Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X
IMPACT FACTOR: 7.056

IJCSMC, Vol. 9, Issue. 10, October 2020, pg.73 – 77

Lung Cancer Prediction Using Machine Learning Algorithms on Big Data: Survey

Dr. M. Kasthuri[1]; M. Riyana Jency[2]

[1]Assistant Professor, Department of Computer Applications, Bishop Heber College (Autonomous), Affiliated to Bharathidasan University, Tiruchirappalli, India

[2]Department of Computer Applications, Bishop Heber College (Autonomous), Affiliated to Bharathidasan University, Tiruchirappalli, India

[1] kasthuri.ca@bhc.edu.in; [2] riyanajency123@gmail.com

DOI: 10.47760/IJCSMC.2020.v09i10.009

Abstract- Lung cancer is a malignant lung tumour that is characterised by the regulated growth of cells in the lung tissue. The most common cancer diagnosed worldwide is lung cancer. More deaths than any other kind of cancer occur due to lung cancer. Early diagnosis and care are very useful and efficient for the survival of cancer patients. Different image processing and soft computing methods may be used for identifying cancer cells from medical images. Classification depends on features extracted from the images. In order to produce better classification results, the focus is on the feature extraction level. In order to distinguish a pattern that can provide some useful insights into what combination of features is most likely to result in an abnormality, this knowledge is then given to machine learning algorithms. The prediction of lung cancer is analysed using various machine learning classification algorithms such as Naive Bayes, Support Vector Machine, Artificial Neural Network and Logistic Regression. The key aim of this paper is to diagnose lung cancer early by examining the performance of exist classification algorithms.

Keywords: Lung cancer, Classification, Machine learning, Support Vector Machines Decision tree, Naïve Bayes: Decision Tree, Artificial Neural Network.

I. BIG DATA CHARACTERISTICS

Big data is a collection of large elements of data that grows exponentially day after day. Big Data is huge in size and complex, so it is not sufficient to store or evaluate conventional tools. Such big data is created from various sources, such as audio, video, photographs, social networking, cell phones, etc. In August 2012, a study was sent to the U.S. "The Congress notes big data as" huge quantities of high-speed, complicated and variable data requiring advanced techniques and technologies to collect, store, transmit, handle and analyse information"[1]. So, the emerging challenges that the different organizations face is the fast-growing data sets of large volume which needs modern technology to organize, store and analyze the proper useful data. Big data has been characterized by 5V's; Volume, Variety, Velocity, Veracity and Value.

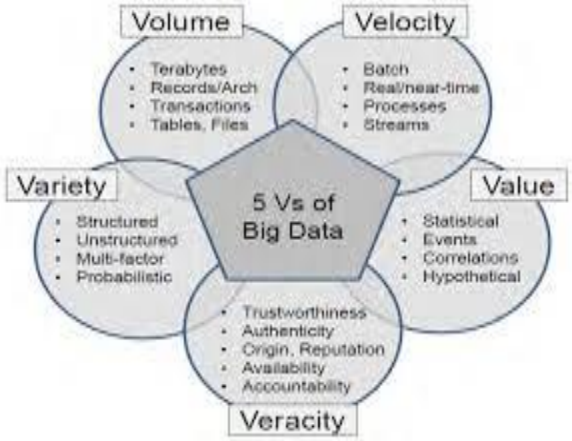

Fig 1: Five Vs of Big Data Analysis [1]

1) Volume: Volume is the size of the data. Today a large number of data has been generated from different fields like healthcare, public sector, retail etc. So, the size of data crosses petabyte and exabytes. Approx 8 zettabytes digital world of data have been reached in 2015 [2]. So, the accurate storage is the first requirement for true analysis.

2) Variety: The collection of data comes from various sources like posts, blogs, audios, images etc. So, the data are of different types such as structured, semi-structured and unstructured. The data that has a defined format like a database is a structured data. The images, videos, audios which have no structure or form are the unstructured data.

3) Velocity: It is the measurement of the speed of the data generation. The real time data flow is very fast and so it is difficult to retrieve the meaning of the data in today's speed. The security also has a vital role as it is very difficult to control the large amount of data. Nowadays two new characteristics have been added i.e. veracity and value. The quality of data is the veracity and the subject contents in the data are the value.

4) Veracity: When dealing with a high number, speed and variety of data, it is not likely that dirty data would be 100 % accurate with all of the data. The output of the data being collected will vary greatly. The accuracy of the analytical data relies on the veracity of the source data.

5) Value: Value is the most important aspect in the big data. The potential benefit of big data, however, is enormous. Access to big data is all well and well, but unless we can turn it into value, it is becoming very expensive to implement IT storage systems to store big data, and business would need a return on investment.

II. MEDICAL IMAGE IN BIG DATA

Medical images play an important role in the medical field. The main objective of medical image analytics is to obtain knowledge from the images derived from devices such as X-Ray, CT scan, MRI, PET-CT etc. The images help in early diagnosis of diseases. Nowadays a large scale of Lung cancer medical data has been generated from different sources. Lung cancer constitutes large proportion of death rates among cancer patients. Lung cancer may initiate in windpipe, main airway, or lungs. It is caused by unregulated growth and spread of some lung cells. People with pulmonary illness such as Emphysema and previous chest issues have a higher risk of being diagnosed with lung cancer. Overuse of tobacco, cigarettes and beedis is the main risk factor that leads to lung cancer in Indian men; however, smoking is not so common among Indian women, which suggests other factors leading to lung cancer. Other risk factors include exposure air pollution, radon gas and chemicals in the workstation. A cancer that begins in the lung is primary cancer of the lung whereas those that begin in the lung and spread to other parts of the body are secondary cancer of the lungs. The stage of cancer is measured by tumour size and how far it has spread. An early stage cancer is a small cancer diagnosed in the lung, and advanced cancer has spread into the surrounding tissue or other body parts. A better understanding of risk factors can help prevent pulmonary cancer. Early detection using machine learning techniques is the key to improving the rate of survival and if we can make the diagnostic process more efficient and effective for radiologists using this, it'll be a key step towards the goal of improved early detection.

III. MACHINE LEARNING ALGORITHMS

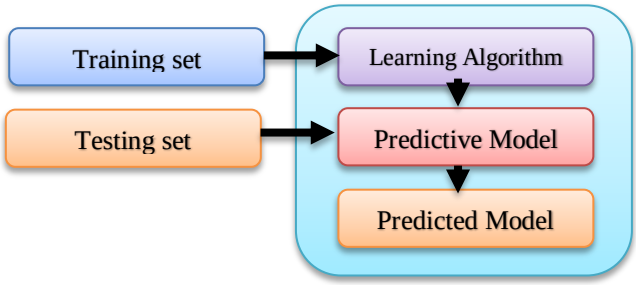

Fig 2 Over all Architecture of Classification.

Almost effective image-based ML systems are developed on SVM, linear regression, decision trees, KNN and so on. [3]

i. Support Vector Machine: For prediction, regression and classification the most prominent method employed is SVM. It classifies the input data set by introducing a boundary called a hyperplane that separates the dataset into two parts. The favourable asset of SVM is, as SVM is a data-driven approach and feasible without a hypothetical scheme that produces an accurate classification. Particularly when the size of the sample is small. SVMs are broadly used for classification when the datasets are biomarkers, to predict and diagnose cancer, neurological and cardiology diseases.

ii. Artificial Neural Network**:** ANN is a computational method that is compromised profoundly with interconnected processing components called neurons, which sort out information as feedback to external stimuli. It is accomplished in two modes, i.e. learning and testing. Learning is used to classify new input. Attesting phase, it receives an input signal from the network and computes it to generate an output. In diversified fields of healthcare, ANN approaches are useful for example in diagnosis and indicating breast cancer, lung cancer, and other ontology prediction, diagnostic systems, drug analysis, etc.

iii. Naive Bayes

A classifier Naive Bayes is a model of probabilistic machine learning which is used for classification tasks. This classifier is very quick and easy to implement but their biggest disadvantage is that they have to be independent of the predictors. In most cases of real life, the predictors are dependent, this impedes the classifier's efficiency. One of the easiest ways to select the most likely hypothesis, given the data we have that we can use as our prior knowledge of the issue. Given our prior knowledge, Bayes 'Theorem provides a way of estimating the likelihood of a hypothesis.

$$P(c|x) = \frac{P(x|c)P(c)}{P(x)}$$

Above,

P(c|x) is the posterior probability of class (c, target) given predictor (x, attributes).

P(c) is the prior probability of class.

P(x|c) is the likelihood which is the probability of predictor given class.

P(x) is the prior probability of predictor.

iv. Logistic Regression

By suitable a linear equation to experiential data, linear regression efforts to perfect the connection among two variables. One variable is measured an explanatory variable, and the other variable is measured as dependent. The linear regression has an equation in the form of Y = a + bX, where X is the independent feature variable and Y is the dependent variable which is also known as label. The angle of the line is b, and a is the intercept on y axis when x=0.

IV. RELATED WORK

This section provides a broad survey on different cancer prediction, diagnosis, and detection using deep learning approaches, techniques, algorithms and datasets used for experimental analysis. Summary of Lung cancer prediction using machine learning by various researchers is shown in Table 1.

TABLE I SUMMARY OF LUNG CANCER PREDICTION

Sno	Paper Title	Methodology	Dataset	Advantages	Results
1	Multi-stage lung cancer detection and prediction using multi-class SVM classifier [4].	SVM	UCI MLDb	Able to detect false positive nodules correctly. 3 Improved Detection of Lung Nodules on Chest Radiographs Using a Commercial Computer- Aided Diagnosis System ANN Large Database with 274 radiographs and 323 lung nodules	97%- identification 87%- prediction
2	A computer aided diagnosis system for detection of lung cancer nodules using extreme learning machine [5].	ELM	Data collected from a reputed hospital with 100 lung images	Compared with the conventional CAD system proposed method gives excelling detection	Able to detect false positive nodules correctly.
3	Improved Detection of Lung Nodules on Chest Radiographs Using a Commercial Computer-Aided Diagnosis System [6].	ANN	Large database with 274 radiographs and 323 lung nodules	The false-positive detections seem to be increased by these pre-existing diseases.	73% detection sensitivity with 4.0 false positive Detections per image.
4	Lung nodule detection on thoracic computed tomography images: Preliminary evaluation of a computer-aided diagnosis system [7].	Rule-based classification, LDA	Institutional Review Board with 1454CT images gathered from 34 diagnosed patients with 63 lung nodules.	Improved the sensitivity of cancer detection, reduce oversight Errors, and decrease inter- and intra-reader variations	84% sensitivity
5	Effect of Computer-Aided Diagnosis on Radiologists' Detection Performance of Subsolid Pulmonary Nodules on CT: Initial Results.[8].	Top-hat transformation method, sieve filter.	An institutional review board of Rinku General Medical Centre with 133 CT	Improve both the residents' and the board-certified radiologists' performance in detecting	80%- true positive rate

6	Computer-Aided Diagnosis for Lung CT Using Artificial Life Models[9].	Artificial life method	CT images composed of series 2D files in DICOM Format	Artificial life models can be more precise virtual ants. If biological aspects are considered.	Improved accuracy
7	Optical Deep Learning Model for Classification of Lung Cancer on CT images [10].	ODNN & LDA	Standard CT database with 50 CT images	Manual labelling time is minimized. Accuracy & precision are enhanced.	Accuracy-94.56%, sensitivity-96.2% & specifity-94.2%
8	Automatic Lung Cancer Prediction From Chest X-Ray Images using Deep Learning approach. [11].	DenseNet-121	Chest X-ray 14 & JSRT dataset	High accuracy is achieved and heatmap is implemented to identify the lung nodule region. Attention-Guided CNN (AG-CNN) can be used to identify the malignancy.	Accuracy-74.43±6.01, sensitivity-74.68±15.33% & specificity-74.96±9.85%.
9	Image-Based Survival Prediction for Lung Cancer Patients using CNNs [12].	ResNet18 & CNN	Lung1 dataset-TCIA	CNN is much harder to interpret when compared with the Cox model	CNN is much harder to interpret when compared with the Cox model. 0.623±0.039 Prediction
10	Lung Cancer Screening with Low-Dose CT Scans using a Deep Learning Approach [13].	Deep Screener algorithm	TCIA-1449 low dose CT images	A false-positive rate is efficiently is reduced	82% of Accuracy
11	Prediction of Lung Cancer Using Machine Learning Classifier [14].	Proposed RBF classifier	lung cancer data	classifier has resulted with a great accuracy and considered as the effective classifier technique for Lung cancer data prediction.	81.25% of Accuracy
12	Classification of lung cancer stages with machine learning over big data healthcare framework [15]	SVM-nonlinear SVM with Radial Basis Function RBF and Multiclass classification (WTA-SVM winner-takes-all with support vector machine) with threshold technique (T-BMSVM)	Sputum color images collected from microscope lab	The proposed method assists medical diagnosis such that early detection of lung cancer stages with accurate results.	Achieves 86% accuracy.
13	Mean Filtering to De-Noising Image Using Various Block Size [16]	Gaussian, Salt & Pepper, Local Var, Speckle, Poisson, Motion	Various type of images	improve the quality performance.	This filter operates by smoothing over a fixed window and it produces artifacts around the object sometimes causes over smoothing

		Blur, Erosion, Dilation			them causing blurring of image.
14	Image of De-Noising Using Linear Mean Filter For Various Block Size [17]	Gaussian, Salt & Pepper, Local Var, Speckle Or Multiplicative, Poisson, Motion Blur, Erosion, Dilation.	Various type of images	improve the quality performance	The proposed work performs well for cameraman image. The Performance measure, Pearson correlation co-efficient is invariant for different block sizes.

One of the major and frequent bases of cancer deaths globally in terms of both instance and transience is lung cancer. The main reason behind the increasing of deaths from it is detecting the disease lately and faults in effective treatment. So, the early detection is needed to save lives from this disease.

V. CONCLUSION

Big data offers an incentive for "big analysis" to advance the quality of life or solve the mysteries of the universe, leading to "big opportunities". The biomedical image of big data, including methods for generating, handling, representing and analysing imaging information for biomedical use, was considered herein. In this study, the principles of machine learning were explored while we outlined their submission in the prediction / prognosis of lung cancer. Most of the studies proposed in recent years concentrate on the development of predictive models using supervised machine learning techniques and classification algorithms to predict valid outcomes of illnesses. It is evident from the review of their findings that the incorporation of multidimensional heterogeneous data, combined with the application of various feature selection and classification techniques, may provide promising tools for cancer domain inference.

REFERENCES

[1] Wu, Dapeng, Hang Shi, Honggang Wang, Ruyan Wang, and Hua Fang. "A Feature-based Learning System for Internet of Things Applications." IEEE Internet of Things Journal (2018).

[2] Elijah, Olakunle, Tharek Abdul Rahman, Igbafe Orikumhi, Chee Yen Leow, and MHD Nour Hindia. "An Overview of Internet of Things (IoT) and Data Analytics in Agriculture: Benefits and Challenges." IEEE Internet of Things Journal (2018).

[3] Chhabra, Anshuman, Vidushi Vashishth, Anirudh Khanna, Deepak Kumar Sharma, and Jyotsna Singh. "An Energy Efficient Routing Protocol for Wireless Internet-of-Things Sensor Networks." arXiv preprint arXiv:1808.01039 (2018).

[4] Heinzelman, Wendi Rabiner, Anantha Chandrakasan, and Hari Balakrishnan. "Energy-efficient communication protocol for wireless microsensor networks." In System sciences, 2000. Proceedings of the 33rd annual Hawaii international conference on, pp. 10-pp. IEEE, 2000.

[5] Tawalbeh, Lo'ai, Fadi Muheidat, Mais Tawalbeh, and Muhannad Quwaider. "IoT Privacy and security: Challenges and solutions." Applied Sciences 10, no. 12, 2020.

[6] Asthana, Shubhi, Aly Megahed, and Ray Strong. "A recommendation system for proactive health monitoring using IoT and wearable technologies." In 2017 IEEE International Conference on AI & Mobile Services (AIMS), pp. 14-21. IEEE, 2017.

[7] Walinjkar, Amit, and John Woods. "ECG classification and prognostic approach towards personalized healthcare." In 2017 International Conference On Social Media, Wearable And Web Analytics (Social Media), pp. 1-8. IEEE, 2017.

[8] Nguyen, Hoa Hong, Farhaan Mirza, M. Asif Naeem, and Minh Nguyen. "A review on IoT healthcare monitoring applications and a vision for transforming sensor data into real-time clinical feedback." In 2017 IEEE 21st International Conference on Computer Supported Cooperative Work in Design (CSCWD), pp. 257-262. IEEE, 2017.

[9] Pandey, Purnendu Shekhar. "Machine learning and IoT for prediction and detection of stress." In 2017 17th International Conference on Computational Science and Its Applications (ICCSA), pp. 1-5. IEEE, 2017.

[10] Siryani, Joseph, Bereket Tanju, and Timothy J. Eveleigh. "A machine learning decision-support system improves the internet of things' smart meter operations." IEEE Internet of Things Journal 4, no. 4, 2017.

[11] Ling, Xiao, Jie Sheng, Orlando Baiocchi, Xing Liu, and Matthew E. Tolentino. "Identifying parking spaces & detecting occupancy using vision-based IoT devices." In 2017 Global Internet of Things Summit (GIoTS), pp. 1-6. IEEE, 2017.

[12] Amit Sagu, Nasib Singh Gill, "Machine Learning Techniques for Securing IoT Environment", International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-9 Issue-4, February 2020.

[13] Uwagbole, Solomon Ogbomon, William J. Buchanan, and Lu Fan. "An applied pattern-driven corpus to predictive analytics in mitigating SQL injection attack." In 2017 Seventh International Conference on Emerging Security Technologies (EST), pp. 12-17. IEEE, 2017.

[14] Ahmed, Muhammad Ejaz, Hyoungshick Kim, and Moosung Park. "Mitigating DNS query-based DDoS attacks with machine learning on software-defined networking." In MILCOM 2017-2017 IEEE Military Communications Conference (MILCOM), pp. 11-16. IEEE, 2017.

[15] Sujitha, R., and V. Seenivasagam. "Classification of lung cancer stages with machine learning over big data healthcare framework." Journal of Ambient Intelligence and Humanized Computing, Volume 2, Issue 10, pp 2652-2063, April 2020.

[16] M.Kasthuri, "Mean Filtering to De-Noising Image Using Various Block Size", International Journal of Advanced Scientific Research and Management, Volume 4 Issue 6, pp. 43 – 47, ISSN 2455-6378, June 2019.

[17] M.Kasthuri, "Image of De-Noising Using Linear Mean Filter For Various Block Size ", International Journal of Scientific & Technology Research, Volume 8, Issue 10, pp. 2606 – 2608, ISSN 2277-8616, October 2019.

