BİNGÖL UNIVERSITY

Turkish Journal of Nature and Science

e-ISSN: 2149-6366
www.dergipark.gov.tr/tdfd

Research Article

Volume 15, Issue 2, Page 196-206, 2026

Received: 25.01.2026
Accepted: 10.05.2026
Published: 30.06.2026

DOI: https://doi.org/10.46810/tdfd.1871343

Comparative Evaluation of YOLOv8, YOLOv11 and ResNet32 for Deep Learning-Based COVID-19 Classification from Chest X-ray Images

Berin BALCI[1*], Fatih BAŞÇİFTÇİ[2]

[1]Tokat Gaziosmanpaşa University, Tokat Vocational School, Department of Computer Technologies, Information Security Technology Program, Tokat, Türkiye.
[2]Selcuk University, Faculty of Technology, Department of Computer Engineering, Konya, Türkiye

**Corresponding author: berin.balci@gop.edu.tr*

Abstract

Respiratory diseases such as COVID-19 and Viral Pneumonia present significant diagnostic challenges due to their similar radiological manifestations. In this context, the rapid and reliable classification of chest X-ray (CXR) images is of critical importance for clinical decision-making. In this study, the classification performance of YOLOv8 and YOLOv11 models was evaluated for a multi-class CXR classification task, with ResNet32 serving as the baseline model. A five-fold cross-validation strategy was applied to a dataset containing COVID-19, Normal, and Viral Pneumonia classes, and model performance was assessed using accuracy, precision, recall, and F1-score metrics. The results indicate that the evaluated models produce effective outcomes for multi-class CXR classification. YOLOv8 achieved the highest performance with an accuracy of 0.9806 and an F1-score of 0.9751, while YOLOv11 yielded comparable results with an accuracy of 0.9780 and an F1-score of 0.9679. The ResNet32 model achieved an accuracy of 0.9713 and an F1-score of 0.9585. However, relatively lower recall values were observed for the Viral Pneumonia class due to class imbalance in the dataset. Overall, the findings suggest that YOLO-based models provide an effective and reliable approach for medical image classification tasks. Future studies should focus on improving the generalizability of these models through the use of more balanced datasets and validation across diverse clinical settings.

Keywords

Chest X-ray,
COVID-19,
Deep Learning,
CNN,
YOLO,
Image classification,
Pneumonia

BİNGÖL ÜNİVERSİTESİ

Türk Doğa ve Fen Dergisi

e-ISSN: 2149-6366
www.dergipark.gov.tr/tdfd

Araştırma Makalesi

Cilt 15, Sayı 2, Sayfa 196-206, 2026

Alınış: 25.01.2026
Kabul: 10.05.2026
Basım: 30.06.2026

DOI: https://doi.org/10.46810/tdfd.1871343

Göğüs Röntgeni Görüntülerinden Derin Öğrenme Tabanlı COVID-19 Sınıflandırması için YOLOv8, YOLOv11 ve ResNet32 Modellerinin Karşılaştırılmalı Değerlendirilmesi

Berin BALCI[1*], Fatih BAŞÇİFTÇİ[2]

[1]Tokat Gaziosmanpaşa Üniversitesi, Tokat Meslek Yüksekokulu, Bilgisayar Teknolojileri Bölümü, Bilişim Güvenliği Teknolojisi Programı, Tokat, Türkiye
[2]Selçuk Üniversitesi, Teknoloji Fakültesi, Bilgisayar Mühendisliği Bölümü, Konya, Türkiye

**Sorumlu yazar: berin.balci@gop.edu.tr*

Özet

COVID-19 ve Viral Pneumonia gibi solunum yolu hastalıkları, benzer radyolojik bulgular sergilemeleri nedeniyle tanı sürecinde önemli zorluklar oluşturmaktadır. Bu bağlamda, göğüs röntgeni (Chest X-ray, CXR) görüntülerinin hızlı ve güvenilir bir şekilde sınıflandırılması, klinik karar verme süreçleri açısından kritik bir gereksinimdir. Bu çalışmada, YOLOv8 ve YOLOv11 modellerinin sınıflandırma başarısı çok sınıflı CXR sınıflandırma problemi kapsamında değerlendirilmiş ve ResNet32 modeli referans (baseline) model olarak kullanılmıştır. COVID-19, Normal ve Viral Pneumonia sınıflarını içeren veri seti üzerinde 5 katlı çapraz doğrulama stratejisi uygulanmış, model performansı doğruluk (accuracy), precision, recall ve F1-skor metrikleri kullanılarak değerlendirilmiştir. Elde edilen sonuçlar, incelenen modellerin çok sınıflı CXR sınıflandırma probleminde etkili sonuçlar ürettiğini ortaya koymaktadır. YOLOv8 modeli 0.9806 doğruluk ve 0.9751 F1-skor değeri ile en yüksek performansı elde ederken, YOLOv11 modeli 0.9780 doğruluk ve 0.9679 F1-skor değeri ile benzer sonuçlar sunmuştur. ResNet32 modeli ise 0.9713 doğruluk ve 0.9585 F1-skor değeri elde etmiştir. Bununla birlikte, veri setindeki sınıf dengesizliği nedeniyle Viral Pneumonia sınıfında görece daha düşük recall değerleri gözlemlenmiştir. Bulgular, YOLO tabanlı modellerin medikal görüntü sınıflandırma görevlerinde etkili ve güvenilir bir yaklaşım sunduğunu göstermektedir. Gelecek çalışmalarda, daha dengeli veri setlerinin kullanılması ve modellerin farklı klinik ortamlarda değerlendirilmesiyle sonuçların genellenebilirliğinin artırılması hedeflenmelidir.

Anahtar kelimeler

Akciğer grafisi,
COVID-19,
Derin öğrenme,
CNN,
YOLO,
Görüntü sınıflandırma,
Pnömoni

1. INTRODUCTION

Respiratory diseases such as COVID-19 and Viral Pneumonia affect millions of people each year and rank among the leading causes of death worldwide [1]. The first cases of COVID-19 were reported on December 31, 2019, in Wuhan, China, as Viral Pneumonia of unknown etiology, and rapidly evolved into a global pandemic [2,3]. Caused by the SARS-CoV-2 virus, this disease spread quickly and imposed a significant burden on healthcare systems [4]. In the United States, the first cases were reported in January 2020, and the number of confirmed cases exceeded 300,000 within a short period [5].

Due to the zoonotic nature of coronaviruses, they can be transmitted across species and cause various symptoms in humans, including fever, cough, sore throat, fatigue, muscle pain, and shortness of breath [6]. In addition to COVID-19, Viral Pneumonia caused by pathogens such as influenza, rhinovirus, and adenovirus also poses serious health risks [7,8]. Although Viral Pneumonia can occur in all age groups, it represents a life-threatening condition particularly for children and elderly individuals [9]. Therefore, early and accurate diagnosis is of critical importance.

In clinical practice, Reverse Transcription Polymerase Chain Reaction (RT-PCR) is considered the reference standard for the diagnosis of COVID-19. However, due to its limited sensitivity (60-70%), imaging modalities play a crucial complementary role [10-12]. Although Computed Tomography (CT) provides high sensitivity in detecting COVID-19 Viral Pneumonia, imaging findings often appear several days after symptom onset, and normal images may be observed in the early stages [13-15]. During the early phase of the pandemic, limitations in test kit availability led clinicians to rely more heavily on CT findings [14,16]. In many countries, including Türkiye, the combined use of imaging and laboratory data improved the accuracy of early diagnosis [17-19].

Lung imaging can be performed using Computed Tomography (CT), Magnetic Resonance Imaging (MRI), and Chest X-ray (CXR). Although MRI and CT provide higher diagnostic accuracy, CXR is the most widely used modality due to its lower cost and reduced radiation exposure [1]. However, the manual interpretation of X-ray images is time-consuming and subject to inter-observer variability. This limitation has increased the need for computer-aided analysis systems. Artificial intelligence, particularly deep learning methods, enables faster and more objective interpretation of medical images.

Deep learning is an approach that can automatically learn hierarchical feature representations from data without requiring manual feature extraction [20]. Consequently, it has been widely used in the field of medical image analysis and has achieved high performance. In the literature, the majority of studies on COVID-19 diagnosis using CXR images are based on Convolutional Neural Network (CNN) architectures. In these studies, CNN models designed with varying depths have achieved high classification performance [21] while higher accuracy rates have been reported with transfer learning-based CNN architectures [22]. Studies employing deeper and more optimized CNN-based models have demonstrated performance levels in the range of 97-98% [23,24]. In addition, hybrid approaches that combine deep learning-based feature extraction with classical machine learning methods have also yielded similarly successful results [25-28]. Studies evaluating large-scale datasets and various model architectures have reported that deep learning-based models provide high accuracy and strong generalization performance [29]. These findings indicate that CNN-based approaches have become a dominant and effective paradigm in the classification of CXR images. However, these models often involve high computational cost and model complexity, which may limit their efficiency in clinical applications requiring rapid decision-making.

You Only Look Once (YOLO) is an efficient deep learning architecture capable of performing simultaneous object detection and class prediction using a single convolutional network [30]. During the training process, YOLO iteratively updates its parameters by predicting object classes and their spatial locations, and it can be applied to various computer vision tasks, including object detection, image segmentation, pose estimation, and classification [31]. In this respect, YOLO offers lower computational cost and faster inference time compared to traditional CNN-based approaches, making it a more efficient alternative, particularly for real-time and clinical applications. YOLO architectures inherently incorporate classification; however, their use for classification tasks has remained relatively limited in the literature compared to object detection applications. Nevertheless, a limited number of studies have demonstrated that YOLO architectures can also produce successful results in image classification tasks [32-34]. However, this application has not yet become widespread, and to the best of our knowledge, no YOLO-based classification study has been reported specifically in the context of CXR images and COVID-19.

Accordingly, in this study, the classification capability of YOLO architectures was comprehensively investigated on a CXR dataset consisting of COVID-19, Normal, and Viral Pneumonia classes. To address the limited use of YOLO architectures in medical image classification, a systematic evaluation of YOLOv8 and YOLOv11 models was conducted and compared with a ResNet32-based baseline model within a 5-fold cross-validation framework. The main contributions of this study can be summarized as follows:

- A systematic evaluation of YOLO-based architectures, originally designed for object detection, is presented for multi-class medical image classification using CXR data.
- A comparative analysis of YOLOv8 and YOLOv11 models is conducted for COVID-19 classification, and the obtained results are benchmarked against a conventional CNN architecture, ResNet32.

• A robust experimental framework based on 5-fold cross-validation is employed to ensure reliable and generalizable results.
• The impact of class imbalance on model performance is analyzed and discussed.

2. MATERIAL AND METHOD

This section describes the CXR dataset, the YOLO-based models and the ResNet32 baseline, the experimental setup, and the performance evaluation metrics used in this study.

2.1. Dataset

In this study, the COVID-19 Radiography Database, publicly available on the Kaggle platform, was used [35]. The dataset consists of Posterior–Anterior (PA) CXR images collected from multiple healthcare institutions and contains a total of 15,153 images. Although the original dataset includes four classes, only three classes—COVID-19, Normal, and Viral Pneumonia—were considered in this study, and the Lung Opacity class was excluded to ensure a more consistent and clearly defined classification setting. The class-wise distribution of the dataset is presented in Table 1.

Table 1. Class distribution of the dataset

Class	Total
COVID-19	3,616
Normal	10,192
Viral Pneumonia	1,345
Total	15,153

The dataset was initially divided into two subsets as 80% training and validation and 20% test. Accordingly, 12,122 images were allocated for training and validation, while 3,031 images were reserved for testing. During the data splitting process, stratified sampling was applied to preserve the class distribution, ensuring that class proportions were consistently represented across all subsets. The class-wise distribution and data splitting details are presented in Table 2.

Table 2. Class-wise distribution of the dataset and data split

Class	Total	Training and Validation (80%)	Test (20%)
COVID-19	3,616	2,893	723
Normal	10,192	8,153	2,039
Viral Pneumonia	1,345	1,076	269
Total	15,153	12,122	3,031

A 5-fold cross-validation strategy was applied to the 80% training and validation subset. In this process, the data were partitioned into five folds, where different subsets were used for training and validation in each iteration, and each sample was used for validation only once. This approach aims to improve the generalizability of the model.

Since the dataset does not contain patient-level identifiers, the splitting process could not be performed at the patient level. This limitation introduces a potential risk of data leakage, as images from the same patient may appear in different subsets, and is therefore considered a limitation of this study. Nevertheless, the use of an independent test set provides a more reliable evaluation of model performance. Example images from the classes included in the dataset are presented in Figure 1.

COVID-19 **Viral Pneumonia** **Normal**

Figure 1. Random sample images from the dataset classes

2.2. You Only Look Once (YOLO)

You Only Look Once (YOLO) is an object detection framework that has gained significant popularity in recent years for real-time image analysis tasks. The YOLO architecture is composed of three major components: the backbone, which functions as a feature extraction network; the neck, responsible for combining and refining features across multiple scales; and the head, which produces the final predictions such as bounding boxes and class labels. Besides these primary modules, the architecture also incorporates supplementary parts, including the stem, downsampling layers, and core building blocks. The stem processes raw input data before passing it through the network, downsampling layers reduce the spatial resolution of feature maps to improve computational efficiency, and the core building blocks extract increasingly abstract and high-level feature representations from the images [36]. In addition, YOLO provides a versatile framework capable of performing various computer vision tasks, including image segmentation, pose estimation, and oriented object detection, in addition to object detection and classification.

2.2.1. YOLOv8

The YOLOv8 framework is defined by the hyperparameters depth_multiple, width_multiple, and max_channels, which determine the capacity of the model. These parameters control the depth and width of the network, as well as the size of the feature maps. The depth parameter determines the number of layers within the C2f blocks, while the width and maximum channel parameters adjust the dimensionality of the feature representations. The model begins with a stem structure that processes the input data, where convolutional layers are used to reduce the resolution while extracting initial low-level features. In the subsequent backbone stage, increasingly abstract and high-level feature representations are extracted through Conv and C2f blocks. Within this structure, Batch Normalization and the SiLU activation function contribute to stabilizing the learning process and improving model performance. The final part of the model is designed as a classification head, consisting of a convolutional layer, Adaptive Average Pooling, dropout, and linear layers. This structure enables the transformation of the extracted feature representations

into class labels, producing the final classification output. The overall architecture of the YOLOv8 classification model is presented in Figure 2.

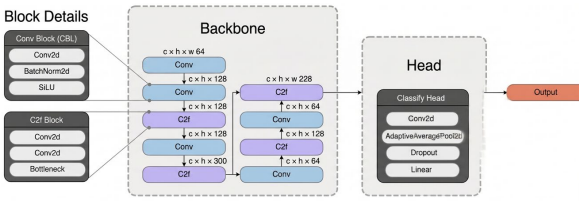


Figure 2. Overall architecture of the YOLOv8 classification model

2.2.2. YOLOv11

The YOLOv11 architecture is defined by the hyperparameters depth_multiple, width_multiple, and max_channels, which determine the capacity of the model. These parameters control the depth and width of the network, as well as the dimensionality of the feature representations. The depth parameter determines the number of layers within the C3k2 blocks, while the width and maximum channel parameters adjust the size of the feature maps. The model begins with a stem structure that processes the input data, where convolutional layers are used to reduce the resolution while extracting initial low-level features. In the subsequent backbone stage, increasingly abstract and high-level feature representations are extracted through Conv and C3k2 blocks. The C3k2 blocks are an improved version of the C2f blocks used in YOLOv8, providing efficient feature extraction with fewer parameters and improving computational efficiency. Within this structure, Batch Normalization and the SiLU activation function contribute to the stability of the learning process.

At the end of the backbone, the Spatial Pyramid Pooling Fast (SPPF) block combines features obtained from different receptive fields, enhancing multi-scale representation capability. The subsequent upsampling and concatenation operations enable the integration of feature maps from different levels, enriching the flow of information within the network [37].

The final part of the model is designed as a classification head, consisting of Adaptive Average Pooling, dropout, and linear layers. This structure transforms the extracted feature representations into class probabilities, producing the final classification output. The architectural structure of the YOLOv11 classification model is presented in Figure 3

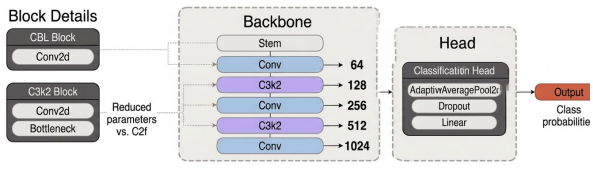


Figure 3. Overall architecture of the YOLOv11 classification model

2.3. ResNet32 (Baseline Model)

ResNet32 is a deep learning architecture incorporating residual connections, developed to mitigate the vanishing gradient problem and performance degradation observed in deep neural networks. These connections enable the direct propagation of input information to deeper layers, stabilizing the learning process and preserving the representational capacity of the model [38]. The ResNet32 architecture consists of 32 layers primarily composed of convolutional layers, organized into residual blocks. Each residual block includes two parallel paths: one performs convolutional transformations, while the other transmits the input directly through an identity connection. This design allows the model to learn residual functions instead of direct mappings, leading to a more efficient optimization process. The overall structure of this architecture is presented in Figure 4 [39].

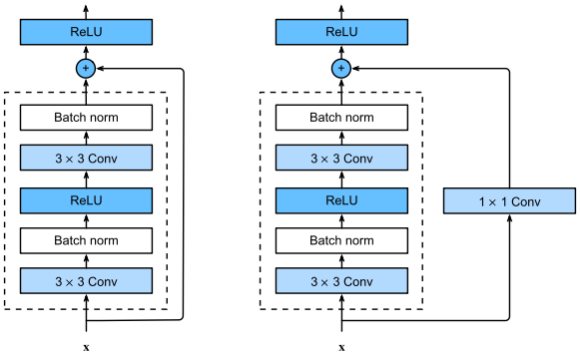


Figure 4. Residual block structure used in the ResNet32 architecture

The input images are fed into the model with a size of 32×32×3, and hierarchical feature representations are extracted through convolutional layers. In the model, the ReLU activation function is used for non-linear transformations, while Batch Normalization is applied to improve training stability. To reduce dimensionality and model complexity, a Global Average Pooling layer is applied in the final stage. This layer transforms the feature maps into a more compact representation, thereby reducing the risk of overfitting. The subsequent fully connected layer and softmax function convert the extracted features into class probabilities, producing the final classification output.

Due to its ability to achieve efficient learning with fewer parameters and its relatively low computational cost, ResNet32 is considered an appropriate reference (baseline) model for comparison with YOLO-based models.

2.4. Experimental Procedure

In this study, different deep learning-based model architectures were compared for the classification of CXR images. Image classification is an approach that aims to assign a single class label and a corresponding confidence score to an input image and provides an effective solution in cases where spatial localization of objects is not required. The main hyperparameters used in this study are presented in Table 3.

Table 3. Key hyperparameters used for model training

Parameter	Value	Description
task	classify	Specifies the classification objective
image size	128	Input resolution of the images
batch	16	Number of samples per training iteration
epochs	100	Total number of training epochs
patience	10	Early stopping threshold
optimizer	auto (Adam/AdamW)	Automatically selected optimization method
learning rate	0.01 / 0.001	Initial learning rate
momentum	0.937 / –	Gradient smoothing parameter (if applicable)
weight decay	0.0005 / 0	Regularization strength

In this study, the performance of models with different architectural designs was evaluated using YOLO-based classification models and the ResNet32 architecture. All models were trained under the same experimental conditions to ensure comparability. The input images were resized to 128×128 pixels, and a batch size of 16 was used during training. The training process was limited to a maximum of 100 epochs. However, an early stopping mechanism with a patience value of 10 was applied to prevent overfitting, and training was terminated when no improvement in validation loss was observed for 10 consecutive epochs. Pretrained weights were utilized for the YOLO-based models, and the optimization process was automatically configured by the Ultralytics framework. For these models, the initial learning rate was set to 0.01, and momentum and weight decay parameters were used to improve training stability and generalization performance. In contrast, the ResNet32 model was trained in the PyTorch environment using the Adam optimizer with a learning rate of 0.001, while other optimization parameters were kept at their default values. This approach allows each model to be evaluated under conditions appropriate to its architecture, providing a more balanced and realistic comparison. All experiments were conducted in the Google Colab environment using an NVIDIA L4 GPU.

2.5. Performance Evaluation Metrics

To evaluate the performance of the classification models, accuracy, precision, recall, F1-score, and macro average metrics were used. These metrics are calculated based on the confusion matrix, which represents the relationship between the predicted labels and the true class labels. The confusion matrix is a structured table consisting of true positive (TP), false positive (FP), true negative (TN), and false negative (FN) values, enabling a detailed analysis of model performance. The overall structure of the confusion matrix is presented in Figure 5.

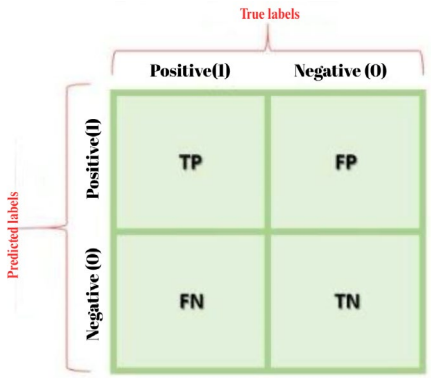


Figure 5. Structure of the confusion matrix

True positive (TP) refers to the case where the model correctly predicts an instance belonging to the positive class, while false positive (FP) denotes the misclassification of a negative instance as positive. True negative (TN) represents the correct classification of instances belonging to the negative class, whereas false negative (FN) refers to the case where a positive instance is incorrectly predicted as negative by the model.

Accuracy represents the proportion of correctly predicted instances to the total number of instances and is defined in Equation (1).

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

Recall represents the proportion of correctly predicted positive instances to the total number of actual positive instances and is defined in Equation (2).

$$Recall = \frac{TP}{TP + FN} \quad (2)$$

Precision represents the proportion of correctly predicted positive instances among all instances predicted as positive by the model and is defined in Equation (3).

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad (3)$$

F1-score is the harmonic mean of precision and recall and is used to provide a balanced evaluation of model performance. It is defined in Equation (4).

$$F1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \quad (4)$$

In this study, the metrics were calculated using the macro average approach to reduce the effect of class imbalance. macro average is defined as the arithmetic mean of the metrics computed for each class.

3. RESULTS

The performance of the evaluated models was assessed using standard classification metrics, including accuracy, precision, recall, and F1-score. To ensure a reliable evaluation, a 5-fold cross-validation strategy was employed, and the results were analyzed on a fold-wise

basis. The corresponding fold-wise performance results are presented in Table 4.

Table 4. Fold-wise performance results of the models based on 5-fold cross-validation

Model	Fold	Accuracy	Macro Precision	Macro Recall	Macro F1-score
YOLOv8	1	0.9732	0.9708	0.9484	0.9593
	2	0.9748	0.9761	0.9493	0.9620
	3	0.9831	0.9892	0.9720	0.9803
	4	0.9831	0.9987	0.9658	0.9764
	5	0.9887	0.9993	0.9953	0.9975
YOLOv11	1	0.9822	0.9815	0.9610	0.9709
	2	0.9736	0.9767	0.9488	0.9622
	3	0.9789	0.9847	0.9597	0.9717
	4	0.9777	0.9721	0.9595	0.9656
	5	0.9777	0.9804	0.9582	0.9689
ResNet32	1	0.9753	0.9611	0.9648	0.9629
	2	0.9604	0.9608	0.9289	0.9441
	3	0.9765	0.9708	0.9620	0.9662
	4	0.9686	0.9667	0.9477	0.9569
	5	0.9757	0.9650	0.9596	0.9622

For the YOLOv8 model, accuracy values ranged between 0.9732 and 0.9887, while macro F1-scores varied from 0.9593 to 0.9975 across the five folds. YOLOv11 achieved accuracy values between 0.9736 and 0.9822 and macro F1-scores between 0.9622 and 0.9717. In comparison, the ResNet32 model yielded accuracy values ranging from 0.9604 to 0.9765 and macro F1-scores between 0.9441 and 0.9662. The final epoch values corresponding to these results are presented in Table 5.

Table 5. Final epoch values of the models based on 5-fold cross-validation

Model	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
YOLOv8	20	31	45	80	33
YOLOv11	69	59	74	61	52
ResNet32	30	33	30	23	49

The values represent the epoch at which training was terminated for each fold. The variation across folds reflects differences in model convergence. The fold-wise results indicate variability in model performance across folds. To provide a more comprehensive evaluation, the mean ± standard deviation and minimum–maximum values derived from these results are presented in Table 6.

Table 6. Comparison of model performance based on 5-fold cross-validation (mean ± standard deviation and minimum–maximum values).

Model	Accuracy (mean ± std / min–max)	Precision (mean ± std / min–max)	Recall (mean ± std / min–max)	F1-score (mean ± std / min–max)
YOLOv8	0.9806 ± 0.0060 0.9732–0.9887	0.9868 ± 0.0113 0.9708–0.9993	0.9662 ± 0.0173 0.9484–0.9953	0.9751 ± 0.0139 0.9593–0.9975
YOLOv11	0.9780 ± 0.0028 0.9736–0.9822	0.9791 ± 0.0042 0.9721–0.9847	0.9574 ± 0.0045 0.9488–0.9610	0.9679 ± 0.0033 0.9622–0.9717
ResNet32	0.9713 ± 0.0061 0.9604–0.9765	0.9649 ± 0.0037 0.9608–0.9708	0.9526 ± 0.0132 0.9289–0.9648	0.9585 ± 0.0078 0.9441–0.9662

For YOLOv8, accuracy ranges between 0.9732 (min) and 0.9887 (max), precision between 0.9708 (min) and 0.9993 (max), recall between 0.9484 (min) and 0.9953 (max), and F1-score between 0.9593 (min) and 0.9975 (max). For YOLOv11, accuracy ranges between 0.9736 (min) and 0.9822 (max), precision between 0.9721 (min) and 0.9847 (max), recall between 0.9488 (min) and 0.9610 (max), and F1-score between 0.9622 (min) and 0.9717 (max). For ResNet32, accuracy ranges between 0.9604 (min) and 0.9765 (max), precision between 0.9608 (min) and 0.9708 (max), recall between 0.9289 (min) and 0.9648 (max), and F1-score between 0.9441 (min) and 0.9662 (max). These values reflect the range of model performance across folds. In addition to the overall performance metrics, class-wise performance results are presented in Table 7.

Table 7. Class-wise performance results (5-fold cross-validation, mean ± standard deviation).

Model	Class	Precision	Recall	F1-score
YOLOv8	COVID-19	0.986 ± 0.008	0.962 ± 0.009	0.974 ± 0.005
	Normal	0.976 ± 0.004	0.994 ± 0.005	0.985 ± 0.004
	Viral Pneumonia	0.984 ± 0.011	0.928 ± 0.027	0.952 ± 0.019
YOLOv11	COVID-19	0.979 ± 0.0047	0.952 ± 0.0098	0.965 ± 0.0038
	Normal	0.979 ± 0.0048	0.989 ± 0.0053	0.984 ± 0.0031
	Viral Pneumonia	0.980 ± 0.0097	0.953 ± 0.0083	0.966 ± 0.0051
ResNet32	COVID-19	0.966 ± 0.0056	0.948 ± 0.0232	0.957 ± 0.0122
	Normal	0.975 ± 0.0104	0.985 ± 0.0023	0.980 ± 0.0046
	Viral Pneumonia	0.952 ± 0.0152	0.923 ± 0.0197	0.937 ± 0.0099

The findings indicate that all models achieved the highest performance for the Normal class. This can be attributed to the class imbalance in the dataset, where the Normal class contains a larger number of samples. For the COVID-19 class, the YOLOv8 model achieved the highest precision and F1-score values compared to the other models, demonstrating superior performance. Although the YOLOv11 model produced comparable results, it showed slightly lower recall values than YOLOv8. The ResNet32 model exhibited lower performance in this class compared to the YOLO-based models. The Viral Pneumonia class emerged as the most challenging category for all models. In particular, the lower recall value observed for YOLOv8 (0.9280 ± 0.0270) indicates that the model missed some instances belonging to this class. In contrast, the YOLOv11 model demonstrated a more balanced performance and achieved the highest F1-score for this class.

Overall, YOLOv8 showed superior performance for the COVID-19 class, while YOLOv11 produced more balanced results, particularly for the Viral Pneumonia class. The ResNet32 model demonstrated lower but consistent performance across all classes. The training and validation loss curves of the models are presented in Figure 6.

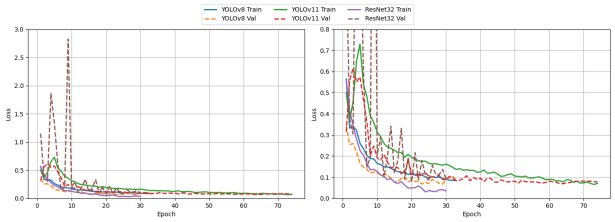


Figure 6. Training and validation loss curves of YOLOv8, YOLOv11, and ResNet32. The right plot shows a zoomed-in view to improve the visibility of loss trends, particularly for models with lower loss values.

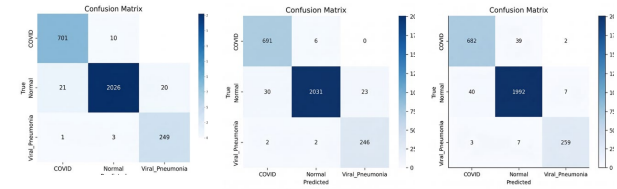
YOLOv8 and YOLOv11 exhibit a smoother and more stable decrease in loss, indicating a more efficient learning process. In contrast, the ResNet32 model shows higher fluctuations, particularly in the early epochs, but reaches a more stable state in the later stages. The small gap between training and validation losses in YOLO-based models suggests a lower risk of overfitting. The training process was terminated based on the early stopping criterion. Overall, YOLO-based models demonstrate a more stable and efficient training process compared to ResNet32.

The generalization performance of the best-performing model was evaluated using an independent test set. The corresponding results are presented in Table 8.

Table 8. Overall performance comparison of the models on the independent test set

Model	Accuracy	Precision	Recall	F1-score
YOLOv8	0.9819	0.978	0.969	0.973
YOLOv11	0.9792	0.983	0.955	0.968
ResNet32	0.967	0.961	0.961	0.961

To provide a detailed analysis of class-wise predictions, the confusion matrix is presented in Figure 7.



YOLOv8 **YOLOv11** **ResNet32**

Figure 7. Confusion matrices of the models on the independent test set

The YOLOv8 model exhibits fewer misclassifications compared to the other models and provides a clearer separation between classes. The YOLOv11 model demonstrates a similar performance, although a limited level of confusion is observed between the COVID-19 and Viral Pneumonia classes. In contrast, the ResNet32 model produces more misclassifications, with COVID-19 samples being more frequently confused with the Normal class. Across all models, the highest classification accuracy is achieved for the Normal class, while Viral Pneumonia remains the most challenging class to classify. The class-wise performance results of the models on the independent test set are presented in Table 9.

Table 9. Class-wise performance results of the models on the independent test set

Model	Class	Precision	Recall	F1-score
YOLOv8	COVID-19	0.969	0.986	0.977
	Normal	0.980	0.994	0.987
	Viral Pneumonia	0.984	0.926	0.954
	Macro Average	**0.978**	**0.969**	**0.973**
YOLOv11	COVID-19	0.991	0.956	0.973
	Normal	0.975	0.996	0.985
	Viral Pneumonia	0.984	0.914	0.948
	Macro Average	**0.983**	**0.955**	**0.968**
ResNet32	COVID-19	0.940	0.943	0.942
	Normal	0.977	0.976	0.977
	Viral Pneumonia	0.966	0.962	0.964
	Macro Average	**0.965**	**0.961**	**0.961**

All models achieved higher performance for the Normal class. For the COVID-19 class, YOLOv8 and YOLOv11 demonstrated strong performance, while the ResNet32 model produced relatively lower results. The Viral Pneumonia class appears to be more challenging for all models, with comparatively lower recall values observed, particularly for YOLOv8 and YOLOv11. Overall, the YOLOv8 model provides a more balanced performance across classes. Representative examples of classification results, including correct and incorrect predictions, are presented in Figure 8.

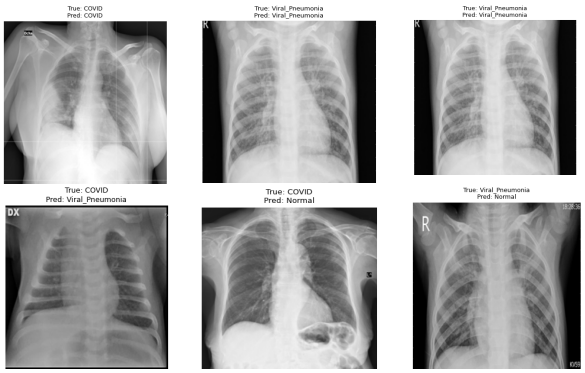


Figure 8. Representative CXR classification results, including correctly classified samples (top row) and misclassified samples (bottom row) across COVID-19, Normal, and Viral Pneumonia classes, with true and predicted labels indicated for each image

4. DISCUSSION AND CONCLUSION

In this study, YOLOv8, YOLOv11, and ResNet32-based deep learning models were used to classify CXR images. The findings indicate that the YOLOv8 model demonstrates a more stable and balanced performance compared to the other models. In particular, the more consistent results obtained during the 5-fold cross-validation process in distinguishing COVID-19, Normal, and Viral Pneumonia classes suggest that the model develops a more robust learning behavior against different data subsets.

The YOLO architecture is widely used in the literature, primarily for object detection tasks, where it has achieved high accuracy [40–42]. The results obtained in this study demonstrate that adapting YOLO-based models to

classification tasks is not only feasible but also capable of producing competitive results. The performance of the YOLOv8 model is consistent with the accuracy and stability trends reported across different datasets [39,43]. However, considering that the existing literature largely focuses on natural images and object detection problems, direct model comparisons on medical images remain limited. This suggests that the findings of this study contribute additional evidence to the limited literature on medical image-based model comparisons. In addition, although the present study includes a comparison with a ResNet32 baseline, future research could incorporate a broader range of CNN-based architectures to provide a more comprehensive evaluation [43].

When the model performances are compared, the higher accuracy and F1-score values achieved by YOLOv8 and YOLOv11 compared to ResNet32 indicate that architectural differences are directly reflected in performance. In particular, the multi-scale feature extraction capability of YOLOv8 provides a significant advantage in representing pathological findings that appear at different sizes and intensities in CXR images. This enables the model to learn more flexible representations that are not limited to specific features but can capture diverse pathological variations. Furthermore, it has been reported in the literature that YOLOv8-based models exhibit more stable and generalizable performance when combined with cross-validation strategies [44]. In this context, the observed performance improvement can be considered not only as an experimental finding but also as a result consistent with the structural properties of the model.

In contrast, the fluctuations observed in the validation loss of the ResNet32 model suggest a less stable optimization process and potentially limited generalization capacity. This may be related to training the model from scratch, as it is well established in the literature that CNN-based models can achieve stronger performance when combined with transfer learning [32]. Therefore, the performance of ResNet32 observed in this study may not fully reflect the full potential of the architecture.

When the class-wise results are examined, the lower recall values observed for the Viral Pneumonia class can be considered a direct consequence of class imbalance in the dataset. The limited number of samples belonging to the minority class may prevent the model from effectively learning class-specific discriminative features. This finding is consistent with studies highlighting the impact of class imbalance on model performance [40]. In addition, it is well established that standard loss functions may lead to biased learning in favor of the majority class, which can particularly result in decreased recall values [46]. Therefore, considering that evaluations based solely on accuracy may be misleading, model performance was assessed using precision, recall, and F1-score metrics. This approach is consistent with studies emphasizing the importance of multi-dimensional performance evaluation in imbalanced datasets [45]. This class imbalance represents an important limitation of the study, as it may have influenced the model's ability to generalize across all classes.

From a clinical perspective, YOLO-based models demonstrate potential for use in decision support systems and rapid screening applications. However, class imbalance, the use of a single-center dataset, and the absence of patient-level separation introduce a potential risk of data leakage, which may limit the generalizability of the results [46]. Therefore, validation with more balanced datasets obtained from multiple centers, as well as the application of data augmentation and class balancing techniques, is recommended. Future studies may also explore the integration of different deep learning architectures and advanced training strategies to further improve classification performance.

Overall, the findings indicate that YOLO-based models provide a competitive alternative to conventional CNN-based approaches in multi-class CXR classification. YOLOv8 achieved an accuracy of 0.9806 and an F1-score of 0.9751, while YOLOv11 achieved an accuracy of 0.9780 and an F1-score of 0.9679. ResNet32 achieved an accuracy of 0.9713 and an F1-score of 0.9585. In class-wise evaluation, lower recall values were observed for the Viral Pneumonia class across all models. One limitation of this study is the class imbalance present in the dataset, which may have affected model performance. In addition, although the models were compared with a ResNet32 baseline, future studies could include comparisons with a wider range of CNN-based architectures. These results indicate that YOLO-based models provide a competitive alternative to conventional CNN-based approaches in CXR Classification.

REFERENCES

[1] Uzen H, Firat H. Göğüs röntgeni görüntülerinden akciğer hastalıklarının sınıflandırılması için farklı derin öznitelikler ile beslenen destek vektör makinesi. Bilisim Teknol Derg. 2024;17(1):11-21.

[2] Wu F, Zhao S, Yu B, Chen YM, Wang W, Song ZG, et al. A new coronavirus associated with human respiratory disease in China. Nature. 2020;579(7798):265-9.

[3] Huang C, Wang Y, Li X, Ren L, Zhao J, Hu Y, et al. Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. Lancet. 2020;395(10223):497-506.

[4] Wu Z, McGoogan JM. Characteristics of and important lessons from the coronavirus disease 2019 (COVID-19) outbreak in China. JAMA. 2020;323(13):1239-42.

[5] Holshue ML, DeBolt C, Lindquist S, Lofy KH, Wiesman J, Bruce H, et al. First case of 2019 novel coronavirus in the United States. N Engl J Med. 2020;382(10):929-36.

[6] Singhal T. A review of coronavirus disease-2019 (COVID-19). Indian J Pediatr. 2020;87(4):281-6.

[7] Pagliano P, Scarpati G, Ascione T. Pneumonia in clinical practice: causes, risks, and management. Clin Respir J. 2021;15(6):567-75.

[8] GBD 2019 Antimicrobial Resistance Collaborators. Global mortality associated with 33 bacterial pathogens in 2019: a systematic analysis for the Global Burden of Disease Study 2019. Lancet. 2022;400(10369):2221-48,

[9] Rambaud-Althaus C, Althaus F, Genton B, D'Acremont V. Clinical features for diagnosis of pneumonia in children younger than 5 years: a systematic review and meta-analysis. Lancet Infect Dis. 2015;15(4):439-50.

[10] Zu ZY, Jiang MD, Xu PP, Chen W, Ni QQ, Lu GM, et al. Coronavirus disease 2019 (COVID-19): a perspective from China. Radiology. 2020;296(2):E15-25.

[11] Kanne JP, Little BP, Chung JH, Elicker BM, Ketai LH. Essentials for radiologists on COVID-19: an update. Radiology. 2020;296(2):200527.

[12] Xie X, Zhong Z, Zhao W, Zheng C, Wang F, Liu J. Chest CT for typical 2019-nCoV pneumonia: relationship to negative RT-PCR testing. Radiology. 2020;296(2):E41-5.

[13] Lee EY, Ng MY, Khong PL. COVID-19 pneumonia: what has CT taught us? Lancet Infect Dis. 2020;20(4):384-5.

[14] Bernheim A, Mei X, Huang M, Yang Y, Fayad ZA, Zhang N, et al. Chest CT findings in coronavirus disease-19: relationship to duration of infection. Radiology. 2020;295(3):200463.

[15] Pan F, Ye T, Sun P, Gui S, Liang B, Li L, et al. Time course of lung changes on chest CT during recovery from COVID-19 pneumonia. Radiology. 2020;295(3):715-21.

[16] Long C, Xu H, Shen Q, Zhang X, Fan B, Wang C, et al. Diagnosis of the coronavirus disease (COVID-19): rRT-PCR or CT? Eur J Radiol. 2020;126:108961.

[17] Li Y, Xia L. Coronavirus disease 2019 (COVID-19): role of chest CT in diagnosis and management. AJR Am J Roentgenol. 2020;214(6):1280-6.

[18] Shi H, Han X, Jiang N, Cao Y, Alwalid O, Gu J, et al. Radiological findings from 81 patients with COVID-19 pneumonia in Wuhan, China: a descriptive study. Lancet Infect Dis. 2020;20(4):425-34.

[19] Chan JFW, Yuan S, Kok KH, To KKW, Chu H, Yang J, et al. A familial cluster of pneumonia associated with the 2019 novel coronavirus. Lancet. 2020;395(10223):514-23.

[20] LeCun Y, Bengio Y, Hinton G. Deep learning. Nature. 2015;521(7553):436-44.

[21] Kaushik VS, Nayyar A, Kataria G, Jain R. Pneumonia detection using convolutional neural networks (CNNs). In: Singh P, Pawłowski W, Tanwar S, Kumar N, Rodrigues J, Obaidat M, editors. Proceedings of First International Conference on Computing, Communications, and Cyber-Security (IC4S 2019). Lecture Notes in Networks and Systems. Singapore: Springer; 2020. p. 471-83.

[22] Narin A, Kaya C, Pamuk Z. Automatic detection of coronavirus disease (COVID-19) using X-ray images and deep convolutional neural networks. Pattern Anal Appl. 2020;24:1207-20.

[23] Pathak Y, Shukla PK, Tiwari A, Stalin S, Singh S. Deep transfer learning based classification model for COVID-19 disease. IRBM. 2020;43(2):87-92. Mijwil MM, Al-Zubaidi EA. Medical image classification for coronavirus disease (COVID-19) using convolutional neural networks. Iraqi J Sci. 2021;62(8):2740-7.

[24] Mijwil MM, Al-Zubaidi EA. Medical image classification for coronavirus disease (COVID-19) using convolutional neural networks. Iraqi J Sci. 2021;62(8):2740-7.

[25] Pereira RM, Bertolini D, Teixeira LO, Silla CN Jr, Costa YMG. COVID-19 identification in chest X-ray images on flat and hierarchical classification scenarios. Comput Methods Programs Biomed. 2020;194:105532.

[26] Barstugan M, Ozkaya U, Ozturk S. Coronavirus (COVID-19) classification using CT images by machine learning methods. arXiv [Preprint]. 2020;2003.09424.

[27] Ozkaya U, Ozturk S, Barstugan M. Coronavirus (COVID-19) classification using deep features fusion and ranking technique. In: Big Data Analytics and Artificial Intelligence Against COVID-19. Studies in Big Data. Springer; 2020. p. 281-95.

[28] Sethy PK, Behera SK, Ratha PK, Biswas P. Detection of coronavirus disease (COVID-19) based on deep features and support vector machine. Int J Math Eng Manag Sci. 2020;5(4):643-51.

[29] Aggarwal P, Mishra NK, Fatimah B, Singh P, Gupta A, Joshi SD. COVID-19 image classification using deep learning: advances, challenges and opportunities. Comput Biol Med. 2022;144:105350.

[30] Shibly KH, Dey SK, Islam MTU, Rahman MM. COVID faster R-CNN: a novel framework to diagnose novel coronavirus disease (COVID-19) in X-ray images. Inform Med Unlocked. 2020;20:100405.

[31] Cai Z, Zhou K, Liao Z. A systematic review of YOLO-based object detection in medical imaging: advances, challenges, and future directions. Comput Mater Continua. 2025;85(2):2255-2303.

[32] Sharma S, Dhakal S, Bhavsar M. Transfer learning for wildlife classification: evaluating YOLOv8 against DenseNet, ResNet, and VGGNet on a custom dataset. J Artif Intell Capsule Netw. 2024;6(4):415-35.

[33] Khairkar AD, Kadam S, Kadam P, Deshpande S. Advancing dental implant classification through YOLO-based deep learning models. Int J Inf Technol. 2025. Epub ahead of print.

[34] Khokhariya A, Sarda J, Pradhan V, Vaghela R, Dave M, Thakkar A. Performance benchmarking of YOLO models for colorectal cancer histopathology image classification. In: Proceedings of the 2025 International Conference on Modeling, Simulation and Intelligent

Computing (MoSICom); 2025; Dubai, UAE. IEEE; 2025. p. 185-190.

[35] Chowdhury MEH, Rahman T, Khandakar A, Mazhar R, Kadir MA, Mahbub ZB, et al. COVID-19 Radiography Database [Internet]. Kaggle; 2021 [cited 2025 Jan 30]. Available from: https://www.kaggle.com/datasets/tawsifurrahman/covid19-radiography-database.

[36] Hidayatullah P, Syakrani N, Sholahuddin MR, Gelar T, Tubagus R. YOLOv8 to YOLO11: a comprehensive architecture in-depth comparative review. arXiv [Preprint]. 2025;2501.13400.

[37] Ultralytics. YOLO11 [Internet]. 2024 [cited 2025 Jan 30]. Available from: https://github.com/ultralytics/ultralytics.

[38] He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2016; Las Vegas, NV, USA. p. 770-8.

[39] Arik K, Agdas MT, Korkmaz A, Kosunalp S, Iliev T. Advanced classification of poxvirus-based skin diseases using deep learning techniques. Trait Signal. 2025;42(5):2777-86.

[40] Suparto A, Pribadi MR. Improving oil palm fruit detection under class imbalance using class-balanced focal loss on YOLOv11. J Sisfokom. 2026;15(2):165-76.

[41] Qu R, Yang Y, Wang Y. COVID-19 detection using CT image-based YOLOv5 network. In: 2021 3rd International Academic Exchange Conference on Science and Technology Innovation (IAECST). IEEE; 2021. p. 622-625.

[42] Xie Q, Zhang L, Chen Y. Fast-YOLO: real-time COVID-19 and pneumonia detection using optimized YOLOv11. Pattern Recognit Lett. 2025;178:57-64.

[43] Karadeniz AT, Başaran E, Çelik Y. Classification of walnut leaf images using a hybrid CNN-based deep learning approach. Turk J Nat Sci. 2026;15(1):192-201.

[44] Joshi K, Yadav Y, Hooda S, Nandal R, Singh B, Singh K, et al. Classification of cotton leaf disease using YOLOv8 based k-fold cross validation deep learning method for precision agriculture. Sci Rep. 2025;15(1):35602.

[45] Paul A, Raj R, Gourisaria MK, Jha AV, Bizon N. HARVEST: a locality-enhanced vision transformer for efficient multi-level grocery classification. Eng Rep. 2026;8(1):e70534.

[46] Islam MP, Hatou K, Shinagawa K, Kondo S, Kadoya Y, Aono M, et al. Hort-YOLO: a multi-crop deep learning model with an integrated semi-automated annotation framework. Comput Electron Agric. 2026;240:111196.