



A survey of fracture detection techniques in bone X-ray images

Deepa Joshi¹ · Thipendra P. Singh¹

Published online: 1 January 2020
© Springer Nature B.V. 2020

Abstract

Radiologists interpret X-ray samples by visually inspecting them to diagnose the presence of fractures in various bones. Interpretation of radiographs is a time-consuming and intense process involving manual examination of fractures. In addition, clinician's shortage in medically under-resourced areas, unavailability of expert radiologists in busy clinical settings or fatigue caused due to demanding workloads could lead to false detection rate and poor recovery of the fractures. A comprehensive study is imparted here covering fracture diagnosis with the aim to assist investigators in developing models that automatically detect fracture in human bones. The paper is presented in five folds. Firstly, we discuss data preparation stage. Second, we present various image-processing techniques used for fracture detection. Third, we analyze conventional and deep learning based techniques for diagnosing bone fractures. Fourth, we make comparative analysis of existing techniques. Fifth, we discuss different issues and challenges faced by researchers while dealing with fracture detection.

Keywords X-rays · Radiology · Fractures · Deep learning · Convolutional networks

1 Introduction

A fracture is a partial or complete break in the bone (Mayne 2013). High impact or force against a bone that it can structurally withstand is the substantial cause of fracture. Traumatic and stress are commonly found bone fractures in human body. Traumatic fracture is a result of automobile accident, serious fall, or intentional causes such as physical abuse whereas stress fractures are associated with repetitive load-carrying pressure to a healthy bone, common among athletes (e.g., gymnasts, dancers, long-distance runners) and military personnel. However, fracture can also occur due to several other reasons like osteoporosis, (a medical condition that weakens the bone), cancers, or a brittle bone disease known as osteogenesis imperfecta. According to the World Health Organization (WHO) report, 1.66

✉ Deepa Joshi
deepa.joshi117@gmail.com

Thipendra P. Singh
tpsingh@ddn.upes.ac.in

¹ Department of Computer Science, UPES, Dehradun, Uttarakhand, India

million people suffer from hip fracture every year throughout the world and the rate is expected to rise by three to four times by year 2050 because of the worldwide increase in the number of older people (World Health Organization 2019). The rate of fracture occurrences are higher by three to four fold among women in countries with a high fracture incidence rate, and is equal among men and women in counties with low fracture prevalence (World Health Organization 2019). All bone fractures are divided into six major categories (Mayne 2013) that is depicted using Fig. 1.

- A. Transverse Fracture: It is the simplest type of fracture where bone is break as a horizontal line.
- B. Oblique Fracture: It is a fracture type where the break is extending in a slanting direction, caused by indirect or rotational force.
- C. Spiral Fracture: It is a fracture type where the break spirals around the bone, common in a twisting injury.
- D. Comminuted Fracture: It is a fracture type where bone breaks into several pieces.
- E. Greenstick Fracture: It is incomplete fracture type where the broken bone is not completely separated.
- F. Impacted Fracture: It is a fracture type where bone breaks but the two ends of fractured bone are forced together. This produces a rather stable fracture that can heal readily but at the cost of some length lost.

X-rays, computed tomography (CT), magnetic resonance imaging (MRI) are various medical imaging modalities used for capturing images of the affected body area. A radiologist expert perform interpretation of these captured images to diagnose the disease and further recommends needed treatment. X-ray is the oldest, fastest, and most frequently used imaging modality, which examines suspected fractures by imaging internal organs of the body (<https://orthoinfo.aaos.org/en/treatment/x-rays-ct-scans-and-mris>). It has become a prime analytic instrument to check patients for fractures, majorly due to its wide-availability across areas where many expensive imaging modalities might not be available. Radiologists or clinicians interpret X-ray samples by visually inspecting them to identify the presence and type of fractures in various bones. The need for advanced level of imaging tools such as MRI and CT scan emerges to obtain more detailed, cross-sectional view of the bone, which might be missed during X-ray examination (<https://orthoinfo.aaos.org/en/treatment/x-rays-ct-scans-and-mris>). Interpretation of radiographic images is a time-consuming and intense process that involves manual examination and classification of fractures.

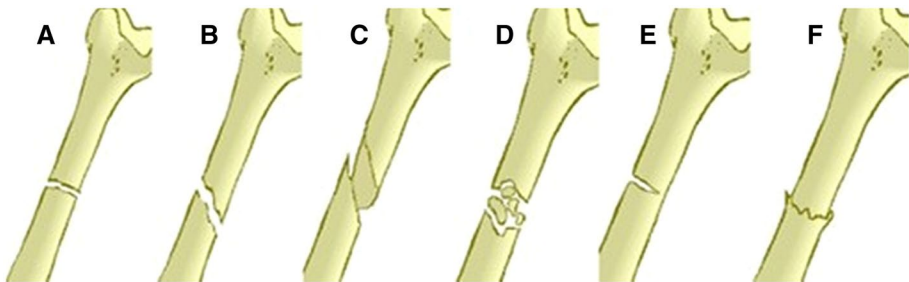


Fig. 1 Major fracture types found in bones (Mayne 2013). **a** Transverse, **b** Oblique, **c** Spiral, **d** Comminuted, **e** Greenstick, and **f** Impacted fracture

Shortage of clinicians in medically under resourced areas, unavailability of expert radiologists in busy clinical settings and fatigue caused due to demanding workloads could lead to false detection rate and poor recovery of fractures. In addition, X-ray image interpretation is mostly performed under the supervision of only one examiner, which increases the risk of inaccurate fracture identification due to unavailability of a second opinion. 41–80% of cases reported error in diagnosing fractures due to inaccurate fracture identification (Guly 2001; Whang et al. 2013). Various studies (Krupinski et al. 2010; Waite et al. 2017; Stec et al. 2018) show how fatigue caused to the examiner because of interpreting several musculoskeletal images could lead to significantly reduced performance in detecting abnormalities. Computer Vision systems could be a potential solution to such problems if it can quickly provide a trustworthy second opinion in identifying suspicious fracture cases.

Traditionally, various low-level pixel-processing techniques such as noise reduction, segmentation and feature extraction were effectively utilized for predicting human bone fractures. Obtaining region of interests by segmenting bone regions from fleshy areas was quite popular before extracting features from the image. The first work on automatic fracture detection used Neck-Shaft Angle (NSA) as the only feature for fracture detection (Tian et al. 2003). However, local disruptions remain undetected using this technique. A new approach is proposed based on texture analysis of trabecular pattern to detect such minor disturbances (Yap et al. 2004; Lim et al. 2004). Texture features such as Gabor Orientation (GO), Intensity Gradient Direction (IGD) and Markov Random Field (MRF) are extracted to detect femur fractures in bone X-ray images (Lim et al. 2004; Lai et al. 2005; He et al. 2007). Hand bone fractures from input X-ray image is detected by locating object boundaries using sobel algorithm followed by extracting texture features using Gray Level Co-occurrence Matrix (GLCM) (Al-Ayyoub et al. 2013). GLCM (entropy, contrast, correlation and homogeneity) is used as the only feature for identifying femur fractures in bone X-rays (Chai et al. 2011). The primary tool to improve classification accuracy involves amalgamation of texture features including GLCM, GO, MRF, IGD and shape features in femur X-rays (Umadevi and Geethalakshmi 2012). Automatic crack detection in X-ray images is complicated and arduous task that demands accuracy and swiftness (Linda and Jiji 2011, 2017). Various fuzzy based approaches are investigated to identify infected sites of a crack by segmenting bone areas using fuzzy index measure (Linda and Jiji 2011). Likewise, hairline fractures are estimated by means of intensity variation concept in multiple bones by segmenting bone areas using Expectation Maximization (EM) algorithm (Zhang et al. 2001).

It is observed that the performance of the classifier is significantly improved with the introduction of multiple-classifier based systems where individual results from base classifiers are fused together (Umadevi and Geethalakshmi 2012; Yap et al. 2004; Lai et al. 2005; Mahendran and Baboo 2011b). The most prominent ensemble based models have used Neural Network (NN), Support Vector Machine (SVM), Naïve Bayes (NB) algorithms for diagnosing fractures in human bones from the year 2003 to 2015. Researchers have developed a new Stacked Random Forests-Feature Fusion (SRF-FF) technique to identify fractures in various regions like hand, foot, ankle, knee, lower leg and arm (Cao et al. 2015). Divide and conquer approach utilizes techniques that combine hierarchical SVM and texture features including GO, MRF and IGD to predict fractures (He et al. 2007). Localizing fractures in bone X-rays is significant in diagnosis, decisions and treatment planning of the patient. GO, Schmid texture feature, and proposed Contextual-Intensity (CI) features are effectively utilized to localize fractures in multiple bones (Cao et al. 2015). Distinct bones such as humerus, radius, ulna, femur, tibia, and fibula are considered as long bone, where the fracture is classified based on its type and the location where it occurred

(Bandyopadhyay et al. 2016a). Based on the intensity of injury, a long bone fracture is categorized into simple and complex types while the location of fracture is recognized through fracture points identified from X-ray images (Bandyopadhyay et al. 2016c). Decision Tree (DT) and K-Nearest Neighbor (KNN) classifiers are used for fracture detection and classification respectively once break points of the leg bone is identified in the processed image (Myint et al. 2018). Various literatures (Bandyopadhyay et al. 2013, 2016a; b; c) have suggested approaches based on digital geometry, which provide a powerful tool for analyzing bone fractures from X-ray images. The digital geometry based techniques such as relaxed straightness and concavity index are effectively utilized for long bone segmentation, rectification of contour imperfections, fracture detection and fracture localization (Bandyopadhyay et al. 2016b). Various approaches like image segmentation, machine learning or deep neural networks have focused majorly on fracture detection techniques but complex fracture, which consists of multiple patterns like transverse, oblique, spiral, and comminuted etc. are either under investigated or difficult to predict. Fractures can vary from minor to complex types. The complex types are severe, which needs to be treated within a time-period to avoid complications. Complex fractures in long bones are identified, localized and visualized using linear structuring elements and Ray Cascade method in CT DICOM images (Linda and Jiji 2018). Hidden Markov Random Field– Expectation Maximization (HMRF-EM) (Zhang et al. 2001) and adaptive thresholding techniques (Singh et al. 2012) for effectively utilized for bone segmentation (Linda and Jiji 2018).

Several features such as textual, shape, edges, horizontal, vertical lines etc. were extracted from the image and supplied to the classifier for predicting the occurrence of fractures. However, this approach is now prevailed by various deep learning methods. A fastest-growing field in artificial intelligence is deep learning, which has gained enormous success in medical imaging by providing better accuracy as compared to traditional approaches (Gale et al. 2017; Lindsey et al. 2018; Chung et al. 2018; Rajpurkar et al. 2017; Kim and MacKinnon 2018). Generating large datasets by collecting and labeling radiographs from expert Radiologists is a challenging and laborious responsibility. The number of radiographs used for diagnosing fractures is increased from 500 to hundreds of thousands from year 2003 to 2018 by collecting more data from hospitals and through data augmentation. Data augmentation is extensively used to amplify datasets during training the deep learning model (Gale et al. 2017; Lindsey et al. 2018; Chung et al. 2018; Rajpurkar et al. 2017; Kim and MacKinnon 2018). A variant of deep learning architecture is Convolutional Neural Networks (CNN) that consists of convolution layer, sub-sampling layer, and fully connected layers. Feature extraction techniques took a tremendous turn when CNN model came into existence which can “learn the features” instead of handcrafting them into the system. Convolution and sub-sampling layers of CNN are part of feature learning process while fully connected layer is used for classification. The capability of the model to learn features on its own rather than handcrafting them into the system has made these architectures a great success (Lakhani and Sundaram 2017; Russakovsky et al. 2015). A 172 layers Deep Convolutional Neural Networks (DCNN) is trained to classify fracture into healthy and fractured classes in pelvis radiographs (Gale et al. 2017). Neer classification (Neer 1970) is used to classify proximal humerus fracture into four categories; greater tuberosity, surgical neck, 3-part, and 4-part by using a pre-trained ResNet-152 classifier (Chung et al. 2018). The performance of the classifiers is compared against evaluations of expert radiologist on similar test images for analyzing the practical adaptability of the model. Transfer learning is effectively utilized by fine-tuning a pre-trained model, where scarcity of resources is an obstacle for creating a successful model from scratch (Lindsey et al. 2018; Kim and MacKinnon 2018). The model is pre-trained on non-medical images (such as ImageNet) and is fine-tuned using Google’s Inception v3 network for diagnosing wrist fractures (Kim and

MacKinnon 2018). However, a model is pre-trained on 100,855 bone images of several other body parts and is fine-tuned on a DCNN for detecting wrist fractures (Lindsey et al. 2018).

Computer-aided design (CAD) systems can assist medical experts by suggesting the type of treatment required for disease diagnosis thereby upgrading patient care. Fracture diagnosis system could facilitate the examination process by providing optimal results upon integration with X-ray machine software. Various reviews (Khatik 2017; Jacob and Wyawahare 2013; Kinnari and Dangar 2017; Mahendran and Baboo 2011a) focusing on fracture diagnosis in plain radiographs exists but a detailed study of traditional as well as deep learning approaches to fracture detection and classification is missing. A comprehensive study is imparted here covering fracture diagnosis with the aim to assist researchers in developing models that automatically detects fracture in several bones of human body.

2 Data preparation

The process of fracture detection and classification is divided into multiple stages. It begins with data collection, which involves collection of datasets from various hospitals or public domain followed by dataset labeling. Dataset must be prepared before feeding it into a classifier for the prediction of fracture occurrence and its corresponding class. Researchers and radiologists are putting many efforts in collecting and labeling radiographic images for research purpose. However, unavailability of the freely available standard dataset is one of the major drawbacks in comparative analysis of the existing systems due to which researchers have shown performance of the proposed model in their private datasets. After dataset collection, next major issue is the labeling process, which is an imperative stage of data pre-processing. Labeling requires annotation of radiographs by experienced radiologists, clinician or orthopedic surgeon that should be done with extreme care else, it will end up reflecting poor dataset quality and might result in reducing overall performance of the model (<https://www.altexsoft.com/blog/datascience/how-to-organize-data-labeling-for-machine-learning-approaches-and-tools>). The person responsible to do the labeling and time taken by him/her is the major challenge in creating a fully-fledged dataset for any classification task. A classification-based algorithm can accurately predict the outcome if the dataset is correctly mapped with extreme care and precision by a team of expertise. Although every industry has, its own regulation and governance challenges in the process of data collection, data management, and labeling that could take several months to finish but struggles of the healthcare industry are unique due to the complexity of data and extreme stringent regulations. A health institute may ask for a waiver of consent from an Institutional Review Board (IRB) study to mitigate some of these concerns, or researchers may process and anonymize DICOM data to strip away any patient health information. The source of data set, dataset split ratio, types of data collected and annotation process involved in classifying the datasets are presented in Table 1. Some freely available datasets for the task of fracture detection is presented in Table 2.

3 Image processing methods for fracture detection

The images collected in previous stage are processed using noise reduction, segmentation and feature extraction techniques for creating a convenient classification system. X-ray is the most frequently used imaging modality for fracture detection due to its painless, economic and non-invasive nature, which has gained enormous popularity in medical imaging.

Table 1 The source of data set, dataset split ratio, types of data collected and annotation process involved in classifying the datasets are shown in the table

Author	Bone type	Total images acquired from hospital	Dataset split ratio	Hospital Name	Fracture prevalence	Annotation process
Tian et al. (2003)	Femur	446	Training set-126 Test set-320	Local hospital, Singapore	Not available	Doctors used Neck-Shaft Angle (NSA) of the femur as the main diagnostic tool to identify fractures in femur X-ray samples Not discussed
Yap et al. (2004)	Femur	432	Training set-324 Test set-108	Local hospital, Singapore	12% in training and test set	Not discussed
Lim et al. (2004)	Femur	432	Training set-324 Test set-108	Local hospital, Singapore	12% in training and test set	Not discussed
	Wrist	145	Training set -71 Test set-74			
Lai et al. (2005)	Femur	432	Training set 324 Test set-108	Local hospital, Singapore	12% in training and test set	Not discussed
	Wrist	145	Training set-71 Test set-74		30% in training and test set	
He et al. (2007)	Femur	420	Training set-200 Validation set-160 Test set-60	Local hospital, Singapore	12% in training, validation and test set	Not discussed
Chat et al. (2011)	Femur	30	Not available	Biomedical research center, Malaysia	50% in input images	Not discussed
Mahendran and Baboo (2011b)	Tibia	1650	Training set-650 Test set-1000	Not available	60% in training and 62% in test set	Not discussed
Umadevi and Geethalakshmi (2012)	Tibia	Not available	Not available	Not available	Not available	Not discussed
Al-Ayyoub et al. (2013)	Hand	98	Training set-116 Test set-29	Jordanian Royal medical services, Jordan, Public domain	50% in input images	Annotations for the X-ray samples are obtained from radiology reports

Table 1 (continued)

Author	Bone type	Total images acquired from hospital	Dataset split ratio	Hospital Name	Fracture prevalence	Annotation process
Al-Ayyoub and Al-Zghool (2013)	Long bones	300	Not available	Hashemite kingdom of Jordan, public domain	33% in training and test set	300 images along with their labels are collected and verified by consulting medical experts
Cao et al. (2015)	Multiple	145	Training set-116 Test set-29	Public domain	Not available	Annotations for the X-ray samples are obtained from radiology reports
Gale et al. (2017)	Pelvis	Dataset collected from hospital is not available, dataset developed after augmentation is 53,278	Training set-45,492 Validation set-4432 Test set-3354	Royal Adelaide Hospital, SA	12% in training and validation set, 19% in test set	Fracture labels from radiographs are collected from radiology and orthopedic surgical unit reports The need for manual labeling of the entire dataset is effectively reduced to 7.4% of the dataset with the help of a radiologist
Dimililer (2017)	Multiple	100	Training set-30 Test set-70	Orthopedics traumatology hospital, Turkey	Not available	100 images are acquired from a benchmark database Annotation techniques are not discussed in the paper

Table 1 (continued)

Author	Bone type	Total images acquired from hospital	Dataset split ratio	Hospital Name	Fracture prevalence	Annotation process
Lindsey et al. (2018)	Wrist	Dataset collected from hospital is not available, dataset developed after augmentation is 31,590	Training set-28,341 Validation set-3149 Test set 1-3500 Test set 2-1400	HSS (Hospital for Special Surgery), United States	Not available	A team of 18 senior orthopedic surgeons were manually involved in annotating 135,409 radiographs out of which 34,990 are wrist images and 100,855 are bone images of 11 different body parts
Chung et al. (2018)	Humerus	1891	Training set-40,000 Test set-181	Multiple Hospitals, Korea	73% in training, validation and test set	Three specialists classify 1891 images collected from multiple hospitals of Korea into 4 categories Two Specialists were having 14–17 years of experience in shoulder orthopedic and one radiologist had 15 years of experience in musculoskeletal diseases

Table 1 (continued)

Author	Bone type	Total images acquired from hospital	Dataset split ratio	Hospital Name	Fracture prevalence	Annotation process
Kim and MacKinnon (2018)	Wrist	1389	T-Training set-8890 Validation set-1111 Test set-1111	Royal Devon & Exeter Hospital, UK	F ~ 50% in training, validation and test set	Trained radiologist was involved in converting images into JPEG format by selecting appropriate region of interest Wrist X-ray images are divided into fractured and healthy category by looking into radiological reports and are verified by radiology registrar having three years of experience
Myint et al. (2018)	Tibia	NA	Training set-40 for detection Test set-52 for classification	Yangon Orthopedic Hospital, Myanmar, radiology websites	Not available	Not available

Table 2 Publicly available bone X-ray dataset that could be used for the task of fracture detection

S No	Name	Source
1	Stanford ML group	MURA: A large dataset of 40,561 bone X-ray images consisting of elbow, finger, hand, humerus, forearm, shoulder, and wrist is publicly available for research use. The dataset is labeled by six board-certified expert radiologists from Stanford hospital and classified it into normal and abnormal cases. https://stanfordmlgroup.github.io/competitions/mura/
2	medpix	Medpix is an online database of medical images https://medpix.nlm.nih.gov/search?allen=true&all=true&query=fracture
3	Radiopaedia	More than 2800 fracture cases are freely available with diagnosis information https://radiopaedia.org/search/?lang=us&q=fracture&scope=cases
4	IIEST, Shibpur Indian Institute of Engineering Science and Technology	49 healthy, 99 fractured and 69 cancerous bone X-ray images are publicly available for sole purpose of conducting research http://oldwww.iiests.ac.in/component/content/article/155-itcategory/3282-medical-image-database
5	MOST: Multicenter Osteoarthritis Study (MOST)	Dataset of knee joint X-ray and MRI images is available with diagnosis report collected from case report forms (questionnaires and examinations) http://most.ucsf.edu/datadocs.asp
6	aylward.org	10,000 images of chest X-rays is available along with diagnosis data https://mhcc.app.box.com/v/ChestXray-NIHCC/folder/37178474737

Poisson, Gaussian, salt and pepper noise are various type of noise artifacts commonly found in radiographs, particularly when collected in a large quantity from public domain such as internet. The need for handling such images rises largely as reducing one type of noise sometimes affects the other. Gradient, Laplacian and Sobel are frequently used methods of edge detection, which is an effective measure to determine boundaries of objects in the image. The shapes and sizes of bone are non-identical in X-ray images due to the difference in age and gender of the patients. Normalization could be used to deal with such size variations but its results are unsatisfactory as it removes important texture information in shrunken images and adds noise and artifacts in case of larger images. Hence, adaptive sampling approach is applied in various literatures to sample X-ray samples instead of scaling them (Yap et al. 2004; Lim et al. 2004; Lai et al. 2005; He et al. 2007). Adaptive sampling does not require accurate segmentation of bone contours as done by Tian et al. (2003), a slight variation of shape is accepted here. Image transforms such as wavelets and curvelets are powerful algorithms to obtain decent quality compressed images with higher PSNR (Peak Signal-to-Noise Ratio) resulting in lesser memory requirements to store medical images. Both wavelet and curvelet transforms (a multi-scale method originated from wavelets) are commonly used for medical image compression, contrast enhancement, edge detection and image registration (Tian and Ha 2004). They are used for extracting enormous set of coefficients from input image, where insignificant features are eliminated via feature selection algorithm for better or faster classification. The primary step in various image-processing applications after smoothing and edge detection is the extraction of essential features (informative representations) from the image. Feature extraction focuses on extracting image characteristics that acquires visual image attributes and the performance of the classifier is highly dependent on the perfect set of features retrieved from the image. Texture can be a useful cue for detecting diseases or tissue types in medical imaging. Visual texture is used for segmenting and discriminating objects from background that has a repeated pattern of elements with some amount of variability in element appearance and relative position. The spatial features of an image is described by its gray level, spatial distribution and amplitude, where amplitude is the simplest feature that discriminates bones tissues from X-ray images (Haralick and Shanmugam 1973). Intensity inhomogeneity and absence of sharp edges are the causes for missed out cracks while detecting fractures from X-ray images (Linda and Jiji 2017). Crack identification is significant for analyzing suspicious cases so that medical experts can suggest possible course of action within time limits (Linda and Jiji 2011). Fuzzy based image segmentation approaches are proposed for identifying cracks by employing fuzzy membership function (Linda and Jiji 2011, 2017). Image histogram is divided into three subsets to produce subset parameters and these parameters act as initial estimates to classify each pixel into one of the subsets by minimizing the fuzzy index (Linda and Jiji 2011). Their result gives promising results in detecting minute cracks from X-ray images. Hairline breakage from X-ray image is recognized by calculating intensity variation over the segmented bone regions (Linda and Jiji 2017). This approach performs better than standard approaches based on fuzzy thresholding (Linda and Jiji 2011; Mansoor et al. 2012) with an overall accuracy of 98%.

With the rise of deep learning neural networks, deep layers of Convolutional Neural Networks (CNN) replaced the task of feature extraction in digital images. CNN is a multi-layered neural network, which consists of convolution layer, sub-sampling layer, and fully connected layers. Convolution and sub-sampling layers of CNN are part of feature learning process while fully connected layers are used for classification. ConvNets or CNN have the ability to learn various low level (minor details of the image e.g. lines, dots or edges etc.) and high level features (built upon low-level features to detect objects and larger shapes)

through abstraction in the layers. Features are extracted using CNN in recent approaches of fracture detection and classification (Gale et al. 2017; Lindsey et al. 2018; Chung et al. 2018; Kim and MacKinnon 2018). Table 3 demonstrates relevant review findings of the image processing methods used for fracture detection in the literatures reviewed.

4 Conventional machine learning based algorithms for fracture detection

Various features including textual, shape, edges, horizontal and vertical lines extracted in previous steps are provided to classification algorithms for predicting the occurrence of bone fractures and thereby classifying it into relevant categories. Once the perfect set of features are fed into the classifier, the accuracy of fracture detection depends on the classifier selected. Hence, proper features must be extracted in order to formulate a powerful classification model. Table 4 demonstrates relevant review findings of the conventional machine learning based algorithms used for fracture detection.

4.1 Primary machine learning based algorithm

Some earlier works of automatic fracture detection reflects the usage of single feature as a classification parameter. For instance, the first work on automatic fracture detection used Neck-Shaft Angle (NSA) as the only feature for fracture detection (Tian et al. 2003). Radiologists considered the image as fractured if the NSA is less than 116° . Using such type of model helped in correctly identifying 94.4% of training and 92.5% of test samples. The major reason behind 7.5% of error rate in test cases is the model's incompetency in detecting minute changes in the femur neck-shaft angle. The upper extremity region of femur is called trabeculae and orientation of trabeculae on femur neck and head significantly changes on the event of fracture. These changes can be detected using neck-shaft angle but local disruptions remain undetected using this approach. Hence, a new approach is proposed which performs texture analysis of trabecular pattern by extracting features in femur X-rays followed by classification to detect such minor disturbances (Yap et al. 2004; Lim et al. 2004). Researchers in Yap et al. (2004) extracted Gabor features while (Lim et al. 2004; Lai et al. 2005) used Gabor orientation extracted by Yap et al. (2004) and additionally acquired Intensity Gradient Direction (IG) and Markov Random Field (MRF) texture features in femur X-rays. These features are then fed to the chosen classifier for diagnosing fractures in X-ray samples. GLCM is used as the only feature in classifying femur fractures in 30 X-ray samples and attain sensitivity and accuracy of 80% and 86.67% respectively (Chai et al. 2011).

4.2 Ensemble based classification system

Ensemble is a machine learning technique, which combines diverse models or classifiers to generate an optimal model that will best predict our wanted outcome. The basic idea behind ensemble model is to use multiple learning algorithms to obtain better predictions as oppose to traditional models that rely on a single classifier's performance. The accuracy of the classifier depends on the perfect set of features extracted or learned from the image. However, the accuracy could further be improved by combining multiple classifiers and by integrating results of all independent classifiers. Ensemble based classification system

Table 3 The table demonstrates relevant review findings of the image processing methods used for fracture detection and classification

Author	Bone type	Relevant Review Findings
Tian et al. (2003)	Femur	Neck shaft angle (NSA) is considered as the only feature for fracture detection
Yap et al. (2004)	Femur	Image features are extracted by performing texture analysis of trabecular pattern in femur X-rays Neck shaft angle (NSA) is considered as the only feature for fracture detection
Lim et al. (2004)	Femur and wrist	Femur contours are extracted using active shape model, supplemented by active appearance models Image features are extracted by performing texture analysis of trabecular pattern in femur X-rays NSA, GO, IGD and MRF are the four different image features extracted from X-rays samples Adaptive sampling is employed to sample the X-ray images instead of scaling them
Lai et al. (2005)	Femur and wrist	Femur fractures are detected by extracting GO, IGD and MRF from X-rays samples Adaptive sampling approach is employed to sample X-ray samples instead of scaling the images
He et al. (2007)	Femur	Femur fractures are detected by extracting GO, IGD and MRF from X-rays samples
Chai et al. (2011)	Femur	The input image is translated into binary form to maximize the calculation speed and ease the computing process Laplacian edge detector is used to detect femur edges followed by removal of fine particles using median filter
Mahendran and Baboo (2011b)	Tibia	K-means unsupervised clustering approach is used to segment shaft and non-shaft regions of femur images SACEN technique is proposed that can simultaneously adjust contrast, enhance edges and remove noise in tibia X-ray images. CLAHE algorithm (Pisano et al. 1998) is used to enhance edges in edge-detected regions of tibia images and WEAD algorithms (Rajan and Kaimal 2006) is used to remove noise in blurry portions of non-edge region
Linda and Jiji (2011)	Multiple bones	Segmentation of the bone region from X-ray image is performed in the second stage followed by extraction of diaphysis region from epiphyses and fleshy regions of tibia X-ray images using region-growing algorithm GLCM, GO, MRF and IGD are various texture features extracted from the processed tibia X-ray images An optimum threshold value is obtained for segmenting bone regions from its surrounding area in grey scale images Infestation sites of crack in X-ray image is detected using fuzzy index measure Various operations involved in quantifying number of cracks includes grid formation, local thresholding, threshold value interpolation, segmentation using fuzzy index measure, background removal, and morphological filtering

Table 3 (continued)

Author	Bone type	Relevant Review Findings
Umadevi and Geethalakshmi (2012)	Tibia	The image quality of X-ray image is enhanced by coupling a hybrid denoising method that uses ICA with wavelets GLCM along with GO, MRF and IGD are used as a principal tool for extracting texture features in tibia images and the performance of the classification system is further improved by combining texture and shape features Shape features are retrieved using fast hough transformation algorithm (Hari et al. 2010) after segmenting diaphysis regions in tibia images
Al-Ayyoub et al. (2013)	Hand bones	Salt and pepper noise is reduced using median filter in bone X-ray images Sobel algorithm is used to detect object boundaries in hand X-rays GLCM entropy, contrast, correlation and homogeneity is used to extract texture features Thousands of features are extracted from hand X-rays and 84 features are finally selected for fracture detection using Weka supervised attribute filter
Al-Ayyoub and Al-Zghool (2013)	Long bones	Intensity variations among X-ray images are handled using histogram equalization process Gaussian noise is handled using a fast and efficient filtering algorithm that gives higher PSNR and lower mean absolute error as compared to mean, weiner, k-means, alpha-trimmed mean and trilateral algorithms A popular filtering algorithm known as haar wavelets is effectively utilized to enhance bone images Salt and pepper noise is handled using k-fill algorithm Edges are detected using modified version of canny algorithm Corner (intersection of two edges) is detected by using a combination of popular harris algorithm and tensor based corner detection algorithm (Teixeira et al. 2008)
Cao et al. (2015)	Multiple bones	Three different types of features are used to localize fractures in X-ray samples. This includes GO, Schmid texture feature, and proposed Contextual-Intensity (CI) feature
Gale et al. (2017)	Pelvis	Preprocessing is done using three CNN based models to deal with variations in medical studies: 1. CNN-frontal is trained to make the model differentiate anatomical features of pelvis images from other images like chest, lateral hip X-rays, spinal images etc. 2. CNN-bounding is a regression-based model trained to identify locations where fracture can occur by localizing the femur neck 3. CNN-metal is trained to include only relevant hip fracture cases by excluding cases where metal implant has occurred Features are extracted using CNN

Table 3 (continued)

Author	Bone type	Relevant Review Findings
Dimililer (2017)	Multiple bones	Haar wavelet transform (Khashman and Dimililer 2008) is applied to enhance the image quality by reducing noise in X-ray images followed by a feature extraction method based on Scale-Invariant Fourier Transform (SIFT) algorithm
Linda and Jiji (2017)	Multiple bones	2D anisotropic diffusion filter (Mahmoodi 2011) and Discrete Wavelet Transform (Sharan et al. 2014) are used for noise removal and edge sharpening in input X-ray samples Expectation Maximization (EM) algorithm (Zhang et al. 2001) is implemented for segmenting bone from flesh and surrounding regions Intensity variation is calculated for detecting hairline breakage Features are extracted using CNN
Lindsey et al. (2018)	Wrist	Features are extracted using CNN
Chung et al. (2018)	Humerus	Features are extracted using CNN
Kim and MacKinnon (2018)	Wrist	An expert radiologist converted X-ray image into JPEG by selecting the appropriate region of interest To reduce the possibility of overfitting, "Red spot" annotations are removed that were applied by the radiologists during image acquisition
Myint et al. (2018)	Tibia	RGB image is converted into grayscale image Bone regions of tibia images are enhanced using USM (Unsharp Mask Filter) Harris corner detection algorithm is used to detect break points of the leg bone in pre-processed image

GLCM Gray level co-occurrence matrix, *GO* Gabor orientation, *IGD* intensity gradient direction, *MRF* Markov random field, *SACEN* simultaneous automatic contrast adjustment, edge enhancement and noise removal, *CLAHE* contrast limited adaptive histogram equalization algorithm, *WEAD* wavelet based anisotropic diffusion algorithm, *ICA* independent component analysis, *PSNR* peak signal-to-noise ratio, *CNN* convolution neural network

Table 4 The table demonstrates relevant review findings of conventional machine learning based algorithms used for fracture detection

Author	Bone type	Relevant review findings	Feature-classifier combination or ensemble approach to detect fractures
Tian et al. (2003)	Femur	Neck-shaft angle of adult femur is considered as the only feature for fracture detection The NSA of a healthy adult femur is about 120 to 130 degrees and a fracture is detected if the NSA is less than 116 degrees The upper extremity region of femur is called trabeculae and the orientation of trabeculae on femur neck and head significantly changes on the event of fracture These changes can be detected using neck-shaft angle but local disruptions remain undetected using this approach	NA
Yap et al. (2004)	Femur	Image features are extracted by performing texture analysis of trabecular pattern in femur images Minor disturbances in the neck-shaft angle are detected by employing ensemble approach Subtle fractures are not detected that neither change, the femur's shape nor disrupt the trabecular pattern significantly	1. NSA + Bayesian 2. NSA + SVM 3. NSA + NB + SVM (<i>ensemble</i>) A fracture is detected if any 'one' out of 'three' or any 'two' out of 'three' combinations detect fracture
Lim et al. (2004)	Femur and Wrist	Feature extraction is followed by classification to detect the fracture in images of femur and wrist A combination of MRF and SVM classifiers have shown the best performance in predicting wrist fractures	1. NSA + Thresholding 2. GO + Bayesian 3. GO + SVM 4. IGD + Bayesian 5. IGD + SVM 6. MRF + SVM A femur fracture is detected if any 'two' out of 'six' or any 'two' out of 'four' combinations detect fracture
Lai et al. (2005)	Femur and wrist	Multiple classifiers are combined with texture features and showed better performance in detecting wrist fractures	1. GO + Gini-SVM 2. MRF + Gini-SVM (performed best in terms of sensitivity and accuracy for femur images) 3. IGD + Gini-SVM

Table 4 (continued)

Author	Bone type	Relevant review findings	Feature-classifier combination or ensemble approach to detect fractures
He et al. (2007)	Femur	Fracture class is divided into three categories: healthy, fractured, and unknown Hierarchical SVM is combined with features such as GO, MRF and IGD to predict fractures based on divide and conquer technique	GO + MRF + IGD + Hierarchical SVM Divide and conquer approach is used for femur fracture prediction
Chai et al. (2011)	Femur	GLCM is used to generate exact borders in input samples by separating fractured and healthy bones from femur X-rays Proposed GLCM classifier has successfully detected fractures with 86.67% accuracy	NA
Mahendran and Baboo (2011b)	Tibia	GLCM features (mean, variance, energy, entropy, homogeneity), GO, MRF and IG are various texture features used Fracture is located using modified Hough transform with gradient analysis	1. Texture features with BPNN 2. Texture features with SVM 3. Texture features with NB 4. Texture features with (BPNN + SVM + NB) A tibia fracture is detected if any 'two' out of 'four' combinations detect fracture
Umadevi and Geethalakshmi (2012)	Tibia	Feature vector is generated by combining texture features and shape features The idea behind selecting BPNN, SVM and NB as base classifiers for designing ensemble system is the presence of high diversity as per Yule's Q method (Kuncheva and Whitaker 2001) The classifiers are trained using boosting approach and the fractures are detected using majority-voting scheme Performance measures such as accuracy, precision recall and F-Measure are calculated from confusion matrix	1. Texture features + (BPNN + SVM + NB) 2. Shape features + (BPNN + SVM + NB) 3. Shape features + texture features + (BPNN + SVM + NB) A combination of (BPNN + SVM + NB) along with shape and texture features reflect significant improvement in detecting tibia fractures

Table 4 (continued)

Author	Bone type	Relevant review findings	Feature-classifier combination or ensemble approach to detect fractures
Al-Ayyoub et al. (2013)	Hand bones	NB, DT, NN and BN are selected as base classifiers to detect bone fractures in hand X-rays The best accuracy is obtained when wavelet features used independently with NB classifier and with combined feature sets	The performance of the individual and ensemble classifiers is reported on wavelet, Curvelet and GLCM feature sets
Cao et al. (2015)	Multiple bones	Fracture is located using Efficient Subwindow Search (ESS) algorithm. The top seven bounding boxes generated as result of fracture localization contains 81% of detection rate The proposed model outperforms SVM and single layer of stacked random forest in detecting and localizing fractures in bone X-ray images	Fracture is detected by utilizing a multi-layer classifier with each layer consisting of various random forests

NSA neck-shaft angle, *NB* Naive Bayes, *SVM* support vector machine, *GO* Gabor orientation, *IGD* intensity gradient direction, *MRF* Markov random field, *GLCM* gray level co-occurrence matrix, *BPNN* back propagation neural network, *KNN* K-nearest neighbor, *DT* decision tree, *NN* neural network

has acquired broad range of attention in various fields such as face recognition (Antipov et al. 2016), geospatial land classification (Minetto et al. 2018), video-based face recognition system (Ding and Tao 2017), medical image segmentation (Kumar et al. 2017), wind power forecasting (Wang et al. 2017) etc. These models have manifested better accuracy (low error) by avoiding overfitting issues and by reducing bias and variance error as compared to its individual constituent classifiers. The importance of such models can be understood by the fact that various prestigious machine-learning competitions like popular Netflix challenge (Andreas et al. 2009), Knowledge Discovery in Databases (KDD) cup 2009 and Kaggle had used ensemble based model to achieve the best accuracy. The most prominent ensemble based models have used Neural Network (NN), Support Vector Machine (SVM), Naïve Bayes (NB) algorithms for diagnosing fractures in human bones from year 2003 to 2015. It is observed that the performance of the classifier is significantly improved with the introduction of multiple-classifier based systems where individual results from base classifiers are fused together (Umadevi and Geethalakshmi 2012; Yap et al. 2004; Lai et al. 2005; Mahendran and Baboo 2011b). Decision of choosing the best classifier among all competing classifiers depend on diversity among models. Choosing classifier merely based on accuracy on training data is fallacious. Some level of diversity must exist among classifiers that are part of ensemble system to make it an effective process, which can be achieved using the approaches given below (Polikar 2009):

1. Using different classification algorithms for ensemble system.
2. Using same classification algorithm with different instantiation or different hyper-parameter settings.
3. Using different feature sets:
 - (a) Random selection
 - (b) Feature selection
4. Using different training sets:
 - (a) Bagging
 - (b) Cross-validation

(a) Bagging or Bootstrap aggregating

A widely accepted technique of ensemble is non-hybrid classifier where same classification algorithm with different instantiation or different hyper-parameter settings is combined to make ensemble model. Bootstrap aggregating is one of the most intuitive and earliest ensemble based algorithms where multiple models consisting of same learning algorithms are trained with subsets of random datasets picked from the original training set with replacement. The output of the multiple-classifier or ensemble is predicted on the basis of majority voting of the constituent classifiers. Several variations of this algorithm exist which tend to enhance the performance of the model. Most popular among them works by increasing diversity among training data for individual classifiers and other is by making use of different classification algorithms (Fig. 2).

(b) Boosting

It is a simple variation of bagging technique, which strives to improve the classification model by converting weak learners into strong learners sequentially, each trying to correct its predecessor. The major difference between bagging and boosting is that bagging follows parallel training stage where each model is built independently and boosting follows sequential approach where current model architecture depends on previous classifier's success. It is a sequential process where similar weights are assigned to the data at the beginning and are redistributed after each training stage, allowing subsequent learners to emphasize more on misclassified cases that are now attached with higher weights (Fig. 3).

(c) Stacked ensembles

Stacking involves multi-layer of learning stages where first layer consists of base learners followed by lower level meta-learner stages, which takes base learners as input to obtain the best combination of first-level base learners. "Super learner" concept was originally developed in the year 1992 (Wolpert 1992) but its implementation with improved performance is shown for the first time in 2007 (Laan et al. 2007), which proves that stacked ensembles helps in creating optimal model for learning. A popular machine-learning algorithm is random forest, which takes a collection of weak learners (e.g. decision tree) and form a single, strong learner by following bagging technique. A fracture detection technique using 145 X-ray images with tenfold cross validation covering various regions like hand, foot, ankle, knew, lower leg and arm is developed by stacked random forests feature fusion (SRF-FF) technique (Cao et al. 2015). A four layer random forests with five decision trees is implemented in the first layer and remaining layers uses fifteen trees as shown in Fig. 4. The classifier is trained to produce confidence score maps which indicates the probability of fractures in X-ray images and then uses Efficient Subwindow Search (ESS) algorithm (Lampert et al. 2008) for localizing regions that has maximum probability of fracture occurrences. The proposed model outperforms SVM and single layer of stacked random forest in detecting and localizing fractures in X-ray images.

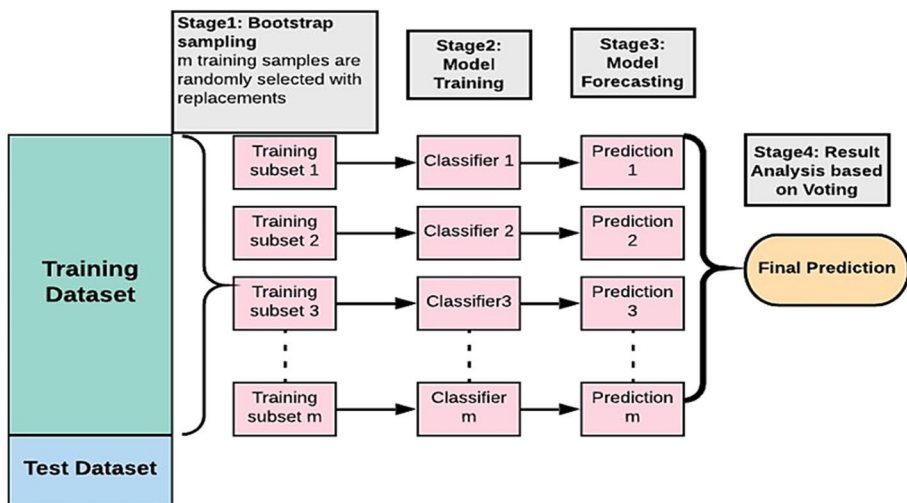


Fig. 2 Bagging process involves parallel execution of individual classification algorithms where training subsets selected randomly with replacement from the training dataset. Outcome of this ensemble model is predicted by majority voting among individual classifiers

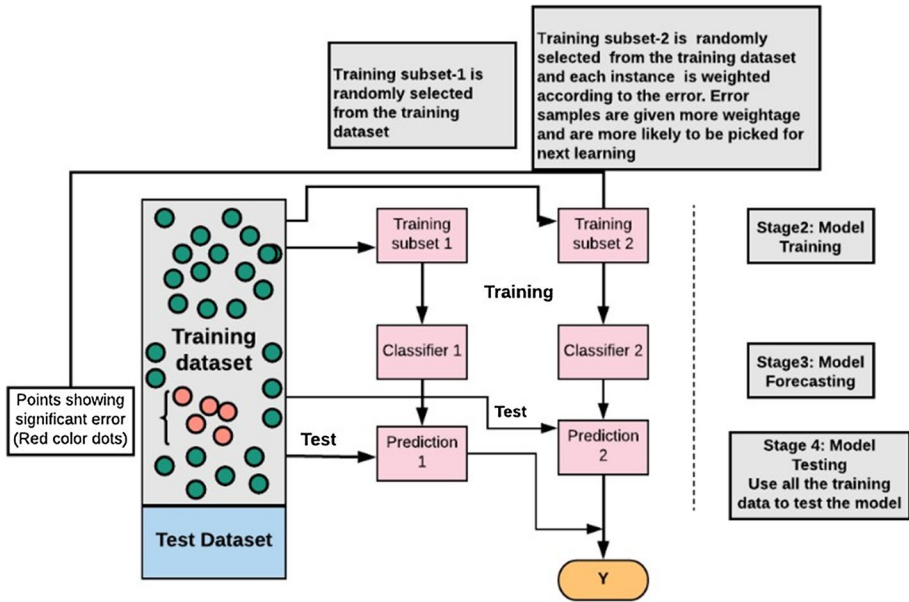


Fig. 3 Boosting is a sequential process where similar weights are assigned to the input points/data at the beginning and are selected randomly from the training set. After each training and testing, misclassified samples are identified and are attached with higher weights. In this manner, it allows subsequent learners to emphasize more on misclassified cases having higher chances to be picked for next classification

Divide and conquer is another type of ensemble technique based on the concept that each sub problem is easier to solve than the whole problem. It demands larger training set and complex problem for it to produce larger clusters and thereby producing effective results. The complex problem of fracture detection is partitioned into Gini SVM’s kernel space instead of feature space as it lacks bigger training set (He et al. 2007). The process starts by training Gini SVM on training set T and calculates error on validation set V. The error obtained is used to select a new validation set V’ (subset of V) which is further classified on the basis of new SVM and training set T’ of T at the next level. This type of architecture enhances the accuracy of the SVM by ensuring that the lower-level SVM (child) always complements the performance of higher level SVM (Parent).

5 Deep learning-based algorithms for fracture detection

Deep learning is a branch of machine learning and artificial Intelligence (AI), which consists of statistical analysis algorithms, that repeatedly trains data to make predictions (Fig. 5). It is the ability of these “trained models” to automatically learn and improves from experiences to make predictions on unknown data (Kim and MacKinnon 2018). Discovering significant features, which depicts abnormalities or pattern in the data, is of great importance in machine learning approach. Traditionally, these features are crafted predominantly with the help of human expertise but, with the advancement of machine learning techniques, the models can automatically learn these features. In Radiology, skilled radiologist extract meaningful information/features from images and then interpret those

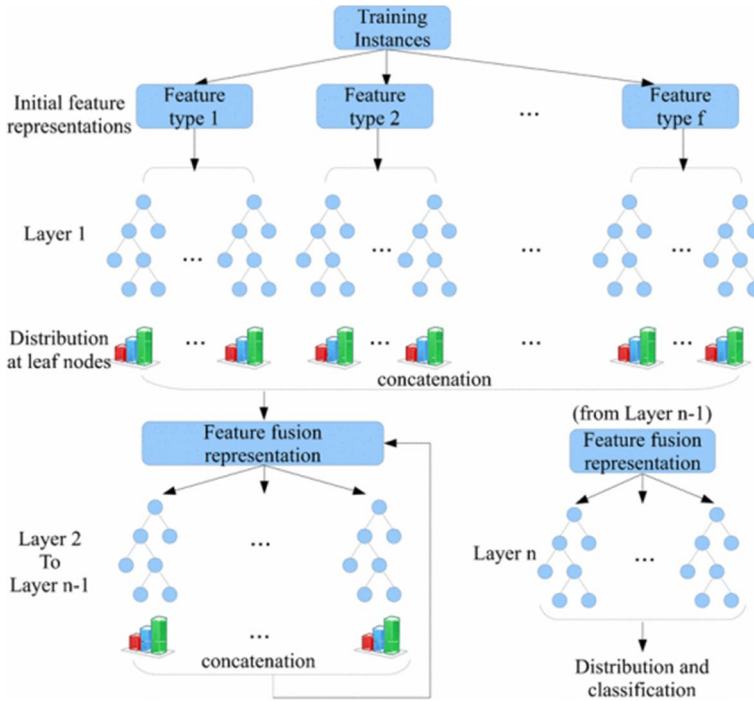
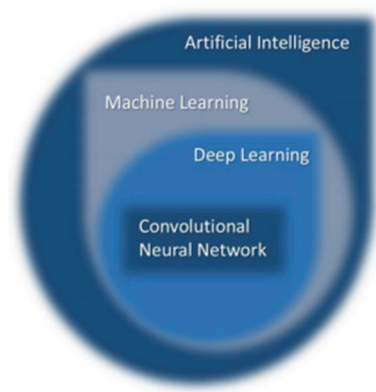


Fig. 4 Flow chart of stacked random forests feature fusion (Cao et al. 2015)

features based on their expertise, experience and knowledge. Thus, it provides tremendous opportunities to apply machine-learning algorithms to make autonomous predictions on the data with similar accuracy as the radiologist expert (Kohli et al. 2017). A widely accepted computational model in the field of machine learning for finding complex patterns in the data is Artificial Neural Networks (ANN). These are the brain-inspired systems, which contemplate to imitate the human's learning process. Neural Networks are also recognized as perceptron's and have existed since the 1940s but have become significant part of artificial intelligence from past few decades. One reason of them become dominant in machine learning area is the advent of a technique called "backpropagation". Backpropagation allows neural network to adjust their weights in hidden layer of neurons according to the desired output (Dormehl 2019). Intelligent Bone Fracture Detection System (IBFDS) combines image processing and neural network techniques for bone fracture detection (Dimililer 2017). Firstly, haar wavelet transform (Khashman and Dimililer 2008) is applied to enhance the image quality by reducing noise in the X-ray image followed by a feature extraction method based on Scale-Invariant Fourier Transform (SIFT) (Lowe 2004). Finally, extracted features are applied to a 3-layer back-propagation ANN, which classifies the image into fracture and non-fracture category.

Deep learning is the advancement of artificial neural networks, which consists of multiple hidden layers and provides greater levels of abstraction. With the rise of deep neural network, the accuracy of predicting a task has improved in a tremendous manner by incorporating deep layers into the model that allows system to learn complex data (Kim 2016). The rise of deep learning in healthcare sector is driven by various factors such as; (1) availability of large datasets, which became possible due to rapid accumulation of electronic

Fig. 5 Deep learning is a branch of machine learning and machine learning is a branch of artificial learning and perform tasks which requires human intelligence (Meyer et al. 2018)



data in the form of Electronic Medical Records (EMRs), (2) GPU advancement providing better performance with graphics and videos, (3) progress in deep learning algorithm due to incorporating multiple layers in deep learning architecture (Shen et al. 2017).

A variant of deep learning architecture is Convolutional Neural Networks (CNN), which consists of convolution layer, sub-sampling layer, and fully connected layers as depicted in Fig. 6. Feature learning techniques took a tremendous turn when CNN model came into existence which can “learn the features” instead of handcrafting them into the system. Convolution and sub-sampling layers of CNN are part of feature learning process while fully connected layer is used for classification. It has emerged as a recent breakthrough in machine learning and accomplished humongous success across various fields of medical image analysis, such as image segmentation, image registration, image fusion, image annotation, genomics etc. Radiologists uses a blend of perception, memory, pattern recognition, and cognitive reasoning for interpreting radiographic studies. Numerous distractors affect their performance, which ultimately leads to increasing workloads and fatigues. Hence, developing systems or tools that automatically detects abnormalities or patterns in the musculoskeletal radiographs without human intervention would improve patient’s security (Waite et al. 2017). A model known as AlexNet developed by Krizhevsky et al. (2017) brought out a revolution in the field of computer vision, which gave a new insight to deep CNN. It was originally developed to compete in the ImageNet competition whose general

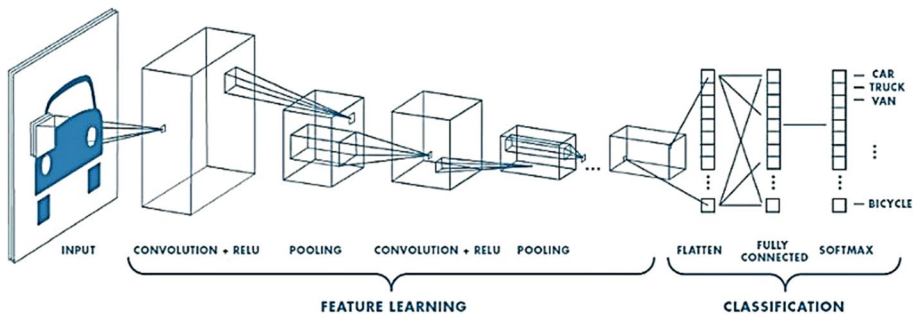


Fig. 6 Schematic representation of the CNN architecture, consisting of convolution layer, sub-sampling layer and fully connected layers where convolution and sub-sampling layers are part of feature learning process and fully connected layer is used for classification (Prabhu 2018)

architecture is similar to LeNet-5 (LeCun et al. 1998) but is substantially larger. It became successful in convincing Computer Vision community for its usefulness after securing first position in the ImageNet competition. Use of effective regularization parameters, data augmentation approaches, rectified linear unit and, the use of graphics processing units for meeting computing requirements helped in bringing out this revolution. It was recognized as one of the top 10 milestones of 2013 in deep learning (Birks et al. 2016). Deep architecture of a CNN is its biggest strength, which allows extraction of discerning features at different layers of abstractions (Krizhevsky et al. 2017; Szegedy et al. 2016).

A large amount of training data is required to train a Deep Convolutional Neural Network (DCNN) from scratch and it becomes difficult to ensure proper convergence of the model due to absence of huge amount of training data especially in medical imaging, where data is kept heavily protected due to privacy concerns. The process of creating new data with minor alterations such as flips, rotation, mirroring, translations etc. from our existing dataset is known as data augmentation that helps in reducing data insufficiency and overfitting problems. The ability of CNN architectures to detect and classify fractures even when they are placed in different orientations is mitigating data scarcity issues, which is the major cause of hindrance to the deep neural network architectures. It is due to data augmentation techniques that the number of radiographs used for diagnosing fractures is amplified from 500 to hundreds of thousands from year 2003 to 2018. Only few published works (shown in Table 5) related to fracture diagnosis have applied augmentation techniques to their private datasets. This approach could successfully be applied in upcoming work to enhance the performance of the classifier. Another promising alternative to data scarcity problem is to fine-tune a CNN, which is pre-trained on a different network architecture (Birks et al. 2016). Pre-trained CNN is the state-of-the-art image classification network trained on millions of images of a particular domain running for several weeks on multiple servers and then used in a different domain of interest. This approach has become extremely useful for researchers where scarcity of resources is an obstacle for creating a successful model from scratch. These large pre-trained models can be utilized to obtain highly sophisticated and powerful set of features needed for required domain of interest (Greenspan et al. 2016). A study is performed where two models of CNN are compared. One model is trained from scratch and the other is a pre-trained model, which is further fine-tuned on the required domain (Birks et al. 2016). The performance of both the model is compared on distinct medical imaging applications such as classification, detection, and segmentation. A fine-tuned CNN model efficiently outperformed the CNN model trained from scratch in the best case and performed similar to CNN model in the worst case. Therefore, fine-tuning of existing model could offer a practical way to reach the best performance for the application at hand based on the amount of available data. Table 6 demonstrates relevant review findings of the deep learning based algorithms used for fracture detection.

Six board-certified expert radiologists from Stanford hospital develop a very large dataset of 40,561 X-ray images consisting of elbow, finger, hand, humerus, forearm, shoulder, and wrist. They manually examined all X-ray samples and labeled them as abnormal and normal (Rajpurkar et al. 2017). The model is pre-trained using ImageNet (Deng et al. 2009) and fine-tuned on 169 layer dense CNN based on (Huang et al. 2017) for end-to-end classification. Hyper-parameters such as Adam parameters (Kingma and Ba 2015) and mini-batches of size 8 is used in achieving similar performance as the best radiologist in detecting abnormalities on finger and wrist bones but the model reflected lower performance than the worst radiologist in forearm, shoulder and humerus samples. Their model is performing similar to the worst radiologist in finger and elbow studies and performing

Table 5 Data augmentation observed in literatures involved in the survey

Author	Images before augmentation	Images after augmentation	Augmentation techniques
Gale et al. (2017)	Not available	45,492	Translation, rotation, histogram matching, shearing
Rajpurkar et al. (2017)	Not available	40,561	Lateral inversion, rotation
Lindsey et al. (2018)	Not available	31,590	cropping, horizontal mirroring, rotation, lighting and contrast adjustment
Chung et al. (2018)	1891	>40,000	Shifting, scale transformations, rotation
Kim and MacKinnon (2018)	1389	11,112	Horizontal flip, rotation, width and height shift, shearing, zooming

Table 6 The table demonstrates relevant review findings of deep learning based algorithms used for fracture detection

Author	Bone type	Relevant review findings
Dimililer (2017)	Multiple bones	Total 100 images are acquired from a benchmark database and are divided into 30 training and 70 test samples. This division of dataset into training and test set is quite unusual from the split ratio adopted heavily across neural networks (Shah 2017)
Gale et al. (2017)	Pelvis	BPNN is used to detect fracture in multiple bones 172 layers deep convolution neural network is trained to classify fracture into healthy and fractured classes The performance of the proposed model is compared against radiologist performance with (CI _{95%}) The performance is also compared against existing published work where both are using dissimilar datasets but similar data distribution in the test set but merely using similar distribution on the test set is not an efficient criterion of doing comparative analysis
Chung et al. (2018)	Humerus	The DCNN model is classified into greater tuberosity, surgical neck, 3-part, and 4-part classes based on Neer (Neer 1970) classification Model is pre-trained using ResNet-152 (https://github.com/kaimingHe/deep-residual-networks) on non-medical images and fine-tuned using Inception v3 on proximal humerus dataset
Kim and MacKinnon (2018)	Wrist	Model is pre-trained using Inception v3 network on non-medical images and fine-tuned using Inception v3 on wrist images
Lindsey et al. (2018)	Wrist	Model is pre-trained using 100, 855 radiographs of all body parts except wrist images and fine-tuned using DCNN on 31,490 wrist images Instead of pre-training the model on non-medical images, the model is pre-trained on medical images comprises of bone images of 12 different body parts

better in remaining samples. An Inception V3 network (Szegedy et al. 2016), which is pre-trained on ImageNet Large Visual Recognition Challenge (ILSVRC) (Russakovsky et al. 2015), is used to automatically detect fractures in 11,112 wrist X-ray images (Kim and MacKinnon 2018). They have additionally used 100 unused wrist images exhibiting 50% fracture prevalence for testing their model and results obtained are comparable with that of the experts with an AUC of .954. However, both approaches are limited to only wrist images, which can further be extended to other bones for its practical adaptability. Proximal humerus fractures are detected and classified into greater tuberosity, surgical neck, 3-part, and 4-part classes based on Neer (Neer 1970) classification (Chung et al. 2018). Microsoft's ResNet-152 is used for pre-training and fine-tuning the model for detecting and classifying fractures in 1891 radiographs collected from various hospitals of Korea. The model outperformed the performance of general physician and general orthopedist in detecting humerus fractures and reflected equivalent performance to shoulder orthopedist (Chung et al. 2018). 100,855 radiographs of 12 different body parts are used to acquire random model parameters by pre-training the model on medical images instead of non-medical images (Lindsey et al. 2018). DCNN is used to fine-tune the model on 31,490 wrist images by initializing the model with the parameters obtained from a pre-trained model. The first output of the model is the probability of the presence of fracture in the image and the other output identifies location and extent of the fracture. Finally, the trained model is tested on two different test sets where first set consists of 3500 images, randomly collected from training and validation set and second set consists of 1400 unseen images collected over a period of 3 months from the same hospital. A controlled experiment is additionally executed to test the accuracy of emergency clinician with and without model's assistance and revealed that the emergency clinician involved in providing X-ray interpretation experienced a reduction rate of 47% by taking assistance of the trained model.

6 Multi-fracture identification techniques for bone X-ray images

The study of multi-fracture classification techniques is extremely significant for speedy recovery of the patient. However, the field of orthopedic surgery and traumatology have investigated scarcity of techniques in classifying fractures despite its huge importance to public health (Chung et al. 2018). Fractures can vary from simple to complex types based on its location and complexity (Mayne 2013). Complex types can become severe for which imperative treatment is required to avoid further complications. Majorly found fractures are addressed in the introduction section and its types are outlined in Fig. 1. Long bone consists of upper and lower extremity region and fracture existing in these bones are identified in several literatures (Chung et al. 2018; Al-Ayyoub and Al-Zghool 2013; Myint et al. 2018; Mansoory et al. 2012; Linda and Jiji 2018; Mahmoodi 2011). Often researchers accomplish the task of fracture classification in two phases where the fracture is identified in first phase, and is classified into multiple categories in the second phase. Authors have categorized long bone fractures into five classes: normal, greenstick, spiral, transverse and comminute using computer vision techniques such as DT, SVM, NB, and NN, out of which SVM outperforms all other classifiers with an accuracy of more than 85% under ten-fold cross validation (Al-Ayyoub and Al-Zghool 2013). Lower leg bones known as tibia is considered for carrying out fracture classification task by utilizing K-Nearest Neighbor (KNN) approach (Myint et al. 2018). The fracture is identified as one of the four possible labels; normal, oblique, transverse and comminute whereas harris corner points are

used to identify fracture locations in tibia images. Cohen's Kappa assessment is effectively used for measuring the correctness of KNN classifier with an accuracy of 82% (Myint et al. 2018). Proximal humerus fractures are detected and classified into greater tuberosity, surgical neck, 3-part, and 4-part classes based on Neer (Neer 1970) classification (Chung et al. 2018). 1891 radiographs collected from various hospitals of Korea are divided into 10 partitions, where one of the partitions is kept for test set and Microsoft's ResNet-152 is used for pre-training and fine-tuning the model for detecting and classifying fractures. The model outperformed the performance of general physician and general orthopedist in detecting humerus fractures and reflected equivalent performance to shoulder orthopedist (Chung et al. 2018). CT DICOM images are taken as input to identify complex fractures by utilizing linear structuring elements (Linda and Jiji 2018). Various 2D slices of 3D input image are enhanced by removing noise and sharpening the edges using 2D anisotropic diffusion filter (Mahmoodi 2011; Mendrik et al. 2009). Hidden Markov Random, Field-Expectation Maximization (HMRF-EM) (Zhang et al. 2001) and adaptive thresholding techniques (Singh et al. 2012) are used for segmenting bone regions from fleshy areas followed by segregation of fractured area using template-matching technique (Jurie and Dhome 2001). Complex fractures are identified and located by means of linear structuring elements in the image. Finally, the fracture is visualized in 3D using Ray Cascade method (Sathik et al. 2015). The performance of the proposed approach is validated against expert radiologist data, providing an overall sensitivity and specificity of 95% and 97% respectively where 57 patients suffered from small to severe bone fractures out of 70 test images (Linda and Jiji 2018).

6.1 Digital geometry based approaches for fracture identification

Senior citizens often suffer from long bone fractures due to conditions such as osteoporosis, stress, and sudden fall. Various literatures (Bandyopadhyay et al. 2013, 2016a; b; c) have suggested approaches based on digital geometry, which provide a powerful tool for analyzing bone fractures from X-ray images. A precise fracture detection approach relies on accurate segmentation of bone areas from the fleshy regions in X-ray image and the quality of fracture detection depends on the sharpness and clarity of the bone contour. However, the intensity based segmentation method becomes inaccurate if the pixel belonging to the bone region and fleshy areas has overlapping intensity regions. Therefore, an entropy based thresholding technique is proposed to accurately segment the bone areas from its surrounding fleshy regions in bone X-ray images (Bandyopadhyay et al. 2016a). The authors successfully implemented spatial filtering techniques based on morphology to remove the noise and spurious edge problem in the image followed by bone contour enhancement using multilevel LOCO (open-close and close-open) (Schulze and Pearce 1993) in scanned analog images (Bandyopadhyay et al. 2016a).

An entropy image is generated by separating the brighter bone regions from relatively darker bone areas in the original image, which is used as a preprocessed image for the task of fracture detection and classification in long bones (Bandyopadhyay et al. 2013, 2016b; c). Relaxed straightness and concavity index techniques of digital geometry are effectively utilized for long bone segmentation, rectification of contour imperfections, fracture detection and fracture localization (Bandyopadhyay et al. 2016b). A rapid change is observed in clock and anticlockwise directions in the chain code while traversing a contour nearby fractured region of long bones. Further, the fracture is categorized into simple and complex types based on diaphyseal, proximal and distal regions in the bone X-ray samples. They

have additionally developed a software tool to make the entire procedure interactive and automated for which the demonstration is available in the given link [<http://oldwww.iist.s.ac.in/it-abiswa.-research>]. Fracture points identified from X-ray images are successfully employed in categorizing fracture into upper (proximal), middle (diaphyseal) and lower (distal) regions of long bones (Bandyopadhyay et al. 2013). The classification of six long bones named as humerus, radius, ulna, femur, tibia, and fibula are based upon the type of fracture and its location in the X-ray image. Müller AO classification guidelines (Müller et al. 1990) are used for the first time in the classification of long bone fractures by utilizing digital geometry based techniques that appeared in their prior works (Bandyopadhyay et al. 2016a, b, c). However, the line of fracture and its location identified in (Bandyopadhyay et al. 2016a, b, c) are prone to certain inaccuracies when a closer view of fractured regions are observed in X-ray images. The authors have introduced a new concept based on relative concavity analysis to overcome this problem (Bandyopadhyay et al. 2013). Fracture is localized by integrating fracture points and chain-code representing digital-geometric concept such as concavity points and relative concavity. A software tool based on MATLAB is developed for classifying fracture into three major classes namely simple, complex and greenstick. 100 test images are taken as input out of which the system correctly classifies 97 and 92 images in the first and second level respectively (Table 7).

7 Performance evaluation

After successfully applying pre-processing, feature engineering, and feature selection approaches, followed by implementing a model and getting some output in forms of a probability or a class, the next step is to find out how effective is the model based on some metric using test datasets. Different performance metrics have been used to evaluate machine-learning algorithms for fracture detection and classification. It includes log-loss, accuracy, sensitivity, specificity, precision, recall, F-measure, AUC, and ROC etc. It is extremely important to estimate the performance of the classifier by finding out the wrong predictions on the test dataset. The major reasons for evaluating the predicting capability of any classification model is 1) to estimate the generalized performance, 2) to increase the predictive capability of a classifier by refining the model parameters and selecting the best performing model from the algorithm's hypothesis space. Several standard performance measures applied in literatures from year 2003 to year 2018 are discussed in this section.

7.1 Confusion matrix

It is one of the simplest and intuitive metrics for determining the correctness and accuracy of a classifier, used in binary or multi-class classification problem. It is a table, which is used to describe the performance of a classifier on a test set for which the true values are known. Various machine-learning algorithms are analyzed to predict the presence or absence of fracture in human bones. Figure 7 represents a confusion matrix where the row corresponds to a value predicted by the machine-learning algorithm and column indicates the known truth. True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN) values of confusion matrix are defined below:

Table 7 The table demonstrates relevant review findings of multi-class fracture detection and classification algorithms

Bandyopadhyay et al. (2013)	Long bones	The fracture is classified into simple, complex and greenstick types where a simple fracture is classified into different subgroups Fracture is localized by integrating fracture points and digital geometric concept such as concavity points and relative concavity Entropy based segmentation and Adaptive thresholding approaches provide better results in precision-recall measure as compared to k-means clustering for contour segmentation in their private test datasets
Bandyopadhyay et al. (2016b)	Long bones	A novel approach based on Relaxed Digital Straight-line Segment (RDSS) is proposed for correcting any false discontinuity that exists in bone contour Fracture is categorized into simple and complex types based on the severity of the injury and the location of fracture in the bone by using fracture points identified from X-ray images
Bandyopadhyay et al. (2016c)	Long bones	Adaptive thresholding approach is used to generate bone-contour region by segmenting bone regions from its surrounding fleshy areas in the long bone X-ray images The property of concavity index is utilized to detect the presence of fracture in long-bones Fracture localization is performed by finding out the exact location and possible fracture points once the fracture is successfully detected
Myint et al. (2018)	Long bones	Bone regions of gray-scaled tibia image are enhanced using USM (Unsharp Mask Filter) Harris corner detection algorithm is used to detect break points of leg bone in the processed image DT and KNN classifiers are used for fracture detection and classification respectively
Linda and Jiji (2018)	Long bones	Complex fractures from long bones are identified using linear structuring elements in DICOM CT images Hidden Markov Random Field-Expectation Maximization (HMR-EM) (Zhang et al. 2001) and adaptive thresholding techniques (Singh et al. 2012) are used for segmenting bone regions from surrounding areas Template-matching technique (Jurie and Dhome 2001) is used to separate fractured area from the image
Chung et al. (2018)	Humerus	The DCNN model is used to classify humerus bone fracture into greater tuberosity, surgical neck, 3-part, and 4-part classes based on Neer (Neer 1970) classification Model is pre-trained using ResNet-152 (https://github.com/kaimingHe/deep-residual-networks) on non-medical images and fine-tuned using Inception v3 on proximal humerus dataset

Fig. 7 Confusion matrix:
Columns correspond to a value predicted by the classification algorithm and row indicates the actual value or known truth

		Predicted	
		Frature not detected	Fracture detected
Actual	Fracture not detected	TN	FP
	Fracture detected	FN	TP

1. True positive: Patients have fracture and correctly identified as “fractured” by the algorithm (TP).
2. True Negative: Patients does not have fracture and correctly identified as “healthy” by the algorithm (TN).
3. False positive: Patients does not have fracture but incorrectly identified as “fracture” by the algorithm (FP).
4. False Negative: Patients do have fracture but incorrectly identified as “healthy” by the algorithm (FN).
5. Accuracy: It is the total number of correct predictions (both fracture and healthy) divided by total number of samples in the dataset.

7.2 Accuracy paradox in classification problem

Choosing the right metrics for classification tasks is extremely important in accessing a model. Accuracy alone is not reliable to select a classifier due to classification imbalance problem where one category represents the overwhelming majority of the data samples (Brownlee 2014). Another imbalanced classification problem occurs when the rate of disease in the public is very low. Consider a classification task where out of 150 patients, 25 patients are diagnosed with fracture and 125 patients are healthy, known as actual truth-value. Following two problems may exists when accuracy alone is used for prediction task—(a) all patients detected healthy: a model that only predicted healthy cases in patients achieves an accuracy of $(125/150)*100$ or 83.3%. This represents high accuracy and if it is used alone for decision support system to inform doctors, it would wrongly inform 25 patients of no occurrence of fracture (high False Negatives) making it a terrible model. (b) All patients detected with fracture: a model that only predicted the presence of fracture in patients achieves an accuracy of $(25/150)*100$ or 16.6%. If it is used alone for decision support system to inform doctors, it would wrongly inform 125 patients of the occurrence of fracture (high False Positives) making it a terrible model. Recall and precision are other performance matrix used for eliminating this accuracy paradox in classification problem (Brownlee 2014).

Recall/Sensitivity

It is the ability of a classifier to find all the relevant cases within a dataset.

$$Recall = \frac{TP}{TP + FN} \tag{1}$$

Precision

It measures the quality of our predictions only based on what our predictor claims to be positive.

$$\text{Precision} = \frac{TP}{TP + FP} \quad (2)$$

Specificity

It is total number of true negative assessments divided by number of all negative assessments.

$$\text{Specificity} = \frac{TN}{TN + FP} \quad (3)$$

Sensitivity measures the ability of the model to predict fractured cases and specificity measures the ability of the model to predict healthy cases.

F1 score

While recall measures the ability to find all relevant instances in a dataset, precision expresses the proportion of the data points the model says relevant actually were relevant. We can maximize either precision or recall at the expense of the other metric. For example, during preliminary fracture detection phase, we would want to correctly detect fractures in all the patients (high recall) and can accept a low precision rate if the cost of the follow-up examination is not significant. However, we may also find an optimal blend of precision and recall using F1 score that is the combination of recall and precision (Koehrsen 2018).

$$P_e = \frac{\left(\frac{SC_1 * SR_1}{T}\right) + \left(\frac{SC_2 * SR_2}{T}\right)}{T} \quad (4)$$

Receiver operating characteristics curve (ROC)

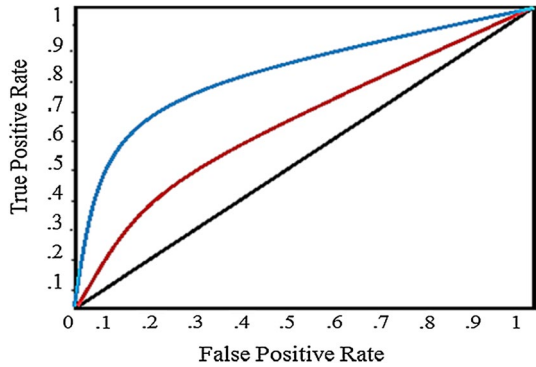
It predicts the probability of a binary classifier. It is the plot of false positive rate (x-axis), which is the same as precision and true positive rate (Y-axis), which is the same as recall. Designing confusion matrices for all possible threshold values is a challenging task, ROC graphs provides a simpler way to summarize all the information. The shape of the ROC curve contains a lot of information (Fig. 8):

1. A point at (1, 1) means that our model correctly classifies all fractured samples but incorrectly classifies all samples that were not fractured.
2. Diagonal line where (True Positive Rate = False Positive Rate), shows that the proportion of correctly classified fractured samples is similar to proportion of incorrectly classified healthy samples.
3. A point at (0, 0) results in zero True Positives and zero False Negatives.

AUROC (Area under the receiver operating characteristics)

Area under curve (AUC–ROC) is a performance measure for classification problems at various thresholds settings. ROC is a probability curve and AUC is the capability of the model in making a distinction between various classes. AUC makes it easier to compare one ROC curve to another, the more the area the better is the performance. Although ROC graphs are drawn using true positive rates and false positives rates, other matrices perform the same task. For example, false positive rates can be replaced with precision. Almost all literature surveyed here used confusion matrix based performance matrix for analyzing the classifier's performance. Table 8 depicts comparative analysis of the model proposed by authors against radiologist's interpretation on the dataset where the performance of the radiologist is assumed to be 100% on almost all the

Fig. 8 Receiver Operating Characteristics Curve (ROC) (Stephanie 2016)



comparisons. Researchers have shown the performance of the proposed model in their private datasets due to the unavailability of the standard dataset.

Cohen’s kappa statistic (k)

Many a times collecting and interpreting research or laboratory data such as collecting and labeling X-Ray images rely on multiple experts or clinicians in healthcare industries. Therefore, it is extremely important to have an agreement among multiple observers working on the interpretations of these samples. One such measurement of calculating the agreement and disagreement is Cohen’s kappa statistic (k). It gives a quantitative measure of the agreement on a situation where two or more independent observers are evaluating the same process (Sim and Wright 2005). For example: two radiologists are interpreting 100 X-ray images for the presence and absence of fracture in a patient simultaneously and mutually agree that fracture is diagnosed in 60 cases and not diagnosed in 15 cases, offering 75% agreement in total. This total observed agreement is calculated by adding diagonal values as shown in Fig. 9.

$$P_0 = \frac{C_1R_1 + C_2R_2}{T} \tag{5}$$

Here, T is the total number of observations, C and R represents column and row respectively.

However, random evaluation of patients by observers would sometimes lead to agreement just by chance and this quantity is measured by using expected or chance agreement P_c or P_e . It is calculated by multiplying positive answers of observer-1 and observer-2 and adding them to the negative answers of observer-1 and observer-2, all divided by total number of observed samples.

$$P_e = \frac{\left(\frac{SC_1 * SR_1}{T}\right) + \left(\frac{SC_2 * SR_2}{T}\right)}{T} \tag{6}$$

Here, S represents summation and P_e is evaluated as .6.

Now, we finally calculate kappa volume using the given formula:

$$k = \frac{P_0 - P_c}{1 - P_c} \tag{7}$$

Table 8 The results of classification model adopted in the literature reviewed are shown in the table

Author	Bone type and number of Test images	No of classes	Classifier selected for training and testing the model	A	S	R	P	F	AUC	CK
Tian et al. (2003)	Femur-320	2 (Healthy and fractured)	NSA (threshold of 116 degree)	–	.97	.45	–	–	–	–
Yap et al. (2004)	Femur-108	2 (Healthy and fractured)	1 of 3 combinations (Table 4)	.926	–	.846	–	–	–	–
			2 of 3 combinations (Table 4)	.954	–	.615	–	–	–	–
Lim et al. (2004)	Femur-108 Wrist-74	2 (Healthy and fractured)	Femur-2 of 6 combinations (Table 4)	.982	–	.92	–	–	–	–
			Wrist-SVM + MRF	.826	–	.77	–	–	–	–
Lai et al. (2005)	Femur-108 Wrist-74	2 (Healthy and fractured)	Femur-Gini SVM + MRF	.981	–	1	–	–	–	–
			Wrist-max, min rule (Table 4)	.959	–	.957	–	–	–	–
He et al. (2007)	Femur-90	2 (Healthy and fractured)	IGD + Hierarchical SVM	.933	–	.857	–	–	–	–
Chai et al. (2011)	Femur- <=30	2 (Healthy and fractured)	GLCM based classifier	.866	.933	.80	–	–	–	–
Mahendran and Baboo (2011b)	Tibia-1000	2 (Healthy and fractured)	2 of 4 combinations (Table 4)	–	.984	.912	–	–	–	–
Umadevi and Geethalakshmi (2012)	Tibia	2 (Healthy and fractured)	Texture features + shape features + ensemble (BPNN + SVM + KNN)	.975	–	.964	.98	.987	–	–
Al-Ayyoub and Al-Zghool (2013)	Long bones	5 (Healthy, Greenstick, Spiral, Comminuted, Transverse)	SVM (tenfolds)	–	–	.855	.85	.854	.893	–
Al-Ayyoub et al. (2013)	Hand bones	2 (Healthy and fractured)	Boosting and then bagging on BN	–	–	.92	.92	.98	–	–

Table 8 (continued)

Author	Bone type and number of Test images	No of classes	Classifier selected for training and testing the model	A	S	R	P	F	AUC	CK
Cao et al. (2015)	Multiple bones	2 (Healthy and fractured)	Stacked ensemble of DT is used for fracture detection	-	-	.812	-	-	-	-
Dimililer (2017)	Multiple bones-100	2 (Healthy and fractured)	BPNN	-	-	-	-	-	.95	-
Gale et al. (2017)	Pelvis-3354	2 (Healthy and fractured)	DenseNet (high precision)	.99	-	.92	.99	.95	-	-
			DenseNet (high sensitivity)	.99	-	.95	.92	.94	-	-
			Est. Radiologist Performance ($CI_{95\%}$)	.99	-	.97	.93	.95	-	-
Chung et al. (2018)	Humerus-181 (approx)	2 (Healthy and fractured) for fracture detection	Model is pre-trained using ResNet-152 on non-medical images and fine-tuned using Inception v3 on proximal humerus dataset	-	.97	.99	-	-	.99	-
		4 (greater tuberosity, surgical neck, 3 part, 4-part, for fracture classification)					A-.94 B-.85 C-.83 D-.85	A-.97 B-.90 C-.88 D-.93	A-.98, B-.94, C-.90, D-.94	
Myint et al. (2018)	Tibia-40 images for fracture detection	2 (healthy and fractured) for fracture detection	Fracture is detected using DT and is classified into three categories using KNN followed by fracture localization	.92	1	.87	1	-	-	.83 for fracture classification
	Tibia-52 images for fracture classification	3 (transverse, oblique, comminuted) for fracture classification								
Kim and MacKinnon (2018)	Wrist-100	2 (Healthy and fractured)	Model is pre-trained using Inception v3 network on non-medical images and fine-tuned using Inception v3 on wrist images	-	-	-	-	-	-	.954

Table 8 (continued)

Author	Bone type and number of Test images	No of classes	Classifier selected for training and testing the model	A	S	R	P	F	AUC	CK
Lindsey et al. (2018)	Wrist-4900 Set-1 (3500) Set-2 (1400)	2 (Healthy and fractured)	Model is pre-trained using 100, 855 radiographs of all body parts except wrist images and fine-tuned using DCNN on 31,490 wrist images	-	-	-	-	-	-	.96 (set-1) .97 (set-2)

Researchers in the published articles are using their private datasets, labeled with the help of expert radiologists. The interpretations given by examiner in the original X-ray samples are compared against the results of the adopted classifier. This comparative analysis is presented in the form of a single table where the performance of all the chosen model is shown along with total number of test samples used. The classifiers are compared against radiologist's performance where the performance of the radiologists is assumed one (if not stated explicitly)

A accuracy, *S* sensitivity, *R* recall, *P* precision, *F* F-measure, *CK* Cohen's kappa statistics, *NSA* neck-shaft angle, *NB* Naive Bayes, *SVM* support vector machine, *DCNN* deep convolution neural network, *BPNN* back propagation neural network, *DT* decision tree, *NV* neural network, *GLCM* gray level co-occurrence matrix, *DCNN* deep convolution neural network, *KNN* K-nearest neighbor, *ANN* artificial neural network

Fig. 9 Cohen kappa statistics for 100 radiographs simultaneously interpreted by two observers

		Observer-1		
		Has fracture	Does not have fracture	
Observer-2	Has fracture	60	10	70
	Does not have fracture	15	15	30
		75	25	

Here, $K = .375$.

According to the commonly used scale presented in Table 9, both the radiologists have a minimal agreement ($k = .375$) in diagnosing fractures for the above example.

Sensitivity of 45.45% is observed when Neck-Shaft Angle (NSA) is used as the only feature in the first ever approach of fracture detection in 2003 (Tian et al. 2003). The poor fracture detection rate is obtained because NSA alone could not detect significant changes observed during the event of fracture. This problem is overcome by using several feature-classifier combinations and ensemble based approaches and resulted in better performance (Yap et al. 2004; Lim et al. 2004; Lai et al. 2005). SVM, NN, DT and NB algorithms are used independently in earlier approaches such as (Lim et al. 2004), which is later combined by using boosting and bagging techniques and successfully achieved the accuracy of 91.8% on 98 hand X-ray images (Al-Ayyoub et al. 2013). Tibia fractures are detected, classified and localized in human bones using Decision Tree and K-Nearest Neighbor algorithms (Myint et al. 2018). This is the first work proposed which performs all three activities after successfully applying Unsharp Masking (USM) in the input image. However, they have used only 40 images to train the model including all class types, which is not sufficient for practical adaptability of the model.

8 Discussions

Fracture detection in radiographs has significant clinical importance. There was an estimated backlog of 200,000 plain radiographs and 12,000 cross-sectional studies in the year 2012 (Cliffe et al. 2016). These figures demand desperate improvements in reporting efficiency and workflow management to mitigate harm caused to patients due to delayed or missed diagnosis. Automatic abnormality detection or localization techniques could help radiologists fight fatigue largely. Radiologists can further prioritize diagnosis and treatment according to the abnormality type detected by the automated system. Most often, the first point of contact for any patient in case of fracture is non-orthopedic surgeons

Table 9 Cohen’s kappa interpretation (McHugh 2012)

Value of kappa	Level of agreement	% of data that are reliable
0–.20	None	0–4
.21–.39	Minimal	4–15
.40–.59	Weak	15–35
.60–.79	Moderate	35–63
.80–.90	Strong	64–81
Above .90	Almost perfect	82–100

or inexperienced clinicians who often lack expertise in detecting fractures (Chung et al. 2018). Hence, it is quite usual that fractures are misdiagnosed during interpretation of X-Ray images. Radiologists or clinicians manually examine the X-rays for the existence of fractures and its type. The interpretation and classification of radiographic images is a time-consuming and intense process, which could be solved using automated fracture classification models.

Majority of the published work of fracture detection and classification is focused on either single anatomical region or a single type of fracture in different anatomical regions. An ideal model would be the one that is able to detect different type of fractures in various anatomical regions. Collecting hundreds of thousands of radiographs, providing correct labels to these X-ray images and feeding enough training data to the models is an arduous task in medical imaging but predicting the fracture using a trained model take less than a second in a modern computer. The major challenge in achieving multi-label classification on different anatomical regions is the availability of the labeled dataset. A promising alternative to solve this problem is to fine-tune a CNN, which is pre-trained on a different network (Birks et al. 2016). Researchers can use these large pre-trained models to obtain highly sophisticated and powerful set of features needed for the domain of interest. Instead of pre-training the model on millions of non-radiology images, the model can be trained on multiple bone X-ray images such as ankle, knee, neck, hip etc. This way we could better initialize model parameters, which can be later used for training the required X-ray images (Lindsey et al. 2018). This approach is similar to transfer learning with the slight difference that we pre-train the model on several types of bone X-ray images instead of non-medical images. This system can also be useful in emergency conditions where such models may provide assistance to the radiologists for diagnosing fracture. The more complex a fracture type is, easier it will become for CNN to classify it as compared to human, as humans face difficulty in interpreting complex bone structures. A machine such as CNN remains consistent in its performance and does not lack concentration while interpreting several complex X-ray structures and it can even surpass human level performance when provided with sufficient amount of training data. In addition, it is a human tendency to predict correct output in familiar shapes as opposed to the one whose fracture configuration is less familiar. Hence, a CNN can potentially be trained with humongous training data with all possible cases ranging from simpler fracture structure to the most complex ones which is more than any orthopedics will even encounter in his/her life.

The decision-making in clinical examination is a complex process that is based on accurate evaluation of clinical findings using diagnostic tests and reference standard data. A gold standard study may refer to an experimental model that has been thoroughly tested and has a reputation in the field as a reliable method (Cardoso et al. 2014). Whenever a classifier is compared against actual or known data provided by radiologist, the performance of radiologist is considered as 100% accurate, which means that the radiologist involved in interpreting hundreds of images correctly detects and classifies fracture in all samples with zero error rate. However, there may be a chance that the radiologist involved in the interpretation of images could not provide the best set of results or do not agree with each other in case of multiple observers. It is extremely important to have an agreement among multiple observers dedicated to interpret the sample images. One such measurement of calculating the agreement and disagreement is Cohen's kappa statistic (k). It gives us a quantitative measure of the agreement in situation where two or more independent observers are evaluating the same thing. It is only in 2017 when a gold standard is created for abnormality detection in X-ray samples including multiple anatomical regions using Cohen's kappa statistic (Rajpurkar et al. 2017). Gold standard is created by randomly choosing three

expert radiologists out of six board certified radiologists and the label is selected based on majority voting. Cohen's kappa statistic is used to compare both radiologists and proposed model's performance with the gold standard (Rajpurkar et al. 2017). Likewise, a gold standard could be created for a labeled fracture dataset for the purpose of fracture detection and classification that is globally available and acceptable. This would enable researchers to compare their proposed model with the Gold standard for better performance analysis. Kappa statistics is used for classifying tibia fractures into four categories with the accuracy of 83% but no such gold standard is observed here for accuracy measurement while calculating kappa statistic (Myint et al. 2018). A similar approach is observed where performance of two models and performance of a group of general orthopedics, general physicians and specialized orthopedics are evaluated on test set for diagnosing and classifying humerus fractures (Chai et al. 2011). The adopted CNN based model outperformed human groups in detecting and classifying fractures except for one category of fracture.

Overfitting issues can be mitigated by including only relevant features necessary for the model. Cropping our region of interest from the input image and exploring several network architectures including dropout could be a solution to overfitting problem. An additional label termed as "indecisive" could be included along with "normal" or "abnormal" fracture classes in the case where radiologist experts indicate indecisiveness in detecting the fracture. This way the examiner would review it for the second time (Kim and MacKinnon 2018). If any result is 'unknown' or 'uncertain' during the examination of fracture then such cases must either be excluded or assigned a new label such as "uncertain" before including them in training data (Gale et al. 2017). Assigning them 'uncertain' would help radiologists and researchers to identify that the patient has to undergo further level of treatment and can go for CT or MRI, as to accurately identify the fracture or fracture type.

9 Conclusion

Various fracture detection and classification approaches have been proposed which includes data preparation, image pre-processing stage followed by feature extraction and classification. The interpretation and classification of radiographic images by expert radiologists is a time-consuming and intense process, which could be solved using automated fracture classification models. The major obstacle in the development of a high performance classification model is the lack of labeled training dataset as pointed by many researchers cited in this survey. We have tried to provide a number of ideas and perspectives to explore that could help in developing an ideal model, which can detect different type of fractures in various anatomical regions.

Funding This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

References

- Al-Ayyoub M, Hmeidi I, Rababah H (2013) Detecting hand bone fractures in X-ray images. *J Multimed Process Technol (JMPT)* 4(3):155–168. <https://doi.org/10.13140/RG.2.1.2645.8327>
- Al-Ayyoub M, Al-Zghool D (2013) Determining the type of long bone fractures in X-ray images. *WSEAS Trans Inf Sci Appl* 10(6):261–270

- Andreas T, Jahrer M, Bell RM, Park F (2009) The BigChaos solution to the Netflix Grand Prize
- Antipov G, Berrani S, Dugelay J (2016) Minimalistic CNN-based ensemble model for gender prediction from face images. *Pattern Recognit Lett*. <https://doi.org/10.1016/j.patrec.2015.11.011>
- Bandyopadhyay O, Biswas A, Chanda B, Bhattacharya BB (2013) Bone contour tracing in digital X-ray images based on adaptive thresholding. In: *Proceedings, 5th international conference on pattern recognition and machine intelligence*, volume LNCS 8251, pp 465–473. https://doi.org/10.1007/978-3-642-45062-4_64
- Bandyopadhyay O, Biswas A, Bhattacharya BB (2016a) Automatic segmentation of bones in X-ray images based on entropy measure. *Int J Image Graph*. <https://doi.org/10.1142/s0219467816500017>
- Bandyopadhyay O, Biswas A, Bhattacharya BB (2016b) Long-bone fracture detection in digital X-ray images based on digital-geometric techniques. *Comput Methods Programs Biomed* 123:2–14. https://doi.org/10.1007/978-3-319-07148-0_19
- Bandyopadhyay O, Biswas A, Bhattacharya BB (2016c) Classification of long-bone fractures based on digital-geometric analysis of X-ray images. *Pattern Recognit Image Anal* 26(4):742–757. <https://doi.org/10.1134/S1054661816040027>
- Birks JS, L'Allier PL, Gingras MA (2016) Convolutional neural networks for medical image analysis: full training or fine tuning? *IEEE Trans Med Imaging* 35(5):1299–1313. <https://doi.org/10.1109/TMI.2016.2535302>
- Brownlee J (2014) Classification accuracy is not enough: more performance measures you can use. <https://machinelearningmastery.com/classification-accuracy-is-not-enough-more-performance-measures-you-can-use>. Accessed 22 Apr 2019
- Cao Y, Wang H, Moradi M, Prasanna P, Syeda-Mahmood TF (2015) Fracture detection in X-ray images through stacked random forests feature fusion. In: *Proceedings—international symposium on biomedical imaging*, pp 801–805. <https://doi.org/10.1109/isbi.2015.7163993>
- Cardoso JR et al (2014) What is gold standard and what is ground truth? *Dental Press J Orthod* 19(5):27–30. <https://doi.org/10.1590/2176-9451.19.5.027-030.ebo>
- Chai HY, Wee LK, Swee TT, Hussain S (2011) Gray-level co-occurrence matrix bone fracture detection. *WSEAS Trans Syst* 10(1):7–16. <https://doi.org/10.3844/ajassp.2011.26.32>
- Chung SW, Han SS, Lee JW (2018) Automated detection and classification of the proximal humerus fracture by using deep learning algorithm. *Acta Orthop* 89(4):468–473. <https://doi.org/10.1080/17453674.2018.1453714>
- Cliffe H et al (2016) Summary of The Royal College of Radiologists' (RCR) reporting backlog surveys and assessment of potential causes and solutions. *Clin Radiol* 71:1–30. <https://doi.org/10.1016/j.crad.2016.06.060>
- Deng J, Dong W, Socher R, Li L, Li K, Fei-fei L (2009) ImageNet: a large-scale hierarchical image database. In: *IEEE conference on computer vision and pattern recognition*, 20–25 June 2009, Miami, FL, USA. <https://doi.org/10.1109/cvpr.2009.5206848>
- Dimililer K (2017) IBFDS: intelligent bone fracture detection system. *Procedia Comput Sci* 120:260–267. <https://doi.org/10.1016/j.procs.2017.11.237>
- Ding C, Tao D (2017) Trunk-Branch Ensemble Convolutional Neural Networks for Video-based Face Recognition. *IEEE Trans Pattern Anal Mach Intell*. <https://doi.org/10.1109/TPAMI.2017.2700390>
- Dormehl L (2019) What is an artificial neural network? Here's everything you need to know. <https://www.digitaltrends.com/cool-tech/what-is-an-artificial-neural-network>. Accessed 13 Feb 2019
- Gale W, Oakden-Rayner L, Carneiro G, Bradley AP, Palmer LJ (2017) Detecting hip fractures with radiologist-level performance using deep neural networks. *Comput Vis Pattern Recognit*. [arXiv:1711.06504](https://arxiv.org/abs/1711.06504)
- Greenspan H, Ginneken BV, Summers RM (2016) Guest editorial deep learning in medical imaging: overview and future promise of an exciting new technique. *IEEE Trans Med Imaging* 35(5):1153–1159. <https://doi.org/10.1109/TMI.2016.2553401>
- Guly H (2001) Diagnostic errors in an accident and emergency department. *Emerg Med J (EMJ)* 18(4):263–269. <https://doi.org/10.1136/emj.18.4.263>
- Haralick RM, Shanmugam K (1973) Textural features for image classification. *IEEE Trans Syst Man Cybern* 3(6):610–621. <https://doi.org/10.1109/TSMC.1973.4309314>
- Hari CV, Jojish JV, Gopi S, Felix VP (2010) Mid-point Hough transform: a fast line detection method. In: *Annual IEEE India conference*, 18–20 Dec. 2009. <https://doi.org/10.1109/indcon.2009.540945>
- He JC, Leou WK, Howe TS (2007) Hierarchical classifiers for detection of fractures in X-ray images. In: *Kropatsch WG, Kampel M, Hanbury A (eds) Computer analysis of images and patterns*. CAIP 2007. *Lecture Notes in Computer Science*, vol 4673. Springer, Berlin. https://doi.org/10.1007/978-3-540-74272-2_119

- How to organize data labeling for machine learning: approaches and tools (2018). <https://www.altexsoft.com/blog/datascience/how-to-organize-data-labeling-for-machine-learning-approaches-and-tools>. Accessed 26 Jan 2019
- Huang G, Liu Z, Maaten LVD, Weinberger KQ (2017) Densely connected convolutional networks. In: IEEE conference on computer vision and pattern recognition (CVPR), 21–26 July 2017, Honolulu, HI, USA, IEEE, pp 2261–2269. <http://doi.org/10.1109/CVPR.2017.243>
- Jacob NE, Wyawahare MV (2013) Survey of bone fracture detection techniques. *Int J Comput Appl* 71(17):31–34. <https://doi.org/10.5120/12452-9342>
- Jurie F, Dhome M (2001) A simple and efficient template matching algorithm. In: Eighth IEEE international conference on computer vision, 7–4 July 2001, Vancouver, BC, Canada, IEEE, pp 544–549. <http://doi.org/10.1109/ICCV.2001.937673>
- Khashman A, Dimililer K (2008) Image compression using neural networks and Haar wavelet. *WSEAS Trans Signal Process* 4(5):330–339
- Khatik I (2017) A study of various bone fracture detection techniques. *Int J Eng Comput Sci* 6(5):21418–21423. <https://doi.org/10.18535/ijecs/v6i5.38>
- Kim KG (2016) Book review: Deep learning, healthcare informatics. *Research* 22(4):351–354. <https://doi.org/10.4258/hir.2016.22.4.351>
- Kim DH, MacKinnon T (2018) Artificial intelligence in fracture detection: transfer learning from deep convolutional neural networks. *Clin Radiol* 73(5):439–445. <https://doi.org/10.1016/j.crad.2017.11.015>
- Kingma D, Ba J (2015) A method for stochastic optimization. In: International conference on learning representations, San Diego
- Kinnari P, Dangar Y (2017) A survey for fracture detection in hand bone X-ray images. *Int J Adv Res Innov Ideas Educ* 3(1):1529–1533. ISSN(O)-2395-4396
- Koehrsen W (2018) Beyond accuracy: precision and recall. <https://towardsdatascience.com/beyond-accuracy-precision-and-recall-3da06bea9f6c>. Accessed 9 Apr 2019
- Kohli M, Prevedello LM, Filice RW, Geis JR (2017) Implementing machine learning in radiology practice and research. *Am J Roentgenol* 208(4):754–760. <https://doi.org/10.2214/ajr.16.17224>
- Krizhevsky A, Sutskever I, Hinton GE (2017) ImageNet classification with deep convolutional neural networks. *Adv Neural Inf Process Syst* 60(6):84–90. <https://doi.org/10.1145/3065386>
- Krupinski EA, Berbaum KS, Caldwell RT, Schartz KM, Kim J (2010) Radiology workdays reduce detection and accommodation accuracy. *J Am Coll Radiol* 7(9):698–704. <https://doi.org/10.1016/j.jacr.2010.03.004>
- Kumar A et al (2017) An ensemble of fine-tuned convolutional neural networks for medical image classification. *IEEE J Biomed Health Inform* 21(1):31–40. <https://doi.org/10.1109/JBHI.2016.2635663>
- Kuncheva L, Whitaker E (2001) Ten measures of diversity in classifier ensembles: limits for two classifiers. In: Proceedings of IEEE workshop on intelligent sensor processing, Birmingham, vol 10, pp 1–10
- Laan MJVD, Polley EC, Hubbard AE (2007) Super Learner. <https://doi.org/10.2202/15446115.1309>
- Lai V, Lum F, Leow WK, Chen Y (2005) Combining classifiers for bone fracture detection in X-ray images. In: IEEE international conference on image processing 2005, Genova, Italy, IEEE. <https://doi.org/10.1109/icip.2005.1529959>
- Lakhani P, Sundaram B (2017) Deep learning at chest radiography: automated classification of pulmonary tuberculosis by using convolutional neural networks. *Radiology* 284(2):574–582
- Lampert CH, Blaschko MB, Hofmann T (2008) Beyond sliding windows: object localization by efficient subwindow search. In: IEEE conference on computer vision and pattern recognition, 23–28 June 2008, Anchorage, AK, USA. <https://doi.org/10.1109/cvpr.2008.4587586>
- LeCun Y et al (1998) Gradient based learning applied to document recognition. *Proc IEEE* 86(11):2278–2324. <https://doi.org/10.1109/5.72679>
- Lim SE, Xing Y, Chen Y, Leow WK, Howe TS, Png MA (2004) Detection of femur and radius fractures in X-ray images. In: Second international conference on advances in medical signal and information processing, pp 399–404
- Linda H, Jiji GW (2011) Crack detection in X-ray images using fuzzy index measure. *Appl Soft Comput* 11:3571–3579. <https://doi.org/10.1016/j.asoc.2011.01.029>
- Linda H, Jiji GW (2017) Hairline breakage detection in X-ray images using data fusion. *Multimedia Tools and Applications*. 77(6):1–16. <https://doi.org/10.1007/s11042-017-5286-5>
- Linda CH, Jiji GW (2018) Hierarchical approach to detect fractures in CT DICOM images, section C: computational intelligence, machine learning and data analytics. *Comput J* 61(7):1099–1108. <https://doi.org/10.1093/comjnl/bxy023>
- Lindsey R, Daluiski A, Chopra S et al (2018) Deep neural network improves fracture detection by clinicians. *Proc Natl Acad Sci* 115(45):11591–11596. <https://doi.org/10.1073/pnas.1806905115>

- Lowe DG (2004) Distinctive image features from scale-invariant keypoints David. *Int J Comput Vision* 60(2):91–110. <https://doi.org/10.1023/B:VISI.0000029664.99615.94>
- Mahendran SK, Baboo SS (2011a) Automatic fracture detection using classifiers: a review. *Int J Comput Sci Issues* 8(6):340–345
- Mahendran SK, Baboo SS (2011b) An enhanced Tibia fracture detection tool using image processing and classification fusion techniques in X-ray images. *Global J Comput Sci Technol* 11(14):23–28
- Mahmoodi S (2011) Anisotropic diffusion for noise removal of band pass signals. *Elsevier Signal Process* 91(5):1298–1307. <https://doi.org/10.1016/j.sigpro.2010.12.018>
- Mansoori MS, Ashtiyani M, Sarabadani H (2012) Automatic crack detection in eggshell based on SUSAN edge detector using fuzzy thresholding. *World Appl Sci J* 18(11):1602–1608. <https://doi.org/10.5539/was.v5n6p117>
- Mayne R (2013) How to identify fracture types. <https://allthingsaafs.com/tag/how-to-identify-fracture-types>. Accessed 7 Apr 2019
- McHugh ML (2012) Interrater reliability: the kappa statistic. *Biochem Med (Zagreb)* 22(3):276–282. <https://doi.org/10.11613/BM.2012.031>
- Mendrik M, Vonken EJ, Rutten A, Viergever MA, Ginneken BV (2009) Noise reduction in computed tomography scans using 3D anisotropic hybrid diffusion with continuous switch. *IEEE Trans Med Imaging* 28(10):1585–1594. <https://doi.org/10.1109/TMI.2009.2022368>
- Meyer P et al (2018) Survey on deep learning for radiotherapy. *Comput Biol Med* 98:126–146. <https://doi.org/10.1016/j.combiomed.2018.05.018O>
- Minetto R, Segundo MP, Sarkar S (2018) Hydra: an ensemble of convolutional neural networks for geospatial land classification. *Comput Vis Pattern Recognit*. [arXiv:1802.03518](https://arxiv.org/abs/1802.03518)
- Müller ME, Koch P, Nazarian S, Schatzker J (1990) The comprehensive classification of fractures of long bones. Springer, Berlin. 73(4):202. <https://doi.org/10.1007/978-3-642-61261-9>
- Myint WW, Tun KS, Tun HM (2018) Analysis on leg bone fracture detection and classification using X-ray. *Mach Learn Res* 3(3):49–59. <https://doi.org/10.11648/j.ml.20180303.11>
- Neer CS (1970) Displaced proximal humeral fractures. I. Classification and evaluation. *J Bone Joint Surg* 52(6):1077–1089
- Pisano ED et al (1998) Contrast limited adaptive histogram equalization image processing to improve the detection of simulated spiculations in dense mammograms. *J Digit Imag* 11(4):193–200. <https://doi.org/10.1007/BF03178082>
- Polikar R (2009) Ensemble learning. *Scholarpedia* 4(1):2776. <https://doi.org/10.4249/scholarpedia.2776>
- Prabhu (2018) Understanding of convolutional neural network (CNN)—deep learning. <https://medium.com/@RaghavPrabhu/understanding-of-convolutional-neural-network-cnn-deep-learning-99760835f148>. Accessed 7 Dec 2018
- Rajan J, Kaimal MR (2006) Image denoising using wavelet embedded anisotropic diffusion (WEAD). In: IET international conference on visual information engineering, pp 26–28. <https://doi.org/10.1049/cp:20060597>
- Rajpurkar P et al (2017) MURA: large dataset for abnormality detection. In: 1st conference on medical imaging with deep learning, MIDL 2018 (2017). [arXiv:1712.06957](https://arxiv.org/abs/1712.06957)
- Russakovsky O, Deng J et al (2015) ImageNet large scale visual recognition challenge. *Int J Comput Vis* 115(3):211–252. <https://doi.org/10.1007/s11263-015-0816-y>
- Sathik MM, Mehaboobathunnisa R, Thasneem AAH, Arumugam S (2015) Ray casting for 3D rendering—a review. *Int J Innov Eng Technol* 5:121–124
- Schulze MA, Pearce JA (1993) Linear combinations of morphological operators: the midrange, pseudomedian, and loco filters. In: Proceedings of IEEE international conference on acoustics, 27–30 April, Minneapolis, USA, IEEE, pp 1520–6149. <https://doi.org/10.1109/icas.1993.319746>
- Shah T (2017) About train, validation and test sets in machine learning. <https://tarangshah.com/blog/2017-12-03/train-validation-and-test-sets>. Accessed 27 Dec 2018
- Sharan V, Keshari N, Mondal T (2014) Biomedical image denoising and compression in wavelet using MATLAB. *Int J Innov Sci Mod Eng (IJISME)* 2(6):9–13
- Shen D, Wu G, Suk HI (2017) Deep learning in medical image analysis. *Annu Rev Biomed Eng* 19(1):221–248. <https://doi.org/10.1146/annurev-bioeng-071516-044442>
- Sim J, Wright CC (2005) The Kappa statistic in reliability studies: use, interpretation, and sample size requirements. *Phys Ther* 85(3):257–268. <https://doi.org/10.1093/ptj/85.3.257>
- Singh TR, Roy S, Singh OI, Sinam T, Singh KM (2012) A new local adaptive thresholding technique in binarization. *Int J Comput Sci* 8(6):271–277. [arXiv:1201.5227](https://arxiv.org/abs/1201.5227)
- Stec N, Arje D, Moody A, Krupinski E, Tyrrell P (2018) A systematic review of fatigue in radiology: is it a problem? *Am J Roentgenol* 210(4):799–806. <https://doi.org/10.2214/AJR.17.18613>

- Stephanie (2016) C-statistic: definition, examples, weighting and significance. <https://www.statisticshowto.datasciencecentral.com/c-statistic>. Accessed 12 Apr 2019
- Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z (2016) Rethinking the inception architecture for computer vision. In: IEEE conference on computer vision and pattern recognition (CVPR), IEEE, pp 2818–2826. <https://doi.org/10.1109/cvpr.2016.308>
- Teixeira L, Celes W, Gattass M (2008) Accelerated corner-detector algorithms. In: Proceedings of the British machine vision conference. <http://doi.org/10.5244/C.22.62>
- Tian D, Ha M (2004) Applications of wavelet transform in medical image processing. In: Proceedings of 2004 international conference on machine learning and cybernetics (IEEE Cat. No.04EX826), pp 26–29. <https://doi.org/10.1109/icmlc.2004.1382071>
- Tian TP et al (2003) Computing neck-shaft angle of femur for X-ray fracture detection. In: Petkov N, Westenberg MA (eds) Computer analysis of images and patterns. CAIP 2003. Lecture Notes in Computer Science, vol 2756. Springer, Berlin
- Umadevi N, Geethalakshmi SN (2012) Multiple classification system for fracture detection in human bone X-ray images. In: Third international conference on computing, communication and networking technologies (ICC-CNT'12), 26–28 July 2012, Coimbatore, India, IEEE. <https://doi.org/10.1109/iccct.2012.6395889>
- Waite S, Scott J, Gale B, Fuchs T, Kolla S, Reede D (2017) Interpretive error in radiology. *Am J Roentgenol* 208(4):739–749. <https://doi.org/10.2214/ajr.16.16963>
- Wang H et al (2017) Deep learning based ensemble approach for probabilistic wind power forecasting. *Appl Energy* 188:56–70. <https://doi.org/10.1016/j.apenergy.2016.11.111>
- Whang J, Baker SR, Patel R, Luk L, Castro A (2013) The causes of medical malpractice suits against radiologists in the United States. *Radiology* 266(2):548–554. <https://doi.org/10.1148/radiol.12111119>
- Wolpert DH (1992) Stacked generalization. *Neural Netw* 5(2):241–259. [https://doi.org/10.1016/S0893-6080\(05\)80023-1](https://doi.org/10.1016/S0893-6080(05)80023-1)
- World Health Organization (2019). https://www.who.int/nutrition/topics/5_population_nutrient/en/index25.html. Accessed 14 Mar 2019
- X-rays, CT Scans and MRIs- OrthoInfo-AAOS (2019). <https://orthoinfo.aaos.org/en/treatment/x-rays-ct-scans-and-mris>. Accessed 14 Jan 2019
- Yap DWH et al (2004) Detecting femur fractures by texture analysis of trabeculae. In: Proceedings—international conference on pattern recognition 3, pp 730–733. <http://doi.org/10.1109/ICPR.2004.1334632>
- Zhang Y, Brady M, Smith S (2001) Segmentation of brain MR images through a hidden Markov random field model and the expectation maximization algorithm. *IEEE Trans Med Imaging* 20:45–57. <https://doi.org/10.1109/42.906424>

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.