Section A-Research paper

Deep Learning on Health Care Big Data Using Apache Spark

Dr G JayaLakshmi,Ph D
Assistant Professor, Department of IT
Velagapudi Ramakrishna Siddhartha Engineering College
Vijayawada, AP, India
jaya1123@vrsiddhartha.ac.in

Medasani Poojitha
Student, Department of IT
Velagapudi Ramakrishna Siddhartha Engineering College
Vijayawada, AP, India
poojimedasani25@gmail.com

Kolasani Sai Sri Lekha
Student, Department of IT
Velagapudi Ramakrishna Siddhartha Engineering College
Vijayawada, AP, India
srilekha.kolasani@gmail.com

Palaparthi Naga Raghavendra
Student, Department of IT
Velagapudi Ramakrishna Siddhartha Engineering College
Vijayawada, AP, India
nagaraghavendra1@gmail.com

***Abstract*— Globally speaking, stroke is a significant health problem. It has stayed the second-leading cause of death in the world since 2000. After the first two major causes of impairment, stroke is the third. Long-term disability has a substantial negative impact on people's capacity to lead productive lives. It is one of the major causes of extreme, ongoing weakness in all countries. Therefore, stroke represents a significant danger to global health. We use machine learning methods on the Healthcare Dataset Stroke to forecast strokes in this. This research utilizes the big data platform Apache Spark. Apache Spark, one of the most popular big data platforms, includes the MLlib library to handle enormous amounts of data. Apache Spark contains an MLlib library to manage huge data. MLlib, an API associated with Spark, offers machine learning methods. MLlib, an API associated with Spark, offers machine learning methods. A Multilayered Perceptron and Decision Trees are used to build the stroke prediction algorithm. These algorithms can help people and healthcare providers focus on health risks and changes in health status, which will eventually improve quality of life**

Keywords Heart Stroke, Prediction, Machine Learning, Big Data, Apache spark, Deep Learning.

I. INTRODUCTION

One of the most important risks to public health today is stroke. After heart disease, stroke is the condition that shortens life the most. The abrupt onset of localized neurological impairments that continue longer than 24 hours is known as a stroke. And atherosclerosis or cerebral artery blockage are the causes. Stroke symptoms might appear suddenly, although they frequently develop gradually. The number of stroke patients in the United States increased dramatically in 2016, placing a significant burden on the healthcare system. Early detection is thought to improve healing and lessen disabilities in stroke victims, their families, and the community, whereas long-term disabilities place a physical, mental, and financial burden on them. Early stroke illness prediction is useful for prevention or early therapeutic action. Data mining and machine learning are crucial to the prediction of stroke. Support vector machines, logistic regression, classifiers based on random forests, and neural networks are a few examples. Artificial intelligence known as "machine learning" tries to give computers the ability to think like people. [1] The purpose of machine learning is to enable computers to perform a specific activity using patterns and interference rather than utilizing explicit instructions. Big data is a term used to describe big and complicated data sets that cannot be handled by conventional analysis techniques. Big data refers to large and complex data sets that can't be processed by traditional analytic methods.

II. PRELIMINARIES

This part explains the fundamental ideas and terms used in this essay.

2.1. Heart Stroke

When plaque fragments break loose, blood clots may form, preventing blood flow to the heart. This deprives the heart muscle of the oxygen and nutrients it needs, which may lead to some cardiac tissue degenerating or even dying. This is a heart attack, also known as a myocardial infarction.

2.2. Prediction

Prediction essentially means creating future predictions, much like machine learning does. It depicts the outcomes of an algorithm that has been trained on some old data. By studying the historical data set, prediction allows one to foresee the outcome of a new data collection. For variables with unknown values, the method generates probable values. Prediction is used to make a form suit the data as closely as possible.

2.3. Machine Learning

Machine learning is a branch of computer science and artificial intelligence (AI) that concentrates on simulating human learning by using data and algorithms to improve the system's accuracy over time.

2.4. Big Data

Big Data is an enormous body of knowledge that is constantly growing exponentially. Due to its size and complexity, no typical data management system can store or process this data efficiently. Big data is a category of exceedingly large data.

2.5. Deep Learning

An artificial neural network, a type of advanced machine learning algorithm, is the foundation of the bulk of deep learning models. The terms deep neural learning and deep neural networking are thus also used to refer to deep learning.

2.6. Decision Tree

One kind of machine learning algorithm used for both classification and regression jobs is the decision tree. Each decision point is represented by a node in the tree, and each potential outcome is represented by a branch or path leading from that node. It is a graphical representation of a set of choices and their potential outcomes.

2.7. MLP

An artificial neural network called a multilayer perceptron (MLP) is made up of numerous layers of interconnected nodes, or neurons. It's a supervised learning method that works well with complex, non-linear relationships between input features and goal variables. It can be applied to both classification and regression tasks.

III. Related Work

The primary focus of this chapter is on the resources that assisted us in comprehending the categorization techniques study field's ideology. The study articles covered the various frameworks and algorithms that could be used to categorize data in great depth.

Sehaa et al., 2020 [1] proposed a big data analytics tool that uses Arabic-language Twitter statistics for healthcare in the Kingdom of Saudi Arabia (KSA). To identify different diseases in the KSA, Sehaa employs Naive Bayes, Logistic Regression, and multiple feature extraction techniques. The top five illnesses in Saudi Arabia, according to Sehaa. More needs to be done in Jeddah and Riyadh to raise consciousness of the major diseases. Since Sehaa is built on top of Apache Spark, it offers real scalability. 18.9 million tweets were gathered between November 2018 and September 2019 and made up the collection. The outcomes are assessed using wellknown numerical criteria (Accuracy and F1-Score) and validated using statistics that are publicly accessible.

Elham et al., 2019. [2] proposes that depending on the requirements, different Big Data analysis and processing tools are used. In other words, it can be said that each technology is complementary, each of which is applicable in a specific field and cannot be separated from one another, and depending on the purpose and the expected expectation, and depending on whether custom tools are designed on these platforms, the platform must be chosen for analysis.

Eman et al., 2019 [3] demonstrated big data, as it has been proven, is a term that describes an excessively large number of datasets that are used to computationally reveal patterns and trends. A processing framework is needed to analyse and extract information from this vast amount of data. There are many different kinds of big data platforms that are frequently used, including Apache Hadoop, Apache Storm, Apache Spark, and Apache Flink. In this article, we discuss the capabilities of Apache Spark for batch and stream processing, use cases, the ecosystem, the architecture, multi-threading and concurrency, and finally, the application of Spark in emerging technologies.

Tahanii et al.,2020 [4] proposed that big data analytics is a cutting-edge field that has demonstrated the ability to make predictions in the healthcare industry. Gaining insightful knowledge from such exceptional and rich data to assist decisionmaking was looked at. Machine learning techniques, which have recently attracted a lot of interest and are extremely important in this age of health data, can be used to provide such value. This paper's addition is to investigate the risk factors for no-shows and to categorize patients in outpatient clinics according to this risk. Present an analysis of five machine learning methods using the Spark platform to forecast patient no-shows. It is difficult to assess the dangers involved and anticipate no-shows. This approach can be applied to enhance care accessibility and clinic resource utilization.

Sujitha et al., 2020 [5] has proposed that Apache Spark has proven to be cutting edge in the big data business, even performing better in ideal conditions. Nevertheless, Pyspark performs better and handles data collections at a range of gigabytes. This study introduces a hybrid methodology to categorise the nodule levels.The technique uses T-BMSVM along with binary classification to differentiate between benign and malignancy, and multiclass classification would be useful to differentiate the stages of lung cancer even better.

IV. SYSTEM ARCHITECTURE

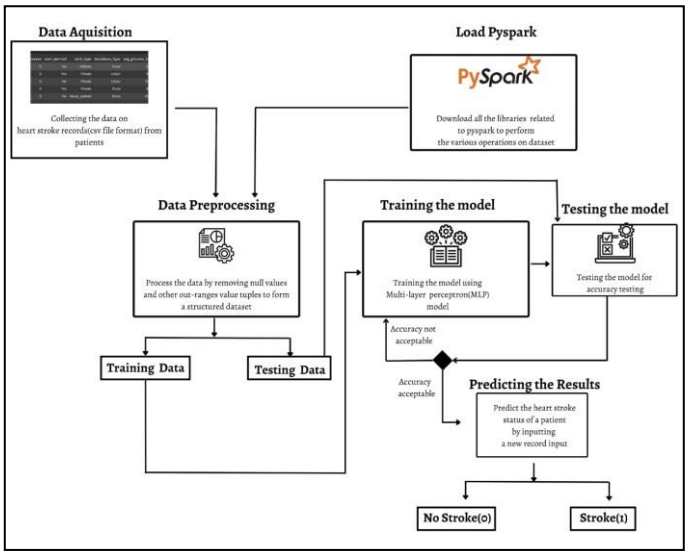

Figure 1. System Architecture

V. DESIGN METHODOLOGY

The dataset is first pre-processed before being divided into train and validation datasets. The dataset is collected from various web sources. using Spark MLlib to build a decision tree and MLP model. With our training dataset, models were trained. With test data, sometimes referred to as the validation dataset, we were validating our model. The model's accuracy is then determined. Lastly, we tested our algorithm using fresh data to forecast stroke.

5.1.1 Decision Tree

Step 1: Choose the most appropriate characteristic to divide thedata.

The best attribute to use to divide the data is chosen as the first stage in creating a decision tree algorithm. The best attribute to split the data is chosen to have the greatest information gain. The information gain quantifies the amount by which the dataset's entropy falls after being divided based on a specific attribute.

Where:

Entropy(parent) is the entropy of the original dataset

Entropy(children) is the entropy of the subsets created by splitting the data based on a particular attribute

[weighted average] is the proportion of the samples thatbelong to each subset

Step 2: Create a node for the selected attribute

Once the best attribute to split the data has been selected, a node is created for that attribute. The node represents a decision point in the decision tree. The attribute is stored in the node and the dataset is split based on the values of that attribute.

Step 3: Recursively repeat the above steps

After creating a node for the selected attribute, the algorithm recursively repeats the above steps for each subset of the data created by splitting the data based on the selected attribute. When there are no more attributes to divide the data into subsets or when all the instances in a subset pertain to the same class, the algorithm terminates.

$$\textbf{Entropy(parent)} = -\sum \textbf{(p\_i * log2(p\_i))}$$

where:

The percentage of cases in the parent dataset that are members of class i is known as p_i.

$$\textbf{Information Gain(attribute)} = \textbf{Entropy(parent)} - \sum \textbf{[(proportion of instances in the subset) * Entropy(subset)]}$$

where:

Entropy(subset) is calculated using the same formula as entropy, but using only the instances in the subset instead of the entire dataset

5.1.2.Multilayered Perceptron

Artificial neural networks, such as the multilayered perceptron (MLP) algorithm, are frequently employed for classification jobs. The algorithm divides incoming data into a number of distinct classes using a feedforward neural network with many hidden layers. The MLP algorithm's stages, with mathematical calculations, are listed below.

Step 1: Initialize the weights and biases

The first step in the MLP algorithm is to initialize the weights and biases for each neuron in the network. The weights and biases are randomly initialized to small values.

Step 2: Feedforward calculation

The input data is passed through the network and the output is calculated in a feedforward computation. Each neuron in the network takes inputs from the neurons in the layer above, weights the inputs, adds a bias term, and then applies an activation function to the outcome. The neurons in the subsequent stratum receive the neuron's output as input.

The feedforward calculation can be represented mathematically as follows:

$$\textbf{z\_j} = \sum \textbf{(w\_ij * x\_i) + b\_j}$$

$$\textbf{a\_j} = \textbf{f(z\_j)}$$

where:

z_j is the weighted sum of the inputs to neuron j

w_ij is the weight connecting neuron i in the previous layer to neuron j

x_i is the input to neuron i in the previous layer b_j is the bias term for neuron j

f(z_j) is the activation function applied to the weighted sum of the inputs to neuron j

a_j is the output of neuron j

Step 3: Calculate the error

By contrasting the network's expected output with the actual output for the training data, the error is determined. Different loss functions, such as mean squared error or cross-entropy loss, can be used to determine the error.

Step 4: Backpropagation

The weights and biases in the network are updated using backpropagation to reduce inaccuracy. The gradient of the error with regard to the weights and biases is used to adjust the weights and biases after the error has been propagated backwards through the network. The chain formula of calculus is used to determine the gradient.

The backpropagation algorithm can be represented mathematically as follows:

$$\delta_j = f'(z_j) * \Sigma (w_{jk} * \delta_k)$$

$$\Delta w_{ij} = \alpha * \delta_j * x_i$$

$$\Delta b_j = \alpha * \delta_j$$

where:

δ_j is the error term for neuron j

$f'(z_j)$ is the derivative of the activation function applied to the weighted sum of the inputs to neuron j

w_{jk} is the weight connecting neuron j to neuron k in the next layer

δ_k is the error term for neuron k in the next layer

α is the learning rate, which controls the step size in the weight and bias updates

Δw_{ij} is the change in weight for the connection between neuron i and neuron j

x_i is the input to neuron i in the previous layer Δb_j is the change in bias for neuron j

Step 5: Repeat

The feedforward and backpropagation steps are repeated for a fixed number of epochs or until the error converges to a minimum value.

VI. DATASET DESCRIPTION

The dataset was gathered from the Kaggle website and consisted of 10 measures for a total of 86,801 individuals. In addition to medical data (hypertension, heart disease, average post-meal glucose level, Body Mass Index (BMI), smoking status, and stroke experience), these metrics included patient demographic data (gender, age, marital status, type of work, and housing) as well as health data.

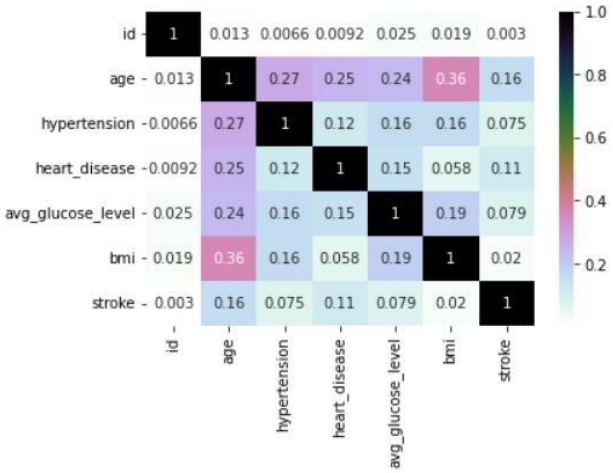

Figure 2. Dataset

VII. IMPLEMENTATION AND RESULTS

In the phases that follow, we'll show how we implemented our proposed system.

Phase-1:

- During this phase, we trained our Decision Tree model using the complete dataset.
- After that, we calculated accuracy to evaluate the performance of our model.
- We also measured the accuracy of our model using validation data, which came out to 97.14 percent.
- The following are the findings from our research using validation and test data:

```
+----------+--------------------+------+--------------------+
|prediction|         probability|stroke|            features|
+----------+--------------------+------+--------------------+
|       0.0|[0.99203258140408...|     0|(16,[0,2,5,6,11,1...|
|       0.0|[0.99203258140408...|     0|(16,[0,2,5,6,10,1...|
|       0.0|[0.93949743839960...|     0|(16,[0,2,5,6,10,1...|
|       0.0|[0.99203258140408...|     0|(16,[0,2,5,6,10,1...|
|       0.0|[0.99203258140408...|     0|(16,[0,2,5,7,10,1...|
+----------+--------------------+------+--------------------+
```

Figure 3. Results of Decision Tree

Below we can observe the plots containing accuracy and loss metrics of decision tree.

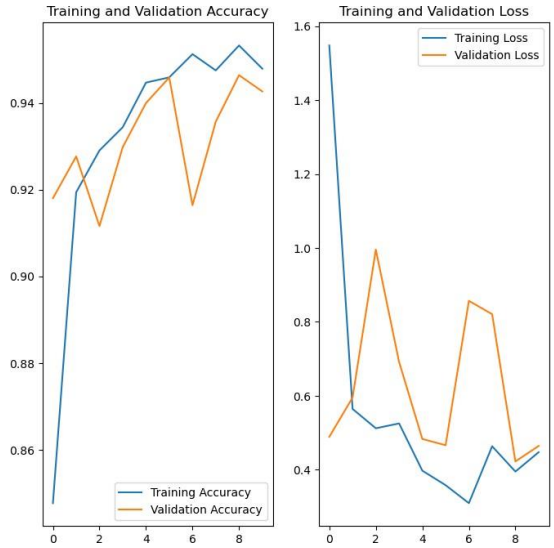

Figure 4. Plots with decision tree

Phase-2:

- During this phase, we trained our Multilayered Perceptron model using the complete dataset.
- After that, we calculated accuracy to evaluate the performance of our model.
- We also measured the precision of our model using validation data, which came out to 98.69 percent.

Below we can observe the plots containing accuracy andloss metrics of Multilayered perceptron.

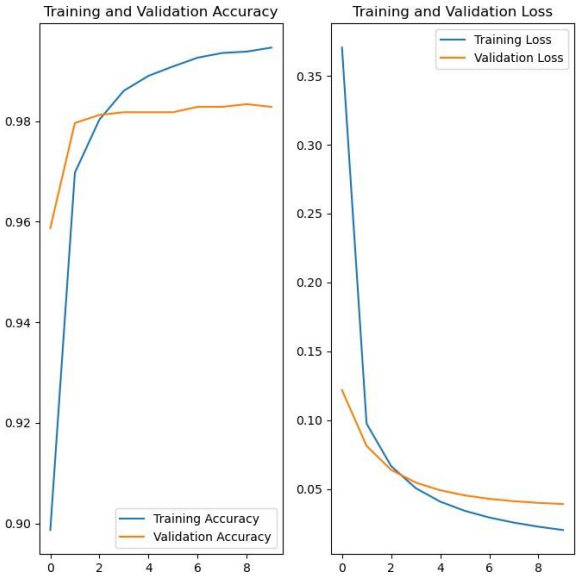

Figure 5. Plots with Multilayered Perceptron

Also, we had observed that avg glucose level plays an important role in determining the stroke status of a patient.

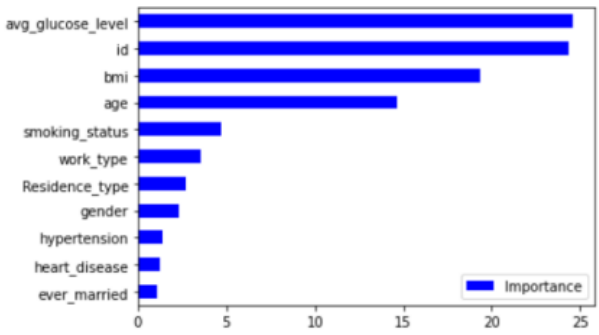

Figure 6. Feature Importance

VIII. CONCLUSION AND FUTURE WORK

ALGORITHM	ACCURACY
Decision Tree	97.14%
MLP	98.69%

There was a total of 7 insights in the stroke dataset. Age and BMI seemed to be positively correlated, but the relationship was tenuous. Compared to younger patients, older individuals had a higher risk of having a stroke. A greater BMI does not raise the risk of stroke. Those with prediabetes are more likely to experience a stroke, and diabetes is one of the risk factors for stroke. Given all other factors being equal, more people with hypertension or heart disease experienced a stroke. No matter their gender or where they stayed, patients are just as likely to experience a stroke. Age and the factor of marital status had a substantial correlation. By enhancing the framework and methods, using more efficient big data technology to analyse large volumes of data, and utilizing the top machine learning algorithms, we can further our project's study in the future and deliver more accurate and helpful results.

REFERENCES

[1] Hermon R, Williams PA. Big data in healthcare: What is it used for? 3rd Australian eHealth Informatics and Security Conference; 2014.

[2] . Chen M, Mao S, Liu Y. Big data: A survey. Mobile NetwAppl. 2014; 19(2): 171-209.

[3] Ristevski B, Chen M. Big data analytics in medicine and healthcare. J Integr Bioinform. 2018; 15(3): 1-5. PMID: 29746254 DOI: 10.1515/jib-2017-0030 [PubMed]

[4] Mooney SJ, Pejaver V. Big data in public health: Terminology, machine learning, and privacy. Annu Rev Public Health. 2018; 39: 95-112. PMID: 29261408 DOI: 10.1146/annurev-publhealth040617-014208 [PubMed]

[5] . Jin X, Wah BW, Cheng X, Wang Y. Significance and challenges of big data research. Big Data Research. 2015; 2(2): 59-64.

[6] . Bello-Orgaz G, Jung JJ, Camacho D. Social big data: Recent achievements and new challenges. Information Fusion. 2016; 28: 45-59.

[7] Arockia Panimalar S, Varnekha Shree S, Veneshia Kathrine A. The 17 V's of big data. International Research Journal of Engineering and Technology. 2017; 4(9): 329-33.

[8] Goga K, Xhafa F, Terzo O. VM deployment methods for DaaS model in clouds. In: Barolli L, Xhafa F, Javaid N, Spaho E, Kolici V. (eds) Advances in internet, data & web technologies. Lecture notes on data engineering and communications technologies, vol 17. Springer, Cham; 2018.

[9] Khan AS, Fleischauer A, Casani J, Groseclose SL. The next public health revolution: Public health information fusion and social networks. Am J Public Health. 2010; 100(7): 1237-42. PMID: 20530760 DOI: 10.2105/AJPH.2009.180489 [PubMed]

[10] Velikova M, Lucas PJF, Samulski M, Karssemeijer N. A probabilistic framework for image information

fusion with an application to mammographic analysis. Medical Image Analysis. 2012; 16(4): 865- 75.
[11] . Antink CH, Leonhardt S, Walter M. A synthesizer framework for multimodal cardiorespiratory signals. Biomedical Physics & Engineering Express. 2017; 3(3): 035028