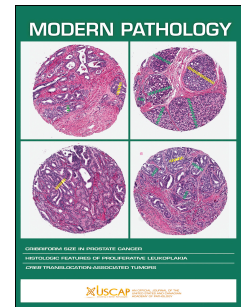


# Journal Pre-proof

Introduction to Artificial Intelligence (AI) and Machine Learning (ML) in Pathology & Medicine: Generative & Non-Generative AI Basics

Hooman H. Rashidi, Joshua Pantanowitz, Mathew Hanna, Ahmad P. Tafti, Parth Sanghani, Adam Buchinsky, Brandon Fennell, Mustafa Deebajah, Sarah Wheeler, Thomas Pearce, Ibrahim Abukhiran, Scott Robertson, Octavia Palmer, Mert Gur, Nam K. Tran, Liron Pantanowitz



PII: S0893-3952(24)00268-0

DOI: <https://doi.org/10.1016/j.modpat.2024.100688>

Reference: MODPAT 100688

To appear in: *Modern Pathology*

Received Date: 16 August 2024

Accepted Date: 27 November 2024

Please cite this article as: Rashidi HH, Pantanowitz J, Hanna M, Tafti AP, Sanghani P, Buchinsky A, Fennell B, Deebajah M, Wheeler S, Pearce T, Abukhiran I, Robertson S, Palmer O, Gur M, Tran NK, Pantanowitz L, Introduction to Artificial Intelligence (AI) and Machine Learning (ML) in Pathology & Medicine: Generative & Non-Generative AI Basics, *Modern Pathology* (2025), doi: <https://doi.org/10.1016/j.modpat.2024.100688>.

This is a PDF file of an article that has undergone enhancements after acceptance, such as the addition of a cover page and metadata, and formatting for readability, but it is not yet the definitive version of record. This version will undergo additional copyediting, typesetting and review before it is published in its final form, but we are providing this version to give early visibility of the article. Please note that, during the production process, errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

© 2024 United States & Canadian Academy of Pathology. Published by Elsevier Inc. All rights are reserved, including those for text and data mining, AI training, and similar technologies.

# Introduction to Artificial Intelligence (AI) and Machine Learning (ML) in Pathology & Medicine: Generative & Non-Generative AI Basics

Hooman H. Rashidi<sup>1,9\*</sup>, Joshua Pantanowitz<sup>2</sup>, Mathew Hanna<sup>1,9</sup>, Ahmad P. Tafti<sup>9</sup>, Parth Sanghani<sup>9</sup>, Adam Buchinsky<sup>9</sup>, Brandon Fennell<sup>8</sup>, Mustafa Deebajah<sup>6</sup>, Sarah Wheeler<sup>1,9</sup>, Thomas Pearce<sup>5,9</sup>, Ibrahim Abukhiran<sup>1,9</sup>, Scott Robertson<sup>6</sup>, Octavia Palmer<sup>1,9</sup>, Mert Gur<sup>3,4</sup>, Nam K. Tran<sup>7</sup>, Liron Pantanowitz<sup>1,9\*</sup>

1. Department of Pathology, University of Pittsburgh Medical Center, PA, USA
2. School of Medicine, University of Pittsburgh, PA, USA
3. Department of Computational and Systems Biology, School of Medicine, University of Pittsburgh, Pittsburgh, PA 15213, USA
4. Department of Mechanical Engineering, Istanbul Technical University, Istanbul, 34437, Turkey
5. Division of Neuropathology, Department of Pathology, University of Pittsburgh School of Medicine and Medical Center
6. Pathology & Laboratory Medicine Institute, Cleveland Clinic, Cleveland, OH
7. Department of Pathology, UC Davis, School of Medicine, Sacramento, CA
8. Department of Medicine, UCSF, School of Medicine, San Francisco, CA
9. Computational Pathology and AI Center of Excellence (CPACE), University of Pittsburgh School of Medicine, Pittsburgh, PA, USA

## **\*Corresponding Authors Details:**

Hooman H. Rashidi  
rashidihh@upmc.edu

Liron Pantanowitz  
pantanowitzl2@upmc.edu

Cancer Pavilion Suite 201  
5150 Centre Ave  
Pittsburgh, PA 15232

**Abstract:**

This manuscript serves as an introduction to a comprehensive seven-part review article series on artificial intelligence (AI) and machine learning (ML) and their current and future influence within pathology and medicine. This introductory review provides a comprehensive grasp of this fast-expanding realm and its potential to transform medical diagnosis, workflow, research, and education. Fundamental terminology employed in AI-ML is covered using an extensive dictionary. The article also provides a broad overview of the main domains in the AI-ML field, encompassing both generative and non-generative (traditional) AI. Thereby serving as a primer to the other six review articles in this series that describe the details about statistics, regulations, bias, ethical dilemmas, and ML-Ops in AI-ML. The intent of these review articles is to better equip individuals who are or will be working in an AI-enabled healthcare system.

**Key words:** Artificial intelligence (AI), machine learning (ML), Generative AI, non-generative AI, predictive analytics, neural network, GPT (generative pretrained transformer), Chat-GPT, DALL-E, stable diffusion, GAN (generative adversarial network), supervised ML, unsupervised ML, reinforcement learning

## Introduction to AI-ML in Pathology & Medicine:

Over the past several decades, the advent of key technology tools such as artificial intelligence (AI) and machine learning (ML) have started to transform the medical arena. In 1950, Alan Turing planted the seeds of this revolution when he developed the Turing Test, a standard test for gauging a machine's capacity for intelligent behavior comparable to that of a person.<sup>1</sup> Since then, there has been an exponential evolution in AI-ML technology, with notable turning points such as Frank Rosenblatt's 1957 creation of the first artificial neural network (ANN) and the 1970s introduction of expert systems. The availability of enormous volumes of digital healthcare data, developments in computer power, and advancements in ML techniques along with the creation of deep neural network (i.e. deep learning or DL) methods have further thrust AI-ML into the forefront of medical research and practice in the twenty-first century. In pathology and medicine today, AI-ML is being increasingly used to enhance diagnostic precision, expedite clinical workflows, customize patient care, and enhance the overall patient experience (**Figure 1**).

These AI-ML platforms are revolutionizing medicine by strengthening medical diagnoses in pathology and other medical disciplines, as well as improving our research and education domains. AI-ML algorithms can now assess large quantities of data (e.g. image, text or tabular data) and achieve unparalleled accuracy and speed in diagnosing various disease processes and to improve operational efficiencies. Specifically, the advances in deep learning-based systems have demonstrated their ability to aid human radiologists and pathologists alike in tasks such as identifying breast cancer within mammograms or detect prostate adenocarcinoma in whole slide imaging (WSI), respectively. Furthermore, clinical decision support systems driven by AI and ML are now capable of analyzing electronic health records (EHRs) and thereby provide individualized treatment suggestions to physicians. AI-ML is also starting to help in the identification of new biomarkers, pharmacological targets, and treatment approaches through the analysis of extensive genomic, transcriptomic, and proteomic data in the research realm.

Besides traditional AI-ML platforms (i.e. non-generative AI), new generative AI capabilities applied to healthcare are also starting to enhance education, diagnostic reporting, searching through large datasets and various other operational needs. These generative AI tools and other advanced technology platforms can now offer interactive simulation-based teaching platforms, tailored learning pathways, as well as virtual patient models and scenarios that accurately replicate real-life clinical situations. Generative AI chatbots and virtual assistants are also beginning to boost patient involvement and empowerment by delivering timely and pertinent health information in more digestible fashion, ultimately improving the overall quality of care. With the continuous increase in volume and intricacy of healthcare data, the need to further integrate such powerful tools in our day-to-day tasks is becoming ever more relevant. The incorporation of such tools within our landscape will not only lead to improved patient outcomes and reduced costs, but hopefully a more enhanced and efficient healthcare system.

This review is part one of our seven-part review series (**Figure 2**) which provides a general introductory overview about the role of AI-ML in pathology and medicine, covering a broad range of essential topics to help all healthcare professionals and trainees navigate this rapidly evolving field. The review begins by clarifying key AI-ML terminology (**Table 1**) and data types

employed within AI-ML platforms, followed by an examination of general (non-generative) ML categories, including supervised, unsupervised, and reinforcement learning studies. The article then distinguishes between generative AI and traditional/non-generative AI platforms, highlighting their similarities and differences. Best practices for designing AI-ML studies are also discussed, as well as common frameworks and libraries used in this field. The review also explores various AI-ML algorithms and neural network models currently being employed in daily practice. Lastly, but most importantly, this article acknowledges the limitations of AI-ML in medicine, including bias, ethical dilemmas, and regulatory challenges, emphasizing the need for a nuanced understanding of these complexities to ensure that there is responsible multi-disciplinary adoption and implementation of such technologies within healthcare.<sup>2-6</sup>

## Basic Concepts and Terminology of AI-ML

### Definition and explanation of key terms (Dictionary of AI-ML terms)

**Table 1** provides a comprehensive glossary of essential terms and definitions along with examples of each, serving as a foundational resource to define commonly used phrases and acronyms within the AI-ML community and tied to key examples within pathology and medicine. The glossary covers a wide range of topics ranging from generative and non-generative AI, to various ML statistical measures to provide readers with a solid foundation of these key terms and concepts that underpin AI-ML applications.

### Data Types in AI-ML

Data types are of utmost importance in the field of AI-ML as they significantly influence the methodology and techniques used. In medicine, the three main data types that we encounter include images (including video), text (including audio), and numerical data, as summarized in **Figure 3**. Each data type requires certain preparation activities and is usually managed by separate AI-ML frameworks. Image data typically employ computer vision applications, such as the analysis of Whole Slide Images (WSIs) in pathology (currently with various proprietary file formats) or the examination of various radiologic images (often in DICOM format). ML methods such as Convolutional Neural Networks (CNNs) are frequently employed for such tasks to ultimately enable the classification of such images (e.g. distinguishing between cancer and normal tissue in pathology images). Textual data on the other hand is commonly processed through the use of Natural Language Processing (NLP) techniques. Nowadays, these techniques primarily utilize transformer-based architectures such as Generative Pretrained Transformer (e.g. GPT4 by OpenAI or Llama3 by Meta) to analyze and generate text responses that resemble realistic human language. Finally, tabular numerical data, which can include continuous and/or discrete values, is a common data type employed in the EHR settings and can be used for various predictive modeling tasks. These tasks include but are not limited to predicting disease states, patient response to therapy or patient outcomes. Traditional ML algorithms such as logistic regression, random forest, or Support Vector Machine (SVM), as well as neural network-based models, are often used within the tabular data task domain.

The selection of an AI-ML framework and preprocessing methods needed is contingent upon the particular data type and problem being addressed. For more detail on this subject, we recommend consulting other review articles on this topic.<sup>7</sup> Irrespective of the data type, it is also crucial to have a comprehensive understanding of various ML categories such as supervised, unsupervised, and reinforcement learning methods. This knowledge is essential as we delve deeper into this field, as these ML categories are applicable to both generative and traditional non-generative AI platforms. Notably, many of the generative AI tools (e.g. ChatGPT/GPT4) that people are now enamored with are a result of a combination of the aforementioned non-generative ML categories (supervised, unsupervised, and reinforcement learning methods). For example, ChatGPT is constructed through a pretraining step (unsupervised/self-supervised), followed by the integration of both a supervised and reinforcement learning approach that enable its seamless chat capabilities (i.e. supervised approach to instruct tune and Reinforcement Learning with Human Feedback or RLHF to do the final fine-tuning).

## Types of Machine Learning: Supervised, Unsupervised and Reinforcement Learning

### Supervised Learning:

In supervised learning (**Figure 4**), machines are trained using "labeled" datasets, where the label provides the "ground truth" for training and/or the "gold-standard". This approach enables the machine to learn from labeled data and make predictions on new, unseen data. Supervised learning can be further divided into two subcategories: classification and regression tasks (see **Table 2**).<sup>8</sup> Classification tasks involve predicting a categorical label or class, such as malignant vs. benign, from the input data. Examples of classification tasks in pathology include image-based diagnosis / classification of tumors based on a WSI's histopathological features. In contrast, regression tasks involve predicting a continuous value or range of values, such as the concentration of a biomarker in blood or the likelihood of disease progression over time. The labeled datasets used in these applications are typically created by human experts (i.e. supervised learning), such as pathologists or clinicians, who annotate the data with the correct diagnoses or classifications. By leveraging these labeled datasets, supervised learning enables machines to learn from human expertise and make accurate predictions on future unseen data, which may ultimately translate to improving diagnostic accuracy and patient outcomes.

### Unsupervised Learning:

In unsupervised learning (**Figure 4**), machines are trained on "unlabeled" datasets, where there is no ground truth or response variable assigned by a human. Some of the most common unsupervised methods used within AI include "Clustering", "Dimensionality Reduction", "Anomaly Detection" and the "Unsupervised part of the Pretraining method" (also known as self-supervised learning) for transformers in Large Language Models (*i.e.* LLMs such as ChatGPT) as shown in **Table 3**. These approaches enable the machine to identify patterns, relationships, and structures within the data without prior knowledge of the expected outcomes. In pathology and medicine, unsupervised learning has numerous applications. For instance, clustering algorithms can be used to group patients based on their gene expression profiles, revealing underlying subtypes of diseases that may not have been apparent through traditional

diagnostic methods. Similarly, dimensionality reduction techniques like principal component analysis (PCA) or t-distributed stochastic neighbor embedding (t-SNE) can be applied to high-dimensional datasets from mass spectrometry-based proteomics, allowing researchers to visualize and identify patterns in protein expression levels associated with specific disease states. Unsupervised learning can also facilitate the discovery of outliers through various anomaly detection methods. By analyzing unlabeled data, unsupervised learning enables machines to uncover hidden insights and relationships that may lead to new diagnostic tools, treatments, and a deeper understanding of disease mechanisms. Additionally, the insights gained from such methods can serve as inputs for supervised learning ML models to enhance their predictive outcomes.

### **Reinforcement Learning (RL):**

Reinforcement Learning (RL) (**Figure 4**) is a form of ML in which an autonomous agent acquires the ability to make decisions through iterative interactions with its environment, where it receives rewards or penalties as feedback for its actions. While the utilization of RL in the field of medicine has been rather restricted, it holds the capability to completely transform healthcare by optimizing treatment approaches and improving the overall quality of care. RL can be utilized to customize cancer treatment strategies by determining the optimal combination of medicines based on each individual patient's distinct attributes and treatment response. In the context of LLMs and chatbots, RL is being utilized through techniques like Reinforcement Learning from Human Feedback (RLHF) and Reinforcement Learning from AI Feedback (RLAF) to optimize a range of LLM models that are being deployed in their fine-tuning stage, amongst others. LLMs, such as but not limited to GPT4, can be refined through the use of Reinforcement Learning from Human Feedback (RLHF) within chatbots like ChatGPT. This involves human assessors offering feedback on the bot's responses, allowing it to acquire knowledge and adjust to user preferences. Similarly, RLAF can be employed to train LLMs in order to produce responses that are more precise and informative by utilizing AI-generated feedback. By utilizing RL, these models can enhance their effectiveness and empathy as conversational partners in healthcare environments, ultimately leading to improved patient engagement and results. The successful incorporation of conventional AI techniques, such as unsupervised learning, supervised instruction-based approaches, and RL-based feedback fine-tuning, into LLM's is crucial for developing generative AI platforms like ChatGPT that exhibit exceptional proficiency and can achieve tasks previously deemed unattainable. Nevertheless, these formidable generative AI platforms are not without their own set of constraints and issues. Hence, gaining a more comprehensive understanding of the similarities and disparities amongst conventional non-generative AI and ML components, as well as recognizing limitations, is crucial.

## **Generative AI versus Traditional ML Platforms**

### **Overview of Generative AI models**

The field of AI has experienced a major boost in light of the rise of generative AI models such as ChatGPT. Leading this charge are LLMs (transformer-based neural networks that can produce language outputs resembling those of humans). These LLMs may also fall under the general umbrella of Foundation Models, which are ML models trained on extensive datasets and capable of performing a wide variety of tasks (natural language processing, computer vision, and other



fields) with minimum need for further adjustment. The transformer architecture, which was introduced in 2017, has played a crucial role in the advancement of LLM foundation models by facilitating parallelization and self-attention mechanisms that enable such models to efficiently handle input text sequences and generate realistic human-like natural language responses.<sup>9</sup> Such development has ultimately led to the creation of extremely popular and capable models as in GPT4 (OpenAI), Claude (Anthropic), Gemini (Alphabet), Mistral (Mistral AI), and Llama (Meta). These models have demonstrated exceptional performance in a range of natural language processing needs, setting new benchmarks in the field and enabling tasks that were previously thought of as improbable to achieve.

In addition to text-based foundation GPT models, other categories of generative AI models are also enabling use of other non-text data types (e.g. image and tabular data) through powerful neural network platforms such as Generative Adversarial Networks (GANs), Variational Autoencoders (VAEs), and Diffusion models.<sup>10</sup> Additionally, the introduction of multimodal models such as GPT4o (able to handle image and text data simultaneously) are now truly revolutionizing the field (e.g. a model that is able to ingest a diagnostic text query and a pathology image and subsequently create a full pathology diagnostic report based on the text combination of that text query input and the contents of the pathology image of interest). Not surprisingly, these incredible models are also saddled with distinct limitations and issues that one needs to be aware of and address before use. Collectively, these models are reshaping our clinical, quality assurance processes, education and research domains and fundamentally changing our workflow.

## Comparison of Generative AI to Traditional Machine Learning Platforms

Generative AI and traditional non-generative AI models share certain similarities, but most importantly they also have distinct differences. As an example, the traditional non-generative AI framework such as in classification models are created to predict outcomes using available data and are commonly assessed using various confusion-matrix-based performance metrics such as but not limited to accuracy, sensitivity, specificity, and ROC-AUC. In contrast, generative AI models, such as LLMs, produce novel and artificial data (i.e. text) that does not possess the same well-defined criteria for evaluation since it usually involves the generation of paragraph outputs instead of making a simple classification call. Additionally, the traditional non-generative AI models are typically dependent on their labeled training data, whereas generative AI models are less constrained and are able to produce new outputs without explicit guidance, as they were inherently designed to do so. Nevertheless, both categories of models are prone to certain key issues which includes bias and ethical constraints. The traditional non-generative AI models have the potential to perpetuate biases found in their original labeled training data while generative AI models can generate biased outputs that can be sometimes challenging to identify and mitigate given our current limited accessibility of these platform's original training data. It is therefore crucial to create novel assessment criteria and protective measures in both approaches to minimize their inappropriate use. Both **Figure 5** and **Table 4** provide an overview of these key similarities and differences. A comprehensive understanding of the individual characteristics in each framework will be crucial for their future responsible and effective implementation and



engagement.

## **AI-ML Tools/Libraries & Algorithms & Neural Networks**

### **Data analysis tools and libraries in generative AI and non-generative AI**

In pathology and various medical AI studies, researchers employ a range of data analysis tools and libraries to extract insights from complex datasets. For generative AI frameworks these include OpenAI's GPT4, Google's Gemini, DALL-E, Stable Diffusion, LangChain, LlamaIndex, AutoGen, and CrewAI that are being increasingly leveraged to generate synthetic images, augment datasets, and develop novel customized chatbots and multimodal models along with multi-agent frameworks. These tools enable researchers to create customized medical foundation models that can serve as virtual assistants for a variety of tasks while also accelerating the development of AI-powered diagnostic tools. In contrast, traditional non-generative AI studies in medicine rely on more established libraries and auto-ML tools like Scikit-learn, PyTorch, TensorFlow, MILO, and STNG to develop predictive models for disease diagnosis, prognosis, and treatment response along with validated synthetic tabular datasets, respectively. These libraries and tools provide the right balance between development and implementation for various ML algorithms, ultimately enabling us to identify patterns that can improve our daily clinical decision support (CDS) workflows. Combining the strengths of both generative and non-generative AI, we are poised to revolutionize the medical field with innovative diagnostic tools, more efficient workflows, and enhanced personalized treatment strategies. However, the engine that runs these tools and libraries is their embedded algorithm or deep neural network model that becomes the true enabler within these frameworks.

### **Algorithms & neural networks in generative AI and non-generative AI**

At the heart of many AI-enabled tools are sophisticated algorithms and neural networks that work in tandem within their associated libraries to help analyze complex datasets and predictive models (e.g. distinguishing tuberculosis positive cases from negative cases in WSI pathology images). For generative AI, classic LLM models (for text generation) such as GPT-4/ChatGPT, Gemini, Llama, or Mistral take advantage of transformer-based neural networks while synthetic image generators such as Stable Diffusion use diffusion-based neural networks, respectively, to accomplish their complex tasks. In addition, Generative Adversarial Networks (GAN) have also been used to generate synthetic images within the generative AI realm. In contrast, traditional non-generative AI studies may also rely on more traditional algorithms besides deep learning options which are typically the gold standard nowadays for image and text analysis (i.e.. use of a non-generative AI convolutional neural network ResNet family of models to classify cancer versus no cancer in images). However, for tabular data non-generative AI tasks (e.g. classification of diabetes versus no diabetes) may also employ non-neural network algorithms which could include random forest (i.e. tree-based), logistic regression, SVM, and KNN (K-Nearest Neighbor) models. The choice of algorithm (neural network versus non-neural network) depends on the task; i.e. image-based studies are most likely to use a deep neural network while tabular data studies are often best suited within non-neural network methods. By combining the

strengths of traditional AI approaches along with the orchestrating capabilities of generative AI models, we can start to leverage the power of both in healthcare.

## **Open-source versus closed-source platforms**

When it comes to AI-ML development in healthcare, the choice between open-source and closed-source AI platforms is an important criterion to consider. Open-source AI frameworks, such as PyTorch, and open-source models such as ResNet50 (amongst others) have shown their capabilities and true worth. Now, with generative AI we are encountering a similar choice to ponder. Although closed source (i.e. proprietary) generative AI models such as GPT4 (i.e. ChatGPT) or Claude have shown to be a revolutionary force, certain open source LLM models are also proving their value and capabilities within this space (e.g. Llama and Mistral family of models). These open-source options offer numerous advantages, including cost-effectiveness, flexibility, and community-driven innovation, especially when they are driven by well-established frameworks such as PyTorch (similar to the non-generative AI ones that we have been used to). Ultimately, the choice between open-source and closed-source AI platforms in healthcare depends on the specific needs and resources of each organization, weighing the benefits of cost, collaboration, security, and customization against the need for support and scalability.

## **Applications of AI-ML in Pathology and Medicine**

### **Image analysis**

The integration of AI and ML in pathology and medicine has revolutionized the field of digital image analysis. Specifically, the application of deep learning-based image analysis tools, that utilize neural network models such as ResNet within stable frameworks as in PyTorch or TensorFlow, has shown tremendous potential for WSI and radiology image analysis classification projects. These AI-ML models can be trained to analyze high-resolution images, enabling accurate detection and classification of various structures and diseases. Other examples include image analysis systems that employ deep learning to provide a differential blood count, automate urine microscopic analysis, or screen liquid-based cytology Pap tests. Deep-learning-enhanced quantitative image analysis (QIA) of biomarkers, for single or multiplex (M-plex) stained samples, also offers a transformative approach to surgical pathology, bringing unprecedented accuracy, efficiency and standardization to the field. Unlike classical image analysis methods that rely on simple thresholding techniques, deep learning models excel in recognizing and quantifying complex patterns within histopathological images. Simple thresholding is limited by its inability to adapt to variations in staining intensity, tissue heterogeneity, and overlapping structures, often resulting in less accurate and less reproducible analyses. In contrast, deep learning models learn from vast datasets to discern subtle differences in tissue architecture and cellular morphology, enabling more precise and reliable identification of pathological features. Moreover, deep learning models can distinguish between subtle features (e.g. invasive vs. in situ malignant components), providing a more detailed and accurate analysis of tumor progression. This level of precision is difficult to achieve with classical image analysis methods, which may struggle with artifacts, overlapping or ambiguous features within the tissue.

By leveraging these advanced image analysis tools, practitioners can improve their diagnostic accuracy, reduce turnaround times, and ultimately enhance patient care.

### **Tabular data analysis**

AI-ML is being increasingly applied to tabular data in pathology and medicine, unlocking valuable insights from within such structured data. Python libraries such as Scikit-learn along with the numerous other readily available algorithms and tools, have enabled the creation of numerous ML supervised classification and regression models using tabular data in order to predict treatment outcomes, disease progression, and patient responses to therapies.<sup>11,12</sup> Use of ML on tabular data within the clinical laboratory can improve quality, workflows, and result interpretation.<sup>13-19</sup> For example, a common source of result errors from blood tests is contamination of blood specimens from intravenous fluids being administered to patients or mislabeled specimens where the blood is drawn from a different patient than listed on the tube's label. This causes erroneous results that may result in significant patient harm. For IV fluid contamination, use of an unsupervised non-linear nearest neighbor-based approach (specifically Uniform Manifold Approximation and Projection) has shown promise in correctly identifying contamination that is missed by current non-ML protocols.<sup>17</sup> Identification of wrong blood in tube errors was assessed using multiple analytes and both logistic regression and SVM models, that significantly outperformed traditional detection methods, though additional work is warranted.<sup>16</sup> Another example was the use of tabular data during the COVID-19 pandemic where automated ML was used to produce algorithms to predict COVID-19 status from mass spectrometry-based proteomic data from anterior nares swab samples.<sup>20</sup> A recent study also showed feasibility of point-of-care AI/ML for use in combat casualty care whereby acute kidney injury can be predicted from a patient's tabular laboratory data findings when using a novel handheld device measuring blood neutrophil gelatinase associated lipocalin levels.<sup>21</sup> As we continue to realize the potential of the application of AI-ML within medicine, it is important that we continue to ensure use of best practices which allow for valid, reproducible models and guidelines for this are emerging.<sup>22-24</sup> By leveraging tabular data analysis, pathologists and medical professionals can move beyond descriptive analytics and into predictive and prescriptive analytics, ultimately enhancing patient care and outcomes.

### **Text data analysis**

In pathology and medicine, AI-ML are starting to get applied to various text data analysis tasks, unlocking insights from unstructured clinical notes, reports, and medical literature. In the past, traditional Natural Language Processing (NLP) techniques based largely on human logic have been used to extract relevant information, classify text, and identifying sentiment. A major drawback of NLP is ambiguity since words strung into sentences may have many meanings depending on the context. There are also cross-cultural challenges with text. However, the emergence of generative AI models is revolutionizing text data analysis in healthcare. These more fluent models can generate contextually appropriate human-like text, summarize complex medical reports, and even create synthetic clinical notes for training and testing purposes. In contrast to traditional NLP, generative AI models can handle ambiguity and uncertainty in clinical text, and learn from limited labeled datasets. For instance, a generative AI model can be

trained to extract relevant information from pathology reports, such as cancer diagnoses and treatment plans, and generate concise summaries for clinicians. Additionally, these models can also assist in generating standardized clinical notes and diagnostic reports, reducing variability, and improving data quality. While traditional NLP techniques are still valuable, the advent of generative AI is poised to transform text data analysis within medicine, enabling more accurate and efficient extraction of insights from our complex unstructured and structured textual clinical data.

### **Multimodal learning**

Modality refers to a single type of (input or output) data (e.g. text, image, video, etc.). Even though single modal AI models are increasingly becoming integrated within our domain, the development of multimodal approaches will undoubtedly further shape the future of AI within medicine. Multimodal learning differs from merely combining different unimodal models that were all trained separately. Instead, this approach simultaneously combines different types of input data to generate content as well as improve performance and accuracy. For example, by leveraging the strengths of various data modalities, such as text data from patient notes and image data from pathology or radiology images, multimodal models can capture a more comprehensive understanding of complex diseases that may not have been as apparent through single-modality approaches. Multimodal models (which include transformers) can be trained from scratch or after fine-tuning. By harnessing the power of multimodal learning, medical professionals can unlock new insights from complex medical data, ultimately enhancing patient care and outcomes.

### **Diversity, Inclusion and Bias**

As AI technology advances, particularly with the advent of generative AI, the importance of diversity, inclusion, and bias mitigation becomes increasingly critical. Generative models like GPT-4o or DALL-E have shown remarkable capabilities in creating content that can mimic human creativity; however, these systems have also shown to have intrinsic bias that can be problematic, especially when applied to healthcare. Such bias can be introduced consciously or unconsciously into the AI lifecycle (e.g., design, development, and deployment stages). As society actively engages in eliminating and rectifying bias in healthcare to achieve health equity, it is imperative to actively assess the performance of AI models for such. The field must have a clear understanding of known health disparities, biases, stereotypes and discriminatory practices and their current impact is imperative and can provide key insights for designing AI models.<sup>25</sup> Creating standardized guidelines for bias detection, assessment, and mitigation should remain a priority.<sup>4,26</sup> Biased data can be informative and used to improve the model's performance.<sup>27,28</sup> Without careful consideration of the training dataset's diversity, AI systems risk perpetuating and amplifying existing societal biases related to gender, race, ethnicity, sexual orientation, disability status, and potentially irreversibly propagating biased healthcare delivery. This can lead to discriminatory outcomes when such technologies are applied in critical areas like healthcare.<sup>6,29-31</sup> Another drawback is that this may limit the generalizability of AI tools when applied to different patient populations. Ensuring that AI systems reflect a diverse range of perspectives is essential for creating equitable technology that serves everyone fairly. By incorporating data

from underrepresented groups and actively seeking to understand the cultural context behind this data, developers can build more inclusive models capable of recognizing and respecting human diversity in their outputs.

## **Explainability in ML and AI Methods**

Explainability is another important aspect of a responsible AI-ML development software life cycle. As the integration of such tools becomes more pronounced, the transparency,, comprehension and accountability of such platforms becomes increasingly vital. Explainable AI (XAI) aims to provide insights into how decisions are made by such tools, helping in identifying and mitigating biases, while also enabling end users to better understand the end results. Techniques such as feature importance analysis, model-agnostic methods like Local Interpretable Model-agnostic Explanations (LIME), and attention mechanisms in neural networks, offer some ways to help within this important process.<sup>32,33</sup> Incorporating explainability into the development of AI algorithms not only reduces the “black box” elements and enhances trust among users and stakeholders, but also facilitates the identification of unintended biases early in the development cycle.

## **Conclusion and Future Directions**

In conclusion, AI-ML applied to healthcare has the potential to revolutionize medical diagnostics, research, and education. However, obstacles such as bias, literacy issues, regulatory hurdles must be continuously addressed as we all start to embrace these platforms within our workflow. The upcoming review articles within this seven-part series will not only help to address many of these obstacles but will also aid in minimizing the AI-ML knowledge gap that currently resides within our healthcare space. As we continue to navigate the intersection of AI-ML and healthcare, it is also our responsibility to not only prioritize collaboration and innovation, but to also establish a responsible development approach as we unlock the full potential of such technology platforms.

**Acknowledgments:** This article series is part of the educational effort of CPACE (Computational Pathology and AI Center of Excellence) at the university of Pittsburgh.

## **Disclosure of Generative AI:**

*During the preparation of this work the authors used DALL-E via ChatGPT-4o, as well as Adobe Express Online, to generate de novo artworks for the creation of certain figures or parts of figures. Additionally, a local large language model was used to generate parts of the initial manuscript bullet outline, but the text content of this manuscript is all human generated and did not include generative AI.*

*Declaration of generative AI and AI-assisted technologies in the writing process:*

*During the preparation of this work the author(s) used [ChatGPT by Open AI] in order to [draft the outline only]. Please also see above about some of the embedded images within our final figures which used DALL-E. After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the publication.*

**Disclosures:**

Liron Pantanowitz is a consultant for Hamamatsu, AiXMed and NTP, serves on the advisory board for Ibex, and is a co-owner of Placenta AI and Lean AP. Hooman Rashidi is a creator of MILO and STNG.

**Author Contributions:** This article series is part of the educational effort of CPACE (Computational Pathology and AI Center of Excellence) at the university of Pittsburgh. The following co-authors (HR, LP, NKT, MG, OP, SR, IA, JP, TP, SW, AT) contributed to the manuscript text contents while HR, LP, TP, BF, PS, AB, MD also contributed to the dictionary of terms (Table 1). All images / figures were constructed by JP.

**Funding:** NA

N/A

**Ethics/Consent:** Not applicable for a review article.



## References:

1. Turing AM. Computing machinery and intelligence. *Mind*. 1950;59(October):433-60.
2. Norori N, Hu Q, Aellen FM, Faraci FD, Tzovara A. Addressing bias in big data and AI for health care: A call for open science. *Patterns (N Y)*. Oct 8 2021;2(10):100347. doi:10.1016/j.patter.2021.100347
3. Char DS, Shah NH, Magnus D. Implementing Machine Learning in Health Care - Addressing Ethical Challenges. *N Engl J Med*. Mar 15 2018;378(11):981-983. doi:10.1056/NEJMp1714229
4. Rajkomar A, Hardt M, Howell MD, Corrado G, Chin MH. Ensuring Fairness in Machine Learning to Advance Health Equity. *Ann Intern Med*. Dec 18 2018;169(12):866-872. doi:10.7326/M18-1990
5. O'Reilly-Shah VN, Gentry KR, Walters AM, Zivot J, Anderson CT, Tighe PJ. Bias and ethical considerations in machine learning and the automation of perioperative risk assessment. *Br J Anaesth*. Dec 2020;125(6):843-846. doi:10.1016/j.bja.2020.07.040
6. Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*. Oct 25 2019;366(6464):447-453. doi:10.1126/science.aax2342
7. Albahra S, Gorbett T, Robertson S, et al. Artificial intelligence and machine learning overview in pathology & laboratory medicine: A general review of data preprocessing and basic supervised concepts. *Semin Diagn Pathol*. Mar 2023;40(2):71-87. doi:10.1053/j.semdp.2023.02.002
8. Rashidi HH, Tran NK, Betts EV, Howell LP, Green R. Artificial Intelligence and Machine Learning in Pathology: The Present Landscape of Supervised Methods. *Acad Pathol*. Jan-Dec 2019;6:2374289519873088. doi:10.1177/2374289519873088
9. Vaswani A, Shazeer N, Parmar N, et al. Attention Is All You Need. 06/12 2017;
10. Pantanowitz J, Manko CD, Pantanowitz L, Rashidi HH. Synthetic Data and its Utility in Pathology and Laboratory Medicine. *Lab Invest*. Jun 24 2024;104(8):102095. doi:10.1016/j.labinv.2024.102095
11. Wertheim GB, Smith C, Lusk M, et al. Validation of DNA methylation to predict outcome in acute myeloid leukemia by use of xMELP. *Clin Chem*. Jan 2015;61(1):249-58. doi:10.1373/clinchem.2014.229781
12. Waljee AK, Sauder K, Patel A, et al. Machine Learning Algorithms for Objective Remission and Clinical Outcomes with Thiopurines. *J Crohns Colitis*. Jul 1 2017;11(7):801-810. doi:10.1093/ecco-jcc/jjx014
13. Herman DS, Rhoads DD, Schulz WL, Durant TJS. Artificial Intelligence and Mapping a New Direction in Laboratory Medicine: A Review. *Clin Chem*. Nov 1 2021;67(11):1466-1482. doi:10.1093/clinchem/hvab165
14. Tran NK, Albahra S, May L, et al. Evolving Applications of Artificial Intelligence and Machine Learning in Infectious Diseases Testing. *Clin Chem*. Dec 30 2021;68(1):125-133. doi:10.1093/clinchem/hvab239
15. Rashidi HH, Bowers KA, Reyes Gil M. Machine learning in the coagulation and hemostasis arena: an overview and evaluation of methods, review of literature, and future directions. *J Thromb Haemost*. Apr 2023;21(4):728-743. doi:10.1016/j.jtha.2022.12.019

16. Rosenbaum MW, Baron JM. Using Machine Learning-Based Multianalyte Delta Checks to Detect Wrong Blood in Tube Errors. *Am J Clin Pathol*. Oct 24 2018;150(6):555-566. doi:10.1093/ajcp/aqy085
17. Spies NC, Hubler Z, Azimi V, et al. Automating the Detection of IV Fluid Contamination Using Unsupervised Machine Learning. *Clin Chem*. Feb 7 2024;70(2):444-452. doi:10.1093/clinchem/hvad207
18. Poole S, Schroeder LF, Shah N. An unsupervised learning method to identify reference intervals from a clinical database. *J Biomed Inform*. Feb 2016;59:276-84. doi:10.1016/j.jbi.2015.12.010
19. Elkhadrawi M, Stevens BA, Wheeler BJ, Akcakaya M, Wheeler S. Machine Learning Classification of False-Positive Human Immunodeficiency Virus Screening Results. *J Pathol Inform*. 2021;12:46. doi:10.4103/jpi.jpi\_7\_21
20. Tran NK, Howard T, Walsh R, et al. Novel application of automated machine learning with MALDI-TOF-MS for rapid high-throughput screening of COVID-19: a proof of concept. *Sci Rep*. Apr 15 2021;11(1):8219. doi:10.1038/s41598-021-87463-w
21. Rashidi HH, Sen S, Palmieri TL, Blackmon T, Wajda J, Tran NK. Early Recognition of Burn- and Trauma-Related Acute Kidney Injury: A Pilot Comparison of Machine Learning Techniques. *Sci Rep*. Jan 14 2020;10(1):205. doi:10.1038/s41598-019-57083-6
22. Norgeot B, Quer G, Beaulieu-Jones BK, et al. Minimum information about clinical artificial intelligence modeling: the MI-CLAIM checklist. *Nat Med*. Sep 2020;26(9):1320-1324. doi:10.1038/s41591-020-1041-y
23. Master SR, Badrick TC, Bietenbeck A, Haymond S. Machine Learning in Laboratory Medicine: Recommendations of the IFCC Working Group. *Clin Chem*. Jul 5 2023;69(7):690-698. doi:10.1093/clinchem/hvad055
24. Haymond S, Master SR. How Can We Ensure Reproducibility and Clinical Translation of Machine Learning Applications in Laboratory Medicine? *Clin Chem*. Mar 4 2022;68(3):392-395. doi:10.1093/clinchem/hvab272
25. Schwartz R, Vassilev A, Greene KK, Perine L, Burt A, Hall P. Towards a Standard for Identifying and Managing Bias in Artificial Intelligence. Special Publication (NIST SP), National Institute of Standards and Technology, Gaithersburg, MD; 2022.
26. Bellamy RKE, Dey K, Hind M, et al. AI Fairness 360: An Extensible Toolkit for Detecting, Understanding, and Mitigating Unwanted Algorithmic Bias. *ArXiv*. 2018;abs/1810.01943
27. Ferryman K, Mackintosh M, Ghassemi M. Considering Biased Data as Informative Artifacts in AI-Assisted Health Care. *N Engl J Med*. Aug 31 2023;389(9):833-838. doi:10.1056/NEJMra2214964
28. Burlina P, Joshi N, Paul W, Pacheco KD, Bressler NM. Addressing Artificial Intelligence Bias in Retinal Diagnostics. *Transl Vis Sci Technol*. Feb 5 2021;10(2):13. doi:10.1167/tvst.10.2.13
29. Raub M. Bots, Bias and Big Data: Artificial Intelligence, Algorithmic Bias and Disparate Impact Liability in Hiring Practices. *Arkansas Law Review*. 2018/12/04/ 2018;71(2):529.
30. Chen Z. Ethics and discrimination in artificial intelligence-enabled recruitment practices. *Humanities and Social Sciences Communications*. 09/13 2023;10doi:10.1057/s41599-023-02079-x
31. Garcia ACB, Garcia MGP, Rigobon R. Algorithmic discrimination in the credit domain: what do we know about it? *AI & SOCIETY*. 2023/05/17 2023;doi:10.1007/s00146-023-01676-3

32. Ribeiro MT, Singh S, Guestrin C. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. presented at: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; 2016; San Francisco, California, USA. <https://doi.org/10.1145/2939672.2939778>
33. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. 2017:618-626.

**Figure Legends:**

**Figure 1.** *Timeline of AI-ML History in Pathology and Medicine.* All artworks were drawn by the authors in combination with generated individual embedded images via DALL-E through ChatGPT- 4o.

**Figure 2.** Your journey through this 7-part review article series.

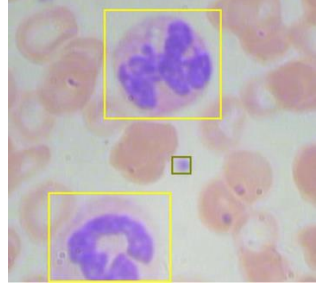
**Figure 3.** *Overview of AI-ML Data Types.* Abbreviations listed above include GPT (Generative Pretrained Transformer), CNN (Convolutional Neural Network) and GAN (Generative Adversarial Networks). All artworks were drawn by the authors in combination with generated individual embedded images via DALL-E through ChatGPT- 4o.

**Figure 4.** *Summary of Machine Learning Categories.* Machine Learning categories include supervised, unsupervised, and reinforcement learning. Semi-supervised learning is where you take advantage of both unsupervised and supervised at the same time by labeling only some of the training data, allowing it to have some human guidance in training but still giving it plenty new unlabeled information to learn from independently. All artworks were drawn by the authors in combination with generated individual embedded images via DALL-E through ChatGPT- 4o.

**Figure 5.** *Comparison of Non-Generative Traditional ML and Generative AI.* The traditional non-generative AI-ML categories include supervised ML (Classification and Regression) along with unsupervised and reinforcement learning. The basis of generative AI is a combination of unsupervised, supervised, and reinforcement learning (i.e., various traditional non-generative machine learning categories are embedded within the generative AI framework).

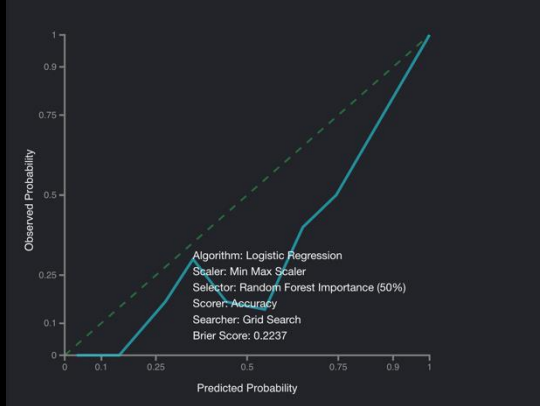
**Table 1: Dictionary of AI-ML commonly used terminology**

| Term                                  | Definition  | Example in Pathology & Medicine   |
|---------------------------------------|---|---|
| Accuracy                              | <p>The proportion of correct predictions (TP + TN) from the total predictions (TP+TN+FP+FN) made by a model.</p> <p>TP: True Positives<br/>TN: True Negatives<br/>FP: False Positives<br/>FN: False Negatives</p>   | <p>In a breast cancer diagnosis model, an accuracy of 95% means that the model correctly diagnosed 95 out of 100 patients (who were either truly positive or truly negative for the disease)</p> <p>Accuracy = (TP+TN) / (TP+TN+FP+FN)</p>  |
| Activation Functions                  | <p>A mathematical function (usually within a neural network) applied to a neuron's output to determine whether that output should be passed to the next layer of the network. It usually introduces non-linearity into the model so that it becomes better able in dealing with more complex tasks.</p> | <p>This Softmax function was used in the output layer of neural networks for the above breast classification task by mapping a vector <math>z</math> (which is usually the weighted sum of inputs plus a bias) to an output (of values to a probability distribution).</p> <p>Common functions include but are not limited to ReLU (Rectified Linear Unit), Softmax, Hyperbolic Tangent (Tanh), and Sigmoid function.</p> |
| Adaboost                              | <p>An ensemble machine learning algorithm that combines multiple weak classifiers to create a stronger classifier.</p>  | <p>Improving cancer diagnosis by boosting the accuracy of weak models that were used in analyzing chemistry laboratory findings.</p>  |
| Adversarial Networks (also see GAN)   | <p>A system of two neural networks where one generates data (generator) and the other evaluates it (discriminator), enhancing data generation, to ensure generated data belong to original data.</p>  | <p>Through the GAN model researchers were able to generate synthetic prostate pathology images that augmented their limited dataset for training their AI model.</p>  |
| Artificial General Intelligence (AGI) | <p>A hypothetical AI system that possesses the ability to understand, learn, and apply knowledge across a wide range of tasks at human-level proficiency. Currently, no such platform exists (but we will likely see this in the near future).</p>  | <p>A future AGI system could potentially diagnose diseases at the same level or better than a human pathologist by analyzing medical images, lab results, and patient histories (i.e. multimodal approach).</p>   |
| Algorithm                             | <p>A finite sequence of well-defined instructions, typically to solve a problem (i.e. set of instructions used to train a model to make predictions).</p>   | <p>A neural network or a Random Forest algorithm can be used for classifying and making predictions for different cancer subtypes.</p>  |
| Annotation                            | <p>The process of labeling data (usually by a subject matter expert) in order to train a machine learning model.</p>  | <p>Marking (annotating) the different cell types in a hematology peripheral blood smear image can be used to train a machine learning model to recognize these different cells in images. The annotated neutrophils are shown below in the yellow bounding box while the platelet is highlighted by the black bounding box.</p>   |

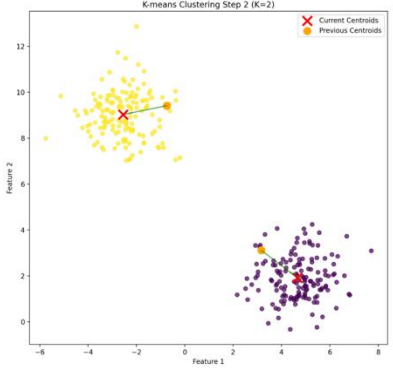
| Term                                 | Definition   | Example in Pathology & Medicine  |
|--------------------------------------|--|--|
|                                      |  |   |
| Anomaly Detection                    | Identifying unusual patterns or outliers in a dataset that do not conform to the expected norm.  | In medical imaging, anomaly detection can be used to detect abnormal patterns in MRI scans.  |
| Artificial Intelligence (AI)         | The development of computer systems that can perform tasks that typically require human intelligence, such as but not limited to visual perception, speech recognition, and decision support.  | A colon cancer image classification AI model is able to make predictions from a patient's pathology H&E images (distinguishing normal colon from cancer as an aid to the pathologist).   |
| Artificial Narrow Intelligence (ANI) | A type of AI designed to perform a specific task, such as cancer prediction, playing chess or recognizing faces. Note: Unlike humans, this AI does not possess the ability to understand, learn, and apply knowledge across a wide range of tasks.   | ANI can be used in a computer vision detection model for identifying specific breast lesions in mammography images.  |
| Attention Mechanism                  | Typically referred to a component in a deep neural network that focuses on specific parts of the input data, ultimately leading to improved model performance.   | This method has become well-known since it is foundational to the new LLM / GPT frameworks where their transformer model's attention mechanism plays a crucial part in encoder and decoder aspects. Ultimately enabling each word in a sentence to attend to every other word, allowing the model to capture word dependencies regardless of their distance in the sequence. |
| Autoencoder                          | A neural network used to compress data into a lower dimension / smaller size (encoding) and then reconstruct it back to the original form (decoding). It is like a zip file: shrinks data down to save space and then expands it back to its original form, preserving as much of the original detail as possible. | Helps reduce the dimensionality of data/information from large pathology images for more efficient storage and analysis.   |
| AutoGen                              | A framework created by Microsoft that enables the use of multiple agents for various LLM applications (i.e. enabling different agents/LLMs to cross talk with each other to solve various tasks).  | A group of LLMs collaborating to streamline a research project (e.g. LLM1/agent1 that gathers manuscripts from PubMed, LLM2/agent2 evaluates the relevance of these manuscripts, and LLM3/agent3 provides a summary of the relevant manuscripts pulled for the given research project).  |
| Automated Machine Learning (Auto-ML) | Automation of the machine learning process, including data preprocessing, feature engineering, and model selection.  | The commercial MILO Auto-ML platform has been used by investigators to quickly and accurately develop predictive models for sepsis prediction in burn patients by automating the analysis of their clinical datasets.  |
| Batch Size                           | Number of samples the ML model processes in each iteration when training a neural network to manage resources and to speed up the process. Like baking   | In deep learning-based image segmentation for tumor detection, a batch size of 32 images is commonly used. i.e. if a model processes 32  |



| Term   | Definition  | Example in Pathology & Medicine   |
|--|---|---|
|  | cookies in batches, the batch size can be important to the final quality of the process.  | pathology slides in a batch, it will adjust its learning after evaluating those 32 images, rather than after each individual image, making the analysis more efficient.   |
| Bagging (Bootstrap Aggregating)              | A technique where multiple models (i.e. ensemble method) are trained on different random samples of data and then combined to improve overall performance. This technique has been shown to improve prediction accuracy and reduce overfitting.   | A Random Forest model (i.e. an ensemble of decision trees where each tree is trained on a different random subset of the data using bagging) can be used to make more accurate disease predictions.   |
| Bayes Theorem                                | A mathematical approach to update the probability of something happening based on new evidence / data.  | If a patient has a positive test result that might indicate sepsis, Bayes' Theorem can help ICU doctors update the likelihood that the patient actually has sepsis, considering how accurate the test is and how common sepsis is given all the other data.   |
| Bias   | A systematic error in a model due to assumptions (such as biased training data), leading to incorrect outcomes. In other words, the model makes systematic errors in one direction due to incorrect or skewed representation assumptions.   | If a machine learning model is trained mostly on data from younger patients, it might systematically misdiagnose older patients, since it has a bias towards data and diagnoses present in the younger age group.   |
| Bias-Variance Tradeoff                       | Since a high bias / low variance can miss important patterns (underfitting) and a high variance / low bias from overtraining can capture too much noise instead of actual trends (overfitting), it is essential to find the right balance between these two zones when training and validating AI models. The ultimate goal here is to enhance the model's ability to generalize to new data. | In predicting patient outcomes, a simple logistic regression model was shown to have a high bias and low variance (underfitted), while a more complex neural network led to a lower bias and higher variance model (overfitted model). In the end, the best (most generalizable) model was shown to be a random forest model that balanced those two. |
| Bing   | A web search engine by Microsoft. Bing also now incorporates LLMs for a chatbot experience in search.   | Using Bing to gather the latest research articles on a topic such as digital pathology with a summary of each.  |
| Bit  | The basic unit of information in most computing systems, representing a binary 0 or 1 value.  | In radiology or pathology imaging, bits are used to represent pixel values in these digital images. For example, an 8-bit image in a radiology image represents 256 different shades of gray in that image ( $2^8 = 256$ possible combinations).  |
| Black Box                                    | Refers to an AI system where the internal workings of a process or system (e.g. decision-making process) is not understood by humans.   | In pathology, the use of a deep learning methods to diagnose cancer within whole slide images can be considered a black box since the inner workings on how their artificial neuronal network derived at such cancer prediction may not be fully understood by humans.  |
| BLEU (Bilingual Evaluation Understudy) Score | Evaluates the precision of n-grams (contiguous sequences of n items from a given sample of text or speech) between generated and reference texts. In other words, it measures the similarity between AI generated text and reference text values, indicating better quality and closer match to the reference.  | Assessing the accuracy of AI-generated pathology reports written in French compared with an equivalent pathology report written in English.   |

| Term              | Definition  | Example in Pathology & Medicine   |
|-------------------|---|---|
| Boosting          | An ensemble machine learning technique that combines the output of several weak models to create a strong model (i.e. improving accuracy).  | For example, in pathology, boosting can be used to enhance the performance of models predicting cancer diagnoses through an algorithm such as Gradient Boosting Machine (GBM), thus improving overall diagnostic accuracy.  |
| Bootstrapping     | A statistical method that involves repeatedly resampling a dataset with replacement to estimate the accuracy of a model (i.e. estimates the variability of a model's performance by repeatedly training and testing it on random subsets of the data which may lead to more generalizable models especially when there is a paucity of data). | Incorporating the bootstrapping technique in a Random Forest machine learning model for breast cancer diagnosis on tabular data allows it to build diverse trees from different subsets of patient data, resulting in a more accurate and reliable overall prediction model.  |
| Brier Score       | Measures the accuracy of probabilistic predictions, calculating the average squared difference between predicted probabilities and actual outcomes. The score ranges from 0 (perfect prediction) to 1 (random guessing).  | When comparing a breast cancer prediction model with a brier score of 0.1 versus another one with a brier score of 0.5, the one with a brier score of 0.1 is very accurate (i.e., if this model predicts that out of 100 patients, 90% are likely to have breast cancer, then about 90 of them actually do), while the one with a brier score of 0.5 is much less reliable from a probability standpoint. |
| Calibration Curve | A graph showing the performance of a model by comparing predicted probabilities to actual probability outcomes.   | <p>The following calibration curve is showing various probability reliability measures at different probability thresholds for this heart disease predictive classification AI model. Note how in this model the probability prediction at 30% threshold is much better than at the 50% threshold.</p>                |
| Categorical Data  | Data that can be divided into discrete groups   | Classifying results into benign or malignant categories in a prostate cancer data cohort.   |
| Causal Inference  | The process of determining whether one thing directly causes another. This is not necessarily the same as correlation, which measures the relationship between two variables, indicating that if they move together or  | If a new drug directly reduces the risk of heart disease for example, rather than just being associated with a reduced risk (possibly due to another confounding factor)  |

| Term   | Definition  | Example in Pathology & Medicine  |
|--|---|--|
|  | not this correlation does not necessarily imply that one causes the other.  |  |
| ChatGPT  | A large language model (LLM) developed by OpenAI, that can generate human-like text responses.  | Using ChatGPT one can get a summary of several kidney disease articles and their highlights, similarities, and differences from each paper summarized nicely within a table format   |
| Classification                                 | The task of assigning labels or categories to data points based on their features.  | Classifying medical images as benign or malignant using a neural network classification machine learning model.  |
| Class Imbalance                                | A situation where some classes (i.e. targets used for the classification task) are underrepresented and others are overrepresented in the dataset.                        | An imbalanced image classification dataset may lead to exaggerated or unrealistic model performance measures (e.g. a 90% model accuracy in a test set with 10% cancer cases and 90% normal cases who's chance/null accuracy is already at 90%, is not as valuable a measure as opposed to a balanced dataset of 50% cancer and 50% normal cases who's 90% accuracy measure becomes a more useful measure to follow since that is being compared to a chance/null accuracy of 50%). |
| Claude   | A family of large language models created by Anthropic, an AI Startup, with similar capabilities to ChatGPT and Gemini.   | Claude can be used to summarize a pathology article that was published so that it can be explained to a lay audience.  |
| CLIP (Contrastive Language-Image Pre-Training) | An AI model developed by OpenAI that learns to understand and relate images and text by training on datasets where images are paired with descriptive text captions.      | For example, in pathology, CLIP can be used to identify and describe abnormalities in pathology H&E images by matching them with relevant textual descriptions.  |
| Closed Model                                   | An AI model whose source code or model architecture is typically not freely and publicly available (generally created by a private company and made available for a fee). | Open AI's ChatGPT, Anthropic's Claude, or Google's Gemini models are all examples of closed models as compared with Meta (Llama) or Mistral's open models.   |
| Clustering                                     | An unsupervised machine learning method that groups similar data points into clusters based on their features.  | Clustering by k-means method of hypothetical diabetic patients, as shown below, based on their responsiveness to therapy reveals two distinct groups/clusters.   |

| Term                               | Definition   | Example in Pathology & Medicine   |  |          |             |             |                    |                     |             |                     |                    |
|------------------------------------|--|---|--|----------|-------------|-------------|--------------------|---------------------|-------------|---------------------|--------------------|
|                                    |  |    |  |          |             |             |                    |                     |             |                     |                    |
| Convolutional Neural Network (CNN) | A type of deep learning model designed especially for recognizing feature patterns in images.  | A prostate cancer CNN model can be used to analyze pathology images to distinguish prostate cancer from normal prostate tissue.   |  |          |             |             |                    |                     |             |                     |                    |
| Confidence Interval (CI)           | A range of values within which a population parameter is likely to lie (a way to compare sample data to true population data). It estimates the true value of a measurement, giving a degree of certainty about where the sample data lies within true values.           | An example would be 95% CI which means that we are 95% certain that the true value lie within this range. For example, if a study finds that a new drug lowers blood pressure by 10-20 points with a 95% CI, it means we are 95% confident within our sample data that the actual reduction is between 10 and 20 points if compared to the entire population.   |  |          |             |             |                    |                     |             |                     |                    |
| Confusion Matrix                   | A table used to evaluate the performance of a classification model by comparing predicted and actual labels.   | <p>In this binary breast cancer prediction model, the confusion matrix shown below can be used to calculate the accuracy, precision, sensitivity, specificity, NPV, and F1 score, etc.</p> <table border="1"> <thead> <tr> <th></th><th>REAL (+)</th><th>RESULTS (-)</th></tr> </thead> <tbody> <tr> <td>Predicted +</td><td>TP (true positive)</td><td>FP (false positive)</td></tr> <tr> <td>Predicted -</td><td>FN (false negative)</td><td>TN (true negative)</td></tr> </tbody> </table> |  | REAL (+) | RESULTS (-) | Predicted + | TP (true positive) | FP (false positive) | Predicted - | FN (false negative) | TN (true negative) |
|                                    | REAL (+)   | RESULTS (-)   |  |          |             |             |                    |                     |             |                     |                    |
| Predicted +                        | TP (true positive)   | FP (false positive)   |  |          |             |             |                    |                     |             |                     |                    |
| Predicted -                        | FN (false negative)  | TN (true negative)  |  |          |             |             |                    |                     |             |                     |                    |
| Correlation                        | A statistical measure of the strength and direction of the linear relationship between two variables. In other words, a measure of how strongly two variables are related to each other.   | A study that showed a correlation between smoking history and lung cancer rates, indicates that as smoking increases, lung cancer rates also tend to increase. These may be represented with "R" (i.e. Pearson Correlation: which ranges from -1 to 1, where 1 means a perfect positive relationship, -1 means a perfect negative relationship, and 0 means no relationship exists).  |  |          |             |             |                    |                     |             |                     |                    |
| Cosine Similarity                  | A measure of similarity between two vectors based on their cosine angle. In other words, how similar two sets of data are by comparing the angle between them, with a value of 1 indicating they are identical and 0 meaning they are completely different/independent.  | When comparing the vector embedding values of the words "king", "queen" and "malignancy", the words "king" and "queen" were shown to have higher cosine similarity scores to each other than to the word "malignancy".  |  |          |             |             |                    |                     |             |                     |                    |
| Cropping                           | The process of selecting a subset of features or data points from a larger dataset for analysis or modeling. In other words, CNN cropping in image analysis refers to cutting out the irrelevant parts of an image to focus on the area of interest for better analysis. | In pathology images, this can be used to extract and analyze specific regions of interest (ROI) in a tissue sample, like a suspected tumor area.  |  |          |             |             |                    |                     |             |                     |                    |

| Term                     | Definition  | Example in Pathology & Medicine  |
|--------------------------|---|--|
| Cross Validation (CV)    | A resampling technique used to evaluate the performance of a model by training and testing it on multiple subsets of data. In machine learning this technique will allow users to test how well a model will perform on new data by dividing the data into different testing parts, training the model on some parts, and testing it on others. | For example, using 10-fold Cross Validation (CV) to evaluate the performance of a machine learning model for predicting patient outcomes in critical care units involves dividing the data into 10 different subsets. The model is trained on 9 subsets and tested on the remaining one, and this process is repeated 10 times, each time with a different subset as the test set. |
| Data Augmentation        | Data Augmentation in machine learning involves creating new training data by making small changes to existing data, helping the model learn better.   | Data augmentation can be applied to create variations in radiology images, like rotating or flipping X-rays, to improve the accuracy of the machine learning model.  |
| Data Preprocessing       | Refers to preparing unclean / raw data for a model by cleaning and transforming it into a usable format for a machine learning platform.  | Preprocessing could involve normalizing patient records, filling in missing values, and converting categorical data like disease types into numerical form so that they become ingestible by the machine learning algorithm.   |
| Decision Tree            | A type of supervised learning algorithm that uses a tree-like structure to classify data points or make predictions.  | A random forest decision tree can be used to distinguish high risk patients from low risk patients (i.e. a simple classification model).   |
| Deep Learning            | A subset of machine learning where computers use many layers of artificial neurons within a neural network to learn complex patterns from large amounts of data. These models have become the gold standard for image and text analysis.  | A deep learning model can be trained to analyze radiology or pathology images to detect abnormalities.   |
| Dependent Variable       | In machine learning this usually refers to the outcome or target variable that the model is trying to predict or explain. Usually performed by using features / patterns from other independent variables.  | An example would be predicting a patient's heart disease risk (i.e. the dependent variable / target) based on their other lab or clinical values such as blood sugar level, blood pressure, heart rate, etc.(i.e. the independent variables).  |
| Dimensionality Reduction | Statistical techniques that can reduce the number of feature dimensions within data (i.e. ability to simplify the data by reducing some feature subsets while still maintaining important information).   | For example, in genomics studies, dimensionality reduction techniques like t-SNE or PCA can reduce thousands of gene expression levels to a few key components. Ultimately, helping to visualize patterns and identify differences between healthy and diseased tissue samples.  |
| Early Stopping           | A technique to prevent overfitting by stopping training when performance on a validation set starts to degrade.   | Ensuring AI models in pathology do not overfit to training data by using early stopping.   |
| Embedding                | Converting any form of data (such as text or images) into a vector representation that can be mapped, transformed and compared in a vector space. Also see "Word Embedding" below.  | Representing patient text data in a pathology diagnostic report as vector embeddings for efficient processing in AI models.  |
| Ensemble Method          | A machine learning technique that combines multiple models to improve prediction accuracy and reduce overfitting.   | A random forest model (an ensemble tree method) can exhibit a high degree of accuracy in predicting the disease state in our patient population.   |

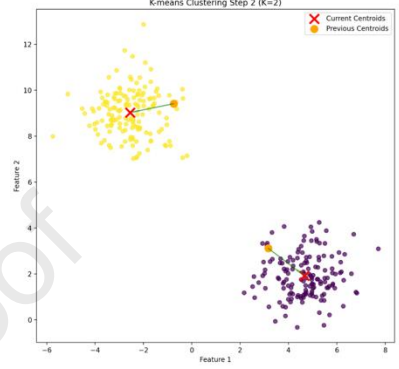


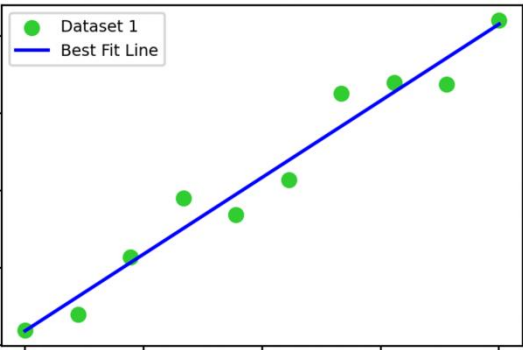
| Term                                 | Definition  | Example in Pathology & Medicine   |
|--------------------------------------|---|---|
| Epoch                                | Refers to one complete pass through the entire training dataset during the learning process (usually a term used in neural networks).   | Trained a deep learning model can involve at least 10 epochs (batches) on a dataset of pathology H&E images.  |
| Expert Systems                       | These are computer programs that mimic the decision-making abilities of a human expert using predefined rules and/or knowledge.   | An expert system can be used to assist doctors by providing diagnostic recommendations based on patient symptoms and laboratory data.   |
| F1 Score                             | The harmonic mean of precision (PPV; positive predictive value) and recall (sensitivity). In summary, a measure that balances the ML model's precision (PPV) and recall (Sensitivity).  | In medicine, an F1 score can be used to evaluate a model that identifies cancer cells, ensuring it accurately finds most cancer cells without including too many false positives. In a breast cancer diagnosis model, an F1 score of 0.9 indicates that the model is making accurate predictions about 90% of the time (balanced between its sensitivity and precision measures). |
| False Negative (FN)                  | A real positive instance that is misclassified/predicted by a model/test as negative.   | A patient with the disease (i.e. Real positive case) is incorrectly diagnosed as healthy (i.e. misclassified as negative) by the ML model   |
| False Positive (FP)                  | A real negative instance that is misclassified/predicted by a model/test as positive.   | A healthy patient (i.e. Real negative case) is incorrectly diagnosed as having the disease (i.e. misclassified as positive) by the ML model   |
| Feature Engineering                  | Creating new input features or modifying existing ones to improve a model's performance measures.   | An example could include combining a patient's weight and height to create a new feature called BMI, which can help a model better predict outcomes.  |
| Feature Extraction                   | Identifying and extracting the most import/relevant features from raw data.   | Extracting key features from an H&E pathology image for tumor classification.   |
| Features                             | The individual input variables (i.e. characteristics) used to train a machine learning model. These may also sometimes be referred to as independent variables (e.g. in a tabular dataset).   | Certain tabular data features in medicine (e.g. gene expression levels and lab findings) can be used to map to a patient's outcomes, creating a machine learning (ML) model that can predict patient outcomes.  |
| Fine-Tuning                          | Adjusting a pre-trained model to better fit a specific task or dataset, by training it on a smaller targeted sample. In other words, making small adjustments to a pre-trained model to improve its performance on a specific task. | Fine-tuning a non-medical general AI model (general image: dog vs wolf vs. cat, etc.) to better recognize specific cancer types in pathology H&E images.  |
| Foundation Model                     | Large, pre-trained AI models that can be adapted for various tasks, usually trained on broad sets of data. Of note, foundation models are large models, but not every large model is a foundational model.                          | The Llama3 foundation model was initially trained on a vast dataset of mostly non-medical texts and later adapted/fine-tuned to better understand medical conditions, with a higher degree of accuracy.   |
| GAN (Generative Adversarial Network) | A type of deep learning machine learning model that consists of two parts, one generating fake data (generator) and the other trying to detect it (discriminator), helping the model to improve over time.                          | GAN approach can be used to generate a large number of synthetic images for training new AI models.   |

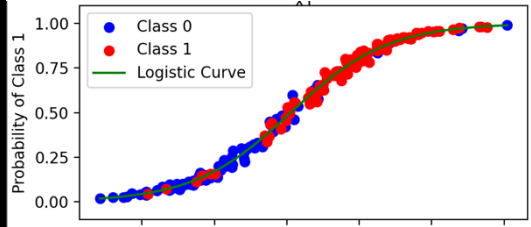


| Term                                     | Definition   | Example in Pathology & Medicine  |
|--|--|--|
| GBM (Gradient Boosting Machine)          | An ensemble machine learning algorithm that typically uses decision trees as its base model. Each tree in the series then focuses on correcting the errors of the previous trees, and together they form a stronger predictive model.  | A GBM can be used to predict the likelihood of readmission to a hospital, by analyzing various health and lab medicine data.   |
| Gemini                                   | A family of closed (proprietary) large language models created by Google. Performs similar functions to ChatGPT.   | Gemini can be used for summarizing and translating several pathology research papers into Spanish.   |
| Generative Pre-trained Transformer (GPT) | A type of large language model that generates human-like text based output (e.g. GPT4 by OpenAI). It uses an attention mechanism, which allows the model to focus on different parts of the input text when generating each word. This helps the model understand the context and relationships between words better, improving the quality and coherence of the generated text. | GPT-4 (i.e. ChatGPT) can be used to provide a bulleted outline for assisting the draft of a medical research paper.  |
| Gemma                                    | A family of open large language models created by Google.  | Gemma can be employed for advanced text analysis of several pathology research papers including their translation from English to Spanish.   |
| Generalizability                         | The ability of a model to perform well on unseen or future data, beyond the training / initial validation test set.  | A generalizable model was shown to not only demonstrate high accuracy on an initial cancer validation test set from one hospital, but also performing well on additional test sets from another hospital system. |
| Generative AI                            | Refers to AI models that create/generate new content, like text, images, or sounds, based on the data they have been trained on.   | Generating synthetic histopathological images can be created through a diffusion model to augment training a machine learning algorithm.   |
| Gradient Descent                         | An optimization method (typically used in neural networks) that helps minimize the loss function in machine learning models, updating weights based on gradients of the error. This helps find the best settings for a model by gradually adjusting them to minimize errors.   | A model with a gradient descent can be used to optimize parameters of its neural network for image classification tasks in a digital pathology study.  |
| Graph Neural Network (GNN)               | A neural network model that processes data structured as graphs, where items are connected by relationships.   | A GNN can be used to predict disease spread through social connections or understanding molecular interactions in a drug discovery study.  |
| Grid Search                              | A technique for hyperparameter tuning by exhaustively searching through a specified subset of hyperparameters.   | A grid search can highlight the best set of hyperparameters for an AI model using laboratory data features to predict a disease.   |
| Ground Truth                             | The actual (true, benchmark or real-world) data or outcome that is the target for training or validating a machine learning algorithm. Ground truth data can be obtained by various means such as labeling datasets, annotation, alternate test results, expert opinion, consensus, among others.  | For an AI algorithm that analyzes digital images of breast carcinoma to score HER2 immunostaining, the ground truth is often represented by the corresponding FISH results for this biomarker.                   |
| Hidden Markov Model (HMM)                | A statistical model that predicts a sequence of hidden states based on observed events.  | HMMs can be used to analyze and predict DNA sequences, identifying patterns and understanding the underlying structure of genomic information.   |

| Term                 | Definition  | Example in Pathology & Medicine   |
|----------------------|---|---|
| Hugging Face         | One of the main hubs for model development, repository and sharing (mostly open models). Typically, the commercial use of which is governed by the creators of each model.  | The Mixtral model (from Mistral) can be downloaded from the Hugging Face site to build a local RAG custom chatbot for medical data use.   |
| Hyperparameter       | The parameters set before training a model, such as learning rate or batch size, that control its performance and behavior.   | Tuning hyperparameters through a grid search and random search approach can enable a model with much improved accuracy for predicting a disease in a subset of patients.  |
| Imbalanced Datasets  | A dataset that has an unequal number of samples in its different classes.   | An imbalanced dataset may become challenging in providing a model/test that performs equally well on both an overrepresented class and underrepresented class. This can also become problematic when designing a model with lab tests since a very low prevalence of a disease state may drastically influence the precision (a prevalence-dependent metric) of a model/test. |
| Imputation           | Filling missing values in a dataset with substituted / inferred values.   | In a breast cancer dataset, the missing values of tumor size can be imputed to provide a complete dataset which can then be used to train a machine learning model.   |
| Independent Variable | A variable that is not dependent on other variables and is used to predict an outcome (or feature).   | In a study on lung cancer, smoking history (independent variable) can be used as one of many variables to predict the risk of developing lung cancer (outcome).   |
| Irreducible Error    | <p>The minimum amount of error that cannot be reduced by improving the model or collecting more data. This is a major part of the basic foundation of supervised machine learning as shown below:</p> $Y = f(X) + e$ <p>Y: output<br/>f: mathematical function/model<br/>X: features / independent variables<br/>e: irreducible error</p> | In a study on prostate cancer diagnosis, the irreducible error in predicting Gleason score using machine learning models may be due to inherent variability in pathology interpretation.  |
| Iterations           | This refers to the repeated cycles of updating a model to improve its performance.  | In a deep learning model for detecting breast cancer from images, iterations can be used to refine the model's performance by adjusting weights and biases.   |
| Kaggle               | A machine learning repository and competition platform with a large source of public domain datasets.   | Kaggle may be used to access various public domain medical datasets, such as segmentation data of specific pathology features.  |
| k-fold               | A cross-validation technique where the dataset is divided into k subsets, with each subset used as a test set once.   | In a study on predicting patient outcomes using electronic health records, a machine learning model can use a 5-fold cross-validation (i.e. randomly divide the dataset into 5 different groups of equal size) to evaluate the model's average performance on the validation test data.   |

| Term                       | Definition   | Example in Pathology & Medicine  |
|----------------------------|--|--|
| K-Means                    | An unsupervised clustering algorithm that groups similar data points into k clusters (see “Clustering” term above).  | <p>K-Means clustering can be used to identify distinct patient clusters/groups. These groups can accordingly be shown to match how these patients responded to certain hypothetical diabetes treatment protocols as shown below.</p>  |
| K-Nearest Neighbors (K-NN) | A supervised learning algorithm that predicts an outcome by finding the k most similar instances in the training dataset.  | In a study on predicting patient response to chemotherapy, K-NN can be used to find patients with similar clinical and genomic characteristics to predict treatment outcomes.  |
| Kernel Function            | A mathematical function that maps input data into a higher-dimensional space for non-linear classification or regression (commonly used in SVM models).  | In a support vector machine (SVM) model for classifying cancer types based on gene expression profiles, the kernel function can be used to transform data into a higher-dimensional space for better separation of classes which translate into a more accurate machine learning model.                                  |
| Keyword Extraction         | A natural language processing technique that identifies important keywords or phrases in unstructured text data.   | In a study on analyzing clinical notes for identifying patients with a certain disease (e.g. chronic renal failure), keyword extraction can be used to identify relevant terms such as "renal failure" and "nephropathy".  |
| Knowledge Graph            | A network of information that links related concepts together, showing how they are connected which may be represented in a graph format.  | A knowledge graph can be used to link symptoms, diseases, and treatments, helping physicians to better understand the relationships of these and make better-informed decisions from this information.   |
| Labeled Data               | Labeling data with a predefined tag (e.g. name, type, diagnosis) is used in the pre-processing step of machine learning that structures the data so that it is meaningful. Such data is used in supervised learning techniques (as opposed to unlabeled data that is employed in unsupervised learning). | Labeling images as cancer cases versus no-cancer pathology diagnosis can be used to train a machine learning model to find this specific diagnosis (cancer vs. no-cancer) in a new image.  |
| Langchain                  | A framework / set of tools for AI developers to interact with LLMs. Langchain does not currently provide its own model but simplifies the process of connecting models to the information they need to provide context, as well as prompting and memory of conversation for the models.                  | Langchain can be used to help build a custom RAG-based medical chatbot framework.  |

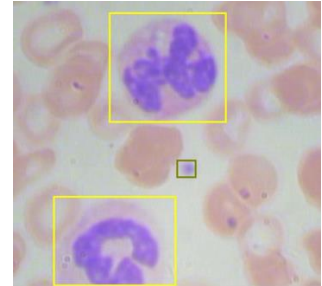
| Term                       | Definition   | Example in Pathology & Medicine  |
|----------------------------|--|--|
| Llama                      | A family of LLMs created by Meta. Some, such as Llama 3, have openly available weights which are now being used relatively freely with limited commercial use.   | The Llama3 model can be finetuned with private medical text data to help build a custom virtual medical transcriptionist chatbot.  |
| Llama Index                | A framework for interacting with LLMs (somewhat similar to what Langchain provides to developers). Llama Index is generally best suited and optimized for retrieving data.   | Llama Index can be used to help build a custom RAG-based medical chatbot.  |
| Large Language Model (LLM) | A deep learning model trained on large amounts of text data to generate language outputs that are coherent and natural-sounding. These are typically based on the transformer neural network architecture  | In a study on generating clinical summaries from EHRs, a LLM can be used to produce concise and accurate summaries of patient encounters.  |
| Learning Rate              | A hyperparameter that controls how much a model's parameters are adjusted during training.   | A suitable learning rate is crucial for training a machine learning model to accurately detect diseases from medical images.   |
| Leave-one-out CV           | A cross validation technique where each instance in the dataset is used as a test set once, with all other instances used for training. Basically, the extreme version of k-fold cross validation  | In a study on predicting cardiovascular disease risk using machine learning models, leave-one-out cross-validation can be used to evaluate the model's performance on individual patients.   |
| Lemmatization              | A natural language processing technique that reduces words to their base or dictionary form (lemma). In other words, the process of reducing words to their base or root form.   | Lemmatization can be used to standardize terms in patient records, like converting "diagnoses" and "diagnosed" to "diagnosis" for more consistent data analysis.   |
| Likelihood                 | A measure of how well a model explains the observed data, with higher values indicating a better fit.  | Likelihood can be used to evaluate how well a model predicts the probability of a patient having a particular disease based on their symptoms or lab test results.   |
| Linear Regression          | A supervised machine learning “regression” approach (i.e. continuous numerical outputs) used to predict the value of one (dependent) variable based on the value of another (independent) variable by using a best fitted line from the data points. | <p>Linear regression can help predict the progression of a disease by analyzing the relationship between the stage of a tumor and the severity of symptoms. The X axis is the dependent variable and the Y axis of the graph (data shown in two dimensions) is the independent variable. The line of regression is the “best fit line”.</p>  |
| Logistic Regression        | A supervised machine learning classification model that functions by predicting the probability from a logistic function with a linear combination of  | Logistic regression could be used to create a model that can help predict kidney injury (0 represents negative for injury and 1 is positive for injury)  |

| Term                                  | Definition  | Example in Pathology & Medicine   |
|---------------------------------------|---|---|
|                                       | independent variables. Probability of the outcome is represented by a logistic function (usually a sigmoid curve that ranges from values 0 to 1).   | status based on factors such as biomarkers, creatinine, and urine output.<br>   |
| Log Loss                              | Measures how well a model's predicted probabilities match actual outcomes, with lower values indicating better performance.   | As a model improves in being able to predict patient complications during a medical procedure, the log-loss of the model should decrease.   |
| Machine Learning (ML)                 | A subfield of AI that involves training models on data to make predictions or decisions without being explicitly programmed.  | For predicting cancer diagnosis from pathology images, machine learning (ML) algorithms can be used to classify tumors as benign or malignant based on their H&E image features.  |
| Mathews Correlation Coefficient (MCC) | A balanced statistical metric used to measure the quality of binary classifications. Ranges between -1 and 1. 1 represents perfect accuracy while -1 perfect inaccuracy. 0 represents no relation. It is also known as Phi coefficient.   | For a machine learning model created to predict whether a patient will develop a certain disease, if the model is evaluated to have an MCC of 0.95, it would be reasonable to assume it has high performance.   |
| Mean Absolute Error (MAE)             | A loss function representing the mean of the absolute values of errors. In other words, MAE calculates the average of the absolute differences between predicted and actual values. A common statistical performance metric in regression models.                                   | For a model that is created to estimate a clinical finding, MAE can be used to judge how close to the real value the estimates tend to be. A MAE score below 10% is good, whereas a score above 50% implies the model is inaccurate (i.e. it is wrong more than it is correct). |
| Mean Squared Error (MSE)              | A loss function representing the mean of the squares of errors. This metric is more effected by rare large errors than Mean Absolute Value. A common statistical performance metric in regression models.   | Mean squared error may be used as a loss function in training a regression model to estimate a clinical finding. The smaller the MSE, the better is the predictive accuracy of the model for this clinical finding.   |
| Metadata                              | This refers to additional data that provides further information (descriptions, attributes) about specific data.  | Pathology images along with their metadata (e.g. patient demographics, diagnosis, etc.) can be used to train machine learning algorithms.   |
| Mistral                               | An AI company specialized in creating LLMs. Mistral has produced models that are open to the public (e.g. Mistral 7B, Mistral 8x7B) and pre-trained closed models (e.g. Mistral Small, Large, Embed) that are only available through their Application Programming Interface (API). | Mistral products may be used to provide LLM capabilities without the need to make an enormous investment necessary to train an LLM from scratch (e.g. a custom chatbot).  |
| Mixtral                               | A mixture-of-experts LLM framework created by Mistral AI. The model and associated trained weights have been made publicly available.   | Mixtral may be used as a base to create a custom chatbot (e.g. RAG model &/or a fine-tuned model).  |

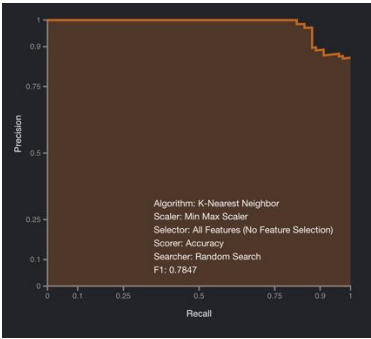


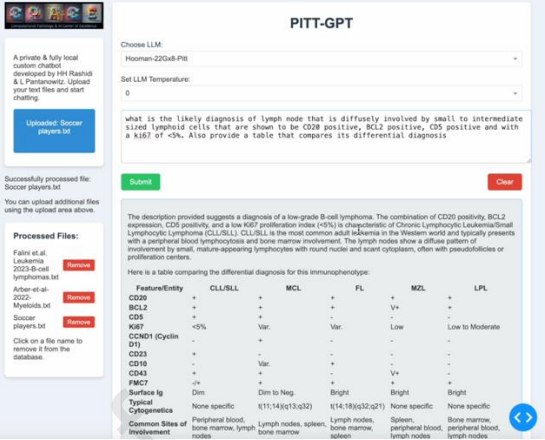
| Term                                 | Definition   | Example in Pathology & Medicine  |
|--------------------------------------|--|--|
| MLP                                  | See Multi-Layer Perceptron.  | -  |
| Model                                | A mathematical representation of a system or process that is trained on data to make predictions or decisions.   | In a study on predicting patient outcomes using EHRs, a machine learning “model” can be developed to predict the risk of hospital readmission based on clinical and demographic variables.   |
| Multicollinearity                    | A situation where two or more independent variables are highly correlated with each other.   | In a study on predicting cardiovascular disease risk using multiple biomarkers, multicollinearity between biomarkers may lead to a less capable machine learning model. For example, the high correlation between hemoglobin and hematocrit may adversely influence the outcome prediction of their associated model (especially if there are a limited number of independent variables assessed which can then further exaggerate the effect of both hemoglobin and hematocrit on the target and dilute the relationship of the other variables). |
| Multi-Layer Perceptron (MLP)         | A type of neural network with multiple layers of artificial neurons that can learn complex patterns in data.   | A MLP model can be built for predicting acute kidney injury based on urine output, biomarkers and Creatinine measures which can thus capture the non-linear relationships between these variables.   |
| Multimodal Model                     | A model that uses multiple types of data such as tabular data, text, images, video, and/or audio.  | ChatGPT’s multimodal GPT4o (i.e. omni) can use multiple modalities, such as both image and text to answer elaborate questions.   |
| Multi Agent Framework                | A set of specialized intelligent agents, normally LLMs, that work together to produce better results than a singular generalized agent would.                                      | A system created to evaluate patient histories for potentially undiagnosed conditions could have an agent to analyze visit notes, an agent to analyze tests and scans, and an agent to combine all assessments for individual fields of medicine. These agents would all communicate to further improve the diagnostic response. AutoGen and CrewAI are commonly used Multi Agent Frameworks for such tasks.   |
| Naïve Bayes                          | A simple and fast algorithm (devoid of hyperparameters) that uses probability to classify data, assuming the simple notion that the features mapped are independent of each other. | The Naïve Bayes model can be used to predict patient response to chemotherapy using clinical and genomic data. The model can thereby be used to classify patients into responders and non-responders based on probability estimates.   |
| Narrow AI                            | See Artificial Narrow Intelligence (ANI)   | -  |
| Natural Language Processing (NLP)    | A subfield of AI that involves developing algorithms and models for processing, understanding, and generating human language.  | In a study on analyzing clinical notes for identifying patients with chronic kidney disease, NLP techniques can be used to extract relevant information from unstructured text data/notes.   |
| Natural Language Understanding (NLU) | A sub-category of Natural Language Processing that enables a machine to understand and interpret human language.   | NLU can help analyze doctor-patient conversations to extract key information (i.e. semantics of the text), such as symptoms and medical histories, for better diagnosis and treatment planning.  |

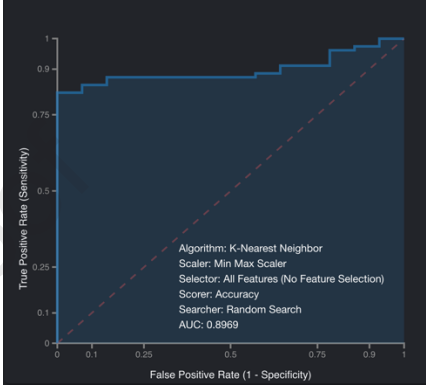


| Term                            | Definition  | Example in Pathology & Medicine  |
|---------------------------------|---|--|
| Negative Predictive Value (NPV) | The proportion of true negatives among all negative predictions made by a model. A prevalence-dependent confusion matrix-based metric.  | A model that has a NPV of 99.7% signifies a reliable call on negative cases (i.e. when the model predicts a negative case, the likelihood of it being truly negative is very high).  |
| Neural Network                  | A type of machine learning model inspired by the structure and function of the human brain/neurons, composed of interconnected artificial neurons, usually comprised of an input, output and hidden layers (although more complex structures also exist). | In a study on image segmentation for tumor detection in medical images, a deep neural network can be trained to identify boundaries between tumors and normal tissue. Because the network can learn and model the relationship between such complex pathology data, the machine learning system can make this intelligent decision with limited human help.                            |
| Normalization                   | The process of scaling numerical data to a common range (e.g., 0-1) to prevent features with larger values from dominating smaller ones within the final model.   | In a study on predicting a disease such as sepsis, normalization can be applied to variables/features such as blood pressure and all lab results to ensure equal weighting of these in the machine learning model.   |
| Object Detection                | A computer vision task that involves identifying and locating objects within images or videos.  | <p>The following object detection model (shown below) is able to identify different white blood cell types and platelets (in this case the objects detected were the 2 neutrophils &amp; the platelet in the center).</p>    |
| One-Hot Encoding                | A technique in which a mutually exclusive categorical variable with more than 2 possible values is encoded as a series of binary variables, one for each possible value.  | A single blood pressure (BP) column that has the following 3 text values within it (low, normal and high) can be converted into 3 separate binary (0s and 1s) columns (i.e. for the new BP-low column the BP-low positive values are shown as 1s, while the non-BP-low values are shown as 0s; the same concept is then repeated for the other two BP columns, BP-normal and BP-high). |
| Open-CLIP                       | Open source Contrastive Language-Image Pre-Training is an open-source implementation of Open AI's CLIP. See CLIP for more details.  | Open-CLIP can be used to create captions for medical images, like describing the findings in an image, assisting doctors in their diagnoses.   |
| Open Model                      | An open-source AI model such as the Llama or Mistral models with the source code, model architecture, and model weights publicly available. Commercial use of these models is determined by each creator.   | A pathology lab could use Meta's Llama models for pathology report generation rather than relying on the paid version of OpenAI's GPT4.  |

| Term                                  | Definition   | Example in Pathology & Medicine  |
|---------------------------------------|--|--|
| Optimizers                            | Algorithms that adjust the model's parameters to minimize errors and improve accuracy during training. Examples include but are not limited to Stochastic Gradient Descent (SGD) and Adam (Adaptive Moment Estimation).  | In a deep learning model for image classification in pathology images, optimizers such as SGD or Adam can be used to update model weights during training.   |
| Outliers                              | A datapoint that varies significantly from other datapoints. These may occur through some error in measurement or as a natural deviation caused by random chance.  | In general, white blood cell count varies between 4,000 and 11,000 cells. An outlier may occur in a sample where the count is 7. This could be due to issues in testing or perhaps transcription of the result.  |
| Overfitting                           | A situation where a model is too complex and performs well on the training data (i.e. initial validation test) but poorly on a new, unseen generalization test dataset.  | A sepsis model in hospital X found to be overfitted could result in it performing poorly on all other secondary and tertiary testing data (i.e. it is not generalizable).  |
| Over-Sampling                         | A data augmentation technique to address class imbalance where samples of data from the minority class are duplicated to even out the distribution.  | For example, to train a model to classify between patients with or without irregular breathing patterns, the raw dataset contains 20% irregular patterns and 80% regular. In order to deal with complications that may occur from using such imbalanced data, oversampling can be used to bring the balance between these two back to 50% (i.e. balanced dataset). |
| Parameter Enhanced Fine Tuning (PEFT) | A method for fine-tuning pre-trained language models by adjusting model parameters using task-specific data. An example of this is Low-Rank Adaptation (LoRA) which is useful for adapting large models to specific tasks without needing extensive computational resources. | In a study on generating clinical summaries from EHRs, PEFT can be used to adapt a pre-trained language model to the specific task of summarization in healthcare through the LoRA method.   |
| Parametric Algorithm                  | A type of algorithm that uses fixed parameters to make predictions or decisions (e.g. logistic regression), as opposed to non-parametric algorithms (e.g. decision trees or KNN) that do not rely on specific distributions or assumptions.                                  | In a study on predicting cardiovascular disease risk the parametric algorithm logistic regression can be used to predict disease and estimate its probability based on the clinical and demographic variables.   |
| PCA (Principal Component Analysis)    | A dimensionality reduction technique that transforms high-dimensional data into lower-dimensional representations while trying to retain most of the useful information.   | In a study on analyzing gene expression profiles in cancer, PCA can be used to reduce the dimensionality of data and identify key genes associated with different cancer subtypes.   |
| Perceptron                            | A type of artificial neuron that serves as a building block for neural networks. It can perform binary classifications by mapping input features to output decisions based on weights and biases.  | Perceptron can be used in the classification layer of developing neural networks for simple tasks such as distinguishing between cancerous and non-cancerous cells in pathology slides.  |
| Perplexity Score                      | Perplexity Score in natural language processing is a way to capture the degree of 'uncertainty' a model has in predicting text. Lower scores indicate better predictions.  | In medicine, a perplexity score could evaluate how accurately a language model summarizes patient notes, with a lower score suggesting more accurate summaries.  |

| Term                                 | Definition  | Example in Pathology & Medicine  |
|--------------------------------------|---|--|
| Pixel                                | The smallest unit of an image or picture, representing a single color value.  | In a study on image analysis for tumor detection in medical images, pixels are analyzed to identify patterns and features associated with malignant lesions.   |
| Pooling Layer                        | A type of layer in neural networks that reduces spatial dimensions to capture larger contextual information.  | In a convolutional neural network (CNN) for image classification in pathology, pooling layers are used to downsample the feature maps and reduce overfitting.  |
| Positive Predictive Value (PPV)      | Also known as Precision. The proportion of true positives among all positive predictions made by a model. A prevalence-dependent confusion matrix-based metric.   | In a cancer diagnosis model, if the PPV is 0.9, this means that out of 100 patients predicted to have cancer, 90 will have confirmed cancer and 10 will not.   |
| Precision                            | See Positive Predictive Value (PPV)   |  |
| Precision-Recall Curve (PR curve)    | A graphical representation of a model's performance, showing the trade-off between precision and recall across different threshold settings. A way to optimize the performance measure of a classification model.   | <p>In a model diagnosing heart disease, the PR curve shown below helps to fine tune the model's ability to correctly identify disease (recall) while minimizing false positives (precision).</p>  <p>Algorithm: K-Nearest Neighbor<br/> Scaler: Min-Max Scaler<br/> Selector: All Features (No Feature Selection)<br/> Scorer: Accuracy<br/> Searcher: Random Search<br/> F1: 0.7847</p> |
| PyTorch                              | PyTorch is an open-source machine learning library used for applications such as computer vision and natural language processing.   | PyTorch can be utilized to develop and train deep learning models for classifying or segmenting histopathological images to identify regions affected by disease.  |
| Quantization                         | A process that reduces the quantity of data needed to represent a model's parameters, thereby decreasing model size and increasing prediction speed, often with minimal loss in model performance. In other words, reducing the dataset can reduce the computation demand without sacrificing the overall accuracy. | The Llama3 Q6 quantized model can perform well in a RAG-based custom chatbot, increasing its speed for generating responses while maintaining relative accuracy of the original unquantized model.   |
| R2 (R-squared)                       | A statistical measure that indicates how well a regression model explains variability in data. The square of Pearson's R (see also Correlation)   | An R2 value of 0.7 in a model predicting patient outcomes based on lab results means that 70% of the variation in outcomes can be explained by the lab results.  |
| RAG (Retrieval Augmented Generation) | Combines the strengths of information retrieval and natural language generation by first finding relevant data (i.e. local/internal or targeted external data) from a database of interest and then using that data generates coherent and more accurate responses. RAG   | An example of a RAG-based custom chatbot is shown below that is retrieving information about lymphoma only from the uploaded manuscripts shown on the left side (i.e. Processed files).  |

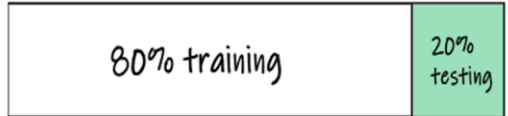
| Term  | Definition   | Example in Pathology & Medicine  |
|---|--|--|
|   | can optimize the output of a Large Language Model (especially when dealing with a targeted or local data).   |    |
| Random Forest                                     | An ensemble tree-based algorithm that combines multiple random decision trees to improve prediction accuracy and reduce overfitting.   | A random forest model can be used to predict a disease, by combining data from electronic health records and lab results.  |
| Recall  | Also known as sensitivity, the proportion of true positive (TP) predictions out of all actual positive instances (TP+FN). See Sensitivity  | In a cancer detection model, a high recall means that the model successfully identifies most of the actual cancer cases (i.e. positive cases), which is crucial for early detection and treatment. A sign of a good screening tool.  |
| Recurrent Neural Network (RNN)                    | A type of neural network designed for sequential data, where connections between nodes form directed cycles, allowing the network to maintain a memory of previous inputs.   | A RNN model can be used to predict the progression of a disease such as diabetes in a patient cohort. By analyzing their sequential medical records, including blood sugar levels, insulin doses, and lifestyle factors over time, the RNN can forecast future blood sugar levels and potential complications. |
| Regression  | A type of supervised learning where the goal is to predict a continuous numerical output variable based on one or more input features.   | A linear regression model can be used to predict an output such as blood pressure levels based on patient age, weight, and medication use.   |
| Regularization                                    | Technique used to prevent overfitting by adding penalties to the loss function for complex models. These include but are not limited to L1 Regularization (Lasso), L2 Regularization (Ridge), or Dropout (temporarily removes a random subset of neurons during training). | L1 regularization (Lasso) can be applied to a logistic regression model predicting cancer risk from genomic data to reduce the impact of noisy or irrelevant features.   |
| Reinforcement Learning                            | A type of machine learning approach where an agent learns to make decisions by taking actions in an environment to maximize cumulative reward.   | Reinforcement learning can be used to optimize Large Language Models that can result in more accurate report generation and summarization, i.e. Reinforcement Learning with Human Feedback (RLHF).   |
| Reinforcement Learning with Human Feedback (RLHF) | A reinforcement learning approach where human feedback is used to guide and improve the learning process of an AI agent such as a Large Language Model.  | RLHF can be used to improve the performance and user friendliness of the ChatGPT framework Large Language Models.  |

| Term   | Definition  | Example in Pathology & Medicine  |
|--|---|--|
| Reinforcement Learning with AI Feedback (RLAF)                       | A reinforcement learning approach where feedback from other AI models is used to guide and improve the learning process of the AI agent.  | RLAF can be used to train Large Language Models to increase the performance in analyzing articles.   |
| Receiver Operating Characteristic (ROC) / Area Under the Curve (AUC) | The Receiver Operating Characteristic (ROC) curve is a global performance metric based on confusion matrix results from multiple thresholds that plots true positive rates (i.e. Sensitivity) against false positive rates (i.e. 1 - Specificity) at different thresholds, with the calculated Area Under the Curve (AUC) representing the overall model performance. For a two-class (i.e. binary) classification task, an AUC of 0.5 indicates chance performance, and an AUC of 1.0 indicates perfect performance at discriminating the two classes. | <p>The ROC-AUC curve for a disease prediction model shows an AUC of ~0.9, indicating good discrimination between disease groups.</p>  |
| Root-Mean-Square Error (RMSE)  | A statistical metric for regression models. This measures the accuracy of a model by calculating the square root of the average squared differences between the predicted values and the actual values, with lower values indicating better performance.  | A RMSE of 2.5 mmHg for a blood pressure prediction model indicates that, on average, the predicted values are 2.5 mmHg away from actual values.  |
| ROUGE Score  | A set of metrics used in LLM assessment to evaluate the similarity between a candidate document and a collection of reference documents, focusing on recall-oriented measures.  | ROUGE scores can assess the performance of a model generating pathology reports.   |
| Segment Anything Model (SAM)   | A computer vision model by Meta that uses deep learning to augment image or video segmentation into objects of interest.  | SAM can be utilized for segmenting regions of interest in pathology images, such as tumors or cells.   |
| Sample   | A subset of data used to train or evaluate a machine learning model.  | A sample of 100 patient records can be randomly selected from a larger dataset to train a machine learning model for disease diagnosis.  |
| Scaling  | The process of transforming numerical features into a common range, often between 0 and 1, to improve model performance.  | Feature scaling can be applied to lab result data (e.g., glucose levels) in a diabetes prediction model to ensure that all features are treated similarly by the algorithm.  |
| Secondary/Tertiary Test Sets   | Additional datasets used to evaluate a machine learning model's performance after training and hyperparameter tuning on primary test sets. A way to confirm model generalizability.   | A secondary test set of 1,500 patient records can be used to further evaluate the performance of a trained cancer diagnosis model before deploying it in clinical practice.  |
| Self-Supervised Learning   | A machine learning approach where the model learns by predicting parts of the input data from other parts, using the data itself as supervision without explicit labels.  | Self-supervised learning can be used to pre-train models on large datasets of unlabeled medical images, improving performance.   |



| Term                         | Definition  | Example in Pathology & Medicine   |
|------------------------------|---|---|
| Semi-Supervised Learning     | A learning approach that combines a small amount of labeled data with a large amount of unlabeled data during training, leveraging both to improve model performance.   | Semi-supervised learning can enhance the performance of classification models in pathology by using a few labeled pathology slides along with many unlabeled ones, reducing the need for extensive manual labeling.   |
| Sensitivity                  | The proportion of true positives among all actual positive instances (e.g., patients with disease). Good screening performance measure.   | A sensitivity of 0.95 for a disease detection model means that 95% of patients who actually have this disease are correctly identified by the model.  |
| Sentiment Analysis           | A type of natural language processing that determines the emotional tone or sentiment behind text data, such as patient feedback or online reviews.   | Sentiment analysis can be applied to online reviews of hospitals to identify areas for improvement in patient satisfaction and care quality.  |
| Specificity                  | Also known as true negative rate, this is the proportion of true negatives among all actual negative instances (e.g., patients without disease).  | A specificity of 0.98 for a cancer screening model means that 98% of patients who do not have cancer are correctly identified as such by the model.   |
| Standardization              | The process of transforming numerical features to have zero mean and unit variance, often used in conjunction with scaling.   | A standard scaler can be applied to a dataset so that all features follow similar scales (i.e. mean of 0 and standard deviation of 1). This assures training of a better performing model.  |
| Stemming                     | A technique used in natural language processing to reduce words to their base or root form by truncating common endings such as “ed” or “ing,” ignoring grammatical variations.                                   | Within clinical notes, words like "diagnoses", "diagnosed" and "diagnosing" are reduced to their root form "diagnosis". This reduces the number of unique terms, and enhances the accuracy and efficiency of modeling algorithms to identify common themes in patient records, such as symptoms and treatment outcomes. |
| Stop Words                   | Common words (e.g., "the," "and") that are ignored in natural language processing tasks due to their lack of meaningful content.  | Stop words can be removed from patient feedback text data before applying sentiment analysis to focus on more informative keywords.   |
| Structured Data              | Structured data is data that has a standardized format for efficient access, often in “tabular” form.   | Structured data in pathology includes lab results, clinical data, or any other data that has been purposefully organized, as in an excel file.  |
| Supervised Learning          | A type of machine learning where the model is trained on “labeled data”, with the goal of making predictions on new, unseen data. The two main types are classification and regression.                           | A supervised learning approach can be used to train a deep learning model for medical image classification, using annotated images as training data.  |
| Support-Vector Machine (SVM) | A type of supervised learning algorithm that finds the decision boundary between classes by maximizing the margin between them.   | A SVM can be used to classify patients into high- or low-risk groups based on their electronic health record data and lab results.  |
| Synthetic Data               | Artificially generated “new” data that mimics real-world data, often used for training machine learning models when real data is scarce or sensitive. These can be synthetic (fake) images, text or tabular data. | A synthetic image dataset can be generated to help to augment training a machine learning model of a rare disease where there are limited images of original cases.   |
| Target                       | The output (dependent) variable or response being predicted by a machine learning model.  | In a cardiovascular disease prediction model for patients hospitalized for acute coronary syndrome, a target variable could be risk of recurrent myocardial infarction in the next 30 days. Another   |



| Term              | Definition  | Example in Pathology & Medicine  |
|-------------------|---|--|
|                   |   | example is the presence or absence of cancer within pathology images.  |
| Tensor            | A multi-dimensional array of numerical values, often used to represent data in deep learning models. Note: tensor and vectors are related. A vector is a one-dimensional array of numerical values, while a tensor is a generalized form of a vector that can exist in more than one dimension. | Tensors can be used to represent medical imaging data (e.g., MRI scans) in a convolutional neural network for lesion segmentation.   |
| TensorFlow        | An open-source machine learning library developed by Google, which is used for building and training machine learning models and deep learning models.  | TensorFlow can be used to develop deep learning models for analyzing medical images, such as detecting cancerous cells in histopathology slides.   |
| Test Set          | A dataset used to evaluate the performance of a machine learning model after training and hyperparameter tuning. It is typically a held-out dataset that is separate from the training and validation datasets to provide an unbiased assessment of the model's generalization ability.         | A test set can be used to evaluate the final performance of a trained model before deploying it in clinical practice.  |
| Threshold         | A value that determines when a prediction is considered positive or negative, often adjusted for optimal trade-offs between sensitivity and specificity.  | The threshold for a disease detection model can be adjusted, such as from 0.5 to 0.3 to increase its sensitivity (i.e., able to better catch true positives).  |
| Tokenization      | The process of breaking down text data into individual units (tokens) which may be words or parts of words, often used in natural language processing tasks.  | Tokenization can be applied to clinical notes text data before applying topic modeling to identify key themes and concepts.  |
| Training Set      | A dataset used to train a machine learning model.   | A training dataset can be used to train a model as part of a supervised (using a labeled set) or unsupervised (using unlabeled data) machine learning approach.  |
| Train-Test Split  | This is the foundation of all supervised machine learning models. The process of dividing a dataset into separate training and testing sets, often with a ratio of 80:20 (80% training : 20% initial validation test) or 70:30 (70% training : 30% initial validation test).                    | <p>A 80-20 train-test split can be applied to a dataset of electronic health records of 1,000 patients, with 800 patient records used for training and 200 for testing a disease prediction model.</p>  |
| Transfer Learning | The process of using pre-trained models (e.g. Resnet50) as a starting point for new machine learning tasks, often leveraging knowledge learned from similar datasets or domains.  | A pre-trained convolutional neural network (such as Resnet50) can be fine-tuned on curated medical imaging data that contains both brain cancer and normal brain images which can then be adapted for a brain tumor detection machine learning model.  |
| Transformer       | These are deep neural network models that are very capable of advanced language processing and able to better understand and generate human-like text. Unlike earlier models, transformer's architecture  | The GPT (Generative Pretrained Transformer) model as seen in Chat-GPT's GPT-4 is an example of a transformer neural network that is capable of answering many complex queries.   |

| Term   | Definition   | Example in Pathology & Medicine   |
|--|--|---|
|  | enables them to better analyze entire sentences at once, focusing on the most important parts using an "attention" mechanism. This ultimately allows them to understand context better, leading to more accurate generated results across a variety of tasks that include but are not limited to question-answering, translation, and summarization.   |   |
| t-SNE  | t-SNE or t-distributed stochastic neighbor embedding is a dimensionality reduction statistical method for visualizing high-dimensional data by giving each datapoint a location in a two or three-dimensional map.   | t-SNE can help cluster genes with similar expression patterns, which thus allows groups of genes to be identified with similar changes in expression within cancerous versus normal tissue.       |
| Turing Test  | A test of a machine's ability to exhibit intelligent behavior equivalent to, or indistinguishable from, that of a human.   | In pathology, a Large Language Model that passes the Turing Test could potentially interact with pathologists and provide suggestions in a manner that feels natural and intuitive.               |
| Uniform Manifold Approximation and Projection (UMAP) | A dimensionality reduction technique that can be used for visualization and general non-linear dimensionality reduction.   | UMAP can be used to visualize features of a multi-dimensional pathology dataset which allows reduction of data to two or three dimensions while preserving the relationships between data points. |
| Underfitting   | Underfitting is a scenario in data science where a data model is unable to capture the relationship between the input and output variables accurately, usually due to lack of data or not enough training (this typically means a high bias and low variance).   | An underfitted model in pathology might fail to distinguish between different types of tumors, resulting in inaccurate diagnoses.   |
| Unstructured Data                                    | Unstructured data is data that either does not have a pre-defined organized manner such as text or image data.   | Unstructured data in pathology includes pathology reports that contain paragraphs of text without synoptic data.  |
| Unsupervised Learning                                | A type of machine learning where the model is trained on "unlabeled data", with the goal of discovering patterns or structure in the data. Common Unsupervised Learning includes, but are not limited to, clustering, dimensionality reduction, anomaly detection or the pretraining phase of a Large Language Model (LLM) which may also be considered a form of self-supervised learning. Note: the LLM pretraining phase is considered both unsupervised (since it does not require external labels) and self-supervised (since the model itself creates its own "labels" from the input data). | An unsupervised clustering algorithm (e.g. K-means) can be applied to patient genomic data to identify subgroups with distinct genetic profiles and potential therapeutic targets.                |
| Up-sampling  | A technique used to increase the resolution or size of data, often employed in image processing to improve the quality of low-resolution images.   | Up-sampling can enhance the quality of medical images, such as images of whole slides, resulting in more accurate analysis and classification.  |
| Validation Test                                      | A dataset used to evaluate a machine learning model's performance during hyperparameter tuning (the test part of the train-test split), often separate from the secondary or holdout test set.   | A validation set of patient records or images can be used to evaluate the performance of a cancer diagnosis model as part of the train-test split   |

| Term                            | Definition   | Example in Pathology & Medicine   |
|---------------------------------|--|---|
|                                 |  | process. These are not ready for deployment since additional holdout testing will also be necessary.  |
| Variance                        | A measure of the spread or dispersion of data values, often used in feature selection and dimensionality reduction techniques. Variance could also refer to the model complexity as the model is being trained (recall the bias-variance trade-off).   | Feature variance is calculated for lab result data (e.g., glucose levels) to identify the most informative features for a diabetes prediction model. Another example would be an overfitted model that has high variance and low bias.  |
| Variance Inflation Factor (VIF) | Measures how much the variance of a regression coefficient is inflated due to multicollinearity with other predictors. A VIF of 1 indicates no correlation, while a VIF above 5 suggests problematic multicollinearity.  | A high VIF value for hemoglobin and hematocrit would indicate unreliable estimates due to their high correlation, potentially leading to inflated variances and less than optimal machine learning model as an end result.  |
| Vector                          | Usually represents a list of numerical values (derived from the original data) in a format that is comprehensible by computers to process. These numbers can encode various features or attributes of an image, text contents or combination of which in a format that can be understood by an AI algorithm. | The vector embeddings (stored in our vector database), representing our local document files, allows us to chat / interact with these text files using our RAG-based large language model framework.  |
| Weighted Average                | An average where each value in the dataset is multiplied by a weight, before summing up and dividing by the total weight.  | <p>Weighted averages can be used to calculate the overall performance metrics of multiclass classification models. For example, for a multiclass model with a total number of 1,400 cases in which there is prostate cancer (100 cases, 90% sensitivity), colon cancer (300 cases, 70% sensitivity), and breast cancer (1000 cases, 75% sensitivity).</p> <p>1<sup>st</sup>: Multiply the sensitivity of each class by the number of cases in that class: Prostate: <math>0.90 \times 100 = 90</math>; Colon: <math>0.70 \times 300 = 210</math>; Breast: <math>0.75 \times 1000 = 750</math>.</p> <p>2<sup>nd</sup>: Sum these weighted sensitivities: <math>90(\text{prostate}) + 210(\text{colon}) + 750(\text{breast}) = 1050</math>.</p> <p>3<sup>rd</sup>: Divide by the total number of cases to get the weighted average sensitivity: <math>\text{Weighted Average Sensitivity} = 1050 / 1400 = \approx 0.75 = 75\%</math>.</p> |
| Word Embedding                  | A technique in natural language processing where words or phrases are converted to and represented as vectors in a vector space.   | Word embeddings can be used by a Large Language Model to analyze and interpret medical literature, pathology reports, and clinical notes.   |

Table 2, Table 3 and Table 4

**Table 2: Supervised Machine Learning categories**

| Supervised Learning Type | Definition   | Examples in Medicine   | Statistical Assessment Measures  |
|--------------------------|--|--|--|
| <b>Classification</b>    | Predicting a categorical label or a discrete “class” | Image-based diagnosis of skin lesions (e.g., melanoma vs. benign nevus) is an example of “Binary Classification” | <b>Confusion matrix-based approach (i.e. based on true positive, true negative, false positive and false negative predictions)</b> to calculate Accuracy, Sensitivity (True Positive Rate or Recall), Specificity (True Negative Rate), Precision (Positive Predictive Value), F1-score, ROC-AUC, etc. |
| <b>Regression</b>        | Predicting a continuous value or range of values     | Predicting biomarker concentrations in blood serum (e.g., PSA level)   | <b>Error-based approaches:</b> Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Coefficient of Determination (R <sup>2</sup> ), etc.  |

**Table 3: Common Unsupervised Machine Learning methods**

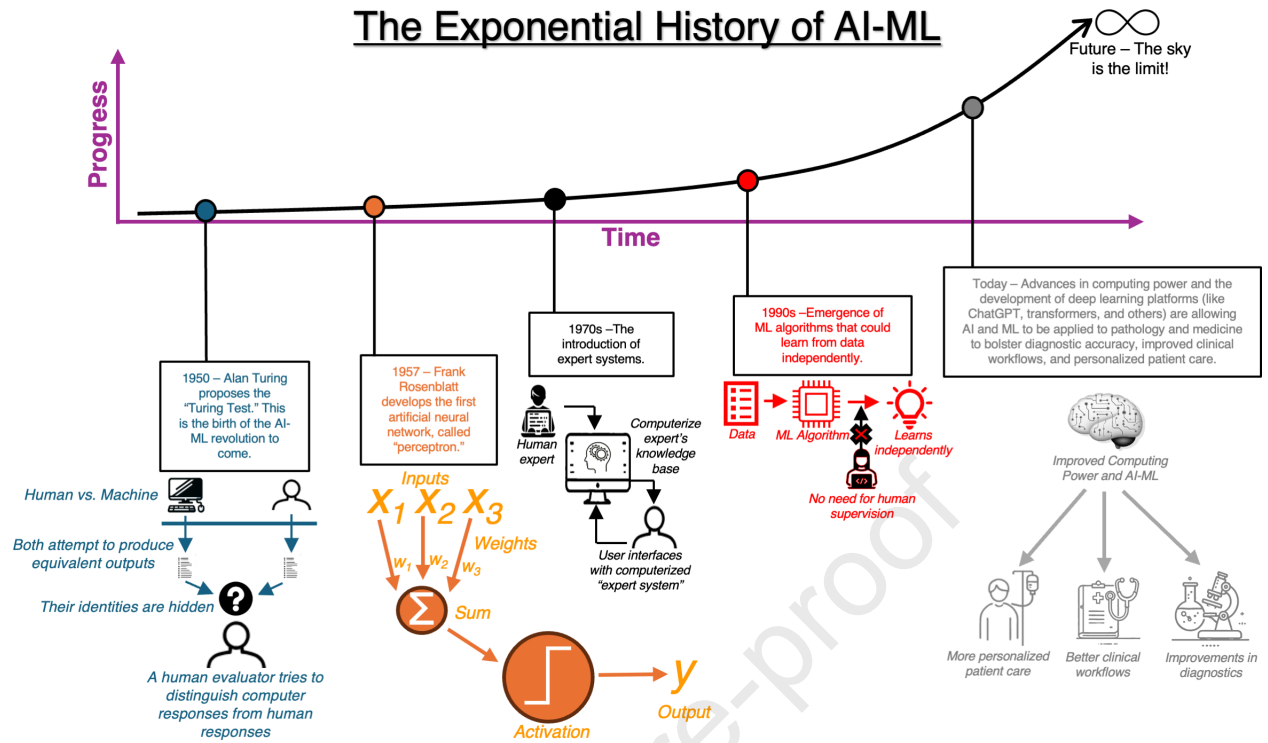
| Unsupervised Method                                       | Definition  | Key Features  | Medical Applications  |
|---|---|---|---|
| <b>Clustering</b>   | Grouping similar data points into clusters based on their characteristics | <ul style="list-style-type: none"> <li>- Identifies patterns and structures in data</li> <li>- No prior knowledge of classes or labels required</li> <li>- Can be hierarchical or non-hierarchical</li> </ul> | <ul style="list-style-type: none"> <li>- Identification of subtypes of diseases (e.g. subtypes of diabetic patients that may uncover their responsiveness to therapy, etc.)</li> </ul>              |
| <b>Dimensionality Reduction (e.g. PCA, UMAP or t-SNE)</b> | Reducing the number of feature dimensions in a dataset while preserving   | <ul style="list-style-type: none"> <li>- Simplifies complex data</li> <li>- Reduces noise and improves visualization</li> <li>- Can be linear or non-linear</li> </ul>  | <ul style="list-style-type: none"> <li>- Visualizing high-dimensional gene expression data using PCA, UMAP or t-SNE</li> <li>- Identifying key biomarkers from large proteomics datasets</li> </ul> |

|  |  |   |   |
|--|--|---|---|
|  | important information  |   | - Reducing feature dimensions in medical imaging data for improved future analysis  |
| <b>Anomaly Detection</b>                           | Identifying outliers or unusual patterns in a dataset that do not conform to the norm          | <ul style="list-style-type: none"> <li>- Detects rare events or errors</li> <li>- Can be used for quality control and monitoring</li> <li>- Often uses density-based or distance-based methods</li> </ul> | <ul style="list-style-type: none"> <li>- Identifying abnormal or out of range lab results</li> <li>- Detecting fraudulent claims or billing errors in healthcare data</li> <li>- Monitoring patient vital signs for early detection of complications</li> </ul>   |
| <b>LLM Unsupervised Pretraining (Transformers)</b> | Pre-training (an unsupervised approach) on a diverse corpus of text from the internet to learn | <ul style="list-style-type: none"> <li>- Learns contextualized vector text embeddings and attention mechanisms</li> </ul>   | <ul style="list-style-type: none"> <li>- Notably, full pretraining is no small task and the more common approach for medical applications is to employ already pretrained large language models and to customize them for our needs with follow up fine-tuning and/or RAG-based approaches (e.g., become customized for pathology or clinical notes, research articles, etc.)</li> <li>- Developing chatbots or virtual assistants for healthcare applications</li> </ul> |

**Table 4: Comparison of Generative AI to Traditional Non-Generative AI Platforms**

|                                  | <b>Traditional Non-Generative AI<br/>(Classification/Regression)</b>   | <b>Generative AI (e.g. LLMs)</b>   |
|----------------------------------|--|--|
| <b>Capabilities</b>              | Predict outcomes based on existing labeled data  | Create new, synthetic data that can augment or compliment human capabilities   |
| <b>Similarities</b>              | Both may be using a similar infrastructure to support their algorithms/neural networks through platforms such as PyTorch   | Both may be using a similar infrastructure to support their neural networks through platforms such as PyTorch  |
| <b>Training Data</b>             | Require labeled training data (supervised)   | Initial pretraining in the LLM is unsupervised, followed by a supervised instruct and later RLHF fine-tuning   |
| <b>Evaluation Metrics</b>        | Classification: Confusion matrix-based performance measures (e.g., accuracy, sensitivity, specificity, F1-score)<br><br>Regression (error based such as RMSE and R2) | Lack clear evaluation metrics due to subjective nature of generated content; may use metrics like perplexity, BLEU, ROUGE, other LLM-based evaluation or human expert evaluation (currently most reliable) |
| <b>Bias and Ethical Concerns</b> | May perpetuate biases present in the labeled training data   | Can create biased or misleading content that may be difficult to detect and raise unique ethical concerns (e.g., misuse or manipulation of generated content)  |
| <b>Applications</b>              | Image classification, object detection, outcome prediction, etc.   | Text generation, image synthesis, data augmentation, etc.  |
| <b>Model Complexity</b>          | Typically simpler models with fewer parameters (may be neural network-based or a non-neural network model)   | Often more complex models with many parameters and layers (neural network-based models such as transformers)   |
| <b>Interpretability</b>          | May be more interpretable (but depends on the type of algorithm employed)  | Less interpretable (black box algorithms) due to complex generative processes (since these are deep learning approaches typically)   |
| <b>Scalability</b>               | Variable, depending on the model, it can be computationally intensive (e.g. pathology WSI tumor identification model)  | Computationally intensive and often require significant computational resources (e.g. fine-tuned custom LLM for hematopathology queries, etc.)   |





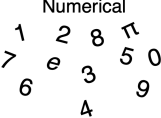


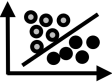


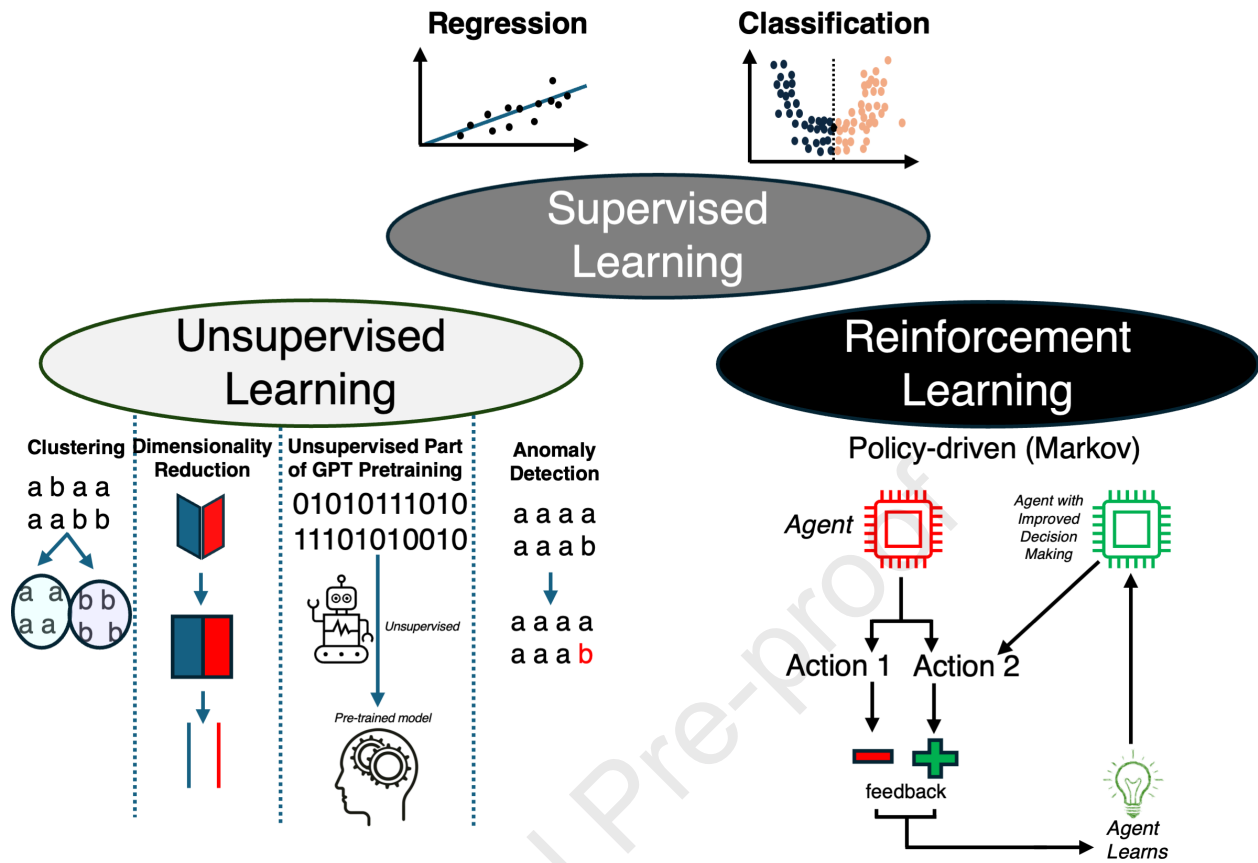
Your Journey Through This 7-Part Review Article Series



Journal Pre-proof

Data Types in AI-ML

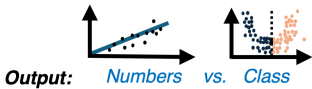
|              |   |   |   |
|--------------|---|---|---|
| Type:        | Images  | Text  | Numerical   |
|              |  |  |  |
| Typical Use: | Computer Vision Applications  | Transformer-based GPT chatbot architectures                                       | Predictive modeling tasks   |
|              |  |  |  |
| AI Examples: | CNN model (ResNet50), GAN model, Diffusion model (Stable Diffusion)               | ChatGPT (GPT 4), Llama, Claude, Mistral   | Predictive modeling (Example Classification)                                      |



Types of AI-ML

Traditional Non-Generative AI

Basis: Regression & Classification



Examples: Lab cost prediction, Cancer classification, etc.

Generative AI

Generating new information



New, albeit subjective, data (image, text, etc.)

ChatGPT/GPT-4 (text generation) or Diffusion model (image generation)