International Journal of Innovative Research in Technology & Science
ISSN: 2321-1156 Volume VII Issue II, March 2019

Image Processing and Machine Learning on Chest X-Ray Dataset

Anamika Gupta, Anurag Joshi
SSCBS, University of Delhi

Abstract:- In this study, we conducted a performance comparison of multiple machine-learning models using a chest X-ray image dataset obtained from Kaggle. The images underwent preprocessing using various image processing techniques. We extracted both first-order and second-order texture features from the images. The extracted data was then standardised, and any outliers were removed. We applied the Redundant Feature Elimination technique to identify the most informative features. Subsequently, we applied several classification models, including Decision Trees, K-nearest neighbours (KNN), Naive Bayes, Neural Networks, and Support Vector Machines (SVM), to the refined dataset. We employed ten-fold cross-validation in our experiments to evaluate the models' performance. Our results indicate that SVM outperformed the other models, achieving an accuracy of 89% and an F1 Score of 91%.

Keywords: Chest X-ray image dataset, Texture features, classification models. SVM

I. INTRODUCTION

Pneumonia is an acute respiratory disease caused by viral or bacterial infections [1]. It can lead to breathing difficulties, especially in younger individuals, and reports suggest that approximately 15% of children under 5 succumb to pneumonia [13]. Fortunately, advancements in medical science have made treating Pneumonia and many other diseases more manageable. Various diagnostic techniques, including chest X-rays, CT scans, chest ultrasounds, needle biopsies of the lung, and chest MRIs, are available for detecting pneumonia [12]. Image processing is a method employed to extract valuable information from images. It involves applying preprocessing techniques to eliminate redundancy, noise, and missing values, ensuring that irrelevant or redundant information does not interfere with the performance of machine learning models. Feature extraction methods are used to obtain relevant information from images, and various methods for this purpose exist. The most common ones are based on first-order statistics and second-order statistics. First-order features, such as Kurtosis, skewness, mean, and variance, focus solely on individual pixels. On the other hand, second-order features consider the spatial relationships between adjacent pixels, with the Grey Level Co-Occurrence Matrix (GLCM) being a prominent technique. GLCM provides a matrix that reveals the frequency of pairs of pixels with specific values and spatial relationships [6]. Haralick et al. describe fourteen texture features that can be extracted using GLCM, including Angular Second Moment, Contrast, Correlation, Variance, Inverse Difference Moment, Sum Average, Sum Variance, Sum Entropy, Entropy, Difference Variance, Difference Entropy, Two Information Measures of Correlation, and Maximal Correlation Coefficient. Researchers have extensively studied chest X-ray images using machine learning and image processing techniques. Various classification methods, such as K-Nearest Neighbors (KNN), Neural Networks, Support Vector Machines (SVM), Naive Bayes, and Decision Trees, have been applied in this context. KNN classifiers are known for their simplicity, classifying new data points based on the majority vote of their neighbours [9]. SVMs optimise image classification [9], while Naive Bayes classifiers are based on the Bayes theorem [14]. Decision Tree Classifiers classify data samples by learning conditional rules from input features [1], and Neural Network models learn complex functions directly from the inputs [6]. The performance of these classification models is typically evaluated using metrics like Precision, Accuracy, Recall, Specificity, Sensitivity, F1 Score, and AUC ROC [13]. Some researchers have employed the lazy learner technique in the realm of Pneumonia detection. For instance, [2] used preprocessing techniques like image resizing and normalisation and applied the KNN algorithm to find the k nearest neighbours. [3] utilised local binary patterns for feature extraction and KNN to compute the model's accuracy. Another group of researchers [4] used feature extraction and the decision tree method for chest X-ray classification. [5] focused on neural network methods for classifying Covid-19, normal, and abnormal cases using chest X-ray data.

Machine learning has greatly benefited the healthcare field, enabling better disease diagnosis from patient

data, including medical images such as X-rays, MRI scans, and CT scans. Achieving higher sensitivity and specificity in models is a challenge. One study involved classical machine learning methods applied to a dataset containing 1100 chest X-ray images, 300 being COVID-19 patients, 400 pneumonia patients, and 400 normal X-ray images [7]. An SVM model was trained using 630 features extracted from the images, and ten-fold cross-validation was employed to evaluate its performance using various metrics. Khan et al. conducted similar work [8], using X-ray radiograph data to identify COVID-19 cases with the SVM technique, performing a three-class classification to identify normal, Pneumonia, and COVID-19 cases [9].

A. Objective of the Proposed Work

This study aims to preprocess a dataset of chest X-ray images, extract features using both first-order and second-order texture features, apply classification models, and subsequently compare the performance of various models for predicting Pneumonia disease.

B. Organization of the Paper

The paper is organised as follows: Section 2 outlines the methodology employed in this study. Section 3 presents the experiments conducted and their respective results. Finally, Section 4 provides the conclusion for the paper.

II. Methodology

The experiments utilised an image dataset of chest X-ray scans for Pneumonia detection. This dataset consists of two classes: Normal and Pneumonia. Several preprocessing steps were applied to this dataset, as outlined below:

1. The images were divided into 16 regions.
2. Both first- and second-order features (GLCM features) were extracted from all images. This process resulted in a conversion of each image into 336 features.
3. All the extracted GLCM features were standardised to ensure they were on the same scale.
4. In order to address the issue of an imbalanced class distribution, the minority class was oversampled using the SMOTE technique.
5. Feature selection was conducted using the Recursive Feature Elimination (RFE) technique, resulting in the selection of 192 features.
6. Outliers were identified and subsequently removed using the DBSCAN algorithm.

After preprocessing the images, various machine learning models are applied to find the efficiency of the model. The experimental section reports the results obtained on the above image dataset.

III. Experiments and Results

A dataset of chest X-ray images for detecting Pneumonia was utilised, containing a total of 5856 images. The distribution of these images is illustrated in Table 1.

Table 1: Distribution of Pneumonia and Normal Chest X-ray Images in Training and Test Sets

	Pneumonia	Normal	Total
Training Set	3883	1349	5232
Test Set	390	234	624

Various machine learning models, including Neural Networks, K-nearest neighbours (KNN), Support Vector Machines (SVM), Naive Bayes, and Decision Trees, were applied to the dataset, and a range of evaluation metrics were computed. These results are presented in Table 2.

Table 2: Performance Metrics for Various Machine Learning Models (NN, KNN, SVM, NB, DT)

Metric/Model	NN	KNN	SVM	NB	DT
Accuracy	0.87	0.87	0.89	0.83	0.77
Precision	0.88	0.88	0.89	0.83	0.74
Recall	0.9	0.91	0.94	0.91	0.96
Specificity	0.81	0.79	0.8	0.68	0.45
F1 Score	0.89	0.89	0.91	0.87	0.84
AUC-ROC	0.92	0.89	0.92	0.84	0.71

The results of the experiments demonstrate that the SVM model outperforms other models on the dataset mentioned above.

IV. Conclusion

In this paper, we conducted a study using an image dataset comprising chest X-rays. The dataset underwent several preprocessing techniques to prepare it for analysis. Texture features were then extracted, and various classification models were employed for

analysis. The results indicate that the SVM model outperforms all other available models. In future research, we intend to explore using deep learning models and conduct a performance comparison. Additionally, we plan to investigate further the application of other data preprocessing and transformation techniques to enhance our models' accuracy.

References

[1]. Pneumonia Chest X-Ray Images" by Paul Mooney. Available online at: https://www.kaggle.com/paultimothymooney/chestxray-pneumonia

[2]. Noronha, L., Tavares, J. M., & Cardoso, J. S. (2019). A KNN-Based Approach for Automatic Detection of Pneumonia in Pediatric Chest Radiographs. In Proceedings of the International Joint Conference on Neural Networks (IJCNN) (pp. 1-8).

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

[4]. Agarwal, C., & Sharma, A.: Image understanding using decision tree-based machine learning. In: ICIMU 2011: Proceedings of the 5th International Conference on Information Technology & Multimedia, pp. 1–8. IEEE, Kuala Lumpur (2011).

[5]. Giacinto, G., & Roli, F.: Design of effective neural network ensembles for image classification purposes. Image and Vision Computing, 19(9-10), 699-707 (2001).

[6]. Aggarwal, N., & Agrawal, R. K.: First and Second Order Statistics Features for Classification of Magnetic Resonance Brain Images. Journal of Signal and Information Processing 3(2), 146–153 (2012).

[7]. Haralick, R. M., Shanmugam, K., & Dinstein, I. H.: Textural features for image classification. IEEE Transactions on systems, man, and cybernetics SMC-3(6), 610-621 (1973).

[8]. NHLBI Website, Pneumonia, https://www.nhlbi.nih.gov/health-topics/pneumonia.

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

[10]. Çelik, M., Dadaşer-Çelik, F., & Dokuz, A. Ş.: Anomaly detection in temperature data using DBSCAN algorithm. In: 2011 International Symposium on Innovations in Intelligent Systems and Applications, pp. 91–95. IEEE, Istanbul (2011).

[11]. Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P.: SMOTE: synthetic minority over-sampling technique. Journal of Artificial Intelligence Research 16, 321–357 (2002).

[12]. Kotsiantis, S. B., Kanellopoulos, D., & Pintelas, P. E.: Data preprocessing for supervised learning. International Journal of Computer Science, 1(2), 111-117 (2006).

[13]. Kim, J. I. N. H. O., Kim, B. S., & Savarese, S.: 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, pp. 133-138. WSEAS, Wisconsin (2012).

[14]. Rish, I.: An Empirical Study of the Naive Bayes Classifier. In: IJCAI 2001 workshop on empirical methods in artificial intelligence, pp. 41-46 (2001).