WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Automated Detection of Lung Cancer Based on Machine Learning Algorithms

Chandrappa S [a], Prithviraj Jain [b]

[a] *Dept. of Information Science and Engg., GSSIEWT, Mysore, India**

[b] *Dept. of Computer Science and Engg., SDMIT, Ujire, India*

A B S T R A C T

Cancer is a disease that can be cured if it is diagnosed at an early stage while it is still treatable. According to the findings of a study, the number of people who pass away each year is rapidly increasing for the same reason—an increase in the prevalence of cancer. There are many different forms of cancer, but a recent survey conducted in India and throughout the world indicated that lung cancer is the second most hazardous illness that is responsible for the most fatalities. If detected and treated too late, lung cancer is the most prevalent kind of cancer that is invariably deadly. If the condition could be discovered at an earlier stage, before it reached its severe degree, then there is a greater chance that it might be effectively treated and identified. Several research has been carried out to investigate the use of machine learning and artificial intelligence in the early detection of cancer. This is being done in the hopes that we may be able to cure patients and, to some degree, save their lives. The combination of a biological image processing technology with knowledge detection of data has led to the development and implementation of a number of different methodologies. As part of this body of research, we have implemented a few machine learning algorithms onto a dataset including information on lung cancer illness. The study of lung cancer in its earliest stages is the primary focus of our current research. To do this, we are evaluating the effectiveness of a variety of machine learning methods. We have summarized the results of our research into the benefits and drawbacks of each methodology, as well as the numerous datasets that were utilized.

Keywords: Lung Cancer, Machine Learning, Image Processing, Natural Language Processing

1. Introduction

To this day, lung cancer remains the largest cause of mortality resulting from a tumor anywhere in the globe. In the meanwhile, it would appear that perhaps the incidence has been climbing at a constant and steady pace [1]. The unchecked expansion of tissues in the lung is what leads to the development of lung cancer. The vast majority of instances are brought on by the usage of tobacco products [2]. Additional risk factors for developing lung cancer include being around asbestos, radon, uranium, and arsenic. There is a high mortality rate associated with lung cancer since the illness has the potential to spread to other areas of the body, including the brain, liver, bone, and bone marrow [3].

Due to the fact that it can make the following clinical care of patients much easier, the early detection and prognosis of lung cancer has become an absolute requirement in the field of cancer research [4]. The primary area in which the machine learning algorithms can be of assistance is in the diagnosis of lung cancer. By extracting a simple and easily comprehensible model for lung cancer from a medical record, the strategy can drastically cut down on the likelihood of developing a condition [5].

The detection of lung cancer through the application of machine learning algorithms is the primary focus of our work. The term "machine learning" refers to the process of programming a computer to have the ability to maximise performance based on a criterion utilising data or previous experience. It is often regarded as the most innovative and forward-thinking type of technological development. The capacity of machine learning to analyse vast volumes of data and identify patterns in real time is assisting in the development of a new methodology for the solution of problems. Cancer is the leading cause of mortality across the board, affecting both men and women equally. Cancer patients who catch the disease in its earlier stages may have a better chance of beating it altogether. Cancer patients who catch the disease in its earlier stages may have a better chance of beating it altogether.

Chest x-rays, which are part of the current diagnostic process, are not very good at identifying lung cancer in its early stages. This is a limitation of the current method. The diagnosis of lung cancer is now being assisted by a use of machine learning within the field of medical science. This project's goals are to lower the number of tumours that are incorrectly identified as being cancerous and to demonstrate how machine learning algorithms may provide significant new insights into the elements that contribute to the development of lung cancer. It has been demonstrated that machine learning algorithms may accurately forecast cases of lung cancer.

2. Literature Survey

A. Bankar et al. [1] illustrated it is possible that the use of techniques from machine learning to healthcare might be of enormous value, with the end goal of treating sickness in millions of people. Researchers have put in a significant amount of work to identify cancer at an early stage and bring new insights into the diagnosis process. In the field of study on machine learning, a variety of algorithms have been used to determine the existence or decisiveness of cancer in correlation with the symptoms displayed by patients. This study will examine the symptoms experienced by three distinct age groups: young adults, working class people, and elderly people. In order to determine the relative value of various features, tree-based algorithms were utilised so that the underlying data patterns could be discovered and analysed. It has been shown that the elements that are responsible for the majority of instances of lung cancer in all age groups, including snoring, coughing up blood, having fingernails that are curled under, having a genetic predisposition to the disease, and passive smoking, are the culprits.

S. Hussein et al. [2] explained using computer-aided diagnostic (CAD) technologies, the process of determining the risk of malignancies based on radiological pictures may be made both more accurate and quicker. As part of precision medicine, the characterization of tumors through the use of such techniques can also facilitate non-invasive cancer staging and prognosis, as well as the development of individualized treatment plans. The first method that authors used is based on supervised learning, and it is for this type of learning that we exhibit considerable increases using deep learning algorithms. In specifically, work demonstrated advantages by employing a three-dimensional convolutional neural network and transfer learning. Following this, authors demonstrated how to incorporate task-dependent feature representations into a CAD system using a graph-regularized sparse multi-task learning framework. This was motivated by the interpretations of the scans provided by the radiologists. The second strategy involves examining an unsupervised learning algorithm as a potential solution to the widespread issue that medical imaging applications frequently face, which is the scarcity of labelled training data. We suggest using proportion-support vector machine as a means of defining tumours, having been motivated to do so by the lessons learned from label proportion techniques in computer vision. Authors are also interested in finding a solution to the fundamental question of whether or not "deep characteristics" are useful for the unsupervised categorization of tumours.

J. Nuhic et al. [3] says that due to the discovery of an excessive number of false negative patients, there are major problems with clinical maltreatment and mismanagement. As a result, the unreliability of lung cancer diagnostics and the strategies that may be utilised to circumvent it in a manner that is only minimally invasive are frequently the focus of research, as is the case with this particular study. This research focuses on the application of machine learning algorithms as a noninvasive method for distinguishing between benign and malignant pleural effusions. The algorithms that have been suggested were picked after careful consideration of the most recent advancements in the field of pulmonary diagnostics. The implementation of machine learning models for the categorization of lung cancer based on expression of tumour markers acquired from pleural fluids and serum is the innovative aspect of this body of work. On data samples from 168 patients, the performance of each model is evaluated and compared before being verified.

Puneet et al. [4] Cancer is caused by the uncontrolled proliferation and division of aberrant cells, which is caused by the creation of a significant number of abnormal cells. This leads to the development of cancer. The objective of this study is to improve the accuracy of previous attempts at predicting lung cancer by applying machine learning methods to analyse regular blood indicators as the primary data source. For the purpose of feature selection, a number of different scikit-learn algorithms have been utilised, and the algorithms have been used to choose just those characteristics that are relevant to our model. Authors concluded that the classification model XGBoost employing GridSearchCV is the one that works best for this assignment.

H. Azzawi et al. [5] illustrated that there are several kinds of lung cancer, each of which has a distinctively different cell size and growth behaviour. It is possible to boost patient survival rates by using more targeted therapies if different subtypes of lung cancer are correctly classified. Using microarray data, the authors of this study presented a novel method known as Structural Binary Classification (SBC), which may be used to categorise the different subtypes of lung cancer. The Gene Expression Programming (GEP) algorithm serves as the foundation for this method. Evaluations of the classification performance of our GEP-based model and comparisons with other prevalent binary decomposition algorithms In terms of accuracy, standard deviation, and area under the receiver operating characteristic curve, the experimental findings demonstrated that the GEP model using our technique performed better than the other models. This work contributes a valuable resource toward the categorization of lung cancer based on information gleaned from tumour structures.

3. Methodology

The architecture that is being considered is shown in Figure 1. The administrator uses a tool to obtain the image, and then dataset and machine learning algorithm approaches are used to the test data in order to extract useful data that is necessary for analysis. After that, categorization is accomplished with the use of machine learning techniques. Illness is discovered, and the outcome is presented to the administrator in terms of the presence or absence of the disease in the population.

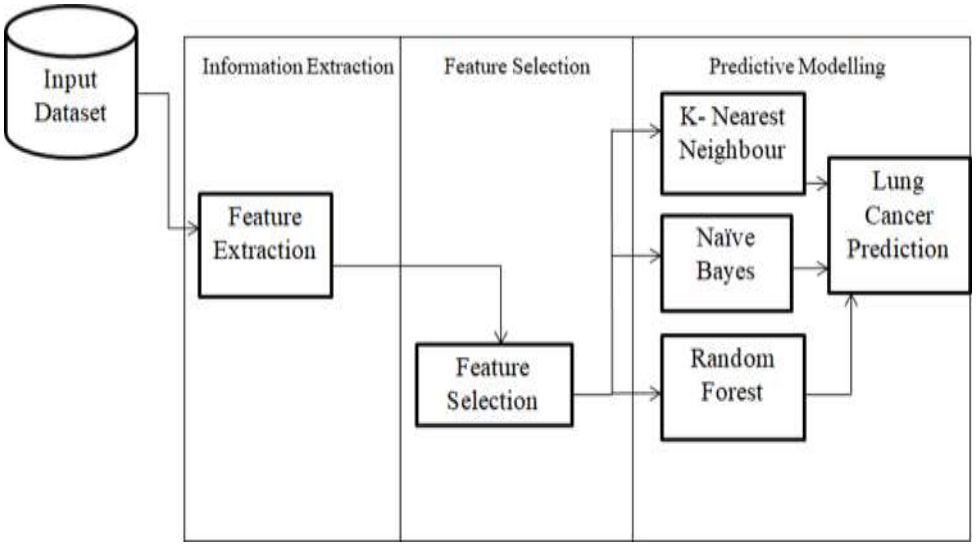

Figure 1. Proposed Architecture

The workflow of the proposed model are as follows:

3.1 Information Extraction

The information analyzed in this report originated in a database. This paper proposes to use the Python programming language to evaluate the efficacy of three algorithms at making prognostications about the likelihood that a patient would survive lung cancer. The RESPIR incidence data files included in three different SEER datasets serve as the basis for the investigations. The process of mechanically gleaning structured data from machine-readable texts that are not themselves structured is known as information extraction (IE). This often involves some form of natural language processing applied to materials written in the English language. Automatic annotation and content extraction from photos are two examples of recent work in multimedia document processing. Beyond its transmission, storage, and presentation, text management entails a variety of other issues, one of which is the development of automatic systems for information extraction. Automatic methods for indexing and categorising texts have been developed in the field of information retrieval, and these methods tend to have a statistical flavour. Natural language processing (NLP) is an additional complementary technique that has successfully modelled human language processing despite the enormous difficulty of the endeavour. IE tackles jobs that fall between IR and NLP in terms of both complexity and attention. IE takes as input a collection of documents where each document follows a template, describing one or more entities or events in a way similar to but different in specifics from documents in the collection.

3.2 Pre-Processing

Due to the non-specific nature of the dataset's entries, several cancers and cancer profiles are represented. Various preprocessing steps were taken to get the data ready for mining. The process is broken down into four phases that all work together to guarantee correct, accurate, and properly structured information. In a preprocessing procedure, the original, unprocessed, "raw," data files serve as input. There are two steps in the first phase (Phase I). The first thing to do is run a structural consistency check to make sure that all the files you're working with have uniform field names and record structures. Due to variations in the data gathering process throughout the years, it is not unusual for data sets from various years to include varying fields. Finding out where there are gaps in the data or mistakes has been identified as the second step. For instance, "null," "N/A," or "0" may be used to indicate empty fields. Missing value labels can be anything, but the convention used in the data must be established. In Phase II, we check if all of the data is relevant and comprehensive. By double-checking for relevance, you can be confident that all the data you need to complete your study is included. One such example would be medical records that did not specify the sort of cancer the patient was diagnosed with. It is necessary to perform additional analysis on the data to see if a synthesis of various fields may be used to infer and replace the missing data. Failure to do so may result in the dataset being dismissed as unsuitable for further analysis. The data granularity needs to be sufficient for the research purpose, and this is another outcome of this stage. Data that lacks the granularity required to establish individual patient-based associations since it was only collected at the group level should be removed. The third step involves checking the data for internal consistency in terms of quality. The goal is to guarantee that the entire dataset serves a purpose. Example: if you're mining for the age at which patients were diagnosed with prostate cancer, a disease that typically only affects men, you could find evidence of female patients if the data is consistent. The dataset would be contaminated and the findings would be skewed if this were to happen. Patients older than 120 years old, for example, or those in situations where the collection only offers information at the group level would result in data that lacks the essential granularity and hence would be better off being eliminated.

3.3 Feature Extraction

Data mining techniques normally operate under the assumption that the input is completely clean and devoid of noise. Several intriguing features of the data are revealed in the output, especially when the third pre-processing phase has been performed. Two fields, for instance, will be removed because they aren't needed in the final analysis. Whether or not a tumour is cancerous can be determined by its Behavior Code. The vast majority of our records have a malignant diagnosis (99.9%), whereas the remaining 0.1% have a null value. Unlike with other cancer kinds, lung malignancies are automatically considered malignant, hence this topic is irrelevant to our research. The Grade field's significance is also debatable. Between checkups, it stores a standardised, qualitative description of the tumor's state. While grades can be ranked from worst to best, the range of differences between them is yet unclear. Preliminary data analysis places nearly half of the records in the Grade IV category. About half of the items in this group are marked as "not available" (N/A). Because it contains evaluative judgments, the Grade field is ignored. Furthermore, the lack of information or clearly defined categories would add noise to the data rather than serve as a useful discriminating characteristic.

3.4 Predictive Modelling

Classifiers are the optimal method to use when building a model for making predictions. Treating a collection of instances as input, where each example belongs to a limited number of classes and is characterised by a predetermined set of characteristics, is the fundamental premise upon which this whole thing is founded. The estimate of the category that a new investigation belongs is the output of the classifier. The accuracy of the prediction shifts dependent on the classifier used as well as the different kinds of features and classes that are contained within the dataset. The process of classifying anything may be thought of as mapping a collection of properties onto a certain class. The data analysis made use of three different classification strategies: the Naive Bayes, the Random Forest, and the KNN. classifier Classifiers are the data mining algorithm of choice when trying to construct a model that can predict anything. Treating a collection of instances as input, where each example belongs to a limited number of classes and is characterized by a predetermined set of characteristics, is the fundamental premise upon which this whole thing is founded. The accuracy of the prediction shifts dependent on the classifier used as well as the different kinds of features and classes that are contained within the dataset. One way to think about classification is as a mapping from a certain set of qualities to a particular class. In the event that it is not feasible, the dataset can be declared improper and thrown away as a result. In addition, as a consequence of this step, the granularity of the data will be ensured to be adequate for the purpose of the study. Verifying that material is relevant to the purpose of the study helps to guarantee that it is included in the investigation. For the purpose of indexing massive document collections and categorizing documents, the field of information retrieval has developed automated approaches, which often take on a statistical flavor. The use of natural language processing is another strategy that complements the others.

4. Results and Discussion

Figure 2 Loading the Training dataset

The training dataset is imported into the programme in the manner depicted in Figure 2, which involves browsing for the source file, which results in the display of the directory path.

Figure 3 Data Preprocessing

After the training dataset has been put through this technique, the notification indicating the preprocessing has been completed effectively is displayed in Figure 3.

Figure 4 Browsing for Test Data

Figure 4 illustrates the process that is used to choose an individual patient for the purpose of conducting a test to determine whether or not they have lung cancer.

Figure 5 Loading the Test Dataset

The process of loading test data into the programme is illustrated in Figure 5, which shows the user looking for the source file, which then displays the directory path.

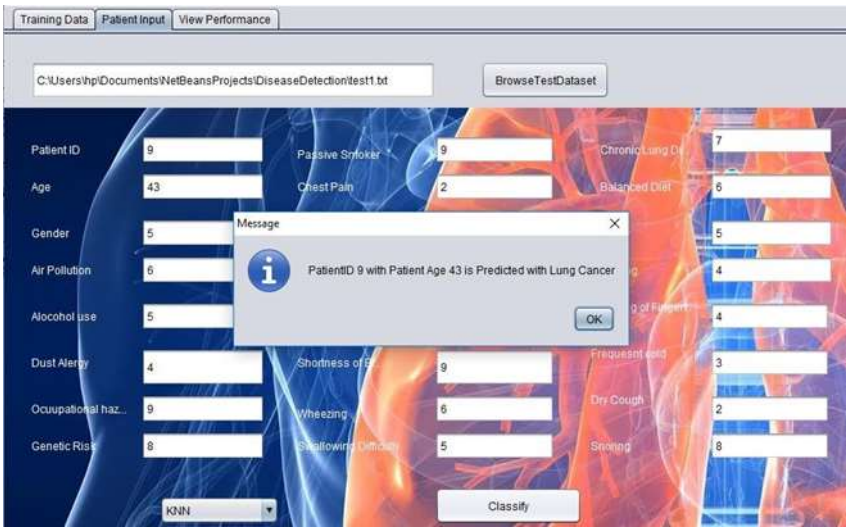

Figure 6 Prediction of Lung Cancer using Machine Leaning Algorithms

The patient depicted in Figure 6 has a patient ID of 9, is 43 years old, and has the above-mentioned attribute values. Using a machine learning classifier, lung cancer has been diagnosed as the patient's primary diagnosis.

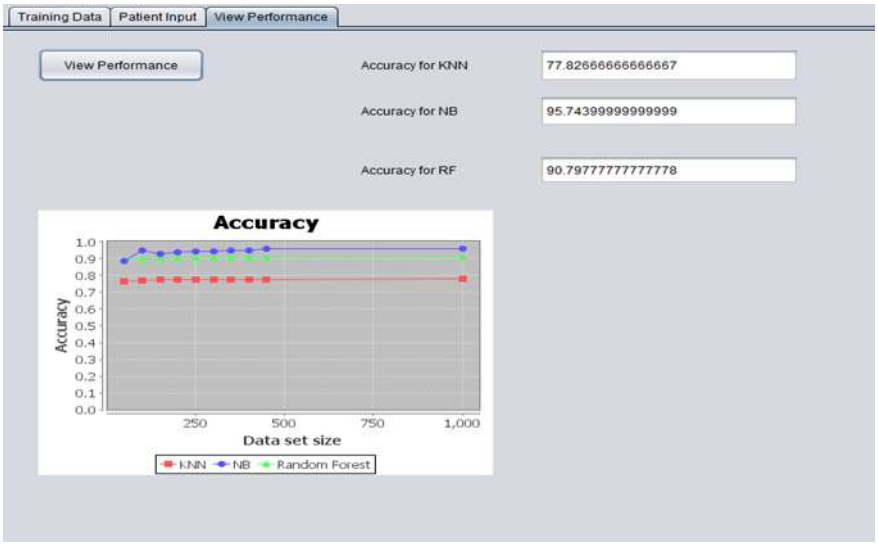

Figure 7 Performance analysis of the three classifiers

The percentages of correct classifications that each of the three classifiers achieved are shown in Figure 7. According to the findings, the Naive Bayes model demonstrates the best accuracy when it comes to forecasting the condition. In addition to this, a comparison of the three classifiers is carried out in the form of a graph according to the size of the dataset.

Conclusion

Because of this research, it is now simpler for physicians and other medical professionals to deal with people who have lung cancer and to diagnose them. This analytical study's primary objective is to identify the important variables that might cause lung cancer in various age groups as well as the key symptoms associated with the disease. When there is a need to find and comprehend the many rules and patterns concealed in complicated data, classifier models can be of tremendous assistance in accomplishing this task. The purpose of employing a variety of classification strategies on the dataset was to evaluate the performance of these three strategies in terms of accurately predicting the proportion of patients who will be diagnosed with lung cancer. A comparison is carried out, and the option with the best score is chosen. It is anticipated that the Naive Bayes model, which has a prediction accuracy of 90%, would be the most successful model for predicting patients who have lung cancer illness, followed by the Random Forests model and the K-NN model. In comparison to Decision Trees, Naive Bayes performed significantly better since it was able to recognize all of the important medical factors. In spite of the high prevalence of lung cancer in India, the illness is often misdiagnosed or left untreated in its early stages due to a lack of public awareness. The primary objective of this research is to develop methods that can detect lung cancer at its early stages, with the goal of reducing both the incidence of the disease and its fatality rate.

References

1. A Bankar, K. Padamwar and A. Jahagirdar, "Symptom Analysis using a Machine Learning approach for Early Stage Lung Cancer," 2020 3rd International Conference on Intelligent Sustainable Systems (ICISS), 2020, pp. 246-250, doi: 10.1109/ICISS49785.2020.9315904.
2. S. Hussein, P. Kandel, C. W. Bolan, M. B. Wallace and U. Bagci, "Lung and Pancreatic Tumor Characterization in the Deep Learning Era: Novel Supervised and Unsupervised Learning Approaches," in IEEE Transactions on Medical Imaging, vol. 38, no. 8, pp. 1777-1787, Aug. 2019, doi: 10.1109/TMI.2019.2894349.
3. J. Nuhic and J. Kevric, "Lung cancer typology classification based on biochemical markers using machine learning techniques," 2020 43rd International Convention on Information, Communication and Electronic Technology (MIPRO), 2020, pp. 292-297, doi: 10.23919/MIPRO48935.2020.9245114.
4. Puneet and A. Chauhan, "Detection of Lung Cancer using Machine Learning Techniques Based on Routine Blood Indices," 2020 IEEE International Conference for Innovation in Technology (INOCON), 2020, pp. 1-6, doi: 10.1109/INOCON50539.2020.9298407.
5. H. Azzawi, J. Hou, R. Alnnni and Y. Xiang, "SBC: A New Strategy for Multiclass Lung Cancer Classification Based on Tumour Structural Information and Microarray Data," 2018 IEEE/ACIS 17th International Conference on Computer and Information Science (ICIS), 2018, pp. 68-73, doi: 10.1109/ICIS.2018.8466448.
6. Chandrappa, S., Guruprasad, M. S., Kumar, H. N., Raju, K., & Kumar, D. S. (2023). An IOT-Based Automotive and Intelligent Toll Gate Using RFID. SN Computer Science, 4(2), 154.

7. Kumar HN, N., Kumar, A. S., Prasad MS, G., & Shah, M. A. (2022). Automatic facial expression recognition combining texture and shape features from prominent facial regions. IET Image Processing.

8. Jagannath, P. G., Marigundanahalli Siddabasappa, G. P., Huthinagadde RamakrishnaBhat, P. K., Jain, A. K., & Singh, P. (2022). FQRS: Farmer Query Redressal System Using Open-Source Framework. Materials Proceedings, 10(1), 9.

9. Kumar, M. A., Abirami, N., Prasad, M. G., & Mohankumar, M. (2022, May). Stroke Disease Prediction based on ECG Signals using Deep Learning Techniques. In 2022 International Conference on Computational Intelligence and Sustainable Engineering Solutions (CISES) (pp. 453-458). IEEE.

10. Singh, P., Guru Prasad, M. S., Taneja, H., Aditya Pai, H., & Sharonchrista. (2022). Statistical analysis of an improved tuning method for optimizing performances of Hadoop applications. Journal of Information and Optimization Sciences, 43(3), 533-541.

11. Guru Prasad, M. S., Singh, P., Taneja, H., Jain, A. K., & Chandrappa, S. (2022). Statistical analysis of multi job processing in Hadoop environment using schedulers. Journal of Information and Optimization Sciences, 43(3), 497-504.

12. Pai, A., Pareek, P. K., Guru Prasad, M. S., Singh, P., & Deshpande, B. K. (2021). Image Encryption Method by Using Chaotic Map and DNA Encoding. NVEO-NATURAL VOLATILES & ESSENTIAL OILS Journal| NVEO, 10391-10400.

13. Prasad, G., Jain, A. K., Jain, P., & Nagesh, H. R. (2019). A Novel Approach to Optimize the Performance of Hadoop Frameworks for Sentiment Analysis. International Journal of Open Source Software and Processes (IJOSSP), 10(4), 44-59.

14. Prasad, G., Nagesh, M. S., & Swathi Prabhu, H. R. (2017). An efficient approach to optimize the performance of massive small files in hadoop MapReduce framework. Int. J. Comput. Sci. Eng. IJCSE, 5(6), 112-120.

15. Nagesh, H. R., & Prabhu, S. (2017). High performance computation of big data: performance optimization approach towards a parallel frequent item set mining algorithm for transaction data based on hadoop MapReduce framework. International Journal of Intelligent Systems and Applications, 9(1), 75.

16. Prasad, G., Nagesh, H. R., & Prabhu, S. (2015). Performance analysis of schedulers to handle multi jobs in Hadoop cluster. Int. J. Mod. Educ. Comput. Sci, 7, 51-56.

17. Prabhu, S., Rodrigues, A. P., Prasad, G., & Nagesh, H. R. (2015, March). Performance enhancement of Hadoop MapReduce framework for analyzing BigData. In 2015 IEEE International Conference on Electrical, Computer and Communication Technologies (ICECCT) (pp. 1-8). IEEE.

18. Prasad, G., Nagesh, H. R., & Deepthi, M. (2014). Improving the performance of processing for small files in Hadoop: A case study of weather data analytics. International Journal of Computer Science and Information Technologies, 5(5), 6436-6439.

19. Prasad, M. G., Nagesh, H. R., & Dharmanna, L. (2013, September). Ensuring data storage in cloud computing for distributed using high security password. In National Conference on Challenges in Research & Technology in the Coming Decades (CRT 2013) (pp. 1-4). IET