

Divyanshu Shrivastava

AI/ML Engineer | ML & NLP (LLMs) | Healthcare Analytics

+91 8982274716

Indore, India

https://github.com/div1996

divyanshushrivastava646@gmail.com

https://www.linkedin.com/in/divyanshu-shrivastava-70aa35164/

https://div1996.github.io/divyanshushrivastava.co.in/

Senior Data Scientist with **5+ years of experience in Healthcare ML, NLP, and GenAI**, specializing in **risk stratification, LLM-based summarization, SDoH modeling, and population health analytics**. Designed and deployed **neural network–based risk models**, SDoH classifiers, and LLaMA-based clinical summarization systems used by US healthcare clients. Skilled at transforming raw clinical/claims data into scalable ML pipelines using **Python, TensorFlow, MLFlow, and AWS**, with a strong focus on clinical accuracy, fairness, and operational impact.

Work Experience

Senior Data/AI Scientist, Vriba Technologies Private Limited — Mar 2026 – Sep 2026 | Pune, India

Summarization and Decision Making for Pre-Authorization

- Built an end-to-end **Pre-Authorization Summarization and Decision Support system** for healthcare claims review using OCR, LLM, and workflow automation.
- Designed a scalable document ingestion pipeline where clinical documents are uploaded through **API**, stored in **S3 Document Storage**, and processed using **SQS-based OCR queues**.
- Implemented **OCR Worker and OCR Runner** components to extract text from medical records, referral notes, treatment documents, lab reports, and prior authorization forms using **AWS Textract / OCR**.
- Developed an automated workflow to store OCR outputs in **S3 OCR Output** and pass extracted clinical text to downstream LLM services for summarization and case review.
- Built **LLM-based Clinical Summary generation** to convert long medical documents into structured summaries including patient history, diagnosis, treatment details, clinical findings, and requested service information.
- Created **LLM Case Details extraction** to identify key information required for pre-authorization review such as procedure codes, diagnosis details, provider notes, medical necessity indicators, and supporting evidence.
- Integrated **Claims History API** to fetch patient claim history and combine it with OCR-extracted clinical data for better decision-making.
- Developed **LLM Determination logic** to support pre-authorization decisions by generating structured outputs such as approval recommendation, denial reason, missing documentation, and review notes.
- Designed **Workflow Runner and Task Worker** services to manage step-by-step processing from OCR completion to summary generation, case extraction, claim history review, and final determination.
- Implemented **Logs / DB tracking** to monitor workflow status, capture processing errors, store audit logs, and support traceability for healthcare operations.
- Improved clinical review efficiency by reducing manual document reading effort and helping reviewers quickly understand patient cases, required treatment, and pre-authorization decision points.
- Applied healthcare domain knowledge in prior authorization, claims, clinical documentation, treatment review, and payer operations to build a practical AI-assisted solution for medical decision support.

Senior Data Scientist, Precognitas Health Private Limited — Jun 2023 – Feb 2026 | Noida

- Built the **Caliper Predictive Cost Risk Model** using neural networks with **130+ features** (HCC diagnoses, demographics, 7–10 SDoH factors), achieving **higher accuracy and completeness** than industry regression-based models.
- Designed **Two-Stage Hurdle Cost Architecture** to separate zero-cost patients and classify cost tiers; improved F1-score for $20K+ segment from **<0.15 → 0.43** and ROC-AUC to **0.87**.
- Applied **LASSO feature optimization** to reduce features from 130 to **50 predictors** while maintaining **>95% of full-model performance**, improving deployment efficiency.
- Built **specialist ensemble models per cost band**, improving high-cost risk precision and weighted accuracy (**~0.73**).
- Implemented **Quality Improvement and Engagement Improvement analytics** to identify care gaps, high-leverage interventions, patient outreach priorities, and avoidable utilization segments.
- Created a **Disparity Index Model** measuring outcomes across race/ethnicity, SDoH clusters, geography, enabling equity-focused care strategies.
- Built production-grade **LLM Summarization Pipelines** for clinical and care-management notes using **LLaMA-3-8B-Instruct** and rule-based prompting, generating structured summaries (Primary Concerns, Findings, Plan/Actions) that reduced clinical review workload by **40–60%**.
- Developed **NLP SDoH Classification** using BART Zero-Shot and fine-tuned models to identify social needs from unstructured text (transportation, housing, food insecurity, financial strain).
- Served as primary ML contact for **US healthcare clients**, translating clinical requirements into ML specifications and delivering population-health insights to leadership teams.

Data Scientist, XALT Analytics Pvt. Ltd. — Sep 2021 – May 2023 | Indore, India

- Led a team of Data Science and Big Data professionals, assigning tasks and driving productivity while fostering a collaborative work culture. Acted as the primary contact for USA-based clients, addressing their needs and queries. Developed and deployed an in-house deep learning image-based object detection product, leading a successful pilot run and demonstration to Indian government officials.

Projects

Summarization and Decision Making for Pre-Authorization — Mar 2026 – Present

This project improved the pre-authorization review process by using automation, OCR, LLM summarization, and claims history analysis to reduce manual clinical document review time by around **50–60%**. It helped reviewers get clear case summaries, key medical details, claims history, and decision support in one workflow. The solution improved decision accuracy by approximately **25–30%** by reducing missed information and making approval, denial, or missing-documentation recommendations more consistent. It was also designed with scalable cloud architecture using S3, SQS, workflow workers, and APIs, so it can handle high-volume healthcare documents and easily scale as case volume increases.

Predictive Cost Risk Model (Population Risk Stratification)

Developed a neural network model using **HCC categories, demographics, and SDoH variables** for predicting patient-level future cost; implemented **two-stage hurdle modeling**, **LASSO feature selection**, and **ensemble architectures**, achieving **high-cost segment F1 = 0.43** and **ROC-AUC = 0.87**.

Quality & Engagement Improvement Models
Designed advanced models for identifying **care gaps, screening needs, chronic care adherence**, and predicting **patient engagement likelihood**, improving prioritization for care managers.

Disparity Index Methodology
Developed a methodology to detect inequities across **race, ethnicity, and SDoH dimensions** using Quality metrics and demographic/clinical factors.

LLM-based EMR & Care Management Summaries
Built production LLM pipelines (LLaMA-3) generating structured JSON summaries for clinical and care-manager notes, reducing review time **40–60%** and improving data usability for downstream risk models.

NLP SDoH Extraction
Developed a zero-shot classification pipeline using **Facebook BART** to identify social determinants of health from unstructured notes including barriers, social needs, and resource gaps.

Skills

Data Science Techniques: Knowledge-Base, **RAG** ,Langchain, Langgraph, Agentic AI, MCP, Agents, Predictive Analytics, **GenAI/LLM**, Data Science , Data Analytics, Machine Learning, Image Analytics, Deep Learning, Natural Language Processing (NLP), Data Visualization,**Text Summarization**,**Zero-Shot Classification, Clinical & Care Manager Summarization, Prompt Engineering**.

IDE Tools & Cloud: AWS,S3,EC2,EMR Studio,bedrock,MLFlow,Power BI,Tableau, GitHub, MS Office, Visual Studio IDE ,Jupyter Notebook , Postman, PowerBI, Docker.

Languages & DB: Python,**PySpark,**SQL,MongoDB,R Programming

Libraries: Pandas,NumPy,Matplotlib,Scikit-Learn,Scipy,Nltk,Spacy,Tensorflow,Keras,Flask,Django,Huggingface

ML & Healthcare Modeling: Risk Stratification, Cost Prediction, Impactability Modeling, Quality & Engagement Analytics, SDoH Modeling

Education

Master of Business Administration (Business Analytics), Devi Ahilya Vishwavidyalaya (NAAC A+ accredited) 2019 – 2021 | Indore, India

Bachelor's of Commerce (Taxation), Devi Ahilya Vishwavidyalaya (NAAC A+ accredited) 2014 – 2017 | Indore, India