Asian Journal of Computer Science and Technology
ISSN: 2249-0701 (P) Vol.11 No.1, 2022, pp.35-39

DOI: https://doi.org/10.51983/ajcst-2022.11.1.3291

Techniques for Diabetes Care Using Artificial Intelligence and Machine Learning: A Review

Ajit R. Patil and Avinash M. Ingole
Department of Computer Engineering, Bharati Vidyapeeth's College of Engineering Lavale, Pune, Maharashtra, India
E-mail: patilajit667@gmail.com, ingoleavi20@gmail.com
(Received 5 February 2022; Revised 20 February 2022; Accepted 10 March 2022; Available online 19 March 2022)

***Abstract* - All aspects of our lives, including healthcare, are being reshaped by AI/ML (Artificial Intelligence/Machine Learning). Diabetic treatment might benefit greatly from the use of AI and ML, which could make it more effective and less time-consuming. In terms of data availability, the large number of diabetics in India brings a unique set of challenges, but it also gives an opportunity. With the use of electronic medical records, India may become a world leader in this field. The use of AI/ML might shed light on our issues and help us come up with solutions that are unique to each.**
***Keywords:* Diabetes Care, Artificial Intelligence, Machine Learning, Techniques**

I. INTRODUCTION

Diabetes Mellitus (DM) is one of the most frequent ailments among the elderly in the United States (World Health Organization, 2020). As of 2017, there were 451 million individuals worldwide with diabetes according to the International Diabetes Federation (IDF). More than 693 million people are predicted to populate our planet in the next 26 years. It is not understood what causes diabetes, although environmental and genetic variables are thought to have a role. Even though it is untreatable, it may be treated with medications and therapies. Diabetics are at risk for additional health issues, such as cardiac arrest and organ damage. Diabetes Mellitus (DM) diagnosis and treatment may also help avoid complications and minimize the risk of major health issues.

The diagnosis of DM may be performed either by a doctor manually or by an automated tool. DM measurement may be done using any one of these approaches. There are several advantages to manual diagnosis, including the fact that it doesn't rely on machines, allowing doctors to focus on their specialty. Even a doctor with years of expertise may not be able to identify all of the symptoms of DM in its earliest stages. An early diagnosis of a disease thanks to advances in machine learning and artificial intelligence (AI) is more probable than a manual DM detection and diagnosis. The risk of human mistake will be lowered since there will be less work for medical professionals to undertake. Effective diagnosis and profitable management may benefit from the use of computer-based decision support systems. There are several sources of data in the DM sector, including as test results, patient reports, treatment and follow-ups, and medication. Compiling all of the data by hand is time-consuming and error-prone. Data organization has deteriorated as a result of poor data management. A more efficient and effective way of extracting and analyzing data is required as the amount of data grows. These devices are used in modern and new hospitals to facilitate the gathering and distribution of massive amounts of data. Automated identification and diagnosis of DM and anomalies is significantly easier and more reliable than human detection and diagnosis. The diagnosis of diabetes mellitus must be automated as a consequence. Automated DM systems may be created using machine learning or artificial intelligence

There are benefits and cons to each ML and AI method. As a consequence, methods for the automatic identification of DM have been developed using both techniques. To perform like a human, AI and ML-based strategies need castability and explainability since many of the best-performing Ml and AI technologies are the least transparent (Holzinger *et al.,* 2019). Physicians have a greater sense of trust in AI systems that are able to explain themselves. Using a causal model, the causability is assessed for efficiency, efficacy in terms of causal knowledge, and user transparency. Machine learning and AI approaches have been widely used by academics in recent years to improve DM control, self- management, and personalization.

II. BIG DATA

Often referred to as SMAC, Big Data is a mash up of social networking, Mobile computing, Analytics, and Clouds. Many areas of study, such as scientific and industrial research, generate a great deal of data. Using traditional technologies, it's difficult to store and analyze such quickly increasing data within the specified timeframe. Data management and analysis systems that can effectively handle the influx of data from a wide range of scientific and commercial applications are in high demand. A company may generate, modify, and manage massive data volumes by using tools, methods, and procedures associated with big data. It's difficult to process huge amounts of Big Data on a human inspection scale, therefore a high-speed system is needed to manage such Big Data. The usage of various Big Data technologies is now being utilized to analyze data more quickly and provide users with the information they need for decision-making and forecasts.

Older data warehouses collect stale data, which is then cleaned up and made current. The Big Data system, on the other hand, includes raw data culled from a variety of sources, including reports, website monitoring, and in-the-moment information. Analytics make advantage of Big Data stored in repositories and computer infrastructure. Other database systems, such as billing, point-of-sale, consumer marketing, and the company's finance systems, are used only to collect conventionally derived data. Big Data systems rely on a variety of sources, including sensor 4 data, e-mail, mobile device data, and other comparable data. Big Data analytics provides more comprehensive solutions to increasingly challenging issues.

III. BIG DATA IN HEALTH CARE

Data volumes have risen dramatically in recent years as a result of the widespread use of digital technologies. In addition to the widespread availability of high-speed mobile networks and the extensive digitization of patient information, a variety of variables contribute to the rapid growth of health care data. One of the most important components in healthcare Big Data is a wide range of data sets. Doctors' prescriptions and images from hospitals are among the sources of these records. It is called the Big Data challenge when researchers try to manually analyze and interpret enormous amounts of complex data. In the healthcare industry, data management is essential. Big data analytics in healthcare is very beneficial to both patients and physicians since it assists in the proper treatment of patients. The four V's - volume, velocity, variety, and veracity - provide challenges when dealing with vast volumes of data. Diabetes patients, in particular, will provide a wealth of information.

With so much healthcare data available, you'll need tools and procedures that can manage massive volumes of data efficiently and reliably. Academics have previously shown that healthcare Big Data may be used to predict disease outbreaks. Large businesses are taking a hard look at the issues of coping with large volumes of data. Health care advancements may be made through gaining a better understanding of physiological processes via the collecting of important data.

A number of Big Data technologies are now being used, but the focus is on enhancing existing Big Data tools and methodologies such as MapReduce in order to analyze and predict diabetic diseases based on vast volumes of gathered data. In addition to Apache Hadoop, other Big Data tools exist, like RStudio and AWS web services. Value-added results may be achieved via the use of big data. Modern medical and health-care services will be provided with the aid of high-performance computers, the Cloud, and Big Data technology. Many data mining and statistical tools are used in predictive analysis, together with both current and historical data, while making predictions about the future. Predictions based on Big Data analytics may be made with great accuracy in the health care system.

IV. BIG DATA ANALYTICS

Dr. notes, prescriptions, clinical reports, machine reading and body sensors like IOT devices generate a large amount of data every day, which becomes Big Data. For healthcare workers, Big Data isn't a big obstacle. It's a struggle to get meaningful information from Big Data. In the healthcare industry, big data analytics will have a major impact. Clinicians may use Big Data analytics on medical data to predict possible outcomes, such as whether a patient will get successful therapy or not. Cloud-based Big Data analytic environments are now required for the analysis of healthcare sector Big Data that is semi-structured, structured, and unstructured.

A. Predications Based on Historical Data

Predictive analytics are becoming important for identifying chronic illnesses before they manifest themselves. The patient's future health state may be predicted using his or her present health indicators. With the use of the MRK-SVM algorithm in Hadoop, large datasets can be processed more quickly, and illnesses such as diabetes may be predicted using current lifestyle characteristics. Elastic Hadoop parallelization in a distributed context is thus critical for implementing the most recent technologies.

V. CLOUD COMPUTING

Because it benefits both present and future generations, the cloud is a valuable resource. IT and scientific professionals rely on computing infrastructure platforms to do their jobs. The internet has caused massive data expansion known as Big Data in a variety of scientific and technical fields. Unstructured data is used to store information gleaned from the internet and scientific research.

Service providers in the data centers provide hardware and software. For parallel data processing in a dispersed environment in a short time and at a low cost, cloud service providers provide Cloud computing services over the Internet. Using Cloud computing, massive amounts of data may be transparently stored, collected, and transferred. When working with huge amounts of data in a distributed Cloud computing environment, MapReduce is a popular choice since it hides the complexities of simultaneous execution over many servers.

In order to store and process large amounts of data using the MapReduce system on the Amazon Cloud, Amazon Web Services (AWS) offers a remote service called EMR, Elastic Map Reduce (EMR). In this research, computation and data storage are handled by Elastic Compute Cloud (EC2) and Simple Storage Service (S3), respectively. EC2 is a Cloud computing platform based on the internet that offers computational capacity. S3 offers an easy-to-use web services interface for storing and retrieving information. As a result, Amazon Web Services (AWS) is a collection of remote infrastructure services that together form the Cloud.

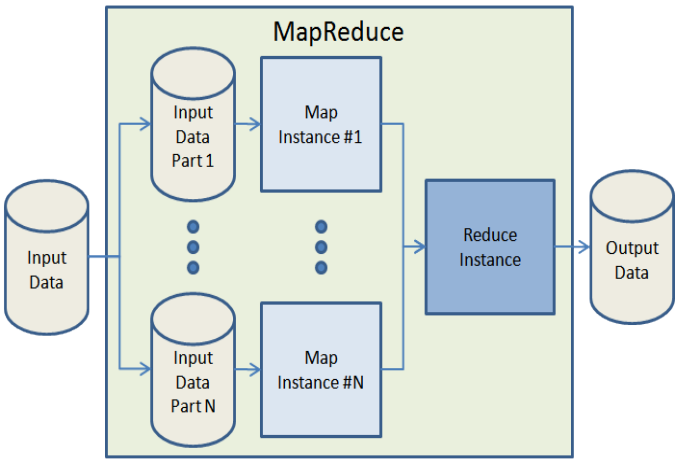

Fig. 1 Amazon Elastic Map Reduce (EMR)

Figure 1 above illustrates an Amazon Web Services (AWS) architecture for large data processing. MapReduce is used for large-scale data processing, and this article describes how it integrates with Cloud infrastructure. Amazon EMR offers a Hadoop framework for Amazon EC2 to run large amounts of data quickly and affordably. Internet-based computing is made simpler for developers thanks to Amazon S3, a storage service on the internet. Amazon EC2 is used for large-scale distributed processing in parallel computing. In the Cloud, it's a web-based service that offers computational power. It is possible to just pay for the capacity that has been utilized on this web-based service platform.

VI. HADOOP MAPREDUCE

Extraction of valuable information from huge quantities of data necessitates a lot of computational power (fast execution speed) and storage space (memory). The use of many CPUs or CPU cores to perform various calculations at the same time, known as parallelism or distributed computing, speeds up computations. The basic goal of parallel and distributed computing is to divide big issues into smaller problems that can be simultaneously, even if they are not identical. Execution speed, memory, and concurrency all play a role in parallelization. Execution speed may be increased by utilizing pipelining or numerous arithmetic logic units to parallelize operations inside a processor, or by having several processors, each of which works on a separate portion of the issue in the background. Parallel or distributed processing may take use of the memory capacity of many computers to solve problems with processing huge data sets on a single system.

To make sense of massive quantities of data, you'll need fast execution and enough of storage capacity. It divides the data into thousands of parts and runs them in parallel on a huge number of computers. Large amounts of data are processed quickly using Hadoop's HDFS and MapReduce, while little amounts of data are stored using MapReduce.

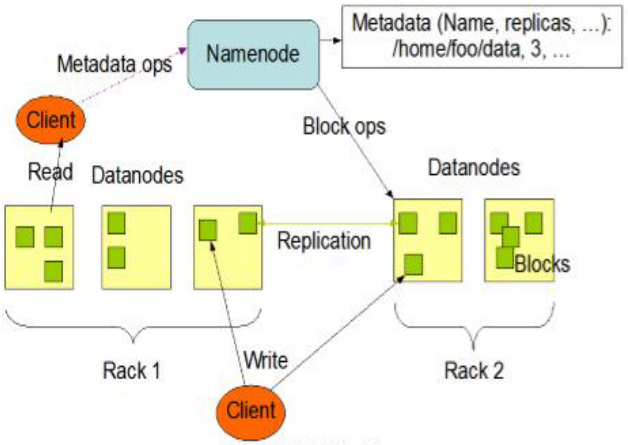

Fig. 2 HDFS Architecture

There are blocks of a predetermined size for each file in the Hadoop Distributed File System (HDFS) shown in the above Figure 2. One or more computers in a cluster store these blocks. It is based on a Master/Slave Architecture, with a single Name Node (Master node) and all other nodes being Data Nodes (Slave nodes). Since these facts have been established, the current study has used the Hadoop platform to organize Big Data in order to address the issue.

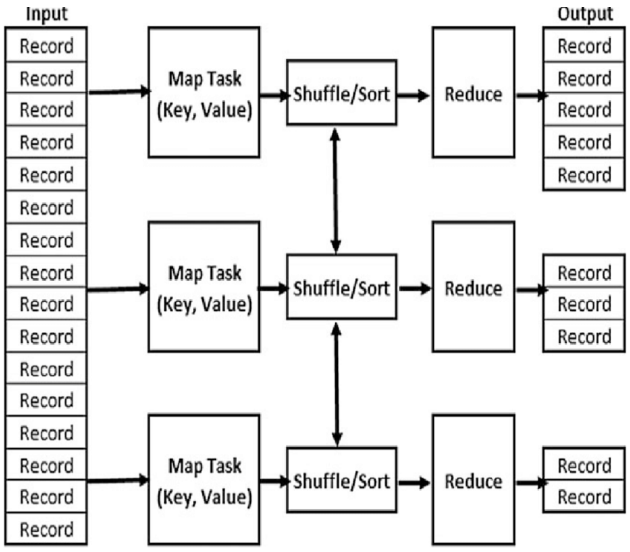

Fig. 3 The MapReduce programming model

Figure 3 depicts the MapReduce paradigm in action, showing the phases of Map and Reduce. The master node splits each dataset into smaller issues and distribute them to each of the worker nodes in the "Map" phase; this process is called "Mapping." This may be repeated by a worker node many times in a multi-level tree structure. As soon as the worker node has finished processing the smaller issue, it returns its findings to the master node for review. It was initially attempting to solve the job process in a more straightforward way when it goes through "Reduce" step: the master node gathers all the solutions to the sub-problems and merges them into one output.

A. Apache Spark

Apache Spark is a fast and dependable general- purpose cluster computing engine. Data can be queried considerably quicker with Spark thanks to its capability for in-memory processing. Spark's main feature is cluster processing and cluster computing in memory. As a result, it'll be faster because of it. It may be used in the cloud or on a standalone MapReduce cluster. Using Spark has two distinct benefits. The primary benefit is speed, which is up to 100 times quicker than MapReduce in performing jobs. The developer-friendliness of Spark is its second selling point. Because of these two major benefits, Spark has supplanted MapReduce as the framework of choice for processing large amounts of big data. Various programming languages, such as Java, Python, and Scala, are supported by this system.

B. R Programming Language

In order to do statistical analysis on extremely big data sets, one may use RStudio, a free and open-source program. RStudio is critical for data analytics since it makes it easier to understand complex datasets. As a result, it's perfect for data mining applications like graphics and software development. To do graphical analysis, the RStudio software connects to the RStudio Server, which is built in C and utilizes the Qt framework. It gathers and runs on a variety of operating systems, including Windows, UNIX, and others, and generates stunning graphs, mathematical symbols, and formulas. Extensibility and data visualization are two critical factors in the success of RStudio. R covers statistical applications in scientific study. Statistical testing, classification, regression, and clustering are among the techniques provided by R. R is used by the majority of statisticians because of its versatility and ease of use. R's biggest drawback is that it is memory- and single-threaded-limited. There are no explicit parallelization constructs in the R language. Because the whole calculation is performed from the computer's main memory, R demands that the full data set fit in RAM.

C. Data Mining and Knowledge Discovery

The goal of the current study is to use data mining to uncover important information from a bigger data collection. In this process, the data is prepared and selected, the data is cleaned, the previous knowledge is incorporated into the data, and solutions are drawn from the observed findings. Discovering meaningful information from vast amounts of data is a task for Knowledge Discovery Databases (KDD). An essential stage in this method is the extraction of data via data mining. As a result, the end goal of this procedure is to extract meaningful information from unstructured data. Data mining makes it possible to find previously undiscovered nuggets of information in huge repositories. Using KDD, you may glean valuable information from a large data collection.

VII. MACHINE LEARNING TECHNIQUES

Automated data set learning is replaced by machine learning techniques. It begins by learning the information and then uses that information to make predictions. There are three types of machine learning algorithms: supervised, semi-supervised, and unsupervised. With the inclusion of new

variables, the model's ability to differentiate between patients with and without illness problems increased significantly. This is because machine learning algorithms could examine many more elements in patient charts than physicians.

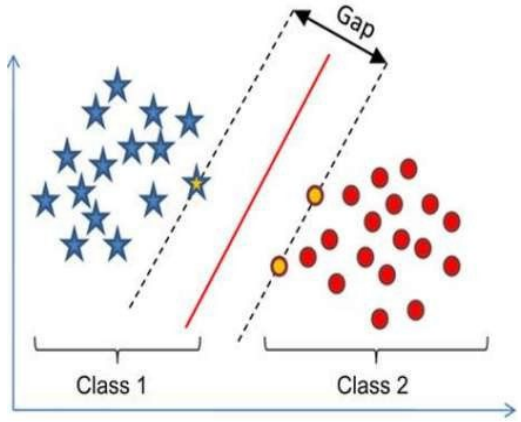

Fig. 4 Support Vector Machine

Support Vector Machine (SVM) is a well-known method for data categorization and regression. For classification and regression, SVMs are supervised learning models that evaluate data using related learning algorithms. To put it another way, Support Vectors are just the individual observations' co- ordinates. One of the finest frontiers for separating the two classes (hyper-plane and line) is the Support Vector Machine. Unsupervised learning techniques such as K-means clustering are employed when there are no labels for the data (i.e., data without defined categories or groups). This algorithm's aim is to identify groups of data, where K represents the number of groups to locate. The method uses an iterative process to allocate each data point to one of K groups depending on the given characteristics. Data points are grouped together depending on how comparable their features are. For data classification, nothing beats using this technique.

A method known as the MRK-SVM algorithm has recently been suggested as a dependable and stable model for dealing with datasets of various sizes. It speeds up the processing of large amounts of data. This hybrid approach integrates Hadoop's clustering and classification methods to handle Big Datasets effectively. This two-phase hybrid approach utilizes K-means clustering as the first step in finding and eradicating instances that were wrongly categorized. Next, a fine-tuned classification is done using SVM, utilizing the right-clustered occurrence from the previous step as a starting point. This hybrid model accurately predicts the onset of chronic illnesses in high-risk diabetic patients while also ensuring that patients get prompt treatment.

VIII. CONCLUSION AND FUTURE SCOPE

Based on the findings of the evaluation of literature and the current research, artificial intelligence and machine learning techniques have practically limitless uses in the healthcare sector. AI and machine learning are now being used to assist hospitals streamline administrative procedures, customize medical care, and cure infectious illnesses. Data science research has the potential to enhance diabetes mellitus diagnosis and diabetes type prediction, which is beneficial to both medical practitioners and patients. The process of developing a machine learning-based model for diabetes detection saves time.

REFERENCES

[1] T. Chardonnens, "Big Data analytics on high velocity streams Specific use cases with Storm eXascale‖," *Information Lab Benoit Perroud Veri Sign Inc,* 2013.

[2] C. Ramirez, M. Nagappan and M. Mirakhorli, "Studying the impact of evolution in R libraries on software engineering research. Software Analytics (SWAN)," *IEEE 1st International Workshop,* pp. 29-30, 2015.

[3] N. Kitcharoen, S. Kamolsantisuk, R. Angsomboon and T. Achalakul, "Rapid Miner framework for manufacturing data analysis on the Cloud," *Computer Science and Software Engineering (JCSSE),* 10th International Join Conference, pp. 149-154, 2013.

[4] Z. A. Vale, C. Ramos, S. Ramos and T. Pinto, "Data mining applications in power systems: Case-studies and future trends," *Transm Distrib Conf Expo Asia Pacific*, pp. 1-4, 2009. DOI: 10.1109/ TDASIA.2009.5356830.

[5] Johann M. Kraus and Hans A. Kestler, "A Highly Efficient Multi-Core Algorithm for Clustering Extremely Large Datasets," *BMC Bioinformatics,* Vol. 11, No. 169, 2010.

[6] Kanchan M. Tarwani, Saleha S. Saudagar and Harshal D. Misalkar, "Machine Learning in Big Data Analytics: An Overview", *International Journal of Advanced Research in Computer Science and Software Engineering,* Vol. 5, No. 4, pp. 270-274, 2015.

[7] Stephen Kaisler, Frank Armour, J. Alberto Espinosa and William Money, "Big Data: Issues and Challenges Moving Forward", *46th Hawaii International Conference, IEEE,* pp. 995-1004, 2013.

[8] Sadhana and Savitha Shetty, "Analysis of Diabetic Data Set Using Hive and R", *International Journal of Emerging Technology and Advanced Engineering,* Vol. 4, No. 7, 2014.

[9] S. Vikram Phaneendra and E. Madhusudhan Reddy, "Big Data-solutions for RDBMS problems - A survey," *In 12th IEEE/IFIP Network Operations & Management Symposium,* 2013.

[10] Kiran kumara Reddi and Dnvsl Indira, "Different Technique to Transfer Big Data: Survey," *IEEE Transactions,* Vol. 52, No. 8, 2013.