Detection of Bone Abnormalities using Generalised Features

Nicholas Kastanos (nk569)
L248 Computer Vision
30 March 2021

***Abstract*—Musculoskeletal injuries and conditions are a common occurrence in emergency departments. One of the most common medical imaging tools, X-rays, are typically analysed by trained radiographers or specialists. However, with the rise of computer vision-based image classification techniques, classification models are being used to detect abnormalities in X-rays. Using SIFT features extracted from X-ray images, various machine learning models are developed and optimised. While some of the classifier models were unable to learn the features, a Multi-layer Perceptron neural network was used to develop a model which could detect musculoskeletal abnormalities. The classifier has an F1 score of 0.65, therefore its performance is low and would not be viable in a high-precision medical environment.**

I. Introduction

Bone fractures and other musculoskeletal injuries are one of the most prevalent emergency department visits [1]. The emergency staff must quickly and accurately complete a diagnosis to determine the severity and appropriate treatment for the patient. Misdiagnoses are costly as they can result in permanent discomfort for the patient.

Bone structure abnormalities can be detected in many ways [2]. The most common of these methods is to use radiograph, or X-ray, images. Classically, detection of bone abnormalities is completed manually by trained professionals, but with the rise of computer imaging and classification technologies, automated detection methods are being developed.

Many existing methods are either tailored to specific bone structures, or make use of deep convolutional neural networks (CNN) [4], [5]. These methods are limiting because they require specific conditions such as orientation and location consistency or are computationally expensive to create. By using classical computer vision techniques and features, a set of flexible features can be extracted from any X-ray to make a fast and accurate classification of musculoskeletal abnormalities. Each X-ray location will be a single classifier; however, the same process can be used to create different abnormality detectors.

A review of existing work in the field is conducted in Section II. A discussion of the image features extracted from the X-rays can be seen in Section III, and a discussion on the classification methods can be seen in Section IV. A critical discussion of the developed models is presented in Section V, and recommendations for future work can be found in Section VI. The code developed during the production of this report can be found at https://github.com/Nicholas-Kastanos/bone-fracture-detection.

II. Literature Review

A. Previous Work

Many existing methods of abnormality and fracture detection are specific to a specific bone structure. Afzal et al. [3] develop a model for automatic deformation detection in the elbow. This method locates the radius and ulna to profile the intensity along the length of the bone. The bones are detected by segmenting the background and soft tissue from the bone structures evident in the X-ray and detecting the capitellum and forearm bones using Canny edge detection, Hough circle detection, and line approximation. The profile of possibly deformed bones can be compared to that of a healthy bone, and a classification can be made. This method is highly effective, reaching accuracies greater than 80% on the MURA dataset [4], however the use of the proposed algorithm is highly restrictive since it can only be applied to elbows and the X-ray must be taken from a side-view.

Donnelley et al. [5] propose a similar method of fracture detection in long bones. The method identifies the long, straight bone segment known as the diaphysis, and detects large gradient changes along the length of the bone as fractures. Additionally, Donnelley et al. make use of the Affine Morphological Scale Space to smooth the image without losing information about the location of boundaries within the image [5], [6]. Similarly to Afzal et al. [3], the model developed by Donnelley et al. [5] is restricted to the type of bones which can be used as input. Additionally, the method only detects fractures in the diaphysis, and not in the bone joints.

Dimililer [7] takes a more generalised approach at bone abnormality classification by extracting Scale-Invariant Feature Transform (SIFT) [8] features of an X-ray image once it has been compressed using the Haar Wavelet transform. These features are used as input to a Back-propagation Neural Network machine learning classifier. This approach, should the model be trained on other bone structures, is able to detect bone abnormalities in other locations. This makes the model flexible in its deployment.

Deep-learning methods are also used to detect bone abnormalities. Rajpurkar et al. [4] developed an ensemble of 5 169-layer DenseNet CNNs which can detect anomalies in bone structure. This classifier was trained on the MURA dataset [4], and was shown to have similar performance to radiologists for finger, wrist, and hand X-rays.

B. MURA Dataset

The MURA dataset [4] is a large upper extremity X-ray dataset. The dataset consists of X-rays belonging to one of seven upper extremity radiographic study types: elbow, finger, forearm, hand, humerus, shoulder, and wrist. The multi-image labelled studies consist of X-rays from the same patient of the same study type. These studies were labelled manually at the time of clinical radiographic interpretation. This method does not provide guarantees of the label correctness as it may have been labelled incorrectly by the initial radiologist. Additionally, the dimensions, position and orientation of the subject within the X-ray, or the X-ray within the image, cannot be guaranteed.

Each study is labelled either as normal or abnormal. Abnormalities include but are not limited to fractures, hardware, degenerative join diseases, and lesions.

The dataset contains publicly available training and validation splits, and the test split is kept secret by the dataset creators to be used as scoring for their challenge. Thus, it will not be used in the evaluation of the developed models.

III. Feature Extraction

A single study in the MURA dataset contains a variable number of images. This property is incompatible with many standard classification methods as they require a constant-size input. To create a compatible feature vector, features must be extracted from each image before they can be combined into a single zero-padded vector.

A. SIFT Features

The Scale-Invariant Feature Transform (SIFT) [8] algorithm detects multi-scale keypoint descriptors, however extracting the features from the raw X-ray image captures noise data. Therefore, before the SIFT descriptors can be extracted, each image must be processed. An example X-ray can be seen in Figure 1, which will be used to illustrate further processing stages.

Fig. 1: Sample X-ray image from the MURA dataset.

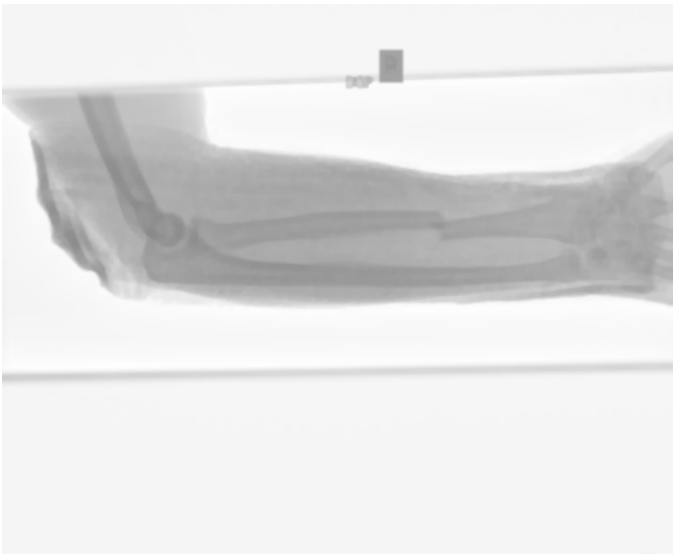


Fig. 2: Sample image after the noise-reduction algorithms have been applied.

Initially, the greyscale white-on-black X-ray images are inverted to facilitate dark-edge detection. Under visual inspection, X-ray images capture the texture of the underlying bone which results in grainy images. The bone texture does not convey information about structural deformations. Therefore, the image is filtered for noise by blurring using a (3×3) Gaussian kernel. Further image filtering is completed using greyscale morphology by opening and closing operations using the same (3×3) kernel. The results of the noise-reduction process can be seen in Figure 2.

X-ray images contain identifying tags to indicate information about the subject matter, however it does not convey information about the bone structure. Therefore, removal of

Fig. 3: Foreground of the sample image.

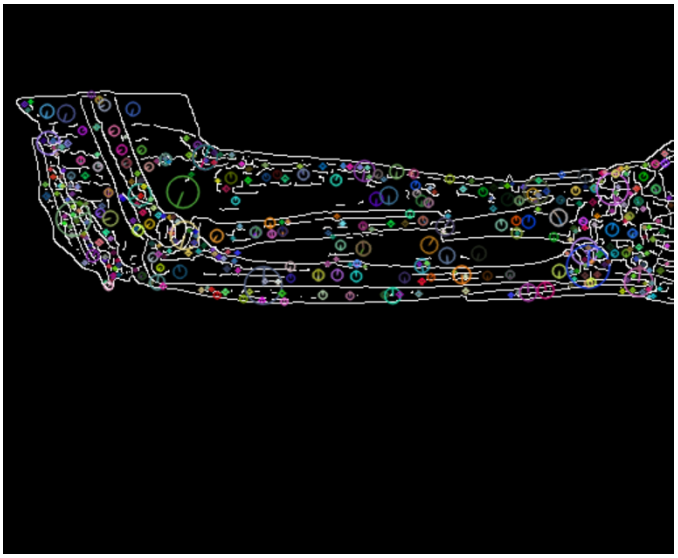


Fig. 4: Image after edge-detection and SIFT feature extraction.

this 'background' area is completed using Otsu's Binarization method [9], followed by contour detection of the largest area continuous contour. This method is effective on X-ray images since the background is distinctly different to the foreground, creating an easily detectable trough in the spectral frequency histogram. An example of the foreground image can be seen in Figure 3

Canny edge detection is completed on the foreground image. This creates an image which highlights any edges in the bone structure. The edges of the image show the outline of the image which reveals dislocations and deformations of the bones. Additionally, fractures within the bone are highlighted by the edge detection algorithm, allowing the feature extraction algorithm to easily focus on these areas.

Finally, n_{SIFT} top SIFT feature descriptors are extracted from the processed image, ranked by the local contrast scores. If the algorithm is unable to detect n_{SIFT} features, the feature vector is padded with zeros to maintain a constant length. An example of the edges image with SIFT features can be seen in Figure 4.

B. Principal Component Analysis

For each image, there are $n_{SIFT} \times 128$ descriptors. Therefore, Principal Component Analysis (PCA) is used to reduce the dimensionality of the feature vector. The PCA transformation is determined using the dataset training split, and the top n_{PCA} components ordered by explained variance. The same PCA transform is then applied to the validation data. This allows the most significant orthogonal components of the dataset to be used during training.

The PCA is completed per image, not per study. This allows the zero-padding introduced to maintain vector length to be ignored during this process.

C. Feature Vector Construction

Once the features have been extracted and the dimensions have been reduced, the image vectors must be combined into a single constant length feature vector for training.

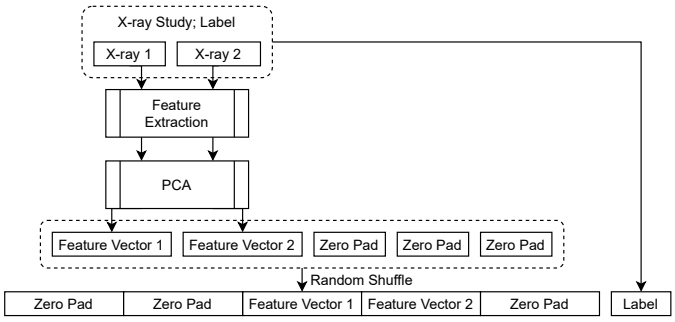


Fig. 5: Processing pipeline for studies.

This is accomplished by concatenating the image vectors and padding the final vector with zeros. However, this process can cause issues with training. Since the maximum number of images in a study is much larger than the mean number of images, the image vectors at the end of the feature vector are frequently zero. Because of this, the classifiers will quickly learn that the inputs at the end of the feature vector have low output variance. Therefore, the position of the image vectors must be shuffled within the feature vector, while maintaining the internal consistency of the image vectors. Figure 5 shows an illustration of the complete feature vector construction process.

IV. Classification

Once the features have been extracted, a model must be developed to complete the classification. Each type of model creates different decision boundaries and learns in different methods resulting if varying strengths and weaknesses. Therefore, the optimal classifier type must be discovered. Three different classifiers are developed and compared to determine which performs the best, and finally a hard-voting classifier is developed using one type of each model.

A. Classification Models

The three individual models under investigation are a Support Vector Machine (SVM), Multi-layer Perceptron (MLP) neural network, and a Random Forest (RF) classifier.

SVMs using the radial basis function kernel are statistical classifiers, which are highly effective when used in higher-dimensional spaces. For this model, the optimal length-scale parameter C must be discovered experimentally.

MLPs are a back-propagation neural network which consist of sequential layers of fully connected neurons. Most neural networks face the issue that large amounts of training data are required to train a well-generalised network, and this problem increases with the size of the network. As a minimum, a MLP must have two layers: the input and the output. Any layers which are not part of this set are called hidden layers, and the number of these layers n_{HL} is the hyper-parameter of interest. The number of neurons per layer is a linearly decreasing range between the feature vector length and the number of output classes, exclusive. This creates a gradually tapering network

architecture. To train the networks, each MLP is trained using the Adam Optimiser for a maximum of 1000 epochs with an early stopping policy. L2 regularisation, ReLU activations, and an initial learning rate of 0.001 are also used. The data is shuffled each iteration and is batched during training.

RF classifiers are an ensemble of randomised decision trees. Each of the trees are trained on slightly different datasets, to increase diversity in the individual trees. The number of tree estimators n_{RF} can be varied to gain consensus from a multitude of different classifiers, increasing the generalisation accuracy of the classifier.

Voting classifiers combine conceptually different learning models into a single classifier to leverage the benefits of each model and balance out the weaknesses. Each of the individual models are trained and make predictions independently. The output of each classifier is counted as a vote towards the combined classification, and the class with the most votes wins.

Each of the discussed classifiers can be sensitive to input scaling. Therefore, each feature is whitened by subtracting the mean and dividing by the standard deviation. This ensures that large-valued features do not dominate the training process.

B. Parameter Search

In order to find the optimal features n_{SIFT} and n_{PCA}, as well as the hyper-parameters for the classifiers, a grid search is performed. From the grid, the parameters used to train the classifier with the highest mean F1 score over 10 iterations on the validation dataset is selected as the optimal parameters. F1 score is preferred over accuracy as the dataset is unbalanced and accuracies greater than 50% can be achieved by always predicting a single class.

The grid search is first performed using a large range and step size. Once the parameters are selected, the further refined optimal parameters are selected by completing smaller-range grid search centered on the previous parameters. This method has a smaller computation requirement than a more exhaustive search at the cost of possibly missing a more optimal configuration.

The range of values explored for each parameter can be seen in Table I.

TABLE I: Bounds for the parameter grid search.

Parameter	Minimum	Maximum	Minimum Step
n_{SIFT}	50	2000	50
n_{PCA}	10	100	10
C	2^0	2^{10}	$2^{0.1}$
n_{HL}	0	5	1
n_{RF}	10	100	1

V. Results and Analysis

The optimal parameters are discovered, and can be seen in Table II along with the performance metrics calculated using the validation dataset. Only X-rays containing images of forearms are used in developing and testing the models. This is because each X-ray location would have different optimal parameters, and using a single X-ray location would provide sufficient proof of concept.

TABLE II: Validation scores of the best performing models

Model	n_{SIFT}	n_{PCA}	C	n_{HL}	n_{RF}	F1	Accuracy
SVM	1550	10	128	–	–	**0.58**	62.4%
MLP	1400	20	–	0	–	**0.65**	64.6%
RF	700	10	–	–	77	**0.56**	68.4%
Voting	1150	10	7.46	0	69	**0.53**	65.4%

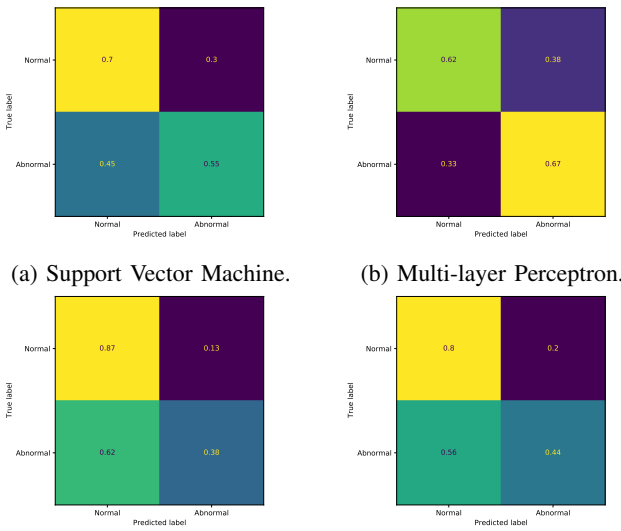
The classifier which produced the highest F1 score is the MLP. Additionally, the MLP architecture which produced the best results has zero hidden layers. This may be attributed to the fact that the dataset is very small compared to typical neural network architectures, and a smaller network would not over-fit the training data.

It can be seen that the number of SIFT keypoints n_{SIFT} has an optimal performance lower than the maximum value. This saturation threshold indicates that the transform cannot gain any more useful information when using additional features. Additionally, the low n_{PCA} values show that the features are highly correlated and there are very few orthogonal components which carry useful information.

Figure 6 shows the confusion matrices for each of the classifiers. It is evident that the RF and Voting classifiers have high accuracies because they are only predicting a single class. This shows that the classifiers have failed to learn how to detect bone abnormalities from the provided features. The SVM and MLP classifiers have been able to learn from the features, however they do not produce reliable results.

The noise reduction processing stage may be removing useful information for the classification small-scale conditions such as degenerative joint diseases, however this cannot be confirmed since the diagnosis of the study is not provided with the MURA dataset.

As a general statement, this method does not produce



(a) Support Vector Machine. (b) Multi-layer Perceptron.

(c) Random Forest. (d) Hard-voting.

Fig. 6: Confusion matrices for each of the models. Each element is normalised by the number of true samples per class.

reliable results. Medical imaging requires high precision implements, and the accuracy of the developed classifiers is not sufficient.

VI. Future Work

More exploration on this topic is necessary to conclude whether a general set of features can be developed for X-ray classification.

Further optimisation of the processing stage can be completed. Canny edge detection has inbuilt thresholds, as well as kernel sizes used for blurring and morphology can be optimised. Additionally, the background removal may not be effective. The MURA dataset provides no guarantees that the image only contains the X-ray contents, and visual inspection shows that some samples contain two-colour backgrounds. Additionally, the dataset contains studies where multiple X-rays are included in a single image. These images would cause issues with the algorithm, and should be removed.

Further features can be explored, such as Oriented FAST and Rotated BRIEF (ORB) features. ORB features are considered to be a faster and as effective alternative to SIFT. Additionally, using Hough and Fourier transforms can yield additional information.

Additionally, evaluation of the models trained on other X-ray types must be conducted to determine whether the methods discussed in this paper are effective on other bone structures.

VII. Conclusion

The features and models developed to detect abnormalities in X-rays of forearms did not produce accuracies comparable to existing methods. The most effective method of abnormality detection through SIFT features makes use of a MLP classifier with no hidden layers, trained on 1400 SIFT features and a 20 component PCA dimension reduction. This model achieves a mean F1 score of 0.65. Further exploration of feature extraction methods is required.

REFERENCES

[1] N. Kozaci, M. O. Ay, M. Akcimen, G. Turhan, I. Sasmaz, S. Turhan, and A. Celik, "Evaluation of the effectiveness of bedside point-of-care ultrasound in the diagnosis and management of distal radius fractures," *The American Journal of Emergency Medicine*, vol. 33, no. 1, pp. 67–71, 2015.

[2] MedlinePlus Medical Encyclopedia. (2021, Mar.) Imaging and radiology. [Online]. Available: https://medlineplus.gov/ency/article/007451.htm

[3] M. Afzal, M. Moazzam, R. Badar, and S. Narejo, "Automatic detection of elbow abnormalities in x-ray imagery," *International Journal of Advanced Computer Science and Applications*, vol. 11, Jan. 2020.

[4] P. Rajpurkar, J. Irvin, A. Bagul, D. Ding, T. Duan, H. Mehta, B. Yang, K. Zhu, D. Laird, R. L. Ball *et al.*, "Mura: Large dataset for abnormality detection in musculoskeletal radiographs," *arXiv preprint arXiv:1712.06957*, 2017.

[5] M. Donnelley and G. Knowles, "Automated bone fracture detection," in *Medical Imaging 2005: Image Processing*, J. M. Fitzpatrick and J. M. Reinhardt, Eds., vol. 5747, International Society for Optics and Photonics. SPIE, 2005, pp. 955 – 966. [Online]. Available: https://doi.org/10.1117/12.594449

[6] G. Koepfler and L. Moisan, "Geometric multiscale representation of numerical images," in *Scale-Space Theories in Computer Vision*, M. Nielsen, P. Johansen, O. F. Olsen, and J. Weickert, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 1999, pp. 339–350.

[7] K. Dimililer, "Ibfds: Intelligent bone fracture detection system," *Procedia Computer Science*, vol. 120, pp. 260–267, 2017, 9th International Conference on Theory and Application of Soft Computing, Computing with Words and Perception, ICSCCW 2017, 22-23 August 2017, Budapest, Hungary. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1877050917324493

[8] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International journal of computer vision*, vol. 60, no. 2, pp. 91–110, 2004.

[9] N. Otsu, "A threshold selection method from gray-level histograms," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 9, no. 1, pp. 62–66, 1979.