Machine Learning Study Plan for Clinical-AI Project

This document lays out a phased, prioritized roadmap of **textbooks**, **lecture series**, and **practice resources**, ordered from foundational concepts to specialized subfields needed for our clinical-AI prototype. You can copy this directly into your `README.md` file.

Phase 1: Fundamental Machine Learning Concepts

Goal: Build a solid grounding in supervised learning (regression, classification), resampling, and basic model evaluation.

1. **An Introduction to Statistical Learning (ISL)** by James, Witten, Hastie & Tibshirani
    - **Phase 1a (Weeks 1–2):**
        - Chapter 2 – Statistical Learning (definitions; supervised vs. unsupervised)
        - Chapter 3 – Linear Regression (simple/multiple, least squares, inference)
        - Chapter 4 – Classification (logistic regression, LDA/QDA basics)
    - **Phase 2 (Weeks 4–5):**
        - Chapter 5 – Resampling Methods (cross-validation, bootstrap)
        - Chapter 6 – Linear Model Selection & Regularization (subset selection, ridge, lasso)
        - Chapter 8 – Tree-Based Methods (decision trees, bagging, random forests)
        - Chapter 9 – Support Vector Machines (margins, kernels)
2. **Coursera: Machine Learning by Andrew Ng**
    - **Phase 1b (Weeks 1–3):**

        - Week 1 – Introduction & Linear Regression
        - Week 2 – Linear Regression with Multiple Variables
        - Week 3 – Logistic Regression
    - **Phase 2 (Weeks 4–6):**
        - Week 5 – Neural Networks (basics)
        - Week 6 – Advice for Applying ML
        - Week 7 – Support Vector Machines
        - Week 8 – Unsupervised Learning (K-means, PCA)
3. **Stanford CS229: Machine Learning (Andrew Ng et al.)**
    - **Phase 1c (Weeks 1–2):**
        - Lecture 1 – Introduction & Linear Regression
        - Lecture 2 – Multivariate Linear Regression
        - Lecture 3 – Logistic Regression
    - **Phase 2 (Weeks 4–5):**
        - Lecture 4 – Neural Networks (basics)
        - Lecture 5 – Advice for Applying ML
4. **Hands-On Machine Learning with Scikit-Learn, Keras & TensorFlow** by Aurélien Géron
    - **Phase 1d (Weeks 2–4):**
        - Chapter 1 – The Machine Learning Landscape (concepts, workflow)
        - Chapter 2 – End-to-End ML Project (data pipeline example)
        - Chapter 3 – Classification (MNIST example, evaluation metrics)

        - Chapter 4 – Training Models (gradient descent, overfitting/underfitting)
    - **Phase 2 (Weeks 4–6):**
        - Chapter 5 – Support Vector Machines
        - Chapter 6 – Decision Trees
        - Chapter 7 – Ensemble Learning & Random Forests
        - Chapter 8 – Dimensionality Reduction (PCA)

Phase 2: Core Supervised & Unsupervised Techniques

Goal: Deepen understanding of model selection, regularization, tree/ensemble methods, SVMs, clustering, and basic neural networks.

1. **Finish ISL (from Phase 1)**
    - Chapter 5 – Resampling Methods
    - Chapter 6 – Linear Model Selection & Regularization
    - Chapter 8 – Tree-Based Methods
    - Chapter 9 – Support Vector Machines
2. **Continue Coursera Machine Learning (Andrew Ng)**
    - Week 5 – Neural Networks
    - Week 6 – Advice for Applying ML
    - Week 7 – Support Vector Machines
    - Week 8 – Unsupervised Learning (K-means, PCA)
3. **Continue Stanford CS229**
    - Lecture 4 – Neural Networks

- Lecture 5 – Advice for Applying ML
- Lecture 6 – Support Vector Machines
- Lecture 7 – Unsupervised Learning (K-means)
- Lecture 9 – Large-Scale ML (practical tips)

4. **Continue Hands-On ML (Géron)**
    - Chapter 5 – Support Vector Machines
    - Chapter 6 – Decision Trees
    - Chapter 7 – Ensemble Learning & Random Forests
    - Chapter 8 – Dimensionality Reduction (PCA)
5. **Pattern Recognition and Machine Learning (PRML)** by Christopher Bishop
    - **Phase 2a (Weeks 3–6):**
        - Chapter 1 – Introduction (probabilistic perspective; generative vs. discriminative)
        - Chapter 2 – Probability Distributions (Gaussian, priors, posteriors)
        - Chapter 3 – Linear Models for Regression (least squares, Gaussian assumptions)
        - Chapter 4 – Linear Models for Classification (logistic, generative LDA)
    - **Phase 3 (Weeks 8–10):**
        - Chapter 6 – Kernel Methods (kernels, dual representations)
        - Chapter 8 – Graphical Models (directed/undirected, factorization)
6. **Practice Resources (begin after Phase 1 basics)**
    - **Kaggle:**
        - “Titanic: Machine Learning from Disaster” (end-to-end pipeline)

        - Free micro-courses: “Intro to ML,” “Feature Engineering,” “Data Visualization”
    - **UCI Machine Learning Repository:**
        - “Heart Disease,” “Breast Cancer Wisconsin” datasets—apply logistic regression, decision trees, random forests
    - **OpenML:**
        - Fetch simple classification datasets; practice scikit-learn pipelines

---

Phase 3: Specialized Algorithms & Deeper Theory

Goal: Cover kernel methods, graphical models, ensemble learning in depth, and bridge toward multimodal and explainable AI.

1. **PRML (continue)**
    - Chapter 6 – Kernel Methods (kernel trick, SVM)
    - Chapter 8 – Graphical Models (Bayesian networks, factor graphs)
2. **Hands-On ML (Géron, finish)**
    - Chapter 9 – (Optional) Support Vector Machines revisited
    - Chapter 10 – Introduction to Artificial Neural Networks with Keras (simple DNNs)
    - Chapter 11 – Training Deep Neural Nets (batch norm, dropout, learning-rate scheduling)
3. **Python Machine Learning (Raschka & Mirjalili)**
    - Chapter 2 – Training ML Models for Classification (scikit-learn basics: logistic regression, k-NN, SVM)
    - Chapter 3 – Tour of ML Classifiers (random forests, gradient boosting)
    - Chapter 4 – Data Preprocessing (handling missing data, scaling)

    - Chapter 5 – Dimensionality Reduction (PCA, t-SNE)
    - Chapter 7 – Ensemble Learning (bagging, boosting)
4. **Stanford CS229 (finish)**
    - Lecture 7 – Unsupervised Learning (K-means)
    - Lecture 9 – Large-Scale ML (distributed or large-data tips)
5. **Coursera (finish)**
    - Weeks 9–10 – Anomaly Detection, Recommender Systems (optional if time permits)
6. **Practice:**
    - **Kaggle Intermediate Competitions:**
        - Medical imaging or tabular healthcare datasets
    - **Google Colab:**
        - Prototype small deep-learning classifiers (e.g., on MNIST or synthetic clinical data)

Phase 4: Natural Language Processing (Clinical Text)

Goal: Acquire the skills to preprocess, tokenize, build embeddings, and perform named-entity recognition (NER) on physician notes.

1. **Speech and Language Processing (Jurafsky & Martin)**
    - **Phase 4a (Weeks 7–9):**
        - Chapter 2 – Regular Expressions, Automata & Language Models (tokenization basics)
        - Chapter 3 – N-grams (statistical language modeling)

        - Chapter 4 – Neural Network Language Models (embeddings)
        - Chapter 5 – Morphology & Word-Level NLP (stemming, lemmatization)
        - Chapter 7 – Neural Sequence Labeling (NER, POS tagging)
2. **Natural Language Processing with Python (NLTK Book)**
    - **Phase 4b (Weeks 7–8):**
        - Chapter 1 – Language Processing and Python (install NLTK, overview)
        - Chapter 2 – Accessing Text Corpora & Lexical Resources (practice)
        - Chapter 3 – Processing Raw Text (tokenization, normalization)
        - Chapter 7 – Categorizing & Tagging Words (POS tagging)
        - Chapter 8 – Learning to Classify Text (text classification pipeline)
        - Chapter 9 – Extracting Information from Text (chunking, NER basics)
3. **Practical Natural Language Processing (Vajjala et al.)**
    - **Phase 4c (Weeks 9–10):**
        - Chapter 1 – Text Preprocessing (tokenization, embeddings)
        - Chapter 3 – Text Classification (supervised methods, evaluation)
        - Chapter 4 – Named Entity Recognition (clinical entity extraction)
        - Chapter 6 – Transformer Models (BERT basics; optional but useful)
4. **Stanford CS224n: NLP with Deep Learning**
    - **Phase 4d (Weeks 9–10):**
        - Lecture 1 – Word Vector Representations (word2vec, GloVe)
        - Lecture 2 – Neural Network Foundations (backprop, softmax)
        - Lecture 5 – RNNs & LSTMs (sequence modeling for notes)

        - Lecture 6 – Transformer Models (BERT, fine-tuning for NER)

5. **spaCy & scispaCy Documentation**
    - **Phase 4e (Hands-On):**
        - Install and experiment with pre-trained biomedical NER models (e.g., `en_core_sci_sm`).
        - Build a small pipeline: tokenization → entity recognition → output entity labels for clinical notes.

Phase 5: Tabular Data & Feature Engineering (Laboratory Values, Vital Signs)

Goal: Master techniques to preprocess, transform, and model numeric lab results alongside NLP features.

1. **Feature Engineering for Machine Learning (Zheng & Casari)**
    - **Phase 5a (Weeks 11–12):**
        1. Chapter 1 – Feature Engineering Basics (why features matter)
        2. Chapter 2 – Data Exploration & Visualization (EDA for lab values)
        3. Chapter 3 – Feature Transformations (normalization, binning—for lab thresholds)
        4. Chapter 4 – Text as Features (TF-IDF on short clinical notes; optional if spaCy is used)
        5. Chapter 5 – Feature Selection (L1-based, tree-based)
2. **Python Machine Learning (Raschka & Mirjalili)**
    - **Phase 5b (Weeks 11–12):**

    1. Chapter 2 – Training ML Models for Classification (logistic regression, k-NN, SVM)
    2. Chapter 3 – Tour of ML Classifiers using scikit-learn (especially RandomForestClassifier)
    3. Chapter 4 – Data Preprocessing (scikit-learn pipelines, missing data imputation)
    4. Chapter 7 – Ensemble Learning (bagging, boosting)

3. **Hands-On ML (Géron, continued)**
   - **Phase 5c (Weeks 12–13):**
     1. Chapter 6 – Decision Trees (apply to lab-value classification)
     2. Chapter 7 – Ensemble Learning & Random Forests (handle heterogeneous features)
     3. Chapter 8 – Dimensionality Reduction (if you end up with many numeric features)
4. **Practice:**
   - Use a UCI "Heart Disease" or "Breast Cancer" dataset—combine lab columns into scikit-learn pipelines.
   - Build an end-to-end model:
     1. Load CSV of labs/vitals →
     2. Impute missing values, scale →
     3. Train Random Forest →
     4. Evaluate accuracy, AUC.

Phase 6: Multimodal Fusion & Explainable AI

Goal: Learn how to combine text-based features (from notes) with numeric features (labs), and produce interpretable outputs.

1. **"A Survey on Multimodal Machine Learning"** (Tsai et al., 2019)
   - **Phase 6a (Week 14):**
     1. Read sections on **Early Fusion vs. Late Fusion** (pp. 4–6).
     2. Understand problem formulations for combining text embeddings + numeric vectors.
2. **Interpretable Machine Learning (Molnar)**
   - **Phase 6b (Weeks 14–15):**
     1. Chapter 1 – Introduction to Interpretable ML (motivations, definitions)
     2. Chapter 2 – White-Box Models (decision trees, rule lists)
     3. Chapter 3 – Post-Hoc Explanations (SHAP, LIME, partial dependence plots)
     4. Chapter 6 – Explaining NLP Models (highlight influential words in predictions)
3. **Hands-On ML (Géron)**
   - **Phase 6c (Week 15):**
     1. Code up a small SHAP example on a RandomForestClassifier trained on combined text + labs.
     2. Visualize feature importance for numeric features and word-level importance for text features.
4. **Practice:**
   - Implement a toy fusion model:
     1. Take a small corpus of clinical note snippets → vectorize using TF-IDF or spaCy embeddings →

2. Concatenate with lab values (WBC, Hemoglobin) →
3. Train a logistic regression or random forest →
4. Use SHAP to explain both numeric and text features for a single patient prediction.

Phase 7: Domain-Specific Medical ML

Goal: Explore resources tailored to clinical text mining, EHR modelling, and medical-image fundamentals (for later expansion).

1. **Deep Learning for Healthcare** (edited volume)
   - **Phase 7a (Weeks 16–17):**
     1. Chapter 2 – Clinical Text Mining (pipelines for extracting structured data from EHR notes)
     2. Chapter 6 – Electronic Health Records: Representations and Modelling (embedding clinical codes, time-series models)
2. **Machine Learning and AI for Healthcare (Springer)**
   - **Phase 7b (Weeks 16–17):**
     1. Chapter 1 – Overview of ML in Healthcare (survey of use cases)
     2. Chapter 3 – NLP for Clinical Text (NER in EHR notes)
     3. Chapter 5 – Predictive Models for Clinical Decision Support (risk scoring, basic survival analysis)
     4. Chapter 7 – Explainability in Healthcare AI (regulatory considerations)
3. **Coursera: AI for Medicine Specialization (DeepLearning.AI)** (Audit mode)
   - **Phase 7c (Weeks 17–18):**

        1. Course 1: AI for Medical Diagnosis (classification on structured data; parallels our lab-value models)
        2. Course 2: AI for Medical Prognosis (time-to-event, survival analysis)
        3. Course 3: AI for Medical Treatment (treatment effect modeling; optional)

4. **Practice:**
    - If you have access to a de-identified clinical note + lab dataset (e.g., MIMIC-III), build a small end-to-end pipeline:
        1. NLP on notes → extract symptom entities
        2. Feature engineering on labs (flag abnormal thresholds)
        3. Train a classifier (e.g., random forest) predicting a mock diagnosis
        4. Evaluate performance and write a brief report

Phase 8 (Optional/Future): Advanced Topics

Goal: Explore deep-learning for medical imaging, federated learning, and compliance frameworks.

1. **fast.ai: Practical Deep Learning for Coders**
    - **Phase 8a:**
        - Lesson 2 – Computer Vision (medical image basics)
        - Lesson 3 – NLP with ULMFiT (transfer learning for text classification)
2. **PRML (finish remaining chapters if you want deeper theory)**
    - Chapter 10 – Approximate Inference (Variational Inference, optional)
3. **Federated Learning Tutorials & Papers**

    - “Federated Learning: Challenges, Methods, and Future Directions” (Kairouz et al., 2019)

4. **Regulatory & Compliance References**

    - FDA 510(k) guidelines for clinical decision-support software

    - ISO 13485 / ISO 14971: Medical device QMS and risk management

---

Quick Reference Table

Phase	Resources & Chapters/Lectures
1	- ISL Ch 2–4- Coursera ML Wk 1–3- CS229 Lect 1–3- Hands-On ML Ch 1–4
2	- ISL Ch 5–6, 8–9- Coursera ML Wk 5–8- CS229 Lect 4–7, 9- Hands-On ML Ch 5–8- PRML Ch 1–4
3	- PRML Ch 6, 8- Python ML Ch 2–4, Ch 7- Hands-On ML Ch 9–11- CS229 Lect 7, 9
4	- Jurafsky & Martin Ch 2–5, 7- NLTK Book Ch 1–3, 7–9- Practical NLP Ch 1–4- CS224n Lect 1–2, 5–6- spaCy/scispaCy tutorials
5	- Feature Eng Ch 1–3, 4–5- Python ML Ch 2–3, Ch 4–5- Hands-On ML Ch 6–8
6	- Tsai et al. (Multimodal survey) pp. 4–6- Molnar Ch 1–3, 6- Hands-On ML SHAP/LIME tutorial
7	- Deep Learning for Healthcare Ch 2, 6- ML & AI for Healthcare Ch 1, 3, 5, 7- AI for Medicine (audit)
8	- fast.ai Lesson 2–3- PRML Ch 10- Federated Learning tutorials- Regulatory QMS references

---

Getting Started

1. **Clone this repository** (or create a new repo and paste this README).

Create subfolders for each resource type (textbooks, videos, practice):

/ml-studyp lan
/textbooks
/videos
/practice
README.md

2.
3. **Populate each folder** as you acquire PDFs or links. For example, under `/textbooks/ISL`, save the PDF or a `links.md` with the download URL.
4. **Follow the Phase 1 plan first**. Only move to the next phase once you feel comfortable with the recommended chapters/lectures from the current phase.
5. **Commit progress regularly**: add notes or small code examples in `/practice` as you work through each resource.

> **Tip:** If full PDFs exceed your repo size limits, store only the relevant chapter files or include direct download links in a `links.md` file. For video playlists, include the YouTube URLs instead of downloading videos.

End of Machine Learning Study Plan