Effectiveness of Automatic Detection of Osteoarthritis using Convolutional Neural Network (CNN) Method with DenseNet201 on Digital Images of Knee Joint Radiography

Dea Nurfadhillah*, Gunawan Santoso, Fatimah, Gatot Murti Wibowo, Darmini, Nuryatno

Abstract. The manual detection of osteoarthritis using Kellgren Lawrence system depends on experience and agreement between doctors. The study was conducted to develop DenseNet201 to assist doctors in making a diagnosis of osteoarthritis grading. This study analyzes the accuracy; sensitivity; specificity; positive predictive value (PPV) and negative predictive value (NPV) of DenseNet201 in grading osteoarthritis and compares the classification results between DenseNet201 and radiologists in detecting osteoarthritis on knee joint images. This study is an applied experiment that compares the classification results of DenseNet201 and radiology specialists. Firstly, DenseNet201 is built with the MATLAB R2021a. Tests are carried out by measuring accuracy, sensitivity, specificity, PPV and NPV of 75 images of knee joint. Lastly, the data is analyzed using the Wilcoxon statistical test. The study has shown that the performance of DenseNet201 was good in detecting osteoarthritis, with accuracy value 91.84%; sensitivity value 76.61%; specificity value 94.32%; PPV 82.60% and NPV 94.32%. There was no significant difference between classification results using DenseNet201 and radiologist with a value (p>0.05) of 0.119. DenseNet201 can be considered as an alternative diagnostic tool for osteoarthritis with the condition that verification of the diagnostic decision still refers to the confirmation and justification of the radiologist.

Keywords: Osteoarthritis, Knee Joint, DenseNet201

*Corresponding author: deanurfadhillah01@email.com
1 Introduction

Osteoarthritis (OA) is a degenerative joint disease that shows damage to the joint cartilage, thickening of the subchondral bone, formation of osteophytes at the edges of the joints and a mild inflammatory process occurring in the nonspecific synovium [1]. OA can attack various joints in the body, but more often affects the weight-bearing joints body, such as the knee joint. MRI, CT scan, ultrasound and X-Ray are radiological examinations that are commonly used to examine OA of the knee joint, but OA examination with X-Ray is still the gold standard in detecting OA because it is more cost-effective, patient safety, wide accessibility, and time efficiency [2]. Examination of osteoarthritis using X-Ray is generally classified based on the severity of knee pain according to the Kellgren-Lawrance (KL) system. The KL system is divided into five levels from grade 0 to grade 4. However, the KL scoring system is influenced by practitioner subjectivity and its accuracy depends on experience and agreement between doctors. Misclassification often occurs when changing grades 3 and 4, so it is necessary to use AI in the detection of OA. The use of AI in this study aims to provide a common ground for doctors around the world and to be able to use CADx (Computer Aided Diagnostic) to assess the severity of the OA knee joint [3].

Deep learning is a branch of machine learning based on Artificial Neural Networks (ANN) or can be said to be a development of ANN. In deep learning, a computer learns to classify directly from images or sounds. Convolutional Neural Network (CNN) is one of the deep learning algorithms used for computer vision use cases such as classifying images or videos and detecting objects in images or even regions in images [4]. CNNs have been widely used in medical imaging, detection classification images and segmentation because it automatically learns all effective and relevant image features. CNN can learn directly from the image thereby reducing the burden of programming. Dense Convolutional Network (DenseNet) is a deep learning architectural model that connects each layer along with feature maps to all subsequent layers. The next layer will receive input feature maps from all previous layers. DenseNet connects each layer or block to each other layer or block in a feed-forward manner. DenseNet has several attractive advantages namely alleviating gradient problems, strengthening feature deployment, encouraging feature reuse, and substantially reducing the number of parameters [5].

Research by Tiwari et al. (2022) showed that identification of OA using DenseNet201 was better than various other models such as ResNet50, VGG-16, InceptionV3, and so on. Based on the performance tests conducted, DenseNet201 was able to obtain an accuracy value of 92.87%, a precision value of 93.69%, a recall of 92.53% and a loss of 0.20. Based on these results, DenseNet201 is the best model for identifying OA [6]. Based on the description above, the deep learning algorithm method proposed by the author to conduct research on automatic OA detection is automatic detection of OA using a Convolutional Neural Network (CNN) with DenseNet201.

2 Materials And Method

The study was approved by health research ethics committee ministry of health, Semarang Health Polytechnic (0433/EA/KEPK/2023). This study involved a diagnostic method based retrospectively collected radiographic knee joint rosenberg view examinations. The examinations had been performed by a DenseNet201, for assessment oh both presence and severity of osteoarthritis using the KL system. DenseNet201 identifies patterns in images based on input data and associated recurring learning. It is fed with both the input (the
radiographic images) and the information of expected output label (grading osteoarthritis) to establish a connection between the features of the different stages osteoarthritis. Before being fed to DenseNet201 for learning, the images are manually classified by a team of radiologist based on KL system.

2.1 Dataset

All the medical images were de-identified, without personal detail of patient. The inclusion exclusion criteria were applied after de-identification.

The following inclusion criteria were used: digital radiographic image from patient with age above 18 years (adult patients); digital radiography image from recorded patient unilateral or bilateral knee pain; digital radiography image of knee joint with result of the radiologist’s expertise that has been classified using the Kellgren Lawrance (KL) system grade 0 to 4 (grade 0 is normal knee joint, grade 1 possible osteoarthritis, grade 2 is mild osteoarthritis, grade 3 is moderate osteoarthritis and grade 4 is severe osteoarthritis); digital radiographic image of knee joint is shown knee joint anatomy included intercondylar fossae, soft tissue and trabeculae bone. The position of knee joint without any rotation.

The following exclusion criteria were used: digital radiography image from patients who had undergone operation for total knee arthroplasty in either of the knees; digital radiography image with poor resolution; digital radiography image with inaccurate collimation area and there are artifacts in digital radiography image of knee joint.

Standardized knee x-rays used in this study is rosenberg view, it consist of posteroanterior radiograph with weight-bearing and 45° of knee flexion. Center ray is at the level of knee joint typically 1.5 cm distal to the apex of the patella, with 10°-20° caudal angle.

2.2 Pre-Processing

The model were trained with 373 X-Ray images. Digital radiographic images of the knee joint available at the hospital are grade 0 = 50 images, grade 1 = 38 images, grade 2 = 165 images, grade 3 = 101 images and grade 4 = 19 images. The dataset was split into two subsets used for training and testing, with split ratio of 80-20. Initially, 80% of the images were used to train the model as input in training programming in the process of making DenseNet201 architecture. 20% of the Images used to test model DenseNet201 already done. An interpretable model took full knee radiographs as input and assessed grading knee joint osteoarthritis according to kellgren lawrance system. (figure 1).

![Fig. 1. Architecture for deep learning algorithms for digital radiography images of knee joint](image)

2.3 Input images

The algorithm processed the images collected for each class. Each DICOM format radiograph was converted into JPG format and downscale images from precious matrix size
2048 x 2500 pixels to 224 x 224 pixels to make it easier for image processing. The training was carried out by cropping and rotating. Figure 2 shows the labelling of the knee joint osteoarthritis images in different classes as 0, 1, 2, 3 and 4, as per the KL grading scale and knee severity from the x-ray images. The data were split into five different classes as per the KL grade shown in figure 2. Table 1 shows the data split in the training and testing subsets, according to KL grades.

![Image of knee joint osteoarthritis images](image)

**Fig. 2.** X-ray Images of different Kellgren Lawrance grades for Knee Joint osteoarthritis

**Table 1.** Data Splits in the Training and Testing Subset according to KL system

<table>
<thead>
<tr>
<th>Osteoarthritis Kellgren grade</th>
<th>Training</th>
<th>Testing</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Samples, n</td>
<td>Proportion, %</td>
</tr>
<tr>
<td>0</td>
<td>40</td>
<td>13.4</td>
</tr>
<tr>
<td>1</td>
<td>30</td>
<td>10.1</td>
</tr>
<tr>
<td>2</td>
<td>132</td>
<td>44.3</td>
</tr>
<tr>
<td>3</td>
<td>81</td>
<td>27.2</td>
</tr>
<tr>
<td>4</td>
<td>15</td>
<td>5</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>298</strong></td>
<td><strong>100</strong></td>
</tr>
</tbody>
</table>

2.4 DenseNet201 Architecture

In this study, transfer deep learning model were developed using MATLAB R2021a application software which is downloaded on the HP Pavilion 13-b127tu and then the model DenseNet201 development has the following steps: (1) Create a CNN training function, (2) Set a folder name, namely specifying the fold that will be used as training data and then renaming the folder with the name of the training data, (3) Save the image data along with the labels that will be trained, (4) Call the DenseNet201 network architecture, (5) Replace the final layer of the DenseNet201 architecture, (6) Read the number of class categories to be transferred, (7) Form a fully connected layer according to the number class category to be transferred, (8) Replace the fully connected layer with a new one, (9) Form a classification layer according to the category class to be transferred, (10) Replace the classification layer with a new one, (11) Replace all layers and connections between layers, (12) Determine the layer to be frozen (13) Read the size of the input layer, (14) Perform data augmentation with the aim of adjusting the size of the input image, (15) Set values that affect DenseNet201 network architecture training, including the max epoch value 5, minibatch 10, learning rate 0.0001 and 0.9 momentum, (16) Perform DenseNet201 network architecture training, (17)
Save DenseNet201 network architecture training, (18) Display DenseNet201 network architecture (19) Read class output of training results, (20) Plot ROC, (21) Form a confusion matrix, (22) Calculate the values of accuracy, sensitivity, specificity, PPV and NPV, (23) Display the evaluation results.

After the “training” syntax is complete, then proceed the preparation of DenseNet201 "test" syntax, with the following stages: (1) Check the name folder that will be tested, (2) Save the image data along with the labels that will be tested, (3) Perform data augmentation with the aim of adjusting the size of the input image, (4) Read the output class of the test results, (5) Plot the ROC, (6) Form a confusion matrix, (7) Calculate the accuracy, sensitivity, specificity, PPV and NPV values, (8) ) Show evaluation results, (9) Close all functions in MATLAB R2021A.

2.5 Test programing

This process is carried out by entering test data that has been divided into certain categories at the previous stage in the test programming (osteoarthritis severity categories, namely: normal, OA symptoms, mild OA, moderate OA and severe OA). After that, the DenseNet201 architecture is called as a result of the training that has been built is done. Furthermore, at this stage, the reading process of the test results is also carried out and then the accuracy of the test is calculated.

2.6 Classification Result

Knee joint OA can be classified based on several pathological features based on the Kellgren Lawrance standard. The purpose of the classification in this study is to detect the severity of osteoarthritis (grading), which means that the more similar the detected grading to the results of the radiologists expertise, the better. The classification results that appear in the deep learning model are the highest values. So, the higher the similarity value of the test data to the training data, the greater the accuracy of the DenseNet201 in detecting osteoarthritis and normal knee joints. The following fig 3 is a description of the detection.

![Fig. 3. Result of automatic Detection of grading of knee joint oa images using DenseNet201](image)

2.7 Performance assessment

After the classification process of the test data is done, the performance evaluation of the deep learning model was carried out using several methods, namely:
1. Cross validation, a test standard that aims to predict the error value of a Machine Learning. The purpose of testing the prediction error value is to obtain an optimal deep learning model.

2. Confusion matrix, the machine learning evaluation process uses a 2 x 2 table based on classification. Then a diagnostic test was carried out to measure the accuracy, specificity, positive predictive value (precision), and negative predictive value of model. Accuracy is a measuring tool for evaluating the performance of the classification algorithm. This shows the ability of the DenseNet201 to classify digital radiographic images of the normal knee joint and the osteoarthritis classification class. Specificity is used to denote the proportion of classes that are classified as negative from negative label. This shows the ability of the DenseNet201 in detecting normal knee joints. Sensitivity is indicating the proportion of positive class of positive class labels, in this case the sensitivity shows the ability of the DenseNet201 in predicting knee joints with osteoarthritis. The positive predictive value indicates the proportion of classes that are classified as positive on all positive prediction labels, in this case the positive predictive value indicates the ability of the DenseNet201 to detect knee joint osteoarthritis and classifying it as osteoarthritis. A negative predictive value indicates the proportion of normal knee joints that predicted as normal knee joints, in this case indicating the ability of the DenseNet201 in predicting normal knee joints and classifying it as normal knee joints.

3. ROC Curve, one of the most effective performance measurement tools for machine learning. In a diagnostic test for classification results that are dichotomous to the gold standard, specificity, and sensitivity and specificity are used to measure inherent validity, the ROC curve illustrates the trade off between specificity and sensitivity across a series of cutoff points when the diagnostic test is performed continuously. To determine the performance level of machine learning, the Area Under Curve (AUC) is calculated. The wider the AUC, the better the machine learning used. The AUC value is in the range of 0.0 to 1.

The test data for 75 knee joint images was interpreted by two radiologists who had more than 5 years of experience. The results of the interpretation of the two doctors' assessments were carried out with the percent agreement and Cohen's kappa tests to see the level of agreement in assessing the indicators on the research instruments given to the radiologists as respondents. Furthermore, the results of the interpretation of these two doctors were used to a diagnostic test to measure the accuracy, sensitivity, specificity, PPV and NPV of the results of the initial doctor's interpretation. After getting the performance results between the two doctors, one doctor was selected with the highest score from the diagnostic test. The performance between the DenseNet201 and the respondents was statistically compared to find out whether there were differences in the results of the classification using the DenseNet and the performance of the respondents using the Wilcoxon test.

3 Result and Discussion
The results of the classification are arranged in a 2 x 2 table according to Table 2-6.

**Table 2. Diagnostic Test of DenseNet201 grade 0 on Radiologist’s Expertise**

<table>
<thead>
<tr>
<th>Deep Learning Classification</th>
<th>Radiologist’s Expertise</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Positive</td>
<td>Negative</td>
</tr>
<tr>
<td>Positive</td>
<td>9</td>
<td>1</td>
</tr>
<tr>
<td>Negative</td>
<td>1</td>
<td>64</td>
</tr>
<tr>
<td>Total</td>
<td>10</td>
<td>65</td>
</tr>
</tbody>
</table>

**Table 3. Diagnostic Test of DenseNet201 grade 1 on Radiologist’s Expertise**

<table>
<thead>
<tr>
<th>Deep Learning Classification</th>
<th>Radiologist’s Expertise</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Positive</td>
<td>Negative</td>
</tr>
<tr>
<td>Positive</td>
<td>5</td>
<td>3</td>
</tr>
<tr>
<td>Negative</td>
<td>3</td>
<td>64</td>
</tr>
<tr>
<td>Total</td>
<td>8</td>
<td>67</td>
</tr>
</tbody>
</table>

**Table 4. Diagnostic Test of DenseNet201 grade 2 on Radiologist’s Expertise**

<table>
<thead>
<tr>
<th>Deep Learning Classification</th>
<th>Radiologist’s Expertise</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Positive</td>
<td>Negative</td>
</tr>
<tr>
<td>Positive</td>
<td>28</td>
<td>4</td>
</tr>
<tr>
<td>Negative</td>
<td>10</td>
<td>32</td>
</tr>
<tr>
<td>Total</td>
<td>38</td>
<td>35</td>
</tr>
</tbody>
</table>

**Table 5. Diagnostic Test of DenseNet201 grade 3 on Radiologist’s Expertise**

<table>
<thead>
<tr>
<th>Deep Learning Classification</th>
<th>Radiologist’s Expertise</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Positive</td>
<td>Negative</td>
</tr>
<tr>
<td>Positive</td>
<td>15</td>
<td>5</td>
</tr>
<tr>
<td>Negative</td>
<td>1</td>
<td>54</td>
</tr>
<tr>
<td>Total</td>
<td>16</td>
<td>59</td>
</tr>
</tbody>
</table>

**Table 6. Diagnostic Test of DenseNet201 grade 4 on Radiologist’s Expertise**

<table>
<thead>
<tr>
<th>Deep Learning Classification</th>
<th>Radiologist’s Expertise</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Positive</td>
<td>Negative</td>
</tr>
<tr>
<td>Positive</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>Negative</td>
<td>0</td>
<td>71</td>
</tr>
<tr>
<td>Total</td>
<td>4</td>
<td>72</td>
</tr>
</tbody>
</table>
The results of the diagnostic test (table 7) showed that DenseNet201 had accuracy and specificity in the very good category, while the positive predictive value (PPV) was in the good category and the sensitivity was in the fairly good category, so it can be said that DenseNet201 is good at detecting osteoarthritis in the knee joint. Based on the research process carried out, there are factors that can affect the accuracy level of DenseNet201, namely the number of images inputted as a training data set, where the more training data used, the higher the accuracy of the deep learning model in classifying. The more image data sets that are trained, the more DenseNet201 will recognize the characteristic characteristics of each image and will add references to the CNN algorithm arrangement. Matrix size also greatly influences the output results of deep learning models. The size of the image will be reduced to 224 x 224 and normalized based on the mean and standard deviation of the ImageNet training data set. In this study the DenseNet201 architecture will automatically resize the input image to 224 x 224. The larger the matrix size will affect the accuracy CNN deep learning models are getting higher. But on the other hand it can slow down DenseNet201's performance in classifying images. Based on the NDP and NDN values, it is known that the knee joint image was not detected as expected by DenseNet201 and there was an error in making the prediction, meaning that DenseNet201 does not have 100% accuracy in image classification and cannot detect osteoarthritis grading as a whole. It can be seen that the error of the grading prediction is from DenseNet201 is in the false negative (the detection of the grading oa is not in accordance with the gold standard). There are several reasons why DenseNet201 is not perfect or gets 100% in accuracy, namely radiographic images of knee joint examinations are often of low resolution so that they can affect the noise removal process, image enhancement to improve deep learning performance is one of the causes of imperfect DenseNet201 in terms of accuracy. Based on the table above, it is known that the DenseNet accuracy rate is 91.74%, the sensitivity is 76.61%, the specificity is 94.32%, the PPV is 82.60% and the NPV is 94.32%. The sensitivity value describes the ability of DenseNet201 to give a positive value where the results of the initial radiologist's expertise also produce a positive value. Meanwhile, the NDP describes DenseNet201's ability to show knee joint OA and classifies it as knee joint OA. The sensitivity and NDP values are lower than the other values, this is due to the small amount of data trained so that the convolution process has not received much reference to the characteristic patterns in pathology. This affects the ability of DenseNet201 to detect and recognize the characteristic features of the OA knee joint image. So it is shown that the deep learning model is less able to recognize the characteristics of knee joint osteoarthritis image. ROC is a performance measurement tool for DenseNet201. The AUC curve is a range of values that determines the performance capability of a deep learning model whether or not it fails based on the area of the curve, the wider the AUC or the closer it is symmetrical to its sensitivity value, the better the deep learning model.
In DenseNet201, the AUC value of grade 0 was 0.995956 and grade 4 was 0.905356 which is above the 0.9 range, meaning that DenseNet201 shows performance in the “very good” category. Grade 1 is 0.856396 above the range 0.8 - 0.9 which means DenseNet201 shows performance in the “good” category. Grade 2 is 0.733124 and grade 3 is 0.751602 indicating a range above 0.7 - 0.8 which means that DenseNet201 shows performance in the “acceptable” category, which means DenseNet201 is feasible and good enough to be used to detect radiographic images of OA and normal knee joints. Based on the AUC value obtained, it can be said that there is a mistake from DenseNet201 in automatically detecting OA grading, while in the process of determining the diagnosis it must be determined on a clinical basis in order to get the right diagnosis. In other words, DenseNet201 can be used under certain conditions.

The respondent’s performance assessment can be seen in the table below. This assessment is intended as a comparison of the feasibility of the classification results from DenseNet201. The percent agreement between the two doctors was 95% and the Cohen’s kappa of the two respondents was 0.928, it was concluded that the level of agreement between the respondents was objective or had a good agreement. Based on the diagnostic test between respondents (table 8), respondent 2 has the highest performance value with an accuracy of 0.890. so that it was chosen as a reference for comparison with the CNN deep learning model of the DenseNet201 architecture.

Table 8. Respondent’s Performance

<table>
<thead>
<tr>
<th>Performance</th>
<th>Respondent</th>
<th>1</th>
<th>2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy</td>
<td></td>
<td>0.867</td>
<td>0.890</td>
</tr>
<tr>
<td>Sensitivity</td>
<td></td>
<td>0.710</td>
<td>0.768</td>
</tr>
<tr>
<td>Specificity</td>
<td></td>
<td>0.912</td>
<td>0.927</td>
</tr>
<tr>
<td>PPV</td>
<td></td>
<td>0.687</td>
<td>0.741</td>
</tr>
<tr>
<td>NPV</td>
<td></td>
<td>0.912</td>
<td>0.927</td>
</tr>
</tbody>
</table>

Comparison of respondents and DenseNet201 is intended to find out whether the DenseNet201 created can be a support for radiologists in hospitals. The results of the comparative test showed that there was no significant difference between the results of the performance of the DenseNet201 architecture CNN deep learning model and the results of the radiologist’s expertise in detecting the severity of OA/grading OA. Based on these results, DenseNet201 can be used as a support for radiologists in detecting OA grading. Based on the Wilcoxon test the performance of DenseNet201 and radiologists have a p-value (p > 0.05) of 0.119, which means that when DenseNet201 is applied to the population, DenseNet201 can provide high accuracy in predicting. Therefore, an alternative to this method is to add more image populations, as well as research funds spent more and use test data that meets the image criteria which in a structured way shows a picture of knee joint osteoarthritis according to the Kellgren Lawrence system and increases the focus of detection on several images on osteophyte area and JSN so that it can increase DenseNet201’s ability to differentiate data between grades.

DenseNet201 is intended to assist radiologists in determining a diagnosis amidst a lack of accuracy in detecting grade of knee joint OA, which is expected to make work easier, improve the diagnostic process and still get fast and accurate results. In practice, the diagnosis decision must still refer to the confirmation and justification of the radiologist.

The limitations of this study are that DenseNet201 cannot produce a 100% figure in detecting osteoarthritis in knee joint images, besides that the data is limited so that it affects
4 Conclusion

The results showed that the performance of DenseNet201 model was quite good in detecting osteoarthritis with an accuracy value 91.84%, sensitivity value 76.61%, specificity value 94.32%, PPV is 82.60% and NPV is 94.32%. There was no significant difference between the classification result using DenseNet201 model and the radiologist with a value (p>0.05) of 0.119. DenseNet201 model can be considered as an alternative diagnostic tool for osteoarthritis, provided that diagnostic verification decisions are based on radiologist confirmation and justification.

References