IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 17 February 2024, accepted 14 March 2024, date of publication 20 March 2024, date of current version 29 April 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3379760

RESEARCH ARTICLE

Detection of Hand Bone Fractures in X-Ray Images Using Hybrid YOLO NAS

SAI CHARAN MEDARAMATLA[1], CHENNUPATI VEDA SAMHITHA[1], SAGAR DHANRAJ PANDE[2], AND SURENDRA REDDY VINTA[1]

[1]School of Computer Science and Engineering (SCOPE), VIT-AP University, Amaravati, Andhra Pradesh 522237, India
[2]School of Engineering and Technology, Pimpri Chinchwad University, Pune, Maharashtra 412106, India

Corresponding author: Surendra Reddy Vinta (surendrareddy.vinta@vitap.ac.in)

This work was supported by the Funding for Open Access Approved by VIT-AP University under Grant 10000679190.

ABSTRACT The majority of bones that have fractured in humans are hand bones. As we use our hands widely, they need early and accurate detection to be diagnosed. Fractures in the hands are most frequently brought on by blunt force trauma, sports injuries, and bone fragility. Getting an X-ray of the affected area of the bone and then discussing the results with a medical practitioner or radiologist is the standard procedure for determining whether or not a fracture exists in the bone. The majority of medical professionals and radiologists use X-rays to diagnose hand fractures; however, in some instances, they might miss small or hairline fractures. Additionally, it might be difficult to find a good radiologist who can detect the fracture properly and in time, because a delay in diagnosis can cause the injury to be more severe, and the bone might not be recovered properly. Therefore, to detect hand bone and joint fractures through X-rays, a hybrid model was developed that uses deep learning algorithms YOLO NAS (You Only Look Once - Neural Architecture Search), Efficient Det, and DETR3 (DEtection TRansformer), which are widely recognized for their exact object detection capabilities. The dataset used for this model is a hybrid dataset of 4736 hand-bone X-ray images, they were further classified into 6 classes based on their types. To evaluate the performance the best method is to compare the proposed model with the existing models, hence, the model was compared with various existing algorithms and result analysis was done.

INDEX TERMS Hand bone fracture, fracture detection, X-rays, YOLO NAS, deep learning, hybrid dataset, hybrid model, DETR3.

I. INTRODUCTION

Every single bone and joint in the hand is necessary for a person to be able to carry out their typical activities. As hands are needed for various purposes every day, the structure of the hand bones is very complex. There are 27 separate bones in the human hand: 8 carpal bones, 5 metacarpal bones, and 14 finger bones (also known as phalanges) that are joined by ligaments and joints. Our hands include about 25% of all the bones in our body [1]. This structure allows you to do a variety of tasks with your hands, including lifting large objects, gripping objects tightly, and threading a small thread through a needle's tiny eye [1]. The bones in our hands provide the hand with stability and make it possible for us to engage in tasks that demand strength. With the help of these flexible bones and joints, we can perform a variety of gestures and movements to convey the meaning of what we are trying to say, the hands play an important role in nonverbal communication too.

As everyone uses their hands for so many different activities, there is a significant risk that someone will hurt their hands. The BLS statistics cited by Occupational Health & Safety magazine say that in a year an average of 1 million people are being considered for emergency hand injuries [2]. Given this information, we may deduce that hand fractures are one of the most frequently occurring injuries. Fractures of the hand can be caused even by something as simple as falling or twisting your body suddenly and forcefully. Fractures in the hands are most frequently brought on by blunt force trauma, sports injuries, and bone fragility. The accident, severity, and

The associate editor coordinating the review of this manuscript and approving it for publication was Ajit Khosla.

the kind of injury will be different for all, so every individual should be treated differently based on their injury. In addition, getting an accurate medical evaluation as soon as possible after a bone fracture is essential for optimum recovery.

As of now the easiest way for radiologists to detect a fracture caused by trauma is to use Radiography [3]. Radiologists are required to have specialized knowledge and training to interpret medical images precisely. When reading the X-rays, there is a potential for false negatives; in some instances, the human eye may be unable to detect the fractures, and when the X-ray is rotated (resulting in an X-ray that is not clear), it may be challenging to detect the fracture. In these times Machine Learning and Deep Learning methods for autonomous actions come into use. As of now, there exist many mechanisms in the medical field using the ML and DL methods for completing tasks accurately within much less time. Some of the popularly used algorithms are CNN (Convolutional Neural Networks), RNN (Recurrent Neural Networks), and SVMs (Support Vector Machines) from Machine Learning and GANs (Generative Adversarial Networks) from Deep Learning. Convolutional Neural Networks (CNNs) have been at the forefront of DL-based object detection, offering remarkable performance in various domains, including medical imaging. These models excel in accurately localizing objects while achieving high detection rates. YOLO (You Only Look Once) is another popular DL architecture for object detection, known for its real-time inference speed and simplicity. YOLO models operate by dividing the input image into a grid and predicting bounding boxes and class probabilities directly from each grid cell, making them efficient for detecting multiple objects in a single pass.

In addition, if there is a lack of radiologists who can identify the fracture, the patient may have to wait longer than necessary to receive a diagnosis and treatment, which poses a significant health risk. For the bone to heal properly, it is essential to treat the injury accurately. We will be able to continue the daily activities and be fit, only if the bone is completely healed, and treating the injury as soon as possible plays a great role in the healing process [4]. This shows the importance of this model for society. This model works by obtaining the input image(X-ray) from the user and searching for fracture in that image using the YOLOv8 algorithm. This invention will be very helpful to all, as the detection of a hand bone fracture becomes easy and the diagnosis can be started at an early stage. This reduces the risk of the bone's incomplete healing. The contributions of this paper are as follows:

- Experimented to find an easy way to detect hand bone fractures.
- The dataset used for this experiment is a novel hybrid dataset of 4736 images.
- Experimented by developing a hybrid model that uses various deep-learning algorithms

The remaining sections will be organized as follows; Section II covers the related task which is the literature review for our proposed plan. Section III describes the proposed methodology, which includes a description of the dataset, preprocessing, and model description. Section IV describes the experiment, analysis of results, and comparison of results. Section V outlines the conclusion and prospective scope of the study.

II. RELATED WORK

The models that use machine learning and deep learning algorithms can never be useless, these algorithms can also help to detect bone fractures and are quite useful for early diagnosis and decreasing the risk to the patient, and it's been increasing every day. Computer vision-based algorithms are the basis for detecting fractures in human bones, and until today, several works have been done in this particular domain.

Ju and Cai [5] performed YOLOv5 and YOLOv8 on children's wrist fractures and compared the results. As they were not satisfied and did not achieve the best results the authors decided to apply data augmentation to the dataset. The model was first trained on 14,204 images and then increased to 28,408 using data augmentation.

Ahmed and Hawezi [6] focused on the detection of bone fractures using machine learning techniques, particularly X-ray image fractures. Various algorithms including Decision tree, Naïve Bayes, Nearest neighbors, Random Forest, and SVM are used and the accuracy of these algorithms lies between 0.64 to 0.92. The authors' aim is to develop a program that helps medical assistants in detecting a fracture in the leg bone.

Karanam et al. [7] used three types of CNN: CNN-frontal, CNN-bounding, and CNN-metal for preprocessing, and Laplacian methodology is used to predict the margins of the injured bones. The authors also proposed an in-depth neural network to assist specialists in distinguishing different types of fractures.

Xue et al. [8] study proposes a method to detect bone injuries. The proposed method uses the guided anchoring method of GA_RPN for better anchor generation. As the model can identify fractures in various parts of the human body it will be useful for medical institutions. The proposed framework showed an accuracy of 97% to 99% outperforming all the existing models of the same domain.

Kandel et al. [9] studied the usage of transfer learning, a deep learning algorithm in computer vision to classify musculoskeletal images to detect bone fractures. Various CNN methods were used such as VGG, Xception, ResNet, GoogleNet, InceptionResNet, and DenseNet, and the metrics accuracy and Kappa were used to evaluate the performance of all networks. The model obtained a confidence score of 95%.

Santos et al. [10] wanted to develop a model that can even detect transverse fractures as thin as 1mm to 13mm deep. They used the concept of feasibility for the model implementation. Reference [10] emphasizes the importance of using the Vivaldi antenna for the accurate detection of bone fractures.

Inui et al. [11] used the deep learning algorithm YOLOv8 to detect Elbow OCD, as the said algorithm is accurate and fast it can be used to perform various tasks such as image classification, object detection, and image segmentation. The authors classified the normal and OCD images into four classes using binary classification. The optimum value of F-measure Confidence is calculated and obtained as 0.701, and 0.781 for the YOLOv8n model, and YOLOv8m model respectively.

Kassem et al. [12] developed a unique deep-learning model based on ResNet50 for the easy identification of pelvis fractures in scanning images. The study investigates the pelvis classification problem using convolutional layers and receptor fields, with a GPU being used to increase performance. The trial results illustrate that the Grad-CAM-based method is effective in detecting pelvic fractures and pelvic morphology.

Zhang et al. [13] performed the study on patients with acute chest injuries who received thin-slice and collected their scans, the scans then were classified into three classes, one assisted by radiologists another assisted by DL as the concurrent reader, another assisted by DL as the second reader, the results have shown that, assisting DL as the concurrent reader improves the detection accuracy of rib fractures.

Bevers et al. [14] did a study with the purpose of examining if SSM-based shape characteristics are linked with fracture presence in patients with a clinically suspected scaphoid fracture. The SSM was created using MATLAB code and a randomly generated template mesh, with some meshes rigidly registered and some non-rigidly registered. The first form mode accounted for 72.1% of the overall shape variance, while the second and third modes explained 6.3% and 4.2%, respectively.

Yadav et al. [15] proposed a methodology to detect bone fractures using hybrid SFNet. The model is also used for the sorting of bone images with the SoftMax activation function. Yadav et al. [15] used a dataset of 34000 images and implemented the model. Considering precision as the metric to evaluate the performance of the model, the least was observed as 45% with MobileNetV2 and the highest was 100

Ma and Luo [16] developed a system that splits all human bones into 20 different types based on human anatomy. For the detection tasks, the Faster R-CNN algorithm was used. The authors compared various methodologies and the results show that the CrackNet is better in terms of recall and specificity. The system achieved 88,39% accuracy outperforming other methods.

Li and Tan [17] conducted a study on the role of Ultrasound images to assess and diagnose various bone fractures. The risk of bias and relevancy of eligible studies was assessed using four domains: selection of patients, index test, reference standard, and flow and timing. The risk of bias calculation was assessed using the QUADAS-2 risk assessment tool. Compared to conventional radiography, the results demonstrated a high degree of accuracy in fracture detection.

Rakesh and Akilandeswari [18] conducted a project to detect bone fractures using the Canny edge detection Bernsen algorithm. The model was performed on a Kaggle-obtained dataset with 15000 samples of wrist fracture RGB images. The Canny detection and Bernsen algorithm achieved 76.85% and 75.29% respectively. Rakesh and Akilandeswari [18] also concluded that Canny edge detection is less sensitive to partial blockage.

Mishra and Mishra [19] conducted a study with the objective of presenting various techniques to detect fractures in bones on X-rays with the usage of trending AI and deep learning algorithms. Types of bone fractures were classified into 7 classes. Machine learning algorithms such as Bagging, Boosting, Stacking, Random Forest, and various deep learning algorithms were performed on the model and the final best accuracy achieved was 90%.

Despite many studies in this particular domain, there are still many factors that can be improved to get the best model, like the fact that most of the studies are based on publicly available unannotated and publicly available annotated datasets which might not be of good quality. In many studies, the classes chosen were very less, and only a particular part of the anatomy is focused. In this study, the images that were selected to experiment are manually chosen based on quality and it focuses on all hand bones and joints.

Table 1 is the analysis after comparing the studies in related work.

III. MATERIALS AND METHODOLOGIES

A. DATASET DESCRIPTION

The data that is necessary to train and develop the model is in the dataset. The performance and generalizability of a model are directly impacted by the excellence and the range of the dataset. A solid dataset is required for a model to understand complex patterns and make accurate predictions. The dataset considered for this model is a self-made hybrid dataset, different datasets from different sites such as Kaggle, Mendeley data are considered and combined to select the best 4736 images. The size and format of the image used for YOLO NAS should align with the model's input requirements, as the dataset contains images of different sizes the squared images of dimensions 608 $\times$ 608 pixels were selected while maintaining the original aspect ratio and using standard image format PNG. These 4736 images were then separated into training, validating, and testing datasets of proportions 70%, 20%, and 10% respectively. The images are classified into 6 different classes: Finger fracture, Wrist fracture, Forearm fracture, Elbow fracture, Humerus fracture, and Shoulder fracture. Each image was annotated according to its class before it was added to the dataset. The curated dataset is highly efficient and simple to comprehend, which is a significant advantage for anyone interested in working in this field. Fig. 1- Fig. 6 belongs to the dataset that is used to train the model, each image represents each class that is considered.

TABLE 1. Comparison table of related work.

Authors	Year	Methods used	Dataset Type	Detection/Classification/Analysis	No. of classes
Rui Yang Ju et al.	2023	YOLOv8	Public	Detection	8 (only wrist)
Kosrat et al.	2023	Various ML techniques	Self-developed	Detection	2
Santoshachandra et al.	2023	SIFT algorithm	Self-developed	Classification	2
Linyan et al.	2021	R-CNN	Self-developed	Detection	2
Kandel et al.	2020	Transfer learning	Public	Classification	2
K. C. Santos et al.	2022	Kirchhoff migration algorithm	Self-developed	Detection	2
Inui et al.	2023	YOLOv8	Self-developed	Detection	5
Mohamed et al.	2023	Transfer Learning	Public	Detection	3
Zhang et al.	2021	Various DL algorithms	Public	Detection	2 (only ribs)
Melissa et al.	2021	SSM construction	Self-developed	Classification	2 (only scaphoid)
Yadav et al.	2022	SFNet and Canny	Public	Detection	2
Yangling et al	2020	CrackNet	Public	Detection	2
Enmiao et al.	2022	QUADAS-2	Self-developed	Analysis	No classes
Rakesh et al.	2023	Bersen algorithm	Public	Detection	2
Mishra et al.	2022	Various CNN methodologies	Public	Analysis	2

FIGURE 1. Humerus positive.

FIGURE 2. Forearm positive.

FIGURE 3. Shoulder positive.

FIGURE 4. Elbow positive.

FIGURE 5. Fingers positive.

FIGURE 6. Wrist positive.

B. DATA PREPROCESSING

The only source of getting the data useful for real-time detections is to get the data in the raw state. The raw data is mostly noisy and inconsistent, so preprocessing will be a sure step to get the raw data into its best form [6]. Data preprocessing is converting the available data into a suitable format for machine learning and deep learning tasks. The main five steps in data preprocessing are; cleaning, that is removing all the noise from the dataset, fixing errors and filling in the missing values, transformation, which is like organizing the images of the dataset for a better understanding, Integration, this step is merging data from different sources, reduction, in which the data that is inefficient and irrelevant is removed from the dataset, and the final step is, normalization, this helps to compare the results of different classes or different features. The dataset that was used is a hybrid one, and the publicly available dataset consists of many unclear and redundant images, all that noise in the image dataset was cleared manually, and the images from different datasets to create one efficient and reliable dataset. No tools were used to annotate the images, all images were annotated manually according to the classes that have been divided.

C. EXISTING METHODOLOGIES

To compare and evaluate the model's performance, the experiment has been conducted on various popularly known existing algorithms. The same dataset that was created for this experiment was used for all the existing algorithms implementations.

1) INCEPTION V3

InceptionV3, a strong CNN architecture, shows a compromise between the two most popularly known metrics accuracy and processing efficiency. This makes the model appropriate for a wide range of image recognition applications. It was designed to address deep learning problems on large image datasets for image classification and recognition. Local and global patterns in input photos can be learned by the network. With this factorization, the network can analyze larger receptive fields while lowering computational costs. The Inception V3 can be trained faster with the help of batch normalization, which normalizes mini-batch activations. The network was trained with the ImageNet dataset, which has millions of classified pictures in hundreds of categories. Pre-training provides the model with a rich set of characteristics that may be fine-tuned or used for other image recognition tasks. The design outperforms other image classification architectures on image classification benchmarks and is widely utilized in research and practice. ReLu transforms all the negative value outputs to Zero and doesn't change the positive value outputs. In Inception V3, ReLu activation is applied after the convolutional layers and dense layers. Mathematically it can be defined as

$$f(x) = max(0, x) \tag{1}$$

here, 'x' denotes the input of the activation function and 'f(x)' denotes the output.

2) VGG 19

VGG19 is a convolutional neural network architecture. It is an upgrade to the VGG16 architecture with 19 layers. The VGG19 architecture consists of nineteen layers, including convolutional, pooling, and three dense layers. The employment of modest 3×3 filters in each convolutional layer increases feature representation acquisition in the architectural design. One of the main aims of the network is to reduce the spatial dimensionality of the input data and it can be achieved by including max-pooling layers to perform feature map-down sampling. The VGG19 architecture may be formally expressed as a series of nonlinear transformations that enable mapping an input picture to a probability distribution across a preset range of classes.

The network's layers may be described as a mathematical function that receives an input feature map and performs a convolution operation with a set of adaptive filters. This is then followed by a non-linear activation function, such as ReLU, and possibly a pooling procedure. The output of each layer is then used as an input for the next layer. In the training phase, backpropagation and stochastic gradient descent algorithms help to update the weights of convolutional layer filters. The difference between the expected probability distribution and the actual distribution is called loss function and the main goal of the network is to minimize it throughout the dataset.

3) EFFICIENTDET

EfficientDet is an architecture-based family of object detection models. The EfficientDet models are intended to attain a high level of precision while conserving computational resources. The models with larger numbers (D7 being the largest) have more parameters and can capture more complex characteristics, but they also require more computational resources. EfficientDet models use a compound scaling technique that simultaneously scales the model's dimension, and resolution to achieve a reasonable balance between accuracy and efficiency. By modulating these dimensions, EfficientDet models can be adapted to various resource constraints and task specifications. EfficientDet models integrate multiple techniques for efficient object detection, such as feature pyramid networks, bi-directional feature pyramid networks, and weighted feature fusion. These techniques allow models to effectively capture multiscale features and enhance the accuracy of object detection for a versatile amount of object sizes. The EfficientDet models have attained SOTA performance on prominent object detection standards such as the COCO dataset (Common Objects in Context) and have demonstrated greater accuracy than previous models while maintaining efficiency. On the basis of the EfficientNet architecture, EfficientDet models are a family of effective and precise object detection models.

4) RESNET50

ResNet-50 is a Convolutional Neural Network that consists of 50 layers of Residual Network architecture. ResNet-50 is distinguished by its residual connections, also known as shortcut or skip connections. Data can flow directly through the network's connections from one layer to the next. The problem of vanishing gradients can be solved through this technique and helps in training multilayer neural networks. ResNet-50 has unused blocks. In the residual block, multiple convolutional layers are followed by batch normalization and ReLU activation. Through element-wise addition between the input and output of the residual block, skip connections are generated. Gradients can propagate effectively during training due to this phenomenon, which creates a condensed information channel. The input image is sampled by a max-pooling layer following a convolutional layer in ResNet-50.

The number of remaining blocks fluctuates across the four phases. Following the initial residual block are three phases with progressively greater spatial resolution. The network successively decreases the spatial dimensions while increasing the number of filters to capture both low-level and high-level data. Global average pooling includes spatial information into a fixed-dimensional feature vector in the network's upper layer. Then, a fully connected layer and a SoftMax layer are used to determine probabilities for photo categorization classes. Mathematically SoftMax can be defined as

$$\mathrm{Softmax}(z(i)) = \frac{\exp(z(i))}{\sum_j \exp(z(j))} \tag{2}$$

For i = 1 to n exp () represents the exponential function and sum () calculates the sum of the exponential values of all elements in the vector.

5) VIT (VISION TRANSFORMER)

ViT categorizes photos using the Transformer design. ViT can learn in the absence of built-in features. ViT, like NLP, interprets images as patches. ViT linearly converts an input image's grid of fixed-size patches into a high-dimensional feature space. Objects are located using transformer decoders utilizing patch and positional embeddings. Transformer encoder patch embeddings are handled by self-attention layers and position-wise feed-forward networks. The overall image and local interactions between patches may provide global context to the model. After numerous Transformer layers, a classification head adds class names to the model. ViT can handle a variety of picture sizes thanks to fixed-size patches. Unlike convolutional neural networks (CNNs), ViT may use photographs without scaling or cropping them. ViT runs admirably on ImageNet. Self-attention methods need more training data than CNNs. Transformer-based systems hold enormous promise in computer vision, particularly for picture categorization. The activation function used here is GELU and is mathematically defined as

$$\mathrm{GELU}(x) = 0.5 \cdot x\left(1 + \tanh\left(\sqrt{\frac{2}{\pi}} \cdot \left(x + 0.04 \cdot x^3\right)\right)\right) \tag{3}$$

Here tanh represents the hyperbolic tangent function, sqrt is the square root and pi is the mathematical constant pi

6) YOLOv8

Ultralytics created an innovative object identification and image segmentation model known as YOLOv8. It is the YOLOv5 model's replacement and offers a lot of enhancements in terms of speed, accuracy, and adaptability. Anchor-free detection of YOLOv8 is the latest approach to object detection that does not use pre-defined anchor boxes to predict object bounding boxes. This increases the efficiency of the performance as it is not just limited to the predefined box sizes. Moreover, YOLOv8 can also achieve high accuracy with a fast inference speed, as it is also flexible and versatile, it can be used for an extensive variety of object detection tasks.

YOLOv8 is a single forward pass neural network and it can predict the bounding boxes and class probabilities in an image with that single pass. The model divides the image into a nxn grid of cells, these cells will be helpful for the prediction of bounding boxes and class probabilities. After the prediction, the model uses a non-maximum suppression (NMS) algorithm to remove overlapping bounding boxes, by this only the most confident bounding boxes will be selected. In this study, after receiving the input image from the user, the fracture can be detected based on the annotated images of the dataset with the help of the convolution and pooling layers in the architecture. YOLOv8 employs a feature pyramid network (FPN) as its backbone architecture. FPN is instrumental in handling objects of various scales by capturing multi-scale features. This allows YOLOv8 to effectively detect both small and large objects within an image. The FPN consists of a series of convolutional layers that extract features at different resolutions, facilitating accurate object localization and classification. This architecture is shown in Fig. 7

Activation functions have a noteworthy influence on the performance of the model. In the YOLOv8 algorithm, Leaky ReLU, Swish, and Tanh are the most commonly used as they give the best results for classification tasks by avoiding the vanishing gradient problem.

Leaky ReLU function:

$$x = max(a * x, x) \tag{4}$$

Swish activation function:

$$x = x * sigmoid(beta * x) \tag{5}$$

Tanh activation function:

$$x = \frac{e^x - e^{-x}}{e^x + e^{-x}} \tag{6}$$

7) YOLO NAS

The YOLO-NAS framework incorporates quantization-aware blocks and selective quantization techniques to achieve optimal performance. The model, upon conversion to its INT8 quantized variant, exhibits a marginal decrease in precision, thereby showcasing a notable enhancement compared to alternative models. The aforementioned advancements ultimately converge to yield an architecture of superior quality, characterized by unparalleled object detection capabilities and exceptional performance. The YOLO-NAS framework presents a novel fundamental unit that exhibits favorable compatibility with quantization, thereby effectively mitigating a notable constraint observed

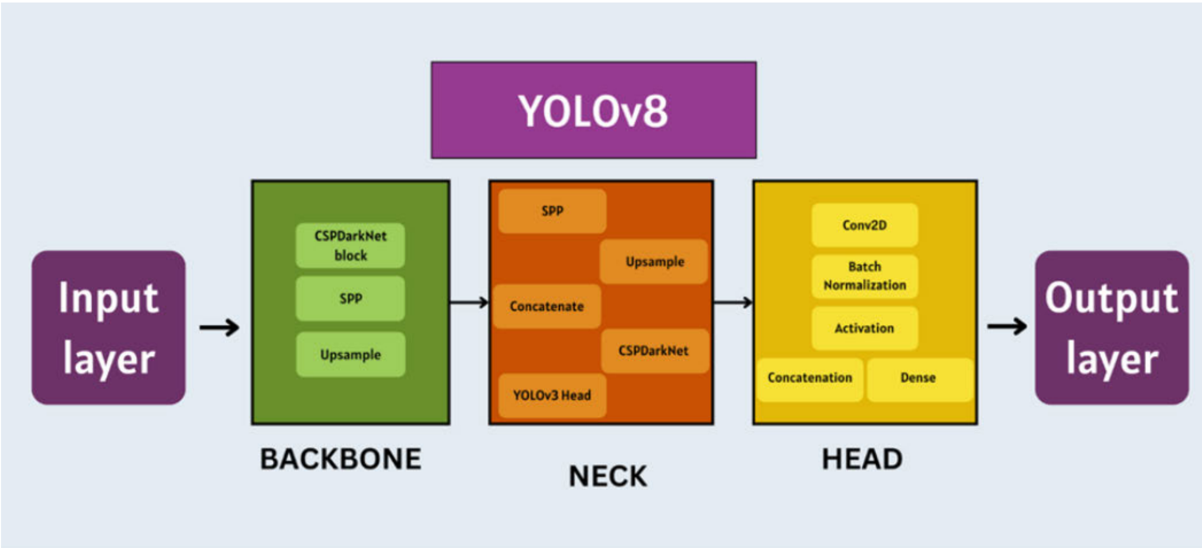

FIGURE 7. **YOLOv8 architecture.**

in prior YOLO architectures. The YOLO-NAS framework effectively utilizes sophisticated training methodologies and subsequent quantization techniques to augment its overall performance. The YOLO-NAS framework employs Auto NAC optimization techniques and has undergone pre-training on well-established datasets including COCO, Objects365, and Robo flow 100. The utilization of pre-training renders it highly conducive for object detection tasks in downstream production environments. The architecture of YOLO NAS is shown in Fig. 8, unlike manually designed architectures, which may not always be optimal for specific tasks or datasets, YOLO NAS can tailor architectures to suit particular requirements. This customization can lead to models that are better suited for real-time applications, resource-constrained environments, or specialized domains. YOLO NAS allows for the exploration of diverse architectural configurations, enabling the discovery of novel designs that might not have been considered through manual design processes.

In YOLO NAS the activations that can be used are:

Mish:

$$f(x) = x \cdot \tanh(\text{softplus}(x)) \tag{7}$$

It provides a smoother activation process

Hard Swish:

$$f(x) = x \cdot \min(\max(0, x + 3), 6) \tag{8}$$

This is computationally effective and is useful in resource-constrained environments

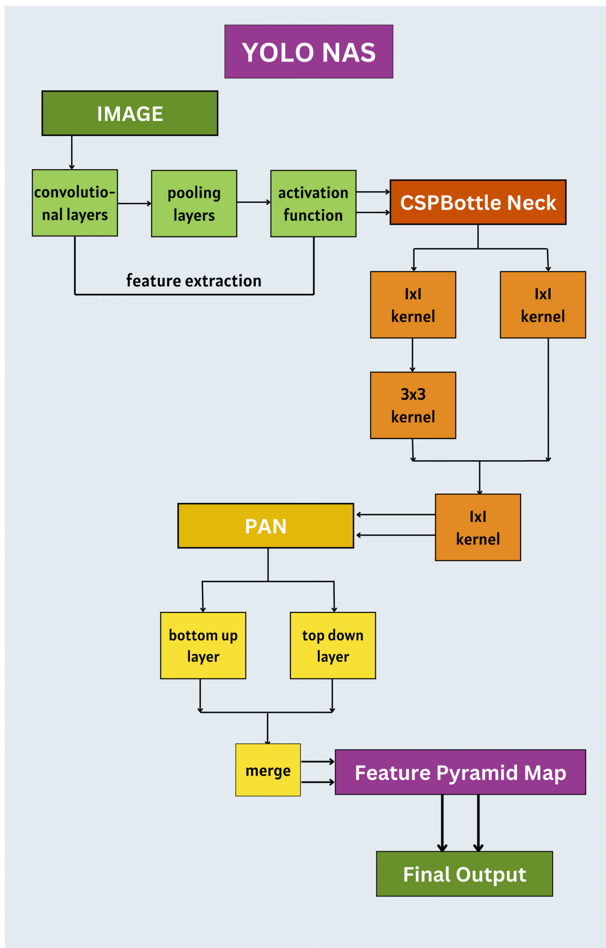

FIGURE 8. **YOLO NAS architecture.**

D. PROPOSED MODEL

1) DESCRIPTION

YOLO NAS is a cutting-edge object detection algorithm, which means that it excels in many ways when compared to traditional algorithms. Some of the many benefits of the YOLO NAS algorithm are precise accuracy, speed, efficiency, versatility, and customization. All these features help us in developing a hybridized YOLO NAS model which will be more efficient and accurate. Hybridizing an algorithm increases the model's strengths. When hybridizing an algorithm there are many ways to do so, in this study, the method used was training two algorithms at the same time, then again training with two other algorithms.

The training images dataset will be provided to the YOLO NAS and Efficient Det algorithms to train the model. The YOLO NAS algorithm helps in extracting the high-level features from the image such as objects, attributes, and relations, and the Efficient Det helps in extracting the low-level features such as edges, local patterns, and histograms. Then both algorithms can merge the features they individually extracted at the fusion station to obtain the first output image. Allowing two algorithms to extract features helps by maintaining preciseness while labeling the bounding boxes. Training epochs are one of the hyperparameter that was considered for the model, the number of epochs trained was 120 and it took about 4 hours for the model to train completely.

Then the first output image will be sent to YOLO NAS to fast screen the image and perform initial object detection, then the image will be sent to the DETR3 transformer which refines the image and confirms the detections for the most precise accuracy. This DETR3 transformer is best known for its efficiency and improved generalizability, as it is also flexible, it is not a complex task to add this transformer in a hybrid model. This step provides the second output image. Then once again the image will be sent to YOLO NAS, where to prioritize the crucial regions and ignore the background, an attention mechanism will be used. This step helps in reducing the noise and providing the best result, that is the final output image.

While the provided text describes the YOLO NAS hybridization process in detail, it doesn't directly offer a single mathematical formula representing the entire algorithm. Hybridization often involves combining different algorithms that each have their complex formulas, making a single encompassing formula quite cumbersome. However, the key components and offer relevant formulas for each:

Feature Extraction:

YOLO NAS: This uses convolutional neural networks (CNNs) for feature extraction. Specific formulas involved depend on the chosen CNN architecture and activation functions. One example of a basic CNN layer:

$$Output[i, j, k] = Activation(\Sigma m, n, l, W[i, j, k, m, n, l] * Input[i - m, j - n, k - l] + b) \quad (9)$$

EfficientDet: This also uses CNNs, potentially with different architectures than YOLO NAS. Similar formulas are applicable here.

Feature Fusion:

The fusion station likely applies element-wise operations like addition or concatenation to combine feature maps from both algorithms. For element-wise addition:

$$FusionOutput[i, j, k] = YOLOFeature[i, j, k] + EfficientDetFeature[i, j, k] \quad (10)$$

Refinement and Detection:

YOLO NAS (fast screening): This could involve filtering potential object locations based on probabilities learned by the YOLO NAS model. Specific formulas depend on the chosen architecture and output representation. DETR3 transformer: This uses a transformer architecture with attention mechanisms for refining detections. Formulas involve complex matrix operations for attention calculations and final predictions.

YOLO NAS uses an attention mechanism to prioritize regions. You can find various attention mechanism formulas, such as:

$$\text{AttentionWeight}[i] = \text{softmax}\left(\sum_j Q[i] \cdot K[j] \cdot V[j]\right) \quad (11)$$

This equation calculates an attention weight for each region based on query, key, and value vectors extracted from the features.

2) HYBRIDIZATION

The proposed method is to utilize three algorithms of deep learning to get the most accurate results and the algorithms are YOLO NAS, Efficient Det, and DETR 3. Initially, the YOLO NAS and Efficient Det algorithms are employed simultaneously to extract high-level and low-level features respectively from training images, ensuring a comprehensive understanding of the visual data. These features are then fused to generate the first output image, preserving precision in labeling bounding boxes. Subsequently, the first output image undergoes a refined detection process. It first passes through YOLO NAS for fast screening and initial object detection, followed by refinement by the DETR3 transformer, renowned for its efficiency and adaptability. This transformer further enhances the accuracy of detections, producing the second output image. Finally, the image is once again processed by YOLO NAS, incorporating an attention mechanism to prioritize crucial regions and suppress background noise, resulting in the final output image. This hybrid model harnesses the capabilities of YOLO NAS, Efficient Det, and DETR3 transformer, utilizing a total of three algorithms to maximize accuracy and efficiency in object detection tasks. Fig. 9 shows the proposed model architecture.

3) DATA AUGMENTATION

In the realm of computer vision, particularly when dealing with medical imaging like hand bone X-rays, data augmentation plays a pivotal role in bolstering the robustness of deep learning models. This expanded dataset fuels the

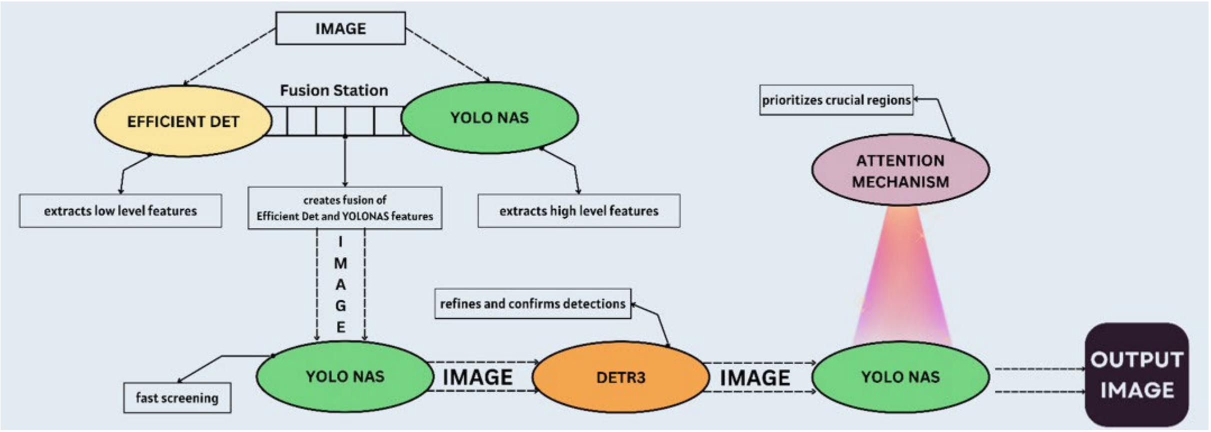

FIGURE 9. YOLO NAS hybridized model architecture.

model's ability to recognize subtle patterns and variations in hand X-rays, potentially leading to improved performance in tasks like fracture detection, bone abnormality classification, or skeletal age assessment. The main motive of the study is to create a model with great performance, so data augmentation is applied. This includes steps like cropping the image and increasing the brightness and saturation of an image. In this study, as all the images of hand bone X-rays were grayscale, through data augmentation, the image's brightness was increased for better understanding. This process helps in bringing versatility to the model's training dataset.

IV. EXPERIMENT AND RESULTS

The dataset used is a hybrid one, which was a mix of various best-quality X-ray images and it has a sufficient number of samples for training and testing, making it appropriate for persistent approval. After model training, the testing dataset was used to judge the functioning of the VGG19, Res Net 50, InceptionV3, Vision Transform, and Efficient Det models. Several commonly used criteria, including accuracy, precision, sensitivity (recall), and F1 score, were utilized to evaluate the performance. The accuracy of the model's predictions as a whole or the classification accuracy of the validation (training) data are both measured. The number of true positives, true negatives, false positives, and false negatives was calculated using a confusion matrix to help assess how well the proposed approach worked. The metrics to evaluate the performance of the model:

Accuracy: It is calculated using the proportion of all successfully predicted photos to all test images [15].

$$\frac{TP + TN}{TP + FP + TN + FN} \tag{12}$$

Precision: The precision can be calculated by dividing actual results by the total number of true positives [15].

$$\frac{TP}{TP + FP} \tag{13}$$

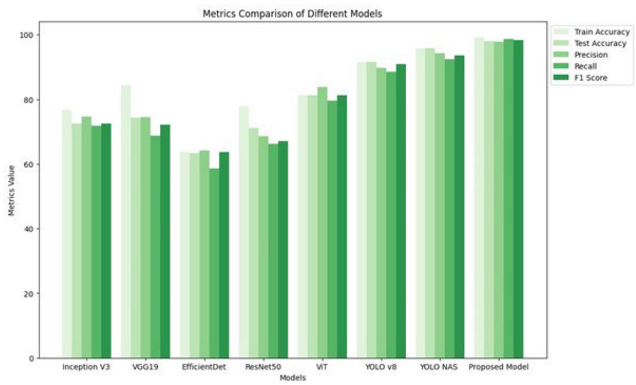

FIGURE 10. Comparing metrics of existing methodologies.

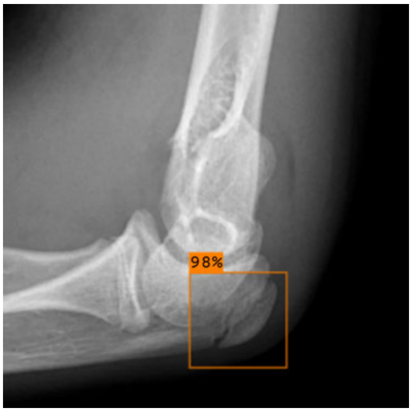

FIGURE 11. Predicted image 1.

Recall: It is calculated by dividing the total number of positive samples by the total number of predictions [15].

$$\frac{TP}{TP + FN} \tag{14}$$

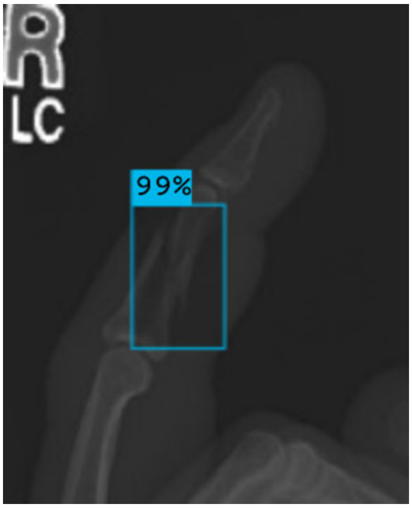

FIGURE 12. Predicted image 2.

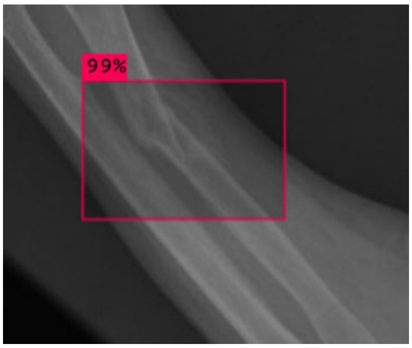

FIGURE 13. Predicted image 3.

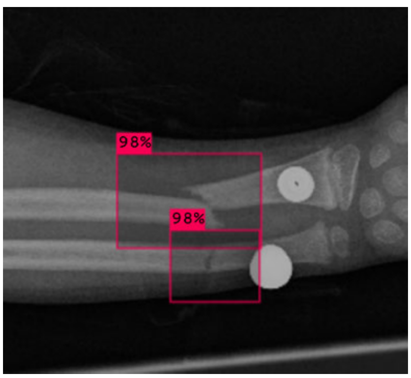

FIGURE 14. Predicted image 4.

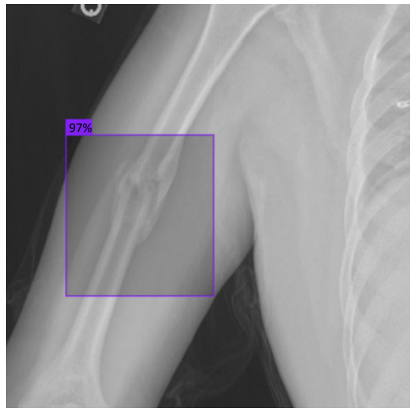

FIGURE 15. Predicted image 5.

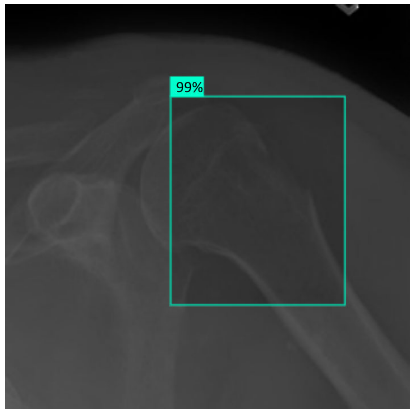

FIGURE 16. Predicted image 6.

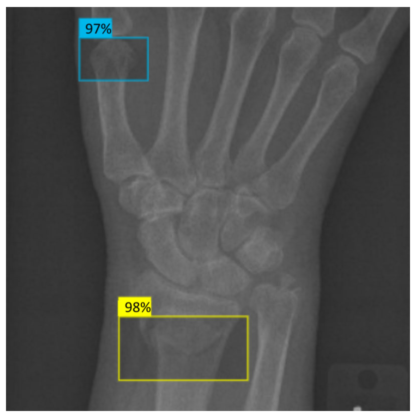

FIGURE 17. Predicted image 7.

F1-Score: The F1-Score measures the harmonic mean of the model performance [15].

$$F1 = \frac{2 \cdot \text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}} \tag{15}$$

Table 2 and Fig. 10 are used to analyze the performance metrics of the seven existing methodologies and the proposed methodology on the dataset. The dataset that is used for all the algorithms is the same dataset that is considered for training the proposed model, and it consists of 6 classes. The ResNet50 algorithm performed comparatively poorly with train and test accuracies of only 77.91 and 71.21. In the case of ResNet50, we can see the overfitting clearly with the differences in train and test accuracies, this goes the same with Efficient Det. Among the existing methodologies, YOLO NAS has the best performance with a training accuracy of 95.84 and a testing accuracy of 95.82, and the

TABLE 2. Performance metrics of various models.

S. No	Model Name	Train Accuracy	Test Accuracy	Precision	Recall	F1 Score
1	Inception V3	76.72	72.50	74.64	71.83	72.52
2	VGG19	84.25	74.30	74.46	68.76	72.20
3	EfficientDet	63.72	63.37	64.19	58.62	63.67
4	ResNet50	77.91	71.21	68.57	66.30	67.16
5	ViT	81.30	81.27	83.35	79.63	81.28
6	YOLO v8	91.65	91.60	89.76	88.59	90.84
7	YOLO NAS	95.84	95.82	94.36	92.47	93.61
8	Proposed Model	99.20	98.10	97.85	98.72	98.39

TABLE 3. Comparison between proposed and existing models.

S. No	Author	Method	Accuracy% / mAP / Sensitivity
1	Sumi et al. [21]	CNN	80%
2	Rui Yang Ju et al. [5]	YOLOv8	0.631 (mAP)
3	Mohamed A Kassem et al. [12]	Transfer learning-based Deep learning	98.5%
4	Sun et al. [22]	Contrastive Learning	0.873 (sensitivity)
5	Srinivas et al. [23]	CNN-Transfer Learning	80%
6	Proposed	Hybridized YOLO NAS	99.20%

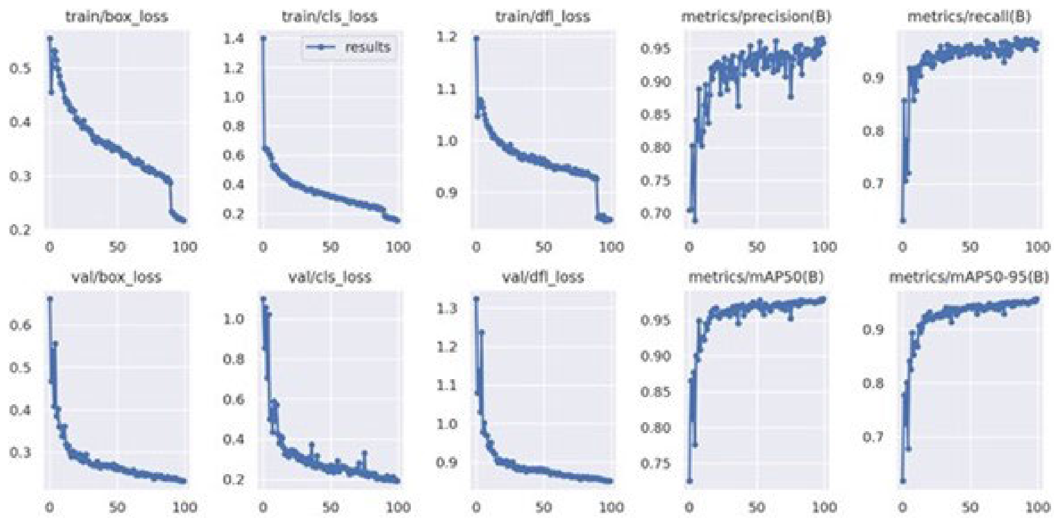

FIGURE 18. Loss function graphs of the proposed scheme.

proposed model showed the best accuracy of 99.20 and 98.10 when compared to all algorithms.

Table. 3 is about the results of various studies in the bone fracture domain. Most of these studies lack novelty in the methodology. The performance of these methods is assessed in various methods, and the experiment that used the YOLOv8 algorithm for fracture detection in the wrist shows the highest mAP of 0.631.

The proposed research employs a formal deep learning approach by leveraging several prominent models, including Inception V3, Vision Transformer, ResNet 50, VGG 19, and EfficientDet, in the context of various applications. Among these models, the proposed hybrid algorithm stood out as it demonstrated superior performance, surpassing the accuracy of all the other models. Notably, it achieved an impressive mean average precision of 97.85 in hand bone fracture identification tasks.

One of the most commonly used metrics for object detection tasks is mean average precision (mAP). It can be calculated by the number of True Positives, False Positives, and False Negatives of each class, and the mean of all classes is defined as the mean Average Precision. In addition to its exceptional precision rate, the proposed model exhibited significantly lower error values than the other models, further emphasizing its effectiveness and reliability in these domains. The findings highlight the algorithm's potential as a versatile and high-performing solution for automated systems, capable of accurate and efficient detection tasks in diverse fields of application.

Fig. 11 to Fig. 17 are the predicted images, which are the output images of the proposed hybrid model and the probability score of the respective class is also marked in the output images to evaluate the performance of a particular class of the model in a specific input image.

TABLE 4. Performance of proposed scheme.

Class	Instances	Precision	Recall	mAP50	mAP(50-95)
All	932	0.989	0.988	0.991	0.988
Elbow Fracture	178	0.975	0.986	0.975	0.984
Fingers Fracture	195	0.985	0.972	0.987	0.977
Forearm Fracture	132	0.973	0.987	0.989	0.981
Humerus Fracture	157	0.965	0.988	0.985	0.986
Shoulder Fracture	136	0.968	0.978	0.978	0.977
Wrist Fracture	134	0.979	0.968	0.973	0.976

Table 4 is made upon analyzing the metric values obtained by the proposed model. The metrics used for the evaluation of the model are Precision, Recall, and, mAP (Mean Average Precision), the Mean Average Precision can be calculated either for a fixed threshold or for a range of thresholds. When compared, the class forearm fracture has the highest mAP of 0.989 and the class Wrist fracture has the lowest mAP value of 0.973. When considered for the whole, the proposed model obtained a good precision rate of 0.989, which shows that the proposed model is great for the detection of hand bone and joint fractures.

Fig. 18 is the graph of the loss functions it illustrates the values of the loss metric throughout the training process for both the classification and regression components of the model. The classification branch and the regression branch use binary cross-entropy loss and distribution focal loss respectively. The loss function graph serves as a valuable tool for monitoring the training process and detecting areas in which the model fails to converge. If the loss values for either branch fail to exhibit a drop, it may become imperative to update the hyperparameters of the model or procure supplementary training data.

V. CONCLUSION

The developed hybrid model was used in this study for the detection of hand bone fractures and it gave promising results. Bone injuries require quick treatment and the algorithm can make real-time detections quick and easy. The performance and the accuracy results can be varied based on different aspects such as the quality of the dataset, hyperparameters, loss function, and optimization. Overall, finding the optimal balance between these factors provided us with the best accuracy.

Further to expand the study, the focus can be on customizing the algorithm more deeply and increasing the dataset. Customizing an algorithm has several benefits, it increases the model's efficiency and can provide novel architectures. It allows the researchers to adapt models and techniques to all scenarios and results in novel solutions to complex problems. The increasing dataset can provide the best results by reducing overfitting and handling rare events. The quality and diversity of the data play a vital role in the performance of a model and training a large dataset allows us to create better deep learning models.

REFERENCES

[1] InformedHealth.org [Internet]. (2006). *How Do Hands Work*. Inst. Quality Efficiency Health Care (IQWiG), Cologne, Germany. Accessed: Aug. 31, 2010. [Online]. Available: https://www.ncbi.nlm.nih.gov/books/NBK279362/

[2] *Take Matters Into Your Own Hands: Understanding Hand Injuries in the Workplace*. Accessed: Aug. 1, 2018. [Online]. Available: https://ohsonline.com/Articles/2018/08/01/Take-Matters-into-Your-Own-Hands.aspx?m=1.&Page=1

[3] A. Pinto, D. Berritto, A. Russo, F. Riccitiello, M. Caruso, M. P. Belfiore, V. R. Papapietro, M. Carotti, F. Pinto, A. Giovagnoni, L. Romano, and R. Grassi, "Traumatic fractures in adults: Missed diagnosis on plain radiographs in the emergency department," *Acta Bio Medica, Atenei Parmensis*, vol. 89, no. 1, pp. 111–123, 2018, doi: 10.23750/abm.v89i1-S.7015.

[4] *Broken Hand Symptoms and Causes by Mayo Clinic*. Accessed: Jun. 30, 2022. [Online]. Available: https://www.mayoclinic.org/diseases-conditions/broken-hand/symptoms-causes/syc-20450240

[5] R.-Y. Ju and W. Cai, "Fracture detection in pediatric wrist trauma X-ray images using YOLOv8 algorithm," 2023, *arXiv:2304.05071*.

[6] K. D. Ahmed and R. Hawezi, "Detection of bone fracture based on machine learning techniques," *Meas., Sensors*, vol. 27, Jun. 2023, Art. no. 100723, doi: 10.1016/j.measen.2023.100723.

[7] S. R. Karanam, Y. Srinivas, and S. Chakravarty, "A systematic review on approach and analysis of bone fracture classification," *Mater. Today, Proc.*, vol. 80, pp. 2557–2562, Jan. 2023, doi: 10.1016/j.matpr.2021.06.408.

[8] L. Xue, W. Yan, P. Luo, X. Zhang, T. Chaikovska, K. Liu, W. Gao, and K. Yang, "Detection and localization of hand fractures based on GA_Faster R-CNN," *Alexandria Eng. J.*, vol. 60, no. 5, pp. 4555–4562, Oct. 2021, doi: 10.1016/j.aej.2021.03.005. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1110016821001605

[9] I. Kandel, M. Castelli, and A. Popovič, "Musculoskeletal images classification for detection of fractures using transfer learning," *J. Imag.*, vol. 6, no. 11, p. 127, Nov. 2020, doi: 10.3390/jimaging6110127.

[10] K. C. Santos, C. A. Fernandes, and J. R. Costa, "Feasibility of bone fracture detection using microwave imaging," *IEEE Open J. Antennas Propag.*, vol. 3, pp. 836–847, 2022, doi: 10.1109/OJAP.2022.3194217.

[11] A. Inui, Y. Mifune, H. Nishimoto, S. Mukohara, S. Fukuda, T. Kato, T. Furukawa, S. Tanaka, M. Kusunose, S. Takigami, Y. Ehara, and R. Kuroda, "Detection of elbow OCD in the ultrasound image by artificial intelligence using YOLOv8," *Appl. Sci.*, vol. 13, no. 13, p. 7623, Jun. 2023, doi: 10.3390/app13137623.

[12] M. A. Kassem, S. M. Naguib, H. M. Hamza, M. M. Fouda, M. K. Saleh, and K. M. Hosny, "Explainable transfer learning-based deep learning model for pelvis fracture detection," *Int. J. Intell. Syst.*, vol. 2023, Jun. 2023, Art. no. 3281998, doi: 10.1155/2023/3281998.

[13] B. Zhang, C. Jia, R. Wu, B. Lv, B. Li, F. Li, G. Du, Z. Sun, and X. Li, "Improving rib fracture detection accuracy and reading efficiency with deep learning-based detection software: A clinical evaluation," *Brit. J. Radiol.*, vol. 94, no. 1118, Feb. 2021, Art. no. 20200870, doi: 10.1259/bjr.20200870.

[14] M. S. A. M. Bevers, C. E. Wyers, A. M. Daniels, E. A. Audenaert, S. M. J. van Kuijk, B. van Rietbergen, P. P. M. M. Geusens, S. Kaarsemaker, H. M. J. Janzing, P. F. W. Hannemann, M. Poeze, and J. P. van den Bergh, "Association between bone shape and the presence of a fracture in patients with a clinically suspected scaphoid fracture," *J. Biomech.*, vol. 128, Nov. 2021, Art. no. 110726, doi: 10.1016/j.jbiomech.2021.110726. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0021929021004930

[15] D. P. Yadav, A. Sharma, S. Athithan, A. Bhola, B. Sharma, and I. B. Dhaou, "Hybrid SFNet model for bone fracture detection and classification using ML/DL," *Sensors*, vol. 22, no. 15, p. 5823, Aug. 2022, doi: 10.3390/s22155823.

[16] Y. Ma and Y. Luo, "Bone fracture detection through the two-stage system of crack-sensitive convolutional neural network," *Informat. Med. Unlocked*, vol. 22, Dec. 2021, Art. no. 100452, doi: 10.1016/j.imu.2020.100452. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S235291482030602X

[17] E. Li and Q. Tan, "Role of ultrasound imaging to assess and diagnose various body fractures: Systemic review and meta-analysis," *J. Radiat. Res. Appl. Sci.*, vol. 15, no. 3, pp. 357–364, Sep. 2022, doi: 10.1016/j.jrras.2022.08.008. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1687850722082243

[18] Y. Rakesh and A. Akilandeswari, "Evaluation of bernsen algorithm for bone fracture detection based on edge detection," *J. Surv. Fisheries Sci.*, vol. 10, no. 1, pp. 2060–2068, 2023.

[19] D. Mishra and A. K. Mishra, "Evolution of artificial intelligence in bone fracture detection," *Int. J. Reliable Quality E-Healthcare*, vol. 11, no. 2, pp. 1–17, Jul. 2022, doi: 10.4018/ijrqeh.299958.

[20] J. Terven and D. Cordova-Esparza, "A comprehensive review of YOLO architectures in computer vision: From YOLOv1 to YOLOv8 and YOLO-NAS," 2023, *arXiv:2304.00501*.

[21] T. A. Sumi, N. Basnin, M. S. Hossain, K. Andersson, and M. S. Hoassain, "Classifying humerus fracture using X-ray images," in *Proc. 4th Ind. Revolution Beyond*, in Lecture Notes in Electrical Engineering, vol. 980, Singapore, M. S. Hossain, S. P. Majumder, N. Siddique, and M. S. Hossain, Eds. Singapore: Springer, 2023, pp. 527–538, doi: 10.1007/978-981-19-8032-9_37.

[22] H. Sun, X. Wang, Z. Li, A. Liu, S. Xu, Q. Jiang, Q. Li, Z. Xue, J. Gong, L. Chen, Y. Xiao, and S. Liu, "Automated rib fracture detection on chest X-ray using contrastive learning," *J. Digit. Imag.*, vol. 36, no. 5, pp. 2138–2147, Jul. 2023, doi: 10.1007/s10278-023-00868-z.

[23] P. Srinivas, J. Ahmed, and M. B. Chander, "Maxillofacial fracture detection system in accident victims using convolution neural network (CNN) transfer learning," *Eur. Chem. Bull.*, vol. 12, no. 5, pp. 800–807, 2023.

SAI CHARAN MEDARAMATLA is currently pursuing the degree with VIT-AP University, with a focus on machine learning.

He is a dedicated Computer Science and Engineering Student with a specialization in artificial intelligence and machine learning. He has demonstrated his research acumen by publishing impactful articles in the field of machine learning. His research article titled *Prognoza: Parkinson's Disease Prediction Using Classification Algorithms*, conducted under the guidance of esteemed faculty showcases his expertise in utilizing classification algorithms for predictive analytics. Additionally, his contribution to *An Enhanced Laryngeal Tissue Classification Model Using Deep Learning Techniques* further exemplifies his proficiency in leveraging deep learning methodologies for medical image analysis. Apart from his academic achievements, he has undertaken several practical projects involving the application of detection and classification techniques in machine learning and deep learning domains. His projects underscore his proficiency in Python programming and his adeptness in AI, ML, and data analysis. His innovative contributions extend beyond academia, as evidenced by his involvement in patenting three inventions in the realm of artificial intelligence and machine learning. These patents reflect his entrepreneurial spirit and his commitment to driving technological advancements in the field.

CHENNUPATI VEDA SAMHITHA is currently pursuing the bachelor's degree in computer science and engineering with VIT-AP University, with a focus interest on artificial intelligence (AI) and machine learning (ML).

Throughout her academic journey, she has consistently demonstrated a passion for advancing the frontiers of AI through practical applications and scholarly pursuits. Her academic achievements are underscored by her involvement in numerous projects that span a diverse range of machine learning and deep learning algorithms. These projects serve as tangible manifestations of her ability to address real-world challenges effectively and contribute meaningfully to the advancement of AI technologies. In her pursuit of continuous learning, she has completed the prestigious Machine Learning Specialization course offered by Stanford University and Deeplearning.AI. This rigorous program has equipped her with a comprehensive understanding of advanced ML techniques and methodologies, further enhancing her capabilities in the field. Moreover, her practical prowess is evident through her contributions to the field, as demonstrated by her publication of three patents under the guidance of her faculty. These patents, focusing on marine animal sound classification, bone fracture detection, and road damage detection, showcase her innovative thinking and her ability to apply ML and DL techniques to solve real-world problems.

SAGAR DHANRAJ PANDE received the Ph.D. degree in computer science and engineering from Lovely Professional University, Phagwara, Punjab, India, in 2021.

He is currently an Assistant Professor with Pimpri Chinchwad University, Pune, Maharashtra, India. He has published and presented more than 80 articles in Springer, Elsevier, CRC, Taylor & Francis, and other reputable journals, which are Scopus-indexed and peer-reviewed journals. Also, he has published more than 30 patents on the topics of data mining, network security, the IoT, and their application. Also, he has published papers at international conferences Springer on the topics of data mining, network security, the IoT, and their application. He has supervised several postgraduate students in cybersecurity, computer networks, communication, and the IoT. He is responsible for teaching artificial intelligence, deep learning, machine learning, cybercrime and security, and Python programming courses to undergraduate and postgraduate students. His research interests include deep learning, machine learning, network attacks, cyber security, and the Internet of Medical Things (IoMT). He received the Emerging Scientist Award from the VDGOOD Professional Association, in 2021, and the Young Researcher Award and the Best Ph.D. Thesis Award from the Universal Innovators, in 2022.

SURENDRA REDDY VINTA received the Ph.D. degree from VBSP University, India, under the supervision of Prof. (Dr.) Sourab Paul and co-supervised by Dr. Raju Bukya.

He is currently an Associate Professor with the School of Computer Science and Engineering, VIT-AP University, Amaravati, Andhra Pradesh, India. Under his guidance four students are working for the Ph.D. degree. A seasoned academician having more than 15 years of experience, he has published four books, such as *Programming in C*, *Programming in C++*, *Machine Learning*, and *Deep Learning* (I. K. International Publishing House Pvt. Ltd.) He has undergone industrial training programs during which he was involved in live projects with companies in the areas of SAP, railway traffic management systems, and visual vehicles counter and classification (used in the Metro rail network design). A prolific writer, he has published twenty patents and authored many research articles in web of science indexed journals. Additionally, he has presented research papers at many conferences in the areas of image processing, machine learning, deep learning, NLP, computer vision, features extraction, and programming, such as digital image processing, feature extraction, machine learning, deep learning, NLP, computer vision, C, Python, data structure, C++, and Java. Also, he has a Professional Memberships of IEEE, ISTE, and IET.