Big Medical Data Analytics Using Apache Spark Framework

by Dragan Stojanović, Dućan Jovanović, Natalija Stojanović

Presented By:

Y. Sai Bhavya	122ad0007
D. Rohith	122ad0013

Overview

1. Problem Statement
2. Introduction
3. Literature Survey
4. Research Gaps/Limitations
5. Methodology
6. Experiments
7. Future Work
8. References

Problem Statement

- The healthcare industry generates vast amounts of data from EHRs, medical imaging, and clinical trials.
- Traditional data processing tools struggle with handling large and diverse medical datasets efficiently.
- There is a need for scalable, distributed computing frameworks for real-time medical data analysis.
- Apache Spark provides a potential solution for efficient big medical data analytics.

Research Paper: Big Medical Data Analytics Using Apache Spark Framework

Introduction

- Big medical data includes structured, unstructured, and semi-structured data from multiple sources.
- Challenges include data privacy, heterogeneous formats, and real-time analytics needs.
- Apache Spark is a distributed computing framework that efficiently processes and analyzes large-scale medical data.
- This study evaluates Apache Spark's effectiveness using three case studies related to COVID-19.

Literature Survey (1/2)

Year	Author(s)	Key Points	Limitations
2019	Dash et al.	Overview of big data challenges and management in healthcare.	Lacks focus on real-time big data analytics.
2020	Hallman et al.	Predictive models for COVID-19 patient resource allocation.	Does not use distributed computing frameworks like Spark.
2021	Rahman et al.	Used deep learning for COVID-19 detection from chest X-rays.	Limited scalability due to single-machine training.

Literature Survey (2/2)

Year	Author(s)	Key Points	Limitations
2023	Berros et al.	Enhancement of digital health services using big data analytics.	Does not integrate real-time streaming frameworks.
2024	Stojanović et al.	Apache Spark for scalable medical data analytics.	Requires further evaluation on privacy and security aspects.

Research Gaps/Limitations

1. Many studies focus on big data in healthcare, but few implement scalable frameworks for real-time analytics.
2. Existing machine learning models for COVID-19 analysis lack integration with distributed computing frameworks like Apache Spark.
3. Challenges in handling heterogeneous medical data formats.
4. Data privacy, security, and ethical concerns limit access to high-quality datasets.

Methodology

Data Collection

- **COVID-19 Radiography Dataset:** Contains 21,165 chest X-ray images categorized into COVID-19, normal, lung opacity, and viral pneumonia.
- **COVID-19 Report Dataset:** Includes case reports, recoveries, and death statistics from global sources like WHO and John Hopkins University.
- **COVID-19 Diagnosis Dataset:** Consists of anonymized medical records from hospitals, including patient demographics and laboratory results.

Data Preprocessing

- Data cleaning to handle missing values and remove inconsistencies.
- Handling imbalanced datasets using oversampling and undersampling techniques.
- Image preprocessing: Resizing, contrast enhancement, and augmentation.
- Feature extraction for both image-based and tabular datasets.
- Normalization and encoding of categorical variables.

Computational Infrastructure and Apache Spark Framework Implementation

- **Computational Infrastructure:**
  - Spark cluster deployed using Docker containers.
  - Experiments conducted on:
    - Intel Xeon E5-2630 v4 (40 cores, 256GB RAM).
    - Intel Core i7-7700HQ (16GB RAM).

- **Apache Spark-Based Implementation:**
  - Spark-based Extract-Transform-Load (ETL) pipeline.
  - Spark MLlib for machine learning tasks.
  - Spark SQL for querying and processing structured medical data.

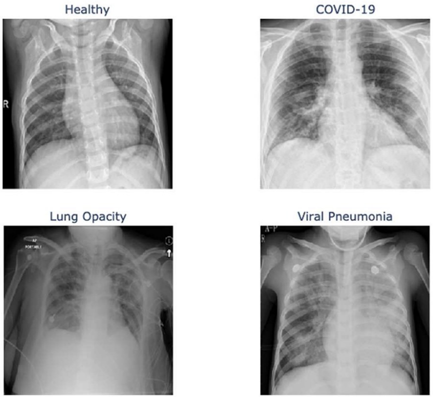
Machine Learning Models Used

- **Radiography Dataset:** CNN with transfer learning (DenseNet-169).
- **Diagnosis Dataset:**
  - Random Forest Classifier (89.03% accuracy).
  - Decision Tree Classifier (88.49%).
  - Logistic Regression (88.81%).
  - Gradient-Boosted Trees (89.97%).

Experiments

Experiment 1: Radiography Data Analysis

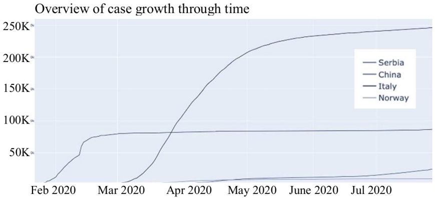
- **Goal:** Classify chest X-ray images into COVID-19, normal, lung opacity, and viral pneumonia.
- **Method:** Convolutional Neural Networks (CNN) using DenseNet-169 model.
- **Results:** Achieved approximately 90% accuracy in COVID-19 detection.
- **Helps in** automating the detection process and reducing diagnostic time for radiologists.



Dataset images

Experiment 2: Daily COVID-19 Report Analysis

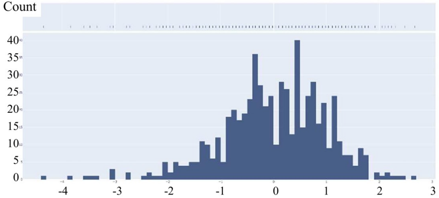
- **Goal:** Analyze the spread of COVID-19 cases over time and make future predictions.
- **Method:** Time-series forecasting using Prophet model and Spark SQL for trend analysis.
- **Results:** Visualization of case growth trends and accurate short-term predictions.
- **Helps in** improving government policies and resource allocation during pandemics.



Overview of case growth through time

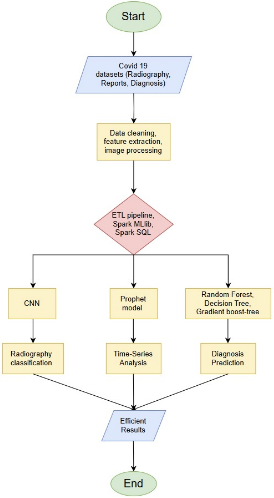
Experiment 3: COVID-19 Diagnosis Prediction

- **Goal:** Predict COVID-19 diagnosis from patient medical records.
- **Method:** Machine learning classification models: Random Forest, Decision Tree, Logistic Regression, Gradient-Boosted Trees.
- **Results:** Gradient-Boosted Trees achieved highest accuracy of 89.97%.
- **Helps in** early detection of COVID-19 cases and assists doctors in decision-making.



Haemoglobin values among the patients

Flowchart


Start
Covid 19 datasets (Radiography, Reports, Diagnosis)
Data cleaning, feature extraction, image processing
ETL pipeline, Spark MLlib, Spark SQL
CNN
Prophet model
Random Forest, Decision Tree, Gradient boost-tree
Radiography classification
Time-Series Analysis
Diagnosis Prediction
Efficient Results
End

Future Work

- Improve real-time analytics using Spark Streaming.
- Integrate deep learning models for enhanced medical image classification.
- Explore federated learning for privacy-preserving medical data processing.
- Optimize Apache Spark for large-scale healthcare applications.
- Investigate hybrid cloud solutions for efficient medical data management.
- Develop automated anomaly detection for early disease diagnosis.

References

Datasets:
1) COVID-19 Radiography Dataset:
https://www.kaggle.com/tawsifurrahman/covid19-radiography-database
2) COVID-19 Report Dataset: https://www.kaggle.com/imdevskp/corona-virus-report
3) COVID-19 Diagnosis Dataset: https://www.kaggle.com/einsteindata4u/covid19

Frameworks:
4) Apache Spark: https://spark.apache.org/
5) Hadoop: https://hadoop.apache.org/

Tools:
6) Apache Spark Docker Deployment: https://github.com/big-data-europe/docker-spark
7) Prophet Time-Series Forecasting: https://facebook.github.io/prophet/
8) Plotly Dash: https://plotly.com/dash/
9) Future SOC Lab: https://hpi.de/forschung/infrastruktur/future-soc-lab.html

Thank You