 ISSN 2723-6471

Image Classifier based on Histogram Matching and Outlier Detection using Hellinger distance

Anamika Gupta [1], Sarabjeet Kaur Kochhar [2,*], Anurag Joshi [3]

[1,3] *S.S College of Business Studies, University of Delhi, India, 110089*
[2] *Indraprastha College for Women, University of Delhi, India, 110054*

(Received: September 9, 2023; Revised: October 12, 2023; Accepted: November 26, 2023; Available online: December 6, 2023)

Abstract

Thoracic diseases, including tuberculosis, pneumonia, lung cancer etc., affect millions of people worldwide every year. Interpretation of chest radiograph for the detection of disease is a time-consuming task and typically requires expert radiologists to interpret the images. In this paper, an effort has been made to develop a prediction model based on histogram matching of Chest X-ray images which can be used as a replacement of human-level interpretation task. Hellinger distance metric is used to match two histograms. The chest x-ray images are pre-processed and converted to histograms. A benchmark histogram is obtained by finding the average of all pixel intensity values. Then outlier images are detected by comparing the histogram of an image with the benchmark histogram using the hellinger metric. Finally, a prediction method is proposed which matches the histogram of unseen images to histograms of nearest neighbor images. Hypertuning of input parameters to the proposed prediction method is performed to get the best set of parameters. The proposed model gives an accuracy of 92.3 % and F1 score of 94.6 % on the training set, accuracy of 86.2% and F1 score of 89.6% on the test set. The performance of the proposed model is better than the existing techniques in this domain.

Keywords: Histogram matching; Image classification; Hellienger distance; lazy classification; Outlier detection

1. Introduction

Acute respiratory illnesses like pneumonia can be brought on by infections with bacteria, viruses, or other microbes. Pneumonia patients experience respiratory difficulties [1]. Around 15% of young children under the age of five pass away from pneumonia each year. Most diseases, including pneumonia, are now easier to treat thanks to advancements in medical technology. Chest X-rays, CT scans of the lungs, chest ultrasounds, needle lung biopsies, and chest MRIs can all be used to diagnose pneumonia [1].

Image processing is a technique for removing important information from images. In order to obtain pertinent information, useful characteristics of the image are extracted utilizing feature extraction methods and pre-processing techniques. The photographs' features are taken out so that machine learning techniques can be applied to them. The image features can be extracted using a variety of techniques [17].

In order to prevent irrelevant and redundant information from impairing the performance of the machine learning models, images are pre-processed to eliminate noise, redundancy, and missing data. Researchers have employed a variety of pre-processing strategies, such as deleting insufficient information, detecting and removing outliers, identifying the pertinent features, and standardizing the features [18].

Chest x-ray images have been explored by several researchers in the past using machine learning techniques. Several researchers have used the lazy learner technique for classification in case of Pneumonia detection. [2] uses preprocessing techniques like image resizing and normalization and then uses KNN algorithm for finding the k nearest neighbors. [3] uses the local binary patterns to extract the features from the images and then apply KNN algorithm to compute the accuracy of the model. Another group of researchers [4] uses feature extraction and KNN method for

***Corresponding author: Sarabjeet Kaur Kochhar (skaur@ip.du.ac.in)**
DOI: https://doi.org/10.47738/jads.v4i4.114

 ISSN 2723-6471

classification purposes on the chest x-ray data. [5] uses KNN method for classifying Covid-19 normal and abnormal cases using the chest x-ray dataset.

A novel method is proposed by [6,19,21] for content-based image retrieval based on histogram matching and KNN classification. An image retrieval approach is proposed in [7] where color histograms are used for searching the images based on color content. Another research uses histogram matching for face recognition which can handle variations in illumination and expression [8]. Further features like histogram are explored by [9][10] where KNN algorithm is used for classification. An image classification technique based on gabor filters is proposed in [11] where local binary patterns are used to represent the histograms.

The Hellinger distance is a popular metric used in histogram matching techniques. It measures the similarity between two probability distributions, which can be used to match histograms. The formula to compute the distance is given below:

$$d(H_1, H_2) = \sqrt{1 - \frac{1}{\sqrt{H_1 H_2 N^2}} \sum_I \quad \sqrt{H_1(I).H_2(I)}} \tag{1}$$

Several researchers have worked on the use of hellinger distance for histogram matching. [15] uses this metric to handle variations in image noise and proposes a robust principal component analysis framework. [14] uses the combination of local binary patterns and hellinger distance for image retrieval systems. [13] utilizes this metric for matching facial histograms resulting in robust feature representations. Illumination and contrast between images is handled using hellinger distance based histogram matching in [12].

1.1. Objective of the Proposed Work

Objective of this work is to design an efficient and accurate prediction model for chest x-ray images. The proposed work pre-processes the image dataset by converting the images to histograms, remove the outliers using histogram matching based on hellinger distance, hypertune the parameters for classification algorithms, and develop a new prediction model using lazy learner technique.

1.2. Organization of the Paper

Section 2 describes the methodology used in the paper which includes preprocessing, outlier detection and the prediction algorithm. Section 3 describes the experiments and their corresponding results. Conclusion is given in section 4.

2. Methodology

Chest-xray-pneumonia dataset from kaggle[16] was used for the study. There are two classes of the images - Pneumonia/Normal. The images were preprocessed, outliers were removed, features (histogram) were extracted, and then a lazy classification method based on histogram matching is used to classify the images. All the steps are mentioned below in detail:

2.1. Preprocessing

The following preprocessing steps are as follows:

1) All the images were converted to grayscale and then cropped to contain only necessary area for classification (refer to Figure 1).
2) Since images in the dataset were of different sizes, All the cropped images were resized to a similar dimension, say d1 × d2.
3) The images were divided into n × n grid (n can be any value experimentally chosen w.r.t. the data used). And hence, each image now can be represented as a 3D array having the shape (n*n × floor(d1/n) × floor(d2/n)) (Refer to Figure 2).
4) Now, for every sub-image of each image (divided into n × n grid), 255 bins were created where, each bin i corresponds to the frequency of the pixel intensity value i occurring in the image (same as in histograms). pixel intensity value 0 should not be taken into account. The transformed image now can be represented as a 2D array having shape (n*n × 255).

 ISSN 2723-6471

All the image data belonging to training & testing set should be transformed following the previous preprocessing steps and the new dataset thus formed should have the shape (m × n*n × 255), representing m images in the dataset, each having the shape (n*n × 255).

Figure 1. Chest X-ray original image

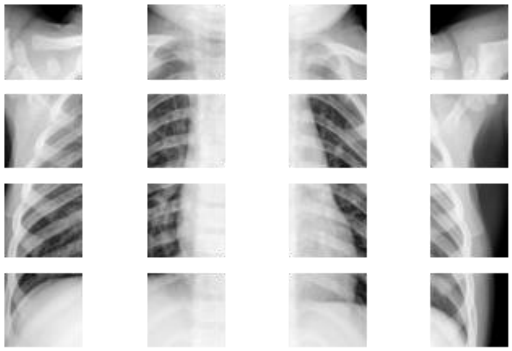

Figure 2. Image after cropping and dividing into 4 x 4 grid

2.2. Outlier Detection & Removal

To remove outliers from the images, we propose a new approach which works by removing outliers from each class of images separately. There are two classes of images in the given dataset Pneumonia/Normal. The algorithm for removing the outliers is given below:

For each class label c of images in the dataset:

1) Find the benchmark histogram for the each class of images:
   a) Create a list, say l, of size equal to 256, each index of the list representing the total count of the respective pixel intensity value in all of the images, initially all set to value 0.
   b) For each image, add the total count of each pixel intensity value found in the image to the corresponding index of the list l.
   c) Divide each item of the list l with n (= total number of images belonging to class label c).
   d) The list l will act as the benchmark histogram for all images belonging to the class label c.
2) Create another list, say image_dist. And, for each image instance that belongs to the class c, calculate the hellinger distance, between the histogram of the image instance and the benchmark histogram (list l) and append the distance to the list image_dist.
3) Calculate the mean, μ, and standard deviation, σ, of the list of distances image_dist.
4) For each image instance, img, do:
   a) Calculate the hellinger distance, d, between the histogram of the image img and the benchmark histogram of the class (created in step 1).
   b) Calculate the z_score = (d - μ)/σ for the distance d calculated in the previous step.
   c) If the z_score for the img comes out to be between a specific threshold (abs(z_score) ≤ 3) then mark that image not an outlier, otherwise mark the image as an outlier.

 ISSN 2723-6471

Figure 3. gives an example of outlier detection. Third image in the set is an outlier.

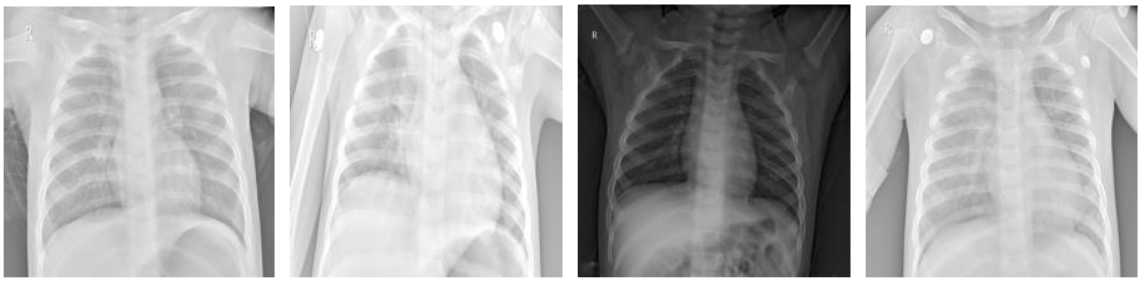

Figure 3. Outlier chest X-ray images detected by the algorithm.

2.3. Proposed machine learning Predictive Technique

The proposed technique is based on lazy classification in which no prior model is developed. The unseen image is tested against all the given images. Best match is predicted amongst the most matching images. Thus, for classifying an image, the algorithm finds the k-nearest image histograms to a given image using the hellinger distance metric between the image histograms. The algorithm is as follows:

For each instance image i in the testing data (represented by a 2D array of shape (n*n × 255)), do:

1) Create a list, say label_scores of size n (n = no. of classes) where element at index c in the list will store the score of the ith image falling into that class c. Initialize the list with 0 values.
2) For each row j of the n*n rows, do:
   a) Create an empty list of pairs, say dist_label.
   b) For each instance p in the training set, do:
      i. Calculate the hellinger distance, d, between the jth row vector (histogram with 254 bins) of the test instance i and corresponding jth row vector of the train instance p.
      ii. Make a pair (d,label) of calculated distance d and class label of the training instance p, append the pair to the list of pairs dist_label (created in step 2a)
   c) Sort the list of pairs, dist_label, of hellinger distances & corresponding class label of training instance, calculated in step 2b, by the distance values in ascending order.
   d) Choose the first k pairs from the sorted dist_label.
   e) Now, calculate the total score for each class label c by adding the inverse of the hellinger distance of all the pairs with the second element (representing label) equal to the class label c.
   f) Add the total score of each class label c to the list label_scores at index position c.
3) Assign the index c, where label_scores[c] = max_element(label_scores), as the class label to the ith instance of the testing data.

3. Results and Discussion

Dataset Description: Chest-xray-pneumonia dataset from kaggle [16] was used by the aforementioned KNN-Histogram based machine learning model to classify between chest X-ray images with and without pneumonia disease. The dataset contained total of 5856 chest X-ray images:

1) Training + Validation set images — 5232
2) Images with pneumonia — 3883
3) Images without pneumonia — 1349
4) Test set images — 624
5) Images with pneumonia — 390
6) Images without pneumonia — 234

Various experiments were conducted to show the efficiency of the proposed model and to find the best parameters for the most accurate results. The evaluation is done on the basis of accuracy, recall, precision and F1 score [20]. The experiments are listed below:

Firstly, the outlier detection & removal algorithm was applied to the dataset with z_score threshold equal to 3.5. Then, pre-processing was done on the images. During the pre-processing steps, all the images from the Chest-xray-pneumonia dataset were converted to grayscale images, resized to dimensions 1024 × 1024 and then each image was cropped by 64 pixels from all 4 directions to include only the area useful for prediction task. After cropping, each image from the dataset was divided into a grid of size n × n (experimental results based on different values of n are shown below). Then step 4 of the pre-processing step was applied to all the images and the final transformed dataset had the shape (5856 × n*n × 1024/n × 1024/n).

Now, the proposed Histogram based ML model with k-neighbours (experimental results based on different values of k are shown below) was applied on the training set by performing a stratified 10-fold cross-validation strategy, and also applied on the test set.

3.1 Results achieved by keeping k=9 & n=4

Table 1. Scores on test set along with stratified 10-fold cross-validation scores on training set (k=9 & n=4)

	Accuracy	Precision	Recall	F1-Score
Training Set	0.923109	0.973842	0.921274	0.946792
Test Set	0.8621794	0.8438914	0.9564102	0.8966346

3.2 Results achieved by keeping k=9 & n=3

Table 2. Scores on test set along with stratified 10-fold Cross-Validation on training set (k=9 & n=3)

	Accuracy	Precision	Recall	F1-Score
Training Set	0.91313	0.971405	0.90991	0.939621
Test Set	0.8461538462	0.8340909091	0.941025641	0.8843373494

3.3 Results achieved by keeping k=9 & n=5

Table 3. Scores on test set along with stratified 10-fold Cross-Validation on training set (k=9 & n=5)

	Accuracy	Precision	Recall	F1-Score
Training Set	0.920805	0.971671	0.92024	0.945216
Test Set	0.8349358974	0.8181818182	0.9461538462	0.8775267539

3.4 Results achieved by keeping k=10 & n=4

Table 4. Scores on test set along with stratified 10-fold Cross-Validation on training set (k=10 & n=4)

	Accuracy	Precision	Recall	F1-Score
Training Set	0.922718	0.972919	0.921777	0.946578
Test Set	0.8605769231	0.841986456	0.9564102564	0.8955582233

3.5 Results achieved by keeping k=8 & n=4

Table 5. Scores on test set along with stratified 10-fold Cross-Validation on training set (k=8 & n=4)

	Accuracy	Precision	Recall	F1-Score
Training Set	0.923297	0.973624	0.921787	0.946966
Test Set	0.8541666667	0.8359550562	0.9538461538	0.8910179641

4. Discussion

The proposed model is tested on different parameters and it was observed that diving the image into grid gives better results than using the original image. Further, tests were performed to check the best value of k (number of neighbours in nearest neighbour algorithm). It was found that k=9 gives the best results.

5. Conclusion and Future Directions

In this paper, a histogram matching lazy classification technique is used to classify the chest x-ray images into normal/abnormal. Hellinger distance is used to measure the proximity between the two histograms. Preprocessing of images is conducted and then outliers are removed. Running the prediction algorithm using various parameters gives us an efficient implementation with best of the input parameters. The research work here establishes the importance of outlier removal and the fine tuning of parameters which improves the performance of the classifier. In future, several other machine learning algorithms can be tested with histogram matching. Other distance measures can be used in histogram matching to improve the performance of the model.

6. Declarations

6.1. Author Contributions

Conceptualization: A.G. and S.K.K.; Methodology: S.K.K. and A.J.; Software: A.G.; Validation: A.G. and S.K.K.; Formal Analysis: A.G. and S.K.K.; Investigation: A.J.; Resources: A.J.; Data Curation: A.J.; Writing Original Draft Preparation: A.G., S.K.K., and A.J.; Writing Review and Editing: A.G., S.K.K., and A.J.; All authors, A.G., S.K.K., and A.J., have read and agreed to the published version of the manuscript.

6.2. Data Availability Statement

The data presented in this study are available on request from the corresponding author.

6.3. Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

6.4. Institutional Review Board Statement

Not applicable.

6.5. Informed Consent Statement

Not applicable.

6.6. Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper

References

[1] World Health Organization (WHO), "Pneumonia facts," [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/pneumonia.

[2] L. Noronha, J. M. Tavares, dan J. S. Cardoso, "A KNN-Based Approach for Automatic Detection of Pneumonia in Pediatric Chest Radiographs," in *Proceedings of the International Joint Conference on Neural Networks (IJCNN)*, vol. 1, no.1, pp. 1-8, 2019.

[3] A. Mahajan, A. Agrawal, dan G. K. Phadke, "Chest X-Ray Classification Using Local Binary Patterns and K-Nearest Neighbor Algorithm," in *Proceedings of the IEEE International Conference on Electrical, Computer and Communication Technologies (ICECCT)*, vol. 1, no. 1, pp. 1-5, 2020.

[4] A. Kaur dan R. Singh, "Chest X-ray Image Classification Using K-Nearest Neighbor," in *Proceedings of the IEEE International Conference on Computing, Communication and Automation (ICCCA)*, vol. 1, no. 1, pp. 1-5, 2021.

[5] S. Doshi dan P. Kulkarni, "Classification of Chest Radiographs for COVID-19 Detection Using K-Nearest Neighbor

Algorithm," in *Proceedings of the International Conference on Data Engineering and Communication Technology (ICDECT)*, vol. 1, no. 1, pp. 1-5, 2021.

[6] J. Wang dan J. Wang, "A novel KNN-based histogram matching method for content-based image retrieval," in *Proceedings of the International Conference on Computational Intelligence and Security*, vol. 1, no. 1, pp. 672-676, 2009.

[7] S. Shinde dan R. Patil, "Image retrieval using color histogram and KNN classification technique," *International Journal of Science, Engineering and Technology Research*, vol. 3, no. 2, pp. 339-342, 2014.

[8] W. Zhang, Z. Hu, dan Z. Zhang, "Face recognition using histogram matching and KNN classification," in *2015 International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery (CyberC)*, vol. 1, no. 1, pp. 76-79, 2015.

[9] Y. Liu, Z. Wang, dan Y. Gu, "Image classification based on a new KNN algorithm using histogram matching," in *Proceedings of the International Conference on Management Science and Engineering*, vol. 1, no. 1, pp. 92-97, 2018.

[10] K. Shah, R. Shah, dan N. Shah, "A novel KNN classification algorithm for image retrieval using histogram matching," in *2020 International Conference on Inventive Research in Computing Applications*, vol. 1, no. 1, pp. 1-6, 2020.

[11] X. Lian, D. Chen, dan H. Zhang, "Image classification using Local Gabor Binary Pattern histogram sequence," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 1, no. 1, pp. 1-8, 2011.

[12] C. Chen, Y. Wang, dan Q. Wu, "A histogram matching method based on Hellinger distance for image matching," in *Proceedings of the 2017 3rd International Conference on Multimedia and Image Processing (ICMIP)*, vol. 1, no. 1, pp. 112-115, 2017.

[13] Y. Wang, D. Zhang, dan P. Liang, "Gabor-based region covariance matrices for face recognition," *Pattern Recognition Letters*, vol. 33, no. 6, pp. 708-717, 2012.

[14] D. Kim, J. Kim, dan J. Kim, "Content-based image retrieval using local binary patterns and Hellinger distance," in *Proceedings of the 13th International Conference on Advanced Communication Technology (ICACT)*, vol. 1, no. 1, pp. 1569-1573, 2011.

[15] F. De La Torre dan M. J. Black, "Robust principal component analysis for computer vision," in *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 1, no. 1, pp. 362-369, 2001.

[16] Chest X-Ray Images (Pneumonia) — https://www.kaggle.com/paultimothymooney/chest-xray-pneumonia.

[17] J. I. N. H. O. Kim, B. S. Kim, dan S. Savarese, "Comparing Image Classification Methods: K-Nearest-Neighbor and Support-Vector-Machines," in *Proceedings of the 6th WSEAS international conference on Computer Engineering and Applications, and Proceedings of the 2012 American conference on Applied Mathematics*, vol. 1001, no. 1, pp. 133-138, 2012.

[18] S. B. Kotsiantis, D. Kanellopoulos, dan P. E. Pintelas, "Data preprocessing for supervised learning," *International Journal of Computer Science*, vol. 1, no. 2, pp. 111-117, 2006.

[19] L. Noronha, J. M. Tavares, dan J. S. Cardoso, "A KNN-Based Approach for Automatic Detection of Pneumonia in Pediatric Chest Radiographs," in *Proceedings of the International Joint Conference on Neural Networks (IJCNN)*, vol. 1, no.1, pp. 1-8, 2019.

[20] M. Sunasra, "Performance Metrics for Classification problems in Machine Learning," [Online]. Available: https://medium.com/@MohammedS/performance-metrics-for-classification-problems-in-machine-learning-part-i-b085d432082b.

[21] C. Agarwal dan A. Sharma, "Image understanding using decision tree based machine learning," in *ICIMU 2011: Proceedings of the 5th International Conference on Information Technology & Multimedia*, vol. 1, no.1, pp. 1–8, 2011.