customary techniques for the administration of diabetes information to deliver legitimate outcomes.

In this paper [4] author describe the Modernizing medicinal services industry's move towards handling huge wellbeing records, and to get to those for examination and put energetically will extraordinarily expands the complexities. Due to developing unstructured nature of Big Data from social insurance ventures, it is important to utilize different enormous information instruments to process the information and get the necessary outcomes. Social insurance industry is confronting a great deal of difficulties which makes it essential to utilize the information investigation in this field. By utilizing prescient investigation calculations in hadoop/Map Reduce to foresee the diabetes type and intricacies related with it and by breaking down it we can pick most appropriate treatment for it in the terms of moderateness and accessibility as well.

In this paper [5] author describe the A Genetic Algorithm based model is produced for patients particulary for diabetic patients to estimate the danger of coronary episode and stroke. To screen the individuals who are experiencing these sicknesses, this model will give a probability chance behind the coronary failure and other neuro productive illnesses. This model can likewise support the doctor with the goal that they can do the treatment of patient's appropriately. At long last, the specialists can ready to discover the affectability and particularity the extent of individuals who don't have the confusion to test negative on the screening trial of the screening procedure.

Principle convergence of this paper [6] on assessment that focuses on this piece of Medical end learning plan through the assembled information of diabetes and to make keen restorative decision sincerely steady system to support the doctors. In this framework, creator has talked about various calculations like Bayesian and KNN (K-Nearest Neighbor) which can be applied to diabetes database and break down them by taking different qualities of diabetes for forecast of diabetes illness.

3. Designs and Implementation

3.1 Architectural Design of the work

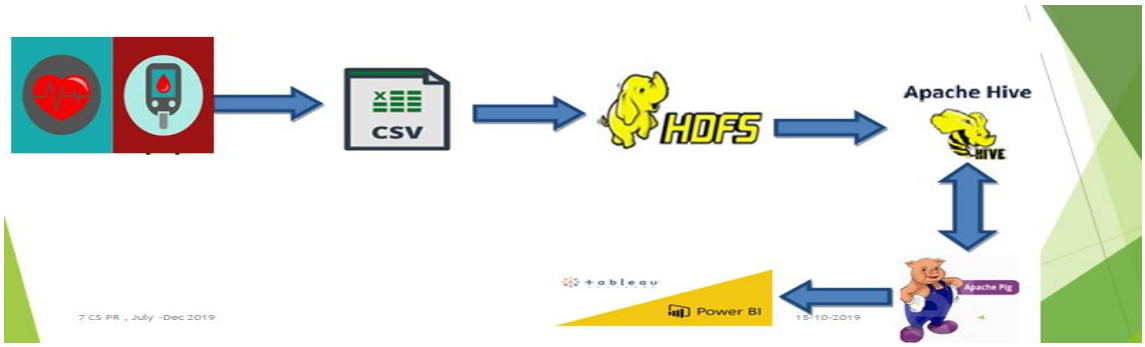

Figure 3 Architecture Design

Details of Inputs/ Data Used

The dataset is collected from PIMA Indian database. Various factors which are present in the dataset are as follows: -

Pregnancies: Number of times pregnant
Glucose: Plasma glucose concentration over 2 hours in an oral glucose tolerance test
Blood Pressure: Diastolic blood pressure (mm Hg)
Skin Thickness: Triceps skin fold thickness (mm)
Insulin: 2-Hour serum insulin (mu U/ml)
BMI: Body mass index (weight in kg/(height in m)2)

Diabetes Pedigree Function: Diabetes pedigree function (a function which scores likelihood of diabetes based on family history)
Age: Age (years)

Table 1: Snapshot of Dataset

Pregnancies	Glucose	Blood Pressure SkinThickness	SkinThickness	Insulin	BMI	Diabetes Age	Pedigr	Outcome
6	148	72	35	0	33.6	0.627	50	1
1	85	66	29	0	26.6	0.351	31	0
8	183	64	0	0	23.3	0.672	32	1

1	89	66	23	94	28.1	0.167	21	0
0	137	40	35	168	43.1	2.288	33	1
5	116	74	0	0	25.6	0.201	30	0
3	78	50	32	88	31	0.248	26	1
8	125	96	0	0	0	0.232	54	1
4	110	92	0	0	37.6	0.191	30	0
10	168	74	0	0	38	0.537	34	1
10	139	80	0	0	27.1	1.441	57	0
1	189	60	23	846	30.1	0.398	59	1
5	166	72	19	175	25.8	0.587	51	1
7	100	0	0	0	30	0.484	32	1
0	118	84	47	230	45.8	0.551	31	1
7	107	74	0	0	29.6	0.254	31	1
1	103	30	38	83	43.3	0.183	33	0
1	115	70	30	96	34.6	0.529	32	1
3	126	88	41	235	39.3	0.704	27	0
8	99	84	0	0	35.4	0.388	50	0
7	196	90	0	0	39.8	0.451	41	1
9	119	80	35	0	29	0.263	29	1

Details of Hardware / Software / Platform to be used by you and used by various researchers:

- Hadoop- Hadoop is an open source framework which is built in JAVA.
- Hue – It is an open source SQL cloud editor.
- Hive – It is a data warehouse built on the top of Hadoop framework.
- Spark– A Data processing tool which provides real time processing.
- Power BI –A data visualization tool provided by Microsoft.

Performance Evaluation

- Quality of data – This analysis totally depends on the variety, veracity of data [7]. Its quality depends on the size of data and number of parameters used for analysis of data.
- Type of data – If data is in digital form it would be more difficult of analyze the data as compared to text form of data.
- Distributive nature – The platform used is distributive in nature which makes it fast, and reliable.
- Nature of analysis – This analysis is done on every minute detail of the data which makes it more reliable.

4 Experimental Results & Analysis mental Results

a) Design specifications and objectives

This analysis is done on Pima Indian Diabetes Dataset to analyse various factors that causes diabetes and their dependencies with diabetes.

Various Steps were used to process [7] the data and obtain the required result.

- Pre-Processing

- Loading data to Hadoop
- Analysing the data using Pig & Hive
- Projecting the result on BI

4.1 Pre-Processing

At first glance, the dataset appears to be clean. On deeper analysis, the dataset revealed abnormal values for biological attributes. Attributes such as Skin Thickness and Glucose had 227, 374 and zero values. The fact that these measures cannot hold zero value indicating that the missing values in the dataset were represented as zero value in the dataset. The missing attributes in the dataset has a place with about 30% of the perceptions in the dataset. As removing these values would result in much information loss[8], andmean value replacement was used to fill the missing values in the data set. Only wrong values in the dataset (zero values) were imputed. Large outliers in the variables were not taken into account.

Table: 2 Pre-processing the data

Pregnancies	Total	Yes	No	Probability
0	111	38	73	34.23423
1	135	29	106	21.48148
2	103	19	84	18.4466
3	75	27	48	36
4	68	23	45	33.82353
5	57	21	36	36.84211
6	50	16	34	32
7	45	25	20	55.55556
8	38	22	16	57.89474
9	28	18	10	64.28571
10	24	10	14	41.66667
11	11	7	4	63.63636
12	9	4	5	44.44444
13	10	5	5	50
14	2	2	0	100
15	1	1	0	100
17	1	1	0	100

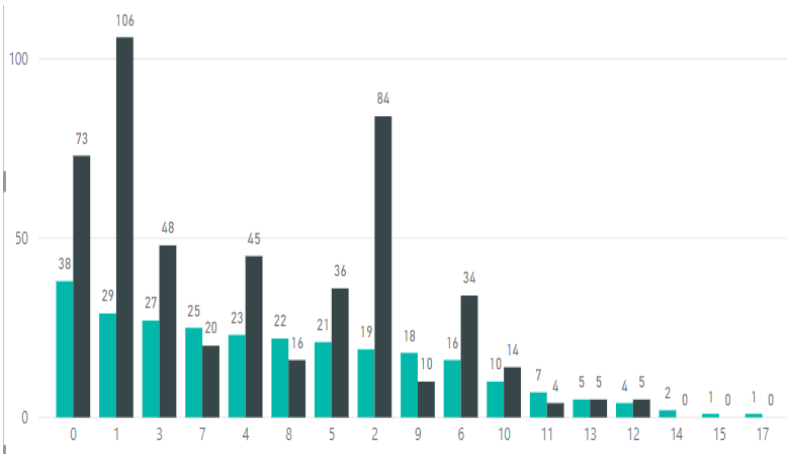

Figure.4 BI Report

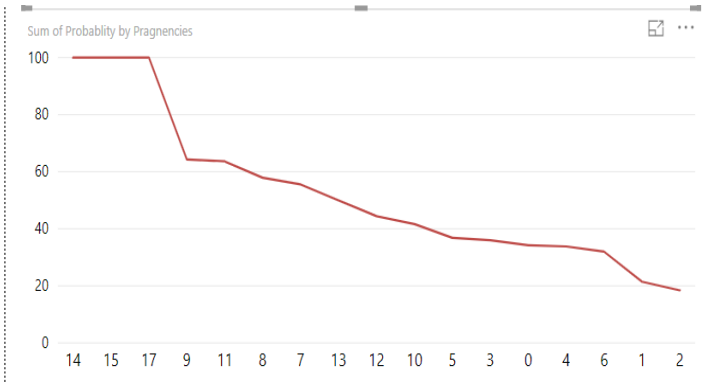

Figure. 5 Graphical Visualization

4.2 Dependency of Diabetes on BMI

The below plot shows that all the women who had been diagnosed [9,10] with Diabetes had a BMI greater than 25, which is above normal levels. On the other hand, women who did not had Diabetes had a BMI ranging from 18 to 60.

$$BMIFormula = (weight(lbs) * 703)/height\ (in^2)$$
$$MetricBMIformula = Weight(kg) * Height(in^2)$$

Query for compute the output [11]

1. Select count(id) from dibetes where BMI>18 and BMI<25;
2. Select count(id) from dibetes where BMI>18 and BMI<25 and outcome=1;

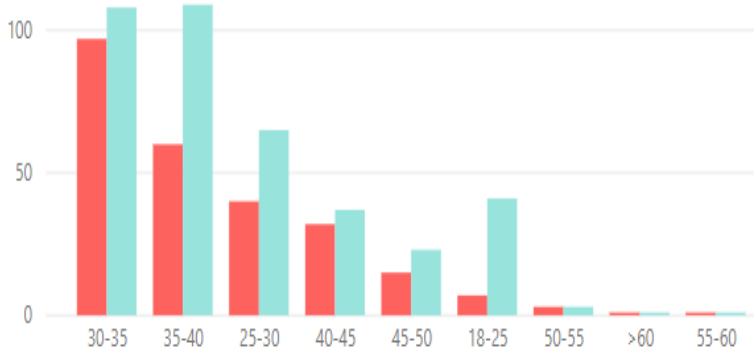

Figure.6 BI Report

5 Conclusion

Enormous information examination in Hadoop is intended for current wellbeing to arrive at great wellbeing results simultaneously as human services accessibility and moderateness. This examination study had led a considerable examination on the treatment of diabetes in the social insurance industry, prompting an immense procedure of data. The system of the introduction screening routine for the treatment of diabetes can give dynamic data and examination that meets the best outcomes in pharmacological administrations. By utilizing Hadoop, we have made this examination increasingly powerful and can be utilized in further investigations on diabetes. This investigation report proposes new viewpoints utilizing Hadoop and Big Data

on how Hadoop and Big Data van are utilized to effectively break down information and make a report that is utilized for further examination should be possible.

A few outcomes that can be drawn from this report:

- Number Women who consider in more prominent numbers are bound to get diabetes.
- Persons having BMI more 40 have more than 50 percentage of chance for getting diabetic.
- Persons having Diastolic Blood Pressure more than 80 have more than 40 percentage of chance of getting diabetic.

Future Scope

Utilizing this model can be created for patients with diabetes to reach and anticipate inclined danger of respiratory failure and stroke. To screen the individuals who are experiencing the malady, the proposed application model can be developed that breaks down different wellbeing factors and will give the probability of expanded danger of cardiovascular failure or other neurode ficiency illnesses.

That model can be utilized to give a final result model to early discovery and hazard identification of cardio and cardio - vascular maladies. A model can be created utilizing the machine that will utilize the subtleties gave by its client and foresee the client how likely the patient is to have a stroke or some other neurological issue.

Referencing and Appendices

1. Guttikonda, Geetha, Madhavi Katamaneni, and MadhaviLatha Pandala. "Diabetes Data Prediction Using Spark and Analysis in Hue Over Big Data." 2019 3rd International Conference on Computing Methodologies and Communication (ICCMC). IEEE, 2019.
2. Eswari, T., P. Sampath, and S. Lavanya. "Predictive methodology for diabetic data analysis in big data." Procedia Computer Science 50 (2015): 203-208.
3. Wang, Lidong, and Cheryl Ann Alexander. "Big data analytics as applied to diabetes management." European Journal of Clinical and Biomedical Sciences 2.5 (2016): 29-38.
4. Eswari, T., P. Sampath, and S. Lavanya. "Predictive methodology for diabetic data analysis in big data." Procedia Computer Science 50 (2015): 203-208.
5. Sabibullah, M., V. Shanmugasundaram, and R. Priya. "Diabetes patient's risk through soft computing model." International Journal of Emerging Trends & Technology in Computer Science (IJETTCS) 2.6 (2013): 60-65.
6. Shetty, Deeraj, et al. "Diabetes disease prediction using data mining." 2017 International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS). IEEE, 2017.
7. Kumar, A., & Sinha, M. (2019). Design and analysis of an improved AODV protocol for black hole and flooding attack in vehicular ad-hoc network (VANET). Journal Of Discrete Mathematical Sciences And Cryptography, 22(4), 453-463.
8. Kumar, A., Dadheech, P., Singh, V., Poonia, R., & Raja, L. (2019). An improved quantum key distribution protocol for verification. Journal Of Discrete Mathematical Sciences And Cryptography, 22(4), 491-498.
9. Dadheech, P., Goyal, D., Srivastava, S., & Kumar, A. (2018). A scalable data processing using Hadoop & MapReduce for big data. J. Adv. Res. Dyn. Control. Syst, 10, 2099-2109.
10. Kumar, A., Goyal, D., & Dadheech, P. (2018). A novel framework for performance optimization of routing protocol in VANET network. J. Adv. Res. Dyn. Control. Syst, 10, 2110-2121.

11. Kumar, A., & Sinha, M. (2019). Design and development of new framework for detection and mitigation of wormhole and black hole attacks in VANET. Journal Of Statistics And Management Systems, 22(4), 753-761.
12. Abhishek Kumar, et al. "The state of the art of deep learning models in medical science and their challenges". Multimedia Systems. (2020).
13. Abhishek Kumar, et al. "Efficient data transfer in edge envisioned environment using artificial intelligence based edge node algorithm". Transactions on Emerging Telecommunications Technologies. (2020).
14. Ambeth Kumar, V.D. et al. "Active volume control in smart phones based on user activity and ambient noise". Sensors (Switzerland) 20. 15(2020): 1-17.
15. Vengatesan, K. et al. "Analysis of Mirai Botnet Malware Issues and Its Prediction Methods in Internet of Things". Lecture Notes on Data Engineering and Communications Technologies 31. (2020): 120-126.
16. Vimal, V. et al. "Artificial intelligence-based novel scheme for location area planning in cellular networks". Computational Intelligence. (2020).
17. Kumar, A. et al. "Comparative Analysis of Data Mining Techniques to Predict Heart Disease for Diabetic Patients". Communications in Computer and Information Science 1244 CCIS. (2020): 507-518.
18. Sayyad, S. et al. "Digital Marketing Framework Strategies Through Big Data". Lecture Notes on Data Engineering and Communications Technologies 31. (2020): 1065-1073.
19. Kumar, V.D.A. et al. "Exploration of an innovative geometric parameter based on performance enhancement for foot print recognition". Journal of Intelligent and Fuzzy Systems 38. 2(2020): 2181-2196.
20. Vengatesan, K. et al. "Secure Data Transmission Through Steganography with Blowfish Algorithm". Lecture Notes on Data Engineering and Communications Technologies 35. (2020): 568-575.
21. Lone, T.A. et al. "Securing communication by attribute-based authentication in HetNet used for medical applications". Eurasip Journal on Wireless Communications and Networking 2020. 1(2020).
22. Vengatesan, K. et al. "Simple Task Implementation of Swarm Robotics in Underwater". Lecture Notes on Data Engineering and Communications Technologies 35. (2020): 1138-1145.
23. Kesavan, S. et al. "An investigation on adaptive HTTP media streaming Quality-of-Experience (QoE) and agility using cloud media services". International Journal of Computers and Applications. (2019)
24. Ankit Kumar, Vijayakumar Varadarajan, Abhishek Kumar, Pankaj Dadheech, Surendra Singh Choudhary, V.D. Ambeth Kumar, B.K. Panigrahi, Kalyana C. Veluvolu, Black Hole Attack Detection in Vehicular Ad-Hoc Network Using Secure AODV Routing Algorithm, Microprocessors and Microsystems, 2020, 103352, https://doi.org/10.1016/j.micpro.2020.103352