

Machine Learning and Augmented Intelligence in Clinical Medicine

Chapter 501 | Part 20: Emerging Topics in Clinical Medicine

KEY CLINICAL POINTS

- 1 Machine learning (ML) is the subfield of artificial intelligence encompassing algorithms that extract generalizable patterns from data, distinct from early rule-based expert systems like INTERNIST-I and MYCIN.
 - 2 Four primary types of ML: supervised (paired input/output), unsupervised (unlabeled data structure), semi-supervised (mixed labeled/unlabeled), and reinforcement learning (reward-based).
 - 3 Supervised learning dominates medical literature; examples include logistic regression, CNNs for imaging, and clinical risk scores like ASCVD equations.
 - 4 Deep learning models learn complex, hierarchical representations of raw data without relying on expert-identified features.
 - 5 Transformer models (e.g., GPT-4) are state-of-the-art for sequential data; GPT-4 has 175 billion parameters.
 - 6 Training involves a 'loss function' to quantify prediction error, optimized via back-propagation and stochastic gradient descent.
 - 7 Hardware like GPUs (e.g., NVIDIA Tesla H100) and frameworks (PyTorch, TensorFlow) are essential for training.
 - 8 AUC (area under receiver operating characteristic curve) is the standard metric; 0.5 = random, 1.0 = perfect.
 - 9 Computer vision applications include diabetic retinopathy detection (IDx-DR), skin cancer classification, and pneumonia detection.
 - 10 Natural language processing (NLP) includes predicting hospital readmission, differential diagnosis from text, and depression detection from speech.
 - 11 Models may be brittle in unfamiliar settings and can amplify biased prior practices; continuous assessment is essential.
 - 12 Transfer learning allows pre-trained models (e.g., from ImageNet) to be adapted to specific medical tasks.
-

1. DEFINITION & OVERVIEW

- **Definition:** Machine learning is the subfield of artificial intelligence encompassing algorithms that extract generalizable patterns from data.
- **Historical Context:**
 - Contrast to 1970s/80s expert systems (e.g., INTERNIST-I, MYCIN) which relied on human-engineered and explicitly programmed rules.
 - Evolution driven by the symbiosis of growing datasets and increasing computational power.

- Shift toward neural networks in deep learning is expected to augment both sensory and diagnostic powers of physicians.
- Potential to free physicians from labor-intensive tasks (e.g., hospital planning, fraud detection, drug monitoring).

1.1 Historical Context & Evolution

- **Evolutionary Path:** From the 1956 Dartmouth Summer Research Project to modern Large Language Models (LLMs).
- **Current State:** Modern machine learning, especially deep learning, represents a sea change in clinical practice.

1.2 Chapter Objectives

- Discuss foundations of modern ML algorithms.
- Discuss emerging applications in clinical practice.
- Provide understanding of how models are developed and their relation to epidemiologic literature.
- Encourage physician involvement in design, evaluation, and continuous assessment.

2. EPIDEMIOLOGY

- **Adoption Trends:**
 - Rapidly increasing across ophthalmology, dermatology, radiology, cardiology, and oncology.
 - Regulatory acceptance evidenced by FDA approvals (e.g., IDx-DR for diabetic retinopathy).
- **Data & Scale:**
 - Large datasets enable models with billions of parameters (e.g., GPT-4: 175 billion).
 - Transfer learning allows adaptation of models from nonmedical tasks (e.g., ImageNet) to medical applications.

2.1 Adoption Trends

- **Dominant Methods:** Supervised learning has dominated the literature; however, unsupervised, semi-supervised, and reinforcement learning are seeing increased use.
 - Computer vision has been dominant in medical ML for the past decade.
 - Natural language processing (NLP) has been transformed by deep learning.

2.2 Data Requirements

- **Model Capacity:** Modern models are capable of learning flexible, rich representations of both clinical data and raw images/text.
 - Ability to exploit spatial (images) and temporal (speech/text) structures.
 - Performance on par with expert physicians in tasks like image interpretation and analysis of unstructured text.
 - Use of semi-supervised learning to mitigate the high cost/difficulty of obtaining large expert-labeled datasets.
 - Multimodal learning allows reasoning across different data types simultaneously (e.g., images + text).

3. ETIOLOGY & PATHOPHYSIOLOGY

- **Learning Mechanism:**
 - Models learn by discovering how to represent raw data.
 - Risk: Models may encode and amplify biased prior practices or be brittle in unfamiliar settings.
- **Architecture:**
 - Deep learning models use hierarchical layer architectures to build complex representations.
 - Neurons calculate weighted sums of inputs followed by a non-linear function (e.g., ReLU).
 - Convolutional neural networks (CNNs) use 'convolutional filters' to exploit spatial structure.
 - VGG-16 (2014) had 140 million parameters.
 - Transformers are designed for sequential data (text, speech).
 - Latent diffusion models can create realistic synthetic data across multiple modalities.

3.1 Model Architecture & Learning Mechanisms

- **Neural Network Structure:**
 - Initial layers learn low-level features (e.g., edges).
 - Higher layers learn motifs and objects for discrimination.
 - Transformer models are used for sequence mapping (e.g., GPT-4).
 - Latent diffusion models can generate realistic synthetic data.

3.2 Training Process

- **Loss Function:** Quantifies discrepancy between predicted label and true label.
 - Examples: Cross-entropy, quadratic (used in linear regression), and 0-1 loss.
- **Optimization:**
 - Stochastic gradient descent combined with back-propagation algorithm to iteratively adjust weights.
- **Infrastructure:**
 - GPUs (e.g., NVIDIA Tesla H100) perform high-volume matrix calculations.
 - Frameworks (PyTorch, TensorFlow) automate training and abstract complexity.
- **Generalization:** Ensured through benchmark datasets and train-test splits.

4. CLINICAL FEATURES

- **Computer Vision (CV):**
 - Uses CNNs and vision transformers (ViT) for spatial data.
 - Applications:
 - Diabetic retinopathy detection (Gulshan et al., AUC 0.99).
 - Retinal OCT scan triage (De Fauw et al., AUC >0.99).
 - Skin cancer classification (Esteva et al.).
 - Pneumonia, pancreatic cancer, and cardiac structure segmentation.
- **Natural Language Processing (NLP):**
 - Uses Transformer models for sequence mapping.
 - Applications:
 - Predicting hospital readmission from EHR notes.
 - Providing differential diagnosis for complex cases.

→ Medical translation and processing of patient speech.

- **Multimodal Learning:** Simultaneous reasoning over different data types (e.g., images + text).

4.1 Computer Vision Applications

- **CNNs & ViTs:** Well-suited for medical imaging.
 - Gulshan et al. (JAMA 2016): CNN trained on 128,175 images; AUC 0.99.
 - De Fauw et al.: Deep learning model for retinal OCT scans; AUC >0.99.
 - Esteva et al. (Nature 2017): CNN to distinguish skin lesions; performed at level of 21 board-certified dermatologists.
 - Other uses: Pneumonia detection, pancreatic cancer identification, cardiac structure segmentation.

4.2 Natural Language Processing Applications

- **Transformer Models:**
 - GPT-4 (175 billion parameters) is state-of-the-art for sequence mapping.
 - Pretraining on massive text corpora allows models to predict the next word.
 - Capabilities: Medical professional exams, general diagnostic reasoning, and medical translation.

5. DIFFERENTIAL DIAGNOSIS

- **Model Limitations:**
 - Risk of encoding/amplifying biased prior practices.
 - Brittleness in unfamiliar or evolving settings.
 - Lack of clinical context in narrow, unrealistic evaluation tasks.
 - Often divorced from actionable recommendations for patient care.

5.1 Model Limitations & Considerations

- **Clinical Context:** Many models are evaluated on tasks that lack the clinical context a physician would incorporate.
 - Requirement: Models must be positioned to address physician/patient needs and undergo continuous assessment.

6. INVESTIGATIONS & DIAGNOSIS

1. Evaluation Metrics:

- AUC (Area Under the Receiver Operating Characteristic Curve): Standard for binary classification.
- 0.5 = random; 1.0 = perfect.

2. Infrastructure Requirements:

- GPU: Specialized hardware for matrix calculations (e.g., NVIDIA Tesla H100).
- Frameworks: PyTorch and TensorFlow automate training processes.

3. Data Management:

- Train-test split: Dividing data to ensure fair estimates of performance after training.

4. Architecture Selection:

- U-Net for biomedical image segmentation.
- Transformer models for sequence mapping (e.g., text, speech).

6.1 Model Evaluation Metrics

- **Loss Function:** Quantifies discrepancy between predicted label and true label (Cross-entropy, quadratic, 0-1).
 - Back-propagation: Algorithm to compute the direction of the loss function with respect to changes in weights.
 - AUC: Standard metric for binary classification; 0.5 = random, 1.0 = perfect.

6.2 Hardware & Software Infrastructure

- **GPU:** Specialized hardware (e.g., NVIDIA Tesla H100) to speed up matrix calculations.
 - Frameworks: PyTorch and TensorFlow automate training and abstract complexity.
-

7. MANAGEMENT & TREATMENT

1. Supervised Learning:

- Definition: Uses paired input and labeled output to learn a generalizable function.
- Examples: Logistic regression, CNNs for histopathology.
- Clinical Use: Risk scores (e.g., ASCVD) based on paired data.

2. Unsupervised Learning:

- Definition: Uses unlabeled data to discover structure/representations.
- Examples: Principal component analysis (PCA), pre-training of LLMs.

3. Semi-Supervised Learning:

- Definition: Hybrid approach using both labeled and unlabeled data.
- Examples: Self-training, consistency regularization.
- Clinical Use: Using unlabeled cardiac MRI with small labeled sets to detect hypertrophic cardiomyopathy.

4. Reinforcement Learning:

- Definition: Agent interacts with environment to optimize a numerical reward.
- Examples: Optimal control, deep reinforcement learning (e.g., AlphaGo Zero).

5. Modern Medical Machine Learning Toolkit:

- Linear and logistic regression: Baseline for continuous or binary outcomes.
- Gradient-boosted trees: Ensemble of decision trees for non-linear functions in supervised settings.
- Convolutional neural network (CNN): Specialized for image structure (e.g., U-Net).
- Transformer models: State-of-the-art for sequence mapping (e.g., GPT-4).
- Latent diffusion models: Generates synthetic data across multiple modalities (e.g., DALL-E2, Stable Diffusion).
- Transfer learning: Adapting models from one task/domain to another (e.g., ImageNet → medical imaging).

7.1 Supervised Learning

- **Definition:** Uses paired input and labeled output examples.
 - Examples: Logistic regression, CNNs.

7.2 Unsupervised Learning

- **Definition:** Uses unlabeled data to discover structure.
 - Examples: PCA, LLM pre-training.

7.3 Semi-Supervised Learning

- **Definition:** Hybrid approach using both labeled and unlabeled examples.
 - Examples: Self-training, consistency regularization.

7.4 Reinforcement Learning

- **Definition:** Agent interacts with environment to optimize a numerical reward.
 - Examples: Optimal control, AlphaGo Zero.

7.5 Modern Medical Machine Learning Toolkit

- **Linear/Logistic Regression:** Baseline for continuous or binary outcomes.
 - Gradient-boosted trees: Ensemble of decision trees for non-linear functions.
 - CNN: Specialized for image structure (e.g., U-Net).
 - Transformer Models: State-of-the-art for sequence mapping (GPT-4).
 - Latent diffusion models: Generates synthetic data across multiple modalities (DALL-E2, Stable Diffusion).
 - Transfer learning: Adapting models from one task to another.
-

8. PROGNOSIS & COMPLICATIONS

- **Bias:** Models may encode and amplify biased prior practices.
- **Brittleness:** Risk of failure in unfamiliar or evolving settings.
- **Context Gap:** Many models lack the clinical context and actionable recommendations required for point-of-care use.

8.1 Model Biases & Limitations

- **Bias:** Models may encode/amplify biased prior practices.
 - **Brittleness:** Risk of failure in unfamiliar settings.

8.2 Clinical Implementation Considerations

- **Requirement:** Models must be positioned to address physician/patient needs and undergo continuous assessment.
-

9. SPECIAL CONSIDERATIONS

- **Data Modalities:**
 - Multimodal learning allows reasoning over images, text, and speech simultaneously.
 - High-dimensional data (e.g., >20,000 protein-coding genes) requires dimensionality reduction (PCA).
- **Hardware & Software:**
 - GPU availability impacts training speed and feasibility.
 - Frameworks (PyTorch/TensorFlow) are essential for standardizing development.
- **Deployment:**
 - Requires continuous assessment to ensure safety and accuracy in clinical practice.

9.1 Data Types & Modalities

- **Multimodal Learning:** Reasoning over images, text, and speech.

- High-dimensional data: Requires dimensionality reduction (e.g., PCA).

9.2 Hardware & Software Considerations

- **GPU:** Specialized hardware for matrix calculations (e.g., NVIDIA Tesla H100).
 - Frameworks: PyTorch, TensorFlow.

9.3 Model Deployment & Continuous Assessment

- **Continuous Assessment:** Essential to ensure models remain safe and accurate in clinical practice.

10. KEY PEARLS & CLINICAL TRAPS

- **Core Distinction:** Supervised learning uses labels; Unsupervised finds structure; Semi-supervised uses both; Reinforcement uses rewards.
- **Metric of Choice:** AUC is the primary metric for binary classification (0.5 = random, 1.0 = perfect).
- **Clinical Utility:** CNNs are preferred for spatial/image data; Transformers are preferred for sequential/text data.
- **Safety Warning:** Models can be 'brittle' and may amplify bias—clinical oversight is mandatory.

10.1 Board-Exam Favorites & Must-Know Concepts

- **AUC:** 0.5 = random, 1.0 = perfect.
- **GPT-4:** 175 billion parameters.
- **VGG-16:** 140 million parameters (2014).

10.2 Common Diagnostic Pitfalls

- **Bias/Brittleness:** Models may amplify bias or fail in unfamiliar settings.
 - **Context Gap:** Models may lack actionable recommendations for point-of-care use.

REFERENCE TABLES

TABLE 501-1

Types of Machine Learning and Clinical Examples

SUBFIELD OF MACHINE LEARNING	DEFINITION	CLASSIC EXAMPLE	CONTEMPORARY EXAMPLE	CLINICAL EXAMPLE
Supervised learning	Methods that use paired input and labeled output examples to learn a generalizable function that predicts output labels from input	Logistic regression	Convolutional neural network (CNN)	Histopathologic detection of lymph node metastases in breast cancer patients

SUBFIELD OF MACHINE LEARNING	DEFINITION	CLASSIC EXAMPLE	CONTEMPORARY EXAMPLE	CLINICAL EXAMPLE
	Methods that use unlabeled input data to discover data structure and learn efficient representations for data (e.g., clustering, dimensionality reduction)	Principle components analysis (PCA)	Pre-training of large language models (LLMs)	
Semi-supervised learning	Methods that use both unlabeled and labeled examples to learn better functions than possible in supervised setting alone	Self-training	Consistency regularization	Use of unlabeled cardiac magnetic resonance images alongside small dataset of labeled examples to detect hypertrophic cardiomyopathy
	Methods to teach an "agent" that iteratively interacts with its environment how to optimize a numerical reward	Optimal control	Deep reinforcement learning (e.g., AlphaGo Zero)	

TABLE 501-2

Select Tools in the Modern Medical Machine Learning Toolkit **METHOD** Linear and logistic regression Gradient-boosted trees

METHOD	DEFINITION	NOTES
Linear and logistic regression	Models a linear relationship between predictors and either a continuous or binary outcome variable; "traditional" statistical modeling	Necessary baseline. In small carefully curated clinical datasets, these methods often perform on par with more sophisticated methods
	Ensemble of "decision trees" with parameters optimized to efficiently learn accurate nonlinear functions in supervised setting	
Convolutional neural network (CNN)	Specialized deep learning architecture with groups of neurons ("convolutional filters") that exploit structure	Widely-used deep learning model for medical imaging tasks (e.g., U-Net architecture for biomedical image segmentation)

METHOD	DEFINITION	NOTES
	Deep learning architecture designed for mapping input sequences to output sequences (e.g., text, speech)	
Latent diffusion models	A generative deep learning model that can create realistic synthetic data across multiple modalities	Can generate realistic, even award-winning art; popular tools include DALL-E2 and Stable Diffusion
	Family of approaches to adapt models trained for one task and apply to another task (across domains)	
j 2		
j 4		
j 5		

TABLE 501-3

Practical Concepts for Training a Deep Learning Model CONCEPT Loss function

CONCEPT	DEFINITION	EXAMPLES
Loss function	Mathematical function that quantifies discrepancy between predicted label and true label	Cross-entropy, quadratic, 0–1
	Algorithm to compute the direction of the loss function with respect to changes in the adjustable weights ("gradient")	
Graphics processing unit (GPU)	Specialized computer hardware to speed up the many matrix calculations involved in training a neural network	NVIDIA Tesla H100
	How data are divided to ensure fair estimates of model performance after training	
Area under the receiver operating characteristic curve (AUC)	Common performance metric for evaluating binary classification models; 0.5 = random, 1.0 = perfect	—
	Computational framework for efficiently performing matrix (tensor) calculations for training deep learning models	