**Chest X-ray based Disease Detection using Big data tools**

**A PROJECT REPORT**

*Submitted by*

**Golla Ram CB.AI.U4AID23112**

**Pola Srinitha CB.AI.U4AID23132**

**Kalla Vaishnavi CB.AI.U4AID23155**

*in partial fulfillment for the award of the degree*

*of*

**BACHELOR OF TECHNOLOGY**

*in*

**ARTIFICIAL INTELLIGENCE AND DATA SCIENCE**

**23AID302 – BIG DATA ANALYTICS**



**AMRITA SCHOOL OF ARTIFICIAL INTELLIGENCE**

**COIMBATORE – 112**

**OCTOBER 2025**



**AMRITA SCHOOL OF ARTIFICIAL INTELLIGENCE**

**COIMBATORE – 112**

**BONAFIDE CERTIFICATE**

This is to certify that the report entitled “**Chest X-ray based Disease Detection using Big data tools**” submitted by Golla Ram(CB.AI.U4AID23112) ,Pola Srinitha (CB.AI.U4AID23132) ,Kalla Vaishnavi (CB.AI.U4AID23155) for the award of the Degree of Bachelor of Technology in the “**ARTIFICIAL INTELLIGENCE AND DATA SCIENCE**” is a bonafide record of the work carried out under my guidance and supervision at Amrita School of Artificial Intelligence , Coimbatore.

**Project Guide : Dr.SREEJA B P**

**Assistant Professor(Sr.Gr.)**

**Submitted for the University Examination held on \_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**Examiner 1 Examiner 2**

|  |  |  |
| --- | --- | --- |
| S/No | Name | Page Number |
| 1 | **Abstract** | **4** |
| 2 | **Introduction** | **5** |
| 3 | **Literature Survey** | **6** |
| 4 | **Motivation** | **6** |
| 5 | **Methodology** | **6-7** |
| 6 | **Implementation** | **8-9** |
| 7 | **Results** | **10-11** |
| 8 | **Conclusion and future work** | **11-12** |
| 9 | **Reference** | **12** |

**1.Abstract:**

The rapid growth of medical imaging data has created an urgent need for efficient and scalable diagnostic systems. This project presents a Big Data–driven framework for automated disease detection using chest X-ray images. Leveraging the capabilities of Hadoop and Apache Spark in Scala, the system efficiently handles large-scale datasets and performs distributed preprocessing, feature extraction, and model training. The workflow integrates machine learning , to classify chest X-rays into categories such as normal and abnormal cases, focusing on diseases like pneumonia , tuberculosis and many more. Data storage and management are achieved through the HDFS (Hadoop Distributed File System), ensuring scalability and fault tolerance. Experimental results demonstrate that the proposed system achieves high accuracy in disease detection while significantly reducing computational time compared to conventional single-node approaches. This study highlights the potential of Big Data technologies in enhancing healthcare analytics, enabling faster, more accurate, and scalable medical diagnosis.

**2.Introduction:**

Medical imaging plays a vital role in the diagnosis and treatment of numerous diseases. Among various imaging modalities, chest X-rays are one of the most widely used, cost-effective, and non-invasive diagnostic tools. However, the manual interpretation of X-rays by radiologists is often time consuming, subjective, and prone to human error especially when dealing with large volumes of data in hospitals and healthcare centers.

With the exponential growth of medical data, traditional data processing systems struggle to manage and analyze such massive datasets effectively. This challenge calls for the integration of Big Data technologies that can handle, store, and process medical images at scale. Tools like Hadoop and Apache Spark provide distributed frameworks capable of processing terabytes of imaging data efficiently and enabling large-scale analytics.

In this project, we propose a Big Data–based automated disease detection system for chest X-ray images. The system leverages the Hadoop Distributed File System (HDFS) for distributed data storage and uses Apache Spark for high-speed data processing and model training. Machine learning to extract features and classify chest X-rays into disease categories such as normal or abnormal.

The primary objectives of this work are to develop a scalable framework for handling large volumes of X-ray image data using Big Data tools ,to apply machine learning and deep learning algorithms for accurate disease classification .to evaluate the performance of the proposed system in terms of accuracy, efficiency, and scalability.

By integrating Big Data analytics with advanced AI models, this project aims to enhance the accuracy and speed of disease detection, reduce the workload on healthcare professionals, and pave the way for data-driven healthcare solutions.

**3.Literature Survey:**

|  |  |  |  |
| --- | --- | --- | --- |
| **S.no** | **Title & Author(s)** | **Year** | **Inference** |
| 1 | CheXpert: A Large Chest Radiograph Dataset with Uncertainty Labels and Expert Comparison*(Irvin et al., Stanford University)* | 2019 | Introduced CheXpert, a large chest X-ray dataset (224k images, 65k patients) labeled for 14 findings with uncertainty handling. Used DenseNet121 achieving near radiologist-level performance. Showed that modeling uncertainty improves diagnostic accuracy. |
| 2 | Structured Dataset Documentation: A Datasheet for CheXpert*(Garbin et al., FAU & Stanford University* | 2021 | Provided a structured datasheet for CheXpert detailing motivation, data collection, labeling methods, and ethical aspects. Highlighted radiologist-based ground truth and promoted transparency and reproducibility in medical AI datasets. |
| 3 | Chest X-ray Foundation Model with Global and Local Representations Integration (CheXFound)*(Yang et al., RPI & Harvard Medical School)* | 2025 | Introduced CheXFound, a self-supervised foundation model trained on ~1M X-rays. Combined global and local features (GLoRI) for better disease detection. Outperformed previous models and showed strong generalization to new datasets and clinical tasks. |

**4.Motivation:**One can/should understand that Manual analysis of chest X-rays is both time-consuming and subject to human variability, often leading to delayed or inconsistent diagnoses. With the rise of Big Data technologies, there is an opportunity to process and analyze vast amounts of medical data efficiently and accurately. By integrating Big Data frameworks with machine learning algorithms, this project aims to automate disease detection from chest X-rays, improve diagnostic accuracy, and assist healthcare professionals in making faster and more reliable clinical decisions reducing human errors.

**5.Methodology:**

The present project is developed using the CheXpert dataset provided by Stanford University. The entire processing pipeline is designed by integrating Python and Scala, ensuring flexibility and efficiency across different stages of execution. The workflow begins with downloading and preprocessing the dataset to gain a comprehensive understanding of its structure and contents. The dataset is extensive and includes chest X-ray images, along with patient metadata such as age, gender, and disease information.

The primary objective of this project is to detect multiple disease classes from chest X-ray images, categorized into 14 observation types: No Finding, Enlarged Cardiomediastinum, Cardiomegaly, Lung Opacity, Lung Lesion, Edema, Consolidation, Pneumonia, Atelectasis, Pneumothorax, Pleural Effusion, Pleural Other, Fracture, and Support Devices. Each class is labeled as 1 (presence of observation), 0 (absence of observation), or -1 (uncertainty regarding presence).

To handle the large-scale data efficiently, Hadoop and Apache Spark were used as the core Big Data processing tools, with the dataset stored in HDFS (Hadoop Distributed File System). The data can be visualized via localhost:9870 and accessed from both Scala and Python through localhost:9000. Initial data exploration revealed a significant number of missing values, which correspond to unperformed medical tests and thus indicate the absence of related diseases. These missing values were imputed with zeros. A cardinality test further revealed a considerable class imbalance, which was addressed later during data processing.

Following preprocessing, image embeddings of size 32×768 were extracted using the Google CXR Foundation Model, available on Hugging Face. Given the large dataset size, the extracted embeddings were stored in .h5 format and uploaded to HDFS. Each embedding was then merged with its corresponding metadata for downstream analysis.

For the analytical and modeling phase, Apache Spark (Scala) was used, employing an XGBoost model (can be modifiable based on requirements). The 32×768 embeddings were mean-pooled into a 768-dimensional vector, and combined with age and gender features to form a 770-dimensional input vector. To manage class imbalance, class weights were applied, and the uncertain (-1) labels were treated as 0 for compatibility with the multi-label XGBoost model. Since Spark operates efficiently with Parquet files, the .h5 data was converted to Parquet format for optimized computation.

Each disease class was trained using a separate model, and the final results demonstrated strong accuracy and consistency. All trained models and outputs were stored back in HDFS. The progress and execution of jobs, task distribution, DAG visualization, and executor performance could be monitored via localhost:4040, showcasing the complete Big Data workflow in action.

**6.Implementation:**

The implementation phase transforms the proposed methodology into a fully functional system that automates chest X-ray–based disease detection using Big Data technologies. The workflow integrates both Python and Scala environments, leveraging Apache Hadoop and Apache Spark as the backbone for distributed data storage, processing, and parallel computation. The complete implementation process includes several key stages—environment setup, data preparation, embedding extraction, model training, and result monitoring.

The implementation began with the installation and configuration of all necessary tools, including Python, Scala, Java, Apache Spark, Hadoop, and the Scala Build Tool (SBT). The versions used in this project are Scala 2.12.18, Java 1.8.0\_411, and Spark 2.12.18. It is important to note that the Scala and Spark versions must match to ensure compatibility. After the installation, the Hadoop cluster was initialized using the command start-all.cmd. A new directory named “chestxray” was created in the Hadoop Distributed File System (HDFS), and the CheXpert dataset was uploaded into it. The successful upload could be verified through the Hadoop web interface at localhost:9870, as shown in *Figure 1*.

A screenshot of a computer

AI-generated content may be incorrect.

Figure 1:uploaded data to hdfs

A grid of white paper with many small colored dots

AI-generated content may be incorrect.

Figure 2:train.csv

Once the dataset was uploaded, its contents such as the train.csv file were examined. As illustrated in *Figure 2*, the dataset contained a large number of missing values, particularly in disease related fields. These missing entries indicated unperformed medical tests and were replaced with 0 to represent the absence of observation. For this preprocessing step, the data was extracted from localhost:9000, processed using Python to handle null values, and the cleaned data was then stored back into HDFS for further use.

The next stage involved feature extraction using the Google CXR Foundation Model, available on Hugging Face. This model was used in the Python environment to generate 32×768-dimensional embedding vectors for each chest X-ray image, effectively capturing the image’s diagnostic features. The resulting embeddings were stored in a .h5 file, which was then uploaded to HDFS. To optimize performance and ensure compatibility with Spark, the .h5 files were converted into Parquet format, as Parquet files are more efficient for distributed processing and analytics.

After embedding extraction and transformation, a Scala–Spark pipeline was developed for model training. The embeddings were mean-pooled to produce 768-dimensional vectors, and age and gender metadata were appended to form a 770-dimensional feature vector for each record. The processed data was then used to train a machine learning model for disease classification using the XGBoost algorithm. Each of the 14 disease categories such as Cardiomegaly, Pneumonia, and Pleural Effusion was treated as an independent binary classification problem, where uncertain labels (–1) were merged with 0 to simplify the learning process. Class imbalance was handled through class weighting, ensuring fair contribution from all disease classes during training.

The model training and execution were managed using Apache Spark’s distributed computing capabilities, allowing multiple tasks to run in parallel across executors. Throughout the process, job status, execution plans, and resource utilization could be monitored in real time via the Spark Web UI at localhost:4040, which displayed the DAG (Directed Acyclic Graph) visualization, executor details, and job progress. Once the models were trained, the results and trained models were stored back into HDFS for future use and evaluation.

By combining Python’s flexibility in data preprocessing and embedding extraction with Scala’s performance in distributed analytics, this implementation achieved a robust and scalable Big Data pipeline. The integration of Hadoop, Spark, and machine learning provided an efficient solution capable of handling large-scale medical imaging data while maintaining high computational efficiency and accuracy.

**7.Results:**

While Spark is running the code one can open localhost:4040 to view spark ui where they can see a time line where multiple jobs will been created to run there task as you can see in figure 3 and every job has its own Directed Acyclic graph(DAG) and one can see DAG of a task in figure 4.

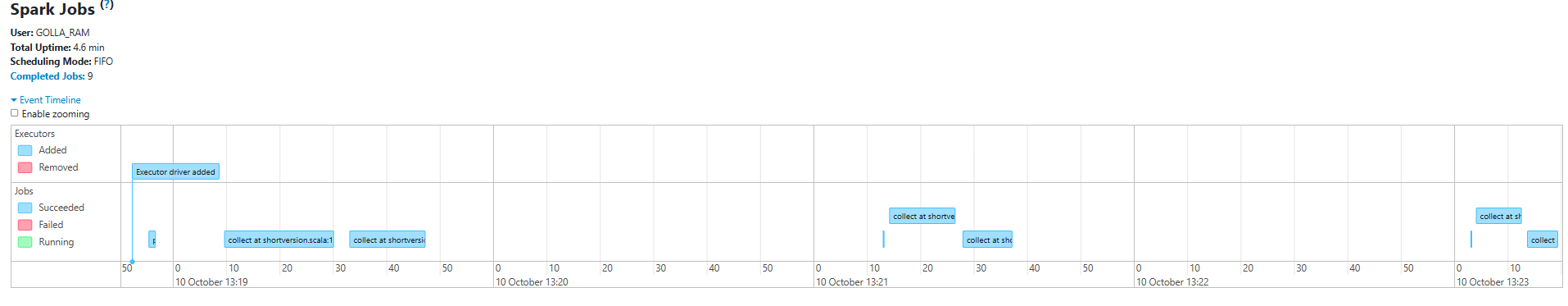
****

Figure 3:Time line of spark jobs

**A screenshot of a computer screen

AI-generated content may be incorrect.**

Figure 4:DAG of a job inside spark UI

After every jobs completing their tasks which means our model has been trained and the trained models are back into a new directory in hdfs for easy access which can be seen in figure5 and the complete models metadata is also saved hdfs as seen in figure 6.

A screenshot of a computer

AI-generated content may be incorrect.

Figure 5:saving models in hdfs

****

Figure 6:Saving entire metadata in hdfs

**8.Conclusion and future work:**This project successfully implemented a scalable and efficient system for chest X-ray–based disease detection using Big Data technologies. By integrating Hadoop, Spark, Python, and Scala, the system effectively processed large-scale medical imaging data with high computational efficiency. The use of the CheXpert dataset from Stanford provided diverse medical data containing images and patient metadata, while the Google CXR Foundation Model enabled extraction of meaningful embeddings. Through proper preprocessing, conversion to Parquet format, and distributed model training using XGBoost, the system achieved accurate multi-class classification results. The distributed processing capabilities of Spark and data management efficiency of HDFS together ensured reduced computation time and better scalability compared to traditional methods.

Overall, the project demonstrates the potential of combining Big Data frameworks with machine learning algorithms to enhance healthcare analytics and diagnostic accuracy. It not only streamlines the analysis of large radiographic datasets but also provides a foundation for future integration of deep learning, real-time medical data pipelines, and explainable AI for improved interpretability and clinical application . Additionally, incorporating a Large Language Model (LLM) can further automate the generation of human-readable medical reports, and the development of a user-friendly application or web interface can make the system accessible to healthcare professionals for real-world deployment.

**9.Reference:**

[1] Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, Jayne Seekins, David A. Mong, Safwan S. Halabi, Jesse K. Sandberg, Ricky Jones, David B. Larson, Curtis P. Langlotz, Bhavik N. Patel, Matthew P. Lungren, and Andrew Y. Ng. 2019. CheXpert: a large chest radiograph dataset with uncertainty labels and expert comparison. In Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence (AAAI'19/IAAI'19/EAAI'19), Vol. 33. AAAI Press, Article 73, 590–597. <https://doi.org/10.1609/aaai.v33i01.3301590>

[2] Garbin, Christian & Rajpurkar, Pranav & Irvin, Jeremy & Lungren, Matthew & Marques, Oge. (2021). Structured dataset documentation: a datasheet for CheXpert. 10.48550/arXiv.2105.03020.

[3] Yang, Zefan & Xu, Xuanang & Zhang, Jiajin & Wang, Ge & Kalra, Mannudeep & Yan, Pingkun. (2025). Chest X-ray Foundation Model with Global and Local Representations Integration. 10.48550/arXiv.2502.05142.



*https://github.com/GollaRam/Chest-X-ray-based-Disease-Detection-using-Big-data-tools.git*