Turkish Online Journal of Qualitative Inquiry (TOJQI)
Volume 12, Issue 6, June, 2021: 7565-7570

DATASET COLLECTION AND VALIDATION USING CORRELATION ANALYSIS

Ruchi, Jimmy Singla

Lovely Professional University, Phagwara, INDIA
jimmy.21733@lpu.co.in

Abstract –
Data collection is critical in the detection of lower spine diseases. Data collected could be in the form of text or image form. In the proposed work data collection is primarily dependent upon image dataset. MRI image dataset collected by visiting different laboratories in Punjab(Pathankot) and through online medium. Dataset verification is accomplished using correlation analysis. Collected dataset contains anomalies and must be eliminated. The pre-processing mechanism involving edge detection and contrast stretching applied and its result is also demonstrated through this proposed work. Comparison of result by applying techniques like histogram equivalence and slicing with contrast stretching yield best possible result of contrast stretching. Dataset contains 5000 samples collected from different labs out of which only 3032 shows positive correlation and were retained for demonstrating lower spine detection.

Keywords: data collection, correlation analysis, edge detection, contrasts stretching, histogram equivalence, slicing.

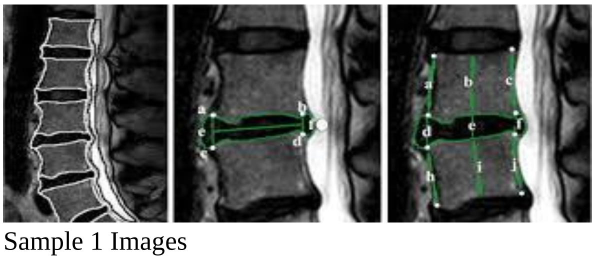
1. Introduction

Data collection is a primary phase associated with the development of expert system. This phase was performed by visiting different laboratories within Punjab (Pathankot) and from online valid sites like kaggle etc. . Authority for the collection of samples was taken from LPU (Lovely professional university, Punjab). All the images collected from the labs may not be usable. To detect the best possible images from the dataset that could yield high classification accuracy, correlation mechanism was applied.[1] Images were passed through the designed filter for feature extraction. Feature extraction was accomplished using PCA [2],[3]. Principal component analysis generates features that were stored within csv file. Correlation between the features and class was calculated. Images having positive correlation were retained and rest of the images yielding negative or 0 correlation was rejected.

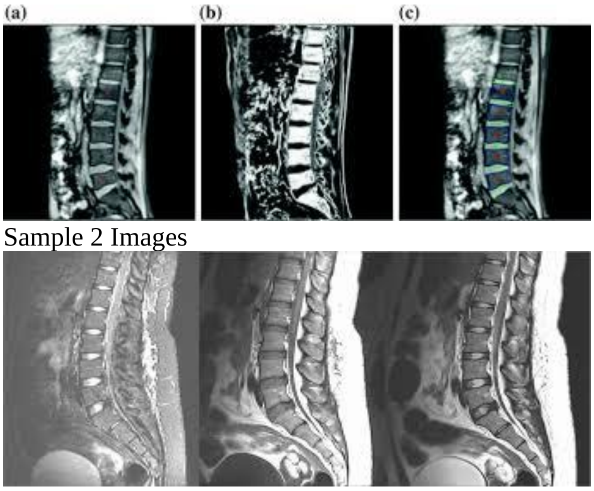
The contrast stretching mechanism was selected for pre-processing the image dataset [16] [17][18][19]. This was done due to best possible result originated from this mechanism as compared to histogram equivalence and slicing mechanism. Rest of the paper is organized as under. Section 2 gives the demonstration of collected dataset, section 3 gives the techniques used for pre-processing, section 4 gives PCA applied on dataset for feature extraction [20], section 5 gives the correlation analysis for retaining highest correlated images and last section gives the conclusion.

2. Collected Dataset

[5]Dataset collection was a physical in our approach. MRI dataset collected by visiting different labs within Punjab region. Total of 5000 images were collected. The size of dataset was sufficiently high to be used along with deep learning approaches. Figure 1 gives the sample of collected dataset.



Sample 1 Images



Sample 2 Images

Sample 3 Images

Figure 1: Sample Dataset

Sample dataset is partitioned into three different sets. In the first sample images MRI imaging is clear and pre-processing mechanism does not have much effect[3][21]. But still accuracy of disease detection could improve. Second sample indicates the clear dataset indicating sample was created with superior machine. These sample are least affected with pre-processing. Third sample requires pre-processing since it was least clear. Without pre-processing, classification accuracy will be hampered especially in the third case.

3. Pre-processing Techniques

Pre-processing mechanisms is compulsory to avoid any errors in the prediction of disease. Within expert system, minimum user interference is desired. To tackle the issue of errors and to design pure expert system with minimum user interference, pre-processing mechanisms are required[6],[7]. Different pre-processing mechanism along with produced results are given as under

- Histogram Equivalence

This mechanism is widely used for pre-processing images. Intensity distribution within the image is adjusted to enhance the contrast using this mechanism. The main objective of this mechanism is to provide linear distribution to the cumulative distribution function associated with images [8][9]. The probability distribution function is represented as under

$$cdf(x) = \sum_{K=-\infty}^{x} P(K)$$

Equation 1: Probability density function

Cdf indicates cumulative density function. P indicates the probabilities lies within the domain indicated with K. x indicates the total pixels present within the image. Following formula is implemented to get the contrast enhancement.

$$T_k = (L-1)Cdf(x)$$

Equation 2: Formula to get new probability distribution function

L indicates the discrete intensity levels within the range [0,L]. The result will be stored within T_k . The result from the operation of histogram equalization is given in figure 2.

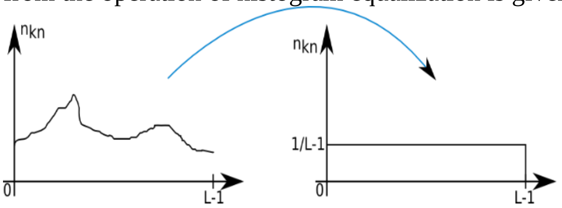


Figure 2: Histogram Equalization for contrast enhancement

As the histogram equalization is applied on the dataset, result obtained is given in figure 3.

Before Histogram Equalization	After Histogram Equalization

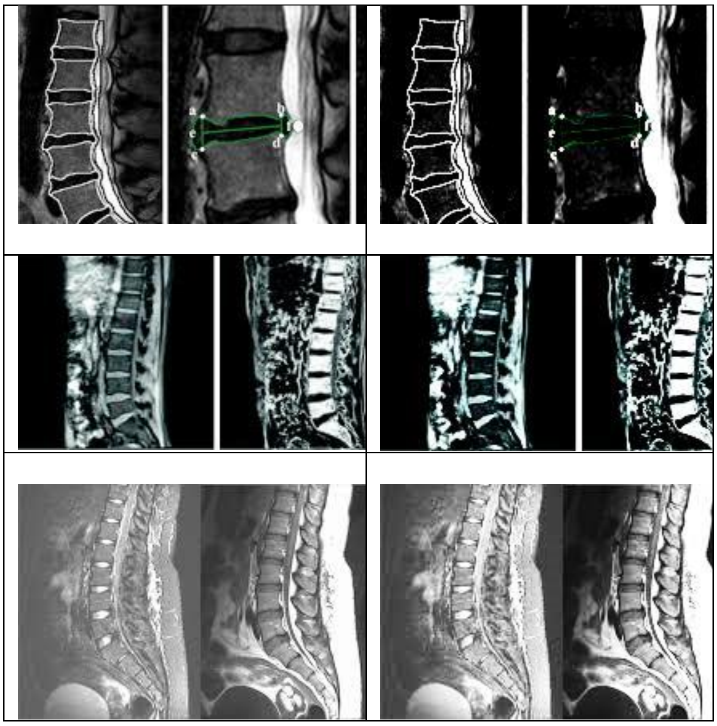


Figure 3: Contrast enhancement with histogram equalization

- Slicing

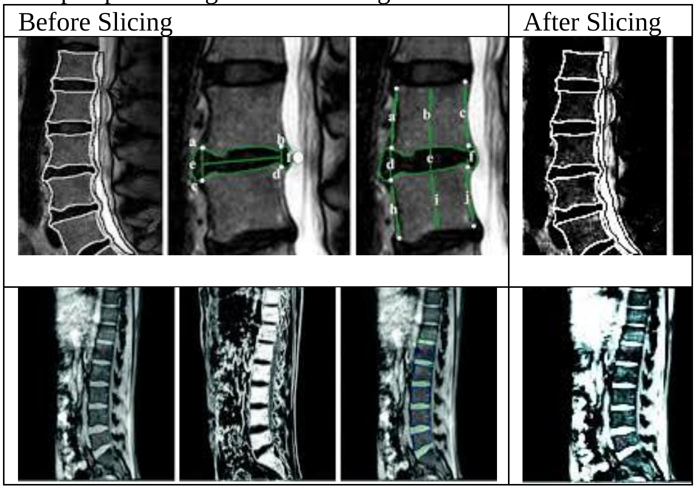
Slicing mechanism first of all crop the image and divide the image into critical and non-critical segments. Non critical segments are eliminated and critical segments are retained. Region of interest is compressed using this mechanism and hence pre-processing is performed much quickly as compared to other approaches[10]. Slicing is performed using the following equation

$$ROI = \sum_{R=1}^{n} K(R)$$

Equation 3: Region of interest equation through slicing

Here R is total number of regions obtained through slicing. K indicates the shrinked region from total area. All the regions extracted through the slicing-based mechanism remove the noise and enhance individual component of colour within the image. The result is better as compared to histogram equalization but size of the image is reduced and hence feature extraction phase could cause an issue.

Result obtained through this pre-processing mechanism is given as under:



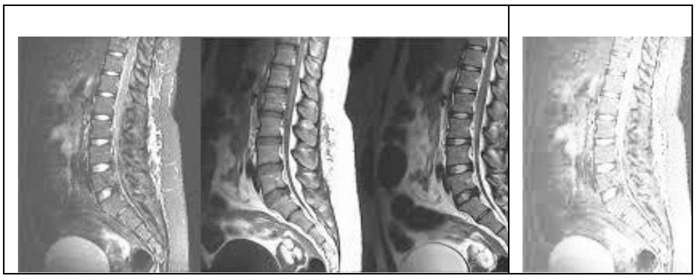


Figure 4: Pre-processing result for slicing

- Contrast Stretching

This is by far the best approach for pre-processing to simply enhance the image. The individual colour component is extracted and enhanced by the fixed margin[3]. Thus, contrast is uniformly distributed throughout the image[11][12][21] and the image intensity range is expanded. The Aim of contrast stretching is to process the images so that the dynamic range of the image will be very high to make the different details present in the image clearly visible. The contrast enhancement through this mechanism is given as under

$$\begin{gathered} R = R * 2 \\ G = G * 2 \\ B = B * 2 \end{gathered}$$

Equation 4: Contrast stretching equation

Here R,G and B indicates Red, Green and Blue components of an image. '2' is a threshold factor by which contrast is enhanced. This approach clearly produces better results as compared to other two approaches. The abnormalities are clearly identified using this approach.

The result obtained using the contrast stretching is given as under

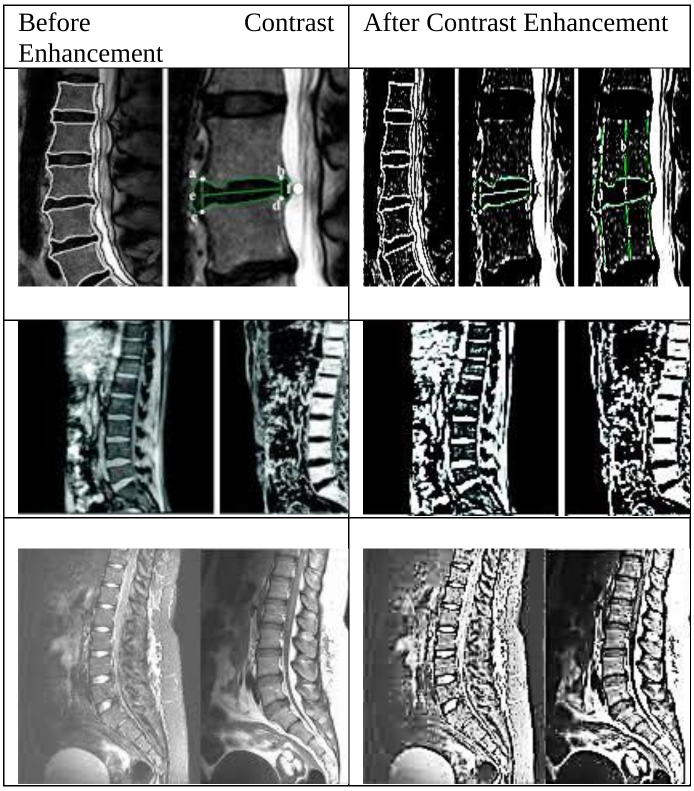


Figure 5: Contrast Enhancement

4. Feature Extraction

Feature extraction is critical phase that will be used to identify the images that are important for result generation. The feature extraction used in this approach is principal component analysis[13], [14]. PCA approach first of all identifies the critical features from the image and then these features are stored within csv file. This is effective approach in dimensionality reduction of dataset for quick segmentation and classification. Overall approach is given as under

PCA (Feature Extraction)

- Standardize the range of initial variables so that they contribute equally to the analysis
- Calculate covariance matrix

$$\begin{cases} Cov(x,x)\ Cov(x,y)\ Cov(x,z) \\ Cov(y,x) Cov(y,y), Cov(y,z) \\ Cov(z,x) Cov(z,y)\ Cov(z,z) \end{cases}$$

- Compute Eigen vector and Eigen values
- Retain the values having highest variance values. This means that highest relationship attributes are retained.

The result obtained using PCA with histogram equalization, slicing and contrast stretching is given in the following table

Table 1: Features extracted using PCA

Approach	Features Extracted
Histogram Equalization with PCA	52
Contrast Stretching	69
Slicing	41

Features extracted using contrast stretching is maximum and hence optimal. After extracting the features next step is to validate the features for selecting appropriate images for segmentation and classification.

5. Correlation Analysis

This is the final step in dataset validation. Correlation analysis is conducted and attributes contributing maximum to selected are retained[15]. Correlation analysis yielding positive values for images are retained and other images are rejected. Table 2 gives the correlation analysis result. Correlation analysis is conducted using following equation.

$$Correlation = \sum \frac{\left(x - mean(x)\right)\left(y - mean(y)\right)}{\sqrt{\sum(\left(x - mean(x)\right)^2 (y - mean(y)^2)}}$$

Equation 5: Correlation Analysis

Table 2: Correlation Analysis

Total Images-5000	
Number of images having Positive correlation	3032
Images having negative correlation	1968
Retained images	3032
Maximum Correlation	0.99
Minimum Correlation	-0.98

6. Conclusion

This paper presents mechanism to validate the dataset collected physically by visiting different labs. Total dataset images are close to 5000. The entire process of dataset acquisition is divided into phases. In the first phase, collected dataset passed through the pre-processing mechanism to enhance the contrast associated with the images. After enhancing the contrast, comparison of result in terms of contrast enhancement is made. Contrast stretching produced best possible result. After this step, second step is to extract features through PCA. With contrast stretching, PCA extract maximum possible features. Correlation analysis is the last phase that indicates the images that are to be retained for the final expert system. The proposed mechanism of dataset validation is effective and gives best possible images that can be used for training and testing an expert system for detecting lower spine diseases.

References

[1] Warne, K., Prasad, G., Siddique, N. H., & Maguire, L. P. (2004, October). Development of a hybrid PCA-ANFIS measurement system for monitoring product quality in the coating industry. In *2004 IEEE International Conference on Systems, Man and Cybernetics (IEEE Cat. No. 04CH37583)* (Vol. 4, pp. 3519-3524). IEEE.

[2] Ali, U. M. E., Hossain, M. A., & Islam, M. R. (2019, December). Analysis of PCA Based Feature Extraction Methods for Classification of Hyperspectral Image. In *2019 2nd International Conference on Innovation in Engineering and Technology (ICIET)* (pp. 1-6). IEEE.

[3] Pravin Kshirsagar et.al (2016), "Brain Tumor classification and Detection using Neural Network", DOI: 10.13140/RG.2.2.26169.72805.

[4] Versaci, M., Morabito, F. C., & Angiulli, G. (2017). Adaptive image contrast enhancement by computing distances into a 4-dimensional fuzzy unit hypercube. *IEEE Access*, *5*, 26922-26931.

[5] Al-Kafri, A. S., Sudirman, S., Hussain, A., Al-Jumeily, D., Natalia, F., Meidia, H., ... & Al-Jumaily, M. (2019). Boundary delineation of MRI images for lumbar spinal stenosis detection through semantic segmentation using deep neural networks. *IEEE Access*, *7*, 43487-43501.

[6] Mishra, Puneet, Alessandra Biancolillo, Jean Michel Roger, Federico Marini, and Douglas N. Rutledge. "New data preprocessing trends based on ensemble of multiple preprocessing techniques." *TrAC Trends in Analytical Chemistry* (2020): 116045.

[7] Ghosh, S., Malgireddy, M. R., Chaudhary, V., & Dhillon, G. (2012, May). A new approach to automatic disc localization in clinical lumbar MRI: combining machine learning with heuristics. In *2012 9th IEEE International Symposium on Biomedical Imaging (ISBI)* (pp. 114-117). IEEE.

[8] Pizer, S. M., Amburn, E. P., Austin, J. D., Cromartie, R., Geselowitz, A., Greer, T., ... & Zuiderveld, K. (1987). Adaptive histogram equalization and its variations. *Computer vision, graphics, and image processing*, *39*(3), 355-368.

[9] Rao, B. S. (2020). Dynamic Histogram Equalization for contrast enhancement for digital images. *Applied Soft Computing*, *89*, 106114.

[10] Hemanth, D. J., & Anitha, J. (2012, December). Image pre-processing and feature extraction techniques for magnetic resonance brain image analysis. In *International Conference on Future Generation Communication and Networking* (pp. 349-356). Springer, Berlin, Heidelberg.

[11] Perumal, S., & Velmurugan, T. (2018). Preprocessing by contrast enhancement techniques for medical images. *International Journal of Pure and Applied Mathematics*, *118*(18), 3681-3688.

[12] Anitha, S., & Radha, V. (2010). Comparison of image preprocessing techniques for textile texture images. *International Journal of Engineering Science and Tec hnology*, *2*(12), 7619-7625.

[13] Ma, J., & Yuan, Y. (2019). Dimension reduction of image deep feature using PCA. *Journal of Visual Communication and Image Representation*, *63*, 102578.

[14] Wang, X., & Paliwal, K. K. (2003). Feature extraction and dimensionality reduction algorithms and their applications in vowel recognition. *Pattern recognition*, *36*(10), 2429-2439.

[15] Senthilnathan, S. (2019). Usefulness of correlation analysis. *Available at SSRN 3416918*.

[16] H. Manoharan, et. al., "Examining the effect of aquaculture using sensor-based technology with machine learning algorithm", Aquaculture Research, 51 (11) (2020), pp. 4748-4758.

[17]. S. Sundaramurthy, S. C and P. Kshirsagar, "Prediction and Classification of Rheumatoid Arthritis using Ensemble Machine Learning Approaches," 2020 International Conference on Decision Aid Sciences and Application (DASA), 2020, pp. 17-21, doi: 10.1109/DASA51403.2020.9317253.

[18] P. R. Kshirsagar, H. Manoharan, F. Al-Turjman and K. Kumar, "DESIGN AND TESTING OF AUTOMATED SMOKE MONITORING SENSORS IN VEHICLES," in IEEE Sensors Journal, doi: 10.1109/JSEN.2020.3044604.

[19] P. R. Kshirsagar and S. G. Akojwar, "Prediction of neurological disorders using optimized neural network," 2016 International Conference on Signal Processing, Communication, Power and Embedded System (SCOPES), 2016, pp. 1695-1699, doi: 10.1109/SCOPES.2016.7955731.

[20] Pravin Kshirsagar et. al.. "Modelling of optimised neural network for classification and prediction of benchmark datasets, Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization, 8:4, 426- DOI: 10.1080/21681163.2019.1711457,2020

[21] PravinR Kshirsagar, Anil N Rakhonde, PranavChippalkatti, " MRI IMAGE BASED BRAIN TUMOR DETECTION USING MACHINE LEARNING", Test Engineering and Management, January-February 2020 ISSN: 0193-4120, Vol. 81, Page No. 3672 –3680.