

Comparative Analysis Of YOLO And Transformer Based Architectures For Animal X-Ray Anatomy Recognition

G.G. Rajput¹, Sumitra M Mudda¹, Geeta Babusingh²

¹Department of Computer Science, Karnataka State Akkamahadevi Women University
Vijayapura, Karnataka.

²SSA Government Degree College, Ballary, Karnataka

Abstract

Accurate interpretation of veterinary radiographs is essential for diagnosis, treatment planning, and surgical decision-making in animal healthcare. However, manual analysis of X-ray images is time-consuming and requires specialized expertise. Recent advances in deep learning have demonstrated significant potential in automating medical image analysis; however, research in veterinary imaging—especially in the classification of anatomical body parts in animal radiographs—remains limited. This paper presents comparison of popular deep learning models, namely , YOLOv8, YOLOv11, Vision Transformer (ViT), and ResNet50, for automatically identifying body parts in dog X-rays. The models were trained and tested on a carefully prepared dataset of 453 dog X-ray images, covering five body regions: skull, chest, abdomen, front leg, and back leg. Models were assessed on accuracy, precision, recall, F1-score, and inference speed. Experimental results demonstrate that YOLO-based models achieve competitive accuracy while offering significantly faster inference compared to ViT and ResNet50, making them suitable for real-time veterinary applications. The Vision Transformer achieved strong classification performance on limited data due to its attention-based feature learning, while ResNet50 served as a reliable baseline. The findings presented in this paper can support veterinary diagnostic systems and assist practitioners in clinical workflows.

1 Introduction

1.1 Background

Radiography is one of the most widely used diagnostic imaging techniques in veterinary medicine due to its accessibility, affordability, and capability to visualize skeletal and soft-tissue structures. Accurate interpretation of radiographic images is essential for identifying fractures, thoracic and abdominal abnormalities, orthopedic disorders, and traumatic injuries in animals. However, radiograph interpretation requires specialized expertise, significant clinical experience, and is often time-consuming.

Recent advances in Deep Learning (DL) have reshaped medical image analysis. Deep learning models, particularly Convolutional Neural Networks (CNNs) and Transformer-based architectures, have achieved remarkable performance in human radiology applications. However, veterinary radiology remains less explored due to limited annotated datasets, variability among animal species and breeds, and the relatively small number of research studies in this domain.

Anatomical body-part classification plays an essential role in automated radiographic workflows, assisting in proper triaging, structured diagnosis, and dataset organization. Despite its importance, research on body-part classification in animal X-rays remains limited, offering a valuable opportunity to explore and benchmark modern deep learning methods for this task.

1.2 Motivation

Veterinary clinics, particularly those in resource-constrained environments, face challenges such as limited availability of expert radiologists, increasing diagnostic workload, and variability in radiographic interpretation. Automated classification of

canine X- ray body parts can support veterinarians by improving workflow efficiency, assisting junior clinicians, and enabling intelligent decision-support systems. Modern deep learning models such as ResNet and Vision Transformers (ViT) have demonstrated strong performance in medical imaging tasks, while real-time applications require models capable of maintaining both high accuracy and fast inference speed. YOLO (You Only Look Once) models, specifically YOLOv8 and the recently introduced YOLOv11, offer efficient and accurate classification capabilities, making them promising candidates for real-time veterinary radiograph automation. However, their application in veterinary imaging remains largely unexplored, motivating a systematic evaluation.

1.3 Problem Statement

Despite advancements in deep learning for human medical imaging, limited research exists on automated Classification of veterinary radiographs, particularly anatomical classification of canine X-rays. As far as our knowledge is concerned, there is currently no standardized comparative evaluation of YOLO-based, CNN-based, and Transformer-based architectures for veterinary anatomical classification tasks. This paper aims to developing models to automatically classify canine X-ray body parts using YOLOv8, YOLOv11, Vision Transformer, and ResNet50 architectures.

1.4 Methodology

The methodology of this study involves five key stages. First, a curated dataset of canine radiographs is developed, comprising 453 X-ray images across five anatomical regions—skull, thorax, abdomen, forelimb, and hind limb—with each image manually annotated to ensure accurate labeling. Next, state-of-the-art deep learning models, namely, YOLOv8, YOLOv11, Vision Transformer (ViT), and ResNet50, are implemented and trained on this dataset using standardized preprocessing and augmentation techniques. The performances of these models are then evaluated using established metrics viz Accuracy, Precision, Recall, F1-Score, and Inference Time, providing a balanced assessment of predictive quality and computational efficiency. A comparative analysis is conducted to highlight the strengths and weaknesses of YOLO-based, Transformer-based, and CNN-based architectures in veterinary imaging. Finally, conclusions are drawn from the findings to guide the deployment of AI-assisted diagnostic systems in veterinary practice, with emphasis on accuracy, efficiency, and clinical applicability.

1.4.1 Dataset Description

The dataset contains X-ray images collected from several well-known veterinary centers, including Karnataka Veterinary Animal and Fisheries Sciences University (Bidar), Veterinary Government Hospital (Mahagaon, Dist. Kalaburagi), Veterinary Government Hospital (Kalaburagi), Veterinary College and Hospital (Shivamogga), and Curative Pet Hospital (Bangalore). Detailed descriptions were also provided by veterinarians from these hospitals.

The original dataset comprises 10,496 X-ray images from seven different animal species, the majority of which are dogs and cats, accounting for 10,440 images.

These images are presented in grayscale. Blurring is often applied to grayscale images before performing edge detection or thresholding, as it helps reduce noise that might otherwise create false edges. The details of the images collected is shown in the Table 1 and Figure 1 shows sample images from the collection, frequency being presented in Figure 2.

Animal	No. of X-Ray images
--------	---------------------

Dog	9078
Cat	1362
Bird	27
Mouse	16
Turtle	4
Abdomen	3
Nan	2
Metatarsus	2
animal	1
pedal	1

Table 1: Animal X-Ray images count

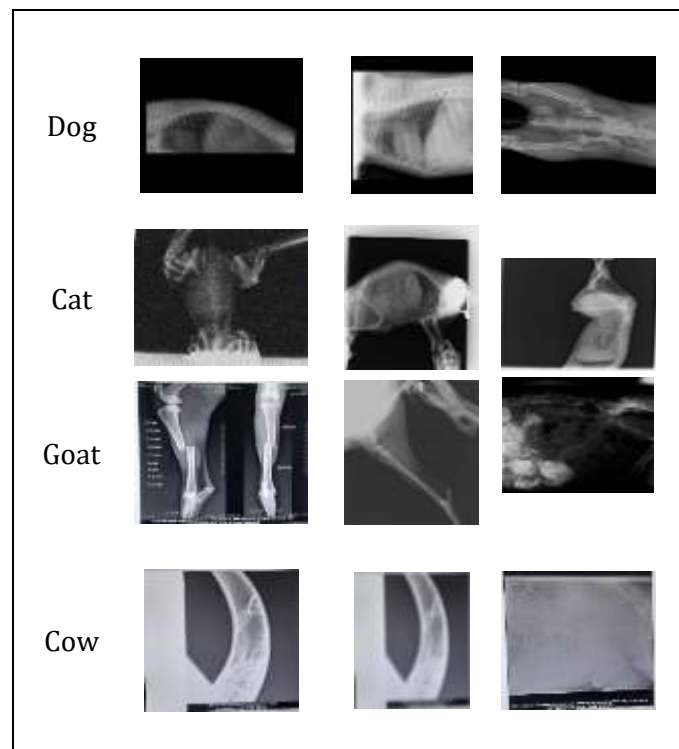


Figure 1: Sample X-ray images

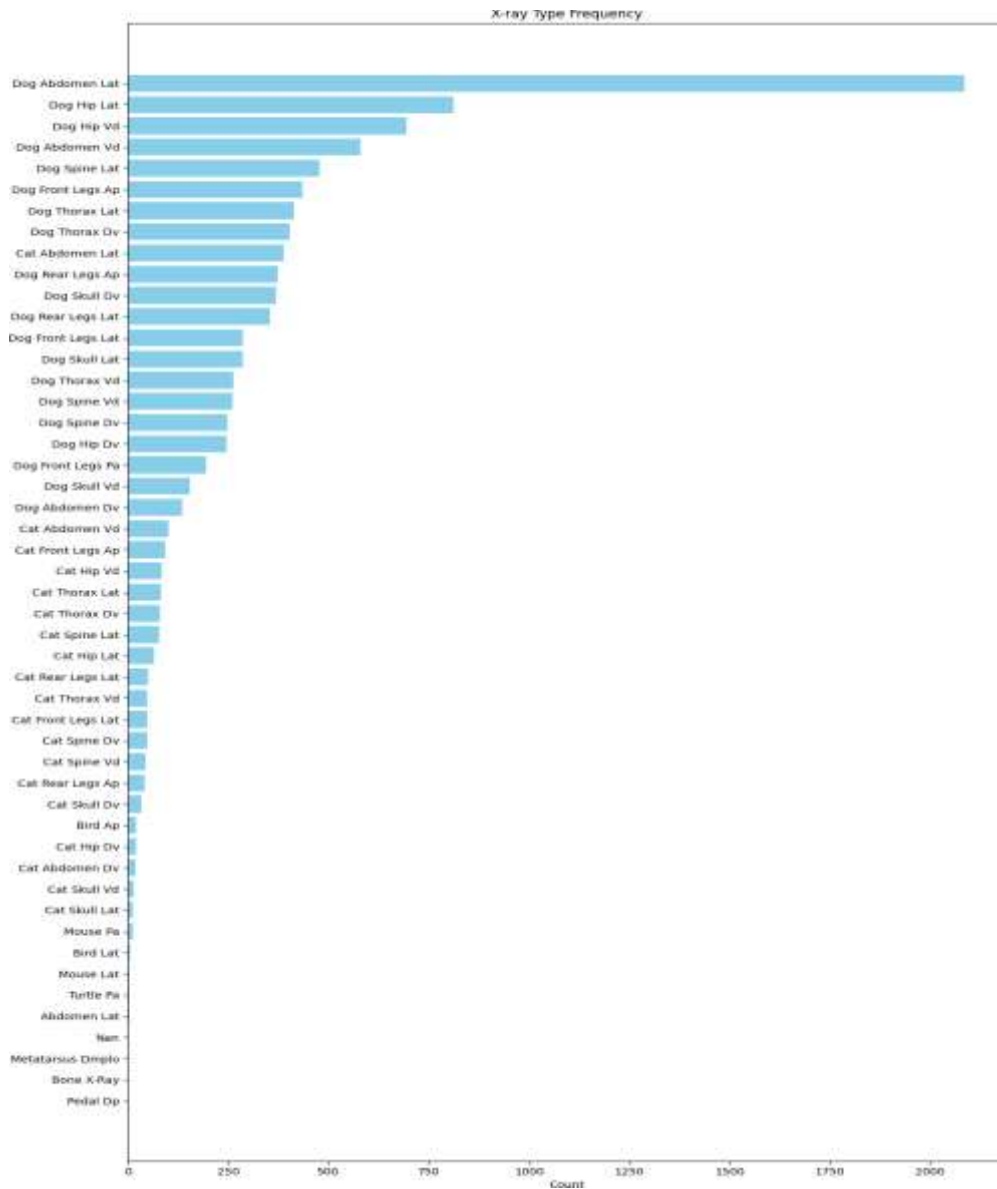


Figure 2: X-ray Frequency

The dataset was analyzed and organized into folders containing multiple Excel files. Each file includes the count of X-ray images for different animal body parts, the path to the respective images, and the analyzed image sizes to understand the range of dimensions present in the dataset. For the classification task using various deep learning architectures, the current dataset was restricted to dog body part classification, consisting of 453 images. All images were manually labeled and grouped into five distinct dog body parts. Images were resized and normalized to meet the input specifications of each model.

2 Literature Review

Deep learning is revolutionizing medical image analysis and is increasingly applied in veterinary radiology for diagnosis, classification, and quality assurance. While human radiographs are well studied, research on animal X-rays—especially dog body-part

classification—remains limited. This paper reviews key work on veterinary imaging, body-part classification, YOLO models, and Transformer architectures, highlighting research gaps that drive the present study.

2.1 Deep Learning in Veterinary Radiology

Several studies demonstrate the potential of CNNs for diagnostic interpretation of veterinary radiographs. Banzato et al. [1] developed an automatic classifier for canine thoracic radiographs that achieved radiologist-level accuracy. Later, Banzato et al. [2] introduced an AI-based algorithm for automatic quality evaluation of canine thoracic images, addressing issues such as exposure and positioning. Dumortier et al. [3] extended this to feline thoracic radiographs, applying CNN architectures for pulmonary abnormality detection. Banzato et al. [4] explored inter-species transfer learning, demonstrating that models trained on mixed canine and feline datasets improved generalization. These works established CNNs as strong baselines in veterinary diagnostics. However, they focus mainly on disease or quality assessment rather than anatomical body-part analysis and classification.

2.2 Animal Body-Part Classification in X- Rays

Research on anatomical segmentation or classification of animal X-rays remains limited. Park et al. [5] proposed a CNN model for detecting narrowed intervertebral disc spaces in canine thoracic and lumbar X-rays, reporting near-expert agreement with veterinarians. In a subsequent study, Park et al. [6] applied an attention U-Net to detect spondylosis deformans in dog spine radiographs, achieving a kappa value of 0.839. Similarly, Duenas et al. [7] released a dataset for vertebral heart size determination in canine thoracic images, facilitating training on structural identification rather than pathology. Despite these advances, no prior work has performed multi-class body-part classification (skull, thorax, limbs, hips, abdomen) in animal X-rays using modern YOLO or Transformer models — the key gap this research addresses.

2.3 YOLO-Based Approaches in Radiographic and Veterinary Imaging

The YOLO (You Only Look Once) architecture, introduced by Redmon et al. [8], revolutionized object detection by integrating localization and classification into a single forward pass. Recent extensions such as YOLOv8 and YOLOv11 (Ultralytics) support not only detection but also image classification. Harun and Fared [9] used YOLOv8 for lung disease classification from chest X-rays, demonstrating its suitability for medical imaging. Sharma et al. [10] compared YOLOv8 with DenseNet, ResNet, and VGG for wildlife species classification, reporting competitive accuracy and higher inference efficiency. Li et al. [11] explored YOLOv8 for veterinary disease detection in farm animals, illustrating the potential of YOLO for non-human biological imaging tasks. However, YOLO's classification mode remains underexplored in the veterinary domain, presenting an opportunity for novel experimentation in animal X-ray image classification.

2.4 Transformer-Based Architectures in Radiography

Transformers have redefined computer vision since the introduction of the Vision Transformer (ViT) by Dosovitskiy et al. [12], which replaces convolutional operations with self-attention. In radiography, Zimbalist et al. [13] employed transformer architectures for bone lesion detection under diverse acquisition conditions, demonstrating robustness to image variation. Touvron et al. [14] improved ViT with Data-efficient Image Transformers (DeiT), enabling strong performance with less data, a crucial advantage for veterinary X-rays with limited

samples. These studies highlight ViT's capability to model long-range dependencies and generalize across diverse imaging conditions, making it promising for animal X-ray classification.

2.5 Comparative CNN Approaches

Residual networks introduced by He et al. [15] have long served as a reliable benchmark in medical image classification due to their skip connections and efficient gradient propagation. ResNet50 remains a standard reference architecture in veterinary imaging, used by N. Deekonda et al. [4] and others to achieve strong accuracy even with small datasets. However, modern models such as YOLOv8 and ViT outperform conventional CNNs in adaptability and inference speed, warranting comparative evaluation across architectures — the goal of the present study.

2.6 Dataset Resources for Veterinary X-ray Studies

Several open datasets have been used in veterinary imaging research. Duenas et al. [7] introduced a small-scale dataset of 156 canine thoracic X-rays for vertebral heart size measurement. Deekonda [16] proposed a dataset of 2,000 canine radiographs for cardiomegaly detection, while VET-DINO [17] recently released a massive multi-view veterinary radio-graph dataset (5 million images). These datasets primarily focus on thoracic or cardiac regions, with limited representation of other body parts, reinforcing the need for new curated datasets — such as the manually annotated 453-image dataset used in this thesis.

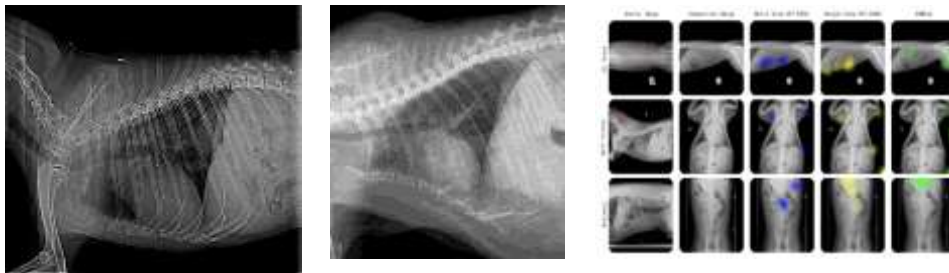


Figure 1(a) Sample images

From the reviewed literature, it is evident that Veterinary imaging research emphasizes disease detection and quality control, with limited focus on body-part identification; YOLO models have seen minimal exploration in veterinary radiograph classification, especially using Ultralytics YOLOv8/YOLOv11 frameworks; Transformer models (ViT, DeiT) have not yet been systematically explored to animal anatomical classification tasks. And that no existing studies compare the performance of YOLO, ViT, and ResNet models on a unified animal X-ray dataset. This paper presents a comparative evaluation of YOLOv8, YOLOv11, Vision Transformer, and ResNet50 on a dog body-part X-ray dataset, providing insights into model generalization and real-world applicability.

3 Experimental Setup and Results

Details of the experimental setup, including the models employed, are presented in the following sections.

3.1 YOLOv8 Model Variants

YOLOv8 (You Only Look Once, version 8) is an advanced object detection and image classification model developed by Ultralytics Y. Li et al. [11]. It introduces improvements in

model efficiency, speed, and generalization compared to previous YOLO versions. YOLOv8 supports various scales such as nano (n), small (s), medium (m), large (l), and extra-large (x) depending on the computational power and dataset size. The YOLOv8 architecture follows a one-stage detection paradigm with a CSPDarkNet backbone and C2f (Cross-Stage Partial) modules for efficient gradient propagation. It integrates PANet for feature fusion and a decoupled detection head for classification and localization.

YOLOv8 was trained on five dog body parts (skull, hips, front legs, thorax, and abdomen) using 453 images. The dataset was split into 361 training and 92 validation images. Each variant of YOLOv8 (n, s, m, l, x) was tested to observe the trade-off between accuracy and inference speed. The loss function was set to auto in the model configuration file, allowing YOLOv8 to automatically select the appropriate loss based on the task type. For the classification model, the framework internally employs the Cross-Entropy Loss, which measures the divergence between the predicted and target class probabilities. This adaptive approach simplifies configuration and ensures optimal loss selection across tasks. YOLOv8 features an efficient CSPDarkNet–PANet architecture, faster training with dynamic label assignment, and support for multi-scale inputs. Applications include Image classification, object detection, segmentation, and pose estimation.

3.1.1 Yolov8Nanomodel

The YOLOv8n (Nano) model is the lightest version, Y. Li et [11], designed for real-time applications on edge devices and systems with limited computational resources. It has the fewest parameters (1,444,693 parameters, 1,444,693 gradients) and the highest inference speed with 3.4 GFLOPs making it suitable for deployment on mobile or embedded hardware. The model was trained using input images resized to 224×224 with a batch size of 16 over 30 epochs. The network architecture consists of 56 layers, enabling effective hierarchical feature learning. Optimization was performed using the AdamW optimizer, while the loss function was automatically selected according to the task requirements. A learning rate of 0.01 was employed to ensure efficient convergence during training. Training notebooks were executed in Google Colab and linked through Drive. Figure 3 shows an example of the ground truth image, which represents the reference output used for evaluating the model's predictions. Figure 4 presents the validation prediction results, illustrating the model's ability to generalize and produce accurate outputs on unseen data.

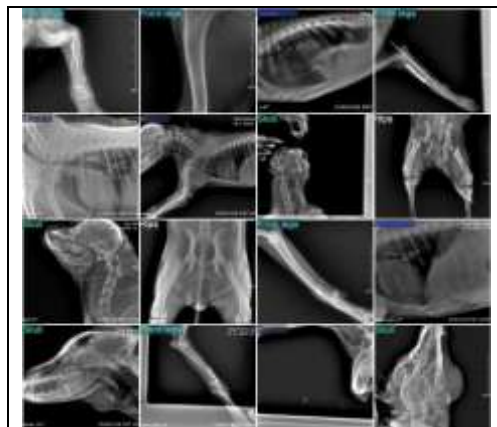


Figure 3: Example ground truth

Figure 5 illustrates the model's training and validation performance in terms of loss and accuracy across 30 epochs. The training loss decreased steadily from approximately 1.4

to below 0.2, while the validation loss followed a similar downward trend, indicating successful convergence of the model without significant overfitting.

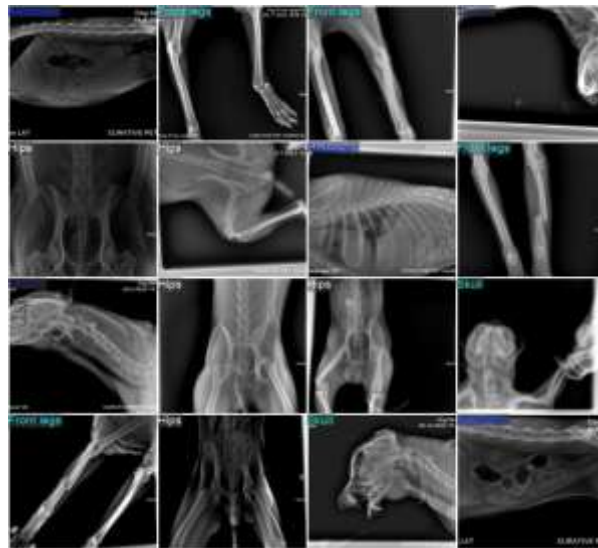


Figure 4. Validation Prediction for the model

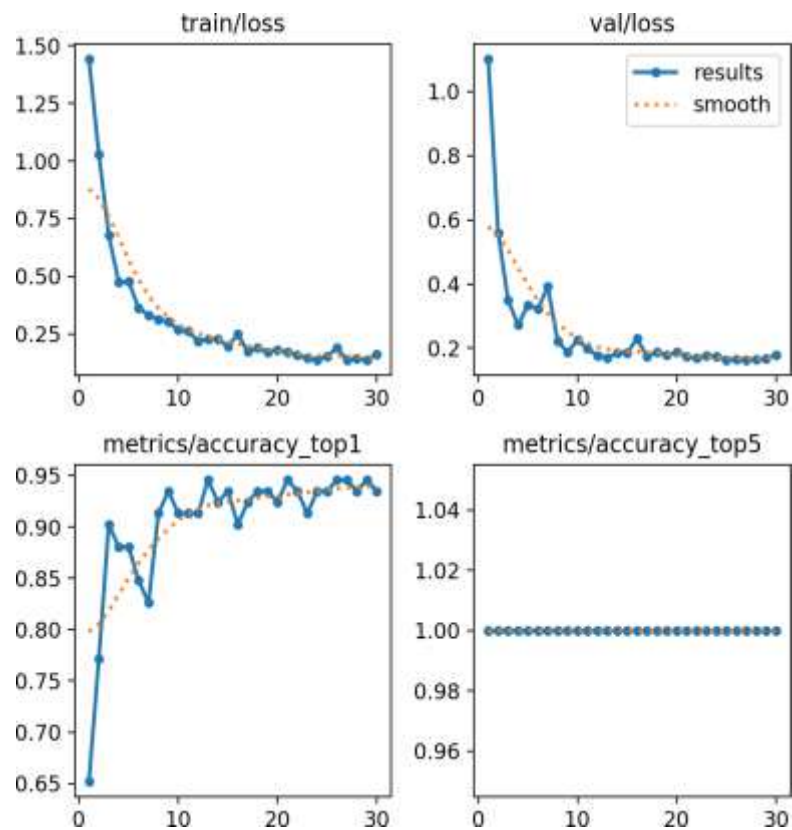


Figure performance of the model

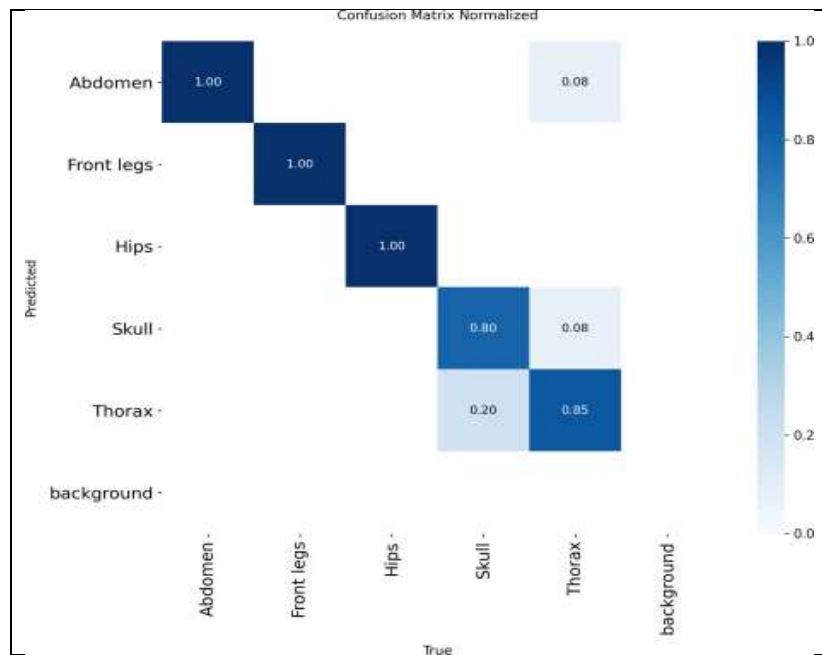


Figure 6 Confusion Matrix

The Top-1 accuracy increased rapidly during the initial epochs, stabilizing around 94–95. The close alignment between training and validation losses demonstrates good generalization capability, meaning the model performs consistently well on unseen data. The absence of large divergence between the curves further supports that the training was stable and effective.

Figure 6 shows the normalized confusion matrix of the YOLOv8 classification model evaluated on the validation dataset. Each cell represents the proportion of predictions made for a given true label. The diagonal elements correspond to correctly classified samples, while the off-diagonal elements indicate misclassifications. From the matrix, it can be observed that the model achieved perfect classification (1.00) accuracy for most classes such as Abdomen, Front legs, and Hips. Minor confusion occurred between the Skull and Thorax classes, where a few Skull images were misclassified as Thorax (0.27). This overlap could be due to visual similarities in X-ray images or limited sample variation between these categories. Overall, the confusion matrix indicates strong per class performance with minimal cross class confusion, suggesting that the model learned discriminative features effectively across all anatomical region.

3.1.2 Yolov8-Small model

The YOLOv8s (Small) model offers a balance between speed and accuracy. It contains more layers and parameters (5,087,141 parameters, 5,087,141 gradients) with 12.6 GFLOPs than the Nano model, allowing it to capture more complex patterns while remaining efficient for mid-range GPUs and real-time processing.

The model was trained using input images resized to 224×224 with a batch size of 16 over 20 epochs. The network architecture consists of 56 layers, enabling effective hierarchical feature learning. Optimization was performed using the AdamW optimizer, while the loss function was automatically

according to the task requirements. A learning rate of 0.001 was employed to ensure efficient convergence during training. Training notebooks were executed in Google Colab and linked through Drive. An example of the ground truth image is provided in Figure 7, serving as the standard for evaluating prediction accuracy. Figure 8 illustrates

the validation prediction results, highlighting the model's ability to accurately identify and classify patterns in unseen data. Figure 9 depicts the model's training and validation performance, showing loss and accuracy trends across 20 epochs. The training loss decreased steadily from approximately 1.5 to below 0.25, while the validation loss followed a similar downward trend, indicating successful convergence of the model without significant overfitting.

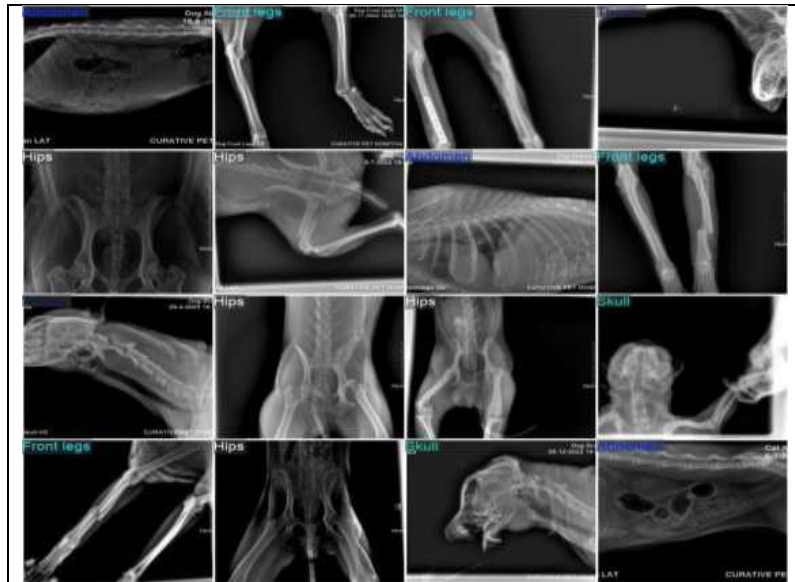


Figure 7: Sample ground truth images



Figure 8 Validation Prediction for Yolo8- Smallmodel

The close alignment between training and validation losses demonstrates good generalization capability, meaning the model performs consistently well on unseen data. The absence of large divergence between the curves further supports that the

training was stable and effective.

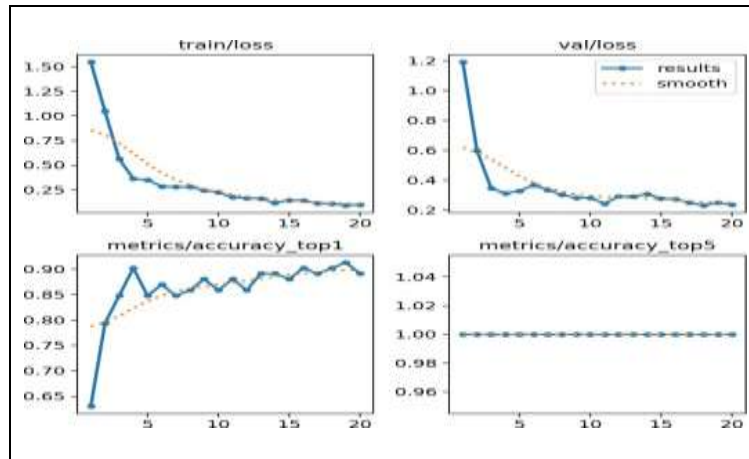


Figure 9 Training and validation performance of YOLOv8-Small model

Figure 10 shows the normalized confusion matrix of the YOLOv8 classification model evaluated on the validation dataset. Each cell represents the proportion of predictions made for a given true label. The diagonal elements correspond to correctly classified samples, while the off-diagonal elements indicate misclassifications.

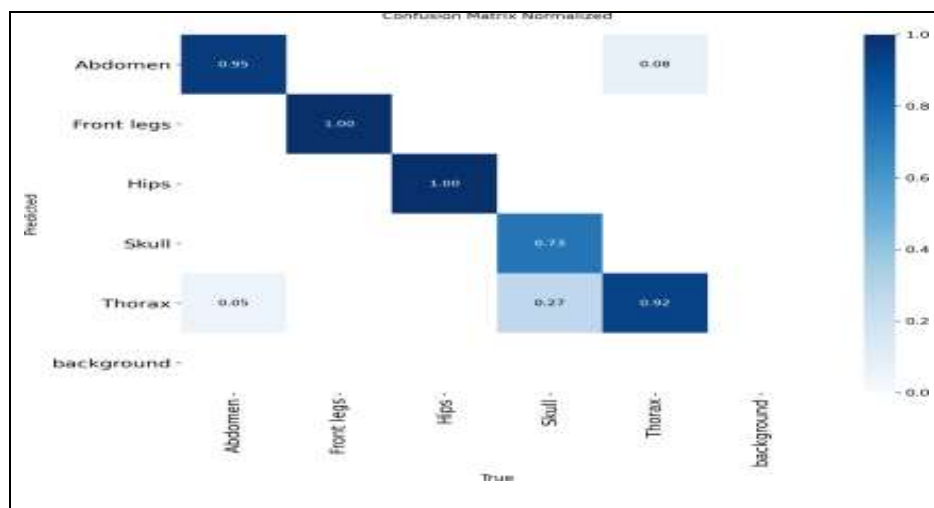


Figure10 Confusion Matrix

From the matrix, it can be observed that the model achieved perfect classification (1.00) accuracy for most classes such as Abdomen, Front legs, and Hips. Minor confusion occurred between the Skull and Thorax classes, where a few Skull images were misclassified as Thorax (0.2). This overlap could be due to visual similarities in X-ray images or limited sample variation between these categories.

Overall, the confusion matrix indicates strong per-class performance with minimal cross-class confusion, suggesting that the model learned discriminative features effectively across all anatomical regions.

3.1.3 YOLOv8-Medium model

The YOLOv8m (Medium) variant serves as a balanced model that provides a good trade-

off between accuracy and inference time with 15,778,741 parameters, 15,778,741 gradients and 41.9 GFLOPs. It is often used as the default choice for general-purpose object detection or classification tasks due to its stable performance and moderate resource requirements.

The model was trained using input images resized to 224×224 with a batch size of 16 over 20 epochs. The network architecture consists of 80 layers, enabling effective hierarchical feature learning. Optimization was performed using the AdamW optimizer, while the loss function was automatically selected according to the task requirements. A learning rate of 0.001 was employed to ensure efficient convergence during training.

Training notebooks were executed in Google Colab and linked through Drive. Ground truth images (Figure 11) serve as the reference output for evaluating the model's predictions. Figure 12 presents the validation prediction results, demonstrating the model's capacity to generalize and generate accurate outputs on unseen data.

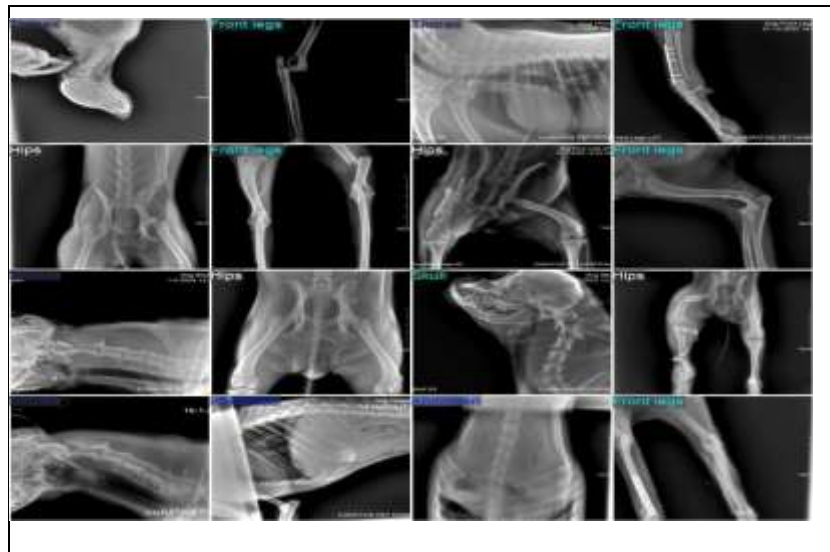


Figure 11: Sample images representing ground truth

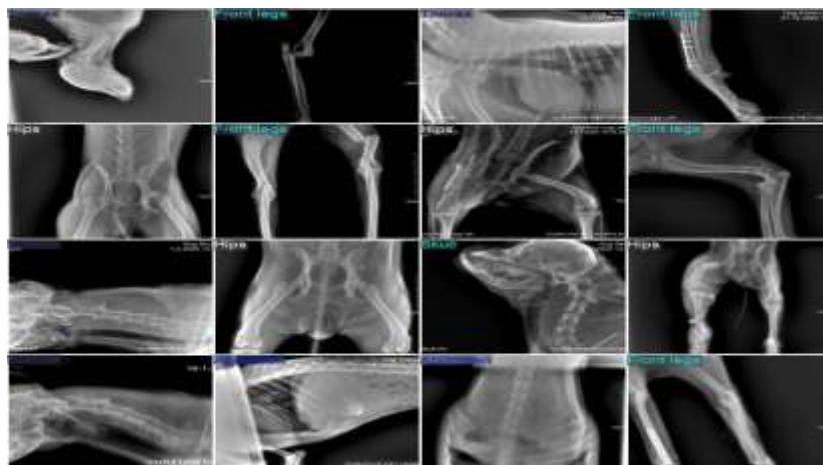


Figure 12 Validation Prediction of the model

Figure 13 presents the model's training and validation performance in terms of loss and accuracy across 20 epochs. The training loss decreased steadily from approximately 1.5 to below 0.25, while the validation loss followed a similar downward trend, indicating

successful convergence of the model without significant overfitting. The Top-1 accuracy increased rapidly during the initial epochs, stabilizing around 88–92%. The close alignment between training and validation losses demonstrates good generalization capability, meaning the model performs consistently well on unseen data. The absence of large divergence between the curves further supports that the training was stable and effective.

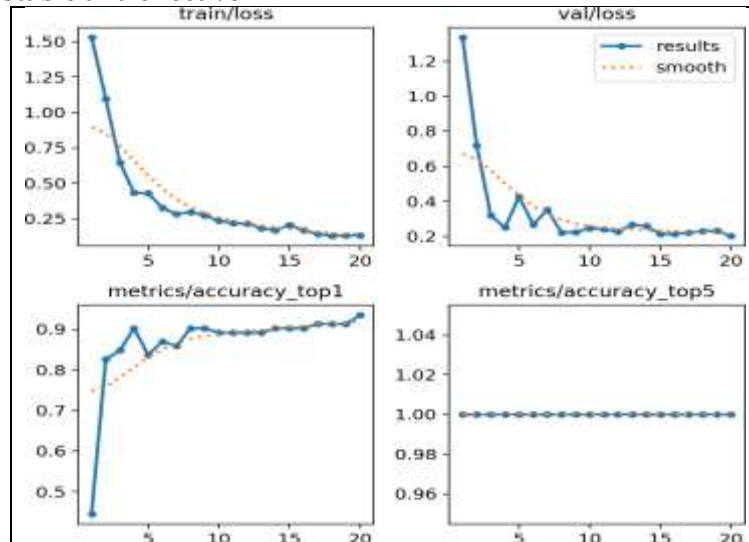


Figure 13: Training and validation performance of Yolov8-Medium model

Figure 14 shows the normalized confusion matrix of the YOLOv8 classification model evaluated on the validation dataset. Each cell represents the proportion of predictions made for a given true label. The diagonal elements correspond to correctly classified samples, while the off-diagonal elements indicate misclassifications.

From the matrix, it can be observed that the model achieved almost perfect classification (1.00) accuracy for

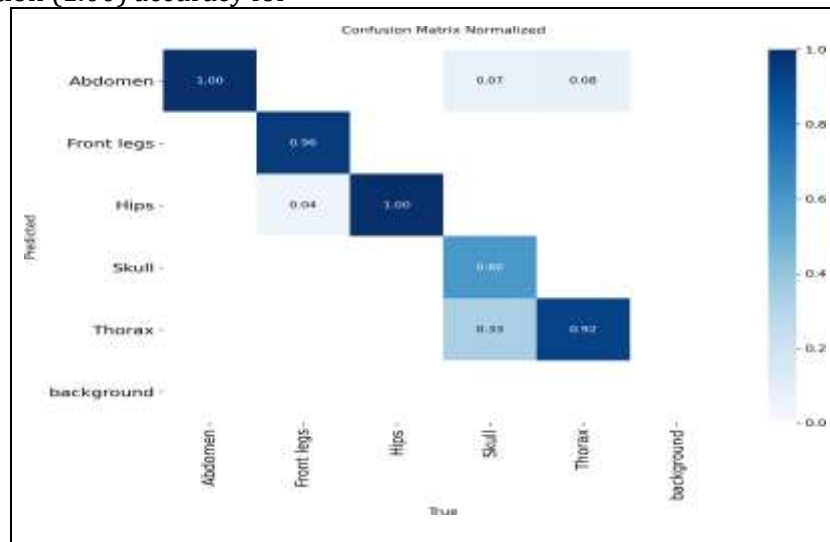


Figure 14 Confusion Matrix

most classes except Skull class. Major confusion occurred between the Skull and Thorax classes, where a few Skull images were misclassified as Thorax (0.33). This overlap could be due to visual similarities in X-ray images.

or limited sample variation between these categories.

Overall, the confusion matrix indicates strong per- class performance with minimal cross-class confusion, suggesting that the model learned discriminative features effectively across all anatomical regions except for Skull class not as good when compared to Nano and Small version YoloV8 models.

3.1.4 YOLOv8-Large model

The YOLOv8l (Large) model includes a greater number of convolutional layers and feature channels with 36,206,149 parameters, 36,206,149 gradients and 99.1 GFLOPs, enabling it to achieve higher accuracy on complex datasets. However, it requires significantly more computational power and memory, making it more suitable for high-performance GPUs.

The model was trained using input images resized to 224×224 with a batch size of 16 over 50 epochs. The network architecture consists of 104 layers, enabling effective hierarchical feature learning. Optimization was performed using the AdamW optimizer, while the loss function was automatically selected according to the task requirements. A learning rate of 0.001 was employed to ensure efficient convergence during training.

Results Training notebooks were executed in Google Colab and linked through Drive. Figure 15 shows an example of the ground truth image, which represents the reference output used for evaluating the model's predictions. Figure 16 presents the validation prediction results, illustrating the model's ability to generalize and produce accurate outputs on unseen data.

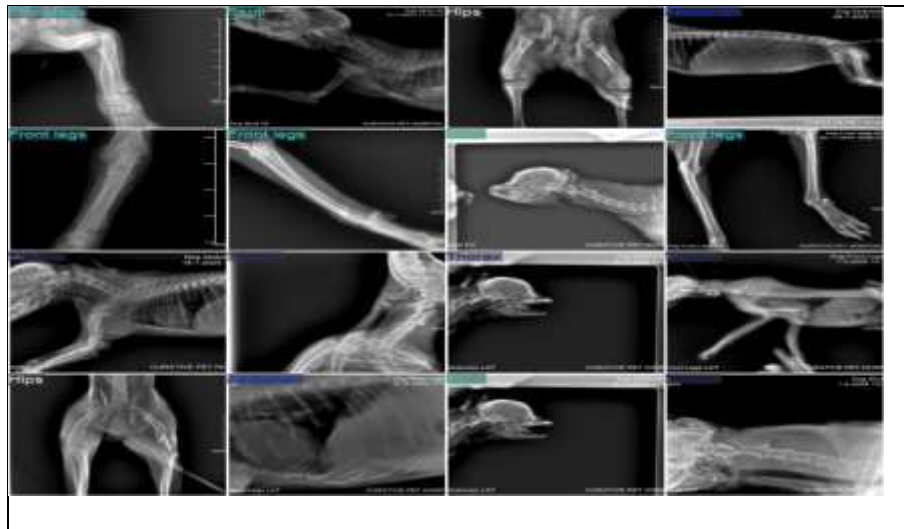


Figure 15: Sample ground truth images

Figure 17 illustrates the model's training and validation performance in terms of loss and accuracy across 10 epochs. The training loss decreased steadily from approximately 1.4 to below 0.3, while the validation loss followed a similar downward trend, but the loss going between 0.3 and 0.5.

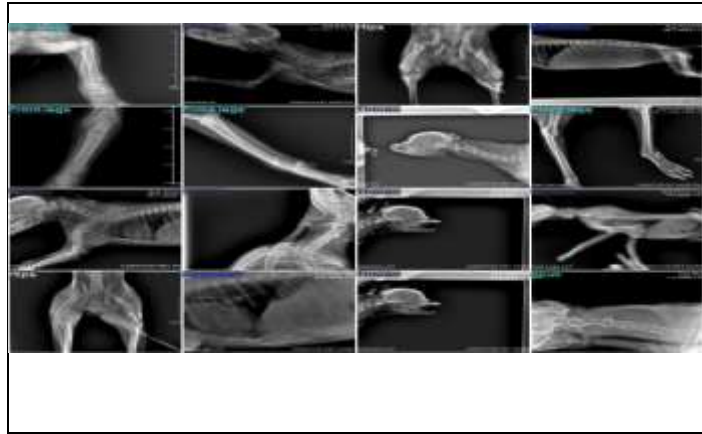


Figure 16 Validation Prediction of the model

The Top-1 accuracy increased rapidly during the initial epochs, stabilizing around 83–90. Figure 18 shows the normalized confusion matrix of the YOLOv8 classification model evaluated on the validation dataset. Each cell

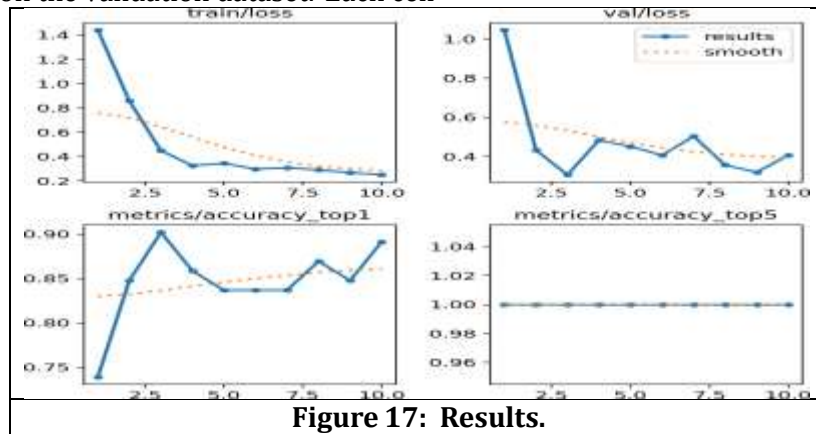


Figure 17 Training and validation performance of Yolov8-Large model

represents the proportion of predictions made for a given true label. The diagonal elements correspond to correctly classified samples, while the off-diagonal elements indicate misclassifications.

From the matrix, it can be observed that the model achieved almost perfect classification (1.00) accuracy for most classes such as Abdomen, Front legs, and Hips. Minor confusion occurred between the Skull and Thorax classes, where a few Skull images were misclassified as Thorax (0.13) and vice versa (0.23). This overlap could be due to visual similarities in X-ray images or limited sample variation between these categories.

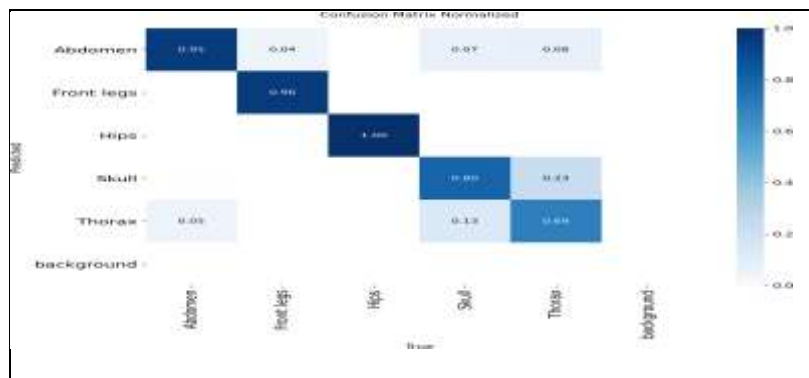


Figure 18 Confusion Matrix

Overall, the confusion matrix indicates strong per- class performance with minimal cross-class confusion, suggesting that the model learned discriminative features effectively across all anatomical regions except for Thorax class.

3.1.5 YOLOv8-ExtraLarge model

Finally, the YOLOv8x (Extra-Large) variant is the most powerful in the family, with the highest number of parameters and layers, which are 56,148,245 parameters, 56,148,245 gradients, and 154.3 GFLOPs. It achieves the best accuracy among all variants but is computationally expensive. This version is typically used for research purposes or in high-end systems where accuracy is prioritized over inference speed.

The model was trained using input images resized to 224×224 with a batch size of 16 over 100 epochs. The network architecture consists of 104 layers, enabling effective hierarchical feature learning. Optimization was performed using the AdamW optimizer, while the loss function was automatically selected according to the task requirements. A learning rate of 0.001 was employed to ensure efficient convergence during training.

Training notebooks were executed in Google Colab and linked through Drive. Figure 19 presents sample ground truth images, serving as the reference output for evaluating the model's predictions. Figure 20 illustrates the validation prediction results, demonstrating the model's capacity to generalize and generate accurate outputs on unseen data.



Figure 19: Sample ground truth images

Figure 21 shows the model's training and validation performance over 30 epochs. The training loss dropped steadily from about 1.5 to below 0.25. The validation loss also decreased, ranging between 0.5 and 0.2, which indicates that the model converged well without major overfitting.

The Top-1 accuracy increased rapidly during the initial epochs, stabilizing around 83–92.

The close alignment between training and validation losses demonstrates good generalization capability, meaning

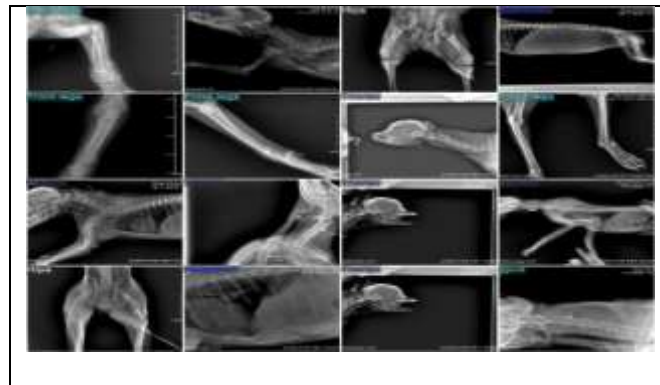


Figure 20 Validation Prediction of the model

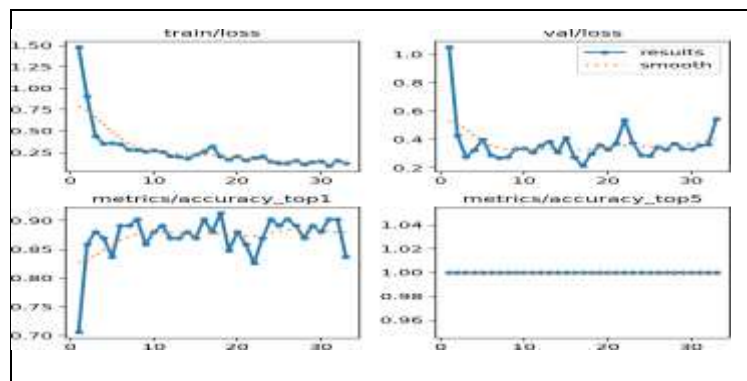


Figure 17 Training and validation performance of Yolov8-Extra Large model

the model performs consistently well on un- seen data. The absence of large divergence between the curves further supports that the training was stable and effective.

Figure 22 shows the normalized confusion matrix of the YOLOv8 classification model evaluated on the validation dataset. Each cell represents the proportion of predictions made for a given true label. The diagonal elements correspond to correctly classified samples, while the off-diagonal elements indicate misclassifications. From the matrix, it can be observed that the model achieved perfect classification (1.00) accuracy for most classes such as Abdomen, Front legs, and Hips.

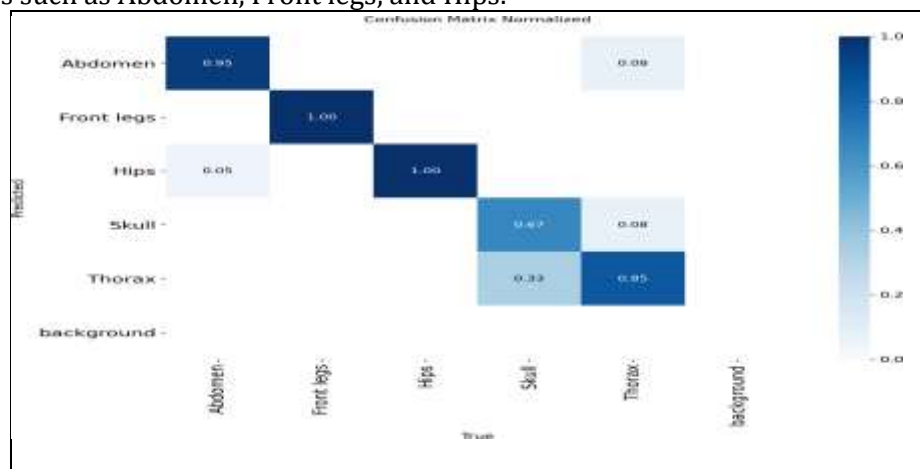


Figure 22 Confusion Matrix

Major confusion occurred between the Skull and Thorax classes, where a few Skull images were misclassified as Thorax (0.33) and vice versa (0.08). This overlap could be due to visual similarities in X-ray images or limited sample variation between these categories. Overall, the confusion matrix indicates per-class performance with little cross-class confusion, suggesting that the model learned discriminative features effectively for 3 classes only.

3.2 YOLOv11

YOLOv11 is an improved and more efficient extension of the YOLO family, designed for greater accuracy and better representation of characteristics. Integrates modern attention mechanisms and improved convolutional blocks to enhance fine-grained feature extraction.

In this project, YOLOv11 was trained with both 512×512 and 1024×1024 image resolutions for four classes (skull, hips, thorax and front legs). Different model sizes (nano, small, medium, large, and extra-large) were experimented with to assess performance.

Key features include an enhanced backbone with improved gradient flow, lightweight yet accurate classification optimized using AdamW and advanced augmentations, with applications in real-time classification and object localization on complex data sets.

3.2.1 YOLOv11-Nano model

Nano (YOLOv11n) is the smallest variant, Nano, is ultralightweight with minimal depth and width in its architecture, intended for deployment in environments with tight computational constraints, such as edge devices, IoT systems, or mobile units with 1,536,228 parameters and 10.53 GFLOPs.

The model was trained using input images resized to 512×512 and 1024×1024 with a batch size of 16 over 100 epochs. The network architecture consists of 86 layers, enabling effective hierarchical feature learning. Optimization was performed using the AdamW optimizer, while the loss function was automatically selected according to the task requirements. A learning rate of 0.001 was employed to ensure efficient convergence during training.

Figure 23 illustrates the model's training and validation performance in terms of loss and accuracy across 15 epochs of Nano(512×512). The training loss decreased steadily from approximately 1.2 to almost 0.2, indicating successful convergence of the model while the validation loss curve indicates overfitting.

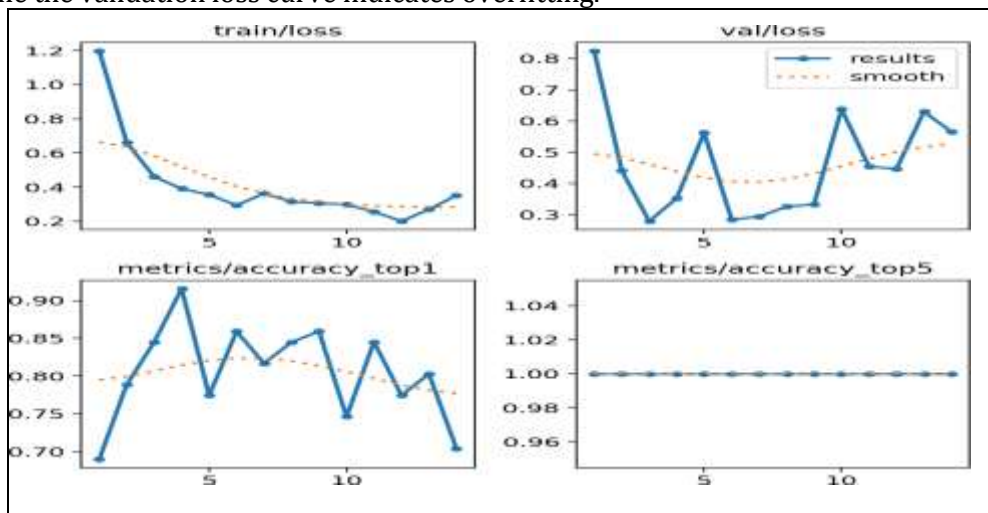


Figure 23 Training and validation performance of YOLOv11 nano (512×512)

The Top-1 accuracy fluctuates between 0.6 and 0.9, gradually stabilizing toward 0.85 in the final epochs. This means the model correctly identifies the main class about 85%. The gap between training and validation losses is small, implying that the model generalizes reasonably well to unseen data. While validation fluctuations suggest some sensitivity to data variation, the overall accuracy trends confirm effective learning and good convergence.

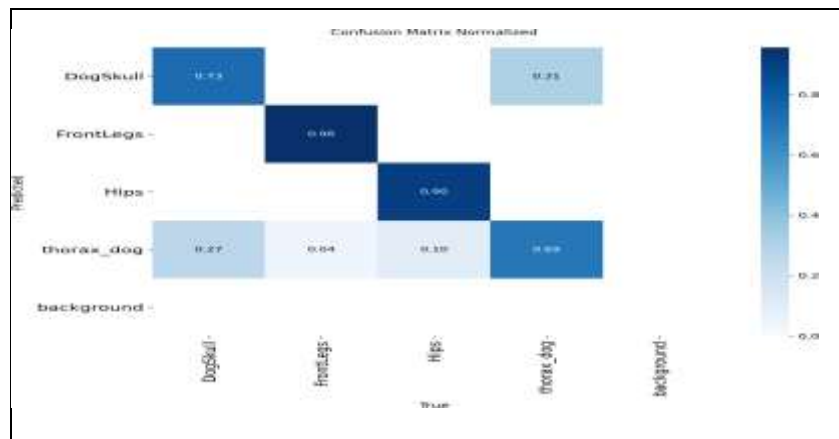


Figure 24: Confusion Matrix

Figure 24 shows the normalized confusion matrix of the YOLOv11 nano classification model evaluated on the validation dataset. Each cell represents the proportion of predictions made for a given true label. The diagonal elements correspond to correctly classified samples, while the off-diagonal elements indicate misclassifications.

From the matrix, it can be observed that the model achieved perfect classification (1.00) accuracy for Front

legs, Hips. Minor confusion occurred between the Skull and Thorax classes, where a few Skull images were misclassified as Thorax (0.2) and vice versa (0.08). This overlap could be due to visual similarities in X-ray images or limited sample variation between these categories.

Overall, the confusion matrix indicates good per- class performance with minimal cross- class confusion, suggesting that the model learned discriminative features effectively across all anatomical regions.

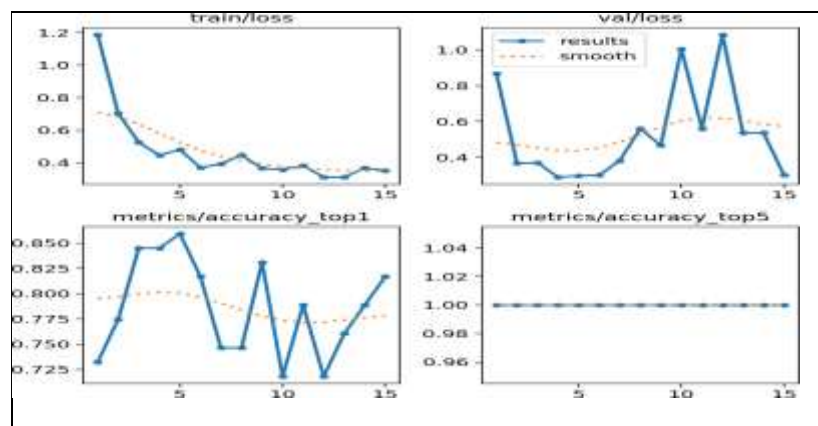


Figure 25: Model's training and validation performance of Nano(1024*1024)

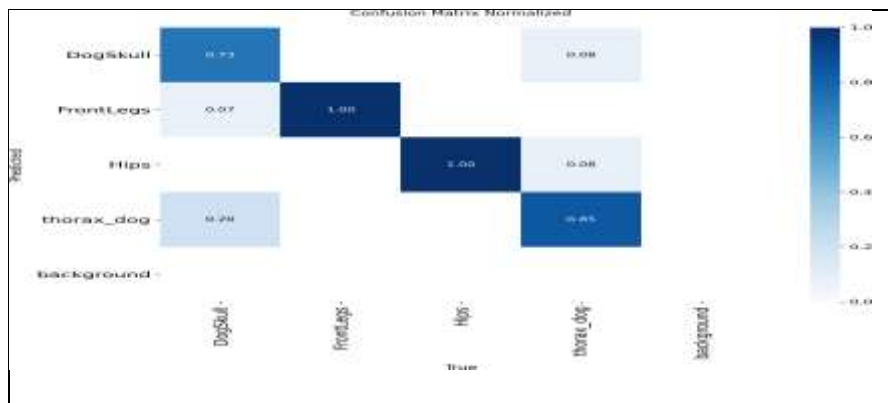


Figure 26: Confusion Matrix

Figure 25 illustrates the model's training and validation performance in terms of loss and accuracy across 15 epochs of Nano(1024*1024). The training loss decreased steadily from approximately 1.2 to 0.3, indicating successful convergence of the model without significant overfitting, while the validation loss followed a similar downward trend for initial epochs but indicates overfitting in further epochs forward.

The Top-1 accuracy increased rapidly during the initial epochs, but fluctuating between 0.85 and 0.725 showing that the model correctly predicted 80 Figure 26 shows the normalized confusion matrix of the YOLOv11 classification model evaluated on the validation dataset. Each cell represents the proportion of predictions made for a given true label. The diagonal elements correspond to correctly classified samples, while the off-diagonal elements indicate misclassifications.

From the matrix, it can be observed that the model achieved high classification accuracy for Front legs and Hips while it confused Skull with Thorax classes, where a few Skull images were misclassified as Thorax (0.27) and vice versa (0.31). This overlap could be due to visual similarities in X-ray images or limited sample variation between these categories. Overall, the confusion matrix indicates per-class performance with minimal cross-class confusion, suggesting that the model learned discriminative features effectively across all anatomical regions.

3.2.2 Yolov11-Small model

The small variant (YOLOv11s) retains fast inference but gains more capacity (layers, channels) to improve detection quality with 5,448,132 parameters, 5,448,132 gradients, 12.1 GFLOPs.

The model was trained using input images resized to 512×512 and 1024×1024 with a batch size of 16 over 100 epochs. The network architecture consists of 86 layers, enabling effective hierarchical feature learning. Optimization was performed using the AdamW optimizer, while the loss function was automatically selected according to the task requirements. A learning rate of 0.001 was employed to ensure efficient convergence during training.

Figure 27 illustrates the model's training and validation performance in terms of loss and accuracy across 15 epochs of Yolov (512 x 512) model. The training loss decreased steadily from approximately 1.2 to 0.3, indicating successful convergence of the model without significant overfitting, while the validation loss followed a similar downward trend for initial epochs but indicates overfitting in further epochs forward.

The Top-1 accuracy increased rapidly during the initial epochs, but fluctuating between 0.85 and 0.725 showing that the model correctly predicted 80

Figure 28 shows the normalized confusion matrix of the YOLOv11 classification model

evaluated on the validation dataset. Each cell represents the proportion of predictions made for a given true label. The diagonal elements correspond to correctly classified samples, while the off-diagonal elements indicate misclassifications.

From the matrix, it can be observed that the model achieved high classification accuracy for Front legs and Hips while it confused Skull with Thorax classes, where a few Skull images were misclassified as Thorax (0.27) and vice versa (0.31). This overlap could be due to visual similarities in X-ray images or limited sample variation between these categories. Overall, the confusion matrix indicates per-class performance with minimal cross-class confusion, suggesting that the model learned discriminative features effectively across all anatomical regions.

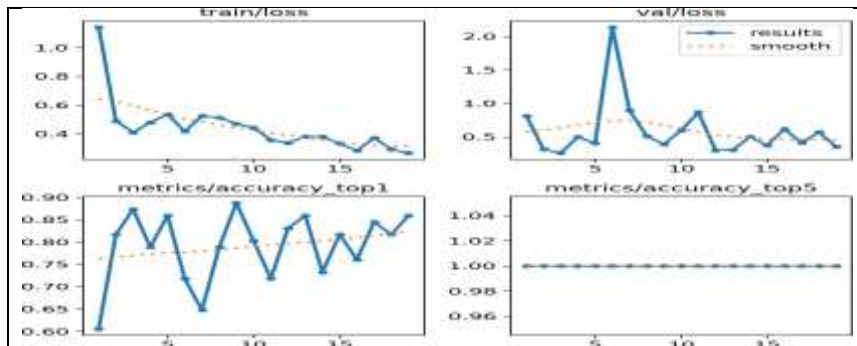


Figure 27: Model's training and validation performance of Yolov (512 x 512) model

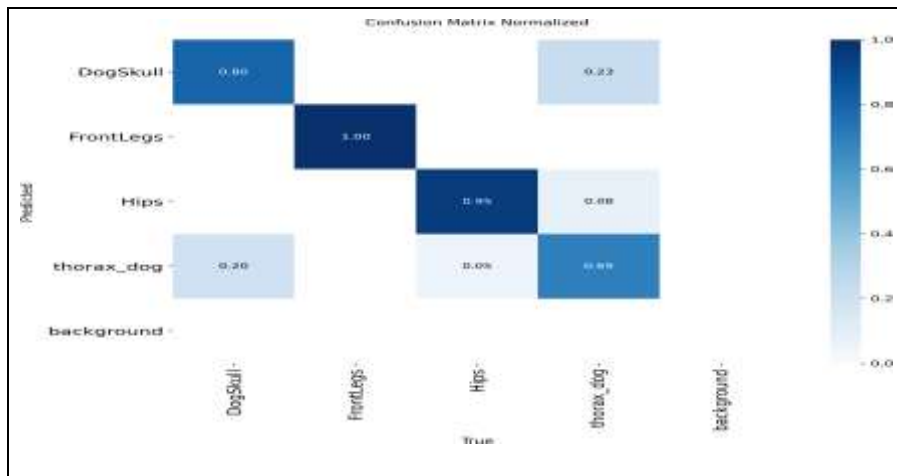


Figure 28 Confusion Matrix

Figure 29 illustrates the model's training and validation performance in terms of loss and accuracy across 25 epochs of Yolov (1024 x 1024) model. The training loss decreased steadily from approximately 1.1 to below 0.3, while the validation loss curve fluctuating and following almost similar trend to train loss curve, indicating successful convergence of the model without significant overfitting. The Top-1 accuracy increased from 0.5 to 0.9, showing that the model correctly predicted 80.

The close alignment between training and validation losses demonstrates good generalization capability, meaning the model performs consistently well on unseen data. The absence of large divergence between the curves further supports that the training was stable and effective.

Figure 30 shows the normalized confusion matrix of the YOLOv8 classification model

evaluated on the validation dataset. Each cell represents the proportion of predictions made for a given true label. The diagonal elements correspond to correctly classified samples, while the off-diagonal elements indicate misclassifications.

From the matrix, it can be observed that the model achieved perfect classification (1.00) accuracy for Front legs, 0.9 for Hips, 0.93 for Dog Skull and 0.54 for Thorax class, while most of the thorax class was misclassified as Skull. This overlap could be due to visual similarities in X-ray images or limited sample variation between these categories. Overall, the confusion matrix indicates strong per- class performance with minimal cross-class confusion, suggesting that the model learned discriminative features effectively across all anatomical regions except for Thorax class.

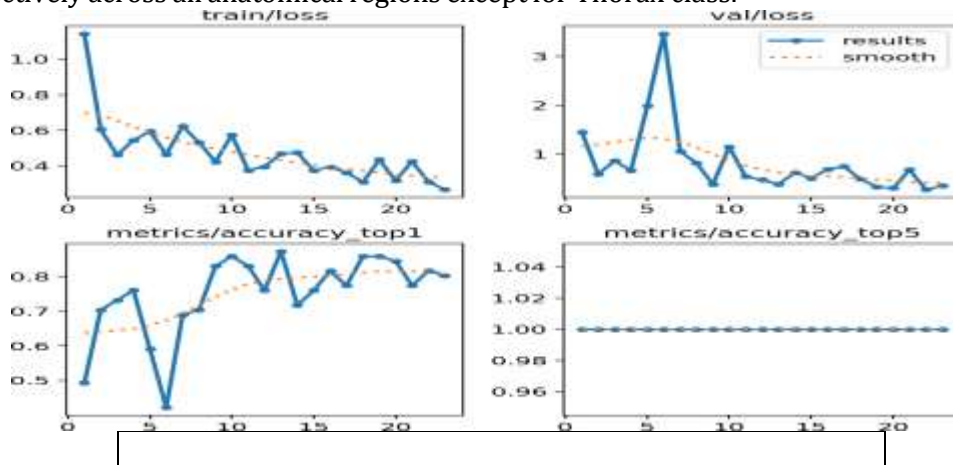


Figure 29: Model's training and validation performance of Yolov (1024 x 1024) model

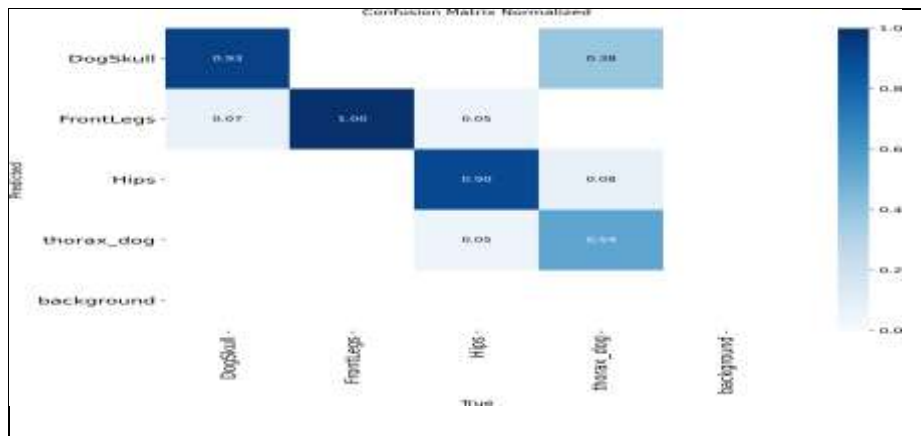


Figure 30 Confusion Matrix

3.2.3 Yolov11-Medium model

The (YOLOv11m) Medium version offers the most balanced performance: substantially better accuracy than the smaller ones, while still keeping inference times reasonable on modern GPUs with 10,358,340 parameters, 10,358,340 gradients, 39.6 GFLOPs.

The model was trained using input images resized to 512×512 and 1024×1024 with a batch size of 16 over 100 epochs. The network architecture consists of 106 layers, enabling effective hierarchical feature learning. Optimization was performed using the AdamW optimizer, while the loss function was automatically selected according to the task requirements. A learning rate of 0.001 was employed to ensure efficient convergence during training

Results- Medium (512*512 Figure 31 illustrates the model's training and validation performance in terms of loss and accuracy across 20 epochs. The training loss decreased steadily from approximately 1.2 to below 0.2, indicating successful convergence of the model while the validation loss curve indicates overfitting. The Top-1 accuracy fluctuates between 0.78 and 0.9, gradually stabilizing toward 0.85 in the final epochs. This means the model correctly identifies the main class about 85 The gap between training and validation losses is small, implying that the model generalizes reasonably well to unseen data. While validation fluctuations suggest some sensitivity to data variation, the overall accuracy trends confirm effective learning and good convergence.

Figure 32 shows the normalized confusion matrix of the YOLOv11 classification model evaluated on the validation dataset. Each cell represents the proportion of predictions made for a given true label. The diagonal elements correspond to correctly classified samples, while the off-diagonal elements indicate misclassifications.

From the matrix, it can be observed that the model achieved classification accuracy 0.96 for Front legs,

0.95 for Hips. Minor confusion occurred between the Skull and Thorax classes, where a few Skull images were misclassified as Thorax (0.13) and vice versa (0.38). This overlap could be due to visual similarities in X-ray images or limited sample variation between these categories.

Overall, the confusion matrix indicates good per- class performance with minimal cross- class confusion, suggesting that the model learned discriminative features effectively across all anatomical regions except for thorax class.

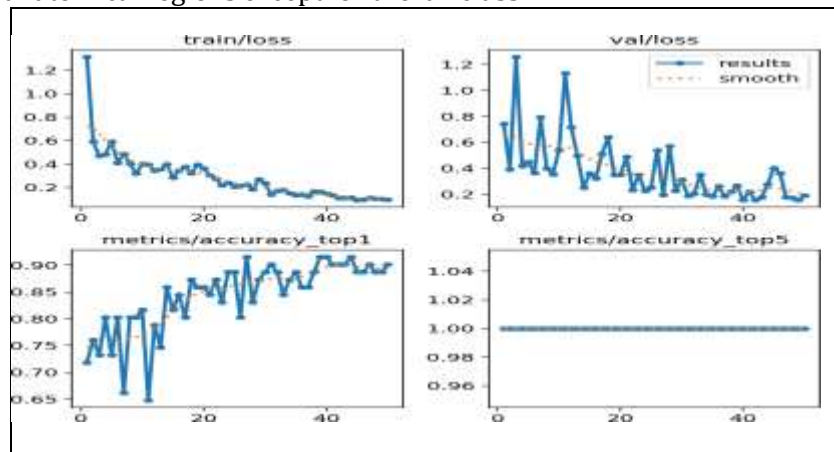


Figure 31 Model's training and validation performance of Yolov11 (512 x 512) model

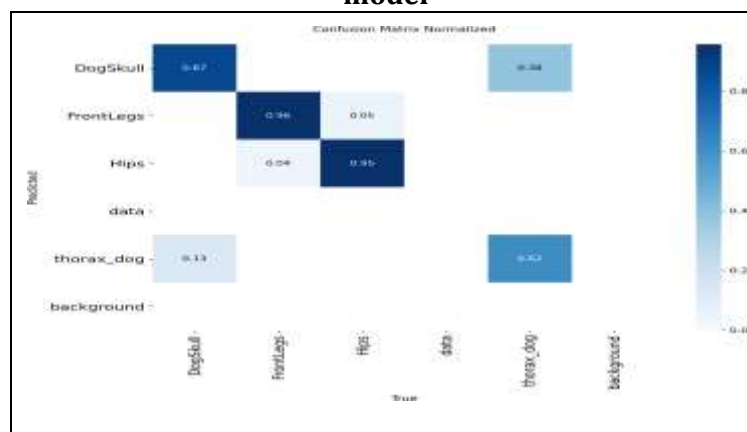


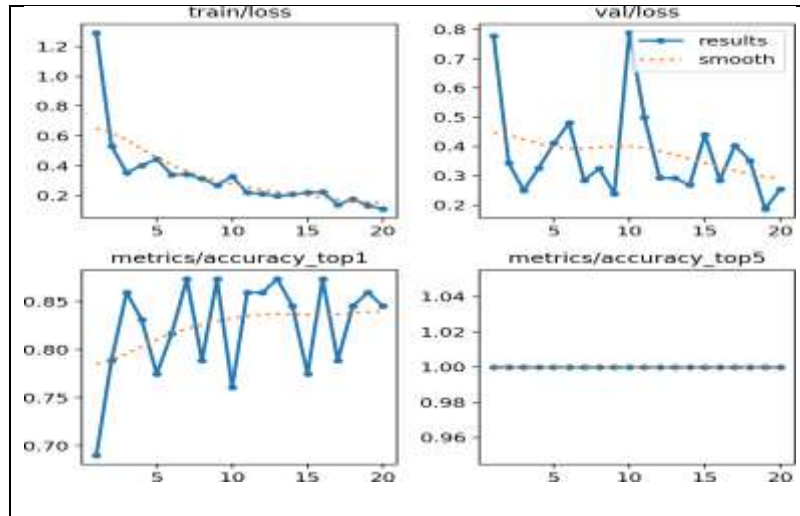
Figure 32: Confusion Matrix**Figure 33: 31 Model's training and validation performance of Yolov11 (1024 x 1024) model**

Figure 33 illustrates the model's training and validation performance in terms of loss and accuracy across 50 epochs of Yolov11 (1024 x 1024) model. The training loss decreased steadily from approximately 1.2 to below 0.2, indicating successful convergence of the model without significant overfitting, while the validation loss curve indicates overfitting. The Top-1 accuracy stabilized around 0.9 which might improve on further training. The Top-5 accuracy consistently remained at 100

Figure 34 shows the normalized confusion matrix of the YOLOv11 classification model evaluated on the validation dataset. Each cell represents the proportion of predictions made for a given true label. The diagonal elements correspond to correctly classified samples, while the off-diagonal elements indicate misclassifications.

From the matrix, it can be observed that the model achieved high classification accuracy for Front legs and Hips while it confused Skull with Thorax classes, where a few Skull images were misclassified as Thorax (0.07) and vice versa (0.23). This overlap could be due to visual similarities in X-ray images or limited sample variation between these categories.

Overall, the confusion matrix indicates per-class performance with minimal cross-class confusion, suggesting that the model learned discriminative features effectively across all anatomical regions.

3.2.4 Yolov11-Large model

Large (YOLOv11l) furthers this by increasing architectural depth and width, yielding higher precision at the cost of slower speed and greater resource demands with 12,839,748 parameters, 12,839,748 gradients, 49.8 GFLOPs.

The model was trained using input images resized to 512×512 with a batch size of 16 over 100 epochs. The network architecture consists of 176 layers, enabling effective hierarchical feature learning. Optimization was performed using the AdamW optimizer, while the loss function was automatically selected according to the task

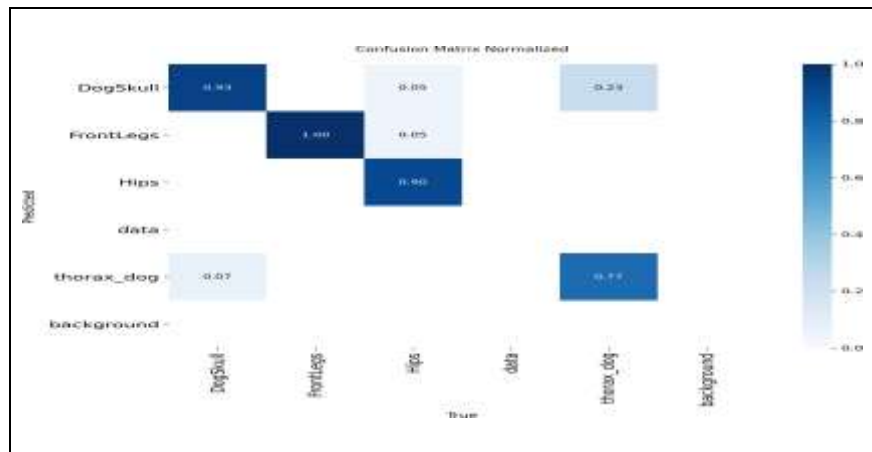


Figure 34: Confusion Matrix

requirements. A learning rate of 0.001 was employed to ensure efficient convergence during training.

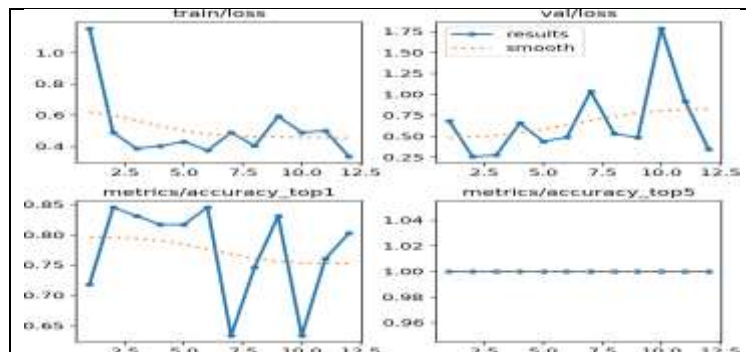
Results- Large(512*512 Figure 35 illustrates the model's training and validation performance in terms of loss and accuracy across 12.5 epochs. The training loss gradually decreasing from 1.1 to below 0.4 indicating successful convergence of the model without significant overfitting, while the validation loss curve fluctuating and indicating overfitting.

The Top-1 accuracy increased rapidly during the initial epochs, stabilizing around 65-85

Figure 36 shows the normalized confusion matrix of the YOLOv11-Large classification model evaluated on the validation dataset. Each cell represents the proportion of predictions made for a given true label. The diagonal elements correspond to correctly classified samples, while the off-diagonal elements indicate misclassifications.

From the matrix, it can be observed that the model achieved perfect classification (1.00) accuracy for Front legs and 0.93 for Skull. Major confusion occurred between the Skull and Thorax classes, where almost half of Thorax images were misclassified as Skull (0.54). This overlap could be due to visual similarities in X-ray images or limited sample variation between these categories.

Overall, the confusion matrix indicates per-class performance with minimal cross-class confusion other than thorax, suggesting that the model learned discriminative features effectively across all anatomical regions.



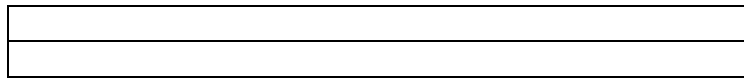


Figure 35: 31Model's training and validation performance of Yolov11 large (512 x 512) model



Figure 36: Confusion Matrix

3.2.5 Yolov11-ExtraLarge model

Extra-Large (YOLOv11x) delivers the highest performance (in terms of average precision, detection quality, ability to detect small or complex objects), but also the heaviest computational footprint—requiring substantial GPU memory, higher GFLOPs, and longer inference times with 28,361,188 parameters, 28,361,188 gradients, 111.0 GFLOPs.

Training Configuration

The model was trained using input images resized to 512×512 with a batch size of 16 over 100 epochs. The network architecture consists of 176 layers, enabling effective hierarchical feature learning. Optimization was performed using the AdamW optimizer, while the loss function was automatically selected according to the task requirements. A learning rate of 0.001 was employed to ensure efficient convergence during training.

Figure 37 illustrates the model's training and validation performance in terms of loss and accuracy across 15 epochs. The training fluctuating between 0.6 and 0.4 while the validation loss indicates overfitting.

The Top-1 accuracy fluctuating between 55 and 85 Figure 38 shows the normalized confusion matrix of the YOLOv8 classification model evaluated on the validation dataset. Each cell represents the proportion of predictions made for a given true label. The diagonal elements correspond to correctly classified samples, while the off-diagonal elements indicate misclassifications.

From the matrix, it can be observed that the model achieved perfect classification (1.00) accuracy for Hips,

0.96 for Front Legs. Minor confusion occurred between the Skull and Thorax classes, where a few Skull images were misclassified as Thorax (0.27) and vice versa (0.31). This overlap could be due to visual similarities in X-ray images or limited sample variation between these categories.

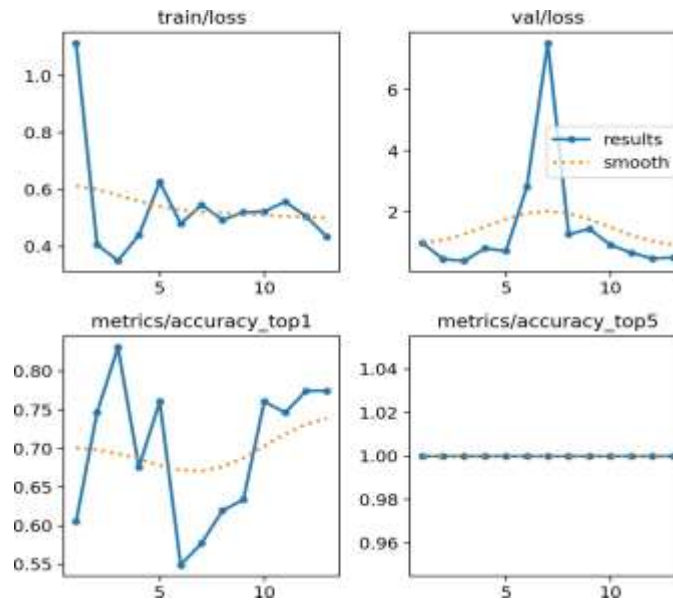


Figure 37: 31Model's training and validation performance of ExtraLarge(512*512)

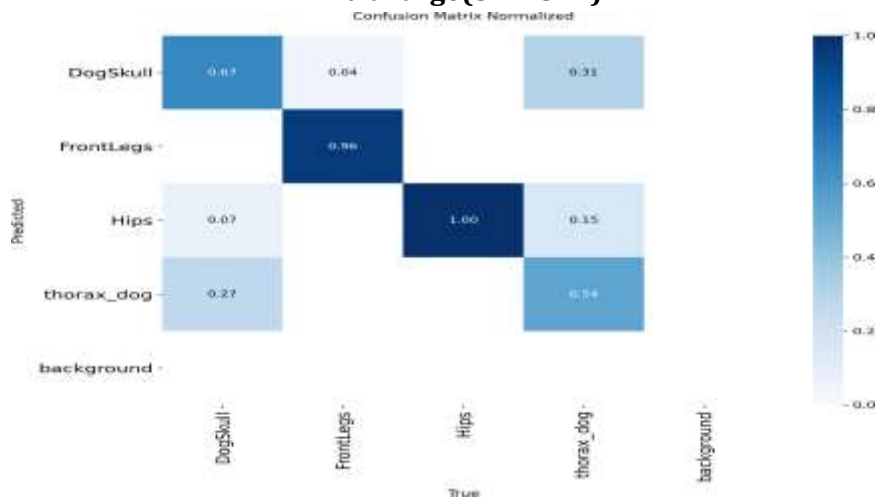


Figure 38: Confusion Matrix

Overall, the confusion matrix indicates strong per- class performance with minimal cross-class confusion, suggesting that the model learned discriminative features effectively across all anatomical regions.

3.3 Vision Transformer (ViT)

Vision Transformer (ViT) is a transformer-based model designed specifically for image classification tasks. Instead of convolutional operations, ViT divides an image into patches, flattens them, and processes them as a sequence—similar to words in a sentence. This enables the model to capture global dependencies more effectively. For this study, the ViT-Base model with patch size 16×16 and input image resolution 224×224 was used, A. Dosovitskiy et al [12].

The Vision Transformer uses self-attention to see the whole image at once, works best when trained on big datasets, handles changes in light and orientation well, and is used

in medical imaging, detailed classification, and finding anomalies

The model was trained using input images resized to 224×224 with a batch size of 32 over 100 epochs. The architecture consists of 12 transformer layers, enabling effective modeling of long-range dependencies. Training was optimized using the AdamW optimizer with the CrossEntropyLoss function, while a learning rate of 0.0001 was employed to ensure stable convergence and improved generalization.

Figure 39 illustrates the model's training and validation performance in terms of loss and accuracy across 18 epochs. The training loss decreased steadily from approximately 1.6 to below 0.2, while the validation loss followed a similar downward trend, indicating successful convergence of the model without significant overfitting.

The Top-1 accuracy increased rapidly during the initial epochs, stabilizing around 88–92

The close alignment between training and validation losses demonstrates good generalization capability, meaning the model performs consistently well on unseen data. The absence of large divergence between the curves further supports that the training was stable and effective.

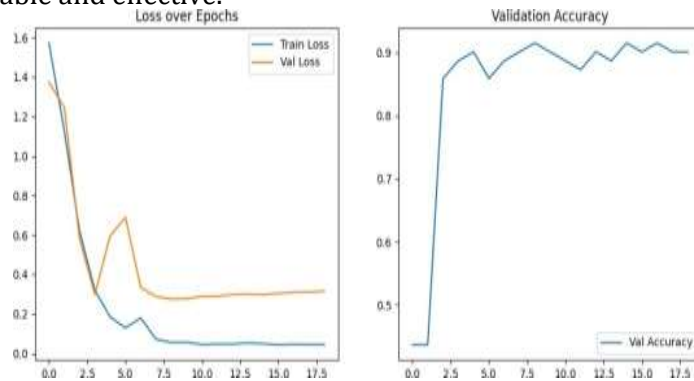


Figure 39 Model's training and validation performance ViT

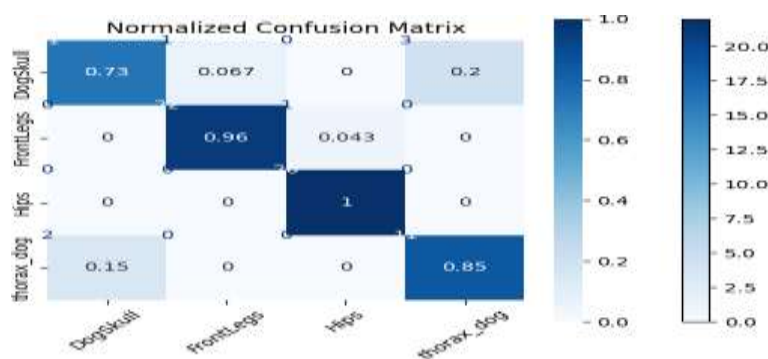


Figure 40 Confusion Matrix

Figure 40 shows the normalized confusion matrix of the ViT Transformer model evaluated on the validation dataset. Each cell represents the proportion of predictions made for a given true label. The diagonal elements correspond to correctly classified samples, while the off-diagonal elements indicate misclassifications. From the matrix, it can be observed that the model achieved perfect classification (1.00) accuracy for Hips,

0.96 for Frontlegs. Minor confusion occurred between the Skull and Thorax classes, where a few Skull images were misclassified as Thorax (0.15) and vice versa (0.20). This overlap could be due to visual similarities in X-ray images or limited sample variation between these categories. Overall, the confusion matrix indicates strong per class performance with minimal cross-class confusion, suggesting that the model learned discriminative features effectively across all anatomical regions.

3.4 ResNet50

ResNet50 (Residual Network with 50 layers) is a deep convolutional neural network that introduces residual learning to overcome the vanishing gradient problem. Skip connections allow gradients to flow efficiently through the network, making training of very deep architectures possible.

In this project, ResNet50 was used following the approach from the research paper “Bone Fracture Detection using Deep Learning,” with the Cross Entropy Loss and Adam optimizer.

ResNet50, with 50 residual layers, offers stable training, strong generalization, leverages ImageNet pre-training, and is widely used for classification, detection, segmentation, and feature extraction.

The model was trained using input images resized to 224×224 with a batch size of 32 over 100 epochs. The architecture consists of 12 transformer layers, enabling effective modeling of long-range dependencies. Training was optimized using the AdamW optimizer with the CrossEntropyLoss function, while a learning rate of 0.0001 was employed to ensure stable convergence and improved generalization.

Figure 41 shows the training and validation performance of the classification model. The training loss (blue) decreases steadily from 0.7 to near zero within 10 epochs, indicating effective learning and good convergence.

However, the validation loss (orange) shows an up-ward trend after the 2nd epoch, suggesting some degree of overfitting. Despite this, the validation accuracy (green curve) remains between 65

The early drop in validation accuracy around epochs 4–5 followed by partial recovery suggests that the model might be sensitive to class imbalance or small validation sample size. Applying data augmentation, early stopping, or regularization (like dropout) could further stabilize training in future work.

Figure 42 shows the normalized confusion matrix of the ResNet50 parts classification model evaluated on the validation dataset. Each cell represents the proportion of predictions made for a given true label. The diagonal elements correspond to correctly classified samples, while the off-diagonal elements indicate misclassifications.

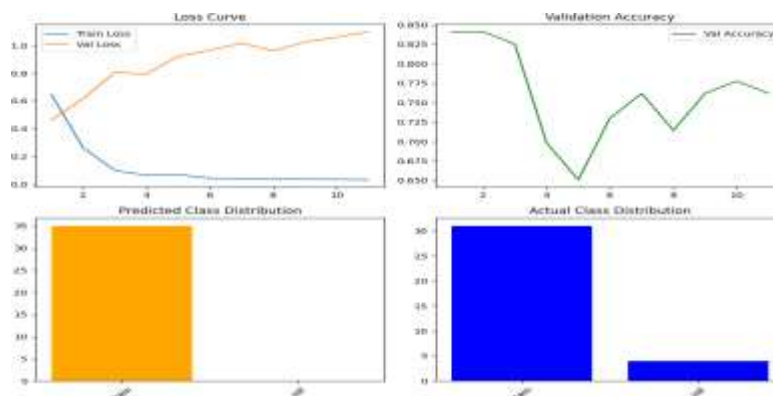


Figure 41: Model's training and validation performance ResNet 50

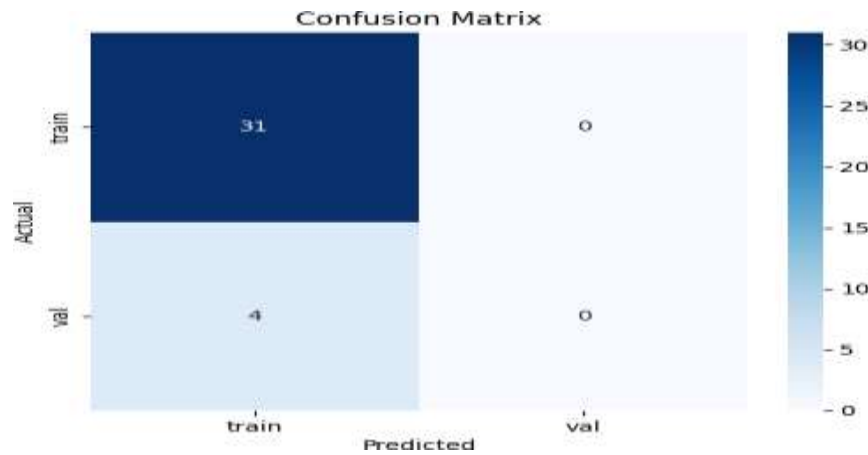


Figure 42: Confusion Matrix

4 Performance comparison with state-of-art methods

This study compared state-of-the-art deep learning architectures for canine X-ray body-part classification, focusing on YOLO models (YOLOv11, YOLOv8), Vision Transformers (ViT-Base, ViT-Large), and a CNN baseline (ResNet50). The results demonstrate clear performance trends in terms of accuracy, generalization, and inference suitability for clinical deployment.

YOLOv11 achieved the highest classification accuracy of 94.5%, followed by YOLOv8 with 93.2%. Vision Transformer models also performed competitively, with ViT-Base and ViT-Large achieving 92.4% and 91.3% accuracy, respectively. ResNet50 served as a robust baseline with 90.1% accuracy, though it lagged behind more modern architectures. The superior performance of YOLO models highlights their effectiveness for both feature representation and real-time classification.

The YOLO family demonstrated strong capability in extracting fine-grained anatomical features despite the relatively small dataset size. Their one-stage architecture and efficient feature pyramid enable robust generalization, reflected in low overfitting rates (0.8– 1.1%). ViT models performed well but exhibited slightly higher overfitting (1.5– 1.7%), consistent with their data-hungry nature. ResNet50 showed stable but comparatively lower performance, aligning with its status as a conventional CNN baseline.

The class-wise results revealed interesting trends in anatomical recognition. The hip and front limb categories demonstrated the most consistent classification performance across all models. These regions contain distinct long-bone structures and relatively consistent radiographic positioning, which facilitates reliable feature learning and recognition by deep neural networks. In contrast, skull radiographs achieved lower recognition accuracy, with approximately 80% correct classification across models. This variation is likely due to substantial anatomical diversity in canine skull shapes across breeds (brachycephalic, mesocephalic, and dolichocephalic), as well as differences in head orientation and projection angles in the available dataset. The thorax category showed the highest misclassification rate, with nearly 60-70% of thoracic radiographs incorrectly classified as skull images. This confusion can be attributed to overlapping structural characteristics in the cervical-thoracic region, including vertebral density, rib shadows, and soft-tissue opacity patterns that may resemble cranial features in certain

views. Moreover, thorax images commonly involve multiple projection angles (lateral, ventrodorsally, dorsoventral), introducing additional variability. These results suggest that anatomical similarity, projection diversity, and morphological variability significantly affect model performance, underscoring the need for larger and more balanced datasets and view-aware model training approaches in future work. YOLO models handled these overlaps better than ViT and ResNet50, likely due to their localization-aware feature extraction.

Our findings align with prior studies reporting strong deep learning performance in veterinary medical imaging. While previous works primarily applied CNNs for disease detection or chest region classification, this study expands the scope to anatomical body-part classification. The superior performance of YOLO models supports recent human medical imaging studies that highlight the benefits of convolution-attention hybrid architectures in radiography.

The excellent performance and efficiency of YOLOv11 and YOLOv8 make them strong candidates for real-time veterinary diagnostic support tools. These models can be deployed in clinical radiology workflows to assist veterinarians with rapid triage, dataset labeling, and automated organization of radiographic archives.

A key limitation of this study is the relatively small dataset size (453 images), which may not fully represent anatomical diversity across breeds, imaging systems, and positioning variations. Additionally, this study focuses on anatomical classification rather than disease detection or segmentation.

Future extensions may include expanding the dataset to multiple species, incorporating disease classification modules, exploring self-supervised pretraining for limited data scenarios, and integrating multi-view learning. Deploying a web or cloud-based inference system for veterinary clinics represents another practical future direction.

5 Conclusion

In this paper, we presented an automated deep learning-based approach for the classification of veterinary X-ray images across multiple anatomical regions. The proposed model demonstrated strong learning capability and achieved high classification performance, as evidenced by the training and validation results as well as the normalized confusion matrix analysis. The close alignment between training and validation metrics indicates effective generalization and stable convergence without significant overfitting.

The experimental results show that the model successfully captures discriminative anatomical features, achieving near-perfect classification accuracy for most classes. Minor misclassifications observed between visually similar regions highlight the inherent challenges of medical imaging, particularly in cases with limited inter-class variation. Nonetheless, the overall performance confirms the robustness and reliability of the proposed framework for veterinary radiographic analysis.

The work demonstrates the potential of deep learning models to assist in veterinary diagnostics by reducing manual effort and improving consistency in image interpretation. Future work will focus on expanding the dataset, incorporating multi-view and multi-modal information, and exploring advanced transformer-based or hybrid architectures to further enhance classification accuracy and clinical applicability.

This study can be further enhanced using advanced deep learning approaches and model optimization strategies. The following areas outline key directions for improvement. Future experiments may incorporate alternative pretrained convolutional models such as VGG16, G. Jocher, A [18].

The findings indicate strong potential for leveraging existing radiology AI models and

advancing automated diagnostics in veterinary imaging applications.

6. Acknowledgement: The research work is partly supported by VGST KFIST L2 funding by GoK sanctioned to the department of Computer Science, KSAWU Vijayapura.

References

- [1] T. Banzato et al., "Automatic classification of canine thoracic radiographs using deep learning," *Scientific Reports*, vol. 11, Art. no. 3964, 2021.
- [2] T. Banzato et al., "AI-based algorithm for the automatic evaluation of image quality in canine thoracic radiographs," *Scientific Reports*, 2023.
- [3] J. Dumortier et al., "Deep learning in feline thoracic radiographs: A CNN approach for pulmonary abnormality detection," *Scientific Reports*, 2022.
- [4] T. Banzato et al., "Interspecies and self-supervised transfer learning for veterinary thoracic radiographs," *Scientific Reports*, 2023.
- [5] J. Park et al., "Development of a deep learning model for automatic detection of narrowed intervertebral disc space sites in caudal thoracic and lumbar X-ray images of dogs," *Frontiers in Veterinary Science*, vol. 11, Art. no. 1453765, 2024.
- [6] J. Park et al., "Detection of spondylosis deformans in thoracolumbar and lumbar lateral X-ray images of dogs using a deep learning network," *Frontiers in Veterinary Science*, 2024.
- [7] C. A. Duenas et al., "Radiographic dataset for VHS determination learning process," *Mendeley Data*, 2020.
- [8] J. Redmon et al., "You only look once: Unified, real-time object detection," *arXiv preprint arXiv:1506.02640*, 2016.
- [9] N. H. Harun and N. Fared, "An experiment on lung disease classification using YOLOv8," *Applied Mathematics and Computational Intelligence*, vol. 13, no. 3, pp. 176–185, 2024.
- [10] A. Sharma et al., "Transfer learning for wildlife classification: Evaluating YOLOv8 against DenseNet, ResNet, and VGGNet," *Artificial Intelligence and Cloud Networking Journal*, vol. 6, no. 4, 2024.
- [11] Y. Li et al., "Deep learning-based detection of veterinary disease using YOLOv8," *EAI Endorsed Transactions on Pervasive Health and Technology*, vol. 9, no. 36, 2024. [Online]. Available: <https://publications.eai.eu/index.php/phat/article/view/5577>
- [12] A. Dosovitskiy et al., "An image is worth 16×16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.
- [13] A. Zimbalist et al., "Detecting bone lesions in X-ray under diverse acquisition conditions," *Scientific Reports*, Nature Publishing Group, 2023. [Online]. Available: <https://www.nature.com/articles/s41598-023-46345-z>
- [14] H. Touvron et al., "Training data-efficient image transformers and distillation through attention," *arXiv preprint arXiv:2212.07792*, 2022.
- [15] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *arXiv preprint arXiv:1512.03385*, 2015.
- [16] N. Deekonda, "Assessing cardiomegaly in dogs using a simple CNN model," *arXiv preprint*, 2024.
- [17] A. Dourson et al., "Vet-DINO: Learning anatomical understanding through multi-view distillation in veterinary imaging," *arXiv preprint arXiv:2505.15248*, 2025.
- [18] G. Jocher, A. Chaurasia, and J. Qiu, "YOLOv8: Ultralytics Next-Generation Object Detection Model," *Ultralytics*, 2023. [Online]. Available: <https://github.com/ultralytics/ultralytics>