Webology (ISSN: 1735-188X)
Volume 18, Number 6, 2021

Enhancing Lung Cancer Detection: A Comparative Efficiency Study Of Machine Learning Supervised And Unsupervised Models

M. Sheik Mansoor[1] and M. Mohamed Sathik[2]

[1]Research Scholar (Reg. No. 17221192161007), Sadakathullah Appa College, Affiliated to Manonmaniam Sundaranar University, Tirunelveli, Tamilnadu, India.
[2]Principal and Research Supervisor, Sadakathullah Appa College, Tirunelveli, Affiliated to Manonmaniam Sundaranar University, Tirunelveli, Tamilnadu, India.

Abstract— Lung cancer is one of the most aggressive types of cancer, known for its rapid spread. It metastasizes through lymphatic fluid and the bloodstream, reaching organs like the bones, glands, and brain. The incidence of lung cancer is rising significantly due to air pollution and industrial contaminants. The World Health Organization (WHO) predicts that lung cancer-related deaths could reach 9.6 million by 2020, highlighting the severity of the issue. Early detection of lung cancer is crucial for effective treatment, but despite the availability of manual CT scan analysis in the medical field, it remains challenging for doctors to accurately determine the stage of the disease from these images. Consequently, to predict lung cancer and its varieties early on, the medical informatics research community has developed a number of machine learning models. This research compares two well-known supervised learning models, Artificial Neural Networks (ANN) and Support Vector Machines (SVM), using data on lung cancer that was taken from the Cancer Image Archive. We also applied the dataset to two unsupervised learning models, Apriori and K-means, to examine the differences in performance. The final results, including performance metrics such as accuracy, precision, and recall, were compared and displayed in a table.

Keywords— Machine Learning; Lung Cancer Prediction; Supervised Learning; Cancer Diagnosis.

1. INTRODUCTION

Cancer that starts in the lung cells and spreads to other areas of the body is known as lung cancer [3]. Similarly, the lymphatic system or bloodstream may allow cancer cells from the mouth, kidneys, and breast to go to the lungs [2,3]. The lungs, located in the chest, are spongy organs responsible for taking in oxygen and expelling carbon dioxide [1]. Breathing involves the passage of air through the trachea, a structure resembling a tube, and the bronchi before reaching the lungs, following the same path in reverse to exit. Tiny sacs in the bronchi, known as alveoli,

transfer oxygen into the bloodstream and remove carbon dioxide from it [4,5].

In the early stages of lung cancer, the patient's DNA undergoes changes or damage, leading to gene mutations. These mutated genes fail to function correctly because they do not receive proper instructions from the DNA. As a result, Lung cancer develops when lung cells start to expand and divide uncontrolled within and around the lungs [6].

According to the Global Cancer Observatory (GCO), 5.4 out of every one million people in India are affected by lung cancer. The alarming concern is that lung cancer has a much lower survival rate compared to other forms of cancer. Each year in India, 25% of cancer patients pass away. Prostate, colorectal, skin, kidney, and breast cancers have much lower fatality rates than lung cancer due to late-stage detection and rapid progression [7]. It is challenging to manually analyze CT images to correctly diagnose lung cancer in its early stages, complicating the ability of medical professionals to accurately assess the disease's precise stage based on these images.

To address these challenges and detect cancer at an early stage, machine learning techniques are applied to patient data. This helps doctors determine the kind and degree of the cancer cells and provides them get a detailed image of the patient's situation [8, 9].

Machine learning techniques can generally be divided into two primary categories based on their applications and functions. In the realm of lung cancer prediction, a variety of research contributions and methodologies have been introduced. In this study, we compared ANN, SVM, Apriori, and K-means. They were trained using open-access Cancer Imaging Archives datasets. The preprocessing, feature extraction, and selection steps were standardized across all four methods.

This comparative study is structured as follows: The differences between supervised and unsupervised learning are discussed in Section 2. The methods for feature extraction, feature selection, and preprocessing are described in Section 3. The performance of the ANN, SVM, Apriori, and K-means models is assessed in Section 4. In conclusion, Section 5 looks at potential future avenues for more comparisons and wraps up the research.

2. SUPERVISED LEARNING AND UNSUPERVISED LEARNING

In supervised learning, certain data points are pre-identified as the right answers, and the machine learning model is trained on labeled data. This allows the algorithm to learn by comparison and predict outcomes for unseen data. In contrast, unsupervised learning does not depend on labeled data. Instead, it is designed to identify patterns and extract information independently. Although unsupervised learning can manage more complex processing tasks than supervised algorithms, its outcomes are often less predictable compared to those of deep learning and natural learning methods.

A. Supervised learning algorithm: Artificial Neural Networks (ANN)

The neurons that make up an Artificial Neural Network (ANN) are networked nodes that find patterns and correlations in data to gather information. It is organized into three layers: the input layer, hidden layer, and output layer. Neurons in each layer receive inputs, process them, and transmit the information to neighboring neurons. Each neuron and the connections between them

have assigned weights that are modified during the learning process. A neural network's learning algorithms use both forward and backward propagation.

Based on the neurons' greatest probability in the output layer, the ANN produces its final result. ANN seems to provide very accurate findings, even if there are other algorithms available for predicting lung cancer in its early stages.

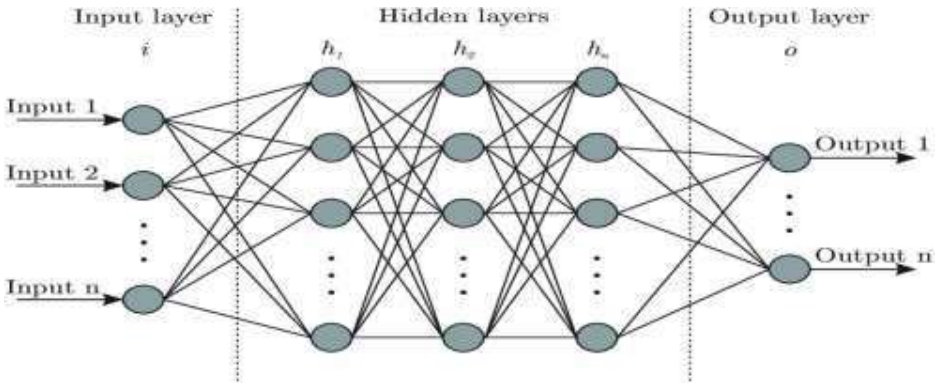

Fig. 1. Structure of Neural Networks

B. Supervised learning algorithm: Support Vector Machine (SVM)

To find correlations within a dataset, supervised learning techniques like SVM are used to regression and classification problems. SVM operates as a discriminative classifier by creating a hyperplane that distinguishes between the output classes. These hyperplanes serve as decision boundaries, and the highest possible accuracy is achieved when the SVM successfully separates all data points into their respective classes using this hyperplane.

In this context, support vectors are the data points that lie nearest to the hyperplane and influence its position and orientation. By leveraging these support vectors, the margin of the classifier can be enhanced for improved distinction. Removing any of these support vectors would alter the position of the hyperplane. These specific points are crucial for constructing an accurate SVM model.

Position and direction of the hyperplane are greatly influenced by the data points that are nearest to it, known as support vectors. Increased margin in the classifier may result in a more distinct division of classes with the use of these support vectors. The hyperplane's location will change if the support vectors are removed. For an appropriate SVM model to be created, these criteria are essential.

C. Unsupervised learning algorithm – Apriori Algorithm

The Apriori algorithm is a foundational method for generating frequent item sets and represents a significant advancement in the field of data mining. For Boolean association rules, it is used to find frequently occurring objects in a collection. The current information about the properties of often occurring objects is used by the algorithm. To locate k+1 item sets, it uses an iterative or level-wise search strategy that uses k-frequent items.

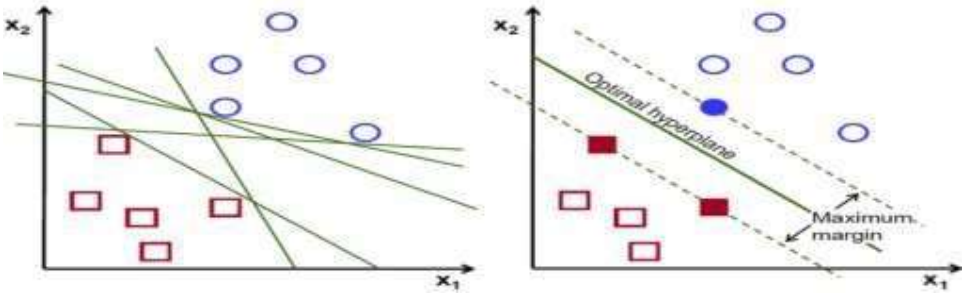

Fig. 2. Hyperplane of SVM

The Apriori attribute is used to narrow the search area to increase the level-wise production of frequent item sets' efficiency. According to this feature, a frequent item set's non-empty subsets must also be regarded as frequent.

D. Unsupervised learning algorithm – K Means

The data points are allocated to individual clusters via the iterative K-means method, which divides a dataset into K separate, non-overlapping clusters. The goal is to maximize similarity within clusters while keeping the clusters as distinct as possible. The sum of the squared distances between each cluster centroids and each data point, which is the average of all data points in that cluster, is reduced while allocating data points to clusters. Lower cluster variation means more related data points.

3. HANDLING DATA

A. Pre-processing

Pre-processing is a crucial step used to convert raw data into a more usable and efficient form. It involves multiple stages, including data cleaning, transformation, and reduction. Data cleaning addresses missing values by either removing or replacing them and eliminates noise using techniques like regression, clustering, or binning. Data transformation involves normalizing the data, selecting relevant attributes, converting continuous datasets into discrete ones through discretization, and generating hierarchies. Data reduction involves various actions, such as aggregating data as necessary, selecting subsets of attributes, applying parametric or non-parametric techniques for reducing numerosity, and decreasing dimensionality.

B. Feature Selection

A dimensionality reduction is feature selection technique aimed at identifying the most significant features for model development. It encompasses four main strategies: wrapper, filter, embedded, and hybrid approaches. The wrapper method is computationally intensive, selecting features by employing classification and utilizing a learning algorithm to assess classification accuracy. On the other hand, the filter approach selects a subset of features without involving any learning algorithms. This type of feature selection approach is suitable for databases with high dimensionality. During the data training phase, the embedded technique chooses features using applied learning algorithms to determine feature relevance. The hybrid approach combines the filter and wrapper methods, first selecting features through the filter method and then testing them using the wrapper approach. This allows it to leverage the strengths of both approaches for more effective feature selection.

C. Feature Extraction

Additional dimensionality reduction technique is feature extraction that transforms raw data into a more manageable form for further analysis. This technique is particularly important in image processing, when image analysis requires the use of numerous parameters. It involves tasks such as low-level extraction, edge detection, curvature extraction, shape recognition, and motion detection. Low-level image processing includes various detection methods, such as edge detection, corner detection, blob detection for identifying regions in images, ridge detection for extracting bright thin lines compared to surrounding areas, and feature transformation using differences in image scales. Curvature extraction aims to identify the direction of edges and detects changes in image intensity as well as autocorrelation. Shape detection focuses on determining image thresholds, extracting regions, and performing template matching. This process also incorporates Hough transformation, it uses the voting system to compare objects inside their class and extract imperfect characteristics. Motion detection involves analyzing the motion in images and assessing optical flow by observing specific areas within the images. Additionally, performance evaluation is conducted for techniques such as ANN, SVM, Apriori, and K-means.

A. Performance comparison of ANN and SVM

Tensor Flow, an open-source program developed by Google Inc., is used in the experiments using ANN and SVM machine learning models. The CT scan data from 1,019 patients with various cancer diagnoses is included in the dataset utilized for this application, which was acquired from the Cancer Imaging Archives. Initially, data pertaining to patients diagnosed with NSCLC cancer is extracted, resulting in approximately 419 records. Training and test datasets are divided 70:30 from NSCLC cancer data. Both datasets are used simultaneously as input for training the ANN and SVM, allowing for concurrent training and computation to achieve optimal prediction results.

Tables 1 and 2 display the predictions generated by the ANN and SVM models, respectively. The ANN model predicted 290 of 320 values with 90.2% accuracy. In comparison, the SVM system made 284 accurate predictions with 88% accuracy.

Predicted	True/Actual			
		Type 'T'	Type 'M'	Type 'N'
	Cancer Type 'T'	96	8	4
	Cancer Type 'M'	5	89	5
	Cancer Type 'N'	4	5	104

Table. 1. Type of cancer prediction using the ANN approach

Predicted	True/Actual			
		Type 'T'	Type 'M'	Type 'N'
	Cancer Type 'T'	101	6	2
	Cancer Type 'M'	7	95	8
	Cancer Type 'N'	8	5	88

Table. 2. Type of cancer prediction using linear SVM

The second key performance metric for machine learning algorithms is precision. This measures the proportion of relevant information retrieved; for instance, in predicting lung cancer types, it indicates the percentage of patients accurately classified as belonging to a specific cancer type.

In the context of lung cancer type prediction, the precision for type 'x' is determined by dividing the number of accurately recognized instances of type 'x' by the total number of predictions made for that type. The precision is determined using this formula:

Precision (Type 'x') = (Correctly identified Type 'x') / (Total predictions for Type 'x' cancer)

Precision values (in percentage)		
	ANN	**SVM**
Cancer Type 'T'	88.8%	92.6%
Cancer Type 'M'	90.8%	86.3%
Cancer Type	92.6%	87.1%

'N'		

Table. 3. Accuracy values of the SVM and ANN algorithms for the specified dataset.

Recall can be calculated using the following formula:
Recall = (Number of correctly predicted type 'x' cancer) / (Number of actual type 'x' cancer patients)
Table 4 presents the precision values for the ANN and SVM algorithms based on the provided dataset.

When comparing the two algorithms, it is evident that ANN generally outperforms SVM in many instances. The precision values indicate that ANN is more effective overall, although SVM shows better precision in certain cases, such as when detecting cancer type T. In terms of recall, ANN tends to yield better results than SVM; however, for predicting cancer type M, SVM demonstrates superior performance with more accurate results.

Comparison of Apriori and K-Means

This research utilized a dataset necessary for extracting valuable insights regarding the impact of the k-means algorithm on the Apriori algorithm, particularly in terms of computation time and the rules generated. The dataset comprises 8,243 records related to disease diagnoses, including variables such as disease diagnosis, age group, gender, and care status. A portion of the data is presented in Table 5.

The first method uses the illness diagnosis, age group, gender, and care status as the four input variables and applies the Apriori algorithm directly to the dataset. This process aims to derive confidence values, rules, and computation times for the Apriori algorithm. The results from this testing can be found in Table 6.

The rule information derived from large item set 4 leads to two specific rules: one concerning the diagnosis of allergic rhinitis in female children of a certain age group with outpatient status, and the other related to the diagnosis of postoperative disease in adult males, also with outpatient status. Both rules exhibit a confidence value of 69%. These findings suggest that the information obtained from the Apriority algorithm is still insufficient.

Recall values (in percentage)		
	ANN	**SVM**
Cancer Type 'T'	91.4%	87%
Cancer Type 'M'	86.4%	89.2%
Cancer Type 'N'	91.2%	89.7%

Information that is more thorough and detailed than that obtained from the Apriori algorithm alone is obtained when the K-Means approach is used with it, as Table 7 shows.

S. No.	Disease Diagnosis	Age Cluster	Gender	Status of Care
1	Observation of Febris	Baby	Male	Outpatient
2	Observation of Febris	Baby	Female	Outpatient
3	Observation of Febris	Baby	Male	Outpatient
4	Observation of Febris	Baby	Female	Outpatient
5	Paronychia	Adult	Female	Outpatient
6	Hnp Lumbalis	Adult	Male	Outpatient
:	:	:	:	:
8243	Disputes with the counselor	Toddlers	Female	Outpatient

Table 5. Sample patient diagnosis data in 2016

Using Apriori							
Full Data							
Disease Diagnose	Cataracts not Specified	Cataracts not Specified	Cataracts not Specified	Cataracts not Specified	Another Allergic Rhinitis	Another Allergic Rhinitis	Post Operation
Age Cluster	--	Elder	Elder	Elder	--	Child	Adult
Gender	Male	--	Female	--	Female	Female	Male
Status of Care	Out	Out	Out	Out	Out	Out	Out
Confidence (%)	69	76	60	66	69	69	69

Table 6. Data processing using Apriori algorithm

K-Means + Apriori				
	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Disease Diagnose	Cataracts not Specified	Cataracts not Specified	Another allergic rhinitis	Post Operation
Age Cluster	Elder	Elder	Child	Adult
Gender	Male	Female	Female	Male
Status of Care	Outpatient	Outpatient	Outpatient	Outpatient
Confidence (%)	66	66	92	93

Table 7. Data processing using K- Means and Aprior

Furthermore, the K-Means + Apriori calculation time is quicker than the Apriori computation time alone. The total time for the combination is 17.41 minutes, compared to 21.93 minutes for the Apriori algorithm by itself.

5. CONCLUSION

This study compared two known machine learning models, ANN and SVM, using data on lung cancer that was taken from the Cancer Imaging Archives. Additionally, another patient dataset was used for unsupervised learning methods, specifically the Apriori and K-means models, to assess performance differences. The final results and performance metrics, such as accuracy, precision, and recall, were compared and presented in a table, allowing for an evaluation of both unsupervised and supervised algorithms.

Reference

1. K. Kancherla and S. Mukkamala," Feature Selection for Lung Cancer Detection Using SVM Based Recursive Feature Elimination Method", Lecture Notes in Computer Science, Vol. 7256, pp. 168-176,2012.
2. S.K. Lakshmanaprabu, S.N. Mohanty,
   K. Shankar, N Arunkumar, and G. Ramirez,"Future Generation Computer System, Vol. 92, pp.374-382, 2018
3. A. Trivedi and P. Shukla, "Lung Cancer Diagnosis by Hybrid Support Vector Machine", Communications in Computer and Information Science, Vol. 628, pp. 177-187, 2016.
4. T. Nadira and Z. Rustama, "Classification of Cancer Data Using Support Vector Machines with Features Selection Method Based on Global Artificial Bee Colony", Vol. 2023(1), pp. 1-7,2018.
5. Arora, P., Boyne, D., Slater, J. J., Gupta, A., Brenner, D. R., and Druzdzel, "Bayesian Networks for Risk Prediction Using Real-World Data: A Tool for Precision Medicine, Value in Health",Vol. 22, pp.437-445, 2019
   M. B. Sesen, T. Kadir, R. B. Alcantara, J. Fox, and M. Brady," Survival Prediction and Treatment Recommendation with Bayesian Techniques in Lung Cancer", AMIA Annual Symposium, pp. 838-847.
6. K. Jayasurya, G. Fung, S. Yu, C. Dehing- Oberije, D. De Ruysscher, A. Hope, W. De Neve, Y. Lievens, P. Lambin , and A. L. A.
   J. Dekker, "Comparison of Bayesian network and support vector machine models for two-year survival prediction in lung cancer patients treated with radiotherapy" Medical Physics, Vol. 37, pp. 1401-1407, 2010.
7. Dharshinni N P, Mawengkang H and Nasution M K M 2018. Mapping of medicine data with k-means and apriori combinations based on patient diagnosis. International Conference on Computing and Applied Informatics.Vol 2 (978).

8. E. Adetiba and O. Olugbara, "Lung Cancer Prediction Using Neural Network Ensemble with Histogram of Oriented Gradient Genomic Features" The Scientific World Journal", Vol.2015, pp. 1-17, 2015.
9. Siegel, R.L., et al. (2021). Cancer statistics, 2021. CA: A Cancer Journal for Clinicians.
10. Schmidhuber, J. (2015). Deep learning in neural networks: An overview. Neural Networks.
11. LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. Nature.
12. Abdar, M., et al. (2020). Review of ANN and SVM techniques in medical image analysis. Artificial Intelligence in Medicine.
13. Zhou, X., et al. (2018). Deep learning in lung cancer detection: A review. IEEE Transactions on Medical Imaging.
14. Tang, C., et al. (2020). SVM for lung cancer diagnosis from medical images. Journal of Machine Learning in Healthcare.