NARAYANA SURAMPUDI, PhD, CAHIMS

Clarksburg, MD 20871 • (847) 975-0153 • lsurampudi@gmail.com • LinkedIn • Blog • GitHub

PROFESSIONAL SUMMARY

Senior Healthcare Data Scientist with 9+ years of experience applying **Generative AI, Large Language Models (LLMs), and Machine Learning** to solve complex clinical informatics challenges. Expertise in developing **clinical NLP pipelines, RAG architectures, and LLM fine-tuning** for healthcare applications including named entity recognition, adverse event classification, and clinical text mining. Proven track record at **NIH Clinical Center** leading OMOP CDM implementations, EHR data integration, and cross-functional collaboration with researchers, clinicians, and engineering teams. **PhD in Chemical & Biomolecular Engineering** with peer-reviewed publications and **CEO Award recipient (2025)** for outstanding contributions to clinical research informatics.

TECHNICAL SKILLS

Generative AI & LLM:

Large Language Models (GPT, Claude, LLaMA), Prompt Engineering, RAG Pipelines, LLM Fine-tuning, LangChain, Vector Databases (ChromaDB, Pinecone), Hugging Face Transformers, Clinical NLP, Named Entity Recognition (NER), Text Classification, Relation Extraction

Clinical NLP & Healthcare AI:

MedSpaCy, ClinicalBERT, BioBERT, RoBERTa, ScispaCy, Adverse Event Classification, Clinical Text Mining, Medical Document Processing, Healthcare AI Ethics & Bias Mitigation

Healthcare Informatics & Standards:

OMOP Common Data Model, HL7 FHIR, SNOMED CT, LOINC, ICD-10, RxNorm, MeSH, UMLS, EHR Systems (AllScripts Sunrise), REDCap, Clinical Data Warehousing

Machine Learning & Data Science:

PyTorch, TensorFlow, Scikit-learn, XGBoost, Survival Analysis, Predictive Modeling, Time Series Forecasting, Clustering, Dimensionality Reduction, Statistical Modeling, A/B Testing

Programming & Data Engineering:

Python, SQL, R, Pandas, NumPy, PySpark, ETL Pipelines, Data Validation, Data Quality Frameworks

Cloud & Infrastructure:

Google Cloud Platform (GCP), Snowflake, Microsoft SQL Server, IBM Netezza, Elasticsearch, Docker, Git, CI/CD

BI & Visualization:

Microsoft Power BI, Tableau, Plotly, Streamlit, Interactive Dashboards, Executive Reporting

PROFESSIONAL EXPERIENCE

Senior Data Scientist   **Mar 2020 – Present**

National Institutes of Health (NIH Clinical Center) – Rockville, MD

- Architect and deploy Generative AI solutions including RAG pipelines with LangChain and vector databases to enable semantic search across 500K+ clinical documents, reducing research query time by 60%.
- Design and implement LLM-based clinical NLP pipelines using ClinicalBERT, BioBERT, and RoBERTa for named entity recognition, adverse event classification, and relation extraction from unstructured clinical notes.
- Fine-tune domain-specific LLMs on clinical corpora to improve NER accuracy by 25% for medication extraction and symptom identification in oncology and nephrology research.
- Develop prompt engineering frameworks and evaluation metrics for clinical text summarization and question-answering systems using GPT and Claude APIs.
- Lead OMOP Common Data Model implementation, mapping 2M+ patient records to standardized vocabularies (SNOMED CT, LOINC, RxNorm) with HL7 FHIR integration for interoperability.
- Build statistical models for Survival Analysis on cancer and End-Stage Renal Disease cohorts, contributing to 6 peer-reviewed publications on prognosis and treatment optimization.

- Lead cross-functional team of data analysts, ETL developers, and software engineers to deliver production-grade data solutions on schedule.
- Develop automated data quality frameworks in Python and SQL, reducing data validation time by 40% and ensuring compliance with clinical research standards.
- Create executive dashboards in Power BI and Tableau for stakeholder reporting on clinical trial metrics and research outcomes.

Technologies: Python, PyTorch, LangChain, Hugging Face, GPT/Claude APIs, GCP, Snowflake, SQL Server, Power BI, Tableau, REDCap, AllScripts Sunrise EHR, Elasticsearch, JIRA

Data Scientist   **Nov 2019 – Feb 2020**

Food and Drug Administration (FDA) – Rockville, MD

- Developed machine learning models for anomaly detection in network infrastructure, achieving 92% accuracy in predicting device failures before critical incidents.
- Designed end-to-end ML pipelines for real-time data processing, model inference, and automated incident reporting.
- Built interactive monitoring dashboards using Plotly Dash, enabling operations team to identify and resolve anomalies 50% faster.

Technologies: Python, Plotly Dash, Scikit-learn, SQL Server

Data Scientist   **Aug 2017 – Oct 2019**

DMI (Mine Safety and Health Administration) – Arlington, VA

- Built predictive models achieving 85% accuracy in forecasting mine safety risks, enabling proactive resource allocation across 50+ US mining districts.
- Applied ensemble methods (Random Forest, XGBoost) with k-fold cross-validation, identifying top 10 risk factors that reduced accident rates by 15%.
- Led modernization of legacy Mine Data Retrieval System (MDRS), improving data accessibility for public stakeholders and regulatory compliance.
- Developed MicroStrategy dashboards for daily monitoring of coal production, inspection history, and safety violations.

Technologies: Python, Scikit-learn, MicroStrategy, Flask, D3.js

Data Scientist – NLP & Big Data   **Apr 2017 – Jul 2017**

DMI – Bethesda, MD

- Led team of 4 (Statistical Analysts & Software Engineers) delivering NLP solutions from POC through production deployment.
- Developed NLP pipelines for sentiment analysis, topic modeling, and opinion mining on large-scale unstructured text data using Spark MLLib.
- Built ML automation pipelines for data ingestion, feature extraction, model training, and deployment in Hadoop ecosystem.

Technologies: Apache Spark, Hadoop, Kafka, Python, Scala, Keras

Sr. Statistical Analyst   **Mar 2016 – Mar 2017**

DMI – Mason, OH

- Developed Marketing Mix Models achieving 14% gross margin increase and 5% ROI improvement for retail clients.
- Performed multivariate regression analysis across 1,000+ variables from 30 datasets, identifying key revenue drivers.

Data Scientist   **Sep 2015 – Feb 2016**

Rang Technologies – Edison, NJ

- Built time series forecasting models using ARIMA and exponential smoothing for demand prediction and inventory optimization.

Computational Research Assistant   **Jan 2010 – Aug 2015**

Tulane University – New Orleans, LA

- Applied machine learning algorithms (regression, PCA, clustering) to molecular simulation data for biomolecular research.

- Developed Python/R automation scripts for HPC cluster job management and large-scale data processing.
- Implemented statistical methods including Monte Carlo simulations, WHAM free energy calculations, and bootstrap error analysis.

Validation Analyst   **2007 – 2009**

Baxter Healthcare & Zimmer Biomet – FDA Regulated Environments
- Validated medical device and pharmaceutical systems under FDA 21 CFR Part 11, GMP, and SDLC compliance frameworks.

EDUCATION

Ph.D., Chemical & Biomolecular Engineering – Tulane University, New Orleans, LA, USA

M.S., Chemical Engineering – University of Louisiana at Lafayette, LA, USA

B.E., Chemical Engineering – Andhra University, India

CERTIFICATIONS & ACHIEVEMENTS

- **Certified Associate in Healthcare Information and Management Systems (CAHIMS)** – HIMSS
- **CEO Award for Outstanding Work at NIH Clinical Center (2025)**
- IBM Corporation Fellowship in Computational Science (2010)
- Distinguished Graduate Student Award (2014)
- Outstanding Graduate Student Teaching Award (2010)

SELECTED PUBLICATIONS

- Chang HE, Hossain MS, Song C, Surampudi N, Nesterova G, Gahl WA. Long-term outcomes in nephropathic cystinosis: a review. Pediatric Nephrology 2025, 1-20.
- Ashbaugh HS, Weiss K, Meng B, Surampudi LN. Temperature and Pressure Dependence of Methane Correlations and Osmotic Second Virial Coefficients in Water. Journal of Physical Chemistry 2015, 119(20), 6280-6294.
- Surampudi LN, Ashbaugh HS. Direct Evaluation of Polypeptide Partial Molar Volumes in Water Using Molecular Dynamics Simulations. Journal of Chemical and Engineering Data 2014, 59, 3130-3135.
- Franklin JM, Surampudi LN, Ashbaugh HS, Pozzo DC. Numerical Validation of IFT in the Analysis of Protein-Surfactant Complexes with SAXS and SANS. Langmuir 2012, 28(34), 12593-12600.
- Ashbaugh HS, Liu LX, Surampudi LN. Optimization of linear and branched alkane interactions with water to simulate hydrophobic hydration. Journal of Chemical Physics 2011, 135(5), 7.
- Surampudi NL, Pesacreta TC, Misra RDK. The determining role of scratch indenter radius on surface deformation of high-density polyethylene and calcium carbonate-reinforced composite. Materials Science and Engineering A 2007, 456(1-2), 218-229.