Optimizing Air Quality Classification: A Comparative Analysis of Large Language Models and Deep Learning Models to Provide An Accessible Health Platform

Morgan Tsai

All figures were made by student unless otherwise stated.

Abstract

Air pollution causes over 6 million premature deaths annually worldwide. High quality air pollution detectors are costly and hefty, making it less feasible for large scale deployment. Traditional air quality monitoring relies heavily on fixed monitoring stations and costly equipment, which limits accessibility. To address this issue, we propose a photo based air quality system, leveraging smartphone technology to estimate air quality through image analysis. Large language models (LLMs) and deep learning models can analyze visual indicators such as haze and visibility in order to approximate air quality levels. Assessing air quality based on photos enables accurate air quality information, crucial to those with health vulnerabilities. In this project we hypothesized that existing deep learning models will provide a more accurate analysis compared to LLMs based on deep learning models being able to be customized for air quality. Using air pollution images to demonstrate how deep learning models can efficiently identify air quality from photographs, deep learning models show a superior performance compared to all large language models tested, offering a cost-effective solution to measuring air quality world wide. We tested multiple LLMs and deep learning models to evaluate their performances in classifying air quality levels. The results showed that deep learning models outperform LLM architectures in accuracy, showing possible potential in deep learning being a reliable method for global air quality assessment in underserved areas.

Introduction

In 2013, the International Agency for Research on Cancer of the World Health Organization classified air pollution as a human carcinogen (Loomis, 2013). Air pollution continues to contribute to premature deaths across the world, with around four to ten millions deaths per year (Cohen, 2005). Pollutants such as particulate matter, nitrogen oxides, sulfur oxides, and volatile organic compounds (VOCs) are major contributors to conditions like cancer, stroke, and heart disease (Burnett 2018; Manisalidis, 2020). As the impact of air pollution grows, the need for an early warning system to minimize its health risks become increasingly critical. A study done in 2017 found that timely alerts can significantly reduce emergency room visits due to air pollution (Solimini and Renzi, 2017). These measures are vital for protecting public health, and as pollution levels in areas rise, the demand for real time monitoring systems becomes even more urgent. However, affordable air quality monitoring systems remain a challenge, especially in underserved communities where resources are limited. Applying air quality research helps the economy as well, which is especially important in underserved communities. Not having to be able to spend money on expensive air quality monitors will allow for resources to be allocated towards other everyday essentials. While current research has focused on improving traditional air quality monitors, there is opportunity to explore how technology such as large language models (LLMs) and deep learning models can contribute to air quality analysis. Little research has explored the potential of LLMs in analyzing images for air quality categorization. This study aims to fill that gap by investigating whether multiple LLMs can effectively assess air pollution levels from photos. In doing so, this research takes on a more holistic approach to tackling the growing global health problem of air pollution while also examining the idea of looking at potential applications for technology in environmental health.

Review of Literature

a. Current Air Quality Monitoring

The IMPROVE (Interagency Monitoring of Protected Virtual Environments) network is used by the U.S. Environmental Protection Agency (EPA) with there being four objectives it aims to look at. They are establishing visibility and aerosol conditions in Class I areas (CIAs), identify chemical species and sources responsible for existing anthropogenic visibility impairment, document long term trends for assessing progress towards the visibility goal, and to provide regional haze monitoring responsible for all protected CIAs (Hand et al., 2023). There have been six reports since IMPROVE was implemented in 1988, yet despite there being these networks are strictly confined to national parks and not to the common public. Geographically, this network gives people in the Midwest less of an advantage compared to those

in the Northeast or South (figure 1). In Maine alone, there are five IMPROVE sites, while in Texas there are only two. The amount of sites in each region greatly vary, and for those who don't live in centers with multiple sites, let alone near a national park, accurate and accessible air quality monitors are severely lacking.

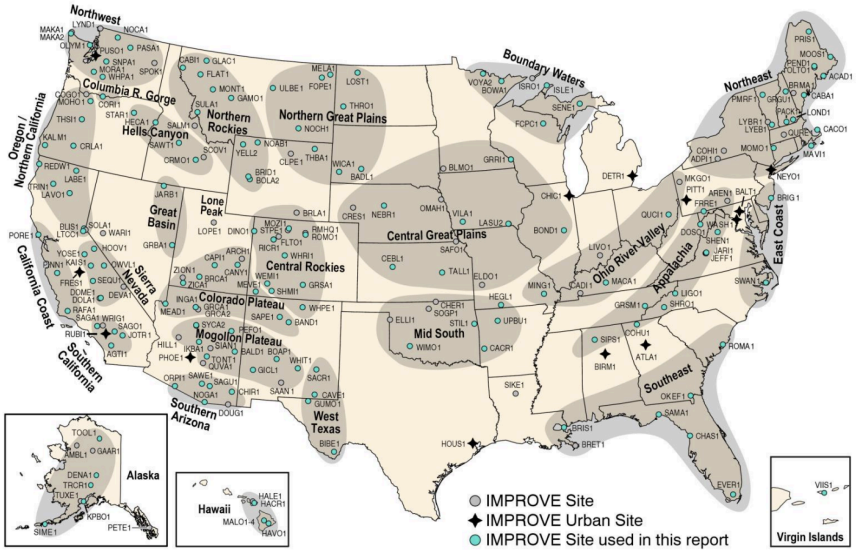

Fig. 1: Location of IMPROVE sites across America (Hand et al., 2023).

Near road monitors are another current type of air monitors available to the general public. Typically stationed 100 meters off a roadway, there's been an upward trend in the stationing of them in high traffic environments (Xie et al., 2024). A study that analyzed 156 studies on roadside air quality monitoring found that there was a high volume of research in the USA, England, and China. Each country has their objective for setting up these stations, with the United States using the roadside monitor data for the national ambient air quality standard (NAAQS) (Rangel et al., 2011) and the United Kingdom using their data to assess the health impacts of traffic (Liang et al., 2018). Despite there being a steady increase in the roadside monitors as high traffic environments become more rampant, this monitor is confined to certain geographic locations. Not living near a research institution that often writes research on air quality would mean that there would have to be other means for communities to get accurate and affordable air quality data.

Aeroqual is an air quality monitoring system that's designed to provide real time data on ozone (O_3), nitrogen dioxide (NO_2), $PM_{2.5,}$ temperature, and relative humidity (Lin et al., 2015). These devices are often used in areas with high pollution levels, such as construction zones or near roadside monitors.While

Aeroqual provides real time air quality data, its bulky design makes it impractical for everyday use. Portable devices like smartphones can be carried around easily and are much more feasible for personal use on a daily basis.

Access to high quality air quality monitoring is limited, especially in developing countries where infrastructure for these systems are lacking. Although air pollution is a global problem, not all nations have the resources of capabilities needed to deploy monitoring systems. However, there are high cell phone penetration rates in many developing countries, opening up opportunities for mobile based solutions. There is a great need to provide accessible and low cost monitoring systems, helping communities gain insight into the environmental health in their community.

b. Deep Learning

Deep Learning has emerged in technology, reshaping the way that multiple fields operate. Deep Learning is a subset of Artificial Intelligence (AI) that uses neural networks to allow computers to learn from data in a similar fashion from a human. There are multiple different applications for deep learning that different models are equipped for, allowing users to pick and choose which one is best fit for their needs. This study examines how deep learning models and LLMs can be leveraged to monitor air quality.

i. Artificial Intelligence (AI) and Its Applications

Artificial intelligence (AI) has been on the rise recently, with an increasing role in enhancing human life (Haleem, 2019). Simulating the features of human intelligence through computer systems, LLMs can be used in multiple different fields. LLMs are notably used in healthcare, where their ability to learn and predict helps doctors interpret medical data and images for better decision making. In drug discovery, LLMs analyze data on proteins and genetic makeup to identify biomarkers, enabling researchers to predict which patients will respond to specific drugs. This improves the efficiency of trial design and drug development, ensuring treatments are tailored to the most responsive patient groups (Arnold, 2023). Beyond healthcare, however, LLMs are also making impacts in environmental research. LLMs can be used to analyze datasets from satellites and weather stations to help scientists predict weather events such as hurricanes and snowstorms (Haupt et al., 2022). Additionally. LLMs can help farmers to optimize their resource usage by taking data of their water and crop yields to recommend strategies to minimize the amount needed. Despite the recent advancements made in LLMs though, most of the models are not openly accessible to the general public. If such models were available, those with limited resources could use LLMs to address the issues in their own lives.

ii. Deep Learning Models

A major reason why we have advancements in LLMs is because of deep learning, a subset of machine learning that uses deep neural networks to make decisions that are similar to those of humans (Naveed, 2024). For instance, deep learning models such as convolutional neural networks (CNNs) can be used for image classification in medical images such as X-rays and MRI's, analyzing them at a much quicker rate than a doctor (Yadav and Jadhav, 2023). Deep learning models extend past healthcare; in a study done by Pu-Yun Kow, CNN's and regression classifiers (RCs) were used to estimate concentrations of pollutants and the Air Quality Index (AQI) based off of an image (Kow et al., 2022). In the study, though, only one model was analyzed. In our research, we aim to look at multiple architectures as comparison to see which is best suited for air quality analysis.

There are multiple different types of architectures available in the deep learning field. Convolutional Neural Networks are one of the most known architectures for image processing. CNNs connect three types of layers with different convolutional layers, analyzing the hidden patterns using pooling layers to capture the spatial features of a photo (Khoei, 2023). Another type of deep learning models are recurrent neural network (RNN) models, which are designed to recognize sequences and patterns within the input data (Mosavi, 2020). This capability allows RNNs to effectively analyze time dependent information and make predictions based on the relationships between data points (Khoei, 2023). While CNNs and RNNs specialize in structured data types such as images, graph neural networks (GNNs) are also used in combination with CNNs for image classification purposes. By representing images as nodes in a graph, GNNs can analyze relationships between regions within an image, which is particularly beneficial for tasks that require spatial understanding of an image (Knyazev, 2019). In this study, we explore multiple deep learning models applied to computer vision tasks, examining and comparing each structure in terms of accuracy.

An example of a CNN is when combating air pollution, researchers developed Eff-AQI, an LLM that was specifically designed for accurate air quality and particulate matter (PM). EFF-AQI used CNN architecture to analyze images in urban environments to detect pollutants through photos (Utomo et al., 2023). In this study, we use the dataset from the same study to look into the different vision and LLMs models that can be used for air quality analysis.

c. Computer Vision Models

Computer vision models are another subset of deep learning that enable devices to interpret visual information from a given dataset. By utilizing neural networks, models can analyze visual data to identify

objects and detect patterns to recognize features within images. There are multiple different types of models available for use, with their own limitations and niches for specific applications. It's crucial to select the appropriate architecture based on the input and desired output to achieve optimal performance. In this work, we compare the accuracy rates of 4 computer vision models in categorizing air quality. In this study the models examined--ResNet, EfficientNet, VGG, and DenseNet--serve as a representative sample of how deep learning can be applied in the environmental field. In other fields, ResNet and EfficientNet have been commonly used in medical imaging, helping identify diseases from scans more efficiently (Rajpurkar et al., 2017; Latha et al., 2024). DenseNet has also been used medically, helping identify Alzheimer's from brain scans (Li and Liu, 2018). VGG has been widely used in facial recognition systems to advance security in smart devices (Parkhi, 2015).

i. ResNet

One of the computer models that we want to look into is ResNet, a deep learning architecture that uses residual learning by adding shortcut connections between layers (He et al., 2015). These connections skip over layers to make it easier for the network to retain and use the information used in earlier layers. In environmental science, ResNet has capabilities to detect and monitor environmental changes such as deforestation in satellite images, aiding environmental scientists to understand and track environmental trends (Cheng et al., 2017). As adept as the ResNet model is in image analysis, it does require much more power from the computer than other models, making it more demanding for those who have limited processing power.

ii. EfficientNet

EfficientNet is another deep learning vision model that uses compound scaling, a method that balances depth, width, and resolution of the model. The model stays true to the name by achieving accuracy with less computational cost, making it versatile for both mobile phones and computers (Tan and Le, 2019). In terms of image analysis, EfficientNet has also been used for radiology purposes to detect breast cancer at a more accurate rate. The model is especially used in places where resources may be limited due to it being less demanding compared to other models (Latha et al., 2024). In terms of the limitations, EfficientNet does require a large amount of data to train it effectively. In cases where labeled data is minimal, the model's performance may not be as accurate as other models.

iii. VGG

Visual Geometry Group (VGG) is one of the more simple architectures that consist of convolutional layers that are connected to perform image classification (Simonyan and Zisserman, 2014). In the

environmental health field, VGG is used to identify crop diseases using images of plant leaves to reduce agricultural losses for farmers (Thakur et al., 2023). Being the more simplistic model, VGG models are much more computationally expensive in terms of memory and time. In a real time situation, VGG models may be much slower in identifying images compared to much more efficient architectures.

iv. DenseNet

Densely Connected Convolutional Networks (DenseNet) is another image recognition neural network that connects each layer to one another to help the network learn better by allowing it to use features from earlier layers more efficiently (Huang et al., 2016). When applied to the environmental field, DenseNet has been used to enhance the classification of Environmental Microorganism Dataset (EMDS) images (Chen, 2022). DenseNet-201 was used for this particular study and was optimized to enhance its ability to classify EMDS at a high accuracy rate of 98.4%. The limitations behind DenseNet also connect with its resources. The architecture requires a significant amount of memory due to its connections, which can be limiting when working with devices with limited resources. Furthermore, because each layer is connected to one another, the training time will be longer compared to ResNet that has skip connections between each layer.

d. Large Language Models

Large Language Models (LLMs) have made immense progress over the years. A type of AI, they've become popular tools in natural language processing (NLP), machine translations, and question answering (Hadi, 2023). In customer service situations, LLMs are often used as chatbots to provide virtual assistance to customers who are in need (Chen et al., 2017). In academic research, researchers can use the summarization modalities of LLMs to efficiently review long papers in a concise manner, allowing them to find research gaps and trends in their respective field (Nallapati et al., 2016). The impact of LLMs also expands into the medical field, where healthcare professionals can use the summarization abilities to sift through patient data, increasing their decision time (Meng et al., 2014). Through the popularization of LLMs in multiple fields, zero shot learning has been popularized where without any prior training is needed. These models are trained on general data effectively enough to perform well without task specific data (Radford et al., 2019). Our proposed goal is to see if given a prompt, if LLMs can categorize air quality based on a photo. The four models that we want to examine are ChatGPT-4o, Gemini, Claude 3.5 Sonnet, Phi-3 Vision, and LlaMa.

ChatGPT-4o is the newest ChatGPT model made by OpenAI. Specifically designed to produce faster responses by training a single new model that combines text, vision, and audio modalites, GPT-4o is the

first of its kind to be suitable for diverse input types (Pang, 2024). There hasn't been much research using ChatGPT-4o, but there has been research with past models of ChatGPT. ChatGPT is commonly used in education for learning and verifying academic answers (Liu et al., 2023) and in the medical field, where it helps users and professionals with diagnosing diseases by answering questions. Despite there being a vast amount of research done with ChatGPT across multiple different fields, there has yet to be much research on the image analysis capabilities of GPT, especially in environmental contexts. In this project, we aim to bridge that gap, looking if ChatGPT can assess the air quality based off of a photo accurately.

Google's Gemini model has been an up and coming LLM for research. Gemini was designed to be a multimodal system from the outset, integrating text, image, and audio capabilities within the model. On the other hand, ChatGPT initially focused on the conversational input vs output and later expanded to include other modalities (Imran, 2024). In a study done comparing Gemini versus ChatGPT, it was found that Gemini holds an edge in delivering factual information due to its connection with Google Search (Rane et al., 2024). It is worth noting however that ChatGPT leads in terms of conversational flow and tends to be more engaging.

Claude 3.5 Sonnet is provided through Anthropic, a U.S. based company. Claude models are trained using hardware from Amazon Web Services and Google Cloud Platform and similar to Gemini, has multimodal input capabilities that includes both text and image (Anthropic, 2024). Compared to Gemini though, yet similar to ChatGPT, Claude lacks the real time search capabilities which means that its responses are based on the dataset that it's trained upon. In terms of giving the latest and most accurate responses, Claude is limited.

Phi-3 Vision is a recent advancement in Microsoft's large language models. Utilizing synthetic data and trained on publicly available web data, the multimodal model has different uses (Abdin, 2024). Microsoft cites that Phi-3 can be used for visual content analysis for educational tools as well as video footage analysis for insurance. Both use image analysis and for looking at the nuances of air quality from a photo, there is great potential for Phi-3 to have success.

LlaMa by Meta is an open sourced model meaning that researchers can adapt the LLM to suit their needs. Trained in a similar fashion with other LLMs such as ChatGPT and Phi-3, the training set consists of multiple different datasets for it to be capable of answering different inputs (Touvron et al., 2023). Unlike Gemini, yet similar to ChatGPT, LlaMa answers solely based on the dataset it's trained on, meaning it's

up to the developers to update and continue to train the model for the most accurate results. (Touvron et al., 2023).

Problem

1) With air quality becoming more and more of a pressing issue that continues to cause millions of premature deaths, there is a significant gap in research in how AI, particularly LLMs, can provide real time analysis of air quality through image assessment.
2) There has been little research in comparing different convolutional neural networks in categorizing air quality. We aim to bridge that gap by comparing four different models to examine which one is the best application for air quality categorization.

Objectives

1) We aim to compare the performance of four deep learning models (ResNet, EfficientNet, VGG and DenseNet) in categorizing air quality from images to assess both the effectiveness of convolutional neural networks in air quality image classification and to examine which model is the best suited for it.
2) We intend to examine the abilities of large language models in accurately assessing air quality.

Hypothesis

In this research, we hypothesize that convolutional learning networks (CNNs), specifically ResNet, EfficientNet, VGG, and DenseNet will outperform large language models in categorizing air quality from images. This is due to the ability of CNNs to be customized and trained on a dataset of labeled air quality images, allowing them to more effectively hone in on the specific patterns that are associated with a specific air quality level.

Methodology

In this study, we utilized an open source dataset provided by Adarsh Rouniyar on GitHub, containing images captured across various regions in India and Nepal. The dataset contains 12,240 images in total, offering a diverse range of pollution levels for the models to be trained on. There were six classes of air pollution that the dataset categorized each photo in: good, moderate, unhealthy for sensitive groups, unhealthy, very unhealthy, and hazardous (**table 1**).

THIS SPACE WAS INTENTIONALLY LEFT BLANK

Table 1: Air quality index (AQI) with corresponding levels and impact on human health that were referenced in this study (Rouinyar, 2023).

Air Quality Index and Activity Guidance

AQI	0-50	51-100	101-150	151-200	201-300	301-500
Air Quality Index Levels of Health Concern	Good	Moderate	Unhealthy for Sensitive Groups	Unhealthy	Very Unhealthy	Hazardous
Status Color	Green	Yellow	Orange	Red	Purple	Maroon
Impact on Human Health	Air quality is considered satisfactory, and air pollution poses little or no risk.	Air quality is acceptable; however, for some pollutants there may be a moderate health concern for a very small number of people who are unusually sensitive to air pollution.	Members of sensitive groups may experience health effects. The general public is not likely to be affected.	Everyone may begin to experience health effects; members of sensitive groups may experience more serious health effects.	Health alert: everyone may experience more serious health effects.	Health warnings of emergency conditions. The entire population is more likely to be affected.

The distribution of images across classes allows models to learn the nuanced differences between each category (**figure 2**). By ensuring that each class is adequately represented, the model can then capture visual cues associated with different air quality levels such as haze and visibility differences. In real world situations, the model can adapt to the weather's natural variability within photos. For instance, it can recognize how fog can influence the appearance of pollution and as a result, the model becomes more accurate in predicting air quality in all contexts.

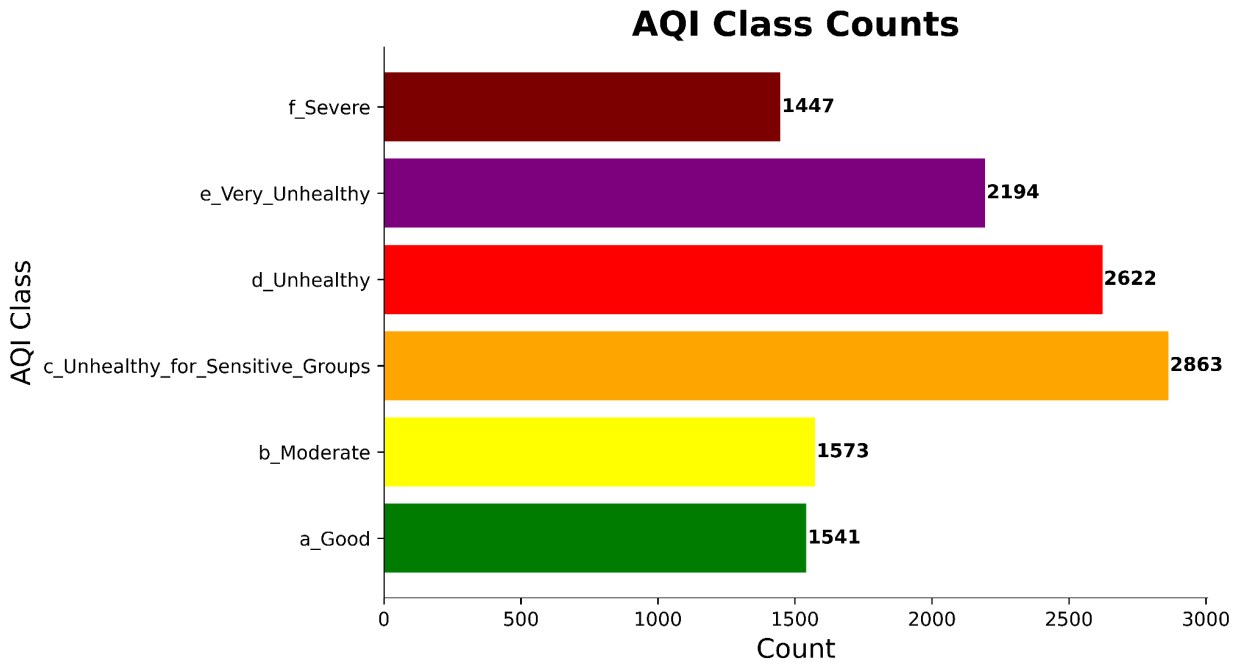

Figure 2: Class count within the dataset ((Rouinyar, 2023)

The nuances in air quality classification are not readily noticeable to the naked eye, as illustrated in figure 3. These visual similarities can lead to misinterpretations of air quality levels, potentially compromising one's health. This subtlety emphasizes the importance of using models to detect and differentiate these differences, as traditional visual assessments aren't as adept at recognizing the critical varions in air quality. Machine learning models can analyze large amounts of data quickly and more effectively, identifying features that are imperceptible to the human eye.

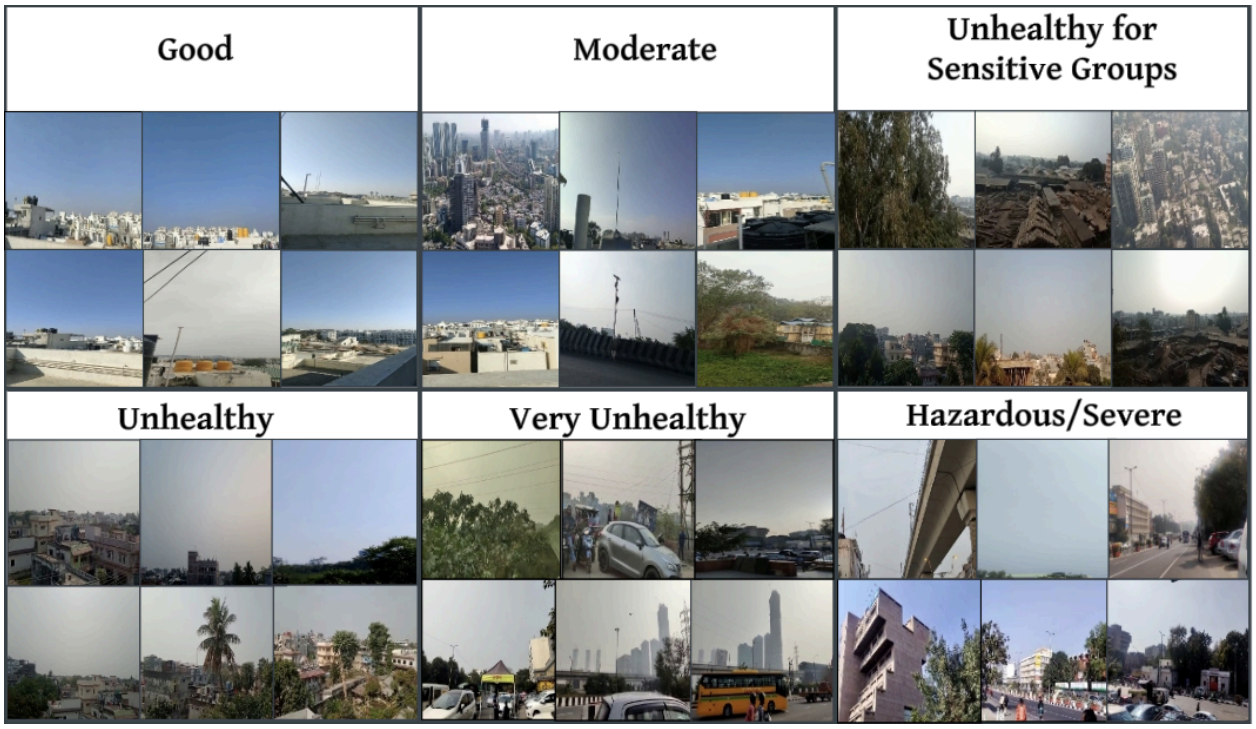

Figure 3: Examples of images in the dataset. The subtle differences can be shown in the similarities between the unhealthy photos versus the moderate photos.

a. Large Language Models

Going back to the objective to see if LLMs can categorize air quality images effectively, multiple LLMs were tested. Namely, ChatGPT-4o from OpenAI, Gemini from Google, Phi-Vision from Microsoft, Claude by Anthropic, and LlaMa by Meta. All LLMs were chosen because of their popularity and multimodal capabilities in image classification tasks. Each model had a chatbot option, which allowed for our query to be sent in an interactive manner, facilitating real time feedback. Each model was evaluated based on its ability to accurately categorize images into distinct air quality classes. Due to the limitations with LLM capacities, we implemented a quantitative testing approach. For each category, we selected 5 representative photos as inputs, which allowed us to evaluate how well each model could interpret and

classify air quality. Following the input, we then record the output for each model. The prompt for each LLM was as follows: “You are an expert in evaluating the air quality of an area based on an image. Here is an image — focus on the background and give an estimate of air quality as Good, Moderate, Unhealthy for Sensitive Groups, Unhealthy or Hazardous.” After each LLM received the prompt, it then processed the input to generate an output classification for air quality.

b. Deep Learning Models

On the deep learning side, different architectures were used as well. EfficientNet, ResNet, DenseNet, and VGG were used due to their popularity in image processing. Each varies greatly in terms of computational power and simplicity. In our study, each model was trained on a dataset comprising an 80-20 split, with the training data containing 80% of the images, and the testing data being the remaining 20%. For each model, we conducted experiments using six different combinations of batch sizes and epochs to determine the optimal settings for training. The combinations included a batch size of 8 and 3 epochs, batch size of 8 and 10 epochs, batch size of 16 and 3 epochs, batch size of 16 and 10 epochs, batch size of 32 and 3 epochs, and batch size of 32 and 10 epochs. By testing these variations, we then analyzed how different batch sizes (the number of training examples processed) and epochs (number of passes through the training dataset) impacted the models’ performance in classifying air quality images. Each analysis for the deep learning model was done through GoogleColab, an open source cloud based platform that’s user friendly for executing Python scripts. Furthermore, the code for this research was adapted from a preexisting image analysis open source code, allowing us to build upon a pre-existing methodology and customize it for our air quality classification needs.

c. Student vs Mentor Role

My role involved reviewing journal articles and gathering prior knowledge in the field of environmental science. I ran the code for the convolutional neural network models and collected data based on the inputs provided to the large language models. Once the data was gathered, I analyzed it to answer the research questions and draw conclusions. My mentor supported me in writing and refining the code for the project, guiding me through any technical challenges I encountered. My mentor also helped me frame my research methodology, especially on the convolutional neural network side, helping me ensure that the approach I took with my research would ensure an equal comparison. Additionally, my mentor provided feedback on analysis and assisted me through interpreting the results.

Results/Discussion

a. Large Language Models

Despite each LLMs models' advanced architecture, our findings indicate that none of the models are capable of accurately classifying the air quality images across each category (**fig 4.**).

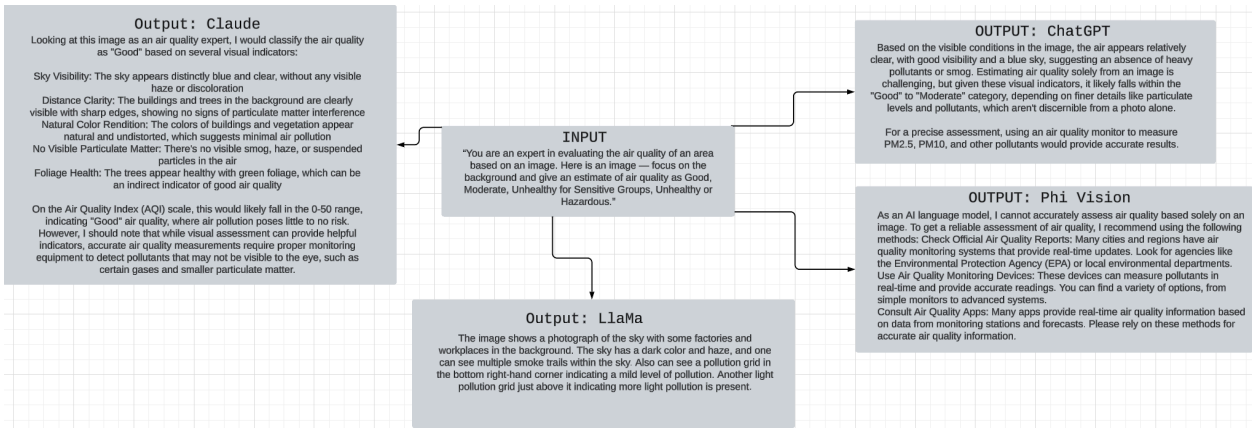

Figure 4: Each output for each LLM. Although the input for each LLM was the same, the output for each model was different.

ChatGPT did give a response, classifying the photo as "good." Despite this, the photo input for this example was in the "unhealthy" category, deeming it incorrect. The model did recognize its limitations and recommended using air quality monitors for more precise evaluations. Claude also classified an "unhealthy" photo as "good," citing several visual indicators such as clear sky visibility, absence of particular matter, and clarity of distance -- all of which were incorrect. It also goes a step further by estimating an AQI range of 0-50, showing that it does understand air quality metrics. However, similar to ChatGPT, it also acknowledges the limitations behind unseen pollutants in the visual input.

Unlike ChatGPT and Claude, Phi-Vision and LlaMa did make the attempt to classify the photo, yet inaccurately categorized the input given. LlaMa did describe the scene as having dark colors and haze, which could be indications of mild pollutants, but did not end up classifying it under the categories of air quality it was given. This response more so recognized the poor air quality conditions instead of an actual classification which left the assessment somewhat ambiguous. Phi-Vision had the most dismissive response out of the four, deeming it impossible to estimate air quality based on an image alone. The model did not attempt to assess any visible indicators or make a definitive classification.

Each LLM had their own unique way of assessing air quality based on an image. Some models prioritized accuracy in their responses, while others attempted to interpret visible cues, although they did risk

misclassification. All models recognized the limitations with each, suggesting specific air quality monitoring devices or apps. In real time use, there is potential for LLMs to act in conjunction with other data sources to gain the most accurate assessment of air quality. It's also worth noting that if a model misclassifies air quality and individuals make decisions based on incorrect information, there could be legal repercussions. Many LLMs also operate based on the data they've been trained on, which may not reflect all real world conditions accurately. Without being equipped for a wide variety of scenarios, LLM outputs may not be the most accurate of making environmental assessments. LLMs are still growing though, so in the future, there are possibilities of an increased accuracy in air quality classification.

b. Deep Learning

In this study, we examined four models–EfficientNet, ResNet, DenseNet, and VGG–to compare their performances in image classification tasks. Each model has their own unique architecture, making these four a representative sample of convolutional neural networks as a whole. The results reflect each model's architectural distinctions, with performance variations that are tied to their own unique design. In each figure below, trials for each model are displayed, with the highest accuracy combination highlighted in yellow for clarity.

i. EfficientNet

There were six different combinations of batch size and epochs tested to see the best combination for efficiency and accuracy. Generally, the more epochs that are used, the more the model can learn from the data and reduce the training error. This was found to be the case for EfficientNet, with the optimal combination being a batch size of 16 and 10 epochs for the best validation accuracy of 75.21% (**table 2**). The runtime for this combination was among one of the longest at 488 seconds. Although this was one of the longest training times, the increased epochs allowed the model more opportunities to learn from the data and adjust its weights more effectively. Beyond a certain number of epochs, the improvements in accuracy tend to diminish, making the 10-epoch mark suitable for EfficientNet's architecture. EfficientNet models are built to achieve high accuracy with a lower computational load compared to other deep learning networks, so using even more epochs would mean increasing the runtime without adding much to the accuracy percentage.

THIS SPACE WAS INTENTIONALLY LEFT BLANK

Table 2: All trials for EfficientNet shown with the metrics of the best validation accuracy and timing being measured.

Batch Size	Epochs	Best Validation Accuracy	Timing (s)
8	3	67.53%	181
8	10	72.31%	604
16	3	67.53%	146
16	10	75.21%	488
32	3	65.32%	132
32	10	74.23%	438

ii. Resnet

ResNet also had the most optimal combination at a 16 batch size and 10 epochs, although the best validation accuracy was slightly lower at 74.64% (**table 3)**. However, ResNet's accuracy was slightly lower than EfficientNet, and the training time was also slightly shorter at 417 seconds. While this combination of a 16 batch size and 10 epochs produced the highest overall validation accuracy, ResNet's peak accuracy was reached at 8 epochs. This suggests that ResNet reaches a close to its maximum learning capacity within fewer epochs. The impact of batch size on ResNet's performance is similar to EfficientNet's results, as smaller batch sizes yielded a higher accuracy, most likely due to variability in each batch size. A smaller batch size will make the model adjust to its learning, preventing it from becoming too fixated on specific details of the training data, in turn making the model more adaptable when it's tested on new data.

Table 3: Impact of batch size and epochs on validation accuracy and training time for ResNet.

Batch Size	Epochs	Best Validation Accuracy	Timing (s)
8	3	69.12%	148
8	10	73.42%	471
16	3	70.84%	126
16	10	74.64%	417
32	3	66.87%	120
32	10	74.60%	402

iii. DenseNet

DenseNet had the most optimal combination at a batch size of 32 and 10 epochs with a best validation accuracy at 77.01%, not far off from the second greatest validation accuracy at a batch size of 16 and 10 epochs (**table 4**). The slight difference in accuracy suggests that DenseNet benefits from larger batch sizes during training, which may provide smoother updates to the architecture. The runtime was 567 seconds, relatively long compared to other architectures, but not as long as the other runtimes for DenseNet. DenseNet demands computational power due to its architectural set up, meaning that a larger batch size helps with the training process without sacrificing accuracy. DenseNet's ability to sustain accuracy with 10 epochs suggests that the densely connected structure benefits from extended training time, allowing it to refine the shared features across layers for greater generalization.

Table 4: Comparison of all combinations of batch sizes and epochs when tested with the DenseNet architecture.

Batch Size	Epochs	Best Validation Accuracy	Timing (s)
8	3	69.78%	226
8	10	73.25%	756
16	3	72.72%	180
16	10	74.85%	589
32	3	69.08%	166
32	10	77.01%	567

iv. VGG

Similarly to ResNet and EfficientNet, the optimal combination for VGG was achieved with a batch size of 16 and 10 epochs, yielding a best validation accuracy of 77.91% at a runtime of 687 seconds (**table 5, overleaf**).This combination provided the highest accuracy for VGG, slightly outperforming the batch size of 8 with 10 epochs, which reached 77.83%. While the validation accuracy between the two combinations is close, the batch size of 16 is a more balanced trade off between accuracy and time as the 8-10 combination lasting 933 seconds. The runtime of 687 seconds, although long compared to other models, is expected of VGG due to it requiring a substantial amount of computational resources. However, the result for this combination is still efficient when compared to other combinations where the longer training times do not yield a significant improvement in accuracy. Overall, the results suggest that VGG benefits from a moderate batch size to achieve a moderate balance between accuracy and runtime.

Table 5: Validation accuracy and training time for VGG with different batch sizes and epochs.

Batch Size	Epochs	Best Validation Accuracy	Timing (s)
8	3	72.80%	281
8	10	77.83%	933
16	3	71.12%	206
16	10	77.91%	687
32	3	67.12%	205
32	10	74.52%	684

v. Comparison

Out of all of the models, the model with the highest validation accuracy was VGG, with an accuracy of 77.91%. The rest of the models accuracy rates are as follows: DenseNet, ResNet, and EfficientNet (**table 6)**.

Table 6: Comparison of all models' most optimal combinations.

Models	Optimal Batch Size	Optimal Epochs	Best Validation Accuracy	Timing (s)
EfficientNet	16	10	75.21%	488
ResNet	16	10	74.64%	417
DenseNet	32	10	71.12%	206
VGG	8	10	77.01%	567

Out of all of the models, the model with the highest validation accuracy was VGG, with an accuracy of 77.91%. The rest of the models accuracy rates are as follows: DenseNet, ResNet, and EfficientNet. However, VGG's accuracy comes with a trade-off: it has the longest average runtime among the models, taking 499.3 seconds to complete its training. This could be due to its high computational demands, which could be a limiting factor in environments where resources are scarce. Following VGG, DenseNet attained a validation accuracy of 77.01% with an average runtime of 412.3 seconds. For applications that want a high accuracy along with a lower runtime, DenseNet could be the best suited model. In contrast to VGG, ResNet recorded the lowest runtime at 280.69 seconds, making it the most efficient model in terms

of training time. While its accuracy is lower than VGG's, ResNet's architecture has residual connections, allowing deeper networks to be trained more. In applications where time is a critical factor, ResNet would be the best suited. Additionally, EfficientNet achieved a validation accuracy of 75.21% and an average runtime of 331.5 seconds. Although its accuracy falls short of VGG and DenseNet, EfficientNet is still a viable option for scenarios that want accuracy along with a runtime similar to ResNet's. The choice of the model should depend on the context and goals of the application. For maximum accuracy, VGG is preferred; but for a balance of accuracy and runtime, DenseNet would be the best option. For the most efficient training time, ResNet would be best suited; and for a combination of a slightly lower accuracy compared to VGG with a lower processing time, EfficientNet is a strong contender. Each model has their own unique architecture that can be used according to the demands of the task at hand.

Conclusion

In this research, 4 LLMs and 4 deep learning model's were tested in determining the effectiveness of categorizing air quality. This study was the first conducted to compare different large language and deep learning models in the context of classifying air quality. All LLMs tested were either incorrect or did not categorize the input at all, but there is still potential for improvement. For the deep learning models, VGG had the highest accuracy rate at 77.91%, but most accuracy rates fell in a similar range of one another. Timing was another factor though, with the shortest time being DenseNet, but the highest was 488. On the deep learning side, most accuracy rates fell in a similar range of one another, indicating that while progress with the architecture has been made, there remains room for enhancement for image classification. To advance the effectiveness of these models, future research should consider hybrid models that could combine models together to provide the most accurate and efficient model. For instance, integrating convolutional neural networks with recurrent neural networks is a popular application that could be applied to image classification. Each convolutional neural network tested had their own strength and the choice of the model depends on the specific goals of the application. Each model's architecture suits different tasks, allowing flexibility for the users in their model selection based on application needs.

As these models evolve and improve, their potential applications in real world scenarios become increasingly significant. LLMs and deep learning models can be utilized to inform public policy decisions, enabling communities to respond to different environmental circumstances. Furthermore, integrating these models into mobile applications such as through an iPhone camera could make the platforms more accessible for communities with fewer resources. This accessibility could make it easier

for users to make decisions about their health and well being based on the conditions. The results from this study is a promising start, providing a solution that is cost effective, contributing to a more sustainable future. The models can also be tested in different environments to help implement the system into someone's everyday lifestyle. Testing this enables a much more concise strategy for air quality management, enhancing its usability. Incorporating user feedback is necessary into bettering the model for future use, helping make a better model and promoting more awareness on air quality's effects on an individual's health. Overall, this research provides the first step in understanding how LLMs and deep learning models can play a role in addressing air quality challenges, ultimately leading to positive changes in the health of the community.

REFERENCES

Abdin, M., Jacobs, S. A., Awan, A. A., Aneja, J., Awadallah, A., Awadalla, H., … Del Giorno, A. (2024, April 23). Phi-3 Technical Report: A Highly Capable Language Model Locally on Your Phone. https://doi.org/10.48550/arXiv.2404.14219

Arnold, C. (2023). Inside the nascent industry of AI-designed drugs. *Nature Medicine*, *29*(6), 1292–1295. https://doi.org/10.1038/s41591-023-02361-0

Burnett, R., Chen, H., Szyszkowicz, M., Fann, N., Hubbell, B., Pope, C. A., … Lim, C. C. (2018). Global estimates of mortality associated with long-term exposure to outdoor fine particulate matter. *Proceedings of the National Academy of Sciences*, *115*(38), 9592–9597. https://doi.org/10.1073/pnas.1803222115

Chen, H., Liu, X., Yin, D., & Tang, J. (2017). A Survey on Dialogue Systems. *ACM SIGKDD Explorations Newsletter*, *19*(2), 25–35. https://doi.org/10.1145/3166054.3166058

Chen, W., Liu, P.-Y., Lai, C.-C., & Lin, Y.-H. (2022). Identification of environmental microorganism using optimally fine-tuned convolutional neural network. *Environmental Research*, *206*, 112610–112610. https://doi.org/10.1016/j.envres.2021.112610

Cheng, G., Han, J., & Lu, X. (2017). Remote Sensing Image Scene Classification: Benchmark and State of the Art. *Proceedings of the IEEE*, *105*(10), 1865–1883. https://doi.org/10.1109/jproc.2017.2675998

Cohen, A. J., Ross Anderson, H., Ostro, B., Pandey, K. D., Krzyzanowski, M., Künzli, N., … Smith, K. (2005). The Global Burden of Disease Due to Outdoor Air Pollution. *Journal of Toxicology and Environmental Health, Part A*, *68*(13-14), 1301–1307. https://doi.org/10.1080/15287390590936166

Hadi, M. U., Tashi, Q. A., Qureshi, R., Shah, A., Muneer, A., Irfan, M., … Mirjalili, S. (2023, July 10). A Survey on Large Language Models: Applications, Challenges, Limitations, and Practical Usage. https://doi.org/10.36227/techrxiv.23589741.v1

Haleem, A., Javaid, M., & Khan, I. H. (2019). Current status and applications of Artificial Intelligence (AI) in medical field: An overview. *Current Medicine Research and Practice*, *9*(6), 231–237. https://doi.org/10.1016/j.cmrp.2019.11.005

Haupt, S. E., Gagne, D. J., Hsieh, W. W., Krasnopolsky, V., McGovern, A., Marzban, C., … Williams, J. K. (2022). The History and Practice of AI in the Environmental Sciences. *Bulletin of the American Meteorological Society*, *103*(5), E1351–E1370. https://doi.org/10.1175/bams-d-20-0234.1

He, K., Zhang, X., Ren, S., & Sun, J. (2015). *Deep Residual Learning for Image Recognition*. Retrieved from https://www.cv-foundation.org/openaccess/content_cvpr_2016/papers/He_Deep_Residual_Learning_CVPR_2016_paper.pdf

Huang, G., Liu, Z., & Weinberger, Kilian Q. (2016). Densely Connected Convolutional Networks. Retrieved from arXiv.org website: https://arxiv.org/abs/1608.06993

Imran, M., & Almusharraf, N. (2024). Google Gemini as a next generation AI educational tool: A review of emerging educational technology. *Smart Learning Environments*, *11*(1). https://doi.org/10.1186/s40561-024-00310-z

Jenny, H., Scott, C., Judy, C., Ann, D., Nicole, H., William, M., … Colorado. (2023). IMPROVE (Interagency Monitoring of Protected Visual Environments): spatial and seasonal patterns and temporal variability of haze and its constituents in the United States: report VI. Retrieved November 6, 2024, from Mountainscholar.org website: https://mountainscholar.org/items/8d422f79-2ae9-4bdb-8171-cdb74d62d2ef

Knyazev, B., Taylor, G. W., & Amer, M. (2019). Understanding Attention and Generalization in Graph Neural Networks. *Advances in Neural Information Processing Systems*, *32*. Retrieved from

https://papers.nips.cc/paper_files/paper/2019/hash/4c5bcfec8584af0d967f1ab10179ca4b-Abstract.html

Kow, P.-Y., Hsia, I-Wen., Chang, L.-C., & Chang, F.-J. (2022). Real-time image-based air quality estimation by deep learning neural networks. *Journal of Environmental Management*, *307*(114560), 114560. https://doi.org/10.1016/j.jenvman.2022.114560

Li, F., & Liu, M. (2018). Alzheimer's disease diagnosis based on multiple cluster dense convolutional networks. *Computerized Medical Imaging and Graphics*, *70*, 101–110. https://doi.org/10.1016/j.compmedimag.2018.09.009

Liang, D., Golan, R., Moutinho, J. L., Chang, H. H., Greenwald, R., Stefanie Ebelt Sarnat, … Sarnat, J. A. (2018). Errors associated with the use of roadside monitoring in the estimation of acute traffic pollutant-related health effects. *Environmental Research*, *165*, 210–219. https://doi.org/10.1016/j.envres.2018.04.013

Lin, C., Gillespie, J., Schuder, M. D., Duberstein, W., Beverland, I. J., & Heal, M. R. (2015). Evaluation and calibration of Aeroqual series 500 portable gas sensors for accurate measurement of ambient ozone and nitrogen dioxide. *Atmospheric Environment*, *100*, 111–116. https://doi.org/10.1016/j.atmosenv.2014.11.002

Liu, Y., Han, T., Ma, S., Zhang, J., Yang, Y., Tian, J., … Ge, B. (2023). Summary of ChatGPT/GPT-4 Research and Perspective Towards the Future of Large Language Models. *ArXiv:2304.01852 [Cs]*. Retrieved from https://arxiv.org/abs/2304.01852

Loomis, D., Grosse, Y., Lauby-Secretan, B., Ghissassi, F. E., Bouvard, V., Benbrahim-Tallaa, L., … Straif, K. (2013). The carcinogenicity of outdoor air pollution. *The Lancet Oncology*, *14*(13), 1262–1263. https://doi.org/10.1016/s1470-2045(13)70487-x

M. Latha, Kumar, P. S., R. Roopa Chandrika, Mahesh, T. R., Kumar, V. V., & Suresh Guluwadi. (2024). Revolutionizing breast ultrasound diagnostics with EfficientNet-B7 and Explainable AI. *BMC Medical Imaging*, *24*(1). https://doi.org/10.1186/s12880-024-01404-3

Meng, X., Yan, X., Zhang, K., Liu, D., Cui, X., Yang, Y., … Guo, Z. (2024). The Application of Large Language Models in Medicine: A Scoping Review. *IScience*, *27*(5), 109713–109713. https://doi.org/10.1016/j.isci.2024.109713

Mosavi, A., Ardabili, S., & Várkonyi-Kóczy, A. R. (2020). List of Deep Learning Models. *Lecture Notes in Networks and Systems*, 202–214. https://doi.org/10.1007/978-3-030-36841-8_20

Nallapati, R., Zhou, B., santos, C. N. dos, Gulcehre, C., & Xiang, B. (2016). Abstractive Text Summarization Using Sequence-to-Sequence RNNs and Beyond. *ArXiv:1602.06023 [Cs]*. Retrieved from https://arxiv.org/abs/1602.06023

Naveed, H., Ullah Khan, A., Qiu, S., Saqib, M., Anwar, S., Usman, M., … Mian, A. (2024). *A Comprehensive Overview of Large Language Models*. Retrieved from https://arxiv.org/pdf/2307.06435

Pang, S., Nol, E., & Heng, K. (2024). ChatGPT-4o for English language teaching and learning: Features, applications, and future prospects. *Cambodian Journal of Educational Research*, *4*(1), 35–56. https://doi.org/10.62037/cjer.2024.04.01.03

Parkhi, O. M., Vedaldi, A., & Zisserman, A. (2015). Deep Face Recognition. *Procedings of the British Machine Vision Conference 2015*. https://doi.org/10.5244/c.29.41

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). *Language Models are Unsupervised Multitask Learners*. Retrieved from https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf

Rajpurkar, P., Irvin, J., Zhu, K., Yang, B., Mehta, H., Duan, T., … Ng, A. Y. (2017). CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning. Retrieved from arXiv.org website: https://arxiv.org/abs/1711.05225

Rane, N., Choudhary, S., & Rane, J. (2024). Gemini Versus ChatGPT: Applications, Performance, Architecture, Capabilities, and Implementation. *Social Science Research Network*. https://doi.org/10.2139/ssrn.4723687

Rangel, A., Raysoni, A. U., Chavez, M. C., Jeon, S., Aguilera, J., Whigham, L. D., & Li, W.-W. (2022). Assessment of traffic-related air pollution (TRAP) at two near-road schools and residence in El Paso, Texas, USA. *Atmospheric Pollution Research*, *13*(2), 101304. https://doi.org/10.1016/j.apr.2021.101304

Sapdo Utomo, Adarsh Rouniyar, Guo Hao Jiang, Chun Hao Chang, Kai Chun Tang, Hsu, H.-C., & Hsiung, P.-A. (2023). Eff-AQI: An Efficient CNN-Based Model for Air Pollution Estimation: A Study Case in India. *Association for Computing Machinery Digital Library* . https://doi.org/10.1145/3582515.3609531

Simonyan, K., & Zisserman, A. (2015, April 10). Very Deep Convolutional Networks for Large-Scale Image Recognition. Retrieved from arXiv.org website: https://arxiv.org/abs/1409.1556

Solimini, A., & Renzi, M. (2017). Association between Air Pollution and Emergency Room Visits for Atrial Fibrillation. *International Journal of Environmental Research and Public Health*, *14*(6), 661. https://doi.org/10.3390/ijerph14060661

Tala Talaei Khoei, Hadjar Ould Slimane, & Naima Kaabouch. (2023). Deep learning: systematic review, models, challenges, and research directions. *Neural Computing and Applications*, *35*. https://doi.org/10.1007/s00521-023-08957-4

Tan, M., & Le, Q. V. (2019). EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. Retrieved from arXiv.org website: https://arxiv.org/abs/1905.11946

Thakur, P. S., Sheorey, T., & Ojha, A. (2022). VGG-ICNN: A Lightweight CNN model for crop disease identification. *Multimedia Tools and Applications*. https://doi.org/10.1007/s11042-022-13144-z

The Claude 3 Model Family: Opus, Sonnet, Haiku Anthropic. (n.d.). Retrieved from https://www-cdn.anthropic.com/de8ba9b01c9ab7cbabf5c33b80b7bbc618857627/Model_Card_Claude_3.pdf

Timlin, A., Hastings, A., & Hardiman, M. (2020). Workbased facilitators as drivers for the development of person-centred cultures: a shared reflection from novice facilitators of person-centred practice. *International Practice Development Journal*, *8*(1), 1–9. https://doi.org/10.19043/ipdj81.008

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., … Lample, G. (2023). *LLaMA: Open and Efficient Foundation Language Models*. Retrieved from https://research.facebook.com/file/1574548786327032/LLaMA--Open-and-Efficient-Foundation-Language-Models.pdf

Xie, P., Zhang, C., Wei, Y., Zhu, R., Chu, Y., Chen, C., … Hu, J. (2024). Status of near-road air quality monitoring stations and data application. *Atmospheric Environment: X*, *23*(100292), 100292. https://doi.org/10.1016/j.aeaoa.2024.100292

Yadav, S. S., & Jadhav, S. M. (2019). Deep convolutional neural network based medical image classification for disease diagnosis. *Journal of Big Data*, *6*(1). https://doi.org/10.1186/s40537-019-0276-2