A Study of Lung Disease Using Image Processing in Big Data Environment

Yogesh Kumar Gupta¹, Saroj Agrawal²

¹Assistant Professor, Department of Computer Science, Banasthali Vidyapith, gyogesh@banasthali.in.
²PhD Scholar, Banasthali Vidyapith, sarojagr708@gmail.com.

Abstract. In the medical field, the Image dispensation techniques are extensively employed for image amelioration in finding and treatment of lung disease in the big data environment, where the point in time feature is very paramount to determine the idiosyncrasy issues in intention images, particularly in lung disease such as cancer, pneumonia, COVID-19 etc, for early detection and treatment stages of lungs disease, Image processing technique are widely used for identification of genetic as well as environmental factors are very important in developing a novel method of lung disease prevention. The core factors of this research are quality, time, and precision of the dataset. The modification and evaluation of image quality depend on the segmentation techniques, an improved area of the object that is utilized as a rudimentary substructure of feature extraction is obtained and comparison is made on relying feature. Medical images are analyzed by different segmentation techniques of image processing. The segmentation techniques are used dataset to find patterns and retrieve information from the dataset for processing. The goal of this study is discussed various image processing techniques and big data analytics tools for lung disease has been given in the tabular form and provides comparative study. This study provides minutiae of big data analytics tools and image processing techniques, specifically discussed in the context of lung disease images.

Keywords— Image Processing; Lungs Disease; Big Data.

1. Introduction

1.1 Big Data

It is a data processing application which originally dealt with five basic concepts: velocity, variety, volume, veracity, and value. It is used for working with image data sets. These datasets are very huge and composite. Nowadays, Big Data is used for analysis of user behavior and prediction. Data is composed in formats like structured, semi-structured, and unstructured. It can also be web-generated like audio, video, text and social media files. The data set grows at an exponential rate and is huge in volume, thus posing problems of storage. Hence, traditional tools and algorithms have not been able to compute and manage this complex and voluminous data. The reason data grows enormously is due to its collection by Internet of Things devices. The techniques accustomed to manage and process big data effectively and efficiently are Hadoop, Map Reduce, Apache Hive and such.

1.1.1 Big Data in Medical. Big Data finds immense utility in this field on account of the number of varieties and the speed of data generation. The data includes medical imaging results of Computerized Tomography Scan, X-Ray, Magnetic Resonance Imaging, etc. These medical records along with information like medical research journals, health insurance details etc. The digitization of medical sector using Big Data has opened vast possibilities for healthcare providers, stakeholders and customers. An IBM report suggests that this comprehensive analysis of data can be used in fields like Genomic analytics, Patient Contour analysis, Scam analysis, etc.

1.1.2 Big Data Pattern and Sources. There are multiple sources available for analysis and in this
research; data will be procured from hospitals, diagnostic centre, and relevant areas.

The three categories of Big Data are as follows:

- **Structured Data:** The data that is well organized structured in the form of tables and can be easily processed is known as structured data. This data easily searched and retrieved from an organized database eg. RDBMS Data.
- **Semi-structured Data:** This type of structured data that is unorganized for e.g. Hypertext markup language (HTML), Extensible markup language (XML formatted data), and RDF data.
- **Unstructured Data:** The data that is unorganized and unstructured becomes difficult to operate and requires advance tools and software to access information eg. Medical imaging data, audios, images, videos etc.

1.1.3 *Characteristics of Big Data.* Big Data has five attributes that are known as 5Vs that provides a comprehensive overview. Each of these refers to the hurdles posed in handling and processing of data. They play a vital role in acquisition, recording, sorting, retrieving, analyzing and visualizing the data set. The characteristics are as follows:

- **Volume:** Big Size data is measured in terms of volume; massive amount of data is stored in the type of KB, MB, GB, TB, PB, YB, EB and ZB.
- **Variety:** Data are generated from different diverse sources in different formats so this Heterogeneity is called variety.
- **Velocity:** The massive amount of data generated at rapidly increasing speed from “internet of things” of healthcare leads to increasing velocity of big data in healthcare.
- **Veracity:** Means authenticity, accuracy or uncertainty of data. Data is uncertain due to the incompleteness and inconsistency.
- **Value:** Large amount of data has different kinds of values such as statistical data, events, correlations, hypothetical data etc. [9]

![Five V's of Big Data](image_url)  
**Figure 1:** Five V'S of Big Data
1.2 Tools and Technology To Analyze Data

1.2.1 Hadoop. Hadoop uses many computers and creates clusters to compute data. It works for both single-server and multiple machine systems and stores significant data. The Map Reduced algorithm is used for running the application, and data is processed parallel with others. The master-slave architecture is followed by it and work with Map-Reduce and HDFS layers.

![Hadoop Architecture](image.png)

**Figure 2: Hadoop Architecture**

1.2.2 HDFS Layer. A single node cannot store such large amount of data, and hence HDFS is used as an alternative. It is based on Google File system and splits the data into many small parts. HDFS adopts the replication of records throughout the scattered nodes so that you can acquire the fault tolerance. It follows the idea of a master-slave structure. A single Hadoop cluster is a concoction of one Name Node (as Master Node) and numeral of Slave Nodes. Master Node is accountable to save the metadata like record attributes, call, places of every block cope with and replicas, and many others. Whereas the facts clearly stored inside the Data Nodes, i.e. divided massive documents that are in the shape of blocks or chunks, stored in Data Nodes. It is considered that each chunk of information is replicated over another server node.

1.2.3 Map Reduce Layer. This layer process the large amount of data in disseminated environment and handle clusters of computers in parallel, which is help in system, for reliability and fault tolerance. Map Reduce is commanded on data which is stored in Hadoop Distributed File Structure and work for synchronization and automatic scheduling.

Map Reduce model works with two functions such as Mapper and Reducer. The mapping process, divide the input dataset the outputs comes from the Mapper are passed to the Reducer and a final outcome is acquired with the process of merging.

1.2.4 Apache Spark. Spark is reliable, speedy based on a cluster computing engine. It has the feature of data parallelism and fault tolerance and provides an API in various programming languages such as Java, Python. Apache Spark has the Resilient Distributed Dataset (RDD), a read-only multiset of data objects disseminated On top of the RDD, the Data frame API was released as an abstraction which is followed by the Dataset API.

1.2.5 Apache Pig. An abstraction layer above the Map-Reduce functions well on any data type, size, or location. It is advantageous overall RDBMS and DBMS, making it the most efficient tool. It uses the ETL (Extract, Transformation, and Load) process. It is customarily work with Hadoop. Using Apache Pig, Hadoop can do manipulation operations. It supports nested data structures and reduces the length of code by using a multi-query approach. The Pig scripts are 50% slower in execution, but they make up by increasing the productivity of data engineers as code writing is easier.
Pig Architecture
The two components are Pig Latin language and environment to Pig programs. Pig architecture consists of parser, optimizer, compiler, and execution engine layered below the pig server. Pig script is inscribed by the programmer utilizing the Pig Latin language and implements them by the execution mechanisms (G Shell, UDFs, and Embedded). After that, these scripts engender the desired output and convert into a sequence of Map-Reduce jobs by Pig. Initially, the Parser is handling the Pig Scripts and engenders a directed acyclic graph as output (Pig Latin verbalizations and logical operators). After that, the logical optimizer accomplishes the task projection, and pushdown. Conclusively, all jobs are submitted to Hadoop in sorted order and engender the desired results after compilation of the optimized logical plan into a series of Map-Reduce jobs.

Figure 3: Pig Architecture

2. Digital Image Processing
It is a technique that has gained wide popularity by playing an imperative role in a number of applications. Image segmentation is a focused theme in the image processing field and has influenced development of many algorithms and techniques. These have to be merged with knowledge of domain to obtain a solution for an image segmentation problem.

2.1 Medical Image Processing
This is a method of obtaining images in digital form and enhancing them to extract the desired meaningful information. Image processing involves three steps, are image acquisition, image manipulation, and analysis of an image. It enhances the image and makes the visualization, restoration, and image retrieval processes more efficient and fruitful. Image analysis yields numeric data rather than image and hence meaningful results can be derived. The following are the phases of the image analysis process: [25]

- Image Accretion
- Image Pre-processing
- Image Segmentation
- Feature Selection
- Feature mining
- Image cataloging
- Evolution
3. Inclusion and Exclusion criteria

Systematic evaluation of papers became carried out through meeting the subsequent inclusion criteria: indigenous papers, full text to be had, focused on especially Big Data, Image Processing, lung diseases; Pneumonia, COVID19, papers from 2008 to 2020, comparing novel assets of information like Google Scholar, Wikipedia access logs, Health Map, WHO, and so on. Moreover, exclusion criteria have been: papers with spurious or bogus content material, without unique facts, no filters have been applied; papers published not before 2008, for lung sicknesses best. Every taken into consideration paper turned into downloaded with full text with desired titles most effective, and a take a look at was made on reference listing to discover the genuineness of the selected papers.

Table 1: Search Procedure For articles

| Input               | Search Procedure Details                                                                 |
|---------------------|------------------------------------------------------------------------------------------|
| Databases           | Google Scholar, Research Gate, Scopus, Springer, IEEE, Science Direct, Scopus,(IEEE Explore Digital Library). |
| Keywords            | Lung diseases, Image processing, Image Segmentation Techniques, Big Data Analysis Tools, healthcare, COVID-19. |
| Time Filter         | 2008-2020                                                                                |
| Inclusion Criteria  | Indigenous articles addressing and targeting the desired criteria like Tools and Techniques of Big data and Image Processing and lung diseases. |
| Exclusion Criteria  | Researches not specifying the included data related Big data and Image Processing and lung diseases. |
| Journals Approached | Int J of Innovative Technology and Exploring Engg., Journal of Digital imaging, Int J of Engng. Trends and Technology, Int J of Computer Science & Information Technology, Int J of Advanced Computer Science and Applications, Int J of ICIC, International Journal of Pure and Applied Mathematics |
4. Literature Survey

In this research, survey is conducted through big data and image processing to study lung disease. Research papers are studied from 2008 to 2020 to prepare list of various features from lungs images in order to easily diagnose the lungs diseases.

Table 2: Summary of Research Paper for Lungs Disease Using Image Processing & Big Data Analytics

| S. No | Author Name | Year of Publication | Algorithm/ Techniques | Results / Observation |
|-------|-------------|---------------------|-----------------------|----------------------|
| 1     | Esra Gülay et al. [7] | 2020 | Active contour model on up to 5 year children’s X Ray images of pneumonia infected | Find lung sizes of healthy individuals and patients with pneumonia with ACM model |
| 2     | Gaspard Harerimana et al. [8] | 2018 | Data Mining Technique on Medical data of IOT And Cyber-Physical (Cpa) Data | To develop a health analytic application and Big data is utilized as substrucoces for health analysis. |
| 3     | Ifeyinwa Angela Ajah et al. [2] | 2018 | Big Data on Different Types Data | Prosperous business strategies It has prosperously availed the organization to achieve cost reductions, more expeditious and better decisions |
| 4     | A. C. Priya Ranjani et al.[19] | 2018 | Apache Pig on WEB LOG DATA | Discovering the hidden patterns in millions of web records |
| 5     | S.Perumal et al.[17] | 2018 | Image filtering techniques like Wiener, Median, Gaussian on Lungs Images | Enhance the quality of the lungs images by use of preprocessing techniques. |
| 6     | Swarna C et al. [5] | 2017 | Apache Pig | Gives the importance of Apache Pig and explore the concept of apache pig. |
| 7     | Fang Yang,1 Murat Hamit et al.[22] | 2017 | Sequential forward selection and principal component analysis methods on Esophageal X-Ray Images data set | Developed to assist medicos in perceiving the feature of digital X-ray image and improving the quality of analysis |
| 8     | D. P. Acharjya et al.[1] | 2016 | Big Data on Different Types Data generated by IOT and Cloud computing. | Research aspect, challenges, and tools are analyze for these big data |
| 9     | Balasubramanian Ramasamy et al.[18] | 2016 | Various Segmentation Techniques on CT and CXR medical images | To diagnose patterns and to repossess information from medical images |
| 10    | A R Amanda et al.[4] | 2016 | Threshold technique on Lungs Images | Threshold segmentation method give better quality by smallest MSE value and the highest PSNR |
| 11    | Abdul Ghaffar Shoro et al. | 2015 | Apache Spark on Twitter twits, | Explore the concept of Big Data |
|   | Authors                          | Year | Techniques/Approach                                           | Description                                                                                                                                                                                                 |
|---|----------------------------------|------|---------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| 12| Ramsingh Jayaraman et al. [14]   | 2015 | Apache Pig tool is used on Library data set                   | The library user frequently analyzed, and the books accessed regularly, authors preferred usually by the students, and most conventional date and time of students are analyzed.   |
| 13| Nida M. Zaitoun et al. [23]      | 2015 | Use Various Segmentation techniques                           | Study of Block based techniques                                                                                                                                                                          |
| 14| Awais Mansoor et al. [15]        | 2015 | Lungs segmentation methods: (a) Thresh holding, (b) Region, (c) Shape, (d) Neighboring anatomy, and (e) machine learning–on CT Scan lungs images | For pulmonary analysis, this is provide the diagnostic approach to medicos for the proper selection of automated segmentation tools                                                                         |
| 15| Shilohi Elizabeth Darmanaygam et al. [6] | 2013 | Segmentation approach for lung parenchyma from chest using iterative thresh holding followed by morphological operations | Thresh holding and morphological operations are proven better                                                                                                                                         |
| 16| Nidhi Singh et al. [21]          | 2012 | The fuzzy C-means algorithm (FCM) on Flower Images            | To analyze appropriate features newly received images                                                                                                                                                    |
| 17| Jude Hemanth D et al. [12]       | 2012 | morphology based techniques on MR brain Images                | Developed to abstract the skull portion circumventing the encephalon tissues and texture predicated feature extraction techniques are utilized                                                                  |
| 18| Ashraf A. Aly et al. [3]         | 2011 | Various Segmentation Techniques on Digital Images             | This study is subsidiary for formative the appropriate utilization of the image segmentation methods for proving their better precision and performance                                                          |
| 19| M. Arfan Jaffar et al. [13]      | 2009 | Image processing algo like Fuzzy C-Mean (FCM) and morphological on lungs images data set | Using these algorithms, System is accomplished to perform automatic segmentation of lung images.                                                                                                           |
| 20| Nisar Ahmed Memon et al. [16]    | 2008 | The algorithms are: (1) the Thresh holding and region growing algorithm (2) The Thresh holding and morphology algorithm (3) The Thresh holding and rolling ball algorithm | Find the limitation of three algorithms is used.                                                                                                                                                     |
Various algorithms used for the analysis on medical images: GLCM, Edge Detection and image Enhancement Techniques, Wavelet Transform, Hadoop Image Processing Interface (HIPI) etc.

Use image processing techniques and hadoop framework author get the result enhanced images. They were get the hidden features that help to identify the early detection of diseases in hadoop environment.

### 5. Challenges

After reviewing copious papers related to lungs diseases using Image Processing and Big Data Analytics, a common issue reported i.e. scantiness of data has now been addressed. It is a challenge for radiologist to diagnose the precise level of infection from lung images. Moreover, the survey also addressed the fact that not every paper included all aspects of data features as some of them focused on Lung infection only while no one took all the factors concurrently to predict the higher accuracy, resulting in incomparable data sets to see the performance of Image processing techniques and Big Data analytics.

### 6. Conclusion

In this paper various big data analytics tools and image processing and methods for lung disease are analyzed, all methods work glowing for different purposes. The algorithms are used for lung disease and their results are discussed in the table. The image segmentation algorithms are satisfied with many features of the lung disease images. The lung disease as a major threat, leveraging the morbidity and mortality scenario, a survey on papers has been undertaken for prediction of lung disease, with 8 papers from the last 12 years. It is also pertinent to note here that the recent outbreak of COVID-19 (coronavirus disease 2019) disease caused by stern sensitive respiratory disease coronavirus 2 (SARS-CoV-2) leads to pneumonia, and infected the tiny air sacs (called alveoli) inside lungs causing pulmonary fibrosis. The cure for this disease is currently under the research phase. However, with the efficient use of the aforementioned data analytics tools may be carried out for corona virus-infected lung images. This study is mainly describing the fundamental concept of Big Data, image processing techniques along with lung disease images.

The paper also focuses on analyzed lung disease images by image processing techniques in a big data environment to take out the significant information from the unattended portion of non-invasive lung disease imaging modularity, with different formats.

### References

[1] Acharjya D. P., Ahmed P, 2016, “A Survey on Big Data Analytics: Challenges”, Open Research Issues, (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 7.

[2] Ajah Ifezinwa Angela, Nweke Henry Friday, (2018), “Big Data and Business Analytics: Trends, Platforms, Success Factors and Applications”.

[3] Aly Ashraf A. et al, 2011, “Research Review For Digital Image Segmentation Techniques, International Journal of Computer Science & Information Technology (IJCSIT) Vol. 3, No 5.

[4] Amanda A R, Widita R, 2016, “Comparison of image segmentation of lungs using methods: connected threshold, neighborhood connected, and threshold level set segmentation”.

[5] C Swarna, Ansari Zahid, 2017, “Apache Pig - A Data Flow Framework Based on Hadoop Map Reduce”, International Journal of Engineering Trends and Technology (IJETT) – Volume 50 Number.

[6] Darmanayagam Shilohah Elizabeth, et al, 2013, “A Novel Supervised Approach for Segmentation of Lung Parenchyma from Chest CT for Computer-Aided Diagnosis”. Springer.

[7] GülEsera, İçerSemra, 2020, “Evaluation of Lung Size in Patients with Pneumonia and Healthy Individuals”, European Journal of Science and Technology, Special Issue, pp. 304-309.

[8] Gaspard Harerimana et al, 2018, “Health Big Data Analytics: A Technology Survey”, Digital Object Identifier 10.1109/ACCESS.2018.287825.
[9] Gupta, Y. K., et al., 2016, “Study of big data with medical imaging communication”, International conference on communication and computing systems (ICCCS-2016), By CRC Press (Taylor & Francis group) ISBN-978-1-138-02952-1 Pp. 993-997.

[10] Gupta Yogesh Kumar, Sharma Shruti, 2019, “Impact of Big Data to Analyze Stock Exchange Data Using Apache Pig”, International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-8 Issue-7.

[11] Hemant D Jude, Anitha J, 2012, “Image Pre-processing and Feature Extraction Techniques For Magnetic Resonance Brain Image Analysis”, Article in Communications in Computer and Information Science DOI: 10.1007/978-3-642-35594-3_47.

[12] Jaffar M. Arfan et al, 2009, “Fuzzy Entropy And Morphology Based Fully Automated Segmentation Of Lungs From Ct Scan Images”, International Journal of Innovative Computing, Information and Control ICICI, Volume 5, Number 12(B) ISSN 1349-4198.

[13] Jayaraman Ramsingh, Velumani Bhuvaneswari, “An Insight on Big Data Analytics Using Pig Script”, International Journal of Emerging Trends & Technology in Computer Science. (2015).

[14] Mansoor Aways et al, 2015, “Segmentation and Image Analysis of Abnormal Lungs at CT Current Approaches, Challenges”, and Future Trends1, Radio Graphics 2015; 35:1056–1076 Published online 10.1148/rfat.2015140232.

[15] Memon Nisar Ahmed et al, 2008, “Deficiencies of Lung Segmentation Techniques using CT Scan Images for CAD”, World Academy of Science, Engineering and Technology International Journal of Computer and Information Engineering Vol: 2, No: 8.

[16] Perumal S, Velmurugan T, 2008, “Preprocessing by Contrast Enhancement Techniques for Medical Images”, International Journal of Pure and Applied Mathematics, Volume 118 No. 18 2018, 3681-3688, ISSN: 1311-8080, (2008).

[17] Ramasamy Balasubramanian et al, 2016, “A Survey: Analysis on Pre-processing and Segmentation Techniques for Medical Images”, International Journal of Research and Scientific Innovation (IURSI), Volume III, Issue VI, ISSN 2321–2705.

[18] Ranjani A. C. Priya, Sridhar M, 2018, “Analysis Of Web Log Data Using Apache Pig In Hadoop”, IJRAR, Volume 5 I Issue 2, ISSN 2348–1269.

[19] Shoro Abdul Ghaffar, Soomro Tariq Rahim, 2015, “Big Data Analytics: Apache Spark Perspective”, Global Journal of Computer Science and Technology: C Software & Data Engineering Volume 15 Issue 1 Version 1.0.

[20] Singh Nidhi et al, 2012, “A Novel Approach for Content Based Image Retrieval”, Elsevier, Procedia Technology 4 (2012) 245 – 250.

[21] Yang Fang, Hamit Murat et al, 2017, “Feature Extraction and Classification on Esophageal X-Ray Images of Xinjiang Kazak Nationality”, Hindawi Journal of Healthcare Engineering Volume 2017, Article ID 4620732, 11 pages.

[22] Zaitoun NidaM et al, 2015, “Survey on Image Segmentation Techniques”, Elsevier, International Conference on Communication, Management and Information Technology (ICCMIT), Procedia Computer Science 65,797 – 806.

[23] Gupta, Y.K., 2020,” Aspect of Big Data in Medical Imaging to Extract the Hidden Information Using HIPI in HDFS Environment. Advancement of Machine Intelligence in Interactive Medical Image Analysis”, Algorithms for Intelligent Systems (AIS). Springer, Singapore, https://doi.org/10.1007/978-981-15-1100-4_2, pp. 19-40.

[24] Gupta, Y. K. et al, 2020, “A Study of Big Data Analytics with Two Fatal Diseases Using Apache Spark Framework”, International Journal of Advanced Science and Technology (IJAST), Vol. 29, No. 5, pp. 2840 - 2851.

[25] Gupta Y.K., 2019, “Empirical Aspect to Extract Hidden Features from Unattended Portion of Neuroimaging Modalities for Early Detection of Diseases”, International Conference Advanced Informatics for Computing Research (ICAICR). Communications in Computer and Information Science (CCIS), DOI: https://doi.org/10.1007/978-981-15-0108-1_21. ISBN: 978-981-15-0108-1. Springer, Singapore. ISSN: 1865-0929, 1075, pp 217-226.

[26] Gupta, Y. K., 2019, “Role of Big Data in Medical Imaging Modalities to Extract the Hidden Patterns Using HIPI in HDFS Environment”, 2019 10th International Conference on Computing, Communication and Networking Technologies (IICCNT), IIT Kanpur, India. IEEE Explore. DOI: 10.1109/IICCNT45670.2019.8944666. pp. 1-6.

[27] Maheshwari, A. & Gupta, Y. K., 2018, “Empirical Aspect of Big Data to Enhance Medical Images Using HIPI”, Proceedings of the Second International Conference on Intelligent Computing and Control Systems (ICICCS), ISBN: 978-1-5386-2842-3 Pp. 215-220.