International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

Machine Learning Algorithms for IOT Services in Big Data and Cloud Computing

[1]Sesha Bhargavi Velagaleti, [2]Dr. Suma T, [3]Dr. Shubhangi N. Ghate, [4]Harendra Singh Negi, [5]Dr. G. Charles Babu, [6]Arun Pratap Srivastava, [7]Navneet Kumar, [8]Dr. Anurag Shrivastava

Submitted: 07/02/2024 **Revised**: 15/03/2024 **Accepted**: 21/03/2024

Abstract: The phrase "cloud computing" refers to a kind of data management system in which mobile devices are not used for either the processing nor the storing of user data. The Internet of Things (IoT), a brand-new technology that is only now entering its formative years, is also becoming more widespread in the networks and telecommunications sectors. The "modern" sector of wireless telecommunications networks is where the majority of the emphasis of application for the Internet of Things is now being directed. In the most recent part of our line of research, we investigated the relationships and interactions that exist between the many different entities and equipment that communicate across wireless networks. They need to achieve the goal that has been set for them as a group in order to make the atmosphere more conducive to the use of big data. This will help create a more favourable environment for the use of big data. This article discusses the Internet of Things (IoT) and Cloud Computing technologies, with a particular focus on the security challenges that each of these technologies has experienced. In the field of medicine, for instance, big data is being put to use in order to bring down the costs of treatment, anticipate the arrival of pandemics, prevent sickness, and carry out a variety of other related activities. This article provides a comprehensive introduction to the approach of big data analytics, which is crucial in a variety of fields of work and businesses. First, we will present a brief overview of the concept of big data, which refers to the quantity of data that is generated on a daily basis, as well as its characteristics and facets.

Keywords: *Internet of Things, Cloud Computing, Big Data, Security, Privacy.*

1. Introduction

The "Internet of Things" is a cutting-edge technology that is now being used in the sector of telecommunications. According to the consensus of a number of authorities in the relevant field, the Internet of Things (IoT) is best described as "the network of devices, vehicles, buildings, and other items that are embedded with sensors and connected to the network, allowing these objects to gather data and exchange it with one another."In the next years, there is expected to be not only a growth in the number of connected devices and places, but also an increase in the number of activities that will be carried out by these devices and locations. It is very necessary to investigate and find solutions to the data privacy and security issues that arise while utilising a wireless network. Through the use of BD analytical tools and services, the problem of data privacy and security in day-to-day life may either be considerably mitigated or eradicated totally. The term "big data" (sometimes abbreviated as "BD") is a relatively new acronym that was developed relatively recently to denote the astonishingly rapid increase in the volume of both organised and unstructured data. CC is often used by BD as an essential piece of technology in the operation of its firm. Computing at the edge is yet another kind of technology that have the potential to be exploited as a foundation in a manner that is analogous to this. Electronics, connection, sensors, software, and a variety of other technologies make it possible to create what is referred to as the Internet of Things (IoT), which is a network of physical commodities or things. It is also sometimes referred to by the acronym IoE, which stands for the Internet of Everything. Because of this, the Internet of Things is able to send data across a wide variety of networked devices at a faster rate and with a greater quality of serviceV. Investigators have devised a plan with

1Assistant Professor, Department of Information Technology, G Narayanamma Institute of Technology & Science, Hyderabad
b.velagaleti@gmail.com
2Professor, Department of Computer Science and Engineering, Sri Venkateshwara College of Engineering, Bengaluru
tsumamurthy.cs@gmail.com
3Assistant Professor, Department of Electronics and Telecommunication Engineering,
Ramrao Adik Institute of Technology, D. Y. Patil Deemed to be University, Nerul, Navi-Mumbai, Maharashtra
shubhangi.ghate@rait.ac.in
4Department of Computer Science & Engineering, Graphic Era Deemed to be University
Dehradun, India
harendrasinghnegi@geu.ac.in
5Professor, Department of CSE, GRIET, Bachupally, Hyderabad, Telangana
charlesbabu.griet@gmail.com
6Lloyd Institute of Engineering &; Technology, Greater Noida
apsvgi@gmail.com
7Lloyd Law College, Greater Noida
navneet.kumar@lloydlawcollege.edu.in
8Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences, Chennai, Tamilnadu
**anuragshri76@gmail.com*

the intention of assisting other investigators working in the field who are interested in matters concerning security. This method provides not only a novel framework for the Internet of Things that has been tested via the use of a case study, but it also examines the Internet of Things-related security of existing research initiatives in the field of information security. On the basis of their own research, the authors of this article have shown how the creation of autonomous devices presents a threat to the country's safety and security [1].

The demand for aid from the "cloud" has grown ineffective as a result of the extensive computations, the massive storage, and the security concerns. Some examples of these limits include limitations on processing power, available energy, the ability to communicate with others, and available storage space. Inefficiencies such as these are one of the primary motivating causes behind our efforts to build a paradigm for the merging of CC and IoT. The term "base" technology refers to the potential of cloud computing to integrate a wide range of technologies and applications in order to improve the capacity and performance of the existing infrastructure. This ability is facilitated by the use of cloud computing.

Additionally, during the last several years, Mobile Cloud Computing (MCC), which is a subset of Cloud Computing, has evolved as a result of developments in the field of "Cloud Computing." The latter seeks to make it possible to have access to data and information whenever and wherever it is necessary by doing away with the requirement to make use of physical equipment at any given moment. The combination of mobile computing and cloud computing is known as MCC, and it is responsible for the enhancement of the capabilities of mobile devices. Additionally, it offers a contemporary method for companies and organisations that are looking for creative services to approach their search. Using CC as a foundation might be beneficial for both the Internet of Things and video surveillance systems since it will improve the functionality of both of these systems [2].

In addition, the purpose of cloud computing is to make it possible to access data and information at any time and from any place without the limitations imposed by the need for physical equipment. This may be done via the use of the internet. Because of the method in which CC operates, it may become an essential technology for the Internet of Things and for other technologies that are connected to telecommunications. Additionally, it may improve the capabilities of other technologies [3].

The remaining parts of this essay are broken up into the sections that follow below. reporting as well as an in-depth study of relevant articles originating from this industry. It is possible to find evaluations, explanations, descriptions, and examples that are unique to Internet of Things applications, communication protocols, Internet of Things architecture, and smart city components. It investigates the creation of enormous data sets, the inclusion of sensor data, the semantic data that has been annotated, and the quality of the data. In this article, machine learning approaches from eight distinct domains are analysed and summarised using the most recent findings from research on Internet of Things data and the pervasiveness of machine learning algorithms. These findings were derived through research conducted by the authors of this article. Once the algorithms have been used in a variety of smart city applications, it will then be possible to make a conclusion, in addition to identifying future research trends and issues that have not yet been resolved. Figure 1 is an illustration of an example of the format of the survey [4].

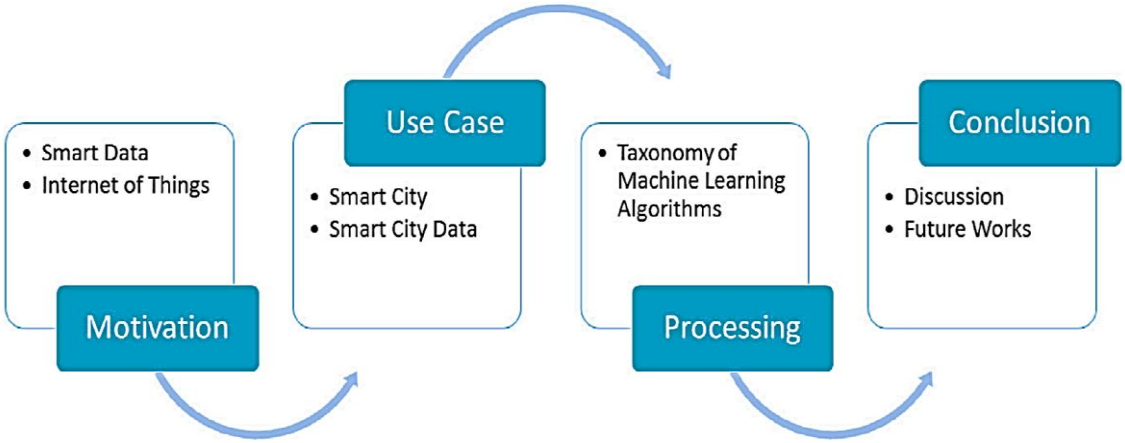

FIG 1: ORGANIZATION OF SURVEY.

2. Review of Literature

The phrase "cloud computing" refers to a kind of data management system in which mobile devices are not used for either the processing nor the storing of user data. The Internet of Things (IoT), a brand-new technology that is only now entering its formative years, is also becoming more widespread in the networks and telecommunications sectors. The "modern" sector of wireless telecommunications networks is where the majority of the emphasis of application for the Internet of Things is now being directed. In the most recent part of our line of research, we investigated the relationships and interactions that exist between the many different entities and equipment that communicate across wireless networks. They are required to work together to complete the task that has been given to them in order to create an environment that is more conducive to the use of Big Dat. Both cloud computing and the internet of things have the potential to see rapid expansion if they are able to take use of the most recent technological improvements that have been made in wireless network technology. This article discusses the Internet of Things (IoT) and Cloud Computing technologies, with a particular focus on the security challenges that each of these technologies has experienced. The two technologies mentioned above, cloud computing and the internet of things, have been particularly compared in order to emphasise their similarities and to research and discover the advantages of their integration. The primary emphasis of this comparison has been on ensuring the usage and transmission of big data. The benefits of combining the two have also been emphasised in this research, which was carried out for that purpose. The technologies of cloud computing and the internet of things have been shown to complement one another and act as the basic technologies for Big Data systems. This has been shown via a number of different studies [5].

The invention of sensing devices that are linked to the internet was made possible by rapid improvements in computer technology, internet connection, and software. These devices have the capability to collect data about the immediate physical environment and transmit that data. By the year 2020, it is projected that there will be between 25 billion and 50 billion internet-connected devices in use throughout the globe. As these numbers continue to rise and as technological advancements are made, an increasing amount of information will be made public. The capabilities of the existing Internet are being improved thanks to a concept known as the "Internet of Things," which is a phrase used to describe the technology of items that have Internet access. This is accomplished by making it easier for people in the digital world and the real world to connect with one another and communicate with one another. The Internet of Things (IoT) is responsible for the generation of big data, which may be identified by its velocity, location, and time dependency, in addition to a variety of different modalities and varying data quality. In addition to that, the amount of data being collected is growing. Processing and analysis of massive amounts of data using intelligence is a prerequisite for developing intelligent applications for the Internet of Things. In this article, the comparison of the many machine learning approaches that may be used to handle the challenges posed by IoT data focuses primarily on the application of such techniques to the context of smart cities. The primary purpose of this article's use case is to compare and contrast these different tactics. The classification of different machine learning algorithms is the aspect of this work that stands out as the most important contribution it makes. This taxonomy provides an explanation of the several approaches that were used in order to examine the data and get more advanced levels of knowledge. In addition, we will discuss the benefits and drawbacks of using machine learning in the context of Internet of Things data analytics. Use cases of applying a Support Vector Machine to traffic data from the smart city of Aarhus are made available to anybody interested in doing a more in-depth examination. Aarhus is a smart city [6].

The Internet of Mechanical Things is rapidly becoming more significant, and it has already evolved into the Internet of Things of the future. Sharing of large amounts of data together with machine learning for wide applications of robotization now in use. the process of linking already existing computer components present in ordinary things to the internet so that such devices may send and receive information. Standard data processing systems are unable to effectively handle business intelligence (BD) informative collections because of their size and complexity. BD relies on these informative collections. "Machine learning," sometimes abbreviated as "ML," is a subset of artificial intelligence that allows computers to "learn" from data even when they are not specifically designed to do so. In machine learning, measurable approaches are used rather often. It is possible that the efficiency of data management and information revelation for large-scale robotization of applications may be improved by combining a few separate innovations, such as sensor technology, the Internet of Things, computational intelligence, machine learning, and big data. IoT and BD are rising as a result of an increasing number of relevant data sources, and the availability of a broad variety of machine learning calculations is opening up new doors for the provision of logical services to businesses [7].

The ability to integrate the present best in class into a system that would assist to cut development costs and allow new kinds of services is now unavailable. However,

in the future, this capability may become available. Over the course of the last decade, a vast quantity of data has been produced as a direct result of the growing scalability of Internet of Things devices. However, such facts are of little utility if they are not supported by data from relevant scientific studies. People are now able to receive information that is pertinent to the vast quantities of data that are made by IoT devices as a result of the many BD and IoT investigation arrangements that have been built. However, due to the fact that these configurations are still in the preliminary stages of development, the discipline has not yet finished conducting an exhaustive investigation of them. In this article, we have made an effort to offer a balanced and comprehensive description of the IoT in BD framework, along with all of the many challenges and roadblocks that it presents. Our primary emphasis has been on developing actionable solutions via the use of a machine learning approach [8].

3. Big Data Challenge Can Be Solved Using ML

The beginning of the maturation process for machine learning as the field makes the transition from experimental labs and proof-of-concept applications to the forefront of commercial partnerships. Along the road, it will assist with the regulation of cutting-edge technology such as self-driving automobiles, precision farming, the creation of medications utilised in the treatment of sicknesses, and the development of sophisticated extortion sites for monetary foundations. Along the way, it will be helpful in regulating developing technology like self-driving vehicles as they become more commonplace. Machine learning, sometimes known as ML, is a method that blends insights, software engineering, and false awareness. It places an emphasis on the creation of quick calculations to allow continuous data preparation. The performance of these machine learning calculations is far better than that of just following to explicitly updated rules [9]. As a consequence, these calculations are now an essential component of computerised reasoning systems.

❖ ML helps to handle IoT data flows

The Internet of Things (IoT), which was one of the most talked-about technical advances of the previous year, presents a challenge that machine learning (ML), which may be able to assist us with, may be able to help us with. According to Vin Sharma, director of ML arrangements in Intel's information Canter cluster, the first iteration of big data analysis developed around the stream of information produced by online social networking, online searching, online recordings, online browsing, and other consumer-generated online behaviours. This was the case in the first iteration of big data analysis. To break down these massive datasets, new innovations, all-mains distributed computing, and virtualization programming were necessary, such as Apache Hadoop and Spark. Additionally, all of the additional powerful computers that were capable of removing data pieces from massive data sets were necessary for it. The systems that are linked to the Internet of Things currently have the upper hand when it comes to the amount of information. As more electronic gadgets and sensors become available on the market, there is a good chance that the amount of data that these gadgets and sensors generate will also expand.

Let's imagine, for the sake of illustration, that a single huge automobile generates 4,000 GB of data every single day. The most recent kind of aircraft, the A380-1000, is fitted with 10,000 sensors that are dispersed over each wing. Because there will be so much information, it won't be able to manage all of the linked devices that are found in intelligent homes and cities, as well as the autonomous systems that are found in intelligent businesses.

❖ New & exciting system requirements

Machine learning (ML) is essential in order to manage the massive volumes of data that are produced by the many and dependable devices that are part of the Internet of Things (IoT). Users of online social networking and web-based search engines are already making use of machine learning, and they see it as being highly natural. Because of this, some people can get the impression that machine learning is a technology from the far future. The recommendation engine on Amazon and the news feed on Facebook both make use of machine learning in order to provide suggestions for the next books and films you should watch.

Machine learning (ML) algorithms are able to comprehend the regular data samples that are shown on Internet of Things (IoT) devices. These algorithms then focus their attention on outliers, which are occurrences that deviate from the norm. Machine learning will distinguish the "flag from the clamour" in vast information flows that originate from billions of information foci so that businesses may concentrate on what is most important. For machine learning algorithms to be relevant and rewarding for associations, they need to be able to run counts on huge numbers in only a few milliseconds, and they also need to be able to do this consistently. The conventional processors and processing stages that are present in data centres are being put under a lot of strain as a result of these ever increasing count numbers [10].

In order to function at scale and maintain consistency, ML frameworks need memory subsystems that are much quicker, processors that can handle numerous linked foci, and strategies that can parallelize the process of preparation for front-line consistent knowledge. These stages are equipped with built-in illustrated maintenance engines as well as the capability to do sophisticated calculations in memory. This ensures that the information

may be utilised immediately and that the results can be trusted.

❖ **Final prediction**

Those individuals who were responsible for processing the primary enrolment will be sought out. ML and false awareness will essentially need more power as they grow closer to reaching a choice regarding the information streams related to IoT and consumer engagement in order to make better arrangements and attempts.

These machines were often used in research laboratories and other supercomputing endeavours as venues for their operations. The sequencing of genomes and the display of atmospheric phenomena were two of these projects. However, machine learning platforms will become more vital as Internet of Things frameworks get noticeably larger and more defined and as organisations dynamically develop their success based on the data that is received via machine-to-machine communication. These processors hand off the execution of the most labor-intensive activities, such as artificial intelligence and fake-cognition sculptures, to other computers. They will never again be relegated to the lofty heights of supercomputing at research centres and universities as they continue to become a necessity for cutting-edge businesses.

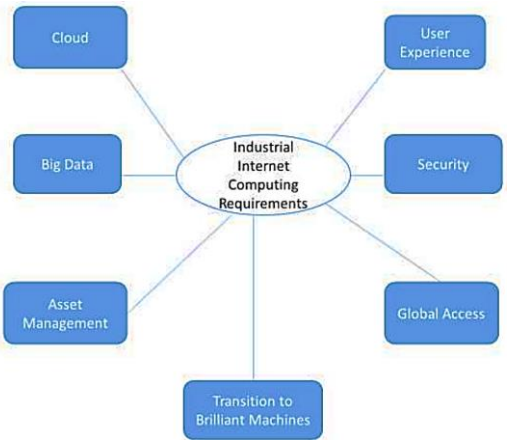

FIG 2: IOT'S COMPUTING REQUIREMENTS

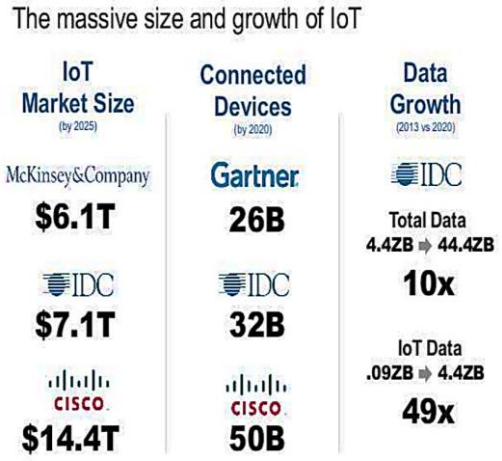

FIG 3: SIZE GROWTH OF IOT

4. Taxonomy of Machine Learning Algorithms

The field of computer science that is most often referred to as machine learning is really one of its subfields. Machine learning, a subfield of artificial intelligence (AI), frees computers from the constraints of explicit programming, allowing them to acquire new skills on their own. The field of study known as machine learning was founded on the principles of pattern recognition and the concept of computational learning. This article delves into some of the most basic concepts pertaining to machine learning as well as some of the most prominent approaches pertaining to machine learning that are used for in-depth data analysis. An input that is taken into consideration by a learning algorithm is referred to as a training set of samples. Learning may typically be broken down into three distinct categories: supervised learning, unsupervised learning, and reinforced learning. When doing supervised learning, the "training set" is nothing more than a collection of relevant input vector samples and target vectors, which are sometimes referred to as labels at times. This collection is referred to by its given name, the label set. In order for the unsupervised learning process to be successful, it is not necessary for the training set to include any labels. The method of reinforcement learning is an effort to find a solution to the problem of determining what action to perform in every given circumstance in order to get the greatest possible amount of rewards. This study focuses mostly on supervised learning and unsupervised learning because of the considerable use of both approaches in the analysis of smart data generated by IoT devices. The utilisation of both techniques is significant. The talent of successfully anticipating the output vector based on the input vector that is being provided is what one hopes to achieve via the process of supervised learning. Applications that include categorization include the creation of objective labels based on a predetermined number of distinct categories. A regression issue is what happens when one or more continuous variables are utilised to construct the target labels. This condition is referred to as a difficulty with regression. The purpose of unsupervised learning may be difficult to pin down, which might provide certain issues. One of the most important objectives is to cluster the input data, which is a word that describes the process of locating appropriate groupings of samples that are similar to one another. When moving the original input variable into a new variable space, one of the purposes of preprocessing the variable may be to find a usable internal representation of the input data. This would allow the variable to be moved into the new variable space. In order to accomplish this goal, the first input variable is copied over into the new variable space. The output of the machine learning algorithm that comes after the feature extraction

preprocessing stage has the potential to be significantly improved thanks to this stage's capabilities.

In order to train CART, the structure of the tree has to be built in a manner that is dependent on the training data. In order to accomplish this goal, you will need to calculate the value for the threshold parameter and determine the split criterion for each node. CART is taught to create the tree from the top down and select the optimum split node by node since picking the optimal tree structure is an NP-complete problem. This allows CART to construct the tree in the most efficient way possible. This is accomplished despite the fact that determining the optimal tree structure is an NP-complete problem. In order to significantly cut down on overfilling and boost the tree's generalizability, the building of the tree has to have certain stopping criteria. Examples of possible stopping points include reaching the maximum depth, determining the purity of the branch's distribution, determining whether or not the benefit of splitting is below a predetermined threshold, and determining whether or not the number of samples in each branch is below the threshold for the criterion. All of these can be accomplished by determining whether or not the benefit of splitting is below a predetermined threshold. After the tree has been constructed, a unique kind of pruning may be performed in order to reduce the amount of overfitting that occurred. The instructions on how to train CART are included in Algorithm 1.

ALGORITHM 1: Algorithm for Training CART

Input: labeled training data set $D = \{(x_i, y_i)\}_{i=1}^{N}$.

Output: Classification or regression tree.

FITTREE(0, D, *node*)

  function FITTREE(*depth*, R, *node*)

    if *the task is classification* **then**

      node.prediction := most common label in R

    else

      node.prediction := mean of the output vector of the data points in R

    end

    (i^*, z^*, R_L, R_R) := SPLIT(R)

    if *worth splitting and stopping criteria is not met* **then**

      node.test := $x_{i^*} < z^*$

      node.left := FITTREE(*depth* + 1, R_L, *node*)

      node.right := FITTREE(*depth* + 1, R_R, *node*)

    end

    return node

The fact that CART is presented in a format that resembles a tree makes it very user-friendly, which is its primary advantage. Additionally, it is fast and scales well for the management of large data sets. Despite this, it is very dependent on the specific training set that is used. This technique has a number of flaws, one of which is that the labelling of the input space is not consistent. This issue arises as a result of the fact that each part of the input space has exactly one label.

❖ Principal component analysis

The purpose of principal component analysis, often known as PCA, is to orthogonally project data points onto the principle subspace, which is an L-dimensional linear subspace that has the highest projected variance. locate a complete orthonormal collection of L linear M-dimensional basis vectors (Wj) and the associated linear projections of data points (Znj) in such a manner that the average reconstruction error is decreased is one approach to describing the aim. Another approach is to say that the goal is to locate an orthonormal collection of L linear M-dimensional basis vectors. In this circumstance, the value shown by x represents the mean of all of the data points.

$$J = \frac{1}{N}\sum\nolimits_{n} \|\tilde{x}_n - x_n\|^2$$

$$\tilde{x}_n = \sum\nolimits_{j=1}^{L} z_{nj} w_j + \bar{x}$$

As seen in Algorithm 3, the PCA approach is used to successfully complete these objectives. PCA is a technique that may have a variety of different run lengths depending on the algorithm that is used to generate w1,...,WL. These run durations can include O(M3), O(LM2), O(NM3), and O(N3). An alternate PCA approach that is similarly accessible is built on iterative expectation maximisation as its foundation. This method may be used to manage data sets that have a significant number of dimensions. The phases that need the most effort have a complexity of O(NML), and the covariance matrix of the dataset is not explicitly produced when applying this technique. This is because the covariance matrix is dependent on the phases that require the most labour. This strategy may also be implemented online, which is beneficial in situations in which M and N are both rather large numbers.

ALGORITHM : PCA Algorithm

Input: L, and input vectors of an unlabeled or labeled data set $\{x_1, \ldots, x_N\}$.

Output: The projected data set $\{z_1, \ldots, z_N\}$, and basis vectors $\{w_j\}$ which form the principal subspace.

$\bar{x} := \frac{1}{N}\sum_n x_n$

$S := \frac{1}{N}\sum_n (x_n - \bar{x})(x_n - \bar{x})^T$

$\{w_j\}$:= the L eigenvectors of S corresponding to the L largest eigenvalues.

for $n := 1$ *to* N **do**

  for $j := 1$ *to* L **do**

    $z_{nj} := (x_n - \bar{x})^T w_j$

  end

end

Principal component analysis, sometimes known as PCA, is widely regarded as one of the most effective preprocessing techniques for machine learning. It has a variety of applications, including data whitening, data compression, and data visualisation, to name a few of the more common ones. It has potential applications in a variety of real-world fields, including neurobiology, neurorecognition, and interest rate derivative portfolios, to name just a few of them. In addition, a kind of PCA known

as the kernelized principal component analysis (KPCA) has the potential to identify nonlinear principal components.

5. Research Methodology

Before delving into more detail about the methodology, this section provides an overview of the dataset.

❖ Dataset and Preprocessing

The proposed method was evaluated while it was carrying out the HAR task by utilising the MHEALTH Mobile Health dataset as the basis for the evaluation. The use of sensors enables the classification of a wide variety of activity types, including but not limited to walking, running, and sitting. The MHEALTH dataset comprises recordings of 10 persons' physical motions captured as they participated in a variety of activities. These activities ranged from walking to running to weight lifting. The information was gathered with the use of three distinct types of sensors, namely an accelerometer, a gyroscope, and a magnetometer. These sensors, taken together, generated three signals, each of which represented one of the three axes. The accelerometer is the only sensor that is located on the chest; the other two sensors are located on the left ankle and the right wrist, respectively.

TABLE 1: PERFORMANCE METRICS FOR DATA PROCESSING

Elapsed Time (s)	Slot Time Consumed (s)	Average Read (ms)	Average Compute (s)	Average Write (ms)	Number Of Rows	Size Of The Dataset
0.3	0.043	22	1.048	2	61,900	0.0175 GB
72	828.547	355	2.3	28,599	41,340	1.2 GB
3.3	3.663	945	4.6	109	1,00,000	2.3 GB
2.1	2.424	118	6.7	77	30,646	2.9 GB
1.6	1.506	237	18.9	145	1,00,000	3.6 GB

Before moving on to the next step of data exploration, let's have a look at the performance impact of using Big Query's basic queries with different amounts of datasets. The results of executing five fundamental select queries on five distinct datasets are shown in Table 1, which lists the relevant information. According to the findings, there is a correlation between the size of the dataset and the average performance of read, write, and compute operations across six important performance categories. The graph illustrates that there is an exponential connection between the size of the dataset and the average compute size, which demonstrates that the average processing time grows in proportion to the size of the dataset.

❖ Data Reduction

In this post, we will investigate edge data reduction strategies that can either be undone or cannot be undone.

❖ *Reversible:*

During the process of carrying out reversible data reduction, three primary scenarios are utilised: all sensors, location-based, and similarity-based. Any one of them may be done with or without sliding windows depending on your preference. Table I displays, using the reversible approach and a 100-point frame, the number of characteristics that were present for each scenario both before and after the data reduction procedure was carried out. This information is shown in relation to each scenario. The term "Direct" is used throughout the table and the rest of the text to refer to the method that does not include a sliding window.

TABLE 2: NUMBER OF FEATURES BEFORE AND AFTER REDUCTION FOR REVERSIBLE APPROACHES (WINDOW SIZE 100)

Scenario	Edge Location	Original Features	66% Reduction	70% Reduction
All Sensors				
Direct	-	21	7	-
S. Window	-	21 X 100	696	635
Location Based				
Direct	L1	4	2	-
	L2	8	5	-
	L3	8	5	-
S. Window	L1	3x100	100	92
	L2	9x100	298	273
	L3	9x 100	299	272
Similarity Based				

Direct	S1	8	4	-
	S2	7	3	-
	S3	7	3	-
Windo w	S1	9x100	295	273
	S2	6x 100	197	185
	S3	6x 100	197	185

Nevertheless, tests were also carried out for reductions of 80%, 90%, and 95%, respectively. Table 1 displays the number of qualities that were present before to a decline of 66 and 70%, and the number of characteristics that remained after the decrease of these percentages, respectively. A starting reduction of 66% was selected since this would allow for the same level of success as the highest reduction that could be achieved with the vector magnitude technique. Due to the fact that everything takes place on a single node, there is no edge location for any sensor. An edge location is a node at which data are collected for strategies that are location-based or similarity-based. We refer to this as an edge location. The location-based method consists of a total of three nodes, each of which corresponds to a sensor that is placed in a different part of the body (the chest, the arm, and the leg, respectively). The accelerometer, gyroscope, and magnetometer are the three distinct kinds of sensors that are associated to the nodes in the similarity-based technique. According to Table 1, the direct technique has a total of nine unique characteristics, which may be translated into three sensors that each have a total of nine axes. As an example, the L2 component of location-based techniques compiles information from a total of three sensors' readings. In circumstances involving windows, the number of features is proportional to the length of the window plus the number of features; for instance, the number of features for a location-based technique and L2 is 9,100. It is very important to keep in mind that the direct option will not be available until there has been a decrease in the amount of data by at least 66%. Because the number of characteristics was already so low at the time, it was agreed that any further decline would be undesired. For example, there were only 7 characteristics for all of the sensors at the time. Because only compressed data is sent to the cloud and buffering takes place on the edge, sliding window approaches make it possible to add a great deal more features than would otherwise be possible.

- ***ML With Reproduced Data:***

There are two actions that need to be taken in order to finish this procedure. Following the completion of the operation for reduction, the data that was transferred will need to be categorised. Obviously, this only becomes relevant when the edge data reduction operations may be reorganised in a different order. Due to the fact that the number of features in this dataset is same to the number of features in the first dataset, this method necessitates the use of more complex models and incurs higher computing costs than ML that operates with fewer data points. However, given that they were created in the first place, the copies of the original data do have the potential to be used in future initiatives. The FFNN technique is utilised to accomplish the classification, despite the fact that there is a much less amount of training data. Additionally, the same strategy that is used with ML to choose the number of layers and neurons in hidden layers is employed here. Because the number of characteristics increases when there are more duplicated versions of the data, the same technique will result in a greater number of layers and neurons than it would have produced if there had been less copies of the data.

6. Analysis and Interpretation

evaluations of precision and recall are typically included for the evaluation in addition to evaluations of accuracy. This is due to the fact that they are utilised extensively in HAR research.

$$\text{Precision} = \frac{TP}{TP + FP}$$

$$\text{Recall} = \frac{TP}{TP + FN}$$

The abbreviations FP and FN stand for "false positive" and "false negative," respectively; the abbreviations TP and TN stand for "true positive" and "true negative," respectively. False positives and false negatives are sometimes combined into a single phrase. When calculating the accuracy of classification for reversible techniques, a data reduction of 66% and a sliding window of 100 points are used. The results are shown in Table 2 below.

TABLE 3: A HIGH LEVEL OF CLASSIFICATION ACCURACY MAY BE ACHIEVED WITH THE USE OF REVERSIBLE TECHNIQUES, A SLIDING WINDOW WITH A VALUE OF 100, AND A DATA REDUCTION OF 66%.

Scenarios	AE		PCA	
	Reduced	Reproduce	Reduced	Reproduce
All Sensors				
Direct	99.28%	99.25%	99%	99.18%
S. Window	100%	100%	100%	100%
Location Based				
Direct	97.15%	98.45%	98.5%	98.38%
S. Window	98.90%	98.93%	98.79%	98.86%
Similarity based				

Direct	97.75%	98.24%	98.32%	98.34%
S. Window	98.90%	98.92%	98.86%	98.85%

It has been shown that the sliding window approach performs better in terms of accuracy than direct strategies. This is quite similar to the way that traditional machine learning works. This is the case regardless of the configuration (whether it be all sensors, location-based, or similarity-based), as well as the algorithm (whether it be AE or PAC), which are both examples of independent variables. The difference between AE and PCA, in addition to the difference between reduced and duplicated approaches, is not one that carries a great deal of weight. Despite a loss in data of 66%, the accuracy was very close to that of traditional cloud-based machine learning. This was an excellent achievement. After additional calculation, the patterns that were shown by precision, recall, and sensitivity were quite similar to those that were displayed by accuracy in Table 2. In contrast to Table 2, which takes into account just reversible processes, this one handles the 66% reduction in the same manner while simultaneously taking into account both reversible and non-reversible methods. The conclusions produced by the vector magnitude technique are much less accurate than those produced by any procedures that can be reversed. This is because the vector magnitude approach does not allow for reversibility. Even though it only takes into account the 100-sliding-window, Table 2 does an analysis of the classification accuracy using a data reduction of 66%.

TABLE 4: PERFORMANCE METRICS FOR DATA OPERATIONS

	Elapsed Time (s)	**Average Read (ms)**	**Average Write (ms)**	**Number Of Rows**	**Total**
(s)					
0.043	0.3	22	2	61900	15481.1
828.547	72	355	28599	41340	17591.5
3.663	3.3	945	109	100000	25264.3
2.424	2.1	118	77	30646	7710.78
1.506	1.6	237	145	100000	25095.9
Total	15.86	335.4	5786.4	66777.2	18228.7

The following metrics are included in the table that can be found above: the total time, the number of rows that were affected, the average amount of time it took to read and write, and the amount of time in seconds that has transpired. The relatively quick operation on the first row was finished in only 0.043 seconds thanks to the average read time of 0.3 milliseconds and the average write time of 22 milliseconds. This surgery had a solitary impact on the first two rows. On the other hand, the total of 15,481.08 seconds for this row is excessive, and more investigation is required as a result. On the other hand, the procedure shown in the second row requires a much longer amount of time and must be completed within 828.5477 seconds. Read and write times for the same number of rows, 28,599, were, on average, 72 and 355 milliseconds, respectively. It came out to a total of 28,599 rows. Once again, the veracity of the assertion that there were a total of 17,591.5 seconds in this string of events is called into doubt. Because the average read and write durations were 3.3 milliseconds and 945 milliseconds, respectively, the third row operation took 3.663 seconds to complete. During the course of this operation, 109 rows were processed, despite the fact that the total period of 25,264.33 seconds seems to be far longer than the time that was actually spent passing. It took 2.424 seconds to get to the fourth row since the average read and write durations were 2.1 milliseconds and 118 milliseconds, respectively. The completion of this task took 7,710.78 seconds, which is about similar to the passage of time. This one movement had a ripple effect that could be felt all the way across 77 rows. When compared to the fifth row, which takes 1.506 seconds to perform an operation, the average read and write operation durations are 1.6 and 237 milliseconds, respectively. Not to add that the process of the fifth row takes a total of 1,506 milliseconds to finish. Although 145 rows were processed, the total length of 25,095.9 seconds is far greater than the amount of time that was actually spent passing.

TABLE 5: SUMMARY STATISTICS FOR PERFORMANCE METRICS

	n	**Mean**	**Median**	**Standard deviation**
Elapsed Time (s)	5	15.86	2.1	31.4
Average Read (ms)	5	335.4	237	363.03
Average Write (ms)	5	5786.4	109	12752.74

Number Of Rows	5	66777.2	61900	32341.19

The summary statistics shown in this table provide a comprehensive overview of the performance metrics by drawing attention to the most important trends as well as the degree of variability or dispersion that can be found within each indicator. It is possible that they will prove to be an invaluable instrument for assessing the efficiency and uniformity of data activities, as well as for locating potential areas for optimisation or improvement.

7. Result and Discussion

During the course of our research into the performance metrics associated with data operations, we compiled summary statistics for four significant parameters. These were the Elapsed Time in seconds, the Average Read Time in milliseconds, the Average Write Time in milliseconds, and the Number of Rows. Each of these times was measured in milliseconds. These statistics, which are based on a sample size of five observations, give essential information on the fundamental patterns and variance in each metric.

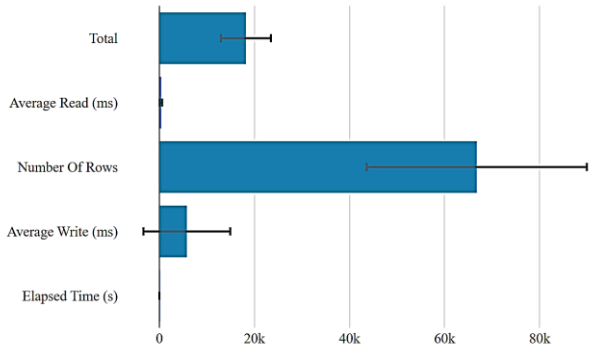

FIG 4: DATA SUMMING UP PERFORMANCE INDICATORS

Metrics, such as those shown in this image, play a crucial role in establishing how effectively a system operates when put to use in actual world conditions. The entire amount of time required for the system to respond may be approximately calculated based on the average amount of time that has elapsed, which is about 15.86 seconds. This figure 4 serves as a starting point for evaluations as well as a benchmark for evaluating the system's effectiveness in managing data operations. In general, a lower number indicates superior performance, while a higher score may indicate that there is room for development in some areas.

The write latency, which varies on average from 5786.4 to 11406.4 milliseconds, and the accompanying standard deviation, which varies from 11406.4 to 5786.4 milliseconds, reflect the consistency and speed of data writing operations. The write latency ranges from 11406.4 to 5786.4 milliseconds. The disparity between the shortest and longest write latencies, which are 2 milliseconds and 28599 milliseconds, respectively, illustrates the range of performance variability that may occur in an application. It is obvious that the system is capable of managing a significant quantity of data given that it can process, on average, 66777.2 rows in one second. On the other hand, drawing emphasis to how long it takes to receive data is done via the use of an average read latency of 335.4 milliseconds. System administrators and software developers need to pay particular attention to this problem since the speed with which data may be retrieved is often critical to the satisfaction of users. The standard deviation for read and write latencies is one metric that may be used to get an idea of how much data access and writing speeds can fluctuate from one instance to the next. By eliminating some of this uncertainty, it could be possible to create a system that is both more predictable and better equipped to react to changing conditions.

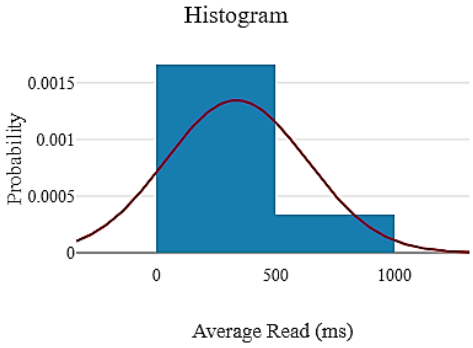

FIG 5: NORMALITY TEST RESULTS FOR DATA DISTRIBUTION

As can be observed in the figure 5 that presents the findings of the normality tests, the dataset in question seems to adhere quite closely to the characteristics of a normal distribution. However, it is vital to keep in mind that normality tests might be sensitive to sample size and other aspects, and that any conclusion should be reached in light of your specific study and the objectives you have set for your research. In other words, it is imperative to keep in mind that normality tests could be sensitive to sample size and other elements. If the assumption that your dataset is normally normal is a crucial presupposition for the statistical tests that you are doing, then our results imply that you may continue to do so with confidence if you are doing what we have shown to be the case.

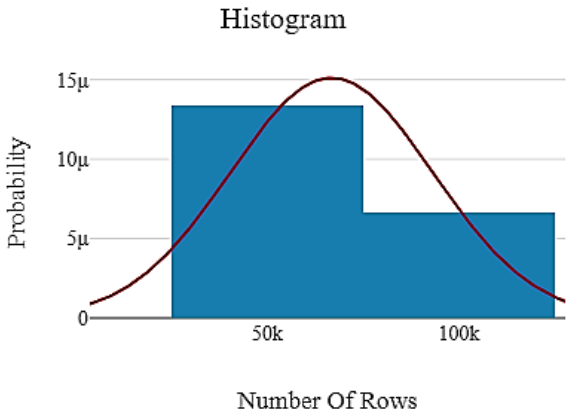

FIG 6: VERIFYING THAT THE ASSUMPTION OF NORMALITY MADE REGARDING THE DATA DISTRIBUTION IS CORRECT

The fact that these normality tests repeatedly produced results that were comparable to one another lends credence to the notion that the dataset under examination is quite near to having a normal distribution (as can be seen in the figure 6). In light of this, it would seem that it is acceptable to presume that normality will be observed when carrying out statistical studies that depend on the premise that normalcy would be observed. If you want to make sure that the normality assumption is acceptable for the work that you are doing, whether it be research or application, you must always make sure to take into consideration the unique context and requirements of your study. You won't know whether or not the assumption was correct until that moment comes around.

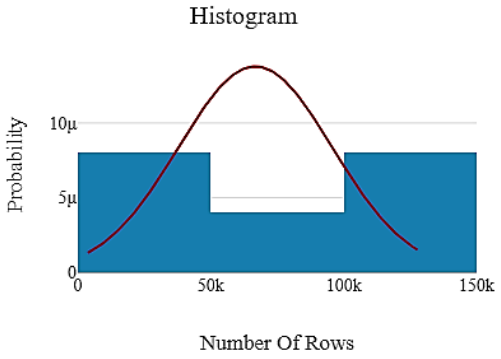

FIG 7: NORMALITY TESTS SUPPORT ASSUMPTION OF DATA NORMALITY

These normality tests demonstrate, time and time again, that the dataset at hand might potentially be properly represented by a normal distribution in the results figure 7. Based on these findings, it seems that doing statistical research based on the assumption of normalcy is something that may be done risk-free. However, if you are determining whether or not the assumption of normality is appropriate for the application or study that you are doing, you should never fail to take into consideration the context and specific requirements of your inquiry. Failing to do so may lead to incorrect conclusions.

8. Conclusions

The technology that enables Creative Commons makes it feasible for there to be a lot of different choices, but there are also a lot of different limits. The processing and storing of data happens remotely, away from mobile devices, at a location referred to as the "cloud." In addition, a brand-new technology that goes by the term Internet of Things (IoT) is making rapid strides in expansion inside the telecommunications market, particularly within the modern wireless telecoms sector. The primary goal of the interaction and cooperation between different objects and items made possible by wireless networks is to enable these entities to perform the duty that has been assigned to them as a single, unified whole. This is being done in order to create an environment that is more suited for the use of big data. In addition, the technology behind wireless networks may be to blame for the rapid development of both cloud computing and the internet of things. In this article, we will discuss the Internet of Things (IoT) as well as cloud computing while putting a strong focus on the many security concerns that are raised by each of these technologies. Combining the aforementioned technologies on purpose is something we do in order to guarantee the safety of both the utilisation and transmission of Big Data. In order to do this, we will investigate the well-known components of both technologies as well as the benefits of integrating them together. This will assist us in understanding how the two of them working together might be beneficial to us. The performance parameters that are analysed in this research provide insights into the degree to which the system is responsive as well as the efficiency with which it processes data. These metrics are available for system administrators and developers to utilise in order to assess and improve their systems, which will ultimately lead to an improvement in user satisfaction and an increase in system reliability. In addition, the results of normality tests carried out on the distributions of the data have repeatedly supported the hypothesis that the data are normal. We now have further proof that statistical analyses that depend on the normality assumption will yield accurate conclusions since we have come to this conclusion. In order to assess whether or not the premise in question is accurate, researchers and analysts should make it a point to make it a habit to constantly take into consideration the specific requirements of the study they are doing.

References

[1] J. Mongay Batalla, P. Krawiec, “Conception of ID layer per-formance at the network level for Internet

of Things", *Springer Journal Personal and Ubiquitous Computing, Vol.18, Issue 2*, pp. 465-480, 2014.

[2] .C. Stergiou, K. E. Psannis, "Recent advances delivered by Mo-bile Cloud Computing and Internet of Things for Big Data ap-plications: a survey", *Wiley, International Journal of Network Management,* pp. 1-12, May 2016.

[3] C. Stergiou, K. E. Psannis, A. P. Plageras, Y. Ishibashi, B.-G. Kim, "Algorithms for efficient digital media transmission over IoT and cloud networking", *Journal of Multimedia Information System, vol. 5, no. 1, pp. 1-10*, March 2018.

[4] A. A. Gnana Singh et al, "A Survey on Big Data and Cloud Computing", *International Journal on Recent and Innovation Trends in Computing and Communication, vol.* 7, no. 4, pp. 273-277, July 2016.

[5] O. Awodele et al, "Big Data and Cloud Computing Issues," *International Journal of Computer Applications, vol.* 12, no. 133, pp. 14-19, January 2016.

[6] J. L. Hernandez-Ramos, M. V. Moreno, J. B. Bernabe, D. G. Carrillo, A. F. Skarmeta, "SAFIR: Secure access framework for IoT-enabled services on smart buildings", *Journal of Computer and System Sciences, vol*. 81, issue: 8, pp. 1452-1463, December 2015.

[7] Bani Ahmad, A. Y. A. ., Kumari, D. K. ., Shukla, A. ., Deepak, A. ., Chandnani, M. ., Pundir, S. ., & Shrivastava, A. . (2023). Framework for Cloud Based Document Management System with Institutional Schema of Database. *International Journal of Intelligent Systems and Applications in Engineering*, *12*(3s), 672–678.

[8] P. William,Anurag Shrivastava, Upendra Singh Aswal, Indradeep Kumar, Framework for Implementation of Android Automation Tool in Agro Business Sector, 2023 4th International Conference on Intelligent Engineering and Management (ICIEM), 10.1109/ICIEM59379.2023.10167328

[9] P. William, Anurag Shrivastava, Venkata Narasimha Rao Inukollu, Viswanathan Ramasamy, Parul Madan, Implementation of Machine Learning Classification Techniques for Intrusion Detection System**,** 2023 4th International Conference on Intelligent Engineering and Management (ICIEM), 10.1109/ICIEM59379.2023.10167390

[10] N Sharma, M Soni, S Kumar, R Kumar, N Deb, A Shrivastava, Supervised Machine Learning Method for Ontology-based Financial Decisions in the Stock Market, ACM Transactions on Asian and Low-Resource Language Information Processing.

[11] Ajay Reddy Yeruva, Esraa Saleh Alomari, S Rashmi, Anurag Shrivastava, Routing in Ad Hoc Networks for Classifying and Predicting Vulnerabilities, Cybernetics and Systems, Taylor & Francis, 2023

[12] P William, OJ Oyebode, G Ramu, M Gupta, D Bordoloi, A Shrivastava, Artificial intelligence based models to support water quality prediction using machine learning approach, 2023 International Conference on Circuit Power and Computing Technologie

[13] J Jose, A Shrivastava, PK Soni, N Hemalatha, S Alshahrani, CA Saleel, An analysis of the effects of nanofluid-based serpentine tube cooling enhancement in solar photovoltaic cells for green cities, Journal of Nanomaterials 2023

[14] K Murali Krishna, Amit Jain, Hardeep Singh Kang, Mithra Venkatesan, Anurag Shrivastava, Sitesh Kumar Singh, Muhammad Arif, Deelopment of the Broadband Multilayer Absorption Materials with Genetic Algorithm up to 8 GHz Frequency, Security and Communication Networks

[15] P Bagane, SG Joseph, A Singh, A Shrivastava, B Prabha, A Shrivastava, Classification of malware using Deep Learning Techniques, 2021 9th International Conference on Cyber and IT Service Management (CITSM).

[16] A Shrivastava, SK Sharma,Various arbitration algorithm for onchip (AMBA) shared bus multi-processor SoC, 2016 IEEE Students' Conference on Electrical, Electronics and Computer Science, SCEECS 509330

[17] A. Gandomi, M. Haider, "Beyond the hype: Big data concepts, methods, and analytics", *International Journal of Information Management,* vol. 35, no. 2, pp. 137-144, 2015.

[18] N. Kaur, S. K. Sood, "Dynamic resource allocation for big data streams based on data characteristics (5Vs)", *International Journal of Network Management, vol*. 27, issue 4, May 2017.

[19] A. Alexandrov, R. Bergmann, S. Ewen, J. C. Freytag, F. Hues-ke, A. Heise, A., F. Naumann, "The Stratosphere platform for big data analytics", *The VLDB Journal, vol*. 23, no. 6, pp. 939-964, 2014.

[20] O. Kwon, N. Lee, B. Shin, "Data quality management, data usage experience and acquisition intention of big data analyt-ics", *International Journal of Information Management, vol.* 34, no. 3, pp. 387-394, 2014.

[21] Shrivastava, A., Chakkaravarthy, M., Shah, M.A..A Novel Approach Using Learning Algorithm for Parkinson's Disease Detection with Handwritten Sketches. In Cybernetics and Systems, 2022

[22] Shrivastava, A., Chakkaravarthy, M., Shah, M.A., A new machine learning method for predicting systolic

and diastolic blood pressure using clinical characteristics. In *Healthcare Analytics*, 2023, 4, 100219

[23] Shrivastava, A., Chakkaravarthy, M., Shah, M.A.,Health Monitoring based Cognitive IoT using Fast Machine Learning Technique. In *International Journal of Intelligent Systems and Applications in Engineering*, 2023, 11(6s), pp. 720–729

[24] Shrivastava, A., Rajput, N., Rajesh, P., Swarnalatha, S.R., IoT-Based Label Distribution Learning Mechanism for Autism Spectrum Disorder for Healthcare Application. In *Practical Artificial Intelligence for Internet of Medical Things: Emerging Trends, Issues, and Challenges*, 2023, pp. 305–321

[25] Boina, R., Ganage, D., Chincholkar, Y.D., .Chinthamu, N., Shrivastava, A., Enhancing Intelligence Diagnostic Accuracy Based on Machine Learning Disease Classification. In *International Journal of Intelligent Systems and Applications in Engineering*, 2023, 11(6s), pp. 765–774

[26] Shrivastava, A., Pundir, S., Sharma, A., ...Kumar, R., Khan, A.K. Control of A Virtual System with Hand Gestures. In *Proceedings - 2023 3rd International Conference on Pervasive Computing and Social Networking, ICPCSN 2023,* 2023, pp. 1716–1721

[27] A. P. Srivastava, P. Choudhary, S. A. Yadav, A. Singh and S. Sharma, A System for Remote Monitoring of Patient Body Parameters**,** International Conference on Technological Advancements and Innovations (ICTAI), 2021, pp. 238-243,