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

Prediction of Heart Attack from Medical Records Using Big Data Mining

[1]M. Sunil Kumar, [2]Vasanthakumari Sundararajan, [3]N. Alangudi Balaji, [4]Sachin Sambhaji Patil, [5]Sudhir Sharma, [6]D. C. Joy Winnie Wise

Submitted: 04/11/2022 **Accepted:** 07/02/2023

Abstract: In recent years, data mining has arisen as a possible new field for identifying insights in the underlying patterns of big datasets. This discovery can be accomplished through the examination of massive amounts of data. The utilization of huge datasets is one method for accomplishing this goal. Data mining is a labor-intensive technique that involves mining enormous databases for buried meaning and searching for prospective applications in those datasets. In a nutshell, it is a method that entails looking at data from several angles in order to achieve a greater knowledge of the data in question. The subject of medicine is one of the many that could benefit from these new understandings, which have a wide range of applications. In this paper, we develop big data mining pattern using deep learning algorithm to predict the rate of heart attack in human beings.

***Keywords*:** *Big data mining, deep learning, heart attack, human beings.*

1. Introduction

We simply would not be able to function if our hearts were removed from our bodies. Existence presupposes that one possesses a heart that is capable of carrying out its regular duties as intended. Problems with the heart can have far-reaching implications on a person health, affecting not only the heart but also the brain, the kidneys, and every other organ in the body in addition to the heart itself [1].

These problems can have a negative influence on the heart itself. A heart attack is the most prevalent cause of unexpected mortality in modern times and it can occur anywhere in the world. It used to be that only people in their middle years or older were at risk for having a heart attack, but modern society reliance on a diet of fast food and lack of exercise has rendered young people equally as vulnerable as older people. In the past, only people in their middle years or older were at risk for having a heart attack. It might be difficult for doctors to anticipate when a patient will have a heart attack because of the complex nature of the disease. As a consequence of this, a technique for forecasting heart attacks that is founded on data mining might prove to be advantageous [2].

According to a research that was not too long ago issued by the World Health Organization cardiovascular diseases (commonly known as CVDs) are the biggest cause of death across the globe. It is estimated that cardiovascular illnesses were the cause of death for 17.9 million persons around the world in 2016, which is equivalent to 31% of the total number of people who passed away worldwide in that same year [3]. According to the findings of recent research, cardiovascular illness is the primary contributor to 85% of these fatalities. A sizeable number of scientists have, for a considerable amount of time now, taken an interest in the modeling and prediction of the risk of cardiovascular disease (CVD).

On the other hand, it has been shown that the predictive potential of these risk scales is fairly limited. This is since the pathophysiology of cardiovascular disease is complex as well as variable. The availability of novel risk biomarkers has been steadily increasing, which has resulted in a concomitant increase in the need for disease prediction models with a higher degree of precision. This makes it extremely difficult to develop risk models that can be relied upon. Because of this, developing risk models that can be depended upon is a challenging endeavor [4].

[1]*Professor & Programme Head, Department of Computer Science and Engineering, School of Computing, Mohan Babu University, (erstwhile Sree Vidyanikethan Engineering College), Tirupathi, AP, India.*
[2]*Associate Professor, Pediatric and Neonatal Nursing Department, Institute of Health Sciences, Wollega University, Ethiopia.*
[3]*Professor, Computer Science and Engineering, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Guntur, Andhra Pradesh, India.*
[4]*Assistant Professor, Department of Computer Engineering, Zeal College of Engineering and Research, Pune, India.*
[5]*Assistant Professor, Department of IT, School of Information Technology, Manipal University Jaipur, Jaipur- Rajasthan, India.*
[6]*Professor, Department of Computer Science and Engineering, Chennai Institute of Technology, Chennai, Tamil Nadu, India.*
[1]*sunilmalchi1@gmail.com,* [2]*vasanhar@yahoo.com,* [3]*alangudibalaji@gmail.com,* [4]*sachin.patil@zealeducation.com,* [5]*sudhir.sharma@jaipur.manipal.edu,* [6]*drdcjoywinniewise@gmail.com*

In addition, recent research has shown that the factors that put people at risk for cardiovascular illnesses differ not only according to the social setting in which they are living, but also according to their ethnic background and their region [5]. As a consequence of this, an adaptive method ought to be applied in order to build more precise models of CVD risk that are able to be effectively adapted to a particular population. This is necessary in order to achieve the desired level of precision.

As a result of recent advancements in national or regional EHR management systems, it is now possible to combine and share EHR data from a number of different institutions. This has made it more simpler and much quicker to collect data from a large population for the goal of utilizing it in retrospective cohort studies to improve one ability to evaluate characteristics that put one at risk for cardiovascular disease. Electronic health records were the source of the information for this investigation [6].

The vast majority of the findings, on the other hand, demonstrated only minor improvements when compared with risk scales that had been established previously. However, despite the fact that the data from EHR are more simpler to obtain than the data that are used in conventional cohort studies, the quality of the data from EHR is noticeably lower [7].

One research that could be followed is identifying whether or not the data from EHRs are inherently inaccurate, which would render them unsuited for the purpose of producing exact forecasts. This is since virtually all of the methods that have been published to far demand that the EHR data be transformed into a single matrix, which removes any dynamism that may exist in the data [7] [8]. One of the main reasons why there are so few studies on this subject is because of this. Because of this, it is anticipated that the development of a more effective modeling strategy, the implementation of more sophisticated machine-learning algorithms, and the provision of new data resources will all positively contribute to the efficacy of the prediction models that are currently being utilized.

2. Related works

It is challenging to gather data from disparate locations due to the sheer volume of data that must be collected as well as the fact that the data must be collected in a variety of formats. Second, the amount of storage space that is required for large datasets that are also diverse is the most significant issue that might occur because of having such datasets [9]. Large amounts of data necessitate stringent criteria, not only for the system storage but also for the component guaranteed levels of performance. The analysis of large amounts of data is particularly challenging in this regard [10]. A new processing paradigm is necessary to handle these challenges because the existing data management systems are inefficient when it comes to dealing with the diverse nature of data or the real-time nature of the data.

On the other hand, the management of structured data is typically handled by more conventional RDBMSs like MySQL. This is because structured data is easier to organize. Sadly, the vast majority of systems are incapable of processing information that is merely semi-structured or unstructured. The fact that traditional RDBMS are unable to scale for fault tolerance and parallel hardware management renders them unsuitable for managing expanding datasets from a scalability point of view [11].

One example of the many works that have been provided by the academic community to solve the issues of storing huge and heterogeneous amounts of data is the NoSQL database management system [12], which is just one example of the many works that have been offered. When dealing with vast amounts of data where a relational model is not required due to the nature of the data, these solutions can be helpful [13] [14].

The Map operation and the Reduce operation are the two components that make up the MapReduce [15] method of performing parallel processing. It is put to use for the purpose of processing enormous quantities of data that are dispersed all throughout a commodity cluster. When employing the MapReduce system, iterative algorithms are carried out at an extremely sluggish pace, which is one of the system most major drawbacks. Iterative computations fall outside the boundaries of what was initially envisioned as being possible with the MapReduce programming model.

Hadoop is a system for processing data in batches that uses the MapReduce programming language for the aim of storing and processing enormous amounts of data in a distributed manner. Hadoop was designed with the intention of achieving these goals. It has a high fault tolerance and provides a distributed storage system through the use of the Hadoop Distributed File System (HDFS) [16].

It is not viable to use Hadoop for real-time stream processing, nor is it possible to use Hadoop for computation that takes place in memory. Additionally, it is not always simple to apply the MapReduce paradigm to all different types of issues. Hadoop can only be utilized for processing data in batches. It possible that the arrival of the results will be greatly delayed depending on how much data is being analyzed [17].

The computing approach that is known as stream computing places a primary emphasis on the rate at which the data, which is continuously input and output, is moved through the system. Processing with low-latency in

addition to real-time computing and high throughput are all provided by BDSC as a result of its utilization of a distributed message bus. Big data analytics in the healthcare business aims to achieve a number of goals, the primary one being the extraction of actionable insights from huge amounts of data. The massively parallel processing architectures that are provided by BDSC are ideally suited to achieving this objective [18].

Research and clinical practice in the healthcare business are beginning to enjoy considerable benefits from the ever-increasing popularity of big data analysis. This is due to the fact that big data analysis is becoming more and more widespread. It has made it feasible to capture, administer, analyze, and assimilate massive quantities of heterogeneous, structured, and unstructured data that are generated by current healthcare systems [19].

In order to accomplish its purpose of real-time discovery of the unrealized potential of big data in the healthcare industry, BDSC has emerged as a significant participant in the field of big data analytics. The application of machine learning to such a vast volume of data creates a difficulty because traditional machine learning architectures were not designed to deal with data streams of this magnitude or speed. The application of machine learning is made more complex as a result of this difficulty. In addition, the processing of analytical data is fraught with a variety of difficulties that must be overcome [20].

In order to perform more in-depth analytical processing, it is important to integrate data between multiple systems. Even though machine learning is utilized in the vast majority of cutting-edge research, the challenge of applying machine learning in real time to enormous streams of streaming data has not yet been solved. This is despite the fact that machine learning is one of the most important components of cutting-edge research. On the other hand, Hadoop is a computer system that operates in batches, and the bulk of healthcare analytics solutions have focused their emphasis on Hadoop as a result [21].

As a whole, the population is becoming older, and as the prevalence of chronic diseases keeps climbing, more and more people are paying attention to the constraints that are imposed by conventional medical care. In addition, the community of people who are afflicted with heart disease is rapidly adopting medical IoT for the purpose of continuous monitoring in order to carry out real-time interventions in the case of an unexpected emergency. This suggests that a considerable amount of data is being generated by the millions of sensors that are currently present in the environment. In scenarios in which there is a risk of someone losing their life, it is not easy to make sense of the available data and decide what actions to take quickly [22].

3. Proposed Method

Internet of Things goal is to make it possible for devices to collect data from one another and share it with one another as well as with the systems that are housed in data centers. This is necessary in order to either comprehend the behavioral patterns of the users or to extract the important information. It is impossible to accomplish either of these goals without first completing this step.

More than any other sector of the economy, healthcare will be the primary user of Internet of Things-related technology in the not too distant future [31], accounting for forty percent of that technology. The fields of medical informatics and other information technology are coming together to improve the healthcare business by reducing costs, eliminating waste, and enhancing the quality of treatment provided to patients. In the event of a medical emergency, such as one brought on by heart disease, diabetes, or any of a broad variety of other chronic conditions, real-time monitoring carried out through the Internet of Things assists in the saving of lives. A plethora of health-related services, many of which measure vital signs in real time, are available online and may be accessed by anybody. In this day and age, the internet offers a wealth of information on a variety of topics, including health.

Spark

Spark is a processing engine, and as such, it runs one master process in addition to many worker processes for each Spark application. This is done so that data can be processed in parallel. The Spark master is analogous to the person behind the wheel of a vehicle, while the Spark employees play the part of the individuals riding in the vehicle.

As a result of the central role it plays, the driver is accountable for carrying out a variety of activities, including assessing, dividing, scheduling, and monitoring the progress achieved by the executors. During the entirety of the application runtime, it is the driver job to make sure that all the data is accurate and up to date. Figure 1 reveals that the primary responsibility of executors is to carry out the instructions that are given to them by the driver.

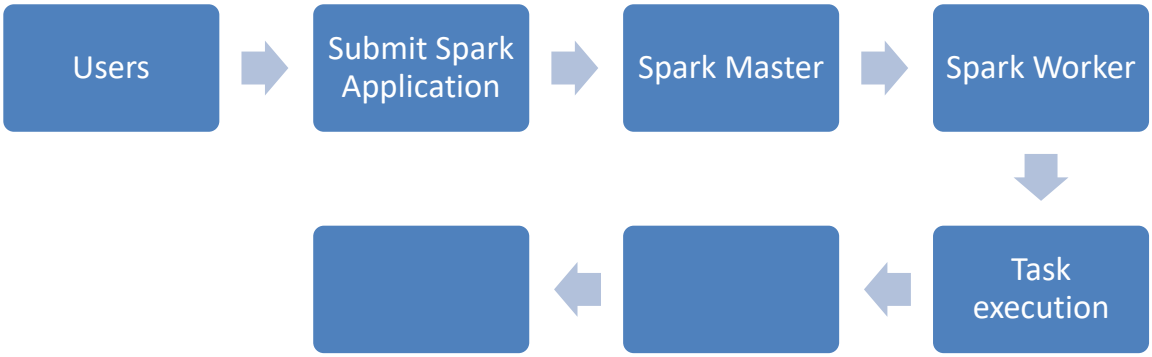

Fig. 1: Proposed Framework

Use case datasets

We placed the suggested model to the test by utilizing two separate collections of data. The diabetes data set that was used in our analysis was obtained for this purpose from Kaggle [36], a website that gives data scientists access to online datasets. This was done for the aim of utilizing the data set in our investigation. Finding and analyzing free data may be made much easier with the help of Kaggle, which is the company primary goal.

The dataset on diabetes includes a total of 15,000 records, each of which includes eight separate attributes, including gestational diabetes, blood sugar, blood pressure, skin thickness, insulin, body mass index, diabetes family history, age, and a binary outcome value of either 1 or 0. The binary outcome value can be either 1 or 0. The value of the binary outcome can be either one or zero.

The second source was a database that has been analyzed before and includes HD patients from Cleveland. A significant percentage of the pieces of writing that were produced regarding machine learning made use of it.

This dataset class label attribute contains labels for two classes: those indicating the existence of heart disease and those indicating the absence of heart disease. The class label property used to have the ability to take one of four alternative values; however, it now only accepts the values 0 and 1. If the value is 1, it indicates that someone has heart disease, but if the value is 0, it shows that someone does not have heart disease.

The Spark environment will be covered in this section, and throughout it, we will learn how to conduct predictive analysis on datasets utilizing that environment. After that, the contents of the CSV file are imported into an RDD that is composed entirely of strings. To train and evaluate our machine learning model that predicts the health of patients, we apply the map transformation to the RDD. This helps us organize the data more efficiently. This transformation makes use of the Parse RDD function so that each string element that is contained within the RDD can be converted into an RDD that is composed of labelled points.

Spark implementation

To classify the features of a user, first data on various diseases must be acquired from a variety of sources, and then a classification model must be developed making use of this information. This must be done to classify a user characteristic, regardless of whether the condition is present. Classification is one of the most important techniques for data mining because it can be used to unearth patterns that were previously concealed.

Classification and regression problems are famous examples of when DT can be employed. Because of their user-friendliness, operational simplicity, interpretability, and scalability in a multiclass classification scenario, DT have been widely adopted as a go-to technique for machine learning classification applications. We made the forecast by utilizing DT from Spark machine, MLlib, which enables DT to be utilized in both binary and multiclass classification. This enabled us to create an accurate prediction.

Decision trees, also known as DTs, are a form of model used in machine learning that are responsible for the creation of subsets of data. The process of partitioning starts with a binary split and continues forever until there is no longer any possibility of further division. A DT is constructed using recursive partitioning in a step-by-step manner. First, it is determined whether to split each node, and then, from among all of the possible splits, the partition that will yield the best results is chosen.

Gini impureness and entropy are two examples of such criteria that are used to separate the group. Both measures can be calculated. The inherent impureness of a label node can be used to quantify the homogeneity of the node. Measurements of classification impurity can now be obtained in the implementation using the Gini and Entropy statistics.

With the help of this approach, the entropy of characteristic S can be calculated:

$$Entropy = \sum_{j=1}^{C_p} (s, j) \log_p (s, j)$$

where,

C - classes and

$p(S, j)$ - instances.

(2)

Calculating the entropy (Sv), where TsTs is the set of values of S in T, Ts is the subset of T induced by S, and Ts,v is the subset of T in which attribute S has a value of v, allows us to determine the amount of data that must be collected following the application of the partitioning by attribute S. we can define the process of acquiring new knowledge as:

$$Gain(S,T)=Entropy(S)-Info(S,T)$$

The formula used to measure the information that was acquired after a partition by attribute S is as follows:

$$GainRatio(S,T)=Gain(S,T)SplitInfo(S,T)$$

The information gain ratio for the attribute set S is defined as follows:

$$SplitInfo(S,T) = -\sum_{v\in Values}(Ts)|T(S,v)||TS|*\log|T(S,v)||TS|$$

For this reason, it is essential to have a reliable parallel Spark model for estimating the state of one health in an environment containing a massive amount of data. Because of this, utilizing the C4.5 model should be prioritized even higher on the list of priorities. The C4.5 code is parallelized here with the help of Spark.

4. Results and Discussions

In this section, we will present an overview of the dataset that was utilized in the experiments that were conducted for this study at a high level. As was said before, for the purpose of our work, we make use of a cutting-edge dataset. This dataset was developed by African medical professionals, and these features are a component of the dataset that they produced. Only 14 of the characteristics that are available in this dataset are used in our analysis to identify whether a person has CHD. The following is a list of algorithmic traits, together with condensed descriptions of their functions and the range of possible values for those roles and values as in Table 1.

Table 1: Attributes

Attribute	Possible values
Age	Valid numbers
Sex	1: male/0: female
trestbps	90–200
Chol	125–565
Fbs	1: true; 0: false
restecg	0: normal 1: with ST-T
thalach	71–202
exang	1 = true; 0 = false
old peak	0–7
Ca	0–3

In this section, we provide a detailed description of the experimental setup, including the experimental settings as well as the hardware and software platforms that were utilized in the studies. To generate estimates of the patient cardiovascular fitness, we searched for trends among the features contained within the dataset.

Experimental Setup

In order to monitor and study the performance of a proposed model on the benchmark dataset obtained from the UCI benchmark repository, the experiment uses the following settings:

To carrying out this research and evaluating the results, an experimental environment was created on a personal computer. This product makes use of Microsoft Windows 10 Pro operating system as its principal means of

computer operation. The effectiveness of the classifier can be evaluated by using a few different performance indicators. There are a variety of metrics that may be used in machine learning to evaluate the performance of a classifier. In the following sections, we will go into further detail on some of these characteristics.

Precision

The precision of a classifier is the yardstick that is used to measure its accuracy when the efficacy of the classifier is being evaluated. When accuracy is high, there are fewer positive findings that are not warranted. As the model precision declines, there will be a rise in the number of false positives.

$$Precision=TP/(TP+FP).$$

Where,

TP - true positive,

FP - false positive.

FN – False Negative

Recall

This is a statistic that can be used to evaluate how comprehensive the classifier is. When there is a high recall, there is a low number of false negatives, and when there is a low recall, the number of false negatives increases. In most circumstances, improving recall will result in a loss of precision.

$$Recall=TP/(TP+FN)$$

F-Score

The F-score is a measurement of how well a test performed in terms of accuracy and recall.

$$F1\ Score=(Recall\times Precision)/(Recall+Precision)$$

F1 score divide the product of your recall score and your precision score by the sum of your recall and precision scores, as stated in the equation below.

As was mentioned earlier, preparing datasets in advance of starting experiments is necessary since they need to be tailored to the kind of investigation that will be carried out. As part of the process of data preparation, the dataset is either saved directly into the file system with the assistance of a Python library called SKLearn or it is fetched into the file system by third-party software. SKLearn is part of the data preparation process.

It was discovered that the dataset had a significant amount of background noise in addition to a few inconsistencies. To clean and standardize the data, this necessitated the creation of multiple programs in the Python programming language. To convert raw data into a feature matrix that can be utilized, specific algorithms need to be developed for a wide variety of jobs. Some of these tasks include data purification, anomaly replacement, mean value computation, normalizers, and many more. These procedures are discussed in greater detail in the section that is devoted to the proposed framework, which can be found above.

After presenting the dataset in the form of a feature matrix, the next step is to separate the dataset into its individual components using a categorization scheme. As an illustration, we separated the records in the dataset into two categories: those with heart disease and those without heart illness. After that, the data is classified utilizing CNN into binary classes, and a classifier model that corresponds to those classes is then crafted and stored on disk.

The overall accuracy of the model and stands at 97% and a classifier was developed that divides heart illness into four distinct categories to accurately reflect the information included in the dataset. We put a program written in Python through its paces so that it could count all possible classes as in Figure 2 - 5.

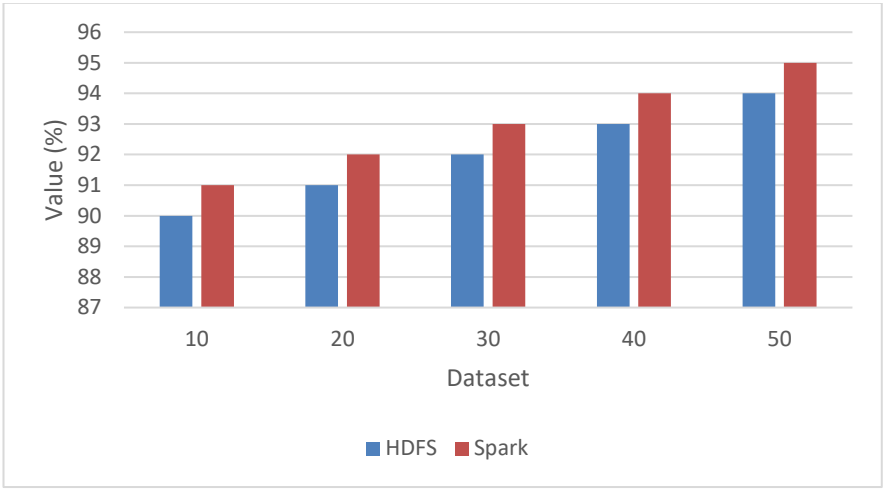

Fig. 2: Accuracy

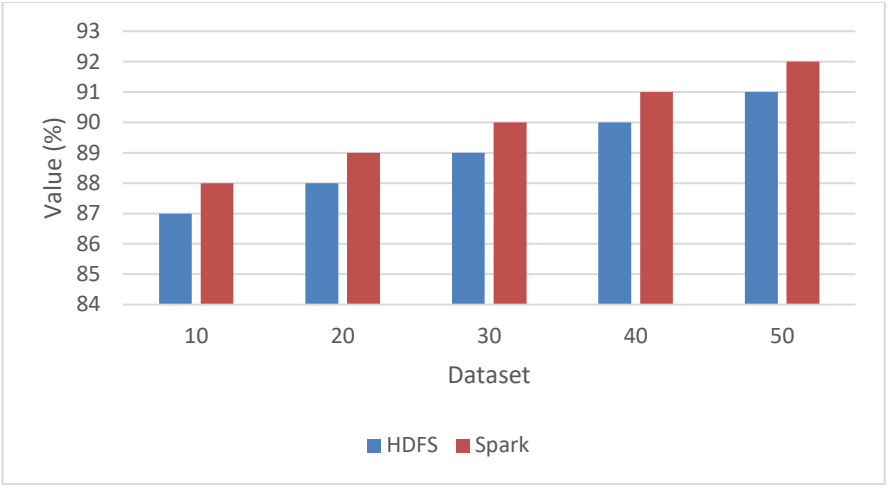

Fig. 3: Precision

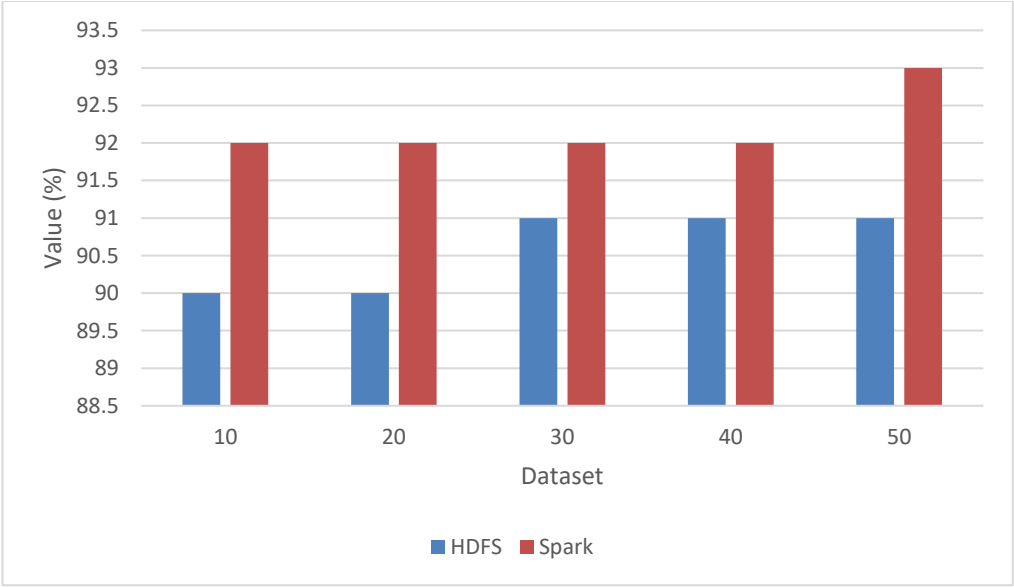

Fig. 4: Recall

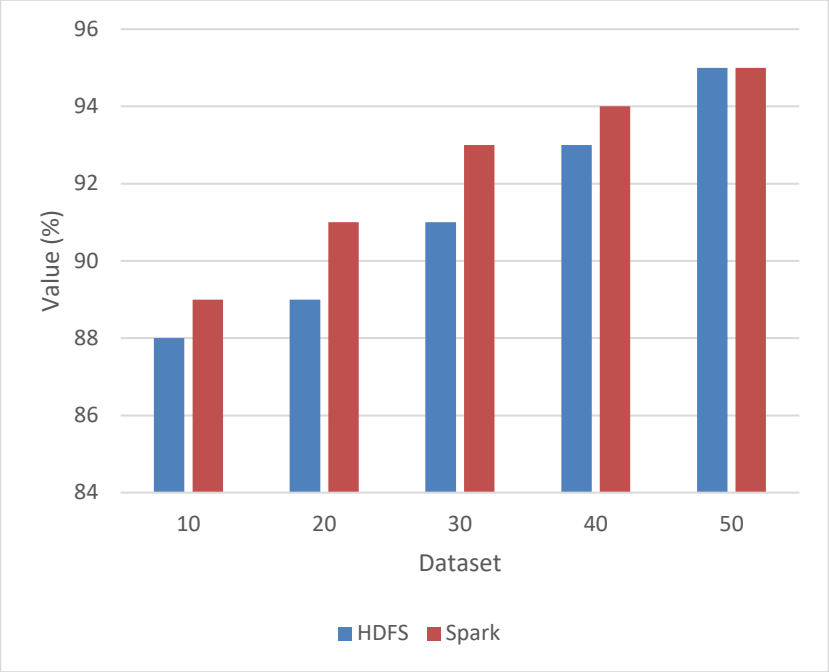

Fig. 5: F-Measure

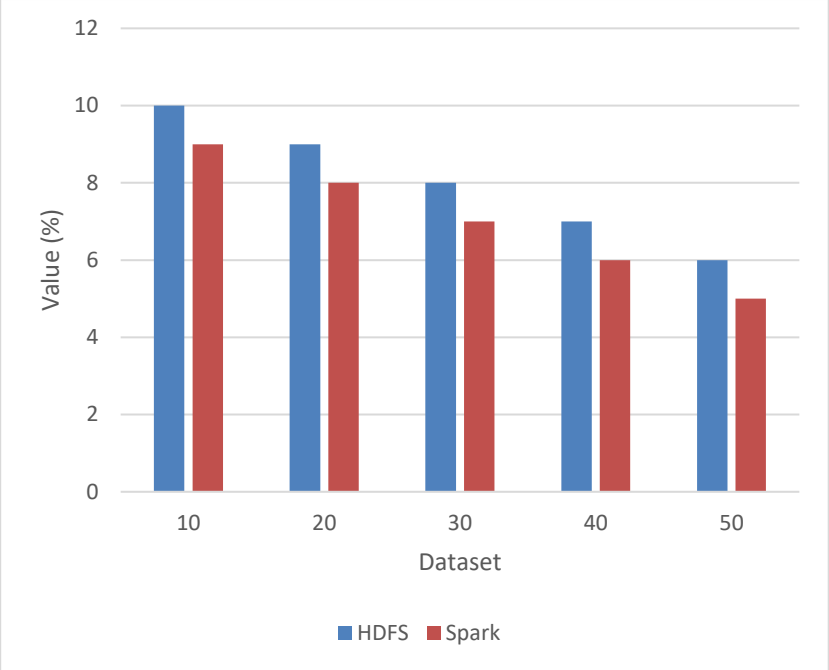

Fig. 6: Classification Error

In addition to that, the Type 4 class contains a total of 50 records. There were 10 records incorrectly classified as having type 1 illness; however, there are 30 records incorrectly classified as having type 2 illness. The sickness of type 3 has a false classification of six, but the disease of type 4 does not have any.

5. Conclusions

If cardiac diseases could be identified at earlier stages, it might be possible to reduce the number of deaths that result from heart attacks. With the help of an accurate classification system, a doctor can determine whether or not a patient has cardiovascular disease before the patient ever shows symptoms.

In this study, an attempt is made to forecast the presence of heart disease by making use of several deep neural networks as well as a cutting-edge dataset obtained from the UCI repository. This data collection includes the results of some heart tests as well as common behaviors associated with humans. The results of the research indicate that the proposed model performs more effectively than the strategies that are currently in use and that are referenced in the study. The accuracy of the

suggested model is exceptionally good, coming in at 97% on average.

References

[1] Kumar, S. R., Gayathri, N., Muthuramalingam, S., Balamurugan, B., Ramesh, C., & Nallakaruppan, M. K. (2019). Medical big data mining and processing in e-healthcare. In *Internet of things in biomedical engineering* (pp. 323-339). Academic Press.

[2] Obasi, T., & Shafiq, M. O. (2019, December). Towards comparing and using Machine Learning techniques for detecting and predicting Heart Attack and Diseases. In *2019 IEEE international conference on big data (big data)* (pp. 2393-2402). IEEE.

[3] Jain, P., & Kaur, A. (2018, August). Big data analysis for prediction of coronary artery disease. In *2018 4th International Conference on Computing Sciences (ICCS)* (pp. 188-193). IEEE.

[4] Eletter, S., Yasmin, T., Elrefae, G., Aliter, H., & Elrefae, A. (2020, November). Building an intelligent telemonitoring system for heart failure: The use of the internet of things, big data, and machine learning. In *2020 21st International Arab Conference on Information Technology (ACIT)* (pp. 1-5). IEEE.

[5] Ishaq, A., Sadiq, S., Umer, M., Ullah, S., Mirjalili, S., Rupapara, V., & Nappi, M. (2021). Improving the prediction of heart failure patients' survival using SMOTE and effective data mining techniques. *IEEE access*, *9*, 39707-39716.

[6] Hemingway, H., Asselbergs, F. W., Danesh, J., Dobson, R., Maniadakis, N., Maggioni, A., ... & Innovative Medicines Initiative 2nd programme, Big Data for Better Outcomes, BigData@ Heart Consortium of 20 academic and industry partners including ESC. (2018). Big data from electronic health records for early and late translational cardiovascular research: challenges and potential. *European heart journal*, *39*(16), 1481-1495.

[7] Munagala, N. V. L., Saravanan, V., Almukhtar, F. H., Jhamat, N., Kafi, N., & Khan, S. (2022). Supervised Approach to Identify Autism Spectrum Neurological Disorder via Label Distribution Learning. *Computational Intelligence and Neuroscience*, *2022*.

[8] Rastogi, R., Chaturvedi, D. K., Satya, S., & Arora, N. (2020). Intelligent heart disease prediction on physical and mental parameters: a ML based IoT and big data application and analysis. In *Machine Learning with Health Care Perspective* (pp. 199-236). Springer, Cham.

[9] Nayak, S., Gourisaria, M. K., Pandey, M., & Rautaray, S. S. (2019, May). Prediction of heart disease by mining frequent items and classification techniques. In *2019 International Conference on Intelligent Computing and Control Systems (ICCS)* (pp. 607-611). IEEE.

[10] Srinivas, K., Rani, B. K., & Govrdhan, A. (2010). Applications of data mining techniques in healthcare and prediction of heart attacks. *International Journal on Computer Science and Engineering (IJCSE)*, *2*(02), 250-255.

[11] Subramanian, B., Saravanan, V., Nayak, R. K., Gunasekaran, T., & Hariprasath, S. (2019). Diabetic Retinopathy–Feature Extraction and Classification using Adaptive Super Pixel Algorithm. Int J Eng Adv Technol, 9, 618-627.

[12] Munagala, N. V. L., Saravanan, V., Almukhtar, F. H., Jhamat, N., Kafi, N., & Khan, S. (2022). Supervised Approach to Identify Autism Spectrum Neurological Disorder via Label Distribution Learning. Computational Intelligence and Neuroscience, 2022.

[13] Tripathy, R., Nayak, R. K., Saravanan, V., Mishra, D., Parasa, G., Das, K., & Das, P. (2021). Spectral Clustering Based Fuzzy C-Means Algorithm for Prediction of Membrane Cholesterol from ATP-Binding Cassette Transporters. In Intelligent and Cloud Computing (pp. 439-448). Springer, Singapore.

[14] Aruna R, D., Surendran S, D., Yuvaraj N, D., & Debtera, B. (2022). An Enhancement on Convolutional Artificial Intelligent Based Diagnosis for Skin Disease Using Nanotechnology Sensors. Computational Intelligence and Neuroscience, 2022.

[15] Manikandan, R., Sara, S. B. V., Yuvaraj, N., Chaturvedi, A., Priscila, S. S., & Ramkumar, M. (2022, May). Sequential pattern mining on chemical bonding database in the bioinformatics field. In AIP Conference Proceedings (Vol. 2393, No. 1, p. 020050). AIP Publishing LLC.

[16] Poornima, S., & Pushpalatha, M. (2018). A survey of predictive analytics using big data with data mining. *International journal of bioinformatics research and applications*, *14*(3), 269-282.

[17] Alonso, S. G., de la Torre Diez, I., Rodrigues, J. J., Hamrioui, S., & Lopez-Coronado, M. (2017). A systematic review of techniques and sources of big data in the healthcare sector. *Journal of medical systems*, *41*(11), 1-9.

[18] Rumsfeld, J. S., Joynt, K. E., & Maddox, T. M. (2016). Big data analytics to improve cardiovascular

care: promise and challenges. *Nature Reviews Cardiology*, *13*(6), 350-359.

[19] Yahaya, L., Oye, N. D., & Garba, E. J. (2020). A comprehensive review on heart disease prediction using data mining and machine learning techniques. *American Journal of Artificial Intelligence*, *4*(1), 20-29.

[20] Athey, S. (2017). Beyond prediction: Using big data for policy problems. *Science*, *355*(6324), 483-485.

[21] Peng, Z. (2019, January). Stocks analysis and prediction using big data analytics. In *2019 international conference on intelligent transportation, big data & smart City (ICITBS)* (pp. 309-312). IEEE.

[22] Reddy, P. C., & Babu, A. S. (2017, February). Survey on weather prediction using big data analystics. In *2017 Second international conference on electrical, computer and communication technologies (ICECCT)* (pp. 1-6). IEEE.