

AI is a code control version of intelligence that is trained to mimic behaviour of a human mind and do tasks on par or better than a human can. Mahmood (no date: 8-9) states that there are 3 main types of widely used AI. Broadly, AI can be split into three categories, machine learning, deep learning and generative AI. These work in different ways and can benefit the healthcare sector in different ways. Machine learning is AI that works on simple algorithm. It analyses records and predicts risks in a way that is simple to understand and derive. Deep learning is more complex. It is specialised machine learning that focuses on a specific task and it learns from a set of input data. This is good as it allows it to make more accurate predictions and give results that may be closer to how a professional worker might do it. Generative AI is the most advanced version of AI. It can generate more responses and work like a human worker, doing tasks that many be task consuming or difficult.

This essay examines AI systems that are used by Ardila *et al* (2019: 954) that use deep learning to diagnose lung nodules, with a particular focus on 3D convolutional neural networks (3D CNNs). This tool helps to spot lung cancer in screenings of low radiation dose chest computed tomography . AI can certainly help to improve healthcare in this area; however it has its limitations and these improvements and limitations will be the focus of this essay.

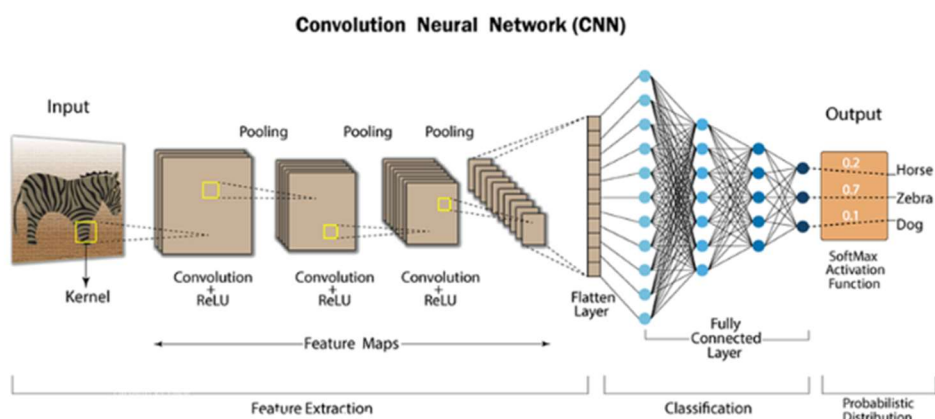


Fig. 1: Diagram depicting how a Convolution Neural Network works. Depicted by Kumar, V. (2025)

An imaging / screening AI analyses images by using a system called “computer vision” to spot deviations in a medical image of a patients lung. The type of computer vision three-dimensional deep learning uses is CNN. This means that it uses 5 main layers to “see” an image, as shown in Fig. 1. Mahmood (no date:23-24) suggests that the first layer is the simplest one, where colours turn into pixels of different values. This value ranges from around 0 to 255 where the higher

the number, the brighter the colour. Then, it passes a filter through the pixel grid to make out edges and curves, making a new grid called a feature map. Afterwards, the AI passes the feature map through an activation function, usually a ReLU. This filter changes negative values into a zero and keeps positive values the same. Then the AI shrinks the feature map so that the most important information is kept and picks out the biggest numbers to pool together an reduce the map even further. Lastly, the AI repeats the steps of the CNN and combines all the relevant information that it needs to understand, leaving a more complex pattern that can help now distinguish cancer from non-cancerous growths. All this information gets compiled into one solid image, in the case for the three-dimensional deep learning model, this image is a 3-dimensional image of a lung. This is done by stacking the vertical images gathered by the CT scan into one model that shows the full intricacies of the lungs.

The AI then runs this image through its database that it was trained on to spot any cancerous growths. The data base that it is trained on is full of images of scans from both healthy and unhealthy patients. The AI runs the most recent scan that it has been given through the test data to see if it matches any image that either is healthy or diseased where than it tells the doctors observing the patients the result via a human computer interface like a monitoring application on a computer. After looking through the data base and seeing if the input matches any result, it will run another set through a different data base that contains treatment plans. It will look at the diagnosis that it has gotten from both its image testing and the input of the doctors using it and

will give a treatment plan that follows how the disease reacts to medicine and to other stimuli and give the doctor a treatment pathway that can be changed for each patient with a similar disease or condition. Another way that the AI helps plan a treatment plan and give diagnosis is by running a risk score through the image it has been given. Every image that has been given to the AI in training has a pre-assigned risk score, either low, medium or high. Then the AI runs the given input image and cross references the image with images in its data-base, if the image matches more images of the high-risk scores than it will be labelled high risk and this will affect both the treatment plan and the diagnosis. If it is ranked high risk, then it will also be sent in for a specialist doctor to check and verify the result. This means that both a trained doctor and an AI have run an image through rigorous checks to see if it is cancer or not. This ensures a better result to most test and will therefore help the patient recover and doctor to help treat other patients with the new data acquired. This data is very important. It contains images of healthy and unhealthy lungs. It can also contain images that are not lungs to make sure that the AI system can differentiate between a lung and something that is not a lung.

AI training needs many types of input data. This input data is called a data set. It contains data from various patients with various medical histories and bodily differences. Having good quality data allows for good AI education that therefore it will give better results out. Mahmood, Z. (no date : p 12) states that AI training follows the policy of “good data to good results” and “garbage in and garbage out”. This is a simple explanation of how AI training works, if good data is given in, good results will follow but if poor data is given in, poor results will be given out, possibly harming someone. Ardila *et al* (2019: 955) base their training on 42290 CT scans from 14581 patients of which 578 had developed biopsy confirmed cancer after a year of initial test. This data came from a publicly available dataset from the National Institutes of Health. Ardila *et al* (2019: 955) separated this data into 3 components the “training set”, the “tuning set” and lastly the “test set”. The testing set was used first to help make the AI recognise what cancers look like and how they differentiate from common scans of healthy patients. In the test set that Ardila *et al* (2019: 956) used, there were 6716 cases of which only 87 were cancer positive . After having the AI trained, it went on to use the “test set” to make sure it gets the best results. Over a series of tests that Ardila *et al* (2019: 955) conducted that contained over 507 patients, the AI managed to get a score of 95.9%. overall. The data used to train the AI was good quality data as it consisted of many cases. However, there weren’t as many cases with cancer as there were without cancer. This can be seen as a good thing as it allowed the AI to have a higher chance to recognise non-cancerous results but it does limit how likely it is to see cancerous results.

The AI I have chosen is a deep learning algorithm that works in three dimensional spaces. It uses a convolution neural network to visualise images. There are many metrics to test how accurate and functional an AI is. @tanisha.digital (2024) depicts that there are mainly 8 metrics for a classification type model like the one used by Ardila *et al* (2019: 958). @tanisha.digital (2024: paragraph 3-10) states that they are:

1. Confusion matrix. This is a matrix that classifies the data. It ranges from true positive, false positive, true negative and false negative. True positive means that the result is true in the case that is applied. False positive is where a negative piece of information is presented as if it were the truth. True negative depicts the information as fully false and false negative presents a positive result as a negative one.
2. Accuracy and precision. This measures how correct the product is. It can be classed into accuracy and precision. Accuracy means that the result is further or closer to a desired result and precision shows how many attempts are at that accuracy.
3. Recall. This is the measure of how well the AI can capture all positive cases.
4. F1 score. This is the mean of the precision and recall. It balances the two and shows either the true positive or true negative results as it blocks out the false values of both of them
5. Threshold. This allows to measure and adjust the sensitivity of the model. It helps to balance the false positives and false negatives and ensures less false negatives.
6. AUC-ROC. This stands for Area Under the Curve – Receiver Operating Characteristic and it distinguishes between all possible thresholds.
7. Log loss. This evaluates the uncertainty of a models prediction by penalizing incorrect classifications based on the confidence of the prediction.

In tests performed, on a dataset of 507 patients, Ardila *et al* (2019: 955) uses mostly confusion matrix, accuracy and precision, threshold and AUC-ROC, noting that it had a 95% confidence interval. This means that it was 95% accurate to if the test patient did or did not have lung cancer. Ardila *et al* (2019: 956) found that it had a AUC of around 92.6% which lead it to yet again have a 95% confidence on results. When Ardila *et al* (2019: 954) pitted the AI system against 6 radiologists, it managed to outperform them, having a 11% decrease in false positives and a 5% decrease in false negatives, showing that it was more successful than the trained radiologists selected for the test, all of whom had over 5 years of experience.

There were limitations too that Ardila *et al* (2019: 959) mentioned. They stated that they were limited due to only using two data-sets. This is limiting for an AIs growth as it means that there is a higher chance of the databases being unfairly biased towards a certain demographic, making the AI struggle more against things that may be more exotic to its given training data. Another limitation mentioned by Ardila *et al* (2019: 958) is the black box nature of their deep learning AI. It was difficult to predict how the AI was reaching its conclusions and therefore made the results it gave have more uncertainty attached to them even if they were correct. This black box nature can also drive patients away from using AI and therefore can possibly lead to higher stress on the workers in the hospital scene. Another limitation Ardila *et al* (2019: 958) mentioned was the AIs tendency to use some data that was subject cluttered. This means that the data was too full of issues and some images were layered on top of each other, leading to faulty information that may lead to issues like incorrect observations and therefore giving incorrect diagnoses for patients, leading to distrust from the radiologists who may have been using it. Ardila *et al* (2019: 958-959) also mention that the data they were using had only subtle cancer cases due to the easier to spot ones being dropped from the data they used as it was already diagnosed. This could lead the deep learning AI to miss easy to spot cancer that may have been quickly noticed by radiologists and therefore possibly delaying a patients correct prognosis to a later date which may build further distrust towards the AI.

AI can be used to improve the health sector in many ways if it is under the supervision of practiced professionals. There are still many issues that using AI can bring, however many of them can be solved with little input from the trainers and users of the AI. Some possible ways to fix the issues that Ardila *et al* (2019: 957-9) faced like using more than 2 data-sets. This would increase the range of trainable information the AI has to its disposal and therefore giving it a higher chance to recognise things that it may not have been able to recognise correctly given only two data sets. Another way to solve the issues that were faced is to possibly have more radiologists peer review the data before it was used in testing, reducing the chance of the cluttering effect mentioned previously. The black box nature is harder to fix however it may be possible to program the deep learning algorithm to give more detailed steps to how it gave out its results so that they are easier to understand for the radiologists who are using it. A way to solve the issue of the poor data set quality is to add in the results that were already diagnosed into the data-set used for training as that would increase the total data-set the AI can use, letting it get more accurate results.

In summary, AI can help to improve the health sector by taking over the more mundane jobs in the hospital that may be taking up a doctors time. It can also help to improve the detection rates of cancer like the deep learning AI that Ardila *et al* (2019: 254-261) used and trained on lung cancer patients. AI in the health sector would be best suited for use in aiding radiologists and doctors to test patients and interpret medical scans. The most optimal conditions for the use of AI in a clinical setting would be if it was under scrutiny and observation of the doctors who would be use of it at all times so that someone is there to correct possible, life-threatening issues that AI might accidentally cause due to data issues from improper data.

Bibliography

Ardila, D., et al. (2019) 'End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography', *Nature Medicine*, 25, p. 954–961.

Mahmood, Z. (no date) 'AI lifelines: technology that saves lives. Key stage 5 programme', *The Scholars Programme*, p.8-42.

@tanisha.digital (2024) 'Key evaluation metrics for AI model performance' Available at:

<https://medium.com/gen-ai-adventures/key-evaluation-metrics-for-ai-model-performance-8e372f17a0a2>

(Accessed : 15 March 2026)

Bibliography - Images

Kumar, V. (2025) 'Understanding the Core Process of a Convolutional Neural Network (CNN) for Image Classification. Available at : https://medium.com/@vishnukumardsvishnu_99381/understanding-the-core-process-of-a-convolutional-neural-network-cnn-for-image-classification-934ecab71045 (Accessed: 15 March 2026).