Diagnosing AI

Evaluation of AI in Clinical Practice

Milan Toma

Diagnosing AI

Evaluation of AI in Clinical Practice

By

Milan Toma, PhD, SMIEEE

2026

© 2026 Dawning Research Press

All rights reserved. No part of this book may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, photocopying, recording, or otherwise, without the prior written permission of the publisher, except in the case of brief quotations embodied in critical articles or reviews.

Published by: Dawning Research Press
Contact: admin@dawningresearch.org
Website: www.dawningresearch.org

ISBN Number

Paperback: 979-8-9998324-8-1

All information has been carefully verified, and sources are cited throughout the text. The content reflects the author's commitment to evidence-based research and is intended for educational and informational purposes. The views expressed are those of the author and do not necessarily reflect those of any affiliated institutions. While every effort has been made to ensure the accuracy of the information presented, errors or omissions may still occur. If you spot a serious inaccuracy, we encourage you to contact us so that we can make corrections in the next edition. The author and publisher assume no responsibility for errors or omissions, or for damages resulting from the use of the information contained herein.

First Edition

Contents

Preface 1

1 The Illusion of Expertise 5

1.1 The Public Discourse 7

1.2 A Brief History of Artificial Intelligence 8

1.3 The Misconception 10

1.3.1 Machine Learning in Radiology 12

1.4 The Critical Distinction 14

2 Chatting Ain't Diagnosing 17

2.1 Materials and Methods 19

2.1.1 Selection of Large Language Models 19

2.1.2 Image Selection and Prompt Design 20

2.1.3 Data Collection and Analysis 22

2.1.4 Cross-Model Evaluation Protocol 22

2.1.4.1 Grading Prompt Design 23

2.1.4.2 Grading Rubric and Scoring Domains 23

2.1.4.3 Stepwise Evaluation Process 25

2.1.4.4 Cross-Evaluation Score Matrices 26

2.1.4.5 Bias Minimization and Quality Assurance 27

2.1.4.6 Data Analysis 28

2.1.5 Reproducibility and Technical Parameters 28

2.2 Results 29

2.2.1 Imaging Modality and Anatomical Region Identification . . . 30

2.2.2 Primary Diagnosis 31

2.2.3 Key Imaging Features 31

2.2.4 Differential Diagnoses 32

2.2.5 Summary of Inter-Model Variability 33

2.2.5.1 Variability in Acuity Characterization 33

2.2.5.2 Variability in Anatomical Extent 33

2.2.5.3 Variability in Pathophysiological Terminology 34

2.2.5.4 Variability in Identification of Incidental Findings . . 34

2.2.5.5 Clinical Implications of Inter-Model Variability . . . 34

2.2.6 Cross-Evaluation Results 35

2.2.6.1 Total Score Matrix 35

2.2.6.2 Domain-Specific Score Matrices 36

2.2.6.3 Inter-Grader Agreement 42

2.2.6.4 Rank-Order Consistency 42

2.2.6.5 Divergent Ground Truth: The ChatBot 2 Anomaly . 43

2.2.6.6 Self-Evaluation Bias Analysis 45

2.2.6.7 Detection of Critical Errors 47

2.2.6.8 Summary of Cross-Evaluation Findings 48

2.2.7 Safety Disclaimers 49

2.3 Discussion 50

2.3.1 Concordance Does Not Imply Clinical Equivalence 50

2.3.2 Anatomical Precision and Prognostic Implications 51

2.3.3 Differential Diagnosis Heterogeneity and Downstream Testing 51

2.3.4 The Insufficiency of Majority Consensus 52

2.3.5 Cross-Evaluation Reveals Deeper Inconsistencies 53

2.3.5.1 Ground Truth Disagreement: The Fundamental Challenge 53

2.3.5.2 Implications for LLM-Based Quality Assurance . . . 54

2.3.6 General-Purpose Systems Versus Task-Specific Machine Learning Models 54

2.3.6.1 Architectural and Training Paradigm Differences . . 55

2.3.6.2 Evidence for Superior Performance of Task-Specific Models 55

2.3.7 Distinguishing Clinical Deployment Contexts 57

2.3.7.1 Autonomous Diagnosis 57

2.3.7.2 Decision Support Under Radiologist Oversight . . . 58

2.3.7.3 Educational Summarization 58

2.3.8 Variability in Safety Disclaimers and Ethical Safeguards . . . 59

2.3.9 Recommendations 60

2.3.10 Limitations 61

2.4 Conclusions 63

3 A Taxonomy of Thinking Machines 67

3.1 The Great Family Tree 68
3.2 Traditional Machine Learning 70
3.2.1 Tree-Based Methods 71
3.2.2 Instance-Based Methods 73
3.2.3 Probabilistic Methods 75
3.3 Neural Networks and Deep Learning 77
3.3.1 Dense Neural Networks 78
3.3.2 Convolutional Neural Networks 79
3.3.3 Recurrent Neural Networks 81
3.4 The Transformer Revolution 82
3.5 Comparing Paradigms 84
3.5.1 The Path Forward 86
3.6 Summary 87

4 Core Algorithms in Practice **91**
4.1 Decision Trees 92
4.1.1 Medical Applications of Decision Trees 94
4.2 Support Vector Machines 95
4.2.1 Medical Applications of Support Vector Machines 96
4.3 K-Nearest Neighbors 98
4.3.1 Medical Applications of K-Nearest Neighbors 99
4.4 Logistic Regression 100
4.4.1 Medical Applications of Logistic Regression 101
4.5 Convolutional Neural Networks 102
4.5.1 Medical Applications of Convolutional Neural Networks . . . 103
4.6 Ensemble Methods: Boosting and Bagging 105
4.7 Light Gradient Boosting Machine 105
4.7.1 Medical Applications of LightGBM 106
4.8 Adaptive Boosting 107
4.8.1 Medical Applications of AdaBoost 108
4.9 Extreme Gradient Boosting 109
4.9.1 Medical Applications of XGBoost 110
4.10 Naïve Bayes 111
4.10.1 Medical Applications of Naïve Bayes 112
4.11 Random Forests 113
4.11.1 Medical Applications of Random Forests 115
4.12 Linear Discriminant Analysis 116
4.12.1 Medical Applications of LDA 116

4.13 Summary 117

5 The Healthy Are Many, The Sick Are Few **119**

5.1 The Nature of Class Imbalance in Medicine 120

5.2 The Accuracy Paradox 122

5.2.1 Alternative Evaluation Metrics 124

5.2.1.1 F1 Score 125

5.2.1.2 Matthews Correlation Coefficient 127

5.2.1.3 Balanced Accuracy 130

5.2.1.4 Youden's J Statistic 131

5.2.1.5 Area Under the ROC Curve 134

5.2.1.6 Area Under the PR Curve 135

5.3 Why Standard Algorithms Struggle 137

5.4 Preprocessing Approaches to Class Imbalance 139

5.4.1 Oversampling Techniques 139

5.4.2 Undersampling Techniques 142

5.4.3 Hybrid Methods 144

5.4.4 Preprocessing in Medical Diagnostics Context 145

5.5 Algorithmic Approaches to Class Imbalance 146

5.5.1 Cost-Sensitive Learning 146

5.5.2 Ensemble Methods for Imbalanced Data 148

5.5.2.1 EasyEnsemble 148

5.5.2.2 BalancedBagging 149

5.5.2.3 SMOTEBoost 149

5.5.2.4 RUSBoost 149

5.5.3 One-Class Learning 150

5.6 Comparative Analysis of Approaches 151

5.7 RUSBoost: A Promising Direction 153

5.8 Summary 154

6 Clinically Oriented Evaluation Protocol **157**

6.1 The Philosophy of Clinical Evaluation 158

6.2 The Clinical Discriminative Performance 160

6.2.1 Mathematical Formulation 161

6.2.2 The Area Under the ROC Curve 161

6.2.3 Sensitivity and Specificity 162

6.2.4 Selecting Weights for Clinical Context 163

6.2.5 Interpretation and Application 165

6.3 The Clinical Predictive Utility Score 166
6.3.1 Mathematical Formulation 166
6.3.2 Positive and Negative Predictive Values 167
6.3.3 The F1 Score in the CPUS Context 168
6.3.4 Prevalence Dependence and Clinical Implications 169
6.3.5 Weight Selection for the CPUS 171
6.4 The Weighted Endpoint Accuracy Score 172
6.4.1 Mathematical Formulation 172
6.4.2 Clinical Rationale for Endpoint Weighting 173
6.4.3 Computing the WEAS: A Worked Example 175
6.5 The Clinical Endpoint Performance Metric 176
6.5.1 The Distance Penalty Function 176
6.5.2 Combining WEAS and Penalty into CEPM 177
6.5.3 Computing the CEPM: A Worked Example 178
6.6 Integrating the Four Metrics 179
6.6.1 Practical Application of the Framework 180
6.7 Selecting Weights Across the Framework 181
6.7.1 General Principles 181
6.7.2 Context-Specific Guidance 182
6.8 Extended Methodological Framework 183
6.8.1 Systematic Weight Elicitation from Experts 183
6.8.1.1 The Challenge of Weight Subjectivity 184
6.8.1.2 Swing Weight Elicitation Methodology 184
6.8.1.3 Aggregating Multiple Expert Opinions 185
6.8.1.4 Documentation and Transparency 186
6.8.2 Uncertainty-Aware Evaluation 187
6.8.2.1 Sources of Uncertainty 188
6.8.2.2 Bootstrap Confidence Intervals for Component Metrics 188
6.8.2.3 Propagating Uncertainty Through Composite Scores . 189
6.8.2.4 Incorporating Weight Uncertainty 189
6.8.2.5 Reporting and Interpretation 191
6.8.3 Temporal Evaluation Framework 191
6.8.3.1 The Challenge of Performance Drift 191
6.8.3.2 Temporal Monitoring Protocol 192
6.8.3.3 Triggered Revalidation 193
6.8.4 Flexible Penalty Structures for Non-Ordinal Problems 194
6.8.4.1 The Limitation of Distance-Based Penalties 194
6.8.4.2 Custom Penalty Matrices 195

6.8.4.3 Eliciting Penalty Matrices from Clinical Experts . . 196
6.8.4.4 Hierarchical Penalty Structures 196
6.8.5 Integration with Decision-Analytic Methods 197
6.8.5.1 Decision-Analytic Foundations 198
6.8.5.2 Utility-Weighted Composite Scores 198
6.8.5.3 Net Benefit and Decision Curve Analysis 199
6.8.5.4 Cost-Effectiveness Integration 200
6.9 Summary of Extended Framework 201
6.10 Summary 202

7 The Primacy of Learning Dynamics **207**
7.1 The Seduction of Aggregate Metrics 208
7.1.1 What Aggregate Metrics Cannot Tell Us 209
7.1.2 The Reproducibility Illusion 210
7.2 Learning Dynamics as Diagnostic Tools 211
7.2.1 The Anatomy of a Learning Curve 212
7.2.2 The Training-Validation Gap 213
7.3 Recognizing Patterns of Pathological Learning 215
7.3.1 The Memorization Pattern 215
7.3.2 The Underfitting Pattern 216
7.3.3 The Instability Pattern 218
7.3.4 The Data Leakage Pattern 218
7.4 The Generalization Challenge 220
7.4.1 Distribution Shift and Its Consequences 220
7.4.2 External Validation as Ground Truth 222
7.5 The Expected Performance Cascade 222
7.5.1 The Healthy Cascade Pattern 223
7.5.2 Interpreting Deviations from the Expected Pattern 224
7.5.2.1 When Performance Rises Instead of Falls 225
7.5.3 Quantifying the Expected Decline 226
7.5.4 Implications for Reporting and Trust 227
7.6 From Training Curves to Clinical Trust 228
7.6.1 The Reporting Gap 228
7.6.2 Toward Transparent Reporting 230
7.7 A Protocol for Dynamic Evaluation 230
7.7.1 Phase 1: Visual Inspection of Learning Curves 231
7.7.2 Phase 2: Quantitative Assessment 232
7.7.3 Phase 3: Contextual Interpretation 233

7.8 Summary 233

8 The Oracle That Cannot Read Its Own Tea Leaves **237**

8.1 Why the Temptation Is So Strong 239

8.2 An Experiment in Interpretation 240

8.3 The Four Assessments 242

8.3.1 Assessment One: Reassurance and Validation 243

8.3.2 Assessment Two: Overfitting Diagnosis with Detailed Recommendations 244

8.3.3 Assessment Three: A Narrative of Three Phases 245

8.3.4 Assessment Four: Categorical Rejection 246

8.4 Quantifying the Disagreements 248

8.5 Analysis of Disagreements 250

8.5.1 Disagreement on Fundamental Diagnosis 251

8.5.2 Disagreement on the Meaning of Validation Fluctuation . . . 251

8.5.3 Disagreement on the Sweet Spot Hypothesis 252

8.5.4 The Impossible Claim of Pipeline Validation 253

8.6 Parallels with the Image Interpretation 254

8.6.1 Categorical Disagreement on Core Diagnosis 255

8.6.2 Confident Claims Beyond Visual Evidence 256

8.6.3 Irreconcilable Ground Truth Disagreement 256

8.7 Why Language Models Fail at Visual Technical Interpretation . 257

8.7.1 Training Objective Misalignment 258

8.7.2 Lack of Grounded Visual Reasoning 258

8.7.3 Statistical Associations Without Causal Understanding . . . 259

8.8 Implications for Machine Learning Practice 260

8.8.1 The Consultation Trap 261

8.8.2 The Illusion of Second Opinion 262

8.8.3 Appropriate Uses and Boundaries 262

8.9 The Irreplaceable Foundation 264

8.10 Summary 265

9 The Economics of Diagnostic Intelligence **267**

9.1 The Economic Imperative 269

9.2 Components of Cost Analysis 271

9.2.1 Direct Operational Costs 271

9.2.2 Error Costs: The Price of Mistakes 272

9.2.3 Implementation and Infrastructure Costs 274

9.3 The Anatomy of Cost-Effectiveness Analysis 276

9.3.1 Modeling Diagnostic Populations 276

9.3.2 Comparing Total Costs 277

9.3.3 The Dominant Role of Error Cost Reduction 278

9.4 Sensitivity Analysis and Uncertainty 280

9.4.1 One-Way Sensitivity Analysis 280

9.4.2 Probabilistic Sensitivity Analysis 281

9.5 From Analysis to Decision 283

9.5.1 Investment Thresholds and Payback Expectations 283

9.5.2 Volume Considerations and Scalability 283

9.5.3 Non-Economic Considerations 285

9.6 Integrating Technical and Economic Evaluation 286

9.6.1 The Sequential Evaluation Paradigm 286

9.6.2 Reporting Standards for Complete Evaluation 288

9.7 Summary 288

Bibliography **293**

Preface

The arrival of large language models into public consciousness has transformed the discourse surrounding artificial intelligence in medicine. Systems that converse with remarkable fluency, that deploy technical terminology with apparent precision, that generate text indistinguishable in style from expert analysis; these have captured imaginations and, more dangerously, inspired misplaced confidence. The temptation to mistake linguistic sophistication for diagnostic competence has never been greater, nor the consequences of such confusion more consequential.

This book is a warning and a guide. It warns against the conflation of chatting with diagnosing, of eloquence with accuracy, of confident prose with genuine understanding. It guides the reader toward rigorous evaluation of machine learning systems intended for clinical deployment, providing frameworks that distinguish models that have truly learned from those that have merely memorized, and that translate clinical priorities into mathematical form.

The chapters proceed as follows:

Chapters 1 and 2 confront the illusion of expertise that large language models create. Chapter 1 examines how these systems exploit the confidence heuristic (our natural tendency to trust confident-sounding sources) generating authoritative medical prose regardless of underlying accuracy. Chapter 2 presents empirical evidence of this unreliability through a systematic evaluation of leading multimodal language models on radiological interpretation, revealing fundamental diagnostic errors, irreconcilable disagreements about basic findings, and the categorical unsuitability of these systems for autonomous medical image interpretation.

Chapters 3 and 4 survey the broader landscape of machine learning, distinguishing the transformer architectures underlying language models from the task-specific systems appropriate for clinical diagnostics. Chapter

3 maps the taxonomy of approaches (from traditional tree-based methods through neural networks to modern transformers) clarifying which paradigms suit which problems. Chapter 4 examines core algorithms as they are actually employed in cardiovascular medicine, demonstrating how decision trees, support vector machines, ensemble methods, and convolutional neural networks have been validated for specific diagnostic tasks.

Chapter 5 addresses the fundamental challenge of class imbalance that pervades medical data: the healthy are many, the sick are few. Standard algorithms, optimizing for overall accuracy, may learn to predict that everyone is healthy while missing every case of disease. This chapter presents preprocessing and algorithmic approaches that restore the balance clinical reality denies.

Chapter 6 develops a clinically oriented evaluation protocol that translates medical priorities into mathematical form. The composite metrics introduced here (e.g., the Clinical Discriminative Performance Score, Clinical Predictive Utility Score, Weighted Endpoint Accuracy Score, and Clinical Endpoint Performance Metric) acknowledge that in medicine, not all errors carry equal weight. A false negative that allows disease to progress unchecked differs fundamentally from a false positive that triggers additional testing.

Chapter 7 establishes the primacy of learning dynamics over aggregate metrics. A model's final accuracy, however impressive, tells us the destination without revealing the journey. The learning curves that plot training and validation performance across epochs reveal whether a model has genuinely learned generalizable patterns or has merely memorized its training data. The training-validation gap, the shape of convergence, the expected performance cascade from internal to external validation; these dynamics predict deployment success in ways that final metrics cannot.

Chapter 8 demonstrates that even the interpretation of learning curves cannot be outsourced to language models. Just as these systems fail at medical image interpretation, they fail at interpreting the diagnostic plots that reveal model quality. Four language models, presented with identical learning curves, arrive at contradictory conclusions about fundamental questions: whether overfitting is present, whether the training pipeline is valid, whether results merit publication. The oracle, it turns out, cannot read its own tea leaves.

Chapter 9 completes the evaluation framework by addressing economics. Technical validation, however rigorous, answers only whether a system works;

not whether it is worth the investment. Cost-effectiveness analysis integrates operational costs, error consequences, and implementation expenses into a comprehensive assessment of whether deployment makes sense.

Throughout, this book emphasizes a distinction that current enthusiasm for language models tends to obscure: the distinction between general-purpose systems that generate plausible text and task-specific systems that have been trained, validated, and in many cases regulatory-cleared for defined diagnostic applications. For medical image interpretation requiring diagnostic accuracy, the appropriate technology remains purpose-built machine learning models; not chatbots, however eloquent.

The reader completing this book should possess a framework for evaluating clinical machine learning systems that extends from initial training dynamics through economic analysis. They should understand why learning curves matter more than final metrics, why class imbalance demands specialized treatment, why clinical priorities must shape evaluation criteria, and why the confident prose of a language model provides no guarantee of accuracy. In medicine, where the stakes are measured in human welfare, such understanding is not merely desirable but essential.

Copyrighted Material

How can we distinguish the promise of artificial intelligence in medicine from its pitfalls? In *Diagnosing AI*, Dr. Toma provides a much-needed roadmap for clinicians, data scientists, and healthcare leaders navigating the surge of large language models and machine learning in clinical medicine. Drawing on rigorous empirical research and clear-eyed analysis, this book dispels the illusion that eloquent AI-generated text equates to diagnostic expertise. Through both case studies and technical frameworks, Dr. Toma demonstrates that while language models can generate confident, fluent medical prose, their outputs often mask fundamental errors and inconsistencies, making them ill-suited for high-stakes clinical deployment. Moving beyond critique, this book guides readers through a robust framework for evaluating AI systems in healthcare.

You'll learn the vital differences between general-purpose language models and purpose-built diagnostic AI, how to interpret learning curves, and why class imbalance in medical data demands specialized strategies. The book introduces clinically oriented evaluation protocols that prioritize patient safety, economic analysis, and real-world utility, ensuring that new technology serves genuine clinical needs rather than hype. With practical examples, clear explanations, and a critical perspective, Diagnosing AI equips its readers with the knowledge to separate innovation from illusion—and to make informed decisions in a field where the stakes are measured in human welfare.

Milan Toma, PhD, SMIEEE, is an expert in artificial intelligence and medical diagnostics, known for his evidence-based approach to evaluating clinical machine learning systems. He has published extensively on the intersection of AI, medicine, and health technology. Dr. Toma's work bridges technical innovation with the realities of clinical practice, focusing on robust evaluation frameworks, learning dynamics, and the safe, effective deployment of AI in healthcare. He is committed to advancing both the science and ethics of medical AI, with a passion for translating complex technical concepts into practical guidance for clinicians, data scientists, and healthcare leaders.

ISBN 9798999832481

90000

9 798999 832481

Copyrighted Material