Pneumonia Detection: Using Data Science to Advance Our Healthcare System
DS4002 Case Study by Kaitlin Blakeslee

Each year, over 2.5 million people die from pneumonia, with the burden especially heavy among the elderly and young children. In the U.S. alone, pneumonia results in over 1 million hospitalizations annually, yet delays in detection remain a persistent challenge. Delays in detection lead to more significant complications and higher risk of mortality. As a rising data scientist, you've been tasked with designing a solution that could transform patient outcomes across the system.

Your mission:
Develop a machine learning model that can detect pneumonia from chest X-rays using real clinical image data. Your team has access to the RSNA Pneumonia Detection Challenge dataset, a widely used public dataset of annotated X-ray images. As artificial intelligence continues to revolutionize healthcare, deep learning tools are becoming essential for fast, scalable, and accurate diagnostics. You will harness this power by implementing a Convolutional Neural Network (CNN), the core technology behind modern image recognition systems used in medical AI. Specifically, you'll fine-tune ResNet50, a deep CNN architecture developed by Microsoft that uses residual connections to train very deep networks without performance degradation. This model is widely used in medical imaging because of its ability to capture complex patterns in visual data. To build your model, you'll apply data augmentation techniques (like rotation, flipping, and zoom) to make your model robust against variability in real-world scans.

But it's not just about building a model. Your challenge is to deliver a high-performance, interpretable model with at least 90% accuracy. You're aiming to achieve a level of precision that could support clinical decision-making. You'll also need to consider other metrics like precision, recall, and AUC to demonstrate real-world effectiveness.

This case study asks you to think like a hospital data scientist. You must weigh technical performance, clinical stakes, and ethical use of AI in medical settings.