Texture Feature Extraction: Impact of Variants on Performance of Machine Learning Classifiers: Study on Chest X-Ray – Pneumonia Images

Anamika Gupta[1], Anshuman Gupta[1](✉), Vaishnavi Verma[1], Aayush Khattar[1], and Devansh Sharma[2]

[1] S.S. College of Business Studies, University of Delhi, New Delhi, India
anamikargupta@sscbsdu.ac.in, {anshuman.18514,vaishnavi.18537, aayush.18576}@sscbs.du.ac.in
[2] Department of Computer Science, University of Delhi, New Delhi, India
devanshsharma80@gmail.com

Abstract. Image textures are a set of image characteristics used for identifying regions of interests (ROIs) in images. These numerical features can thus be used to classify images in various classifiers. This paper introduces the task of classifying Chest X-ray images with Machine Learning Classifiers and to see the impact of variations on the result of classification. For this purpose, second-order statistical features (GLCM texture features) are extracted from all the images with preprocessing and classification is performed using these features. Various variants are applied for image processing. First-order features are included, the image is divided into multiple regions, different values of distance for GLCM are used. Several evaluation metrics are used to judge the performance of the classifiers. Results on Chest X-ray (Pneumonia) dataset shows remarkable improvements in the accuracy, F1-Score, and the AUC of the classifier.

Keywords: Texture extraction · GLCM · First order statistics · Pre-processing

1 Introduction and Related Work

Pneumonia is a form of acute respiratory disease which can be caused by an infection due to a virus, bacteria, and sometimes by other microorganisms. Other causes of pneumonia include allergic reactions from medications or conditions such as autoimmune diseases. Pneumonia is a fatal illness in which the air sacs get filled with pus and other liquid. A person can get infected with pneumonia but conditions like cystic fibrosis, asthma, diabetes, heart failure, a history of smoking, a poor ability to cough, and a weak immune system can be risk factors making someone more prone to pneumonia [15]. People infected with pneumonia feel difficulty in breathing. Every year, around 15% of children under the age of 5 die due to pneumonia [19]. With the advancement in medical

L. Bellatreche et al. (Eds.): BDA 2020, LNCS 12581, pp. 151–163, 2020.
https://doi.org/10.1007/978-3-030-66665-1_11

technology, the treatment of most of the diseases has become easier, including pneumonia. Diagnosis of pneumonia can be done by observing the chest X-ray, CT scan of the lungs, ultrasound of the chest, needle biopsy of the lung, and MRI of the chest [15]. X-ray images are preferred over CT scan images, as the technology for X-ray imaging is easily available and widespread. Therefore, the use of Chest X-ray images, along with computer-aided technology, is becoming popular today as this approach is more cost-efficient and can benefit a larger audience. Figure 1 shows the chest X-ray scans of a normal person and pneumonia infected person.

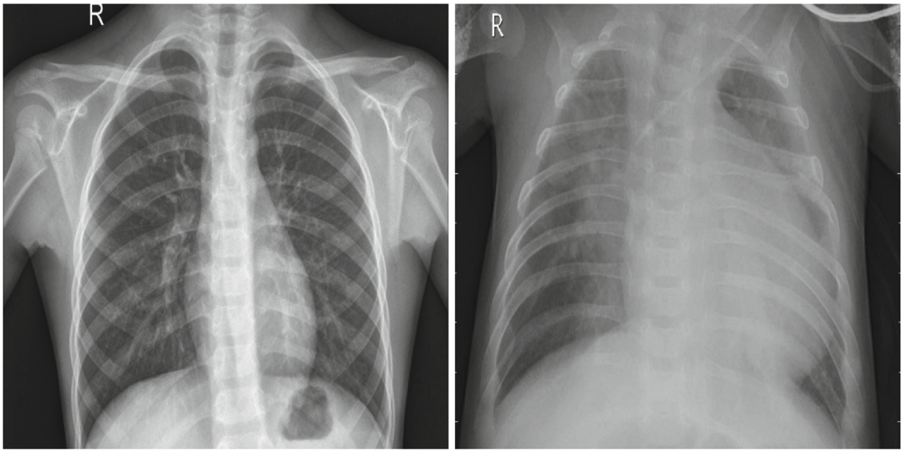

Fig. 1. Chest X-ray images of a normal person (left) and pneumonia infected patient (right) (Source: Kaggle [11]).

Image processing is the manipulation of an image to extract some meaningful information from the image. The first part of image processing is extracting some useful features from the image known as feature extraction. Feature extraction involves the careful calculation of textural features in images so as to get sufficient information for further classification and analysis while cutting back on any redundant information that might be misleading for the same. Texture feature extraction can thus be used to save memory, time, and computation costs. In this paper, we use GLCM (Grey level co-occurrence matrix) for feature extraction. It is a traditional approach to texture analysis with a number of applications, especially in medical image analysis. Haralick's fourteen GLCM features are calculated and used for the classification of Pneumonia chest X-ray images [10].

The pre-processing is necessary to resolve several types of problems including noisy data, redundancy in data, missing data values, etc. as a lot of irrelevant and redundant information can heavily affect the performance of Machine Learning models. To produce more understandable and accurate results, several inadequate information is removed [6]. To detect extreme points in the dataset also known as outliers, DBSCAN is used as the outlier detection technique [4]. Redundant and irrelevant features are removed using RFE [13]. The normal class has relatively fewer samples as compared to the

Pneumonia class in the dataset. Hence, SMOTE is used for oversampling the minority class. Standardisation is used to scale the features in the dataset [7].

In several earlier studies, GLCM technique has been used for diagnosing various medical conditions. Pugalenthi et al. [16] used the technique for the detection of tumors using brain MRIs, where features were extracted using GLCM, followed by machine learning classifier to classify the MRI images using the textural features extracted by GLCM [16]. It has also been used to classify the bone X-ray images into fractured and non-fractured categories, where the second-order statistical information of gray levels between neighboring pixels extracted by GLCM are fed to different classifiers such as Logistic Regression and Decision trees for abnormality detection [14]. Ankita et al. used the technique to classify lung cancer datasets. Features were extracted using GLCM and classification was done on the images using SVM classifier [3]. Although GLCM has been prominently used in many pieces of research for feature extraction, there are several instances where various other methods have been used for feature extraction from radiographs like chest X-ray images and brain MRIs. The Wavelet transform and Curvelet transform methods have been used for the classification of early-stage lung cancer diagnosis of chest X-ray images [9]. Chandra et al. also proposed a Hierarchical feature extraction method, which has been used to extract features from chest X-ray images for detection of tuberculosis related abnormalities in chest X-ray images [5].

2 Background

Feature extraction can be performed using First Order Statistical features and Second-order Statistical features. Kurtosis, skewness, mean, variance, etc. are some of the first-order features which only consider the pixel under observance. GLCM is one of the most popular approaches for computing second-order statistical features. GLCM considers the spatial relationship of adjacent pixels in the texture analysis. GLCM provides a matrix that shows how often pairs of pixels with a specific value and in a specified spatial relationship occur [2]. Haralick et al. [10] described fourteen textural features - 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 - that can be extracted using GLCM for texture analysis of images.

Various classification techniques have been popular in literature for predicting the behaviour of unknown tuples. Some of the most popular ones are KNN, Neural networks, Support vector machine, Bayes classifier, and Decision tree classifier. KNN classifiers use data and classify new data points based on a majority vote of its neighbours [12]. SVM can be utilized to perform image classification and to optimize the classification of images [12]. Naive Bayes classifiers are probability driven ML models based on the Bayes theorem [17]. Decision Tree Classifiers classify the data samples by learning some conditional rules of the input features and are also used to train the image understanding system to accomplish supervised machine learning [1]. The neural network model is made up of neurons, which are small computational units that learn some function of the input features. Many such neurons are combined in a neural network for learning some complex functions directly from the inputs. A neural network has the ability to learn complex features from simpler ones and thus reduce the need for hand-engineered features for classification tasks of raw data, like images in various image recognition and classification tasks, and get the output directly [8].

Various evaluation metrics have been proposed in the literature for comparing the performance of the classifiers such as Confusion Matrix, Precision, Classification Accuracy, Recall, Specificity, F1 Score, and AUC ROC Curve. Confusion Matrix is a matrix that shows a visualized representation of the performance of a machine learning model [18]. It describes 4 metrics as its entries as shown in Table 1. True Positive (TP) are those samples in which both the actual and predicted class were true, that is, the object classified was true in reality and the model also classified it as true; True Negatives (TN) are those samples which the model classified as false and the target was also false; False Positives (FP) are when the actual class of a sample is false but the predicted class is true, that is, the target of the sample was false but the model predicted it as true; False Negatives (FN) are those samples for which the actual class is true but the predicted class is false, that is, the target of the sample was true but the model predicted it as false [18]. Table 2 describes other classification metrics used in the study.

Table 1. Confusion matrix and its entries for a binary classification problem.

	Predicted Negative (0)	Predicted Positive (1)
Actual Negative (0)	TN	FP
Actual Positive (1)	FN	TP

Various pre-processing techniques such as Standardisation, Oversampling, Outlier Detection, and Feature Ranking have been used in order to improve the performance of various Machine Learning algorithms such as KNN, SVM, Naive Bayes, Decision tree, and Neural network. The process of scaling one or more features/attributes of a dataset such that their mean is equal to zero and the standard deviation is equal to one is called standardisation. If x_i is the value of a feature for the i^{th} sample in the dataset consisting of m samples, $\bar{x}$ is the mean value of that feature, σ is the standard deviation, and z_i is the standardised feature value for that sample then for each feature in the dataset standardisation can be performed using Eq. (1).

$$z_i = \frac{x_i - \bar{x}}{\sigma}, \ where\ i = 1, 2, \ldots, m \tag{1}$$

The process of increasing the number of samples for the minority class by creating new samples that are similar to the existing data for that class is called oversampling. Synthetic Minority Oversampling Technique (SMOTE) is a technique that draws a line between existing samples that are close in the feature space and then randomly picks a new sample along the line [7]. Outlier detection is the process of identifying samples that are very different from other samples and differ from the overall pattern in the dataset.

Table 2. Description of classification metrics along with their formulas.

Metric	Definition	Formula
Precision	It is defined as the ratio of the number of correct positive predictions to the total number of samples that were predicted positive	$\frac{TP}{TP+FP}$
Classification accuracy	It is defined as the proportion of the total samples correctly predicted out of all the samples	$\frac{TP+TN}{TP+FP+FN+TN}$
Recall	It is defined as the proportion of positives correctly predicted by our model out of all the samples that were actually positive	$\frac{TP}{TP+FN}$
Specificity	It is defined as the proportion of negatives correctly predicted by our model out of all the samples that were actually negative	$\frac{TN}{TN+FP}$
F1 score	The harmonic mean of recall and precision is known as F1 Score	$2 * \frac{Precision*Recall}{Precision+Recall}$
AUC-ROC curve	It is defined as the area under the Receiver Operator Characteristic (ROC) curve. It tells about how capable our model is in distinguishing the classes	

Density-Based Spatial Clustering of Applications with Noise (DBSCAN) is an outlier detection technique based on the density-based clustering algorithm. It classifies various samples into three categories, namely core points, border points, and outliers using the clustering method and uses the concept of reachability for categorizing various points [4]. Machine Learning models are prone to overfitting when the number of features becomes very large. To mitigate this, features are ranked according to their importance and the least important features are discarded. In Recursive Feature Elimination (RFE), the model is fitted on the entire set of features and the importance of each feature is calculated, following which the weakest features are eliminated. This process is repeated recursively until the required number of features are selected [6].

3 Methodology and Experiments

The study in this paper focuses on analyzing the impact of applying various techniques that affect the performance of the classification on various evaluation metrics. The basic model of extracting features using the GLCM method has been developed. Standardisation, outlier handling, feature selection, and imbalance handling have been experimented with on the Chest X-Ray dataset. Experiments regarding the division of the image into multiple parts, incorporating first-order statistical features, changing the distance of pixels have been performed to check the suitable combinations. Observations have been reported in all the experiments.

3.1 Experimental Setup

Experiments are performed on the Chest X-ray images of Pneumonia patients obtained from the Kaggle website. The dataset consists of a total of 5,856 Chest X-Ray images in 3 different folders dividing the data into 3 separate sets for training, testing, and validation. The training set consisted of 5216 samples, test set consisted of 624 samples and validation set consisted of 16 samples. Since, the validation set was very small, it was not used for evaluation. All the folders are further subdivided into 2 more additional folders containing different categories of images: Normal and Pneumonia. The dataset consists of X-Rays of patients of age 1-5 as part of their routine clinical checkup from the Guangzhou Women and Children's Medical Center, Guangzhou [11].

The size of the images given in the dataset was not uniform. Since comparison cannot be performed on different sizes, the images have been resized to 1024×1024 pixels. Fourteen Haralick's GLCM Features are extracted from the images with four different orientations (0, 45, 90, 135) and a distance of two pixels. The mean of four orientations is computed and used as one feature. Five classifiers namely, K-Nearest Neighbours (KNN), Support Vector Machine (SVM), Naive Bayes, Decision Tree, and Neural Network are used to classify the obtained features.

Seven evaluation metrics namely, train accuracy, test accuracy, precision, recall, specificity, F1 score, and AUC-ROC are computed to judge the performance of the classifiers. Running the above experiment, we got the results as shown in Table 3.

Table 3. Classification results for the base experiment.

	Neural Network	*KNN*	*SVM*	*Naïve bayes*	*Decision Tree*
Train accuracy	0.5805	0.8288	0.9988	0.7977	1.0
Test accuracy	0.5544	0.6555	0.6266	0.7676	0.7484
Precision	0.7478	0.6866	0.6260	0.7980	0.7484
Recall	0.4333	0.8256	1	0.8410	0.9
Specificity	0.7564	0.3718	0.0043	0.6453	0.4957
F1 score	0.5487	0.7497	0.77	0.8190	0.8172
AUC-ROC	0.6185	0.6787	0.6375	0.8248	0.6979

The results have been shown in the form of a bar graph (Fig. 2) for better understanding. Here, we can observe that the highest test accuracy is 76.76% for Naïve- Bayes classifier, highest precision is 79.8% for Naïve-Bayes classifier, highest recall is 100% for SVM, highest specificity is 75.6% for neural network, the highest F1 score is 81.89% for Naïve-Bayes and highest AUC-ROC is 82.48% for Naïve-Bayes classifiers.

3.2 Variants

We experimented with various variants of the above experiment and compared the results. Variants used are listed below:

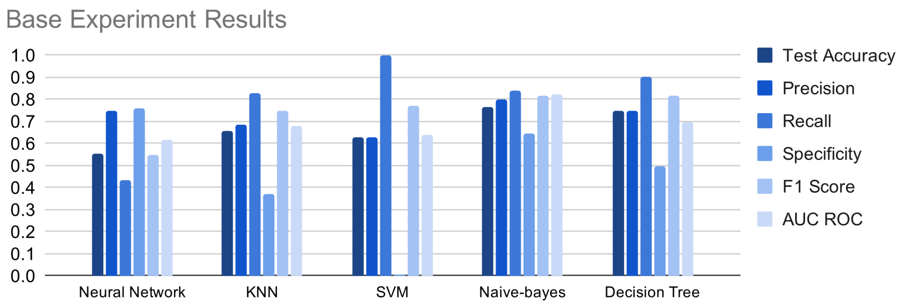

Fig. 2. Base experiment results (with 14 GLCM features).

Variant 1. Study the impact of standardization, Outlier removal, oversampling, and feature ranking. The sub-variants are as follows:

1.0. Fourteen GLCM features, distance = 2 (Base Experiment).
1.1. Standardise the features: StandardScaler is used to scale the feature vectors.
1.2. Standardise and remove the outliers: The outliers are removed using the DBSCAN technique.
1.3. Standardise and remove the outliers, oversample the minority class.
1.4. Standardise and remove the outliers, oversample the minority class, feature ranking: Selected the best features using the RFE feature ranking method.

The results of the above experiments are shown in Table 4. We observe here that standardisation of the features increases the test accuracy, F1 score, and AUC but precision, recall, and specificity go down. Removing the outliers improves the test accuracy and F1 score further, but precision, recall, and specificity are going down. However, oversampling the minority class or feature ranking is not improving any of the evaluation metrics. It remains almost the same as variant 1.3. Since the difference between normal and abnormal cases is not much, the dataset can be considered as a more or less balanced dataset. Hence, oversampling doesn't seem to improve the results. Further, the number of features is very less, hence feature ranking doesn't improve the result.

Table 4. Classification results for Variant 1.0 to Variant 1.4

	Variant 1.0 (Base)	*Variant 1.1*	*Variant 1.2*	*Variant 1.3*	*Variant 1.4*
Train accuracy	100%	100%	100%	100%	100%
Test accuracy	76.76%	77.56%	**78.52%**	78.52%	78.52%
Precision	**79.8%**	78.55%	77.23%	78.57%	78.57%
Recall	**100%**	94.6%	96.1%	94.1%	94.1%
Specificity	**75.64%**	61.9%	54.27%	58.9%	58.9
F1 score	81.89%	83.76%	**84.4**	84.0	84.0
AUC-ROC	82.4%	**85.89%**	82.9%	82.4%	81.2%

Variant 2. Study the impact of dividing into multiple regions. The sub-variants are as follows:

2.0. Fourteen GLCM features, normalised, outliers removed, d = 2 (Base Experiment).
2.1. Base experiment and Divide the image into 4 regions.
2.2. Base experiment and Divide the image into 16 regions.
2.3. Base experiment and divide the image into 64 regions.

The results for the above experiments are shown in Table 5.

Table 5. Classification results of Variants 2.0 to 2.3.

	Variant 2.0 (Base)	*Variant 2.1*	*Variant 2.2*	*Variant 2.3*
Train accuracy	100%	100	100	100
Test accuracy	78.52%	79.4	**87.6**	86.6
Precision	78.57%	79.6	**87.3**	86.5
Recall	94.1%	94.3	**97.1**	95.8
Specificity	58.9	61.9	**77.3**	75.9
F1 score	84.0	85.1	**90.4**	89.7
AUC	81.2%	83.9	**92.0**	90.4

From Table 5 we observe that dividing the image into 4 regions improves all the evaluation metrics. Dividing the image into 16 regions improves all the evaluation metrics drastically. Dividing into 64 regions yields better than dividing into 4 quadrants but not better than 16 regions. So, we conclude that dividing the image into smaller parts for feature extraction certainly improves the performance but after a certain limit, the performance starts deteriorating.

Variant 3. Study the impact of including first-order features. The sub-variants are as follows:

3.0. Only GLCM features.
3.1. GLCM features and first-order features.

The results of the above experiments are as illustrated in Table 6.

We observe from Table 6 that most of the evaluation metrics yield better results when first-order features are also extracted along with GLCM features (second-order features).

Table 6. Classification results of Variants 3.0 and 3.1

	Variant 3.0	*Variant 3.1*
Train accuracy	100%	**100**
Test accuracy	78.52%	**79.6**
Precision	78.57%	**79.9**
Recall	**94.1%**	91.7
Specificity	58.9	**63.6**
F1 score	84.0	**84.9**
AUC	81.2%	**84.3**

Variant 4. Study the impact of including first-order features and dividing the image into sub-images. The sub-variants are as follows:

4.0. GLCM and first-order features (Base Experiment).
4.1. Base experiment and divide the image into 4 regions.
4.2. Base experiment and divide the image into 16 regions.
4.3. Base experiment and divide the image into 64 regions.

The results of the above experiments are shown in Table 7.

Table 7. Classification results of Variants 4.0 to 4.3

	Variant 4.0	*Variant 4.1*	*Variant 4.2*	*Variant 4.3*
Train accuracy	100	100	**100**	100
Test accuracy	79.6	81.4	**88.4**	85.7
Precision	79.9	82.7	**89.7**	85.5
Recall	91.7	95.8	**96.1**	96.1
Specificity	63.6	69.2	**83.3**	73.9
F1 score	84.9	86.5	**90.0**	89.0
AUC	84.3	91.1	**92.4**	91.0

Dividing the image into 4 regions improves the performance of the classifiers on all evaluation metrics as shown in Table 7. Dividing into 16 regions improves the results further. However, dividing into 64 regions doesn't yield results better than 16 regions.

Variant 5. Study the effect of changing the pixel distance for the GLCM. The sub-variants are as follows:

5.0. 14 GLCM features, 5 first-order features and dividing the image into 16 regions with d = 2 (Base Experiment).

5.1. Change the distance d = 4.
5.2. Change the distance d = 8.
5.3. Change the distance d = 16.

The results of the above experiments are as shown in Table 8.

Table 8. Classification results of Variants 5.0 to 5.3

	Variant 5.0	*Variant 5.1*	*Variant 5.2*	*Variant 5.3*
Train accuracy	100	100	100	100
Test accuracy	88.4	**89.5**	88.4	87.2
Precision	**89.7**	89.7	87.8	87.7
Recall	**96.1**	94.3	94.61	94.35
Specificity	**83.3**	82.0	78.2	78.6
F1 score	90.0	**91.8**	91.1	90.1
AUC	92.4	**93.2**	93.2	92.6

It can be observed from Table 8, that considering more distance between the pixels to compute GLCM features doesn't improve the performance of the classifiers. Some metrics improved while others' performance degraded.

3.3 Inter-variant Comparative Analysis

Comparison of Variant 1 and Variant 2. Variant 1 consisted of the application of several preprocessing techniques after the extraction of GLCM features. In Variant 2, the x-ray images were divided into subregions and GLCM features were extracted for each region followed by the application of preprocessing techniques as used in Variant 1.

On studying the results obtained in both variants (Table 4 and Table 5) we observe that results majorly improved from Variant 1 to Variant 2 as shown in Fig. 3. The test accuracy changed from 78.52% to 87.6% (11.56% increase), precision from 79.8% to 87.3% (9.4% increase), recall from 100% to 97.1% (2.9% decrease), specificity from 75.64% to 77.3% (2.19% increase), F1-score from 84.4% to 90.4% (7.1% increase) and AUC-ROC from 85.89% to 92% (7.11% increase).

Comparison of Variant 1, 3, 4 and 5. Variant 3 uses first-order features along with the fourteen GLCM features extracted for the image. Variant 4 combines Variant 2 (dividing image into subregions) with Variant 3. In Variant 5, the best experiment settings from Variant 4 are selected and the distance parameter for the GLCM is varied. We compared the best results from all the four variants with respect to test accuracy, F1-score, and AUC-ROC metrics as shown in Fig. 4. From Variant 1 to Variant 3, the test accuracy slightly changed from 78.52% to 79.6% (1.37% increase), F1-score from 84.4% to 84.9% (0.6% increase) and AUC-ROC from 85.89% to 84.3% (1.85% decrease). From Variant 3 to Variant 4, the test accuracy increased from 79.6% to 88.4% (11.05% increase), F1-score from 84.9% to 90% (5.67% increase) and AUC-ROC from 84.3% to 92.4% (9.6%

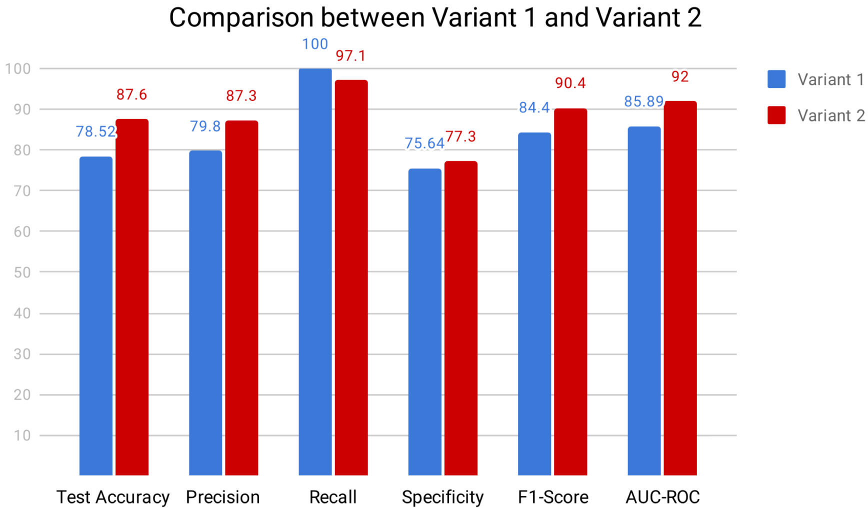

Fig. 3. Comparison of Variant 1 and Variant 2 to study the individual impact of first-order features and division of the image into regions.

increase). From Variant 4 to Variant 5, the test accuracy improved from 88.4% to 89.5% (1.24% increase), F1-score from 90% to 91.8% (2% increase) and AUC-ROC from 92.4% to 93.2% (0.86% increase). Overall, there was a 14% increase in test accuracy, 9.24% increase in F1-score and 8.5% increase in AUC-ROC from Variant 1 to Variant 5, highlighting the effectiveness of the experiments performed which can also be observed from the upward trend shown in Fig. 4.

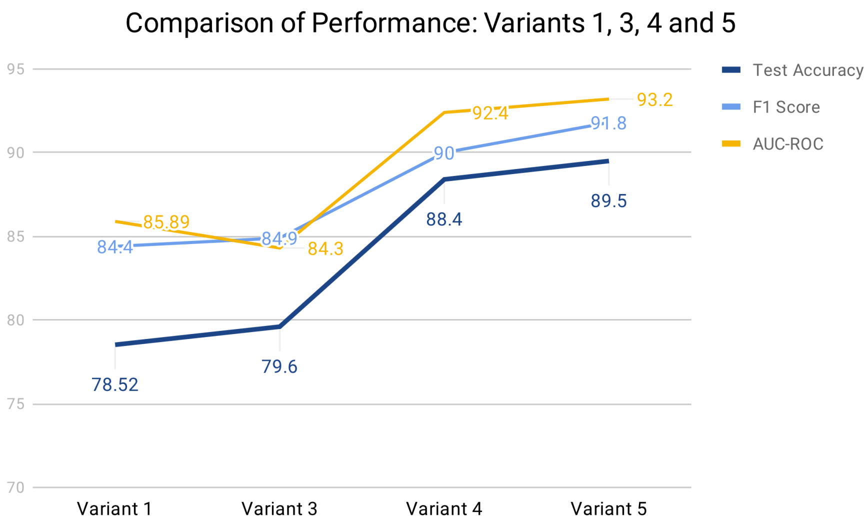

Fig. 4. Comparison of the best results from Variants 1, 3, 4, and 5 with respect to test accuracy, F1-score, and AUC-ROC.

4 Conclusion and Future Scope

The study was mainly focused on the use of the GLCM technique for extracting second-order statistical features (GLCM texture features) and the impact of variants on the performance of Machine Learning classifiers for the purpose of identifying Pneumonia patients from the Chest X-ray.

The base experiment showed the highest test accuracy of 76.76%. From variant 1 we concluded that the pre-processing techniques such as standardisation, outlier removal, oversampling, and feature ranking slightly improved the accuracy. The pre-processed images were then divided into smaller regions in variant 2 and then the GLCM features were extracted for each region which increased the accuracy. We then combined the 14 GLCM features with four first-order features in variant 3 and found that even without division of the image the inclusion of the first-order features improved our performance. Thus, we combined both the approaches of variant 2 and variant 3 in variant 4 and found that the combined results were even better. In order to increase the accuracy further, we observed the effect of changing the pixel distance in variant 5 and it can be concluded that increasing the pixel distance increased the test accuracy to some extent but after a certain limit the accuracy started deteriorating. The maximum accuracy achieved after applying all these variants was 89.5%.

Hence, GLCM based analysis of Chest X-ray images for the extraction of textural features provides good accuracy and can be used for the detection of Pneumonia disease.

References

1. 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). https://doi.org/10.1109/icimu.2011.6122757
2. Aggarwal, N., Agrawal, R.K.: First and second order statistics features for classification of magnetic resonance brain images. J. Sig. Inf. Process. **3**(2), 146–153 (2012). https://doi.org/10.4236/jsip.2012.32019
3. Ankita, R., Kumari, C.U., Mehdi, M.J., Tejashwini, N., Pavani, T.: Lung cancer image-feature extraction and classification using GLCM and SVM classifier. Int. J. Innov. Technol. Explor. Eng. **8**(11), 2211–2215 (2019). https://doi.org/10.35940/ijitee.K2044.0981119
4. 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). https://doi.org/10.1109/inista.2011.5946052
5. Chandra, T.B., Verma, K., Singh, B.K., Jain, D., Netam, S.S.: Automatic detection of tuberculosis related abnormalities in chest x-ray images using hierarchical feature extraction scheme. Expert Syst. Appl. **158**, 113514 (2020). https://doi.org/10.1016/j.eswa.2020.113514
6. Chatterjee, S., Dey, D., Munshi, S.: Integration of morphological preprocessing and fractal-based feature extraction with recursive feature elimination for skin lesion types classification. Comput. Methods Programs Biomed. **178**, 201–218 (2019). https://doi.org/10.1016/j.cmpb.2019.06.018
7. Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P.: SMOTE: synthetic minority over-sampling technique. J. Artif. Intell. Res. **16**, 321–357 (2002). https://doi.org/10.1613/jair.953

8. Giacinto, G., Roli, F.: Design of effective neural network ensembles for image classification purposes. Image Vis. Comput. **19**(9–10), 699–707 (2001). https://doi.org/10.1016/S0262-8856(01)00045-2
9. Gindi, A., Attiatalla, T.A., Sami, M.M.: A comparative study for comparing two feature extraction methods and two classifiers in classification of early stage lung cancer diagnosis of chest x-ray images. J. Am. Sci. **10**(6), 13–22 (2014). https://doi.org/10.7537/marsjas100614.03
10. Haralick, R.M., Shanmugam, K., Dinstein, I.H.: Textural features for image classification. IEEE Trans. Syst. Man Cybern. SMC **3**(6), 610–621 (1973). https://doi.org/10.1109/TSMC.1973.4309314
11. Kaggle, Chest X-Ray Images (Pneumonia) Dataset Page. https://www.kaggle.com/paultimothymooney/chest-xray-pneumonia. Accessed 13 Sept 2020
12. 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)
13. Kotsiantis, S.B., Kanellopoulos, D., Pintelas, P.E.: Data preprocessing for supervised learning. Int. J. Comput. Sci. **1**(2), 111–117 (2006)
14. Mall, P.K., Singh, P.K., Yadav, D.: GLCM based feature extraction and medical X-RAY image classification using machine learning techniques. In: 2019 IEEE Conference on Information and Communication Technology, pp. 1–6. IEEE, Allahabad (2019). https://doi.org/10.1109/cict48419.2019.9066263
15. NHLBI Website, Pneumonia. https://www.nhlbi.nih.gov/health-topics/pneumonia. Accessed 13 Sept 2020
16. Pugalenthi, R., Rajakumar, M.P., Ramya, J., Rajinikanth, V.: Evaluation and classification of the brain tumor MRI using machine learning technique. J. Control Eng. Appl. Inform. **21**(4), 12–21 (2019)
17. Rish, I.: An empirical study of the Naive Bayes classifier. In: IJCAI 2001 Workshop on Empirical Methods in Artificial Intelligence, pp. 41–46 (2001)
18. 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. Accessed 15 Sept 2020
19. World Health Organization (WHO), Pneumonia facts. https://www.who.int/news-room/fact-sheets/detail/pneumonia. Accessed 13 Sept 2020