Comparing Machine Learning Algorithm Detection for Development of an Artificially Intelligent Assistant Radiologist

AP Research 2023

Word Count: 4243

4/14/2023

Introduction

X-ray imaging is the most common form of radiographic imaging in hospitals and care facilities around the world (Bindman et al., 2008). Due to the large amount of medical case interpretation performed each day by radiologists, radiography is a target for the development of artificial intelligence (AI) solutions to improve the efficiency and quality of healthcare. Advancements in AI such as progress in deep learning research have ignited efforts that focus on conventional radiographs (x-rays) due to their importance in radiology practices, the large amounts of image data available for training algorithms, and their simplicity as images. (Hinton, 2018; Lecun, 2015)

Research Question

Which type of Machine Learning Algorithm is most applicable when detecting fractures in MSK (Musculoskeletal) scans of the human humerus to develop an AI assistant radiologist?

Definitions

Artificial Intelligence, according to Stanford's John McCarthy (2007), is defined as "...the science and engineering of making intelligent machines, especially intelligent computer programs" (pg. 2).

According to International Business Machines Corporation IBM (2023), a company that is a forerunner in the development of AI, machine learning can be defined as a subset of artificial intelligence that describes the plethora of methods for optimization based on a given set of data. Methods such as backpropagation allow computers to create neural networks that are trained and then tested on their ability to analyze datasets and identify patterns. Neural networks are a form of machine learning, referred to as deep learning. These networks are named for their many nodes and connections, which are similar to biological neurons and axons. This paper implements different instances of a specific type of neural network, called a Convolutional Neural Network (CNN), which is optimized for binary image classification and pattern recognition.

Binary image classification can be defined as the process of predicting a label's value for each given image. For example, a machine sorting normal and abnormal chest radiographs is a good example of binary image classification (Cheng et al., 2021). A bounding box is a coordinate plane the machine defines based on the pixel dimensions of an image, and typically, image classification algorithms detect the location and spatial extent of objects within a defined bounding box.

Literature Review

Artificial intelligence's use for the interpretation of conventional radiographs has achieved high performance for a number of use cases, (Lakhani et al., 2017; Lindsey et al., 2018; Rajpurkar et al., "CheXNet: radiologist-level pneumonia detection on chest x-rays with deep learning.", 2017) and AI products are currently being produced for medical image analysis (GE Healthcare, 2019; Zebra Medical Vision, 2019). These commercialized programs are providing opportunities for hospitals to use AI to improve image analysis and patient care.

Humerus Fracture

While over the past few decades, technology has been able to rapidly improve the quality of healthcare in our society, hospitals today still have issues. A major cause of dissatisfaction for patients with healthcare providers is lengthy waiting time before getting diagnoses and treatment (Anderson et al., 2007; Georgievskiy et al., 2008; Huang et al., 1994; Soriano, 1996). Patient anxiety while waiting for test results is known to have negative effects for individual patients. Given a dataset of patients with ranging medical histories and backgrounds, it was found that a majority of patients experience anxiety, and 45% experiencing emotional change while waiting for image test results (Woolen et al., 2018). Because of this anxiety, and based on the fact that there is great variation in waiting times for radiology, it has been concluded that there is room for improvement in the radiological service organization, as it is not optimal (Nuti & Vainieri, 2012). Improving the efficiency of MSK analysis can give patients an overall better experience in hospitals.

Image Classification

Preparing medical image data for machine learning tasks is a highly-complicated process that can be completed in many different ways, depending on a network's intended purpose (Montagnon et al., 2020; Willemink et al., 2020). In the case of neural networks and deep learning, it is essential that the training images that are generally representative of the task to be solved. With limited data, a model can easily be trained to the point of 100% accuracy when predicting labels on given training data, but then it may label new data extremely poorly. These models are considered to overfit the data in the training set (Salman & Liu, 2019), also known as exhibiting poor generalization. Images from a single source, medical center may indicate specific trends, and cannot be applied to all data in existence. A diverse set of data is important to train a model for a given task, as insufficient information can lead to biased predictions caused by a biased sample.

Currently, there are many efficient methods of image classification, but there are a few models that are most optimal, according to recent studies. In the current medical studies, Densenets are

one of the most common models used to detect abnormalities in MSKs, and are used in many modern studies in the field (Chauhan et al., 2021). A Densenet is a deep learning model that has denser connections to improve accuracy. Dense connections are similar to feed-forward networks, or FFNs. Both FFNs and Densenets use non-cyclical inputs that all consider a specific weight, making them more accurate for data with large amounts of inputs. Because image classification considers each pixel in an image, it can better consider the high amounts of data, giving Densenets a large advantage in image classification (Zhang et al., 2021). Another form of neural network that is commonly used in image classification in MSKs are residual networks, also known as Resnets. Resnets are also a common form of neural network that uses less layers than Densenets. These algorithms are extremely prevalent in the image classification of MSK images in other literature. They are known for requiring low computing power, but having high accuracy.

Artificial Intelligence in Hospitals

With the rapid development of new technologies, well-developed Artificial Intelligence models and deep learning networks are beginning to outperform humans in complicated tasks, such as the post-processing of CT, MR, and nuclear medicine studies. Current research indicates that AI will soon be able to automate the tasks of a radiologist and complete them at the same or greater speed, accuracy, and scale (Akkus et al., 2018; Nie et al., 2016; Wang et al., 2019; Yoon et al., 2015). Early research also indicates the potential for the creation of CT images from MRI scans or vice versa, called synthetic modality transfer, (Nie et al., 2016). This would remove the need for an expensive and separate imaging procedure. While AI has yet to reach a level where its analysis could replace human doctors', deep learning technology can improve the quality of care by assisting other doctors. Technology has allowed radiologists to have more optimized workspaces and easier access to data, allowing for an increase in efficiency (Hawnaur, 1999; van Lent et al., 2012). Evidence also suggests that these positive consequences of automation have led to greater workloads and examination speeds, causing a negative impact on radiographer morale, role satisfaction and "burn out" (Hutton et al., 2014; Lohikoski et al., 2019; Sheth et al., 2010; Singh et al., 2017). Despite fears of burn out, having an artificially intelligent assistant radiologist will support doctors, allowing them to have better accuracies when analyzing images. This will improve the quality of healthcare provided in hospitals.

As AI becomes more powerful, it is important to consider how it can be used in hospitals, as well as the many problems that could be experienced by hospital employees and patients. Deep learning in hospitals has generally great benefits, if used correctly. For example, it could help combat radiologist fatigue, caused by high demand for their services. As confidence in the security of the radiology field lessens, the amount of new radiologists per year is decreasing. This has caused each radiologist to read an increasing amount of cases per day. Physician shortages further contribute to the issue, especially for areas that are considered to be medically

unserved (Nakajima et al., 2008). Physician fatigue is a well-known problem that affects many healthcare professionals, but research has shown that radiologists are particularly affected. Fatigue is so impactful that radiologists' abilities to read cases accurately sometimes suffer. (Bhargavan & Sunshine, 2005; Lu et al., 2008; Berlin, 2000; Fitzgerald, 2001). Research that compared radiologist workload tiredness during their analysis of MSK images found a statistically significant decrease in the detection of fractures after an eight-hour workday compared to the start of the day (Krupinski et al., 2010). Thus, a program that implements a network to highlight abnormalities locally in an image can draw the attention of a clinician. If implemented correctly, a form of this program could lead to more accurate and efficient analysis of images, reduce errors, and help to standardize the quality of healthcare around the world.

Methods

Data preparation

Stanford's MURA dataset contains 40,561 images of Musculoskeletal (MSK) radiographs from 14,863 studies of the upper extremity. Data was segregated to focus on humerus scans, and further segregated into test data and training data, with 436 images used for training and 291 images reserved for testing. All of the images were labeled as normal or abnormal by doctors before any machine interaction.

Data augmentation

In preparation for model creation, images were scaled to 224 pixels by 224 pixels. Using bilinear interpolation, pixel values were normalized. Throughout training, the images were transformed multiple times in various manners: rotation up to 30°, reflection of images across both horizontal and vertical axes, and enlargement of up to 110%.

Model training

Three models were compared and developed, with Densenet-169, Densenet-201, and Resnet-101 architectures. These models were selected because they represent some of the most successful and common models compared to previous studies despite having vastly differing technology requirements. These models were all created using version 3.10.10 of the Python programming language (python.org), and the Tensorflow (tensorflow.org), Keras (keras.io), and OpenCV (opencv.org) libraries. The training was done using a workstation running on Windows 10 and 1 Nvidia GeForce RTX 2080 Ti graphics card (11 GB of RAM). My method was based upon Chauhan T et. al's study that analyzed MSKs with AI (2021). I implemented methods similar to that of this study to compare the efficacy of Densenet-169, Densenet-201, and Resnet-101, but I did not compare them in such depth, or using humerus fractures.

Because I needed to accurately compare each model, and each model required slightly different processing power, technical factors, such as batch size and epochs must be considered. Put simply, the batch size is a number of samples processed before the model is updated and the number of epochs is the number of complete passes through the training dataset. The size of a batch must be more than or equal to one and less than or equal to the number of samples in the training dataset. Because my maximum batch size was limited mainly by the available GPU-RAM and therefore could have only increased to a limited amount in larger networks, batch size was not varied, and set as constant throughout the models at 32. Especially with increased image resolution, lowering the batch size would be a major limitation to network performance. The different models were trained for 10 and 12 epochs. Since the performance of

a neural network can be subject to minor random fluctuations, the training was repeated for a total of three times.

Evaluation

Predictions on the validation dataset of the models for each of the three network architectures were combined and averaged so that the models could be evaluated together. For each of these trials, receiver operating characteristic (ROC) curves were graphed and values for the areas underneath each of these curves were calculated (AUROC). I chose to use AUROC because it is a value to compare each model to each other, but is also standardized, so this data can easily be compared to other research papers.

Diagram

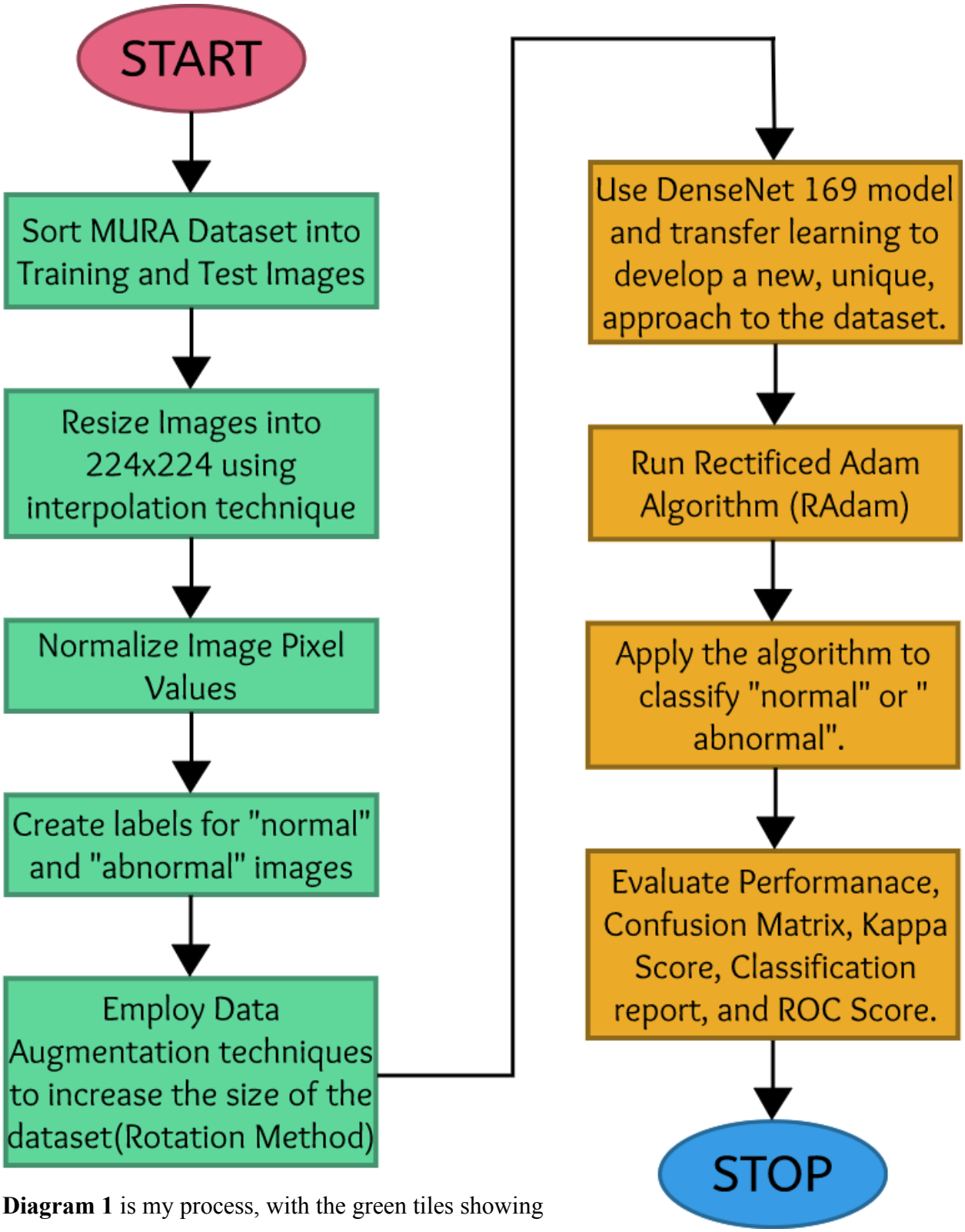


Diagram 1 is my process, with the green tiles showing data preparation and augmentation, and the yellow panels showing training and augmentation

Results

Measurements

After training, each model was tested on the remaining half of the MURA dataset not used for training. During these tests, four values were measured: the amount of True Positive readings, True Negative readings, False Positive readings, and False Negative readings. True Positive refers to instances where the algorithm correctly identified an image that was positive for an abnormality, indicating a fracture, while a True Negative reading indicated a correct identification for the lack of a fracture. Conversely, a False Positive reading occurred when the machine labeled a case as positive even though it was negative, and a False Negative would be a positive diagnosis for a negative case.

Table 1

	Avg. True Positive (TP)	Avg. True Negative (TN)	Avg. False Positive (FP)	Avg. False Negative (FN)
Resnet-101	118.67	117.33	34.33	21.67
Densenet-169	123.00	116.00	32.67	20.33
Densenet-201	127.00	119.00	27.67	18.33

Because they have variation in training data, it is expected that each model produced a slightly different amount of correct values in every instance, each model was trained three times, and then experimentally tested, producing the averages shown in the table above. Three major values were derived from these four measurements shown in the table above: Cohen's Kappa Score, Accuracy, and AUROC (Area Under the Receiver Operating Characteristic curve).

Cohen's Kappa Score

Cohen's Kappa (κ) is a statistic that measures interrater reliability. The statistic's purpose is to measure the consistency of each model when it makes decisions. The standard value ranges from negative one to one, with zero being the amount of agreement expected by random chance. Cohen suggested the Kappa result be interpreted as follows: values ≤ 0 as indicating no agreement and 0.01–0.20 as none to slight, 0.21–0.40 as fair, 0.41– 0.60 as moderate, 0.61–0.80 as substantial, and 0.81–1.00 as nearly perfect agreement. Cohen's Kappa can be calculated via the following formula:

$$\kappa = \frac{Pr(a) - Pr(e)}{1 - Pr(e)}$$

In this case, Pr(a) represents the observed agreement in the study, while Pr(e) represents the agreement that occurred by chance. Given a value n that represents the number of cases (which is 292 for all models), Pr(a) and Pr(e) can be represented via the following:

$$Pr(a) = TP + TN + FP + FN$$

$$Pr(e) = \frac{\left(\frac{(TP + FN) \cdot (TP + FP)}{n} + \frac{(TN + FN) \cdot (TN + FP)}{n} \right)}{n}$$

Table 2

Model	Avg. Kappa Score
Resnet-101	0.6587347649448636
Densenet-169	0.6736740597878496
Densenet-201	0.6800618238021638
Human Radiologist	0.847333333

Because these values of the Kappa Score fall between 0.61 and 0.80, by definition, they provide substantial values of agreement. This proves that each of these neural networks are not purely random, and further proves their correlations are not due to randomness. The human radiologist values are provided by Stanford in their initial case study done in conjunction with the release of the MURA dataset (Rajpurkar et al., "Mura: Large dataset for abnormality detection in musculoskeletal radiographs.", 2017). The average radiologist's Kappa Score is much higher than the ML models, however, demonstrating that the experienced radiologists that Stanford tested on the same dataset were much better at classifying the images than the CNNs.

Accuracy (Raw Agreement %)

As the true value of artificial intelligence comes from measuring how accurately the model can analyze X-ray images, it is important to compare each model's readings to the true values. To calculate the accuracy of each model based on the four measurements provided, the following formula was used:

$$ACC = \frac{TP + TN}{TP + TN + FP + FN}$$

Table 3

Model	Avg. Accuracy Percentage
Resnet-101	80.82%
Densenet-169	81.85%
Densenet-201	84.24%

The averages of the accuracies of each network are close in value, but show a slight trend with higher-layer neural networks having slightly higher accuracies.

ROC Curve

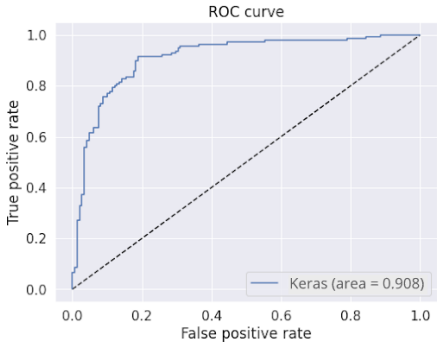
The Receiver Operator Characteristic Curve (ROC Curve) compares the True Positive Rate (TPR) and the False Positive Rate (FPR). This graph is useful when comparing which models do a better job at classifying positive data. The diagonal line splitting the middle of the graph represents a random classifier, so the farther above this random classifier the ROC Curve is, the better the model. The graph is useful for directly comparing the efficacy of each model. Modeling the ROC Curve compares TPR and FPR, which were calculated using the following formulas:

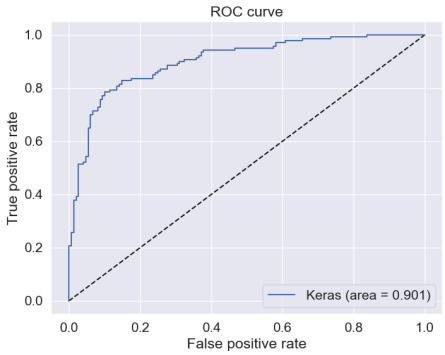
$$TPR = \frac{TP}{(TP + FN)}$$

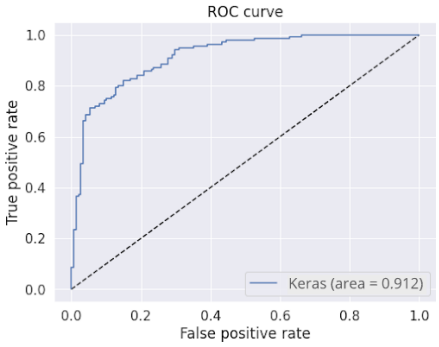
$$FPR = \frac{FP}{(FP + TN)}$$

Receiver Operator Character Graphs

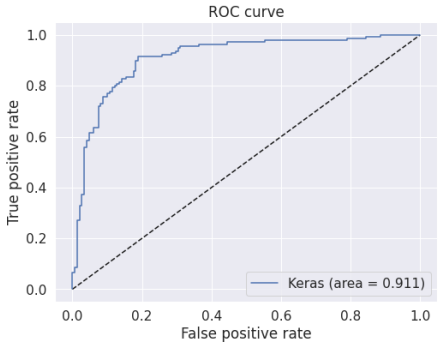
Resnet-101

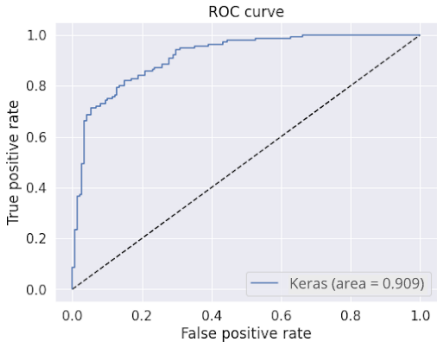


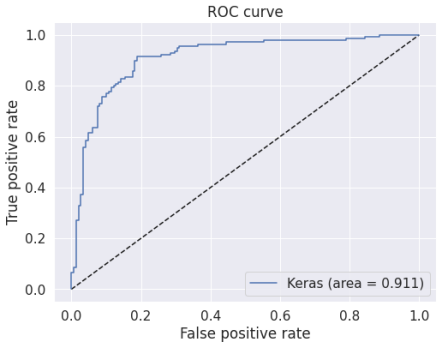




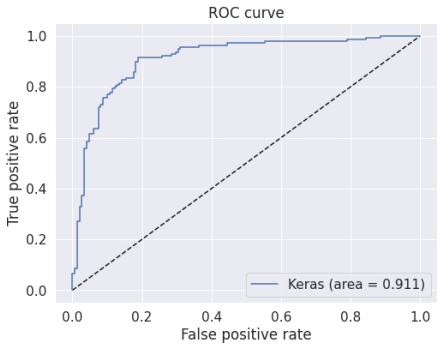
Densenet-169

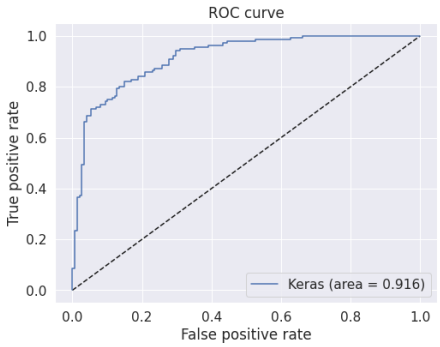


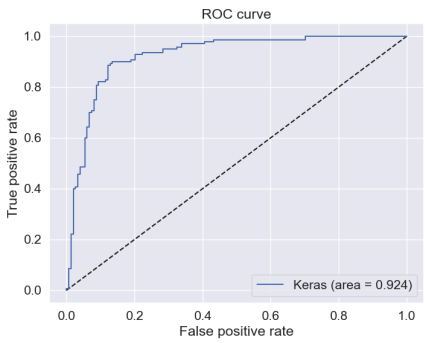




Densenet-201







All of the curves above demonstrate a correlation that is much greater than pure randomness, which is shown by the black dotted line cutting through each graph. The ROC Curves make comparison between models simple, as a blue line that is farther from the dotted black line shows greater accuracy, while one closer to the line shows worse-decision making. As shown by the

curves, each of these models has a similar degree of accuracy, with Densenet-201 having a slightly better accuracy overall, as it is closer to perfect accuracy than the other curves.

AUROC Values

A way to quantify the ROC Curve is to integrate the curve, generating a value called the AUROC (Area Under the Receiver Operator Characteristic Curve). The value of AUROC corresponds to the model's performance at distinguishing between the positive and negative classes. The better the model is at classifying binary data, the higher the AUROC value. An AUROC value of one indicates a perfect classifier, while a value of zero represents a model that only generates false positives or false negatives. AUROC was generated using the following equation:

$$TPR(T) : T \rightarrow y(x)$$

$$FPR(T) : T \rightarrow x$$

$$A = \int_{x=0}^{1} TPR\left(FPR^{-1}(x)\right) \mathrm{d}x = \int_{\infty}^{-\infty} TPR(T)\, FPR'(T)\, \mathrm{d}T$$

Table 4

Model	Average AUROC Value
Resnet-101	0.907
Densenet-169	0.910
Densenet-201	0.917
Radiologists	0.929

Similarly to the average accuracies, the AUROC values of each network are similar, but show a slight trend with higher-layer neural networks having slightly higher accuracies. Also correlated to the average accuracies, the average AUROC value of Stanford's radiologists is larger than that of the ML models.

McNemar's Test

Thomas G. Dietterich's MIT press article, *Approximate Statistical Tests for Comparing Supervised Classification Learning Algorithms*, was used to determine that McNemar's Test is optimal for the analysis of data found in this study. According to Dietterich, McNemar's test is best until "there are situations in which the learning algorithms are efficient enough to run ten times" (Dietterich, 1998). McNemar's Test is a nonparametric test that is used with repeated measures in two related groups, similar to dependent-sample t-test, except it uses dichotomous outcome variables. The test measures the changes in discordancy, or disagreements, in each case. One independent variable with two-related groups within subjects design is related to a dependent dichotomous variable that has a nominal level of measurement. Effectively, McNemar's test compares the sum of False Negatives (FN) and False Positives (FP) between algorithms. McNemar's test also assumes that each image in the dataset is randomly selected, independent from the rest in the dataset, the number of discordant pairs is greater than or equal to thirty, and there is no assumption of normality.

The Null Hypothesis (H_0) for the McNemar test is that the number of my discordant cells remains the same in both studies. The p-values were calculated using McNemar's test with the continuity correction given the data. A p-value less than 0.05 would be considered statistically significant, proving that the model did not alter the number of discordant cells.

Table 5

Models	Two-Tailed P-Value
Resnet-101 and Densenet-169	0.0616438356
Densenet-169 and Densenet-201	0.335616438
Densenet-201 and Resnet-101	0.684931507

The image in each dataset, by definition, was randomly chosen from the set, and each image is independent of one another, as they are from different cases. In each case, the number of discordant pairs is greater than 30, and no normal distribution was assumed.

Based on the results of **Table 5**, as the p-value is greater than 0.05 in all cases, and therefore, H_0 can be rejected. Under the null hypothesis, each algorithm should have a similar or the same disagreement, but the McNemar test has proved that each algorithm has a significantly different performance on the same training. This proves that there is a significant difference in the disagreement of each classifier.

Discussion

In any hospital, implementing Resnet-101, Densenet-169, or Densenet-201 could be advantageous, as all of these models are proven to have a high accuracy, and based on Cohen's Kappa, they have consistently and substantially accurate diagnoses. Densenet-201 was the most accurate model out of these three, and should be considered as the technically best option of the group to be used as a radiologist's assistant. However, solely having the highest accuracy does not make a CNN the best option for a hospital to use. Considering the limited hardware available for hospitals and users, for some locations, it may be beneficial to implement Resnet-101 or Densenet-169.

While accuracy and Cohen's Kappa are important measures of the effectiveness of an algorithm, another important aspect to consider is the training time of each algorithm. By having shorter training times, processes take less time, allowing algorithms to test more hyperparameters. Furthermore, shorter training times might simplify the implementation of improved training methods. For example, a shorter training time would allow for the implementation of 'human in the loop' annotations, which is when the training of a network is supervised by a human expert. This expert may intervene and correct the network at critical steps, allowing for a model, with human support. In these instances, the human expert may check incorrect classifications of images, and therefore reduce high loss error. In order to best train a network on the MURA dataset, implementing a human-in-the loop approach would help to correct label noise and overall improve an algorithm's effectiveness. Neural Networks such as Resnet-101 with fewer layers require experience shorter training times, and could overall benefit hospitals in the training process.

In addition to having shorter training times, lower intensity algorithms with less layers have lower hardware requirements. This is especially important in humerus radiographs because it allows for superior image resolution. Because information can be lost due to downscaling and algorithms need to find small patterns in many pixels, having a higher image resolution can be crucial information. A standard resolution of 2,048 × 2,048 can be improved to to 4,280 × 4,280 px, allowing for better results.

Despite the benefits of lower layer networks, deeper ones remain the most powerful. When hospitals have access to powerful computers and machines, Densenet-201 is the best option for them. If a healthcare network does not have the ability to host a deep learning application program on their machines, then Densenet-169 or Resnet-101 may be a strong alternative.

My findings differ from applied techniques used in previous literature, where deeper network architectures, mainly a Densenet-169, were used to classify other data from the MURA dataset. Densenet-169 is rarely compared to other networks, and knowing the most powerful model is

important for the development of future programs. The authors of the MURA dataset achieved an average overall AUROC of 0.929, using a 169 layer Densenet, which was not surpassed by any of the models used in my analysis, despite differences between the best performing networks and the MURA baseline were smaller than 0.02. With greater optimization, research, resources, experience, and time, I feel as though my models could improve, and using Densenet-201, I could definitely build stronger models than those built with Densenet-169. As a high school student, I also understand my models may not reach the high standards of sophistication held by modern day deep learning research.

A common problem in deep learning research is the development of indeliberate human biases based on the training data provided, however, steps were taken to prevent this bias (Brady & Neri, 2020). For example, when gathering data for their MURA dataset Stanford focused on acquiring a set of HIPAA-compliant images that included people of a variety of ethnicities and sexes. By recognizing the issue at hand and designing a minimally biased dataset and model, this bias can be reduced, or even eliminated.

Limitations

A limitation of this study is the general lack of data, the entirety of the data came from the MURA dataset. As a consequence, this data is overgeneralized, and cannot be applied to all datasets or algorithms, which should be considered when interpreting my results. Lack of standardized data remains a common problem in deep learning research, especially in medical fields (Kim et al., 2019).

Another substantial limitation is the quality of computing power. With a more powerful computer, I would have been able to test a greater range of deeper neural networks, and test if those accuracies were significantly different than Densenet-201. Because hospitals are also limited in their computing power, this limitation should not be highly considered, because even if a more powerful CNN would present a statistically significant change in accuracy, ROC, or Kappa score, a hospital may not be able to host such a model.

Conclusion

In this paper, it has been shown that there is a statistically significant difference between the difference in disagreement of Densenet-161, Densenet-201, and Resnet-101. It also demonstrated a clear trend in the values for accuracy, AUROC, and Cohen's Kappa score, given multiple trials of testing on the same dataset. Densenet-201 consistently had the best performance, Resnet-101 consistently performed the worst and Densenet-161 consistently performed between the two. As compared with experienced radiologists's analyses, the depth and strength of each of these neural networks has been proven to be not yet comparable to human doctors, however, neural networks

remain a valuable assistant to healthcare providers regardless, as they provide high accuracy results. Previous studies have found varying degrees of success using models with different layers or structures, even achieving similar accuracies using networks consisting of fewer layers (e.g. Resnet-101). This study has shown that these differing models still have many differences. With the benefits that come with lower-level models, such as Resnet-101 or Densenet-169, hospitals and computer scientists should consider the marginal benefits and costs of implementing many-layered CNNs into radiology software. It should be noted that more powerful networks will remain the best at classifying information, even if they are not the most easily accessible.

Bibliography

Akkus Z, Kostandy P, Philbrick KA, Erickson BJ. Extraction of brain tissue from CT head images using fully convolutional neural networks : Proceedings of Medical Imaging 2018: Image Processing. Houston, USA; 2018.

Anderson RT, Feldman BA. What Patient's Want: A Content Analysis of Key Qualities That Influence Patient Satisfaction. Journal of Medical Practice Management. 2007;In Press.

Bindman RS, Miglioretti DL, Larson EB. Rising use of diagnostic medical imaging in a large integrated health system. Health Aff. 2008;27(6):1491–1502. doi:10.1377/hlthaff.27.6.1491

Brady, A. P., & Neri, E. (2020). Artificial Intelligence in radiology—ethical considerations. Diagnostics, 10(4), 231. https://doi.org/10.3390/diagnostics10040231

Chauhan, T., Palivela, H., & Tiwari, S. (2021). Optimization and fine-tuning of DenseNet model for classification of COVID-19 cases in medical imaging. International Journal of Information Management Data Insights, 1(2), 100020. https://doi.org/10.1016/j.jjimei.2021.100020

Cheng, P. M., Montagnon, E., Yamashita, R., Pan, I., Cadrin-Chênevert, A., Perdigón Romero, F., Chartrand, G., Kadoury, S., & Tang, A. (2021). Deep learning: An update for radiologists. RadioGraphics, 41(5), 1427–1445. https://doi.org/10.1148/rg.2021200210

Dietterich, T. G. (1998). Approximate statistical tests for comparing supervised classification learning algorithms. Neural Computation, 10(7), 1895–1923. https://doi.org/10.1162/089976698300017197

GE Healthcare. 510(k) Summary K183182. 2019. https://www.accessdata.fda.gov/cdrh_docs/pdf18/K183182.pdf

Georgievskiy I, Georgievskaya Z, Pinney W. Using Computer Simulation Modeling to Reduce Waiting Times in Emergency. 2008 [cited 2012 Oct 22]; Available from: http://www.business.alcorn.edu/Stuff/PUBS/CPWPN-2008-01.pdf

Hawnaur J. Recent advances: diagnostic radiology. BMJ 1999; 319: 168–71. doi: 10.1136/bmj.319.7203.168

Hinton G. Deep learning—a technology with the potential to transform health care. JAMA. 2018;320(11):1101–1102. doi:10.1001/jama.2018.11100

Huang XM. Patient attitude towards waiting in an outpatient clinic and its applications. Health Serv Manag Res Off J Assoc Univ Programs Health Adm AUPHA. 1994;7(1):2.

Hutton D, Beardmore C, Patel I, Massey J, Wong H, Probst H. Audit of the job satisfaction levels of the UK radiography and physics workforce in UK radiotherapy centres 2012. Br J Radiol 2014; 87: 20130742. doi: 10.1259/bjr.20130742

Kim, D. W., Jang, H. Y., Kim, K. W., Shin, Y., & Park, S. H. (2019). Design characteristics of studies reporting the performance of Artificial Intelligence algorithms for diagnostic analysis of Medical Images: Results from recently published papers. Korean Journal of Radiology, 20(3), 405. https://doi.org/10.3348/kjr.2019.0025

Lakhani P, Sundaram B. Deep learning at chest radiography: automated classification of pulmonary tuberculosis by using convolutional neural networks. Radiology. 2017;284(2):574–582. doi:10.1148/radiol.2017162326

Lecun Y, Bengio Y, Hinton G. Deep learning. Nature. 2015;521:436–444. doi:10.1038/nature14539

Lindsey R, Daluiski A, Chopra S, et al. Deep neural network improves fracture detection by clinicians. Proc Natl Acad Sci U S A. 2018;115(45):11591–11596. doi:10.1073/pnas.1806905115

Lohikoski K, Roos M, Suominen T;in press Workplace culture assessed by radiographers in Finland. Radiography 2019; 25: e113–8. doi: 10.1016/j.radi.2019.05.003

McCarthy, J. (2007, November 12). Extinguished philosophies lie about the cradle of every science as the ... WHAT IS ARTIFICIAL INTELLIGENCE? Retrieved April 14, 2023, from https://www-formal.stanford.edu/jmc/whatisai.pdf

Montagnon E, Cerny M, Cadrin-Chênevert A, et al. Deep learning workflow in radiology: a primer. Insights Imaging 2020;11(1):22.

Nie D, Cao X, Gao Y, Wang L, Shen D. Estimating CT image from MRI data using 3D fully Convolutional networks. Deep Learn Data Label Med Appl 2016; 2016: 170–8.

Nuti, S., & Vainieri, M. (2012). Managing waiting times in diagnostic medical imaging. BMJ Open, 2(6). https://doi.org/10.1136/bmjopen-2012-001255

Rajpurkar P, Irvin J, Zhu K, et al. CheXNet: radiologist-level pneumonia detection on chest x-rays with deep learning. 2017;arXiv:1711.05225.

Rajpurkar, P., Irvin, J., Bagul, A., Ding, D., Duan, T., Mehta, H., ... & Ng, A. Y. (2017). Mura: Large dataset for abnormality detection in musculoskeletal radiographs. arXiv preprint arXiv:1712.06957.

Salman S, Liu X. Overfitting Mechanism and Avoidance in Deep Neural Networks. CoRR. 2019;abs/1901.06566. http://arxiv.org/abs/1901.06566. Published January 19, 2019.

Sheth S. The working radiological technologist: on the path for burnout? 2010 HealtheCareers https://www.healthecareers.com/article/career/the-working-radiological-technologist-on-the-path-for-burnout.

Singh N, Wright C, Knight K, Baird M, Akroyd D, Adams RD, et al.. Occupational burnout among radiation therapists in Australia: findings from a mixed methods study. Radiography 2017; 23: 216–21. doi: 10.1016/j.radi.2017.03.016

Soriano A. Comparison of two scheduling systems. Oper Res. 1966;14(3):388–97.

van Lent WAM, Deetman JW, Teertstra HJ, Muller SH, Hans EW, van Harten WH. Reducing the throughput time of the diagnostic track involving CT scanning with computer simulation. Eur J Radiol 2012; 81: 3131–40. doi: 10.1016/j.ejrad.2012.03.012

Wang S, He K, Nie D, Zhou S, Gao Y, Shen D. Ct male pelvic organ segmentation using fully convolutional networks with boundary sensitive representation. Med Image Anal 2019; 54: 168–78. doi: 10.1016/j.media.2019.03.003

What is machine learning? IBM. (n.d.). Retrieved April 14, 2023, from https://www.ibm.com/topics/machine-learning

Willemink MJ, Koszek WA, Hardell C, et al. Preparing Medical Imaging Data for Machine Learning. Radiology 2020;295(1):4–15

Woolen, S., Kazerooni, E. A., Wall, A., Parent, K., Cahalan, S., Alameddine, M., & Davenport, M. S. (2018). Waiting for radiology test results: Patient expectations and emotional disutility. Journal of the American College of Radiology, 15(2), 274–281. https://doi.org/10.1016/j.jacr.2017.09.017

Yoon Y, Jeon H-G, Yoo D, Lee J-Y, Kweon IS. Learning a Deep Convolutional Network for Light-Field Image Super-Resolution : IEEE International Conference on Computer Vision Workshop (ICCVW. Santiago, Chile; 2015.

Zebra Medical Vision. 510(K) Summary—HealthPNX. 2019. https://www.accessdata.fda.gov/cdrh_docs/pdf19/K190362.pdf

Zhang C., Benz P., Argaw D.M., Lee S., Kim J., Rameau F., et al. Proceedings of the IEEE/CVF winter conference on applications of computer vision. 2021. Resnet or densenet? Introducing dense shortcuts to ResNet; pp. 3550–3559. [Google Scholar]