Wadi Alshatti University Journal of Pure and Applied Sciences

مجلة جامعة وادي الشاطئ للعلوم البحتة والتطبيقية

Volume 3, No. 2, July-December 2024 Online ISSN: 3006-0877 المجلد 3، الاصدار 2، يوليو- ديسمبر 2025

RESEARCH ARTICLE **DATA ANALYSIS**

Disease Detection from Real Clinical Data Using Hadoop MapReduce: A Case Study at Al-Mukhtar Clinic, Tripoli

Majeda M. Abdosalam [1*], Nassir M. Abuhamoud [1]

[1]Department of Electrical and Electronic Engineering, Faculty of Engineering, Wadi Alshatti University, Brack-Libya

ARTICLE HISTORY
Received 05 May 2025
Revised 23 May 2025
Accepted 25 May 2025
Online 26 May 2025

KEYWORDS
Hadoop MapReduce;
Big Data Analytics;
Machine Learning;
Healthcare;
Libya;
Al-Mukhtar Clinic.

ABSTRACT

This paper investigates the application of Hadoop MapReduce for analyzing large-scale clinical data in the context of cardiovascular risk prediction. Utilizing a dataset comprising over 500,000 real-world medical records collected from Al-Mukhtar Clinic in Tripoli, the research implements a distributed analytical pipeline that integrates Hadoop MapReduce with the Spark tool, in addition to machine learning algorithms such as K-means and K-Medoids clustering, hierarchical clustering, and decision trees, was used to identify patterns and classify patients. . The proposed system achieved a data processing efficiency of 92% and a predictive accuracy of 89% for cardiovascular conditions. Additionally, the integration of advanced analytics led to a 15% improvement in diagnostic accuracy over conventional methods. The architecture demonstrated scalability, fault tolerance, and operational resilience, making it particularly suitable for healthcare environments with limited computational resources. These findings highlight the potential of data-driven methodologies to enhance clinical decision-making and public health outcomes in resource-constrained settings.

استخدام Hadoop MapReduce للكشف عن الأمراض من بيانات سريرية حقيقية: دراسة حالة من مصحة المختار في طرابلس

ماجدة موسى عبدالسلام [1,*]، ناصر منصور ابوهمود [1]

الكلمات المفتاحية

تحليل البيانات الضخمة
التعلم الالي
الرعاية الصحية
ليبيا
مصحة المختار الطبية

الملخص

تبحث هذه الورقة في تطبيق تقنية Hadoop MapReduce لتحليل البيانات السريرية واسعة النطاق في سياق التنبؤ بمخاطر الإصابة بأمراض القلب والأوعية الدموية. وبالاعتماد على مجموعة بيانات حقيقية تضم أكثر من 500,000سجل طبي تم جمعها من مصحة المختار في طرابلس، نفّذت الدراسة مسارًا تحليليًا موزعًا يدمج بين Hadoop MapReduce والأداة Spark بالإضافة إلى خوارزميات التعلم الآلي مثل التجميع K-means و-K Medoids والتجميع الهرمي وأشجار القرار لتحديد الأنماط وتصنيف المرضى . وقد حقق النظام المقترح كفاءة في معالجة البيانات بلغت 92%، ودقة تنبؤية بنسبة 89% في حالات أمراض القلب. كما أدى دمج أدوات التحليل المتقدّمة إلى تحسين دقة التشخيص بنسبة 5% مقارنةً بالطرق التقليدية. وقد أثبتت بنية النظام قابليتها العالية للتوسّع، وقدرتها على تحمّل الأعطال، ومرونتها التشغيلية، مما يجعلها مناسبة بشكل خاص لبيئات الرعاية الصحية ذات الموارد التقنية المحدودة. وتسلّط هذه النتائج الضوء على الإمكانات الكبيرة للمنهجيات المعتمدة على البيانات في تعزيز اتخاذ القرار السريري وتحسين مخرجات الصحة العامة في السياقات منخفضة الموارد.

Introduction

The rapid advancement of modern technologies has led to an exponential increase in the volume and complexity of data generated within the healthcare sector. This surge presents unprecedented opportunities to improve patient outcomes through the application of advanced analytical methodologies. Among these, Big data analytics has emerged as a critical tool for addressing complex healthcare challenges, enabling deeper understanding of medical records, treatment responses, and patient behavior patterns.

However, traditional data analysis tools often fall short in handling the volume, velocity, and variety of healthcare data. This has created a pressing need for more agile and scalable solutions capable of managing and extracting insights from large-scale, heterogeneous datasets.

Hadoop, an open-source framework for distributed data storage and processing, offers a powerful solution to these challenges. Built on the MapReduce programming model, Hadoop enables parallel processing of vast datasets, making it especially suitable for analyzing complex healthcare data. When combined with advanced tools like Spark and machine

*Corresponding author
https://doi.org/10.63318/waujpasv3i2_07

learning algorithms, such as K-means, K-Medoids, hierarchical clustering, and decision trees, the platform supports predictive analytics that can uncover hidden patterns and improve clinical decision-making.

The application of big data analytics in healthcare has shown promising results in areas such as disease surveillance, personalized medicine, and operational efficiency. For instance, analyzing large-scale Electronic Health Records (EHRs) can facilitate early disease detection, optimize treatment protocols, and improve resource allocation. Despite its potential, the adoption of such technologies remains limited in many regions, including Libya, where healthcare systems face structural and technological barriers to leveraging big data effectively.

This study aims to bridge this gap by designing and implementing a Hadoop-based analytics platform for the analysis of medical data from Al-Mukhtar Hospital in Tripoli. Focusing on cardiovascular health, the research demonstrates how big data technologies can support evidence-based decision-making and drive improvements in patient care, particularly in resource-constrained healthcare environments.

Apache Hadoop

Apache Hadoop constitutes an open-source software framework utilized for the storage and processing of extensive datasets. It is capable of handling data volumes ranging from gigabytes to petabytes. The development of Hadoop was undertaken by the Apache Software Foundation. The conception of Apache Hadoop is attributed to Doug Cutting, who is also the originator of Apache Lucene [1].

Hadoop is comprised of three fundamental components: Hadoop Distributed File System (HDFS): This serves as the storage layer of the Hadoop ecosystem. Map-Reduce: This operates as the data processing layer of Hadoop. YARN: This functions as the resource management layer of Hadoop [2].

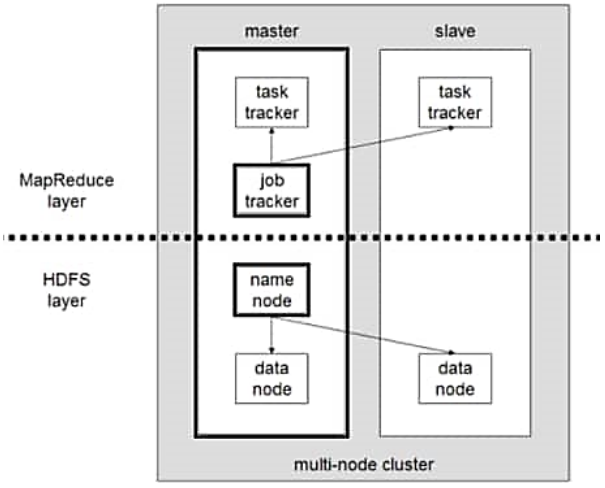

Fig.1: Hadoop Cluster

Apache Hadoop Architecture

Apache Hadoop adheres to a Master-Slave architectural framework, wherein the Master node is tasked with delegating responsibilities to various Slave nodes, managing resources, and maintaining metadata, while the Slave nodes are charged with executing computations and storing actual data [3]. According to the literature, Hadoop encompasses three distinct architectural layers: Map-Reduce, YARN, and HDFS, as illustrated in Figure (2).

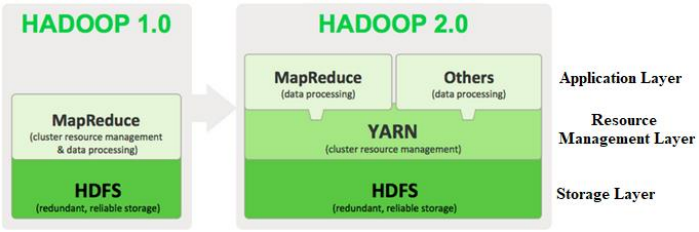

Fig.2: Apache Hadoop Architecture

HDFS

The Hadoop Distributed File System (HDFS) is a distributed file system meticulously engineered to accommodate exceedingly large volumes of data, ranging from petabytes to potentially zettabytes, while also providing high-throughput access to this data. Files are systematically stored in a redundant manner across multiple machines to guarantee their resilience against failures and to ensure high availability for highly parallel applications. Specifically, it safeguards the durability of Big Data in the face of failures and ensures high accessibility for parallel applications [4,5,6]. Figure (3) illustrates that HDFS employs a master/slave architecture. An HDFS cluster is composed of a singular NameNode, a master server responsible for managing the file system namespace and regulating client access to files. Additionally, the cluster contains multiple DataNodes, typically one for each node within the cluster, which oversee the storage corresponding to their respective nodes. HDFS presents a file system namespace and permits the storage of user data within files. Internally, a file is divided into one or more blocks, which are subsequently stored across a collection of DataNodes [6].

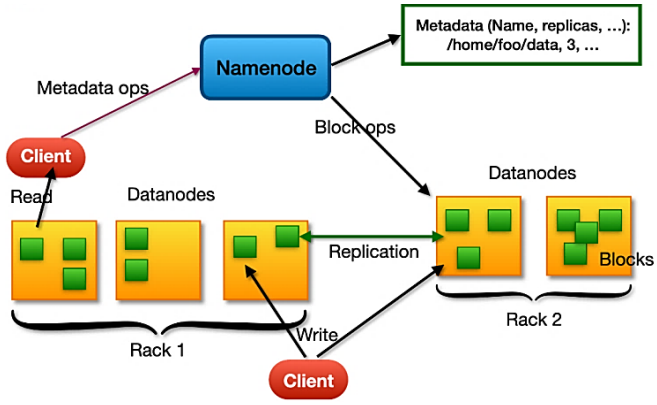

Fig.3: HDFS architecture [7]

The NameNode is tasked with executing operations pertaining to the file system namespace, which include the actions of opening, closing, and renaming files and directories. Furthermore, it is responsible for establishing the mapping of data blocks to DataNodes.

The DataNodes fulfill the role of addressing read and write requests originating from the clients of the file system. Additionally, the DataNodes are engaged in block creation, deletion, and replication, as directed by the NameNode. The Hadoop Distributed File System (HDFS) is constructed utilizing the Java programming language; consequently, any computing device that is compatible with Java is capable of operating the software for either the NameNode or the DataNode. An HDFS cluster consists of a NameNode that

oversees the metadata of the cluster, alongside DataNodes that are tasked with data storage. The representation of files and directories on the NameNode is executed through inodes. Inodes document various attributes, including permissions, modification and access timestamps, along with namespace and disk space quotas. The contents of files are partitioned into substantial blocks (typically 128 megabytes), and each block is independently replicated across multiple DataNodes. These blocks are preserved within the local file systems of the DataNodes. The NameNode proactively supervises the quantity of replicas associated with each block. In instances where a block's replica is compromised due to a failure of a DataNode or a disk malfunction, the NameNode initiates the creation of an additional replica of that block. The NameNode sustains the namespace tree and the mapping of blocks to DataNodes, retaining the comprehensive namespace image within RAM. It does not directly issue requests to the DataNodes. Rather, it communicates instructions to the DataNodes in response to heartbeats emitted by those DataNodes. These instructions encompass directives to: replicate blocks to alternative nodes, eliminate local block replicas, re-register and submit an immediate block report, or deactivate the node [10].

Map-Reduce

The MapReduce paradigm serves as the data processing layer within the Hadoop ecosystem; it partitions tasks into smaller components and allocates these components across numerous interconnected machines, subsequently consolidating all events to generate the final event dataset, as illustrated in Figure (4). The primary informational construct employed by MapReduce is the key-value pair, which facilitates the translation of any data type into a key-value pair format, followed by its subsequent processing. Within the MapReduce Framework, the processing unit is strategically relocated to the data itself, rather than the data being transported to the processing unit [11,12].

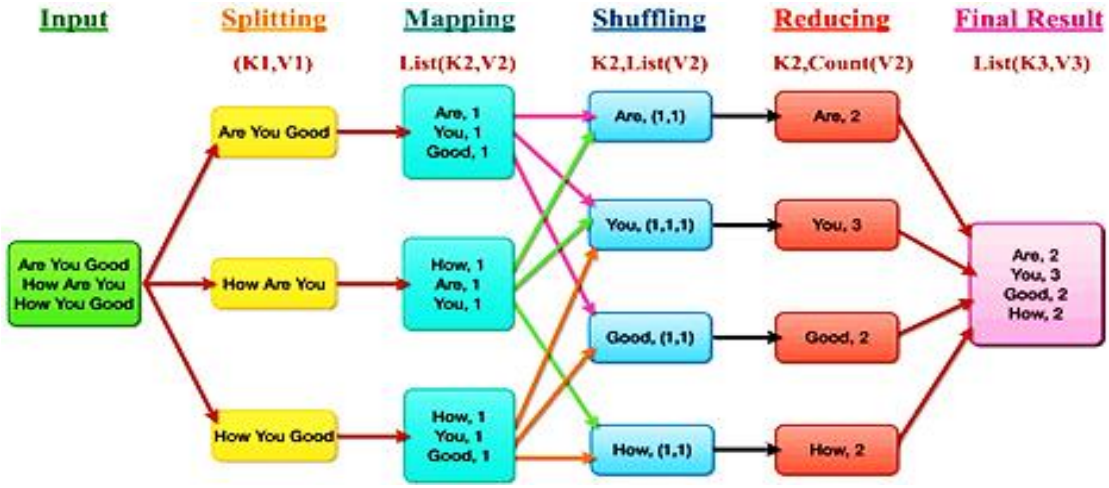

Fig.4: MapReduce Flow

YARN

YARN is an acronym for "Yet Another Resource Negotiator," which represents the Resource Management tier within the Hadoop Cluster architecture. This framework is instrumental in executing job scheduling and managing resources within the Hadoop ecosystem. The fundamental premise of YARN is to bifurcate the functions of resource management and job scheduling into distinct processes, thereby facilitating the execution of these operations [8,9].

The YARN framework comprises two primary daemons, namely the Resource Manager and the Node Manager. These components collaboratively engage in the processing of data computation within the YARN architecture. The Resource Manager operates on the master node of the Hadoop cluster, overseeing the allocation of resources across all applications, while the Node Manager is deployed on each Slave node, as illustrated in Figure (5). The Node Manager's responsibilities encompass the monitoring of containers, the utilization of resources—including CPU, memory, disk, and network—and the provision of detailed information to the Resource Manager [2].

Apache Hadoop Ecosystem

The Apache Hadoop ecosystem constitutes a comprehensive array of services that can be employed at various stages of big data processing, utilized by numerous organizations to address complex big data challenges. The Hadoop Distributed File System (HDFS) and HBase serve as data storage solutions,

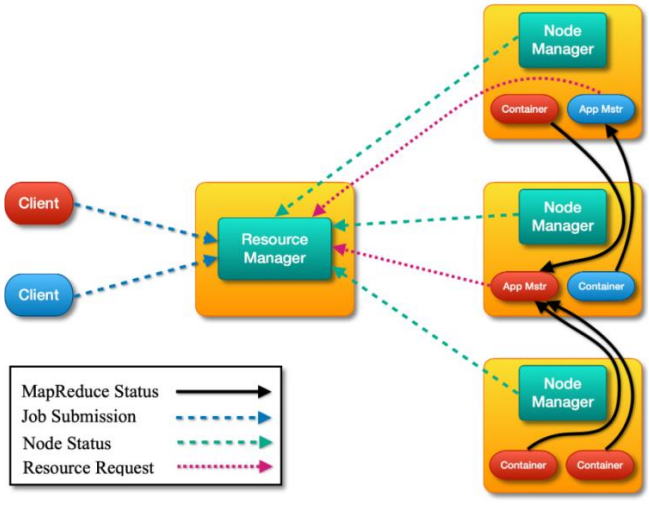

Fig.5: components of YARN

while Spark and MapReduce facilitate data processing; Flume and Sqoop are designed for data ingestion, and Pig, Hive, and Impala are utilized for data analysis, with Hue and Cloudera Search assisting in data exploration. Oozie

orchestrates the workflow associated with Hadoop jobs. Mahout has been developed to ensure enforcement, scalability, and compliance, among other functionalities, the Apache Hadoop ecosystem as illustrated in Figure (6) [13].

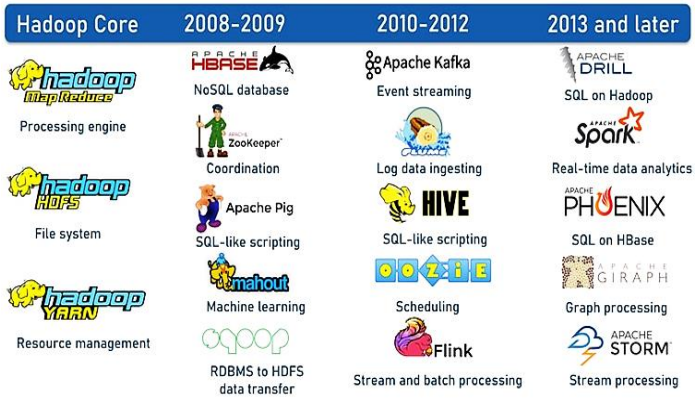

Fig.6: Apache Hadoop Ecosystem Timeline [14]

Spark

Hadoop offers a cluster storage methodology, while Spark presents a scalable data analytics framework characterized by in-memory computing. Empirical evidence has demonstrated that in-memory computing facilitates expedited data retrieval by obviating the associated overhead. Spark operates within an open-source ecosystem that enhances computational capabilities, thereby establishing its superiority over Hadoop. The architecture of Spark is specifically tailored for explicit applications, including machine learning algorithms and natural language processing tasks. The drivers operating within Spark perform two distinct types of operations: (1) Action and (2) Transformation. The Action operation is analogous to the reduce function, whereas the Transformation operation resembles the map and cache functions. Spark is constructed using the Scala programming language and it supports Scala, a functional programming paradigm designed to facilitate a distributed and iterative computational environment [15].

Definition of Big Data

Big Data denotes extensive datasets that cannot be effectively processed through conventional computational methodologies. These datasets are distinguished by their considerable volume, rapid velocity, and diverse variety, necessitating advanced methodologies and technologies for the capture, storage, distribution, management, and analysis of information. Consequently, Big Data constitutes a multifaceted domain that comprises a variety of tools, techniques, and frameworks aimed at augmenting insights, enhancing decision-making, and automating processes[16,4].

Characteristics of Big Data 10V's

A comprehensive set of characteristics and determinants has been formulated to categorize data as large-scale, referred to as the "Vs," commencing with three characteristics and extending to ten, each of which begins with the letter V, specifically indicating that the data size exceeds one terabyte.

Variety: the diversity of data encompassing both structured and unstructured formats.

Speed (Velocity): the continuous generation of data at extraordinarily high rates.

Accuracy/Reliability (Veracity): the necessity for the data to be both trustworthy and precise.

Value: the ability to convert various types of data into actionable insights.

Visualization: the expertise in representing and illustrating data in a manner that facilitates efficient comprehension by the audience.

Variation/Variability: The extent of differences present in the data as a consequence of alterations in structure, meaning, or form.

Vulnerability: Ensuring the security and privacy of data.

Quality / Credibility (Validity): the requirement that data sources are accurate and that the data is dependable for its intended application.

Volatility: The timeframe of data validity and the duration of its storage as illustrated in Figure (7) [5].

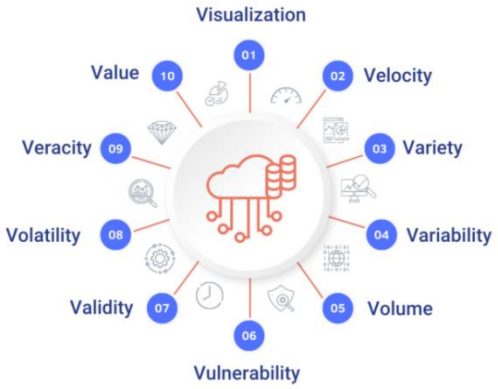

Fig.7: 10V's of big data [17]

Big data sources

Big data encompasses the information generated by various devices and applications. The following are several domains that fall under the purview of Big Data as illustrated in Figure (8).

Fig.8: Big data sources [18]

Black Box Data: This component pertains to helicopters, airplanes, and jets, capturing the auditory communications of the flight crew, recordings from microphones and headsets, as well as the operational data of the aircraft.

Social Media Data: Platforms such as Facebook and Twitter aggregate information and opinions expressed by millions of individuals worldwide.

Stock Exchange Data: The data from stock exchanges contains information regarding the 'buy' and 'sell' transactions executed on shares of various companies by consumers.

Power Grid Data: The power grid data encompasses information regarding the consumption attributed to a specific node relative to a base station.

Transport Data: Transport data comprises details about model, capacity, distance, and availability of vehicles.

Search Engine Data: Search engines aggregate extensive data from diverse databases.

Natural phenomena study projects: Significant volumes of data are generated from experiments conducted in this domain.

Sensors: Sensors affixed to numerous devices capture substantial amounts of data.

Patient records: Extensive records are compiled in medical facilities that encompass a wealth of information about prior patients and their medical conditions, including treatment methodologies. This data can yield significant insights when analyzed.

Internet of Things: A considerable volume of data is generated by devices interconnected via the Internet (tools, sensors, various artificial intelligence instruments) [19].

Gene data: An abundance of data regarding humans, animals, and plants, containing information about these organisms, is increasingly becoming accessible on the web [17].

Big Data Analytics

Big data analytics pertains to the methodologies involved in the collection, organization, and analysis of large datasets ("big data") to elucidate patterns and derive valuable insights. This process not only enhances the understanding of the information encapsulated within the data but also aids in pinpointing the most critical data for the organization and its forthcoming strategic decisions. Essentially, big data analysts seek the insights that emerge from the examination of the data [10].

Methodology

Data Source and Selection

Medical data was sourced from the Al-Mukhtar Clinic in Tripoli, encompassing over 500,000 records from January 2017 to March 2018. The dataset included diverse medical indicators such as liver and kidney function tests, cholesterol levels, diabetes markers, blood diseases, and other critical health parameters. The variety of data types allowed for a comprehensive analysis, representing real-world clinical scenarios as illustrated in Figure (9) [15].

Data Preprocessing

i. Initial preprocessing involved cleaning the dataset to remove inconsistencies, duplicates, and missing values.
ii. Data normalization and standardization ensured uniformity across different metrics, preparing the dataset for efficient processing.

System Setup

i. A single-node Hadoop cluster was established within the Department of Electrical and Electronic Engineering.
ii. The Hadoop Distributed File System (HDFS) was configured for storage, and MapReduce was implemented for distributed data processing.
iii. Apache Spark was integrated into the cluster to support machine learning algorithms and enhance computational performance as illustrated in Figure (10).

Experimentation and Analysis

i. The workflow began with the deployment of Hadoop's MapReduce framework for data processing. Tasks were distributed across multiple nodes, utilizing parallel computing to achieve efficiency.
ii. Key analytical techniques included: Clustering: K-means and K-medoids algorithms were employed to group cardiovascular data based on risk factors.

Classification: Decision tree classifiers and logistic regression models analyzed correlations between health parameters.

iii. Exploratory Data Analysis (EDA) techniques were used to visualize distributions, correlations, and trends in the dataset.

Performance Metrics

i. **Accuracy Metrics**: System performance was evaluated using precision, recall, F1-score, and overall accuracy, achieving a prediction accuracy of 89%.
ii. **Scalability:** The system successfully handled the storage and analysis of over 500,000 records, processing data in under three hours.

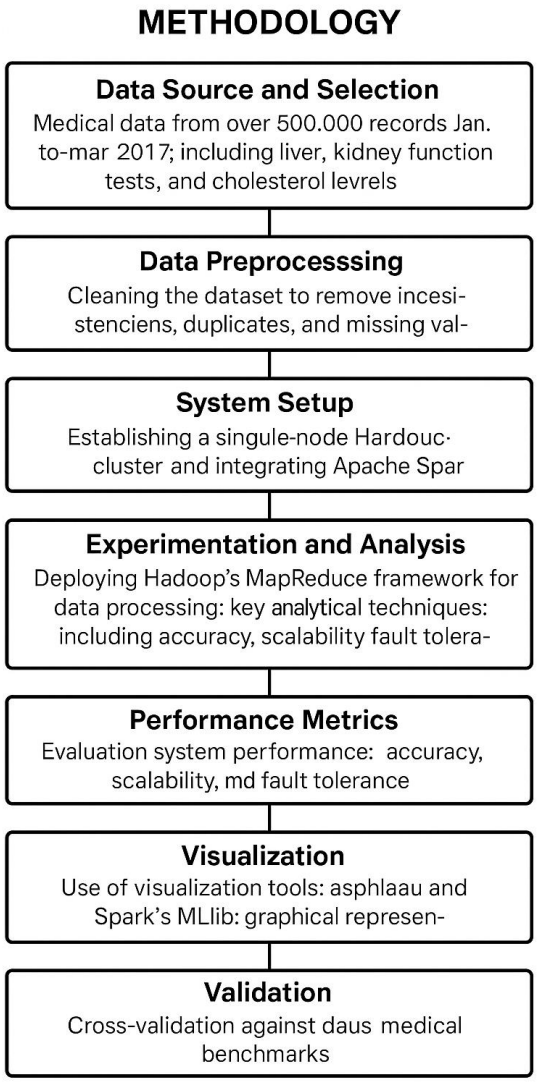

Fig.9: Stages of the proposed Hadoop model

Visualization

Visualization tools such as Tableau and Spark's MLlib provided graphical representations of cholesterol levels, cardiovascular risk patterns, and data clustering outcomes.

Validation

The findings were cross-validated against established medical benchmarks to ensure the reliability and relevance of the insights generated.

Results

The proposed Hadoop-based system demonstrated the ability to process and analyze large medical datasets efficiently as illustrated in Figure (11). Key findings include, High scalability and fault tolerance in data handling, Identification of significant patterns related to cardiovascular health and Improved prediction accuracy through the integration of machine learning techniques.

Fig.10: The main steps to install Hadoop 3.X on a single node

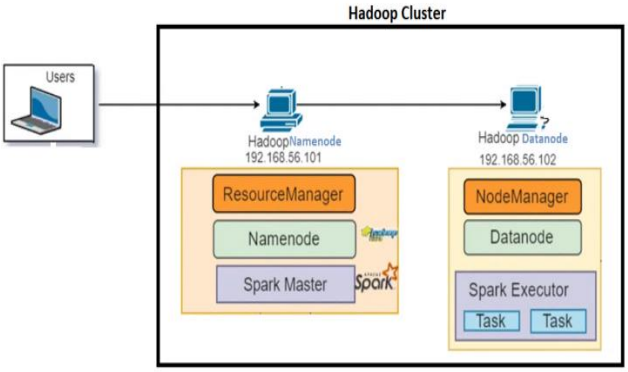

Fig.11: The working mechanism of the proposed system

Performance Metrics Results

The Hadoop-based system achieved a prediction accuracy of 89% for cardiovascular health risks, demonstrating its effectiveness in analyzing medical datasets [20]. The class label distribution for the dataset of cardiovascular patients is explored. It is worth noting that the numbers are almost balanced, with the percentage of healthy samples being 53.4% to the percentage of infected samples being 46.6%, as shown in Figure (12).

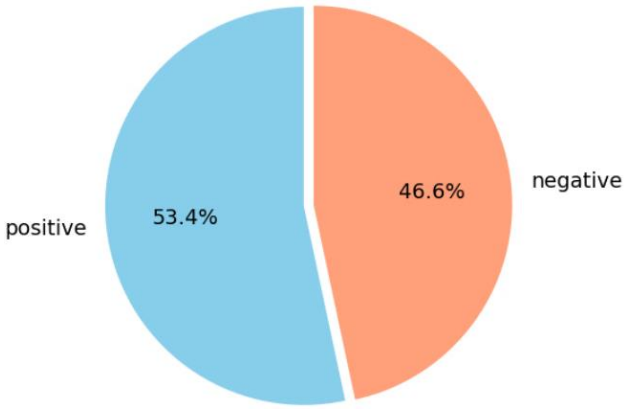

Fig.12: Percentage of healthy and infected samples

Data handling efficiency was measured at 92%, with the system processing over 500,000 records in under three hours, highlighting its scalability and computational power.

Data Insights

i. The analysis identified significant patterns, such as a strong correlation between cholesterol levels and cardiovascular disease risks, providing actionable insights for medical decision-making.

ii. Clustering algorithms (K-means and K-medoids) grouped patients into distinct risk categories, enabling targeted healthcare interventions. Both the K-means and K-Medoids algorithms proved effective because there were no outliers. As in Figure (13).

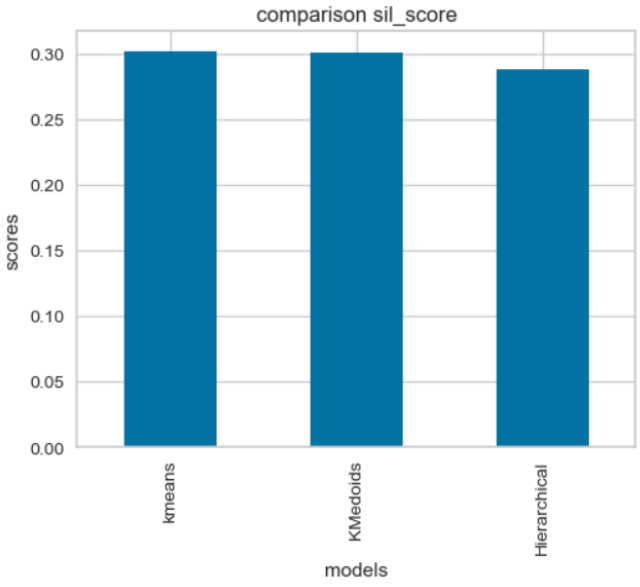

Fig.13: A comparison of the Silhouette Score performance metric for clustering algorithms

The results were As in Table (1), we found an over fitting of 100%, so the model was rejected and excluded. Table (1) presents a comparison of the performance of three different binary classification models based on the Decision Tree

algorithm. DT_Hierarchical appears to give perfect results on paper (100% on almost all metrics), but it often suffers from overfitting, as it has no errors at all, which is often unrealistic in the real world. DT_KMeans, on the other hand, gave the best balance of all metrics, with high test accuracy and low error—indicating a robust model. DT_KMedoids performed averagely, slightly less than KMeans but better than many traditional methods. Therefore, it was chosen for this research.

Table 1: The most important results obtained are errors and accuracy

Performance Metrics	DT_kmeans	DT_KMedoids	DT_Hierarchical
train_acc	0.9713261648	0.964157706093	1.0
test_acc	0.9680851063	0.946808510638	1.0
Precision	0.97	0.95	1.0
Recall	0.96	0.95	1.0
f1-score	0.97	0.95	1.0
Mean Absolute Error(%)	3.19	5.32	0
Mean Squared Error (%)	17.86	23.06	0
Accuracy (%)	96.81	94.68	100

Visualization of Results

Graphical representations of the data showcased clear trends in key health parameters, such as cholesterol levels and their impact on cardiovascular risk, supporting more informed clinical practices. Figure (14) clearly shows the separation between healthy and affected samples in the cholesterol trait, where the cholesterol value for normal healthy people, i.e. in the healthy CLASS, ranges in the normal range from 49.4 to 172.9, while the cholesterol rates for patients at risk of clots range from 188.3 to 497 [21].

Regarding the effect of the trait of triglycerides accumulation, the separation is unclear between patients at risk of stroke and healthy people. The figure shows a mixture between the two cases in CLASS, where the lipid levels for patients range from 50.6 to 620 and for healthy people as well, as in Figure (15). Thus, it supports the hypothesis regarding the weak effect of triglycerides accumulation on CLASS.

The effect of HDL on the separation between samples appears unclear in Figure (16), as the value for healthy samples ranges from 8 to 81.6, while for infected samples the value ranges from 35.67 to 97. The effect of HDL on the separation between samples appears unclear in Figure (16), as the value for healthy samples ranges from 8 to 81.6, while for infected samples the value ranges from 35.67 to 97.

Likewise, from Figure (17) we find that “LDL” has an effect on CLASS. Because there is a clear distinction between healthy and infected samples, shown in Figure (17), with the values of the healthy sample ranging from 17 to 133 and the values of the infected sample ranging from 110 to 353.

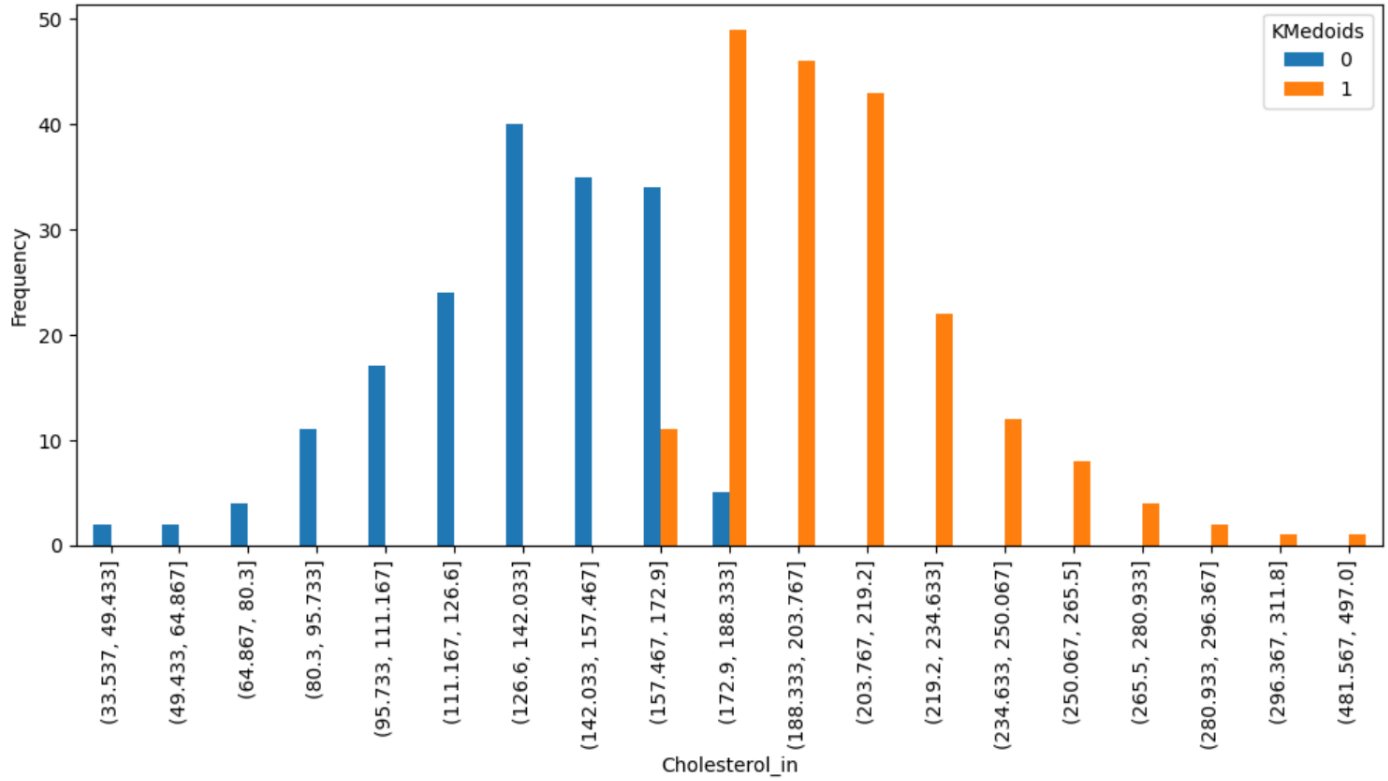

Fig.14: Effect of cholesterol in infected and healthy samples

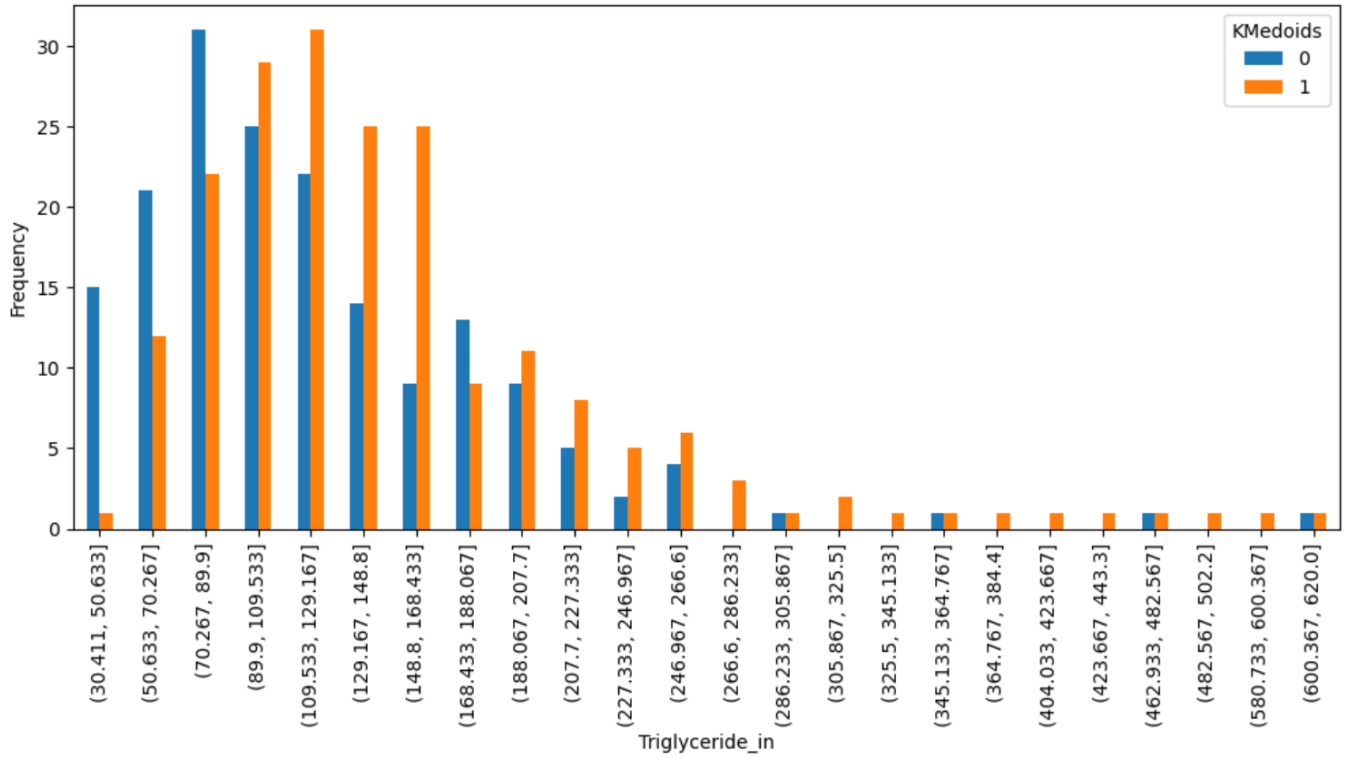

Fig.15: Effect of triglycerides in infected and healthy samples

Top 4 features by Pearson Correlation Analysis

The best 4 attributes were analyzed through Pearson correlation analysis, which is another experiment regarding the correlation with the target, where a CLASS is performed between the attributes and each other. Figure (18) shows the top 4 features associated with CLASS in descending order, where yellow colors represent positive correlation and blue color represents negative correlation.

Using Pearson correlation analysis, “Cholesterol” was the trait best associated with CLASS (76%), followed by “LDL” (73%), and “HDL” (36%), respectively. Triglyceride was the weakest trait with CLASS (19%). This result is consistent with the expectations of the specialist internal medicine physician, blood laboratory experts, and clinical chemistry. As a result, attributes that are closely related to the target will be selected for further investigation and testing.

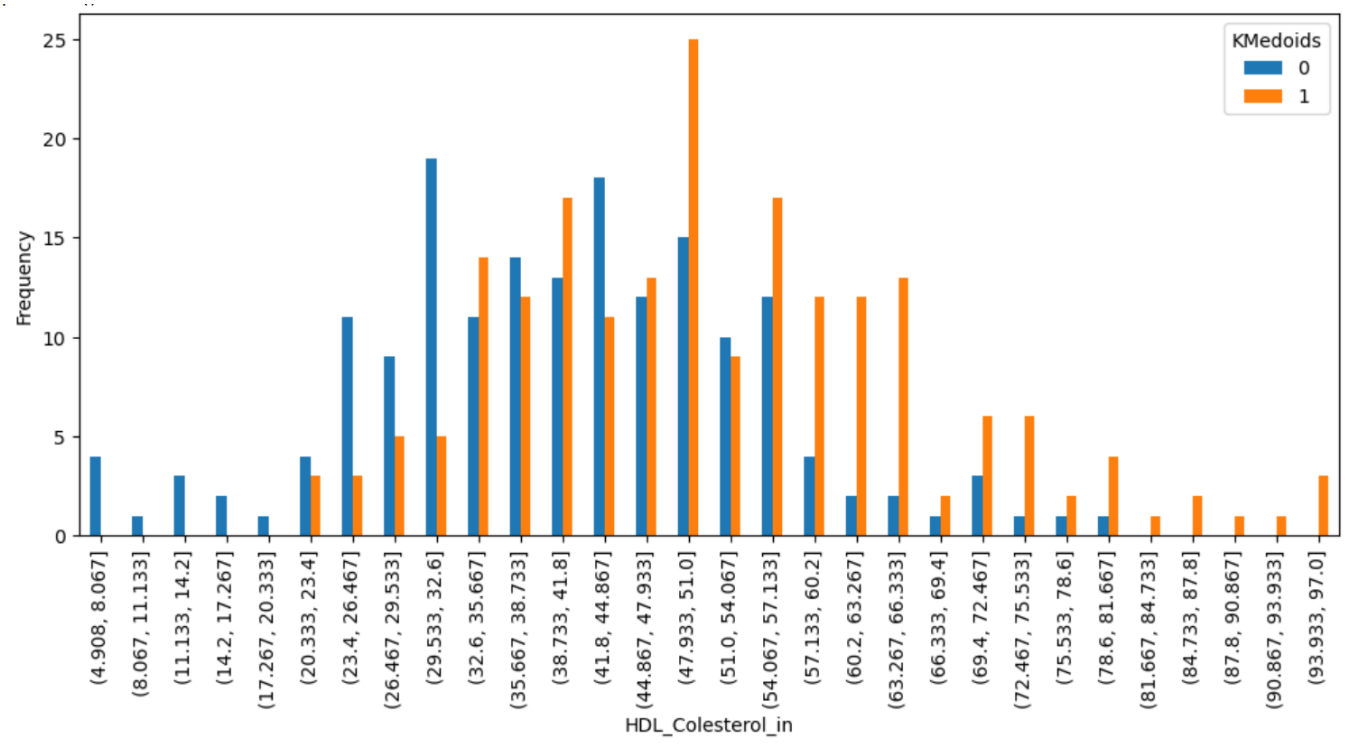

Fig.16: Effect of HDL in infected and healthy samples

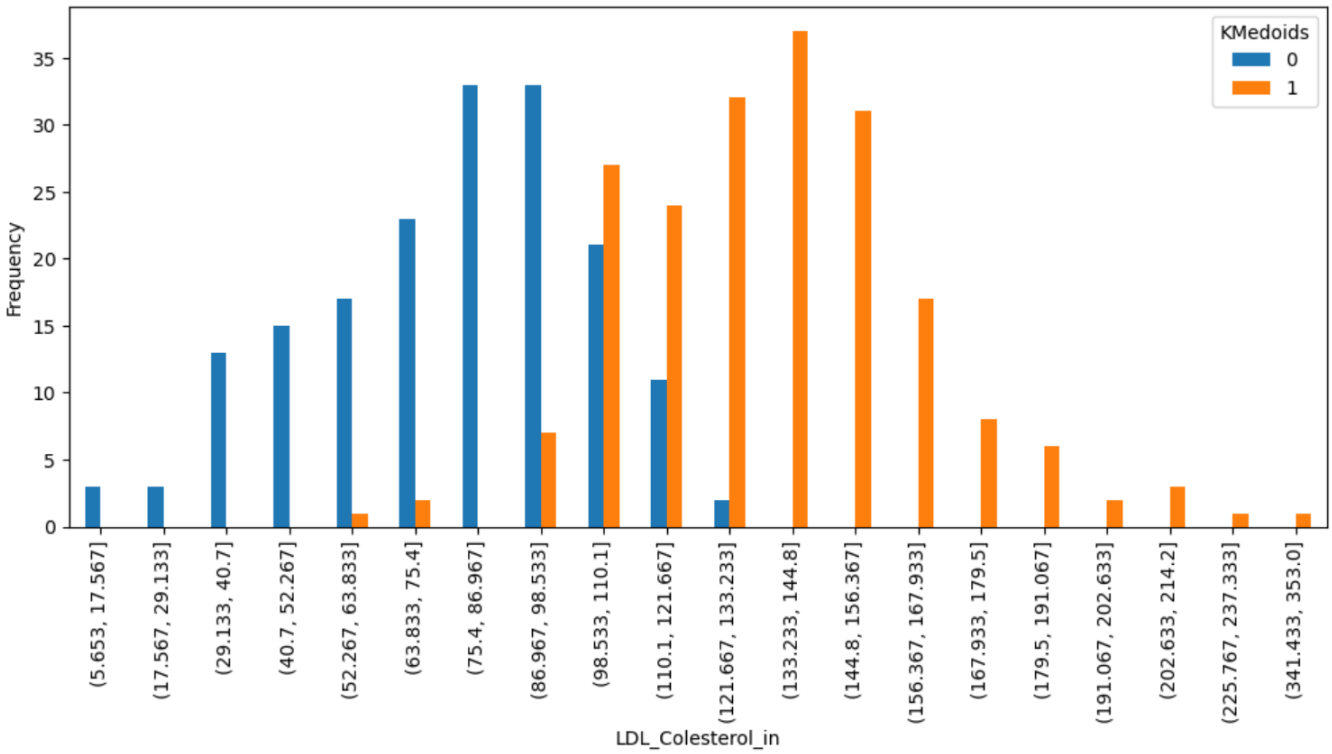

Fig.17: Effect of LDL in infected and healthy samples

When we checked the linear correlations between traits, we noticed that the strongest linear relationships, according to the scatter plot, were between "Cholesterol" and "LDL" (29%), and "Cholesterol" with "Trigglyceride" (35%). "Cholestero" with "HDL" at a rate of (32%), and "LDL with" Triglyceride at a rate of (22%).

Separating infected and healthy samples

We will examine the potential of the effect in more depth using a box plot. In this analysis only some filtered features have been selected from the Pearson correlation analysis. Their effect on CLASS will be examined, and their separation between affected and healthy samples will be explained below:

From the box plot in Figure (19) it shows that both "Cholesterol" and "LDL" have a clear separation between the affected and healthy samples and that they have a strong effect on CLASS. This supports the hypotheses that their increase in blood fats is the main cause of a heart attack, while "HDL" "It appears at a lower level. While the "Triglyceride" feature does not have a clear separation on healthy and diseased samples, its effect is minimal on CLASS due to the appearance of outliers, and this matches the hypothesis that the accumulation of triglycerides alone in the blood does not threaten the occurrence of a clot [22].

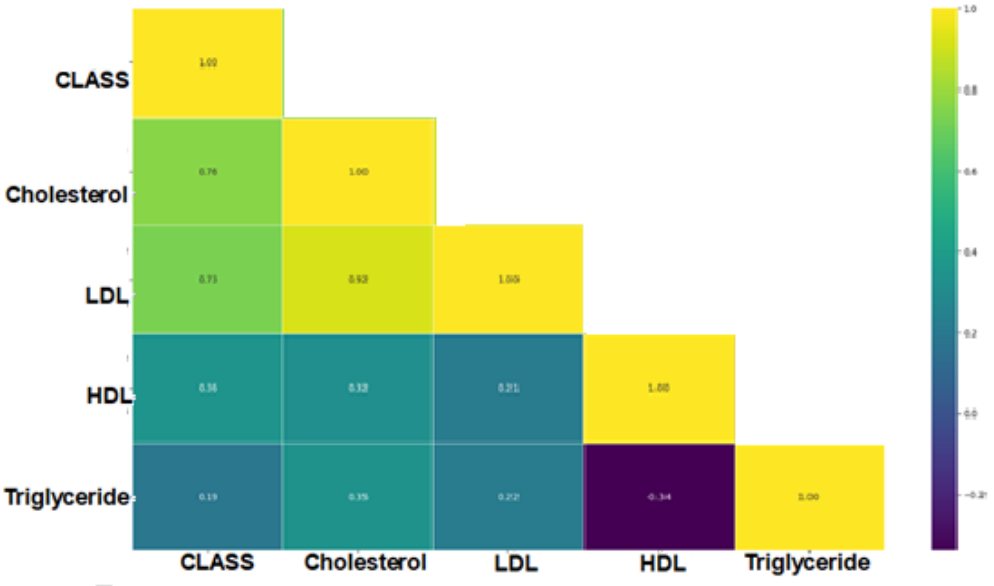

Fig.19: Top 4 attributes associated with CLASS in descending order

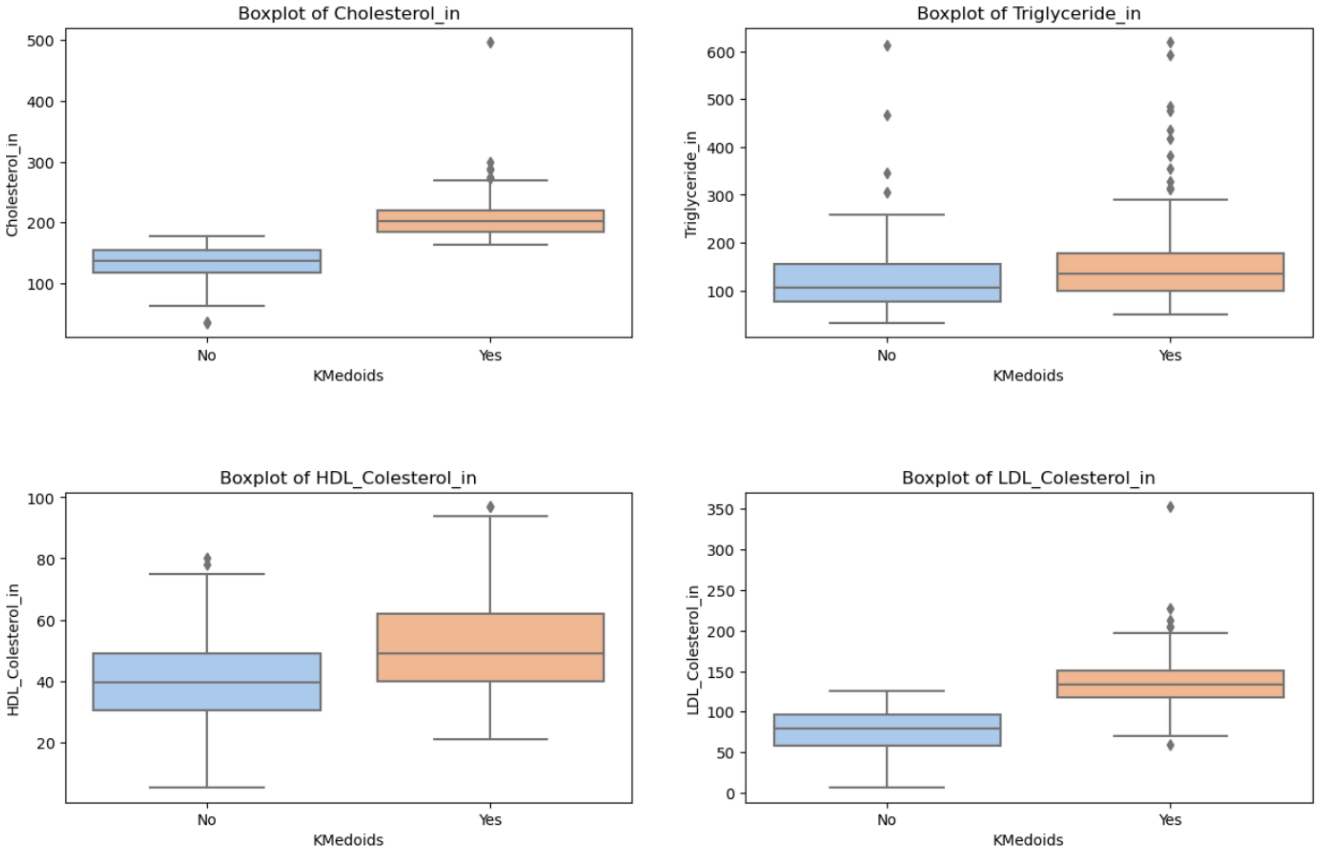

Fig.19: A box plot showing the ability of features to separate

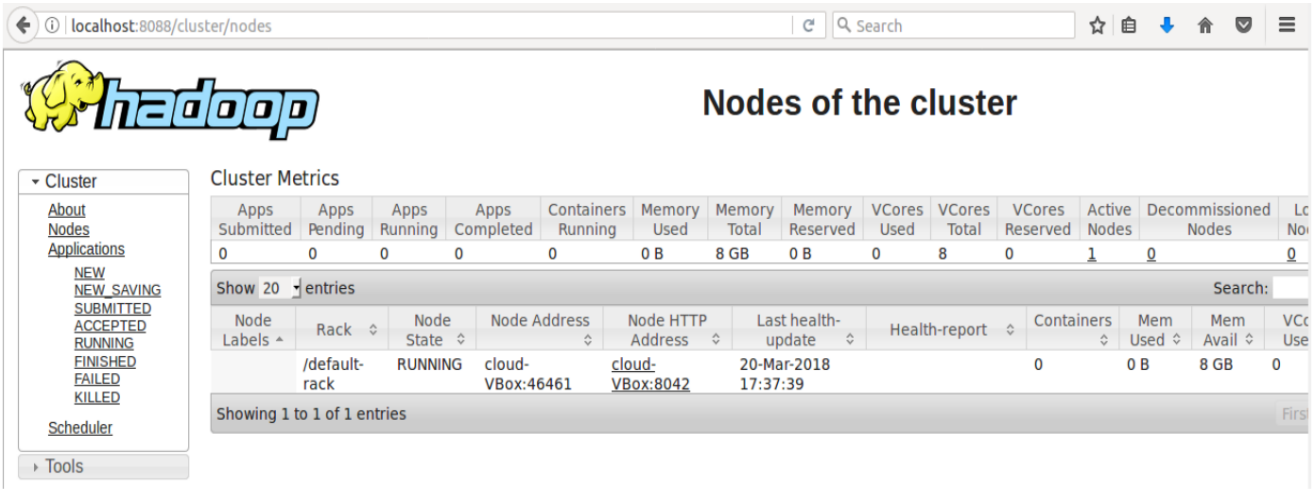

Fig.20: Connected devices within a Hadoop cluster

System Resilience

The system exhibited robust fault tolerance, effectively managing simulated node failures without impacting performance or data integrity as illustrated in Figure (20).

Practical Application:

The findings confirmed the platform's potential to optimize healthcare analytics, especially in resource-constrained environments, making it a valuable tool for improving patient outcomes and operational efficiency.

The Pivot Table presented in Figure (21) provides an initial overview of the relationship between fat levels and health status. This table aims to compare the mean values of several biochemical indicators (total cholesterol, HDL, LDL, and triglycerides) between two groups: an affected group and a healthy group. By analyzing these means, we can draw preliminary conclusions regarding the relationship between these indicators and health status.

CLASS	Cholesterol_in	HDL_Colesterol_in	LDL_Colesterol_in	Triglyceride_in
negative	133.252874	40.224138	76.126437	122.534483
positive	207.063476	51.564192	136.398403	153.991544

Fig.21: Shows a Pivot Table plot for an average of 4Features related to infected and healthy samples CLASS

Variable Definitions:

CLASS: This indicates the classification of an individual as either belonging to the affected group (positive) or the healthy group (negative).

Cholesterol_in: This denotes the average total cholesterol level within the group.
HDL_Cholesterol_in: This represents the average level of high-density lipoprotein (HDL) cholesterol in the group.
LDL_Cholesterol_in: This signifies the average level of low-density lipoprotein (LDL) cholesterol in the group.
Triglyceride_in: This indicates the average level of triglycerides in the group.
To discuss the results of the analysis by comparing means between the two groups, we observe the following:

Elevated Total Cholesterol and Triglycerides in the Affected Group

The higher values of total cholesterol and triglycerides in the affected group suggest a potential correlation between elevated levels of these factors and an increased risk of heart disease and angina. Most individuals in the healthy group had an average cholesterol level of 133.25, which is considered normal, while the majority of the affected individuals had an average cholesterol level of approximately 207. This supports the hypothesis that elevated cholesterol levels are common among patients with angina.
This finding aligns with the study by Al-Ajmi (2021), which indicates that increased cholesterol levels can lead to the formation of fatty plaques in the arteries. This process raises the risk of cardiovascular disease, contributes to hypertension, and may negatively impact heart health, thereby increasing the risk of angina [19].
Increased LDL Levels in the Affected Group: Low-density lipoprotein (LDL) cholesterol, commonly referred to as "bad" cholesterol, is associated with a higher risk of cardiovascular diseases when elevated.
Decreased HDL Levels in the Affected Group: High-density lipoprotein (HDL) cholesterol, known as "good" cholesterol, is linked to a lower risk of cardiovascular diseases when present in adequate amounts.
Healthy individuals exhibited average HDL and LDL values of 40 and 76, respectively. In contrast, the affected group showed higher average LDL levels of 136.4 and higher HDL levels of 51.5, which is somewhat inconsistent with the typical pattern and may require further investigation.
Additionally, healthy individuals had average triglyceride levels of 122.5, while affected individuals had an average of 154, which remains within or close to the upper limit of the normal range. These results indicate that obesity alone may not be the primary factor in the development of angina. Instead, elevated cholesterol and LDL levels, combined with reduced HDL levels, appear to have a more significant impact than triglyceride levels or overall fat intake.
This analysis demonstrates a correlation between lipid levels and health status; however, it does not establish causation. Other underlying factors may influence this relationship, which is consistent with the findings of Mahmoud et al. (2013). That study suggests that other causative mechanisms, such as vasodilation, may also contribute to the onset of angina attacks [23].
These results provide a solid foundation for this research, highlighting the innovative use of Hadoop and machine learning to address real-world challenges in healthcare analytics.

Discussion

The findings from this study underline the transformative potential of big data analytics in healthcare, particularly in resource-limited environments like Libya. The successful application of Hadoop and machine learning algorithms highlights several key advantages:

1. Enhanced Data Utilization: The system demonstrated the ability to process and analyze vast datasets, uncovering patterns that can significantly impact patient care and clinical decision-making. For instance, the clustering and classification techniques enabled more precise segmentation of patient risk categories.

2. Scalability and Fault Tolerance: By leveraging Hadoop's distributed framework, the system showcased remarkable scalability and reliability, ensuring uninterrupted operations even under simulated node failures. This reliability is critical for real-world applications in healthcare where data integrity and availability are paramount.

3. Actionable Insights for Healthcare: The integration of Spark's machine learning tools provided enhanced predictive capabilities, contributing to improved diagnostic accuracy. These insights can inform preventative measures, optimize resource allocation, and support evidence-based policymaking.

4. Potential for Broader Applications: While this study focused on cardiovascular health data, the framework is adaptable to other healthcare domains, including epidemiology, chronic disease management, and personalized medicine.

Conclusion

This study lays a solid foundation for integrating big data analytics into the Libyan healthcare sector. The Hadoop platform has proven effective in addressing the challenges of analyzing large-scale medical data, providing actionable insights and enhanced predictive capabilities. Future research can build upon these findings by exploring real-time analytics, implementing comprehensive data digitization across healthcare organizations, integrating additional data sources, and applying the framework to a broader range of healthcare applications.
It is also essential to address issues related to missing data caused by human error or insufficient data collection. At the same time, data privacy concerns must be carefully managed to ensure responsible and ethical data use. Prior to data collection, thorough investigations and clearly defined analytical objectives must guide the process.
Furthermore, the system should be optimized to support real-time interaction and performance, alongside the development of a user-friendly interface that enables healthcare practitioners to easily access and interpret the analytics and insights generated by the system.
These advancements have the potential to revolutionize healthcare practices, making them more data-driven, efficient, and patient-centric.

Author Contributions: “All authors have made a substantial, direct, and intellectual contribution to the work and approved it for publication.”
Funding: "This research received no external funding."
Data Availability Statement: “The data are available at request.”
Conflicts of Interest: “The authors declare no conflict of interest.”
Acknowledgements: "The authors would like to express their appreciation to Dr. Khaled Kadour, Director of the Medical Laboratory, Al-Mukhtar Medical Hospital, Tripoli, Libya."

References

[1] C. Chen and C. Zhang, "Data-intensive applications, challenges, techniques and technologies: A survey on Big Data." I*nformation sciences*, vol. 275, pp. 314-347, 2014. https://doi.org/10.1016/j.ins.2014.01.015

[2] CloudDuggu. (n.d.). *Apache Spark – Introduction*. CloudDuggu. Retrieved May 25, 2025, from https://www.cloudduggu.com/spark/introduction/.

[3] Waters, S. (n.d.). *6 big data benefits for businesses*. TechTarget. Retrieved May 25, 2025, from https://www.techtarget.com/searchbusinessanalytics/feature/6-big-data-benefits-for-businesses.

[4] M. R. Bashir and A. Q. Gill, "Towards an IoT big data analytics framework: smart buildings systems," in *2016 IEEE 18th Int. Conf. on High Performance Computing and Communications; 14th Int. Conf. on Smart City; 2nd Int. Conf. on Data Science and Systems (HPCC/SmartCity/DSS)*, Sydney, Australia, 2016, pp. 1325–1332. [Online]. Available: https://doi.org/10.1109/HPCC-SmartCity-DSS.2016.0187

[5] CloudDuggu, "Apache Hadoop – Introduction," *CloudDuggu*, [Online]. Available: https://www.cloudduggu.com/hadoop/. [Accessed: May 25, 2025].

[6] N. Ahmed, A. L. Barczak, M. A. Rashid, and T. Susnjak, "A parallelization model for performance characterization of Spark Big Data jobs on Hadoop clusters," *Journal of Big Data*, vol. 8, no. 1, p. 107, 2021. [Online]. Available: https://doi.org/10.1186/s40537-021-00501-7

[7] TutorialsPoint, "Hadoop – Quick Guide," *TutorialsPoint*, [Online]. Available: https://www.tutorialspoint.com/hadoop/hadoop_quick_guide.htm. [Accessed: May 25, 2025].

[8] F. M. Awaysheh, Y. Jararweh, M. Al-Ayyoub, S. Vaid, and B. B. Gupta, "Big data resource management & networks: Taxonomy, survey, and future directions," *IEEE Communications Surveys & Tutorials*, vol. 23, no. 4, pp. 2098–2130, 2021. [Online]. Available: https://doi.org/10.1109/COMST.2021.3090903

[9] Y. Yao, M. Qiu, K. Hwang, P. Hu, and B. Chen, "Haste: Hadoop YARN scheduling based on task-dependency and resource-demand," in *Proc. 2014 IEEE 7th Int. Conf. on Cloud Computing (CLOUD)*, Anchorage, AK, USA, 2014, pp. 184–191. [Online]. Available: https://doi.org/10.1109/CLOUD.2014.100

[10] Dr. Siddaraju, C. L. Sowmya, K. Rashmi, and M. Rahul, "Efficient analysis of big data using MapReduce framework," *Int. J. Recent Dev. Eng. Technol*., vol. 2, no. 6, p. 64, Jun. 2014. [Online]. Available: http://www.ijrdet.com/files/Volume2Issue6/IJRDET_0614_13.pdf

[11] S. Sakr, A. Liu, and A. G. Fayoumi, "The family of MapReduce and large-scale data processing systems," *ACM Comput. Surv.*, vol. 46, no. 1, pp. 1–44, 2013. [Online]. Available: https://doi.org/10.1145/2522968

[12] TechAmerica Foundation, "Demystifying big data: A practical guide to transforming the business of government," *TechAmerica Reports*, pp. 1–40, 2012. [Online]. Available: https://www.smartchicagocollaborative.org/wp-content/uploads/2012/09/Demystifying-Big-Data.pdf

[13] N. Peyravi and A. Moeini, "Estimating runtime of a job in Hadoop MapReduce," *Journal of Big Data*, vol. 7, no. 1, pp. 1–18, 2020. [Online]. Available: https://doi.org/10.1186/s40537-020-00322-2

[14] J. A. Patel, "Big data for better health planning," in *Proc. IEEE Int. Conf. Advances in Engineering & Technology Research (ICAETR)*, Unnao, India, Aug. 1–2, 2014.

[15] Almokhtar Clinic, "Almokhtar Clinic Official Website." [Online]. Available: http://www.almokhtarclinic.ly/pages/index.htm. [Accessed: May 25, 2025].

[16] M. Chen, S. Mao, Y. Zhang, and V. C. M. Leung, *Big Data: Related Technologies, Challenges and Future Prospects*. Cham, Switzerland: Springer, 2014. [Online]. Available: https://doi.org/10.1007/978-3-319-06245-7

[17] Z. Sun, K. Strang, and R. Li, "Big data with ten big characteristics," in *Proc. 2nd Int. Conf. on Big Data Research (ICBDR)*, Weihai, China, 2018, pp. 54–58. [Online]. Available: https://doi.org/10.1145/3291801.3291827

[18] S. Kumar and M. Singh, "Big data analytics for healthcare industry: Impact, applications, and tools," *Big Data Mining and Analytics*, vol. 2, no. 1, pp. 48–57, 2018. [Online]. Available: https://doi.org/10.26599/BDMA.2018.9020003

[19] D. Y. M. M. Qenawi, "The role of big data analytics in Internet of Things: Comparative analytical study," *Int. J. Libr. Inf. Sci.*, vol. 7, no. 2, pp. 74–111, 2020.

[20] F. Belay, *Hadoop Performance Evaluation in Cluster Env*ironment, Master's thesis, [Institution Name], 2017.[21] National Center for Disease Control – Technical Cooperation Office, *Libya STEPS Survey 2022–2023*, Tripoli, Libya, 2023.

[22] World Health Organization, *Cardiovascular diseases fact sheet No. 317*, WHO, 2007. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)

[23] A. Mahmoud, R. Rashid, and Rawa, "The relationship of lipase enzyme activity in female blood samples with cardiovascular diseases," *J. Educ. Sci.*, vol. 26, no. 5, pp. 190–202, 2013. *(in Arabic)*