JOURNAL OF
EMERGING INVESTIGATORS

Article

How artificial intelligence deep learning models can be used to accurately determine lung cancers

Siri Cherukuri[1], Munib Mesinovic[2]
[1]The Hockaday School, Dallas, Texas
[2] University of Oxford, Oxford, United Kingdom

SUMMARY

Disease is a highly stressed topic in current society, especially the importance of early diagnosis and disease prevention with the help of new technology. Doctors can diagnose diseases like lung cancers, but the process can be more time-consuming, expensive, and inaccurate compared to relying on technology alone. Although previous applications of artificial intelligence exist in the clinical world, our goal was to find which deep learning model would be most accurate in predicting lung cancer. We hypothesized that autoencoders would perform better than other deep learning models such as convolutional neural networks (CNN) and pre-trained models based on the strength and reliability of the data created by the encoder for classification. We studied Kaggle dataset "Chest CT-Scan Images Dataset," which contains images of both benign and malignant tumors, with malignant categories of lung squamous cell carcinoma, large cell carcinoma, and adenocarcinoma. We trained and tested each model with this dataset on the machine learning tool Google Colaboratory. Based on the results, the pre-trained CNN model had the accuracy at correctly detecting lung cancers, while the autoencoder had the lowest compared to the others. Overall, our results do not support our original hypothesis and instead show that pre-trained models may be more applicable in lung cancer diagnosis.

INTRODUCTION

Disease is an important part of human evolution and society as well as in the medical field. Medical professionals can use effective interventions in health, and possibly improve it, while also considering ethical responsibilities (1). Disease can be slowed and even reversed with the inventions of new medications and through health interventions. With the recent COVID-19 pandemic, disease has created a sense of uncertainty and anxiety in how people perceive life (2). With the development of technology, the past century has been filled with advances that ensure safety and good health for all members of society. Recently, with the invention of artificial intelligence (AI), medicine will continue to see massive development in disease detection, which will transform the future of medicine (3).

In the past, AI has contributed to multiple advancements in medicine, such as medical imaging, drug development, predictive analytics, and precision medicine (4). By learning to use complex algorithms with multiple layers, machine learning can detect images and create predictions (5). Image detection can be applied to analyze images from X-rays, magnetic resonance imaging (MRI) scans, and computer topography (CT) scans to create more accurate diagnoses for patients than radiologists (6). Machine learning-based technologies allow for earlier detection and prevention. In addition to using algorithms to detect disease, AI can also use algorithms to identify potential drug treatments for patients as well as develop new variants of drugs that can revolutionize the medical industry (7). Using patient data, AI can also predict diseases in patients that potentially result in better patient outcomes. For example, AI can be used to make personalized treatment plans based on past patient history and genetic makeup, allowing for more precise and effective treatments (8). Lastly, AI is more accessible than a doctor as online chatbots and virtual assistants create opportunities for those in remote areas to self-diagnose and have access to healthcare services.

Deep learning employs algorithms inspired by the human brain with multi-layer neural networks being a prominent technique (9). Earlier approaches like Random Forests did not use these algorithms and, thus, were less effective for image detection due to their focus on tabular data, limited accuracy, and slow predictions (10). In contrast, deep learning includes advanced models like convolutional neural networks (CNNs) and autoencoders for image recognition. CNNs use convolutional layers, which contain filters to identify features like edges and shapes in images and create a map using this data. CNNs excel in image classification, object detection, and segmentation (11). Autoencoders, on the other hand, utilize an encoder-decoder architecture to reduce noise and capture essential features. Autoencoders, including convolutional autoencoders, learn from data by encoding it into a lower-dimensional form and then decoding it. Limitations of autoencoders include a potentially higher loss function due to their dual encoding-decoding process, making them more suitable for cases with limited computational resources (12). Finally, both CNNs and autoencoders contribute significantly to modern image recognition and generative modeling. Some examples of image recognition include classifying an image as a "cat" or "dog," or being able to detect pedestrians and other cars in autonomous driving applications. Generative modeling entails removing noise from an image, such as restoring clarity to pixelated or corrupted photos, or compressing and decompressing images for efficient memory storage. CNNs, unlike traditional neural networks, use convolutional layers that apply filters, known as kernels, to the input image to extract features, such as edges and shapes, to classify objects in the input image. On the other hand, autoencoders

https://doi.org/10.59720/23-232

are a type of neural network that works with an encoder and decoder, allowing the model to be trained to learn a lower-dimensional space image and capture the most important features in the input image. CNNs can be used as the encoder process in an autoencoder, a model commonly known as a convolutional autoencoder (13). CNN-based encoders, like the original CNN model, work well at detecting features and generative modelling, making it a useful tool for the encoding process of the autoencoder.

InceptionV3 is a pre-trained CNN designed by Google for image detection (14). A pre-trained CNN, unlike a regular CNN, is already trained on large datasets, allowing it to have already learned detection of relevant features in images. The original Inception architecture was designed to improve the efficiency of CNNs to lower the number of parameters. The model first functions with a convolutional layer by extracting features from the input using filters known as Inception modules with multiple branches. These modules are designed to capture features at differing levels of abstraction, and with multiple branches, allowing the model to create a single feature map. The features are concatenated, or linked together, to form this feature map. Similar to a non-pretrained CNN model, InceptionV3 also utilizes pooling layers, which reduces the dimensionality of the feature map and results in reduced parameters and overfitting prevention. The main feature that makes InceptionV3 different from CNN are its pre-trained weights, which are determined based on a wide range of images. Because of its range of knowledge and easier ability to adapt to newer data, this model performs well on smaller, specific datasets with minimal training data, making it a popular model for imaging (**Figure 1**) (15).

Lastly, autoencoder can learn features at a higher level and learn through clustering. Autoencoders are made up of three main parts: an encoder, bottleneck, and decoder. The encoder takes the input image and uses a deep learning neural network with neurons to learn from the image and convert it into a lower-dimension representation. In this study, our autoencoder converted a 255 x 255-pixel image to both 160 x 160-pixel images and 40 x 40-pixel images using the encoder. Once converted, the new lower-dimensional image is classified as latent space representation, or a learned space representation. This section is known as a bottleneck as a larger image is transforming to a smaller size. After the bottleneck portion, the process reverses and goes through the decoder; the lower-dimension representation reconstructs the original image using latent, or learned, factors, creating the reconstructed image as a final product. The encoding process allows the model to focus on significant and essential features only, making malignant scans easier to identify through abnormal patterns in feature detection. The autoencoder can be trained from end-to-end or layer-by-layer, which is where they are stacked together to create a deeper encoder (**Figure 2**) (16). A deeper encoder can act as a pre-training step, allowing the learned features from encoding to enhance each CNN's convolutional layer's ability to detect objects in images.

Our goal was to determine if autoencoders can be used to make more accurate predictions of lung cancer from CT scans compared to other more common machine learning models and how successfully each model performs at detecting cancer v. no-cancer scans. We hypothesized that autoencoders would perform the best due to their higher

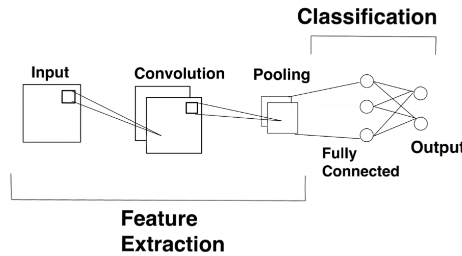

Figure 1: Convolutional neural network (CNN) model architecture. The model consists of five layers, each falling into two sections: input, convolutional layers, which apply filters to the input data to extract features such as edges and textures, and pooling layers, which reduce the dimensions of the feature maps by summarizing the information and retaining the most important features. These layers fall under the feature extraction section, as well as the fully connected layer and the output, which fall under the classification section.

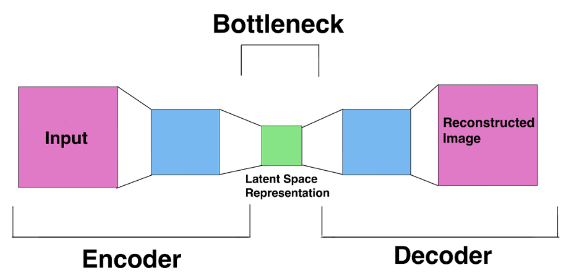

Figure 2: Autoencoder model architecture. The model consists of three sections: the encoder, bottleneck, and decoder. Together, these deconstruct the input image into a latent space representation, which is a compressed encoding of the input that captures its most essential features, and then with the decoder reconstructs the representation into a new output image.

ability to detect features compared to other deep learning models. We hypothesized that an autoencoder offers better prediction accuracy for diagnosing lung cancer through CT scans than CNNs because autoencoders can detect more abstract and varying features compared to a CNN, which can generally only detect low-level features such as edges and textures (17). To compare the performance of each model, we used the area under the curve value (AUC), which represents the performance between classes in the model. In other words, the AUC determines the model's ability to correctly determine if a case is either a positive or negative case of cancer (**Figure 3**) (18). Our results disproved our hypothesis, with the CNN (AUC ~76%) and pre-trained CNN (InceptionV3, AUC ~87%) having a higher AUC than the autoencoder (AUC ~74%). Based on previous research, many CNN models have an average AUC) of 95%, meaning there is still room for improvement in these models. Our study highlights the potential of utilizing autoencoders and CNNs for lung cancer detection from CT scans, emphasizing the need for further research to improve both model architectures and training processes. By advancing these technologies, this work can aid in enhancing early cancer detection and diagnosis, leading to better patient outcomes and more effective healthcare solutions.

https://doi.org/10.59720/23-232

RESULTS

To test our hypothesis of whether autoencoders will outperform other deep learning models when detecting lung cancer from CT scans, we tested four models and collected the following metrics metrics: AUC, accuracy, sensitivity, and specificity. AUC measures the model's ability to distinguish between positive (cancerous) and negative (benign) cases, with higher values indicating better performance (**Figure 4**). AUC is derived by plotting the receiver-operating characteristic (ROC) curve, which is drawn by calculating the true positive rate (malignant cases which are correctly identified by the model) and false positive rate (benign cases which are correctly identified) at every possible threshold. The ROC curve is a visual representation of model representation across all thresholds and is used to assess the overall diagnostic performance of a test or model (19). Accuracy calculates the proportion of correct predictions by comparing the model's predicted values to the true values. To calculate accuracy, we found the proportion of the number of true positives and true negatives compared to the entire sample size. Sensitivity shows how well the model identifies positive cases, representing the proportion of true positives detected out of all actual positives. Specificity measures how well the model identifies negative cases, indicating the proportion of true negatives detected out of all actual negatives. These metrics provide a comprehensive view of the model's effectiveness in detecting lung cancer from CT scans.

We began by testing the CNN, and when we found that the AUC was somewhat low to our expectation, we included regularization as well to test the model fairly. Regularization is a tool in machine learning used to prevent overfitting, which occurs when a model performs well on training data but not as well on new, unseen data. This entails adding a penalty during the training process, which discourages the model from fitting too closely to the training data and allowing the model to be more generalizable. We then tested the autoencoder, which used parts of our CNN code in its model, and then finally pre-trained model InceptionV3.

Based on the results collected from computation, the InceptionV3 had the highest AUC (~87%), while Autoencoder + Regularization had the lowest (~68%). The InceptionV3 model clearly had the best performance in determining the difference between the positive and negative, or cancerous and benign, classes. However, CNN had the highest overall accuracy (~79%), while the autoencoder was the lowest (~76%), meaning it had a lower correct prediction rate than the other models. The sensitivity is lower for all models, showing that they had more trouble detecting true positives. On the other hand, the specificity of the models was distinctively higher, showing that they were more easily able to determine true negatives. Due to the sharp difference between the validation loss (measures the error on unseen validation data) and training loss (measures the error on the training data) in the autoencoder learning ability, which is a sign of overfitting (when the data is too close to the training data and therefore unable to generalize to larger sets of data), we tested the autoencoder with regularization, which did not provide the expected results of improved AUC or accuracy (**Figure 5, Table 1**). Based on the learning ability functions of the CNN and autoencoder, the CNN learns at a better rate as the learning ability of the training loss and the validation loss are much closer than that of the autoencoder, meaning the CNN model has more accuracy (**Figure 6, Table 1**).

DISCUSSION

We expected to find that autoencoders would work best at accurately predicting lung cancer due to the lower-dimensional space extracted from the encoder being an efficient representation of the input as suggested in earlier applications (20). However, we found that the InceptionV3 model has a more accurate prediction rate, which did not support our hypothesis. CNNs also outperformed autoencoders, except with regularization. Across the board,

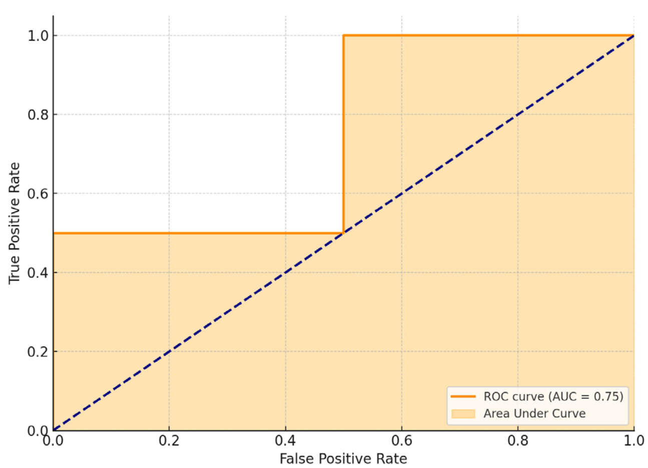

Figure 3: Example of a Receiver-Operating Characteristic (ROC) curve illustrating the Area Under the Curve (AUC) used to evaluate model performance. The AUC represents the model's ability to distinguish between positive (cancerous) and negative (benign) cases. A higher AUC indicates a better-performing model, as it is more capable of correctly identifying true positives and true negatives across various thresholds.

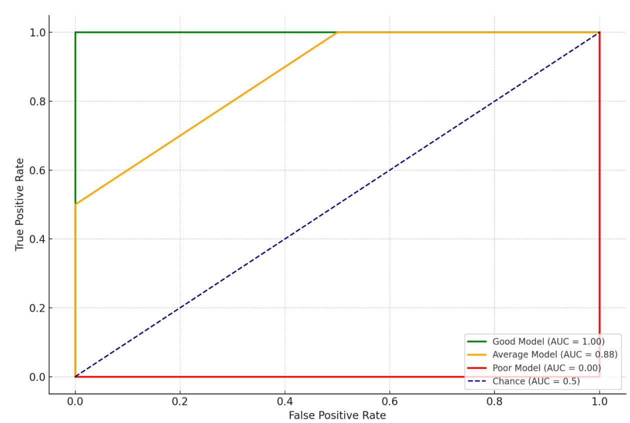

Figure 4: Comparison of ROC curves for multiple AI models using AUC to assess classification performance. The CNN demonstrates superior learning ability, as indicated by a higher AUC and closer alignment between training and validation loss curves. In contrast, the autoencoder shows a lower AUC, suggesting reduced accuracy and greater overfitting. This figure supports the finding that the CNN is more effective at distinguishing between cancerous and benign cases.

https://doi.org/10.59720/23-232

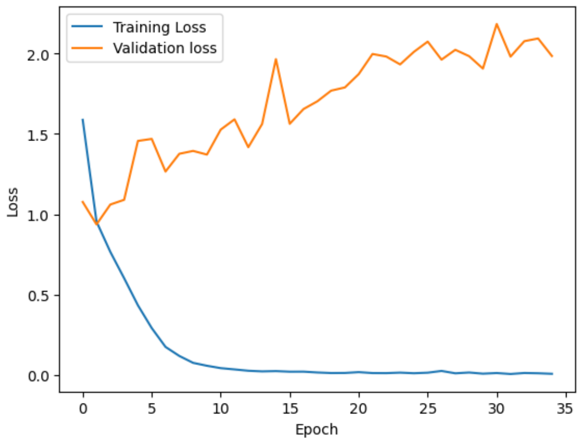

Figure 5: Autoencoder learning ability. Comparison of training loss (blue) and validation loss (orange). Due to the weak correlation between the validation and training loss, the autoencoder demonstrates a weaker learning ability in comparison to a convolutional neural network model.

Model	AUC	Accuracy	Sensitivity	Specificity
CNN	75.99%	79.37%	60.21%	85.21%
CNN + Regularization	68.35%	76.83%	47.69%	81.28%
InceptionV3	86.54%	79.21%	64.03%	85.48%
Autoencoder	74.12%	75.56%	52.25%	82.32%

Table 1: Artificial intelligence (AI) metrics for each tested model. Includes the percentage of area under curve (AUC), accuracy, sensitivity (true positive cases), and specificity (true negative cases).

the pre-trained InceptionV3 model greatly outperformed the other models, even at detecting true positives and true negatives, making it a more accurate model overall.

These results show that a pre-trained model, which is trained using a wide range of images, can result in a better prediction rate for detecting lung cancer. Although the AUC and accuracy can be improved to a higher level with more normalization and dropout layers to achieve an excellent AUC value, the current AUC ranges are still able to achieve an acceptable standing. Because the specificity rate, or true negatives detection rate, is higher in all models, they can determine benign cases at a higher accuracy rate. However, the sensitivity rate is significantly lower, which would need to be increased to have an impact on accurate diagnoses to determine true cases of malignancy, or true positives.

Limitations in data collection, computational costs, and hardware are a disadvantage to the AI approach as the required GPU and memory to conduct these state-of-the-art-level coding processes are not accessible, making it hard to collect accurate and useful data. Computing time is another important limitation, as each epoch—which refers to one complete pass through the entire training dataset—takes nearly 4-5 minutes, meaning that with more layers, the computing time would increase drastically. Lastly, there is an issue of intractability: there may be no efficient enough algorithm, making it difficult to accurately collect results. Autoencoders generally work better when restricted in computational resources. However, because there are two processes instead of one like in a CNN, there is a higher loss function, which is the sum of the errors from

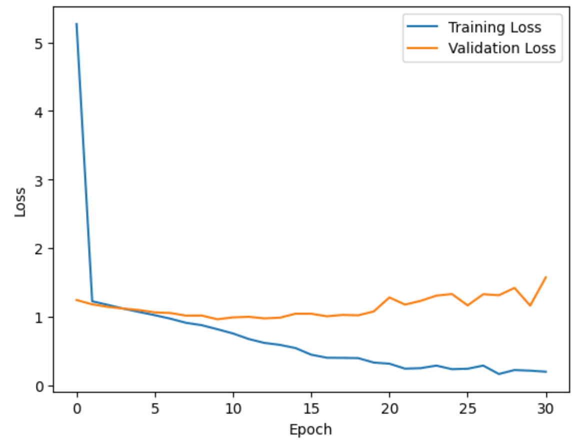

Figure 6: Convolutional neural network (CNN) learning ability. Comparison of training loss (blue) and validation loss (orange) for the CNNs model.

training and validation, which is not beneficial to the model's accuracy. While there are significant limitations in AI for cancer detection and other medical applications, researchers are actively developing new methods and technologies to overcome these barriers. Advances in increasing the training dataset's size and autoencoder algorithms which are more efficient are all contributing to the improvement of deep learning models.

MATERIALS AND METHODS

The dataset used for data collection is the Kaggle "Chest CT-Scan Images Dataset" by Mohamed Hany (21). This dataset includes different classifications of lung cancer, including those of malignant cancers (adenocarcinoma, large cell carcinoma, and squamous cell carcinoma) as well as benign cancer scans. Within the dataset, there were 120 adenocarcinoma, 51 large cell carcinoma, 90 squamous cell carcinoma, and 54 benign scans. To create the models, Pytorch, a machine learning library, and Google Colaboratory, a machine learning tool, were the main resources used. We used Pytorch version 2.0 to run our models.

The dataset was imported and pre-processed using a combination of pandas and other Python libraries. This process included flipping and rotating the images of the lung CT scans to ensure the model retains the most accurate images for image detection. The pre-processing process was followed with the training process, which included normalizations to uproot images that may negatively influence the training and testing process if they differ from accurate representations. During the training process, each model went through multiple epochs, or complete passes of training through the model, to allow the model to learn from new data (22). We collected both training loss (loss from the training set after each epoch) and validation loss (loss from the validation set of the data, which tests the performance of the model after training) to determine the error produced by the model. A higher loss value means the model is providing false output, while a lower loss value means there are fewer errors in the model (23).

Metrics included in this data collection include AUC, accuracy, sensitivity, and specificity. AUC is a measure of performance across classification thresholds (19). It measures

https://doi.org/10.59720/23-232

the accuracy of the difference between positive (cancerous) and negative (benign) classes. AUC values between 90-100% are considered excellent, while those above but not including 50% are considered acceptable. Anything below these AUC values is considered a failed model. Accuracy is a metric used to evaluate models based on the correct predictions. This metric uses the count of predictions to determine when the prediction value is equal to the true value. A test's sensitivity shows how many positive cases are detected out of a pool of positive cases and specificity shows how often a test comes back negative when it is truly negative. In this dataset, the data contains a 70% train, 20% validation, and 10% test split.

Github Code: https://github.com/siricher/LungCancerModel/blob/master/README.md

Received: April 24, 2023
Accepted: September 8, 2024
Published: June 7, 2025

REFERENCES

1. Scully, Jackie Leach. "What Is a Disease?" *EMBO Reports*, vol. 5, no. 7, July 2004, pp. 650–653, https://doi.org/10.1038/sj.embor.7400195.
2. "Covid-19: How to Manage Your Mental Health during the Pandemic." *Mayo Clinic, Mayo Foundation for Medical Education and Research*, 13 Dec. 2022, https://www.mayoclinic.org/diseases-conditions/coronavirus/in-depth/mental-health-covid-19/art-20482731#:~:text=The%20COVID%2D19%20pandemic%20may,what%20the%20future%20will%20bring. Accessed 25 Feb. 2023.
3. "The Evolution of AI in Healthcare." *Healthcare AI & Utilization Management*, https://www.xsolis.com/blog/the-evolution-of-ai-in-healthcare#:~:text=The%20Origins%20of%20AI%20in,identify%20blood%20infections%20treatments. Accessed 25 Feb. 2023.
4. Pinto-Coelho, Luís. "How Artificial Intelligence Is Shaping Medical Imaging Technology: A Survey of Innovations and Applications." *Bioengineering*, vol. 10, no. 12, 18 Dec. 2023, p. 1435, https://doi.org/10.3390/bioengineering10121435.
5. "What Is Deep Learning?" *IBM*, 17 June 2024, www.ibm.com/topics/deep-learning. Accessed 15 Aug. 2024.
6. Hosny, Ahmed, et al. "Artificial Intelligence in Radiology." *Nature Reviews Cancer*, vol. 18, no. 8, 17 May 2018, pp. 500–510, https://doi.org/10.1038/s41568-018-0016-5.
7. Alowais, Shuroug A., et al. "Revolutionizing Healthcare: The Role of Artificial Intelligence in Clinical Practice." *BMC Medical Education*, vol. 23, no. 1, 22 Sept. 2023, https://doi.org/10.1186/s12909-023-04698-z.
8. Manning. *LiveBook*, https://livebook.manning.com/concept/deep-learning/feature-map#:~:text=In%20CNNs%2C%20the%20feature%20map,%2C%20edges%2C%20or%20even%20objects. Accessed 3 Mar. 2023.
9. Mishra, Shambhavi. "An Introduction to VAE-Gans." *W&B*, 30 Oct. 2021, https://wandb.ai/shambhavicodes/vae-gan/reports/An-Introduction-to-VAE-GANs--VmlldzoxMTcxMjM5. Accessed 3 Mar. 2023.
10. *Neural Networks vs. Random Forests – Does it always have to be Deep Learning?, https://*blog.frankfurt-school.de/wp-content/uploads/2018/10/Neural-Networks-vs-Random-Forests.pdf. Accessed 15 Aug. 2024.
11. Alzubaidi, Laith, et al. "Review of Deep Learning: Concepts, CNN Architectures, Challenges, Applications, Future Directions." *Journal of Big Data*, vol. 8, no. 1, 31 Mar. 2021, https://doi.org/10.1186/s40537-021-00444-8.
12. GeeksforGeeks. "Autoencoders -Machine Learning." *GeeksforGeeks*, 6 Dec. 2023, https://www.geeksforgeeks.org/auto-encoders/. Accessed 15 Aug. 2024. (NOW 12)
13. "Keras Documentation: Convolutional Autoencoder for Image Denoising." *Keras*, https://keras.io/examples/vision/autoencoder/. Accessed 27 Feb. 2023.
14. "Advanced Guide to Inception V3 | Cloud TPU | Google Cloud." *Google*, https://cloud.google.com/tpu/docs/inception-v3-advanced. Accessed 12 May 2023.
15. "Inception V3." *Inception V3 - an Overview | Science Direct*, www.sciencedirect.com/topics/computer-science/inception-v3. Accessed 15 Aug. 2024.
16. Bank, Dor, et al. "Autoencoders." *ArXiv.org*, 3 Apr. 2021, https://arxiv.org/abs/2003.05991. Accessed 4 Mar. 2023.
17. Tate, Andrew. "Using Autoencoders for Feature Selection." *Hex,* 9 Oct. 2023, https://hex.tech/blog/autoencoders-for-feature-selection/. Accessed 15 Aug. 2024.
18. Bhandari, Aniruddha. "Guide to AUC ROC Curve in Machine Learning : What Is Specificity?" *Analytics Vidhya*, 6 Aug. 2024, https://www.analyticsvidhya.com/blog/2020/06/auc-roc-curve-machine-learning/. Accessed 15 Aug. 2024.
19. "Classification: Roc Curve and AUC | Machine Learning | Google Developers." *Google*, https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc#:~:text=AUC%20. Accessed 12 May 2023.
20. Wang, Jun, et al. "Denoising Autoencoder, a Deep Learning Algorithm, Aids the Identification of a Novel Molecular Signature of Lung Adenocarcinoma." *Genomics, Proteomics & Bioinformatics*, vol. 18, no. 4, 1 Aug. 2020, pp. 468–480, https://doi.org/10.1016/j.gpb.2019.02.003.
21. Hany, Mohamed. "Chest CT-Scan Images Dataset." *Kaggle*, 20 Aug. 2020, https://www.kaggle.com/datasets/mohamedhanyyy/chest-ctscan-images. Accessed 12 May 2023.
22. "What Is Epoch in Machine Learning?: Unext." *UNext*, 7 Mar. 2023, https://u-next.com/blogs/machine-learning/epoch-in-machine-learning/. Accessed 15 Aug. 2024.
23. "Training and Validation Loss in Deep Learning." *Baeldung on Computer Science*, 10 July 2024, https://www.baeldung.com/cs/training-validation-loss-deep-learning. Accessed 15 Aug. 2024.