ChatGPT

Telemedicine AI Project: Datasets and Models

Symptom Checker

- **Kaggle Disease–Symptoms Data** – A public dataset (Choong Qian Zheng's *"Disease and Symptoms"* dataset) containing ≈5,000 records linking about 800 diseases to ~600 symptoms [1]. This is a general-medicine symptom–disease mapping ideal for training a classifier. It can be downloaded from Kaggle, and used to train a lightweight model (e.g. a small neural net or scikit-learn classifier).
- **Mendeley "Disease and Symptoms" (2023)** – An open CC BY-4.0 dataset with 773 diseases and 377 symptoms (≈246k rows) that preserves symptom severity and disease occurrence probabilities [2]. This large mapping of symptoms→diseases (with probabilistic weights) is suited for dynamic question logic (e.g. Bayesian or decision-tree inference). Use it to estimate likelihoods or to drive follow-up questions based on symptom severity.
- **SymbiPredict (SDPD) Dataset** – A curated CC BY-4.0 symptom–disease dataset ("Symptom–Disease Prediction Dataset") linking symptoms to diseases with expert sourcing [3]. It focuses on general/internal medicine and is designed for predictive modeling. (Data can be downloaded from Mendeley.) Combining it with NLP symptom extraction (e.g. mapping free-text symptoms to these standard terms) lets you handle free-text input.

Pill Identifier

- **NLM RxIMAGE (C3PI) Dataset** – A large open pill-image collection from the NIH's RxIMAGE project [4] [5]. It includes 4,000 "reference" images (high-res lab photos) and 133,000+ consumer-grade pill photos (front/back of pills under varied lighting) [5]. These cover thousands of branded and generic tablets. Use this for training an image-based pill classifier or detector. (Download from the NIH data portal via the links under "Downloads & Resources" on [C3PI Data.gov][31].)
- **NLM Pillbox Metadata & Images** – The retired **Pillbox** database provides CSV metadata for solid medications (shape, color, imprint, active ingredients, NDC codes, etc.) [6]. Pillbox's final image library (over 66,000 pill images) is still available via NLM's FTP [7]. You can combine Pillbox CSV data with RxIMAGE photos or use the images themselves (front/back scans). These resources cover English-brand and generic pills. They support both image lookup and text-based ID (e.g. match a prescription imprint string to Pillbox entries).
- **Usage Tips:** A deep-learning model (e.g. small CNN) can be trained on the NLM images, but for the RTX 3050 you'll want a compact backbone. For example, YOLOv5s (small) has been used successfully for real-time pill detection [8]. Lightweight CNNs (MobileNet, EfficientNet-B0, or a custom 6-layer net) can classify pill images and imprint text. For text labels, use OCR (Tesseract/EasyOCR) or a retrained text-spotter, as suggested by recent work [8]. The Pillbox metadata (CSV/JSON) can be used to verify a pill once the image or imprint is recognized.

Medical Report Analyzer

- **Medical Lab Report Images (Kaggle)** – A publicly shared collection of scanned lab-report images (blood test reports, etc.). For example, the "Medical Lab Report Dataset" on Kaggle contains a few

hundred real lab report images (≥426 images) [9]. It has been used in studies of OCR on bloodwork. This can be used to train or test OCR and data extraction pipelines on clinical report formats.

- **MIMIC-IV Clinical Notes (PhysioNet)** – A massive de-identified text corpus of general medical documents. MIMIC-IV-Note provides **331,794** English discharge summaries (hospital notes) plus **2.32 million** radiology reports [10]. While these are already digital text, they can train NLP models (NER, summarization, question-answering) on real medical language. These are free for researchers (after signing data-use terms via PhysioNet). Combine with OCR datasets to handle scanned input.
- **OCR Training Data:** For scanned PDF processing, you can supplement with general document OCR corpora (e.g. the UW III dataset, RVL-CDIP, or synthetic text images). However, the Kaggle lab-report images and real hospital report scans (if available) are most domain-relevant. Tools like Tesseract or EasyOCR can be trained/enhanced on these images. After OCR, use NLP datasets like MIMIC to interpret lab values and summaries.

Model Training & Pretrained Models

- **Lightweight Classification Models:** With only a 6GB GPU, focus on small models. For the **symptom checker**, classical ML works well: for example, SymbiPredict used a **Linear SVM (LinearSVC)** with TF-IDF–encoded symptom text [11]. This kind of model runs quickly on CPU/GPU and can predict diseases from selected symptoms. A small feedforward neural net (as in [22†L163-L170]) with tens of neurons per layer also fits within memory. Even Naive Bayes or logistic regression on symptom vectors can be effective.
- **Pill Identification Models:** Use compact CNNs. For object detection/recognition, **YOLOv5s** (the "small" YOLOv5) is explicitly designed for high speed/low resource and has been applied to pill detection [8]. For pure image classification, consider MobileNetV2, EfficientNet-B0, or other mobile-friendly networks; these can run on a 6GB GPU if batches are small. Even ResNet-18 or ResNet-34 (half the size of ResNet-50) can often fit. Pretrained backbones (on ImageNet) can be fine-tuned on the pill images for quick convergence.
- **Text/Report Models:** For processing lab/discharge reports, a smaller language model or rule-based system is best. If using transformers, try *distilled* or *small* versions (e.g. DistilBERT, MobileBERT) trained on medical text (BioBERT or PubMedBERT variants have lighter siblings). Otherwise, classical NLP (regex for lab values, spaCy NER on medical terms) can run fast. Pretrained OCR (Tesseract) and pre-trained NLP (scispaCy en_core_sci_sm or Meta's Llama-7B for Q&A on CPU) are options. Always prefer smaller models (e.g. Llama-2-7B or GPT-4o if using API) or classical ML to meet RAM/VRAM limits.

AI Chat Assistant (Fallback/Advisor)

- **Google Gemini/Vertex AI (MedLM):** Use Google's Generative AI APIs (Gemini via Vertex or the MedLM model). For example, Vertex AI MedLM accepts a JSON POST with a "Question:" prompt [12]. An example request body is:

```
{
  "instances": [{"content": "Question: What causes ringworm?"}],
  "parameters": {"temperature":0,"maxOutputTokens":256}
}
```

(This uses the MedLM-Medium model endpoint on Google Cloud [12].) The response will be a medically-oriented answer. In code, you'd call the `projects/…/models/medlm-medium:predict` endpoint with proper auth [12]. This allows a fall-back "talk to chat" mode if the symptom/pill modules can't decide, or to provide extra medical info. *Prompt tip:* Always include a strict health-system prompt. For example: "You are an AI health assistant. Only answer health-related questions; if non-medical, respond 'I'm here to assist with health questions only.'" [13]. This ensures safe, focused responses.

- **OpenRouter API (GPT-style models):** OpenRouter offers a unified API (OpenAI-compatible) to many LLMs (GPT-4o, Llama, etc.) [14]. You can simply make an OpenAI-format chat completion call to `https://api.openrouter.ai` with your key, selecting a model like `openai/gpt-4o`. For example (Python):

```
response = openai.ChatCompletion.create(
    model="openai/gpt-4o",
    api_base="https://api.openrouter.ai",
    api_key="YOUR_KEY",
    messages=[
      {"role": "system", "content": "You are a helpful medical advisor."},
      {"role": "user", "content": "I have headache and fever; what might
this be?"}
    ]
)
```

  This will forward the request to the specified model. OpenRouter normalizes the payload so you can use the same code for different LLMs [14].
- **Prompt Examples:** For health guidance, use clear role prompts. E.g. system prompts like the one above [13]. For user prompts, use straightforward questions or commands. Example: **"Explain [medical condition] in simple language for a patient."** or **"Summarize the causes and risk factors for [disease]"** [15]. Another example: **"List 3 FAQs with answers for someone diagnosed with [condition]."** [16]. Tailoring prompts with context and word limits (as shown in [61]) yields focused, patient-friendly answers.

References: We recommend the cited open datasets and APIs above for all components. For lightweight models, scikit-learn and small TensorFlow/Keras models have been used in research [11] [8]. Always choose models and batch sizes that fit 6GB VRAM (e.g. MobileNet or tiny YOLO for images, small transformers or rule-based NLP for text).

---

[1] Building a Sequential Neural Network to Predict Disease Development | by Anna Ekmekci | Medium
https://medium.com/@anna.ekmekci/building-a-sequential-neural-network-to-predict-disease-development-776520be39e7

[2] Disease and symptoms dataset 2023 - Mendeley Data
https://data.mendeley.com/datasets/2cxccsxydc/1

[3] SymbiPredict - Mendeley Data
https://data.mendeley.com/datasets/dv5z3v2xyd/1

4 5 Computational Photography Project for Pill Identification (C3PI) - Catalog
https://catalog.data.gov/dataset/computational-photography-project-for-pill-identification-c3pi-82201

6 7 Pillbox (retired January 28, 2021) - Catalog
https://catalog.data.gov/dataset/pillbox-retired-2f7f9

8 Real-time pill identification and classification using deep learning framework for medicine inspection systems | Discover Electronics
https://link.springer.com/article/10.1007/s44291-025-00122-6

9 arxiv.org
https://www.arxiv.org/pdf/2509.06033

10 MIMIC-IV-Note: Deidentified free-text clinical notes v2.2
https://www.physionet.org/content/mimic-iv-note/2.2/

11 GitHub - rchrdwllm/predictive_healthcare_model
https://github.com/rchrdwllm/predictive_healthcare_model

12 Create MedLM prompts | Generative AI on Vertex AI | Google Cloud Documentation
https://docs.cloud.google.com/vertex-ai/generative-ai/docs/medlm/medlm-prompts

13 Creating a Health Chatbot Using Twilio and Gemini: A Step-by-Step Guide | Twilio
https://www.twilio.com/en-us/blog/developers/community/create-health-chatbot-using-twilio-gemini-step-by-step-guide

14 OpenAI: GPT-5 Chat – API Quickstart | OpenRouter
https://openrouter.ai/openai/gpt-5-chat/api

15 16 100+ ChatGPT prompts for healthcare professionals
https://www.paubox.com/blog/100-chatgpt-prompts-for-healthcare-professionals