Hindawi
Mathematical Problems in Engineering
Volume 2022, Article ID 9457730, 9 pages
https://doi.org/10.1155/2022/9457730

Research Article

Classification of Thoracic Diseases Based on Chest X-ray Images Using Kernel Support Vector Machine

Rijah Khan and Tahir Mehmood

School of Natural Sciences (SNS), National University of Sciences and Technology (NUST), Islamabad, Pakistan

Correspondence should be addressed to Rijah Khan; rijahkhan2013@gmail.com

Received 8 August 2022; Revised 29 September 2022; Accepted 1 October 2022; Published 14 November 2022

Academic Editor: A. M. Bastos Pereira

Machine learning is the leading field of artificial intelligence that has achieved expert-level performance. Diagnosis and treatment of various medical diseases have led to advancements in medical imaging. Chest X-ray-based thoracic disease classification or identification is one of the potential applications in medical imaging based on machine learning. The study consists of 112,120 images of 30,804 individual patients with fourteen thoracic disease labels, which encapsulated the support vector machine (SVM). We have considered 04 kernels in SVM, namely, linear (L-SVM), polynomial (P-SVM), radial basis (R-SVM), and hyperbolic tangent (H-SVM) for classification of thoracic diseases based on X-ray images. To reduce the dimensionality and outliers from the SVM, variants are coupled with novel fast principal component analysis (FPCA). It appears that there is a significant ($p \leq 0.05$) difference between SVM variants where P-SVM and R-SVM next in order outperforms on most of the disease identification models with average validated classification accuracy ranging from 92% to 98%. The average calibrated accuracy ranges from 99.5% and reaches to 100% in most of the cases. The study is worth investigating as it is good for radiologists as they will be able to classify the diseases and it will help in improving and enhancing different medical techniques.

1. Introduction

Machine learning is one of the developing fields of artificial intelligence that has been seen significant success in recent years. It is playing an important role in solving complex algorithms and cognitive problems of computer vision. Classification is predictive modeling in which a class label is predicted for the input data. Classification algorithms are mostly used in hospitals for the accurate classification of diseases. It is commonly used in the screening and prevention of many life-threatening thoracic diseases. In medical diagnostics, the most challenging part is a multilabel classification [1, 2].

Chest radiography is an important step in radiology workflow as it needs timely reporting of possible findings and the diagnosis of diseases [3]. Radiological examination of the lungs is carried out by radiography (X-rays), computed tomography (CT), and magnetic resonance imaging (MRI). Among all of them, the X-rays comprise an innocuous and relatively inexpensive examination [4]. Retrieving all X-ray images leads to enormous amount of data. Then, this massive amount of X-ray images can be modeled using machine learning algorithms. Thoracic disorders are circumstances of the heart, lungs, esophagus, mediastinum, the chest wall, and great vessels. They are most widespread in underdeveloped and developing countries where medical facilities are inadequate. Many other factors such as overpopulation, pollution, and unhygienic conditions also increase the risk of these diseases. Therefore, inhibiting the disease from becoming mortal, early diagnosis, and controlling can play vital roles.

Identifying abnormalities in patients having thoracic diseases permits the doctor to recommend the treatment. Primary diagnosis needs a thorough evaluation of chest imaging analyses and systematic assessment of the anatomical sections of the thorax. There are high chances of no findings being reported in chest radiography. Consequently, lung function is harmed. The lung disease spreads easily such

as asthma, pleural effusion, fibrosis, and pneumonia, causing loss of versatility in the lungs, which leads to the reduction in the volume of air. Pneumonia is the single largest cause of death in children in the world as has been notified. Various research papers were taken into consideration. Some were solely based on image processing techniques while other papers involved the use of artificial neural networks for the prediction of diseases from chest X-Rays. Ke [5] used image descriptors based on the spatial distribution of Hue, saturation and brightness values in X-ray images, and a neural network co-working with heuristic algorithms to detect degenerated lung tissues in X-ray image. Poap et al. [6] presented research results on the application of heuristic method for detection over aggregated X-ray image that comes from implemented segmentation. Capizzi et al. [7] presented an evaluation model based on a composition of fuzzy system combined with a neural network. The new methodology lowers the computational demands considerably and increases detection performances. Khan et al. [8] proposed the framework VGG-SegNet which supported nodule mining and pretrained DL-based classification to support automated lung nodule detection.

To alleviate the shortcomings of previous approaches, which require more computational resources and were time-consuming, we devised a novel strategy fast principal component analysis (FPCA) which speeds up the whole process, as an alternative of standard PCA [9–11]. Also, we applied an effective strategy named support vector machine (SVM) [12] with different kernels, namely, linear (L-SVM), polynomial (P-SVM) [13], radial basis (R-SVM) [14], and hyperbolic tangent (H-SVM) [15] to get the best accuracy for all the diseases. We have classified the fourteen diseases at once. Classification accuracy is used to calculate the performance of a model based on the predicted class labels.

The residual part of the paper is assembled as follows: Section 2 discusses the data set. Section 3 represents the dimension reduction and classification method of the chest X-ray images. Experimental results are stated in Section 4.

2. Chest X-ray Data

Database used is published by the National Institutes of Health (NIH) Clinical Center, namely "ChestX-ray14." The data used to support the findings of this study can be accessed through the link https://paperswithcode.com/dataset/chestx-ray14. This database is embodied of 112120 frontal-view images collected with disease labels from 30,805 unique patients. Images are classified as thoracic diseases, that is, atelectasis, consolidation, infiltration, pneumothorax, edema, emphysema, fibrosis, effusion, pneumonia, pleural thickening, cardiomegaly, nodule, mass, and hernia. The 60,412 images do not hold the listed 14 diseases, and they are labeled as "no findings." The distribution of the number of X-ray images related to fourteen thoracic diseases is presented in Table 1 and in Figure 1.

Table 1: Distribution of the X-ray images related to fourteen thoracic diseases.

Atelectasis	4215	Cardiomegaly	1093	Consolidation	1310
Edema	628	Effusion	3955	Emphysema	892
Fibrosis	727	Hernia	110	Infiltration	9547
Mass	2139	No finding	60412	Nodule	2705
Pleural thickening	1126	Pneumonia	322	Pneumothorax	2194

3. Methods

In this paper, support vector machine and fast principal component Analysis are the algorithms that are applied to the chest X-rays. In that vein, the effect of the kernel function on diseases is investigated. The flow chart indicates the opted methodology in Figure 2.

3.1. Fast Principal Component Analysis (FAST-PCA). In order to address the dimensionality in X-ray image data, principal component analysis (PCA) [9–11] is a basic data compression tool. We have opted Fast PCA which works more efficiently for dimension reduction with uncorrelated features.

Consider a data set Y containing N observations of P features $(P < N)$.

$$\frac{1}{N}\sum_{n=1}^{N}\left\{u_1^T y_n - u_1^T \overline{y}\right\} = u_1^T S u_1, \tag{1}$$

where $\overline{y} = 1/N\sum_{n=1}^{N} y_i$ is the mean.

$S = 1/N\sum_{n=1}^{N}(y_n - \overline{y})(y_n - \overline{y})^T S$ is the covariance matrix of data set Y.

3.2. Support Vector Machines. For classification and regression problems, a supervised machine learning algorithm was originally proposed by Cortes and Vapnik, commonly known as support vector machine (SVM) [12]. Kernels are the mathematical transformation functions that are used to transform two-dimensional points into n-dimensional space, and a hyperplane is created which divides the data set into distinct sets.

3.2.1. Linear Kernel (L-SVM). The linear kernel is the simplest kernel function that deals with linearly separable data. The kernel function is as follows:

$$K\left(\overrightarrow{x}_i, \overrightarrow{x}_j\right) = \overrightarrow{x}_i^T \cdot \overrightarrow{x}_j + C. \tag{2}$$

The mathematical formula for linear kernel is as follows:

Optimization of the C regularization parameter is only required for linear kernel.

3.2.2. Polynomial Kernel (P-SVM). The polynomial kernel is a nonstationary kernel, and it works for both the hard margin and soft margin. Polynomial kernels are suitable for problems where all the training data are normalized.

$$K\left(\overrightarrow{x}_i, \overrightarrow{x}_j\right) = \left(\alpha + \overrightarrow{x}_i \cdot \overrightarrow{x}_j + C\right)^p. \tag{3}$$

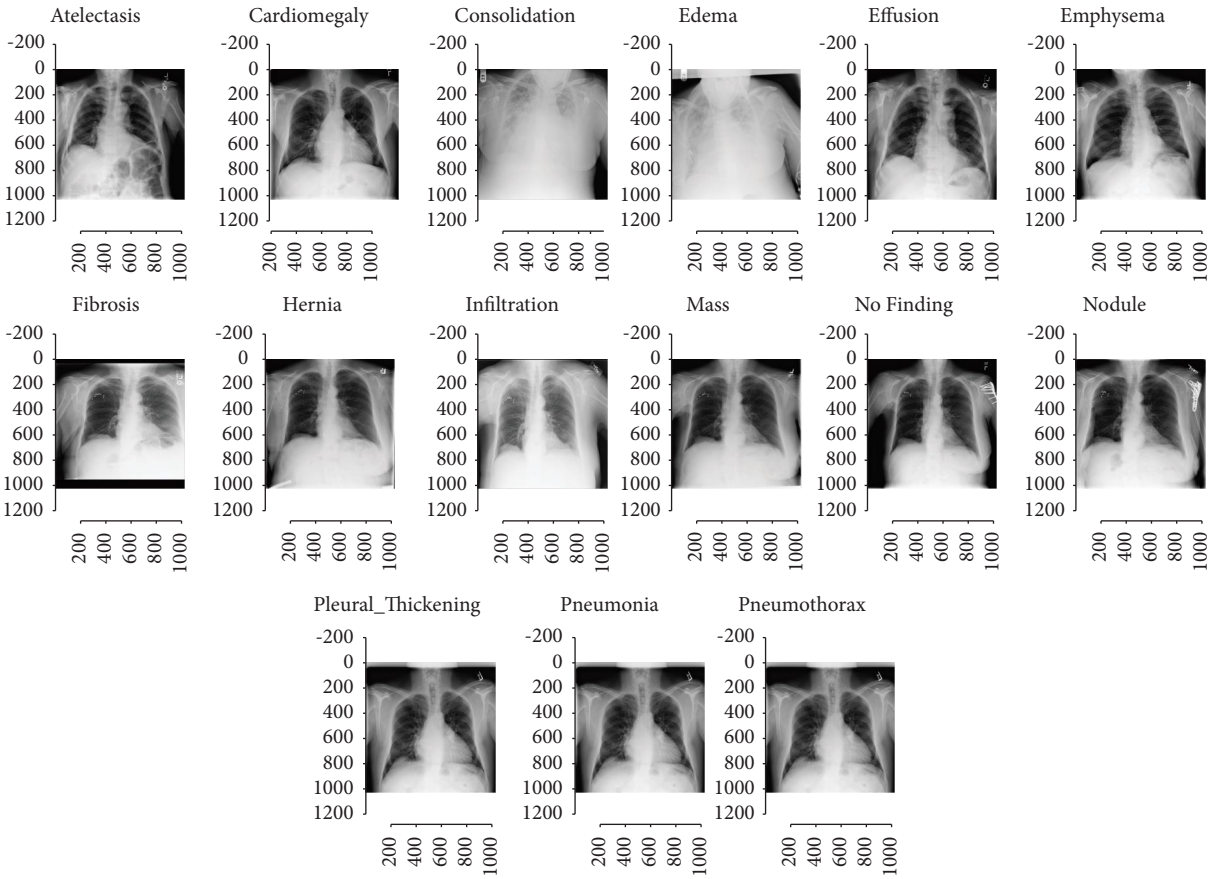

Figure 1: X-ray images of fourteen thoracic diseases and no finding.

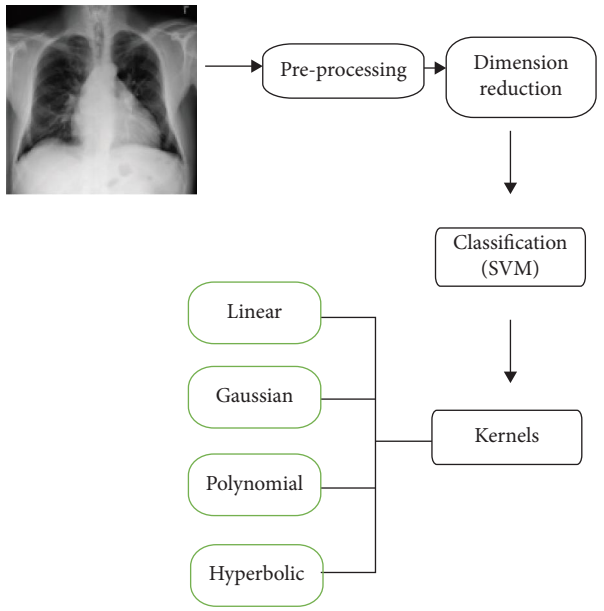

Figure 2: Flow chart indicating the opted methodology.

The kernel parameters are alpha, c, and d (polynomial degree) which are adjustable.

3.2.3. Gaussian Radial Basis Function Kernel (R-SVM). When there is no prior knowledge of data, radial basis function (RBF) kernel is used to perform a transformation. It has two parameters that must be considered: C and gamma. The C parameter is common for all SVM kernels. The low estimate of the C parameter creates the smooth decision surface, while a high value aims at classifying all training sets correctly. The gamma value defines how much a single training example has influence.

$$K\left(\vec{x}_i, \vec{x}_j\right) = e^{-\frac{\left\|\left(\vec{x}_{i-} \vec{x}_j\right)\right\|^2}{2\sigma^2}}. \tag{4}$$

3.2.4. Hyperbolic Tangent Kernel (H-SVM). The hyperbolic tangent kernel has two other names, namely, sigmoid kernel and multilayer perceptron kernel. Basically, it comes from the field of neural networks. The bipolar sigmoid function is applied as an activation function for artificial neurons.

$$K\left(\vec{x}_i, \vec{x}_j\right) = \tanh\left(\alpha \vec{x_i} \cdot \vec{x_j} + C\right). \tag{5}$$

The two parameters of the sigmoid kernel, alpha, and c are slope and intercept, respectively.

4. Results and Discussion

The work is implemented by removing the multiple disease images from the chest X-ray data set as they were making 836 classes which was difficult to handle. So, the 20796 multiple disease images were removed and remaining 91324 images of data were used for further work. The actual images of the database have high resolution, and this imposes challenges for computer hardware. Also, designing the machine learning algorithm is very difficult in this case. The size of the original images is $1024 \times\times 1024$ pixels. After rescaling the images, the size becomes 256×256 pixels, which makes the region of interest properly identified. Rescaling has been carried out on the whole database images without losing significant information. Figure 3 illustrates the rescaling of the image. From two images, one can see the original image and the resized image.

Our data set has a multiclass classification problem, and now we have fifteen classes. One-vs-one approach splits our multiclass classification data set into binary classification problem. It classifies our data sets whether the patient has disease or not and then labels each class to −1 and 1. The purpose of fast PCA was to remove the outliers and reduce the dimensions of the chest X-ray images, which helps to find the most related symptom subsets. The results for fast principal component analysis of first thoracic disease are shown in Figure 4. PCA biplot represents the distance between the observations and variables. The X-axis in biplot describes the feature that contributes towards PC1. On the other hand, the Y-axis describes the feature that contributes towards PC2. The points lying out of the outer circle are marked as outliers. Similarly, the points lying inside of the inner circle are marked as colinear.

Before applying SVM, it is important to get the data in the right shape through standardization. PCA is very sensitive to variables that have different value ranges. The variables with larger ranges will predominate over those with smaller ranges if there are significant differences in the initial variable ranges. To prevent this, we must standardize the range of the initial variables in order to ensure that each variable contributes equally to the analysis. Mathematically,

$$z = \frac{\text{value} - \text{mean}}{\text{standarddeviation}}. \tag{6}$$

Standardization is also named a z-score standardization because it has the characteristics of a standard normal distribution with a mean zero and a standard deviation one. When the standardization is done for all the variables, they are on the same scale from range 0 to 1. The histogram shows the rescaled distributions of the first thoracic disease. It has a negatively skewed distribution as more values are concentrated on the right-hand side of the distribution as shown in Figure 5.

The findings of the Monte Carlo cross-validation are obtained through the repetitive random selection and statistical methods. In Monte Carlo simulation, each data point is tested arbitrary times. Multiple probability simulation and repeated random subsampling cross-validation are another names for Monte Carlo simulation. This method is comparable to random experiments in which the precise outcome is uncertain. The result has high bias and low variance. It splits the training data set at random maybe 70-30 percent or 60-40 percent. We have used R software package "e1071," which is developed for the support vector machine to help us to do the classification. For the comparison of different methods, we have used Monte Carlo simulation with 12 runs. When the runs were increased, the computer hanged from processing. It was assumed that the data were huge and required more computer hardware space. However, SVM, like other machine learning methods, has a weakness, that it is time processing. The data are divided into 70% training and 30% testing. Accuracy of all fitted models is computed for both testing and training data. The calibration accuracy of each disease is computed and is presented in Figure 6. All kernels show nearly 99.8% accuracy for most of the diseases. The pleural thickening and pneumothorax diseases have 100% accuracy for polynomial kernel. The polynomial kernel has the highest accuracy for all the diseases while the hyperbolic tangent kernel has the lowest accuracy among these kernels. The next graph compares the classifier accuracy over testing data as shown in Figure 7. All four kernels show accuracy around 95% except hyperbolic tangent kernel. The radial basis function and polynomial kernel out perform all kernels by showing around 96% accuracy for all diseases, whereas hyperbolic tangent kernel has nearly 92% accuracy, which appears the worst classifier. The edema and effusion diseases beat other thoracic diseases as their accuracy for all models ranges from 94% to 98. PCA has slightly improved the results of all four kernels by removing the outliers. The parameter being chosen from each run is presented in

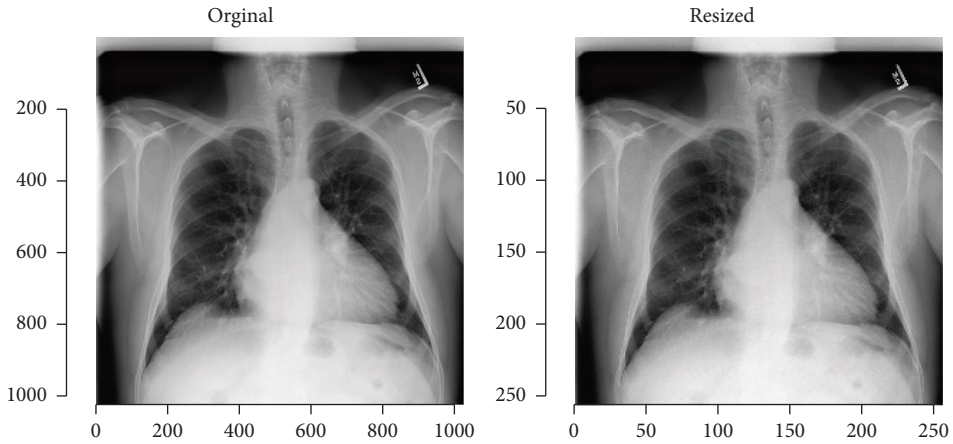

Figure 3: Rescaling of the chest X-ray image compared to original image.

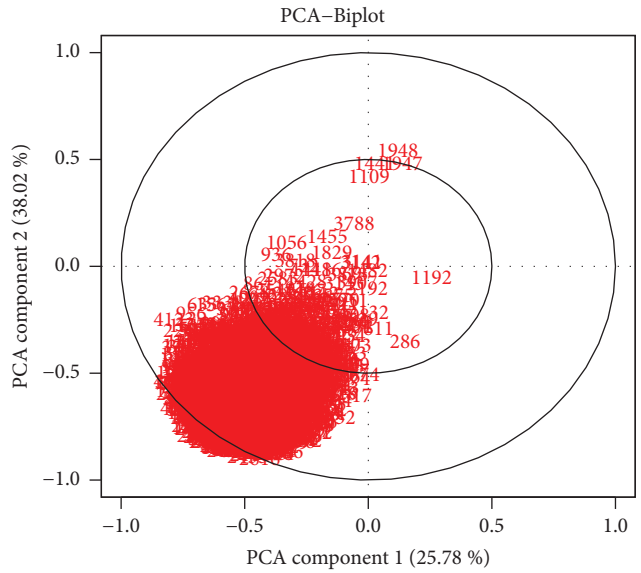

Figure 4: PCA biplot presents the patients having atelectasis disease. The first PC captures 28.61% of the variation, and the second PC captures 38.45% of the variation. As a whole, PC1 and PC2 cover over 64% of the variation.

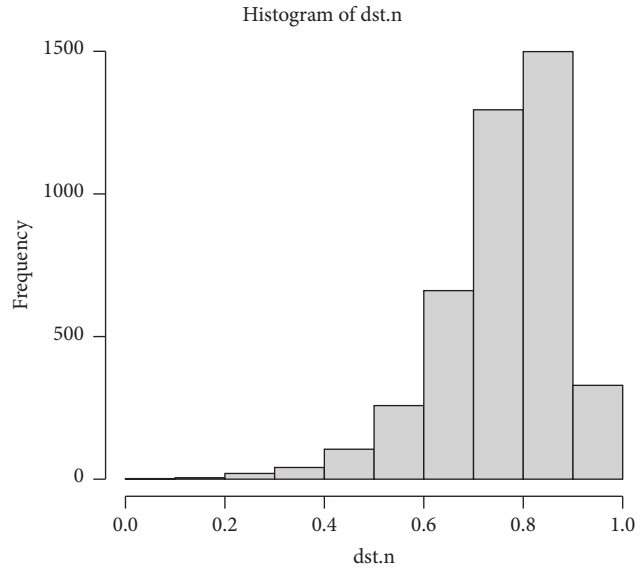

Figure 5: Histogram shows the distribution of atelectasis disease. The atelectasis disease has a frequency range from 0 to 1500.

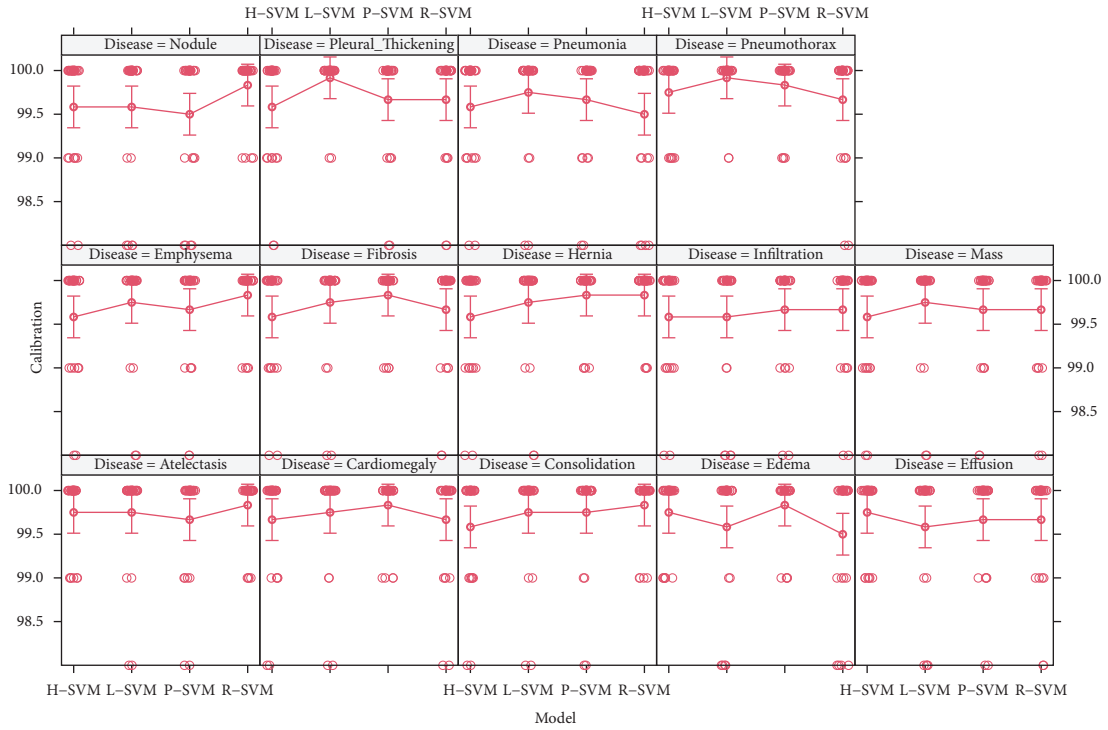

Figure 6: Calibration accuracy of each disease for all considered SVM kernels.

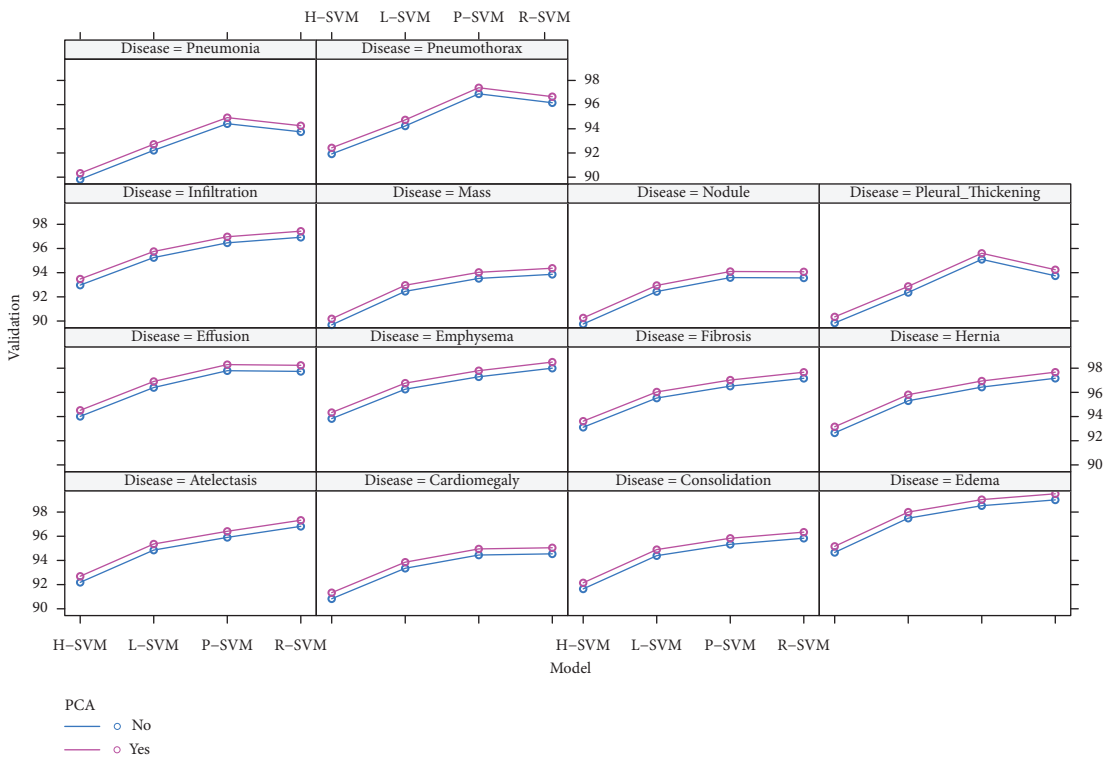

Figure 7: Validation accuracy of each disease for all considered SVM kernels.

Figure 8. The graph presents the distribution of C for different kernels. C parameter is used for controlling the outliers; lower value of C implies allowing more outliers while higher value of C implies allowing fewer outliers. We have used C value ranges from 0 to 3.8. The Polynomial kernel is the best kernel because it had the highest C value which is allowing lowest outliers. The diseases such as cardiomegaly and pneumothorax have highest value of C around 3 for polynomial and linear kernels, respectively. Most of the diseases show leptokurtic distribution while

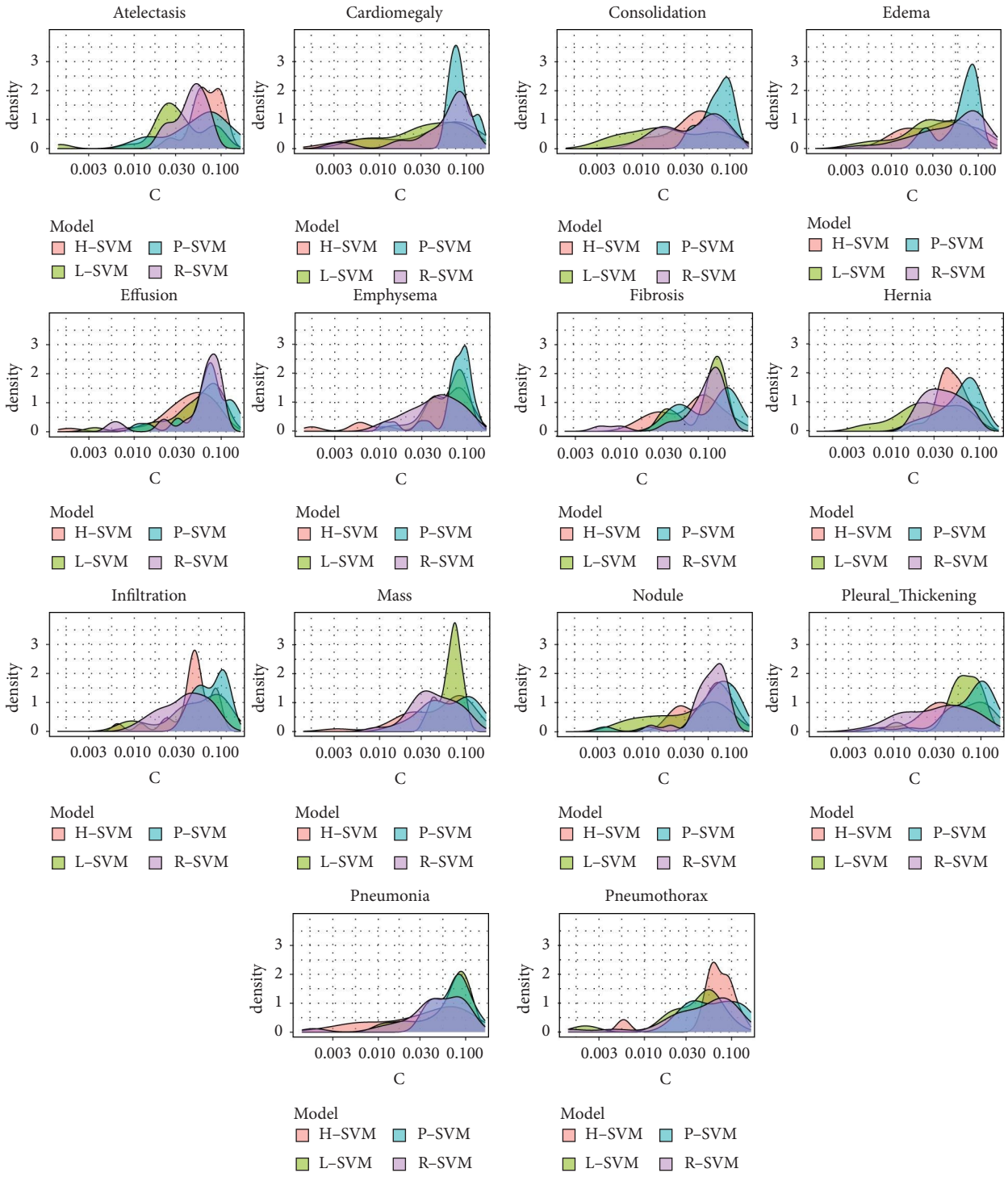

Figure 8: Distribution of C of each disease for respective kernels.

pleural thickening and pneumonia diseases have platykurtic distribution. Finally, to statistically describe the impact of kernels, we have used the analysis of variance (ANOVA) approach. It is a statistical method that separates observed variance data into different components that are used for additional tests. The response is taken as validated accuracy, whereas the linear, polynomial, Gaussian kernels, thirteen thoracic diseases of chest X-rays, and PCA are taken as factors. The ANOVA findings are presented in Table 2. It shows that all models and diseases are significantly varying, $p - value \leq 0.001$, except pneumothorax, which is not showing significance. The t-value is used to compare the means of all models and diseases. The standard error shows the uncertainty of all models which is 0.8 and for all diseases it is 0.14. When outliers are removed from the data set, then the standard error is 0.5, which is significant.

Table 2: ANOVA results are presented, where the response is taken as validated accuracy, whereas the linear, polynomial, and Gaussian kernels, thirteen thoracic diseases of chest X-rays, and PCA are taken as factors. It appears that all methods are significantly varying $p - value \leq 0.001$.

Factor	Level	Estimate	Std. error	t-Value	p-Value
(Intercept)		92.29	0.11	809.98	0.001
Model	L-SVM	2.54	0.08	33.49	0.001
	P-SVM	3.95	0.08	52.00	0.001
	R-SVM	4.10	0.08	53.92	0.001
Disease	Cardiomegaly	−1.65	0.14	−11.60	0.001
	Consolidation	−0.64	0.14	−4.51	0.001
	Edema	2.48	0.14	17.46	0.001
	Effusion	1.54	0.14	10.86	0.001
	Emphysema	1.40	0.14	9.88	0.001
	Fibrosis	0.64	0.14	4.49	0.001
	Hernia	0.45	0.14	3.15	0.001
	Infiltration	0.46	0.14	3.25	0.001
	Mass	−2.56	0.14	−18.01	0.001
	Nodule	−2.60	0.14	−18.32	0.001
	Pleural_thickening	−2.18	0.14	−15.34	0.001
	Pneumonia	−2.39	0.14	−16.81	0.001
	Pneumothorax	−0.14	0.14	−0.99	0.32
PCA	Yes	0.50	0.05	9.38	0.001

5. Conclusion

Based on experiments conducted, it can be concluded that the combination of SVM and PCA method has obtained good results. The average calibrated accuracy for all methods is 100%. The Gaussian radial basis function and polynomial kernel produce the best performance of correct classification for most of the diseases as compared to the other types of kernel functions. This has verified the need for proper kernel trick function choice which would yield more accurate results. The limitation in this study was the multilabel classification problem which was not resolved. The model of SVM training still has room for improvement. The new kernels could create more smooth results.

Data Availability

The data used to support the findings of this study can be accessed through the link https://paperswithcode.com/dataset/chestx-ray14.

Ethical Approval

The study was conducted in compliance with ethical standards.

Conflicts of Interest

The authors declare that there are no conflicts of interest.

References

[1] A. Bohr and K. Memarzadeh, "The rise of artificial intelligence in healthcare applications," in *Artificial Intelligence in healthcare*, pp. 25–60, Elsevier, Amsterdam, Netherlands, 2020.

[2] K. Arulkumaran, M. P. Deisenroth, M. Brundage, and A. A. Bharath, "Deep reinforcement learning: a brief survey," *IEEE Signal Processing Magazine*, vol. 34, no. 6, pp. 26–38, 2017.

[3] Y.-X. Tang, Y. B. Tang, Y. Peng et al., "Automated abnormality classification of chest radiographs using deep convolutional neural networks," *NPJ digital medicine*, vol. 3, pp. 70–78, 2020.

[4] A. B. Wolbarst, *Looking within: How X-ray, CT, MRI, Ultrasound, and Other Medical Images Are Created, and How They Help Physicians Save Lives*, Univ of California Press, Berkeley, California, 1999.

[5] Q. Ke, "A neuro-heuristic approach for recognition of lung diseases from x-ray images," *Expert Systems with Applications*, vol. 126, 2019.

[6] D. Poap, M. Wozniak, R. Damaševičius, and W. Wei, "Chest radiographs segmentation by the use of nature-inspired algorithm for lung disease detection," in *Proceedings of the 2018 IEEE Symposium Series on Computational Intelligence (SSCI)*, pp. 2298–2303, IEEE, Bangalore, India, November 2018.

[7] G. Capizzi, G. L. Sciuto, C. Napoli, D. Połap, and M. Woźniak, "Small lung nodules detection based on fuzzy-logic and probabilistic neural network with bioinspired reinforcement learning," *IEEE Transactions on Fuzzy Systems*, vol. 28, no. 6, pp. 1178–1189, 2020.

[8] M. A. Khan, V. Rajinikanth, S. C. Satapathy et al., "Vgg19 network assisted joint segmentation and classification of lung nodules in ct images," *Diagnostics*, vol. 11, no. 12, p. 2208, 2021.

[9] K. R. Gabriel, "The biplot graphic display of matrices with application to principal component analysis," *Biometrika*, vol. 58, no. 3, pp. 453–467, 1971.

[10] A. Maćkiewicz and W. Ratajczak, "Principal components analysis (pca)," *Computers & Geosciences*, vol. 19, no. 3, pp. 303–342, 1993.

[11] P. Geladi, H. Isaksson, L. Lindqvist, S. Wold, and K. Esbensen, "Principal component analysis of multivariate images," *Chemometrics and Intelligent Laboratory Systems*, vol. 5, no. 3, pp. 209–220, 1989.

[12] W. S. Noble, "What is a support vector machine?" *Nature Biotechnology*, vol. 24, no. 12, pp. 1565–1567, 2006.

[13] D.-X. Zhou and K. Jetter, "Approximation with polynomial kernels and svm classifiers," *Advances in Computational Mathematics*, vol. 25, no. 1-3, pp. 323–344, 2006.

[14] L. M. Manevitz and M. Yousef, "One-class svms for document classification," *Journal of Machine Learning Research*, vol. 2, pp. 139–154, 2001.

[15] M. Sellathurai and S. Haykin, "The separability theory of hyperbolic tangent kernels and support vector machines for pattern classification," in *Proceedings of the 1999 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings. ICASSP99 (Cat. No. 99CH36258)*, vol. 2, pp. 1021–1024, IEEE, Phoenix, Arizona, USA, March 1999.