JCRES
Journal of Current Research in Engineering and Science
Bi-Annual Online Journal (ISSN : 2581 - 611X)

Volume7-Issue1, January 2024
Paper: 9

Lung Nodule Detection Using Modified KNN

Premanand Ghadekar
Information Technology
Vishwakarma Institute of Technology
Pune, India
premanand.ghadekar@vit.edu

Harshada Giri
Information Technology
Vishwakarma Institute of Technology
Pune, India
harshada.giri21@vit.edu

Vaishnavi Chaudhari
Information Technology
Vishwakarma Institute of Technology
Pune, India
vaishnavi.chaudhari21@vit.edu

Shreyas Manwadkar
Information Technology
Vishwakarma Institute of Technology
Pune, India
shreyas.manwadkar21@vit.edu

Priti Patil
Information Technology
Vishwakarma Institute of Technology
Pune, India
priti.patil21@vit.edu

Ishan Gawali
Information Technology
Vishwakarma Institute of Technology
Pune, India
ishan.gawali21@vit.edu

Abstract—**In this paper, a system for detecting lung nodules is proposed. The lungs are a vital component of the body and any condition affecting them can have a significant impact on overall respiratory function. Lung cancer has a high fatality rate, making it one of the most dangerous forms of cancer. Early detection of lung nodules is crucial for accurate diagnosis of lung cancer. The proposed system employs a technique using support vector machines, gradient boosting, random forest, and a modified k-nearest neighbor algorithm to detect lung nodules. The modified k-nearest neighbor algorithm employs the use of Minkowski, Manhattan distance, and Gaussian kernel as hyperparameters to enhance the accuracy of nodule detection. The proposed model has been shown to have greater accuracy than existing models through calculations.**

Keywords—Lung Nodule, Cancer, CT, KNN, Diagnosis, Gaussian.

I. INTRODUCTION

Cancer is a condition as soon as a few of the body's tissues get larger and are out of control and migrate to other bodily regions. In the many cells that generate up the human body, cancer can develop practically anywhere[1]. Individual cells often divide to create new cells as the body needs them. New cells return old ones when they die as a consequence of ageing or damage[2].

Lung cancer starts in the lungs and can extend to the lymph nodes or even other bodily functions, like the brain. The lungs may potentially become infected with cancer from other tissues[2]. Metastases are the term used to describe the spread of cancer cells from one tissue to the next. Lung nodules in CT images can now be seen using computer-aided detection (CAD) techniques, which have recently been created which might result in incorrect descriptive statistical and incorrect diagnostics by doctors sometimes[3].

By tuning the hyperparameters, the modified KNN algorithm can improve its performance compared to the traditional KNN algorithm. This is because the optimal values of the hyperparameters can be different for different data sets and problems. By tuning the hyperparameters, the modified KNN algorithm can adapt to the specific characteristics of the data, resulting in improved performance.

In this system, the aim is to decrease the number of hyperparameters in order to shorten the execution time for lung nodule detection. By simplifying the algorithm and leaving out certain hyperparameters such as k, the study aims to improve the performance of the detection method. The accuracy and time detection of the optimized algorithm is found to be quite low [4].

II. LITERATURE REVIEW

In 2019 to more accurately detect lung nodules, they used a two-stage convolutional neural network architecture in this study. For example, in the first stage, they developed a sampling method for the nodule and using the ResDense-based U-Net segmentation architecture to roughly detect nodules. According to the current voxel point's location in relation to the nodule and its information about intensity, the method classifies the sampling area into three groups:area inwhat location the nodule is in, the region to the background closely associated with the nodule and the low-correlation background region far from the nodule.[5].

In this paper [1], the authors present the development of several experiments utilizing convolution type neural networks. The study demonstrated that radiologists' perceptions can be translated into computer systems. They addressed many of these problems and improved their methods in this study by using (a) an imitation vision-type network convolution neural network (b) the creation of a good background removal method to improve the "observation" of the image block by the neural network, and (c) They addressed several of these problems and enhanced their approaches in a radiologists' rating scale was used in this study to train the neural network. In addition to the training techniques mentioned above, several improved techniques are being tested in their lab. The development of this research will continue to be reported.

In this paper [6] they used a two-level artificial neural network (ANN) architecture; they have created a computer-aided diagnosis system in this work. This was specifically trained on the issue of identifying lung cancer nodules identified on digital chest radiographs and tested, reviewed, and evaluated.

This study proposes an autonomous computer-aided detection (CAD) method for the identification and 3D visualisation of lung nodules. The process consists of four steps. First, they use morphological techniques to improve the images and a median filter to remove slice noise. Next, they employ an adaptive threshold technique and active contour modelling to separate lung regions from the CT data. Nodule detection is the second stage, which consists of the two steps of feature extraction and classification. In order to extract features, they use 3D anatomical characteristics to lower the value of false positives and 2D stochastic features to accurately detect nodules (FP). The nodule contours are extracted by active contour modelling in the third step. The segmented nodule is subjected to a 3D visualisation technique in the final step to produce better visual results[7].

The classification scheme for normal and problematic CT scan lung pictures is suggested in this paper [8].For doctors and researchers, determining the presence of cancer cells from a lung picture is extremely challenging.As a novel approach for diagnosing lung tumours, the Grey Wolf Optimised and Whale Optimization Algorithm-Support Vector Machine (GWO & WOA-SVM) is presented.

In this paper [2] they suggest a new computer-aided identification method for lung nodules in low dose computed tomography that makes use of 3D convolutional neural networks (CNN). Both a priori knowledge of lung nodules and confusing anatomical components as well as data-driven machine-learned features and a classifier are used by the system.

This paper [9] addresses the issue of the variety of lung nodules and their proximity to the background. The development of a multi branch ensemble learning architecture called MBEL-3D-CNN is built on 3D CNN. The strategy includes three key ideas. Creating a 3D-CNN to fully utilise ensemble learning to significantly improve the generalisation performance of the 3D-CNN model, spatial data of lung nodules in 3D space, and incorporating a multi-branch network architecture that is ideal for the heterogeneity of lung nodules are examples of these three steps.

III. Proposed Methodology

In this section modified k-nearest neighbour approach is proposed for lung nodule detection. In this approach different 5 parameters are used for detection purpose. At first the dataset is given as an input. That dataset is taken from the Kagglewebsite. The dataset gives the coordinates of an image and using that coordinate the nodules are mapped and detected. There are main 5 steps which are followed in the project. The first one is image acquisition then it is sent for the preprocessing purpose. For image pre-processing standard scaler technique is used. Next step is segmentation. After that model will extract the features of that image and then accordingly the class will be identified.

1. The first step is image acquisition, which involves obtaining an image dataset.
2. The dataset is then divided into two parts: training and testing.
3. The dataset is further normalized using the Min-Max scalar algorithm and is divided into negative and positive cases.
4. Image preprocessing is done using the standard scalar algorithm, which standardizes the hyperparameters by subtracting the mean and scaling to unit variance.
5. The model then downsamples the training and testing data.
6. .Apply CV2 BRISK Algorithm. CV2 BRISK algorithm is used to select features of an image in a given dataset which is turned into grayscale. And then random images are taken to obtain binary features.
7. Applying the modified K-NN algorithm , In this model 5 hyperparameters are used which are , n-neighbours , weight , leaf-size, algorithm , metric. The Gaussian kernel is designed in weight.Algorithm can be auto, kd-tree , ball-tree , brute.
8. This 5 parameters are converted to dictionary. It is provided to grid search then grid search will select the best parameters.
9. Now this will fit the model and train the dataset after that Model can calculate the validation and training , testing score.

The process of detecting lung nodules begins with obtaining an input dataset which is then pre-processed. The data is split for training and testing purposes, and then the k-nearest neighbors algorithm is used[11]. Hyperparameter tuning is implemented in this algorithm to enhance the accuracy. The k-nearest neighbors algorithm is a lazy learning algorithm that is mainly used for classification purposes. This algorithm uses non-parametric methods and is known for its simplicity [7].

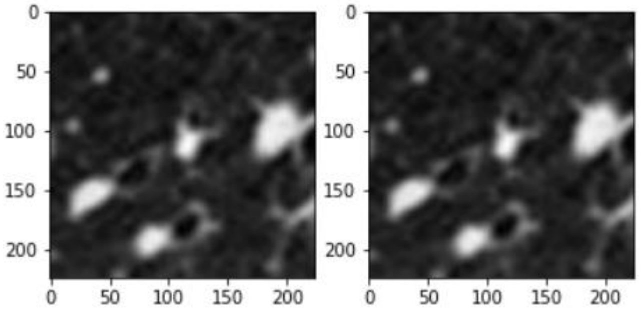


Fig 1. After pre-processing the data

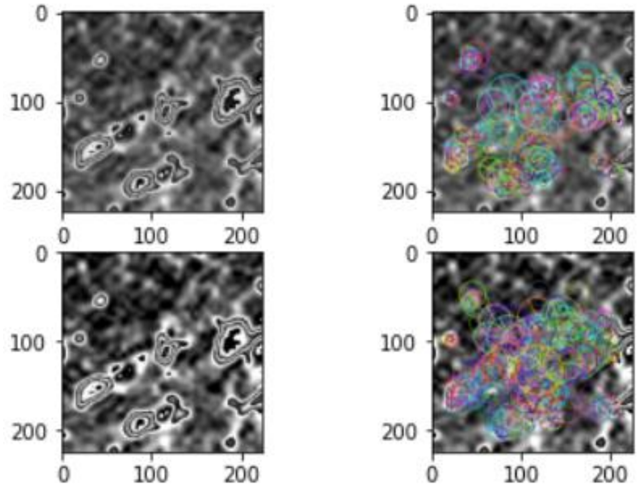


Fig 2Nodules in Lungs

K-NN contains studies about k neighbours and then classifies the test point[10].

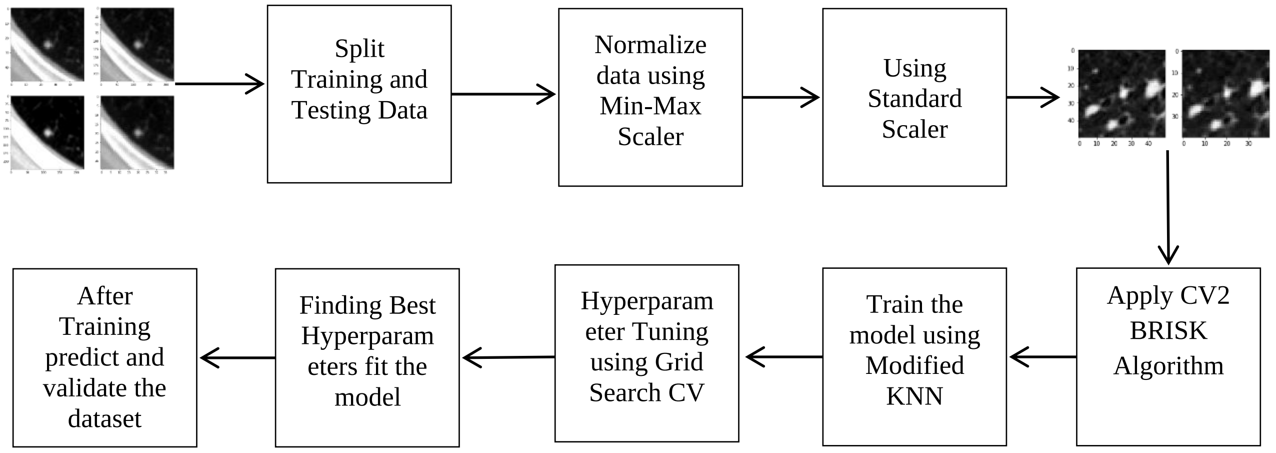

Split Training and Testing Data
Normalize data using Min-Max Scaler
Using Standard Scaler
Apply CV2 BRISK Algorithm
Train the model using Modified KNN
Hyperparameter Tuning using Grid Search CV
Finding Best Hyperparameters fit the model
After Training predict and validate the dataset

Fig 3. Project Flow Diagram for Lung Nodule Detection using Modified KNN

K-NN Algorithm [10]-
1. Decide the value of k.

2. After considering all points in a n-dimensional space the distance is calculated.

There are 3 different ways to calculate the distance: Manhattan distance metric:

$$d_{manhattan} = \sum_{i=1}^{n} |x_i - y_i| \qquad (1)$$

a. Minkowski distance metric:

$$d_{minkowski} = \left(\sum_{i=1}^{n} |x_i - y_i|^p\right)^{1/2} \qquad (2)$$

b. Euclidean distance metric :

$$d_{euclidean} = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \qquad (3)$$

1. Now the distance is sorted for all of the data points after that select the k neighbours which are nearest from the data point.
2. Now predict the class by taking the weighted average of its neighbours.
3. At last calculates the accuracy of predictions.

The k-nearest neighbors (K-NN) algorithm is a simple but powerful method for classification [11]. However, its performance can be low, to tackle this, hyperparameter tuning is applied to the K-NN algorithm to improve its performance which is called modified k-nn [12]. One of the controlling properties of the model, such as its complexity and the speed of learning [12].

In the case of k-nn algorithm, there is a research gap in the area of hyperparameter tuning. The traditional k-nn algorithm has certain fixed parameters, such as the number of nearest

neighbors (k) and the distance metric used. However, these parameters may not always be optimal for a given dataset.

Recent research has focused on developing methods to automatically tune these hyperparameters to improve the performance of the k-nn algorithm. This includes using techniques such as grid search, random search, and bayesian optimization to systematically explore the hyperparameter space and find the best combination of parameters for a given dataset.[12]

Hyperparameter Tuning into K-NN:

1. First list out the parameters that you want to tune. Here 5 different parameters are used as follows: weight, n-neighbours, algorithm, leaf-size and metric.
2. Convert those parameters to a dictionary.
3. Create the new object of k-nn using kNeighborsClassifier class.
4. Use the gridsearch.Grid search will select the best parameters among the grid of the parameters.
5. Now fit the model using the training data
6. And now you can print the value of the best hyperparameters and be able to classify the data.

In modified k-nn the model uses the manhattan distance formula:

$$d = \sum_{i=0}^{n} |x_i - y_i| \tag{4}$$

This method of modified k-nn algorithm is used for lung nodule detection, and it has been shown to achieve results that are higher than the traditional k-nn approach. Using this approach, an accuracy of 0.87 is obtained, while the normal k-nn approach achieves an accuracy of 0.84.

There are two major benefits to this approach. Firstly, it provides high accuracy with less complexity [11]. Secondly, various algorithms such as auto, ball-tree, k-d tree, and brute can be used in the hyperparameter tuning process [11]. These algorithms provide different levels of accuracy. Among these, the ball-tree and k-d tree methods have O(k * log(n)) prediction time complexity and O(k*n) training time complexity [2].

Evaluating the performance of a lung nodules dataset using multiple machine learning algorithms.

The supervised learning technique Support Vector Machines (SVM) may be applied to classification and regression issues. It determines the ideal line (or "hyperplane") dividing the dataset's various classes.

- K-Nearest Neighbors (KNN) is a straightforward, non-parametric technique for regression and classification. A new data point is categorised depending on how close it is to existing data points in the training set.
- Modified KNN is a variation of the traditional KNN algorithm, it is not a standard machine learning algorithm, Hyperparameter tuning is done to find the best parameters for the dataset.
- Gradient Boosting (GBoost) is an ensemble machine learning algorithm that can be used for classification and regression problems. It combines multiple weak models (such as decision trees) to create a stronger model.

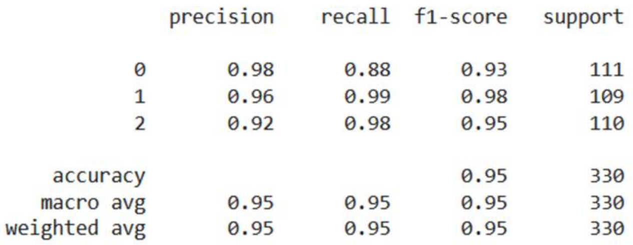
IV. Results and Conclusion

With changing lifestyle and changing climatic conditions more people are vulnerable to fatal diseases like cancer. But the ever-growing technology and use of technologies in the medical field gives these people a second chance. Use of Image Processing, Data Science and Deep Learning has allowed the medical field to achieve these heights.

The main objective was to improve the performance of the system, providing a system which will deliver more accurate results consuming less time and resources [13].

The proposed systems performance without using the hyperparameter and the results are as follows:

Using normal k-nn:



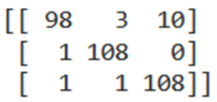
```
              precision    recall  f1-score   support

           0       0.98      0.88      0.93       111
           1       0.96      0.99      0.98       109
           2       0.92      0.98      0.95       110

    accuracy                           0.95       330
   macro avg       0.95      0.95      0.95       330
weighted avg       0.95      0.95      0.95       330
```

Fig 4 Performance Metrics

The Confusion Matrix of the System without using the hyperparameters:

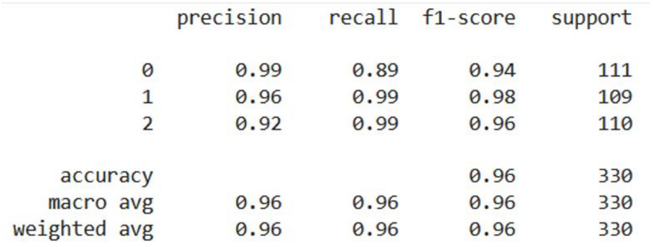


```
[[ 98   3  10]
 [  1 108   0]
 [  1   1 108]]
```

Fig 5 Confusion Matrix

The proposed systems performance using the hyperparameter are as follows:

Using k-nn with hyperparameter tuning:



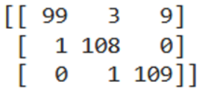
```
              precision    recall  f1-score   support

           0       0.99      0.89      0.94       111
           1       0.96      0.99      0.98       109
           2       0.92      0.99      0.96       110

    accuracy                           0.96       330
   macro avg       0.96      0.96      0.96       330
weighted avg       0.96      0.96      0.96       330
```

Fig 6 Performance Metrics

Confusion matrix using hyperparameter tuning:



```
[[ 99   3   9]
 [  1 108   0]
 [  0   1 109]]
```

Fig 7 Confusion Matrix

An analysis of the results obtained by running both systems was conducted and it was found that the implementation of hyperparameter tuning in the modified k-nn algorithm was able to significantly increase the accuracy of lung nodule detection from 84% to 95%. Additionally, this approach also resulted in a reduction of complexity and time consumption. While this increase may not seem significant, in the medical field, even small improvements in accuracy can have a significant impact, as early detection is crucial in the treatment of lung cancer and can often be the difference between life and death.

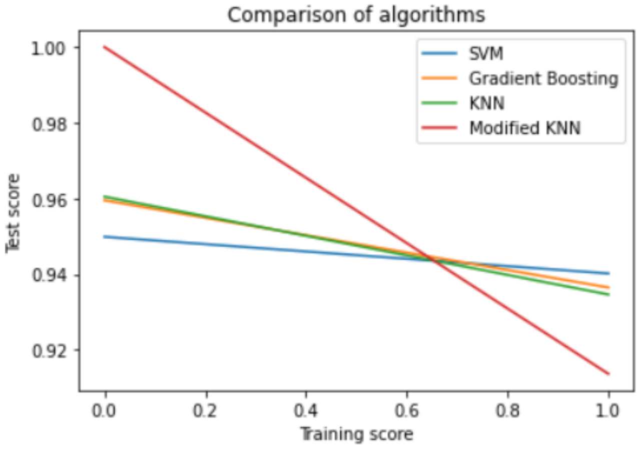


Fig 8 Comparison of Algorithms

Table 1 Comparison of Accuracy of Different Algorithms

Algorithm	Training Accuracy	Testing Accuracy
SVM	0.977058	0.955610
Gradient Boosting	0.980721	0.953144
KNN	0.991517	0.9469790
Modified KNN	0.997108	0.954946

The dataset of lung nodules was used to train a modified k-nearest neighbors (KNN) algorithm, and the algorithm achieved an accuracy of 0.997108 after 15 training epochs. This high accuracy suggests that the algorithm is able to accurately identify lung nodules in images. Using SVM algorithm we achieved an accuracy of 0.977058 after 15 training epochs, using Gradient Boosting algorithm, the algorithm achieved an accuracy of 0.980721 after 15 epochs and using Normal KNN.

V. Future Scope

Additionally, there are also methods which are trying to incorporate the hyperparameters tuning within the training process which are known as Meta-learning, Neural Architecture search. These methods are trying to make the model to learn the hyperparameters along with the actual weights, this way we are making the model more robust to different kind of data and settings.

Ensemble methods such as bagging, and boosting can be used to improve the performance of the KNN algorithm. These methods combine multiple KNN models to produce a more accurate and robust final model.

Incorporating more information such as context information, time series information, and domain knowledge can improve the performance of the KNN algorithm.

VI. References

[1] S.-C.B. Lo, S.-L.A. Lou, Jyh-Shyan Lin, M.T. Freedman, M.V. Chien,“Artificial Convolution Neural Network Techniques and Applications for Lung Nodule Detection”in the Proceedings of IEEE Trans Med Imaging, 1995 DOI: 10.1109/42.476112.

[2] Xiaojie Huang, Junjie Shan, Vivek Vaidya, “LUNG NODULE DETECTION IN CT USING 3D CONVOLUTIONAL NEURAL NETWORKS” in the Proceeding of IEEE 14th International Symposium on Biomedical Imaging, 2017 DOI: 10.1109/ISBI.2017.7950542

[3] Asma Jamil, Anup Kasi, “Lung metastases” in proceeding of StatPearls Publishing, 2022 PMID: 32310606 Bookshelf ID: NBK556146

[4] WANGXIA ZUO FUQIANG, ZHOU ZUOXIN , LILIN WANG ,”Multi-Resolution CNN and Knowledge Transfer for Candidate Classification in Lung Nodule Detection” in the proceeding of IEEE Access,2019 DOI: 10.1109/ACCESS.2019.2903587

[5] Haichao Cao, Hong Liu, Enmin Song, Guangzhi Ma, Xiangyang Xu, Renchao Jin, Tengying Liu, Chih-Cheng Hung ,”A Two-Stage Convolutional Neural Networks for Lung Nodule Detection” in the proceeding of IEEE J Biomed Health Inform, 2020 DOI: 10.1109/JBHI.2019.2963720

[6] M.G. Penedo, M.J. Carreira, A. Mosquera, D. Cabello“Computer-Aided Diagnosis: A Neural-Network-Based Approach to Lung Nodule Detection” in proceeding of IEEE Transactions on Medical Imaging Volume:7 Issue:6 DOI: 10.1109/42.746620

[7] Sarah Soltaninejad, Mohsen Keshani, Farshad Tajeripour“Lung Nodule Detection by KNN Classifier and Active Contour Modelling and 3D Visualisation” in proceeding of The 16th CSI International Symposium on Artificial Intelligence and Signal Processing, 2012 DOI: 10.1109/AISP.2012.6313788

[8] K. Vijila Rani, S. Joseph Jawhar, “Lung Lesion Classification Scheme Using Optimization Techniques and Hybrid (KNN-SVM) Classifier” in proceeding of IETE Journal of Research Volume:68, 2022 DOI: 10.1080/03772063.2019.1654935Corpus

[9] Haichao Cao, Hong Liu, Enmin Song, Guangzhi Ma, Xiangyang Xu, Renchao Jin, Tengying Liu, Chih-Cheng Hung“Multi-Branch Ensemble Learning Architecture Based on 3D CNN for False Positive Reduction in Lung Nodule Detection” in proceeding of IEEE Access DOI: 10.1109/access.2019.2906116

[10] Hamid Parvin, Hosein Alizadeh, *and* Behrouz ,Minaei Bidgoli,“Validation Based Modified K Nearest Neighbour” in proceeding of AIP Conference 1127,2009 DOI: 10.1063/1.3146187

[11] PatrickSchratz, JannesMuenchow, EugeniaIturritxa, JakobRichter, AlexanderBrenning, “Hyperparameter tuning and performance assessment of statistical and machine-learning algorithms using spatial data” in proceeding of Ecological Modelling Volume 406, 24 August 2019 DOI: 10.1016/ J.ECOLMODEL.2019.06.002

[12] Furqan Shaukat, Gulistan Raja, Ali Gooya, Alejandro F. Frangi“Fully automatic detection of lung nodules in CT images using a hybrid feature set” in proceeding of The International Journal of Medical Physis Research and Practice DOI: 10.1002/mp.12273

[13] Syed Muhammad Naqi , Muhammad Sharif ,Ikram Ullah Lali,“A 3D nodule candidate detection method supported by hybrid features to reduce false positives in lung nodule detection” in proceeding of springer link multimedia tool and application,Volume:78 Issue:18 2019 DOI: 10.1007/s11042-019