Novel transfer learning based bone fracture detection using radiographic images

Aneeza Alam[1], Ahmad Sami Al-Shamayleh[2], Nisrean Thalji[3], Ali Raza[4], Edgar Anibal Morales Barajas[5,6,7], Ernesto Bautista Thompson[5,6,8], Isabel de la Torre Diez[9] and Imran Ashraf[10]*

Abstract

A bone fracture is a medical condition characterized by a partial or complete break in the continuity of the bone. Fractures are primarily caused by injuries and accidents, affecting millions of people worldwide. The healing process for a fracture can take anywhere from one month to one year, leading to significant economic and psychological challenges for patients. The detection of bone fractures is crucial, and radiographic images are often relied on for accurate assessment. An efficient neural network method is essential for the early detection and timely treatment of fractures. In this study, we propose a novel transfer learning-based approach called MobLG-Net for feature engineering purposes. Initially, the spatial features are extracted from bone X-ray images using a transfer model, MobileNet, and then input into a tree-based light gradient boosting machine (LGBM) model for the generation of class probability features. Several machine learning (ML) techniques are applied to the subsets of newly generated transfer features to compare the results. K-nearest neighbor (KNN), LGBM, logistic regression (LR), and random forest (RF) are implemented using the novel features with optimized hyperparameters. The LGBM and LR models trained on proposed MobLG-Net (MobileNet-LGBM) based features outperformed others, achieving an accuracy of 99% in predicting bone fractures. A cross-validation mechanism is used to evaluate the performance of each model. The proposed study can improve the detection of bone fractures using X-ray images.

Keywords Radiographic images, Bone fractures, Deep learning, Transfer learning, Image processing

*Correspondence:
Imran Ashraf
ashrafimran@live.com
[1] Faculty of Computer Science and Information Technology, Khwaja Fareed University of Engineering & Information Technology, Rahim Yar Khan, Pakistan
[2] Department of Data Science and Artificial Intelligence, Faculty of Information Technology, Al-Ahliyya Amman University, Amman 19328, Jordan
[3] Faculty of Computer Studies, Arab Open University, Amman, Jordan
[4] Department of Software Engineering, University of Lahore, Lahore 54000, Pakistan
[5] Universidad Europea del Atlantico, Santander 39011, Spain
[6] Universidad Internacional Iberoamericana, Campeche 24560, Mexico
[7] Universidad de La Romana, La Romana, República Dominicana
[8] Universidade Internacional do Cuanza, Cuito, Bie, Angola
[9] Department of Signal Theory, Communications and Telematics Engineering, Unviersity of Valladolid, Paseo de Belen, 15, 47011 Valladolid, Spain
[10] Department of Information and Communication Engineering, Yeungnam University, Gyeongsan 38541, Republic of Korea

Alam *et al. BMC Medical Imaging*   (2025) 25:5

Page 2 of 16

Introduction

The human body contains 206 bones of different shapes, sizes, and strengths. Detecting internal injuries, especially minor ones, is a challenging task for orthopedics. Machine learning (ML) is becoming prominent in analyzing radiographic images to detect bone injuries. The features extracted from X-ray images help identify the fracture and healthy bones [1]. These features help doctors diagnose and treat ailments in a timely manner.

In 2019, the Global Burden of Diseases(GBD) reported 178 million new fractures only, an average increase of 33.4% more than in 1990. There were 445 million cases of acute or prevalent cases of fractures [2]. Fractures impact patients' quality of life and put an economic burden on the health system. The average global cost for one year of health and social care for hip fracture is $43,669 [3]. These high costs of treatment emphasize the importance of timely recovery, as prolonged diseases have a significant financial burden on patients.

Traditionally, X-ray reports are examined by trained medical practitioners, mostly radiologists [4]. This way of fracture diagnosis is prone to human error and inconsistencies. There is a viable chance that a fracture may be overlooked. The traditional method is time-consuming, and delayed diagnosis may lead to improper handling. The accuracy of these diagnoses depends entirely on the medical professional's experience, and years of practice are needed for accurate interpretation. Also, one person can attend to a limited number of patients, delaying the diagnosis and treatment process. So, there is a need for a more reliable and quicker way of radiographic fracture identification.

ML in medical fields has revolutionized healthcare by improving accuracy, treatment, and personal care [5, 6]. Image processing and ML techniques have eased disease and drug prediction, improving clinical predictions [7]. Using ML techniques promises a bright future with better diagnoses, personalized treatments, and improved patient outcomes [8]. Recently, transfer learning-based feature engineering has become popular for improving the prediction accuracy of image datasets [9]. Pre-trained neural networks are used in transfer learning to get the benefits of the best features from the image dataset.

The main contributions of this study using transfer learning-based feature extraction are given as:

- A novel transfer learning-based MobLG-Net feature extraction approach is introduced, which gives the most relevant features that are further used for training ML models.
- Five ML models are applied to both spatial and MobLG-Net features to get an unbiased comparison of both techniques.

- Hyperparameter tuning is done to get the most optimized results from ML models from both sets of features.
- A comparison of applied models and state-of-the-art techniques is presented to give a better view of how important the research findings are.

The remaining paper is organized as follows: Literature review section covers the previous investigations done to detect fractures in bones using X-ray images. The proposed novel feature extraction and applied machine-learning models are discussed in Proposed methodology section. Results and discussions section has detailed performance metrics of the applied methods on both sets of features, along with a comparison with previous studies. The conclusion and the prospects of the study are discussed in the final section.

Literature review

Artificial intelligence has been used in the past to detect fractures using clinical and publicly available data. The literature contains fracture detection studies using artificial intelligence [10] on open-source and clinical radiographic bone image datasets.

An average precision of 62.04% is achieved by Guan et al. [11] using 4,000 arm x-ray images. They handled the low-quality, noisy dataset, which had images with a dark background. However, the precision achieved is not very high, it is better than many modern deep learning (DL) methods.

Analysis performed by Kim and MacKinnon [12] achieved an accuracy of 95%. They used a dataset of 11,112 radiographic images of wrists. The dataset was divided into 80:10:10 training, validation, and testing groups. They trained CNN based on Tensorflow (1.0) using 20,000 iterations. The testing phase concluded that transfer learning using deep CNN can be applied for radiographic fracture detection. Training on plain X-ray images is highly transferable and can be used in medical imaging to reduce clinical risks.

Tanzi et al. [13] studied bone fracture detection and classification using 95 images augmented to create a dataset of 4476 images. They applied R-CNN to classify and find the location of the fracture. They achieved an accuracy of 96% with a precision of 0.866 in finding the precise location of the fracture. They used the VGG16 neural network to predict and locate fractures. The accuracy to identify the fracture is 95%.

An average accuracy of 86.76% is achieved by Lee et al. using a dataset of 786 images [14]. The images were resized to 512 x 512 pixels using Bi-LSTM decoded. The models applied were GoogLeNet-inception v3 and two proposed M1 and M2 models. The accuracy achieved by

Alam *et al. BMC Medical Imaging*   (2025) 25:5

Page 3 of 16

the proposed methods is not that high, and the dataset size is small.

Wrist fracture detection using X-ray images was done by Hardalac et al. [15] utilizing 26 different artificial intelligence models. Twenty distinct fracture identification processes were carried out on the wrist X-ray images dataset from Gazi University Hospital. The experiment is conducted using object detection models with different backbones. An average precision of 86.39% is achieved with a uniquely designed wrist fracture detection-combo (WFD-C) model.

The study [16] proposed DL techniques for fracture detection. A DL-based convolution neural network was developed using plain X-ray images of a biplane. The training and validation datasets consisted of 3245 fractured and 3210 normal wrist images, respectively. An accuracy of 98.0% is achieved using CNN based on the VGG16 model. The diagnosis identifies the fracture and points to its location using a heat map. This diagnosis method reduced the error of fracture detection to 47% in the trauma centers. As the dataset used for the study is only from adults above 18 years, fracture detection in children is impossible as their bones have intact epiphyseal lines.

The survey conducted by Sahin [17] used 176 radiographic images with 105 normal and 71 cracked ones. They used color space conversion to convert the images to grayscale format. Canny edge detection was performed for edge detection. For the classification of normal and fractured bone, 12 different machine-learning models were utilized. The highest accuracy of 89% was achieved by the LDA classifier, followed by 86% for logistic regression and 85% for the RF classifier, respectively.

The study conducted by Ahmed et al. [18] used a small dataset of 270 lower-leg X-ray images. The dataset was pre-processed via various stages, i.e., noise cancellation, contrast improvement, and edge detection. Gray Level Co-occurrence Matrix (GLCM) used five properties (correlation, dissimilarity, energy, homogeneity, and contrast), four distances, and seven angles for feature extraction. This technique extracted 140 features for each image. Five different machine-learning algorithms were used after splitting the dataset into 80:20. Highest accuracy, i.e. 92.85, in the current scenario is achieved by SVM.

A recent study [19] on long bone fracture detection uses a dataset of 3000 X-ray images of fractured and normal bones. The dataset was divided into 7:2:1 training, validation, and test datasets for applying machine-learning algorithms. The highest accuracy 96.5% is achieved by the ResNet50 Fine Tune model with a loss score of 0.16. Binary classification achieved an accuracy of 87.7% for four classes, which is low compared to other models. This can be due to the reduced dataset size, which was split into four classes.

Table 1 shows a summary of the analyzed works. Numerous studies have been conducted to recognize fractures using radiographic images. A thorough look into the literature has revealed some limitations that need to be addressed.

We found that most of the studies used a specific region of the body to identify the fracture. Despite existing studies, the fracture detection accuracy is low. In addition, feature engineering part using transfer learning is under investigated. There are minimal studies that can generalize fractures from all body parts. So, there is a need for a study that can identify common fractures so that findings can be practically implemented in medical care centers.

Proposed methodology

Figure 1 shows the methodology followed for fracture detection using radiographic images. Two types of features are extracted from the existing X-ray images dataset. We used CNN and a novel proposed MobLG-Net

Table 1 The analyzed literature review comparison

Ref.	Year	Dataset used	Technique used	Accuracy
[12]	2018	11,112 wrist images	CNN based on TensorFlow	95%
[11]	2020	4,000 arm images	Stochastic Gradient Descent	62.04% AP
[14]	2020	786 X-ray images	GoogLeNet-inception v3, M1, M2	86.78%
[16]	2021	6,455 wrist images	CNN	98.0%
[13]	2020	4476	CNN, VGG16	96%
[15]	2022	Not defined	wrist fracture detection-combo (WFD-C)	86.39% AP
[17]	2023	176 images	LDA classifier, LR, RF classifier	89%, 86%, 85%
[18]	2023	270 lower leg images	Naive Bayes, Nearest Neighbors, Decision Tree, SVM, Random Forest	64.2%, 83.9%, 80.3%, 92.8%, 85.7%
[19]	2024	3000 radiographs of bones	ResNet50	96.5%
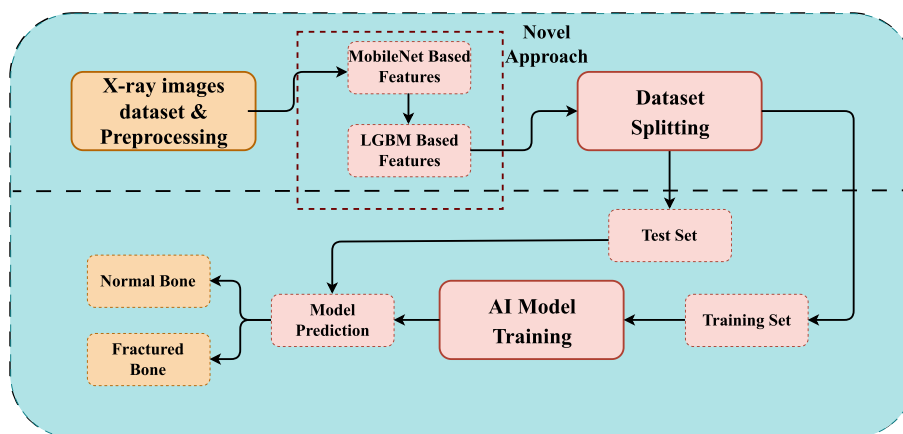
Fig. 1 The architectural illustrations of the proposed methodology for fracture detection

method to get useful features from the given dataset. Each set of features is utilized for training machine learning models. The extracted features are split into 80% training and 20% test datasets. The outperformed approach is used to detect fractures in bones using an X-ray image dataset.

Radiographic images dataset

The publically available benchmark bone fracture multi-region x-ray images dataset is utilized in this study [20]. The dataset contains 9,463 X-ray images of fractured and normal bones from all body parts, including knees, limbs, hips, lumber, etc. The dataset consists of two classes, i.e., fractured and non-fractured, as shown in Fig. 2. The number of training and test dataset images is 8,863 and 600, respectively.

Image preprocessing and formation

To get reliable results from machine learning models, it is necessary to have a clean dataset without any noise or null values [21]. The dataset is imported first using OpenCV, and basic pre-processing is applied to standardize the input data for further analysis that includes image resizing. The first step is to resize images to uniform 224 x 224 pixels to ensure they are compatible with the neural network architecture. Then, the dataset labels are encoded as 'fractured':0 and 'non-fractured':1 to ease supervised learning. Using a balanced dataset for training is advisable. Figure 3 shows the number of images in each class. The dataset is balanced as the number of images in both classes is almost similar.

Exploratory data analysis

The dataset used for training and testing the ensemble fracture detection model consists of 10,580 images having a uniform resolution of 224x224. The average size

of images is 12 kb with 160 dpi average resolution. The shape of the training set is 8863,224,224,3.
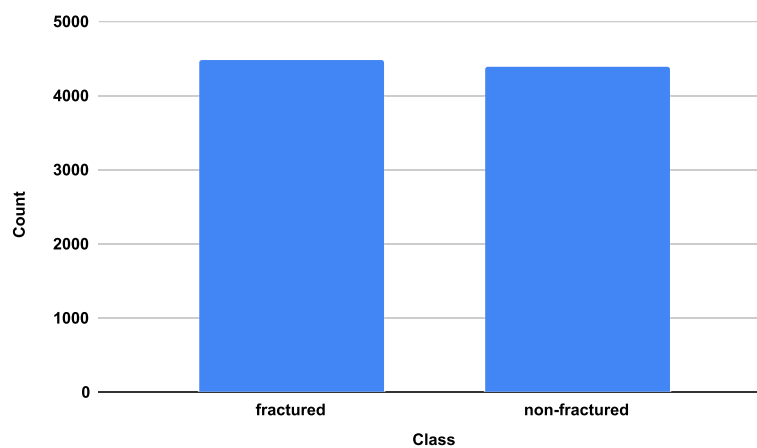
Novel transfer learning-based feature engineering

Using the radiographic images dataset, we have suggested a novel transfer learning-based feature engineering method, i.e. MobLG-Net, to detect fractured and healthy bone. The feature extraction mechanism is illustrated in Fig. 4. The novel MobLG-Net approach used three methods to get spatial features from the radiographic images dataset. First, the sequential model is implemented with the first layer of convo2D having 64 filters of 3x3. The linear stack of the sequential model is followed by max pooling 2D, dropout, flatten, and dense layers. A total of 790,337 trainable features are extracted. Then we applied MobileNet, which is based on CNN, and used pre-trained 'Imagenet' weights [22]. The use of a pre-trained model gave better results than the sequential model. Finally, a novel MobLG-Net model, a sequential model with the first layer of MobileNet (based on CNN), is implemented. Figure 4 illustrates the feature engineering techniques used.

The uniqueness of the proposed MobLG-Net lies in its innovative approach to feature extraction. Unlike traditional methods, MobLG-Net combines the strengths of sequential modeling and transfer learning. Initially, a sequential model extracts a rich set of 790,337 trainable features using convolutional layers, max pooling, dropout, and dense layers. While effective, this approach is outperformed by MobileNet, a lightweight convolutional neural network pre-trained on the 'Imagenet' dataset, which leveraged pre-trained weights to capture high-level spatial features with greater precision. By integrating the first layer of MobileNet with a custom sequential model, MobLG-Net maximizes the strengths of both feature extraction techniques. This novel combination

Fig. 2 The x-ray images examination using the target label

Fig. 3 The radiographic images data distribution analysis using encoded target label

enables the extraction of more nuanced spatial features from radiographic images, leading to superior performance across various machine learning classifiers.

Algorithm 1 shows the sequential flow of the proposed transfer learning method. Algorithm 2 contains the Pseudo Code for MobLG-Net method.
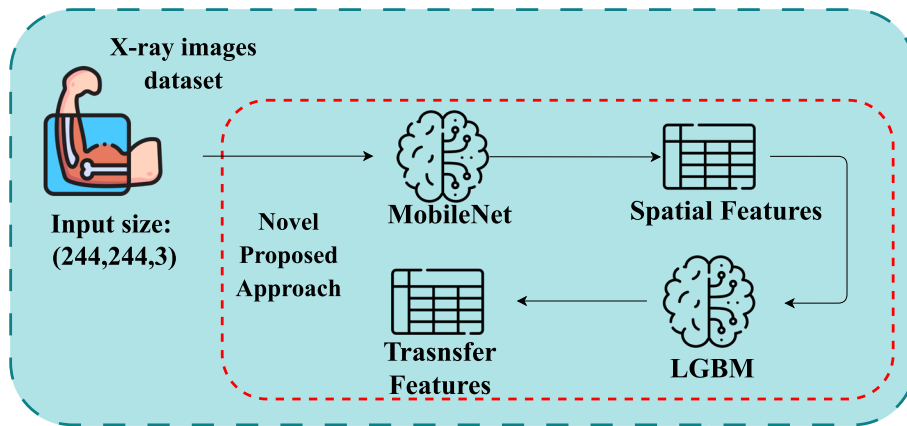
Fig. 4 The workflow of proposed transfer learning-based feature engineering approach

Algorithm 1 MobLG-Net algorithm

Input: X-ray images dataset of the bone.
Output: Transfer learning-based features generation.
initiate;

1: $F_{MobileNet} \longleftarrow MobileNet_{prediction}(XiS)$ // $XiS \ \epsilon \ X-ray\ images\ set$,   here XiS is original xrays data and F_{cnn} is extracted spatial features set.

2: $F_{lgbm} \longleftarrow LGBM_{probabilities\ prediction}(F_{cnn})$ // here F_{lgbm} is the generated probabilistic based features set.

3: $F_{Prob} \longleftarrow F_{lgbm}$ // here F_{Prob} is the final probabilistic-based generated features set used for bone fracture detection.

end;

Algorithm 2 Pseudo code for MobLG-Net model

<u>Applied artificial intelligence approaches</u>

Input: X-ray images dataset (X)

Output: Transfer learning-based features for fracture detection

1: Initialize the system and load the X-ray dataset (X).
2: Preprocess the dataset:
3:   (a) Resize images to the input shape required by the model.
4:   (b) Normalize pixel values for better performance.
5: Feature Extraction Phase:
6:   (a) **Sequential Model:**
7:     i. Create a Sequential Model with layers:
8:       - Convo2D layer with 64 filters of size 3x3.
9:       - MaxPooling2D layer.
10:       - Dropout layer.
11:       - Flatten layer.
12:       - Dense layer.
13:     ii. Extract spatial features (F_seq) using the sequential model.
14:     iii. Store 790,337 trainable features from the Sequential Model.
15:   (b) **MobileNet Model:**
16:     i. Load pre-trained MobileNet with 'Imagenet' weights.
17:     ii. Extract features ($F\_MobileNet$) from X-ray images.
18:   (c) **MobLG-Net Model:**
19:     i. Create a hybrid Sequential Model:
20:       - Use the first layer from MobileNet.
21:       - Add additional layers (Convo2D, MaxPooling2D, Dropout, Flatten, Dense).
22:     ii. Train MobLG-Net on the dataset.
23:     iii. Extract final transfer learning features ($F\_MobLGNet$).
24: Probabilistic Feature Generation Phase:
25:   (a) Pass $F\_MobLGNet$ to the LightGBM (LGBM) model.
26:   (b) Generate probabilistic predictions (F_LGBM).
27: Final Features:
28:   (a) Combine and refine F_LGBM for fracture detection.
29: **Return** the final probabilistic-based features (F_Prob).

Most of the studies use ML and DL methods in image classification jobs [23]. Convolution neural networks are the most adopted DL algorithm for image detection and prediction [24]. They apply numerous layers of filters to take out beneficial features from the dataset and understand hidden patterns in the data. Besides classical CNNs, transfer-based feature extraction is used to get the most useful features from the radiographic images. We have applied various ML techniques to both spatial and deep features extracted using CNNs and transfer learning. The advanced ML models have performed well in the detection and prediction of cracks in bones when trained and tested on image datasets [25].

- **CNN:** is a famous neural network whose architecture has been designed to read and absorb visual data such as images by learning the same way humans do. The network captures different levels of abstraction by using a sequence of convolutional layers, from the simple edges and textures in early layers to more complex patterns, such as bone structures or fractures in deeper ones. CNNs are well suited for medical imaging where we benefit from this hierarchical feature extraction, which is useful as a bone fracture in X-ray image detection [26]. When properly trained, the CNN can classify new images correctly by identifying familiar patterns, thus helping in decision-making for diagnosis.

Alam *et al. BMC Medical Imaging*   (2025) 25:5

Page 8 of 16

- **MobileNet:** is the proposed novel feature extraction approach. It uses a sequential implementation of MobileNet and LGBM to get the most relevant and effective features from the radiographic images dataset. The proposed method is lightweight and is beneficial where resources like processing power and memory are limited. This is done using a depthwise separable convolution introduced by MobileNet. A standard convolution has a filter layer that receives the input feature, reshapes it based on filter size, and then applies operation for each location. The input channels are convolved with filters separately to produce the output of each filter in a channel, and then pointwise convolution combines these outputs. This has the effect of decently reducing both parameters and computational complexity while only losing very little accuracy at all.

- **KNN** is widely applied for classification tasks, especially involving image datasets [27]. KNN classifies data points based on the proximity to their closest neighbor. KNN is a non-parametric ML model based on Euclidean distance to calculate the nearest neighbors. Fracture detection involves local patterns or irregularities within X-ray images. KNN can easily identify such irregular patterns as the classification relies on local neighbors [28]. This makes KNN a good choice for fracture detection in bones using X-ray images.

- **LGBM** is an influential ML algorithm with excellent speed and prediction power. LightGBM uses gradient boosting to detect complex patterns within the data without losing efficiency. LightGBM can handle large datasets and scattered features, which makes it an excellent choice for detecting fractures in bones.

- **LR** is mostly used in binary classification tasks. LR models the possibility of a certain class by analyzing the input features like the intensity of pixels or extracted features [29]. This helps identify if a fracture is present or not in the given dataset. LR is well suited for classifying fractures from X-ray images due to its simplicity and capability to handle linearly separable data.

- **RF** is an ensemble model that combines the estimations of several decision trees and makes a final prediction. By constructing several decision trees on different portions of the training dataset, RF averages their output, reducing the risk of overfitting [30]. Using multiple trees enhances the generalization ability of the ML model, making it highly effective for fracture classification tasks. RF is prominent for its capability to tackle nosy data and complex patterns, giving more accurate results.

Hyperparameter tuning

Choosing the optimized parameters for training ML models is crucial as parameters can affect the reliability of the research findings [23]. Hyperparameter tuning involves trying a combination of parameters that give the best performance while testing the trained models [31]. Parameter optimization is performed using the randomized search parameters and cross-validation-based mechanism, which improved the predictive capabilities of ML models [32]. We have used various techniques twice: once with spatial features extracted using CNNs and once with the novel MobLG-Net features. The hyperparameters used in the applied ML techniques are given in Table 2.

Results and discussions

This section gives a review of the performance of applied ML techniques. The Results and discussions section details the experimental setup and presents a comparison of results from the classical CNN approach and the proposed MobLG-Net approach. The accuracy, precision, recall, and f1-score give a comparative analysis of models, and these performance metrics are included in the Results and discussions section. The Results and discussions section is the essence of the whole experimental setup and extensively describes the effectiveness of ML

Table 2 The optimized hyperparameters used for DL and ML techniques

Model	Hyperparameter values
CNN	activation='sigmoid', optimizer='adam', loss='binary_crossentropy'
MobLG-Net	weights = 'imagenet', activation='softmax',
LGB	n_estimators=10, boosting_type='gbdt', max_depth=−1, num_leaves=31
KNN	n_neighbors=2, weights='uniformn', metric='minkowski', leaf_size=30, p=2
LR	penalty='l2', max_iter=100, solver='lbfgs', fit_intercept=True
RF	max_depth=300, criterion='gini', min_samples_split=2, n_estimators=300, random_state=0

Alam *et al. BMC Medical Imaging*   (2025) 25:5

Page 9 of 16

and DL models in detecting fractures using radiographic images.

Experimental setup

The research experiment is conducted using state-of-the-art Python libraries like sklearn, and TensorFlow. ML models are trained using a powerful GPU with Intel(R) Core(TM) i5-8350U CPU @ 1.70GHz 1.90 GHz processor and 15.9 GB RAM. Programs are written using Google Colaboratory (a free cloud-based platform for Python) in Python 3 for training and validation of ML models. The performance indicators like Accuracy (Acc), recall, precision, and F1 score are extracted to compare the results of the applied ML model.

In the context of machine learning for bone fracture detection, Acc measures the overall correctness of the model by calculating the proportion of correctly classified instances among all predictions. Recall focuses on the model's ability to correctly identify fractures, ensuring that as many true cases as possible are detected. Precision assesses the reliability of fracture predictions, indicating the proportion of correct fracture predictions out of all predicted fractures. Finally, the F1 score combines precision and recall into a single metric, providing a balanced measure of the model's performance, especially useful in cases of class imbalance.

Results with ML models

The training and testing of ML models is done on two sets of features, i.e. spatial features and novel MobileNet extracted features. Time series analysis shows the variation in learning rates during the training time [33]. Spatial features are extracted using CNNs, and Fig. 5 shows the training and validation accuracy using CNNs. The training accuracy improved with each epoch, and the highest accuracy of 91% was achieved. On the other hand, validation results are not that good. The highest value of validation accuracy is 81%, which is reached in epoch seven and decreases after that.

Figure 6 shows the training and validation loss across 10 epochs during the training of CNNs. Figure 6a shows that training loss sharply drops from around 60 to almost zero after just one epoch and remains zero throughout the remaining process. When training loss reduces rapidly to zero and becomes low, indicating that the model is overfitting and training too well.

The validation loss starts at 1.6 and decreases up to around epoch 5. After this, the value of validation loss fluctuates between 1 and 1.2 but does not drop significantly as training loss. The model initially improved in handling unseen data, but fluctuations after epoch 5 indicate that the model started to overfit the training data.
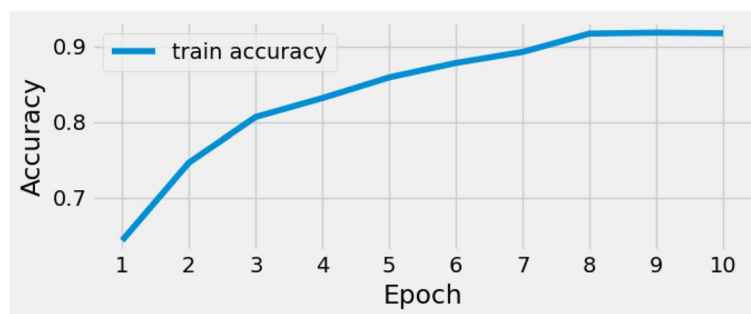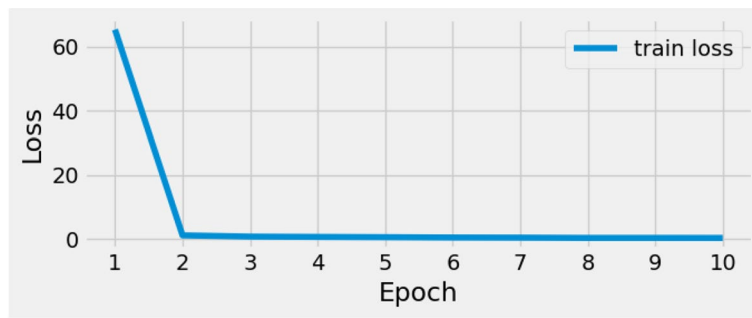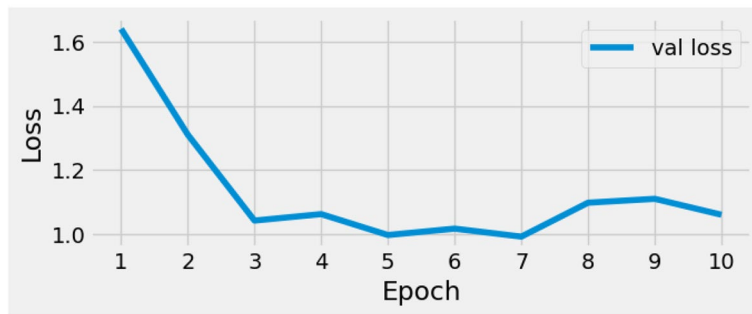

(a)
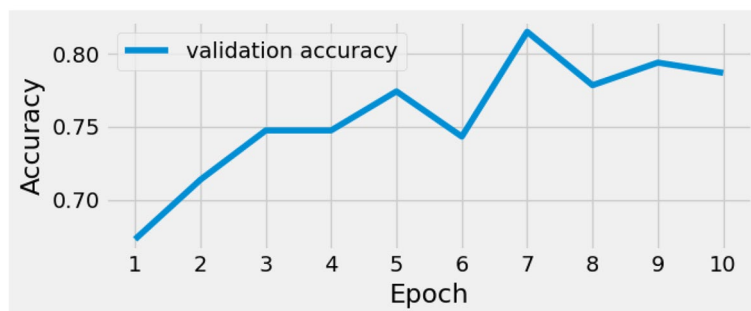
(b)

Fig. 5 Shows (**a**) train accuracy and (**b**) validation accuracy for classical CNN approach

Fig. 6 Shows (**a**) train loss and (**b**) validation loss for classical CNN approach

Results with MobileNet

The results obtained with the MobileNet approach are given in Fig. 7. The training accuracy is initially between 78% and 87%. The accuracy improved with time and reached 90% during epoch 2. The training accuracy reached 95% during epoch nine and remained constant during epoch 10. The validation accuracy shows fluctuations throughout the validations. The value of accuracy started from 91% and improved to 95% from epoch 1 to epoch 4. The validation accuracy sharply dropped to 86% at epoch 5, indicating that model performance decreased significantly on the validation data. After epoch 5, the validation accuracy again started rising up to 96%, which is 1% higher than the training accuracy.

The training and validation loss scores are illustrated in Fig. 8. The training loss starts at a high value of 0.85 and gradually decreases from epoch 1 to 4. The value of training loss then remains low between 0.4 and 0.2 throughout up to the end. This consistently low value indicates that the model is learning well from the training data.

The validation loss is given in Fig. 8b, and its value is 0.3, which is not very high compared to the starting value of train loss. The loss gradually decreases from epoch 1 to 4. Just like fluctuations in validation accuracy, the validation loss also started to rise during epoch 5. The validation loss started to decrease after epoch five and reached a minimum value of 0.0 around epochs 9 and 10.

## Classification results using CNN and MobileNet

Table 3 gives the classification of fractured and normal bone images. The CNN model achieves an accuracy of 81%, showing a suitable performance. The proposed MobileNet shows a high classification accuracy of 97%. MobLG-Net clearly outperforms CNN across all performance metrics. High accuracy, precision, recall, and f1-score values suggest that MobileNet is suitable for classifying fractured and non-fractured bones. The ability of MobileNet to classify correctly makes it reliable. However, still, there is room for performance improvement.

Results with only spatial features

The performance of ML models on these classically extracted spatial features is given in Table 4. The highest accuracy of 93% is achieved with the LGBM model, followed by random forest having 89% accuracy. The results of the classical approach are acceptable, and reasonable values of accuracies have been achieved, yet there is room for improvement. So, we applied the ML models to the features extracted from the proposed novel approach. The results are discussed in the next section.
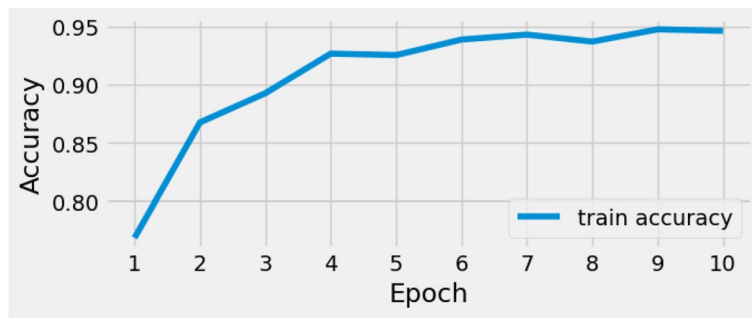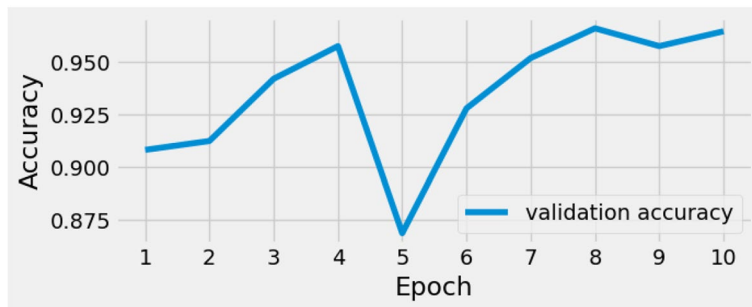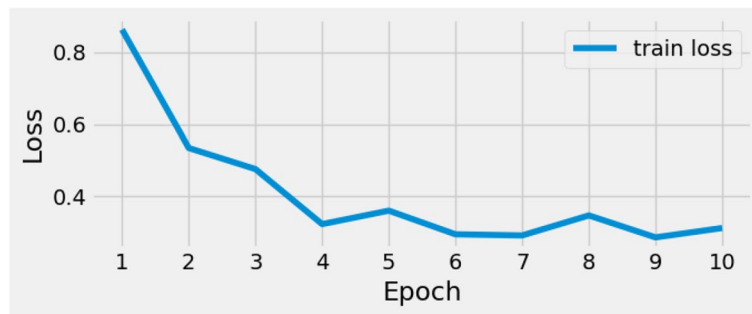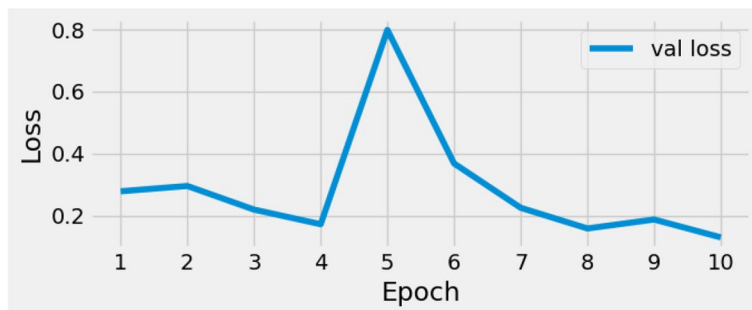
Fig. 7 Shows (**a**) train accuracy and (**b**) validation accuracy for the MobileNet approach

Fig. 8 Shows (**a**) train accuracy and (**b**) validation accuracy for the MobileNet approach

Table 3 The classification of fractured and non-fractured bone images using CNN and MobileNet

Model	Accuracy	Label	Precision	Recall	F1
CNN	0.81	Fractured	0.83	0.78	0.80
		Non-fractured	0.79	0.84	0.82
		Average	0.81	0.81	0.81
MobileNet	0.96	Fractured	0.95	0.99	0.97
		Non-fractured	0.99	0.95	0.97
		Average	0.97	0.97	0.97

Results with proposed approach

Most optimized features, extracted from the proposed MobLG-Net, are used for training machine learning models. With the same hyperparameters used for the classical approach, novel transfer learning-based features performed well. The light gradient boosting and logistic regression outperformed the KNN and random forest by 1%. An accuracy of 99% is achieved by both algorithms, followed by 98% accuracy of KNN and RF. The precision, recall, and f1-scores are also above 97, showing the generalization ability of applied models on novel extracted features. Table 5 details each target label's performance metrics. The proposed transfer learning approach extracted better features than the classical CNN approach. The proposed method outperformed the classical approach by 5%, proving the pre-trained models' effectiveness.

Figure 9 provides the confusion matrix study of the applied methods. The given matrix shows the truth table of predictions made by the applied models. The confusion matrix analysis illustrates the strengths and weaknesses of the models. KNC and LGBMC have a lesser number of false predictions of 22 samples than RF and LR, although

Table 4 The evaluation of the effectiveness of ML techniques using spatial features on test data

Model	Accuracy	Target Class	Precision	Recall	F1 Score	Support
KNN	0.72	fractured	0.67	0.87	0.76	892
		non-fractured	0.81	0.57	0.67	881
		Average	0.74	0.72	0.72	1773
LGB	0.93	fractured	0.94	0.91	0.93	892
		non-fractured	0.91	0.95	0.93	881
		Average	0.93	0.93	0.93	1773
LR	0.64	fractured	0.63	0.70	0.66	892
		non-fractured	0.66	0.59	0.62	881
		Average	0.64	0.64	0.64	1773
RF	0.89	fractured	0.91	0.87	0.89	892
		non-fractured	0.87	0.91	0.89	881
		Average	0.89	0.89	0.89	1773

Table 5 The performance evaluation of applied ML techniques on features extracted using novel MobLG-Net

Model	Accuracy	Target class	Precision	Recall	F1 Score	Support
KNN	0.98	fractured	0.97	0.99	0.98	892
		non-fractured	0.99	0.97	0.98	881
		Average	0.98	0.98	0.98	1773
LGB	0.99	fractured	0.99	0.99	0.99	892
		non-fractured	0.99	0.99	0.99	881
		Average	0.99	0.99	0.99	1773
LR	0.99	fractured	0.99	0.99	0.99	892
		non-fractured	0.99	0.99	0.99	881
		Average	0.99	0.99	0.99	1773
RF	0.98	fractured	0.98	0.98	0.98	892
		non-fractured	0.98	0.98	0.98	881
		Average	0.98	0.98	0.98	1773
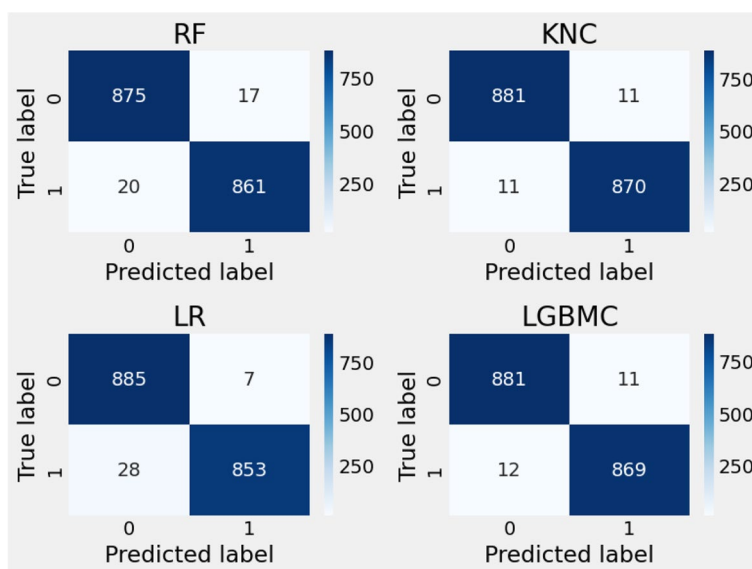
Fig. 9 Confusion matrix study of ML approaches with proposed features

Table 6 The runtime computational complexity analysis of applied ML models

Model	Runtime computational cost (in seconds)	
	With spatial features	With MobLG-Net features
KNN	0.016	0.007
LGB	29.46	0.0415
LR	0.245	0.035
RF	70.00	3.910

Table 7 The cross-validation-based performance analysis of applied methods

Model	K-fold accuracy	Standard deviation
RF	0.977	0.0043
LR	0.985	0.0034
KNC	0.976	0.0042
LGBM	0.985	0.0035

the accuracy of LR is high. The RF model achieved a high false prediction error of 37 samples. The same is the case with the LR model. This confusion matrix analysis shows that applied ML models have performed well on transfer-based features.

Computational complexity analysis

The time taken by each applied ML model to train on spatial and MobLG-Net features is the computational complexity. The time taken by RF is 70 seconds, as shown in Table 6, which is the highest time for training on spatial features. Also, the KNN and LR took less than a second to predict fractures in radiographic images. An average performance is shown with the spatial features, with LGB and RF taking too much time. When the time comparison is made with the novel extracted features, the proposed approach also outperformed in computational cost. The highest time taken is by random forest, which is 3.9 seconds, followed by KNN, RGB, and LR, taking a fraction of a second. Moreover, the accuracy of

LDB and LR is the highest, with the lowest computation cost, making the best models for generalization.

Performance analysis using cross-validation

To ensure the reliability and robustness of the proposed methods, we conducted a comprehensive cross-validation-based performance analysis. The results, summarized in Table 7, demonstrate that the models achieved high accuracy with minimal standard deviation. Among the applied methods, LR and LGBM emerged as the top performers, both achieving a k-fold accuracy of 0.985 with standard deviations of 0.0034 and 0.0035, respectively. The RF and KNC models also exhibited strong performance, with accuracies of 0.977 and 0.976 and standard deviations of 0.0043 and 0.0042, respectively. These results highlight the effectiveness of the selected models and the robustness of our approach to detecting bone fractures.

Comparison with state-of-the-art approaches

The comparative analysis of the proposed approach with the previous state-of-the-art studies is presented

in Table 8. For an honest comparison, we have taken the most recent studies between the years 2020 and 2024. The highest accuracy achieved by previous research is 98%, which used only hand-wrist images and has very limited application. Also, the classical CNN model is used in the study to get this high accuracy. Following this, the remaining studies have not shown extraordinary performance to be noted. The proposed novel feature extraction has yielded good results, and with 99% accuracy, the approach stood prominent in detecting fracture in radiographic images.

Ablation study

The ablation study provides a comprehensive analysis of the performance improvements achieved by the proposed MobLG-Net method compared to classical approaches. Table 9 highlights the accuracy achieved using both techniques across various machine learning methods. The classical combination of MobileNet and LGBM demonstrated solid performance, with accuracy values ranging from 0.64 to 0.93 depending on the method. However, the proposed MobLG-Net approach significantly enhanced these results, achieving near-perfect accuracy values between 0.98 and 0.99. Notably, methods like LR and LGBM saw the most dramatic improvement, with accuracy jumping from 0.64 to 0.99 and 0.93 to 0.99, respectively. This analysis clearly underscores the efficacy of MobLG-Net, demonstrating its potential to set a new benchmark for robust and accurate feature extraction in machine learning pipelines.

Study limitations

While our study demonstrates the effectiveness of the proposed MobLG-Net method for classifying fractured and normal bones using a dataset of 9,463 X-ray images, it is not without limitations. One significant challenge is the dataset size, which, while substantial, may not fully capture the diversity of fracture patterns across different populations. This limitation could potentially affect the model's generalizability, particularly when applied

Table 9 Performance comparison of ML models and proposed approach as ablation study analysis

Model	Performance accuracy using classical approaches (MobileNet and LGBM)	Performance accuracy using proposed MobLG-Net method
KNN	0.72	0.98
LGB	0.93	0.99
LR	0.64	0.99
RF	0.89	0.98

to underrepresented groups, such as pediatric patients, whose bone structure and fracture characteristics differ from adults. Additionally, our study focuses on binary classification, distinguishing between fractured and normal bones. Expanding this work to include multiclass classification for different types of fractures such as greenstick, comminuted, and spiral fractures-could enhance the clinical applicability of the model.

Conclusion and future work

Fracture detection using X-ray images and transfer learning-based feature extraction is presented in this study. The spatial and MobLG-Net features are extracted using CNNs and the proposed MobLG-Net method. During the research experiments, both sets of extracted features are fed to various ML methods, and their performance is compared. The results indicate that a low performance is observed with the spatial features. However, the models trained on the proposed MobLG-Net feature outperformed and gave a high-performance score. We analyzed the performance matrices of applied ML methods and noted that the proposed approach has satisfactory results in all matrices. The time complexity analysis of the applied models is done to check the time taken by each. Performance is also validated using a cross-validation mechanism. The study shows that the models trained using MobLG-Net features took less learning time. The research results are compared with other state-of-the-art studies.

Table 8 The contrast between the proposed and other state-of-the-art studies in predicting fractures using radiographic images

Reference	Year	Technique	Accuracy
[13]	2020	CNN, VGG16	96
[14]	2020	GoogLeNet-inception v3, M1, M2	86.78%
[16]	2021	CNN	98%
[15]	2022	wrist fracture detection-combo (WFD-C)	86.39%
[18]	2023	Naive Bayes,, Nearest Neighbors,SVM, Random Forest, Decision Tree	92.85%
[19]	2024	Binary Classification	96.5%
This study	**2024**	**LGB, LR, RF, KNN**	**99%, 99%, 98%, 98%**

The study to detect fractures yielded satisfactory results and generalized the unseen data well. In future work, we aim to extend the research dataset size, including more radiographic images, so that a more robust prediction can be made. Further, instead of binary classification (i.e., fracture or normal), we will introduce multi-class classification, which will identify the types of fractures. By addressing this area, we will be able to implement this research practically and improve the clinical diagnosis of fractures.

Data availability

The dataset used in this study is publicly available at the following link: https://www.kaggle.com/datasets/devbatrax/fracture-detection-using-x-ray-images/data.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

References

1. AlGhaithi A, Al Maskari S. Artificial intelligence application in bone fracture detection. J Musculoskelet Surg Res. 2021;5:4.
2. Wu AM, Bisignano C, James SL, Abady GG, Abedi A, Abu-Gharbieh E, et al. Global, regional, and national burden of bone fractures in 204 countries and territories, 1990–2019: a systematic analysis from the Global Burden of Disease Study 2019. Lancet Health Longev. 2021;2(9):e580–92.
3. Williamson S, Landeiro F, McConnell T, Fulford-Smith L, Javaid MK, Judge A, et al. Costs of fragility hip fractures globally: a systematic review and meta-regression analysis. Osteoporos Int. 2017;28:2791–800.
4. Ayesa SL, Katelaris AG, Brennan PC, Grieve SM. Medical imaging education opportunities for junior doctors and non-radiologist clinicians: A review. J Med Imaging Radiat Oncol. 2021;65(6):710–8.
5. Kwakernaak S, van Mens K, Cahn W, Janssen R, Investigators G, et al. Using machine learning to predict mental healthcare consumption in non-affective psychosis. Schizophr Res. 2020;218:166–72.
6. Ahammed M, Mamun MA, Uddin MS. A machine learning approach for skin disease detection and classification using image segmentation. Healthc Analytics. 2022;2:100122. https://doi.org/10.1016/j.health.2022.100122.
7. Ramesh J, Aburukba R, Sagahyroon A. A remote healthcare monitoring framework for diabetes prediction using machine learning. Healthc Technol Lett. 2021;8(3):45–57.
8. Thalji N, Aljarrah E, Almomani MH, Raza A, Migdady H, Abualigah L. Segmented X-ray image data for diagnosing dental periapical diseases using deep learning. Data Brief. 2024;54:110539.
9. Das NN, Kumar N, Kaur M, Kumar V, Singh D. Automated deep transfer learning-based approach for detection of COVID-19 infection in chest X-rays. Irbm. 2022;43(2):114–9.
10. Kaya O, Taşcı B. A Pyramid Deep Feature Extraction Model for the Automatic Classification of Upper Extremity Fractures. Diagnostics. 2023;13(21):3317.
11. Guan B, Zhang G, Yao J, Wang X, Wang M. Arm fracture detection in X-rays based on improved deep convolutional neural network. Comput Electr Eng. 2020;81:106530.
12. Kim D, MacKinnon T. Artificial intelligence in fracture detection: transfer learning from deep convolutional neural networks. Clin Radiol. 2018;73(5):439–45.
13. Tanzi L, Vezzetti E, Moreno R, Moos S. X-ray bone fracture classification using deep learning: a baseline for designing a reliable approach. Appl Sci. 2020;10(4):1507.
14. Lee C, Jang J, Lee S, Kim YS, Jo HJ, Kim Y. Classification of femur fracture in pelvic X-ray images using meta-learned deep neural network. Sci Rep. 2020;10(1):13694.
15. Hardalaç F, Uysal F, Peker O, Çiçeklidağ M, Tolunay T, Tokgöz N, et al. Fracture detection in wrist X-ray images using deep learning-based object detection models. Sensors. 2022;22(3):1285.
16. Oka K, Shiode R, Yoshii Y, Tanaka H, Iwahashi T, Murase T. Artificial intelligence to diagnosis distal radius fracture using biplane plain X-rays. J Orthop Surg Res. 2021;16:1–7.
17. Sahin ME. Image processing and machine learning-based bone fracture detection and classification using X-ray images. Int J Imaging Syst Technol. 2023;33(3):853–65.
18. Ahmed KD, Hawezi R. Detection of bone fracture based on machine learning techniques. Meas Sensors. 2023;27:100723.
19. Ali SNE, Sherif HM, Hassan SM, El Marakby AAER. Long bones x-ray fracture classification using machine learning. J Al-Azhar Univ Eng Sect. 2024;19:121–33.
20. Batra D. Fracture detection using x-ray images. https://www.kaggle.com/datasets/devbatrax/fracture-detection-using-x-ray-images/data. Accessed 01 Aug 2024.
21. Khalid M, Raza A, Younas F, Rustam F, Villar MG, Ashraf I, et al. Novel Sentiment Majority Voting Classifier and Transfer Learning-based Feature Engineering for Sentiment Analysis of Deepfake Tweets. IEEE Access; 2024;12:67117–29.
22. Simonyan K. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556. 2014.
23. Raza A, Younas F, Siddiqui HUR, Rustam F, Villar MG, Alvarado ES, et al. An improved deep convolutional neural network-based YouTube video classification using textual features. Heliyon. 2024;10:1–16.
24. Terziyan V, Vitko O. Causality-aware convolutional neural networks for advanced image classification and generation. Procedia Comput Sci. 2023;217:495–506.
25. Meena T, Roy S. Bone fracture detection using deep supervised learning from radiological images: A paradigm shift. Diagnostics. 2022;12(10):2420.
26. Hussein HI, Mohammed AO, Hassan MM, Mstafa RJ. Lightweight deep CNN-based models for early detection of COVID-19 patients from chest X-ray images. Expert Syst Appl. 2023;223:119900.
27. Shah AA, Parah SA, Rashid M, Elhoseny M. Efficient image encryption scheme based on generalized logistic map for real time image processing. J Real-Time Image Proc. 2020;17(6):2139–51.
28. Nasim S, Al-Shamayleh AS, Thalji N, Raza A, Abualigah L, Alzahrani AI, et al. Novel Meta Learning Approach for Detecting Postpartum Depression Disorder Using Questionnaire Data. IEEE Access. 2024.
29. Younas F, Raza A, Thalji N, Abualigah L, Zitar RA, Jia H. An efficient artificial intelligence approach for early detection of cross-site scripting attacks. Decis Analytics J. 2024;11:100466.
30. Tanveer MU, Munir K, Raza A, Almutairi MS. Novel artificial intelligence assisted Landsat-8 imagery analysis for mango orchard detection and area mapping. PLoS ONE. 2024;19(6):e0304450.

31. Haider M, Hashmi MSA, Raza A, Ibrahim M, Fitriyani NL, Syafrudin M, et al. Novel Ensemble Learning Algorithm for Early Detection of Lower Back Pain Using Spinal Anomalies. Mathematics. 2024;12(13):1955.

32. Weerts HJ, Mueller AC, Vanschoren J. Importance of tuning hyperparameters of machine learning algorithms. arXiv preprint arXiv:2007.07588. 2020.

33. Sayed MS, Rony MAT, Islam MS, Raza A, Tabassum S, Daoud MS, et al. A Novel Deep Learning Approach for Forecasting Myocardial Infarction Occurrences with Time Series Patient Data. J Med Syst. 2024;48(1):53.

Publisher's Note