STEPHEN MANOHAR GURRAM
Senior AI/ML Engineer
Ph: +1-9137033873 | Gmail: stephen.m.gurram@gmail.com

PROFESSIONAL SUMMARY

- Senior AI/ML Engineer with **10+ years of experience** designing scalable machine learning, deep learning, and AI-driven solutions across **banking, fintech, insurance, and telecom domains** in the U.S.
- Proven ability to translate complex business challenges into **production-grade AI/ML systems**, consistently delivering measurable impact and improving key business KPIs by **25–35%**.
- Strong expertise in **Python (3.9+)** and core ML frameworks including **TensorFlow (2.x), PyTorch (2.x), and Scikit-learn**, with hands-on experience across the complete ML lifecycle—**data preprocessing, feature engineering, model training, evaluation, deployment, and optimization**.
- Extensive experience in building and optimizing large-scale data pipelines using **Apache Spark (3.x), Databricks, and distributed computing frameworks**, enabling high-performance data processing and reducing training time significantly.
- Proficient in **MLOps and production ML systems**, leveraging **MLflow, Docker, Kubernetes (AKS/EKS), Jenkins, and CI/CD pipelines** for automated model deployment, monitoring, versioning, and lifecycle management.
- Strong cloud expertise across **Azure (Azure Machine Learning, ADLS, Data Factory, AKS, Azure OpenAI)** and **AWS (S3, EC2, Lambda, EMR, SageMaker)**, building scalable, secure, and cost-efficient ML solutions in enterprise environments.
- Specialized in **Natural Language Processing (NLP) and Generative AI**, with hands-on experience in **BERT, RoBERTa, Hugging Face Transformers, and GPT-based architectures**, including development of **chatbots, RAG-based systems, document intelligence solutions, and conversational AI platforms**.
- Experience designing and deploying **Generative AI applications using LLMs**, including prompt engineering, fine-tuning strategies, and integration with enterprise systems for automation and knowledge retrieval.
- Strong background in **data engineering and ETL pipelines** using **SQL, Snowflake, Apache Airflow, and Azure Data Factory**, ensuring reliable data ingestion, transformation, and availability for ML workflows.
- Skilled in developing **high-performance REST APIs** using **FastAPI and Flask**, enabling real-time, low-latency inference and seamless integration of ML models into production systems.
- Hands-on experience with **feature engineering, model explainability (SHAP, LIME), and advanced hyperparameter tuning** techniques (Grid Search, Random Search, Bayesian Optimization) to improve model performance and interpretability.
- Expertise in **model monitoring, drift detection, logging, and observability** using tools such as **Prometheus, Azure Monitor, and custom solutions**, ensuring long-term reliability and stability of deployed models.
- Experience building and optimizing **recommendation systems, fraud detection systems, predictive analytics models, and time-series forecasting solutions**, driving improved decision-making and operational efficiency.
- Strong understanding of **data governance, security, and regulatory compliance** (GDPR, HIPAA, PCI-DSS), ensuring secure handling of sensitive enterprise data.
- Proven ability to work in **cross-functional, fast-paced environments**, collaborating with data engineers, DevOps teams, product managers, and business stakeholders to deliver scalable AI solutions.
- Recognized for strong **analytical thinking, problem-solving, and system design skills**, consistently reducing model error rates, improving system performance, and enabling data-driven decision-making at scale.

TECHNICAL SKILLS

Programming	Python (3.9+), SQL, Scala (basic)
ML/DL Frameworks	TensorFlow (2.x), PyTorch (1.12+), Scikit-learn (1.x), Keras
NLP	Hugging Face Transformers, BERT, GPT, spaCy (3.x), NLTK
Big Data	Apache Spark (3.x), Hadoop, Databricks
MLOps	MLflow (2.x), Kubeflow, Docker, Kubernetes (1.25+), Jenkins
Cloud	AWS (SageMaker, S3, Lambda, EC2), Azure ML
Data Tools	Snowflake, Airflow (2.x), Pandas, NumPy
Visualization	Tableau, Power BI, Matplotlib, Seaborn
APIs	FastAPI, Flask
Version Control	Git, GitHub
Methodologies	Agile, Scrum, ITIL

PROFESSIONAL EXPERIENCE

Client: KeyBank – Pittsburgh, PA
Role: Senior AI/ML Engineer
Duration: May 2024 – Present | USA

- Designed highly scalable machine learning and deep learning models using **PyTorch (2.x)** and **TensorFlow (2.11+)** for critical banking use cases including fraud detection, credit risk scoring, and transaction anomaly detection, improving detection accuracy by over **30%** and strengthening enterprise risk mitigation.
- Architected and implemented end-to-end machine learning pipelines using **Azure Machine Learning, Azure Data Lake Storage (ADLS Gen2), Azure Functions, and Azure Data Factory**, enabling automated workflows for model training, validation, deployment, and orchestration, reducing deployment cycle time by approximately **40%**.
- Built and deployed high-performance, real-time inference systems using **FastAPI, Docker, and Azure Kubernetes Service (AKS)**, supporting millions of financial transactions with low latency, high throughput, and high availability.
- Developed advanced NLP solutions using **BERT, RoBERTa, and Hugging Face Transformers (v4.x)** for financial document classification, KYC/AML text analysis, and customer sentiment modeling, improving operational efficiency and decision-making.
- Designed and implemented **Retrieval-Augmented Generation (RAG) pipelines** using **Azure OpenAI Service, LangChain, and Azure AI Search**, enabling secure, context-aware responses over enterprise financial data and internal knowledge systems.
- Built scalable **vector search solutions** using **Azure AI Search and FAISS**, enabling semantic document retrieval for applications such as regulatory Q&A, policy search, and internal knowledge assistants.
- Developed **LLM-powered intelligent assistants** for banking use cases including customer support automation, document summarization, and compliance reporting, significantly reducing manual effort and improving response accuracy.
- Engineered end-to-end **RAG workflows**, including document ingestion, embedding generation, indexing, retrieval, and prompt orchestration, ensuring accurate and grounded LLM outputs.
- Implemented **prompt engineering and response optimization techniques**, reducing hallucinations and improving factual accuracy of LLM-generated outputs in regulated financial environments.
- Integrated **security, governance, and access control mechanisms** within AI systems, ensuring compliance with **PCI-DSS, GDPR, and internal banking regulations**.
- Implemented enterprise-grade **MLOps pipelines** using **MLflow (2.x), Azure Machine Learning, Kubernetes (AKS), and Jenkins/Azure DevOps**, enabling automated CI/CD, model versioning, monitoring, and lifecycle management.

- Engineered scalable data processing pipelines using **Apache Spark (3.x)** on **Azure Databricks** and **Snowflake**, optimizing feature engineering workflows and reducing data processing time by approximately **35%**.
- Designed and deployed model monitoring and drift detection frameworks using **Azure Monitor, Application Insights, Prometheus**, and custom Python-based solutions to ensure model reliability and performance.
- Built and optimized recommendation systems using collaborative filtering and deep learning techniques, increasing customer engagement and product adoption rates by approximately **25%**.
- Applied advanced statistical modeling, hypothesis testing, and **A/B testing methodologies**, ensuring measurable, data-driven business impact across banking use cases.
- Performed extensive hyperparameter optimization using **Grid Search, Random Search, and Bayesian Optimization**, improving model precision, recall, and overall predictive performance.
- Designed and implemented **feature stores (Azure ML / Databricks Feature Store)** and reusable ML components, improving model reusability, consistency, and accelerating development cycles.
- Collaborated with cross-functional teams including data engineering, risk analytics, product, and business stakeholders to deliver scalable AI/ML solutions aligned with enterprise goals.
- Integrated CI/CD pipelines using **Git, Docker, Jenkins, and Azure DevOps**, enabling faster, reliable, and auditable deployment of machine learning models.
- Supported and maintained **mission-critical 24/7 production ML systems**, ensuring high availability, rapid incident response, and minimal downtime in high-stakes financial environments.

Client: CVS Health- Atlanta, GA
Role: AI/ML Engineer
Duration: Jan 2022 – Apr 2024 | USA

- Developed predictive machine learning models using **Scikit-learn (1.x)** and **TensorFlow (2.10+)** for healthcare use cases such as **patient risk stratification, hospital readmission prediction, and claims analytics**, improving prediction accuracy by approximately **28%** and enabling proactive clinical decision-making.
- Designed and implemented scalable, production-grade data pipelines using **Azure Data Factory**, **Azure Databricks**, and **Azure Data Lake Storage (ADLS)**, enabling efficient ingestion, validation, transformation, and orchestration of large-scale healthcare datasets for machine learning workflows.
- Built and optimized recommendation systems for **personalized healthcare interventions and treatment suggestions** using collaborative filtering and deep learning techniques, improving patient engagement and care outcomes.
- Engineered and deployed end-to-end machine learning systems using **Docker**, **Azure Kubernetes Service (AKS)**, and **Azure Machine Learning**, ensuring high availability, fault tolerance, and scalability for real-time healthcare applications.
- Performed extensive feature engineering and dataset construction using **Pandas, NumPy, and SQL**, including handling missing clinical data, normalization, encoding, and feature selection to improve model robustness and predictive performance.
- Developed and exposed RESTful APIs using **Flask** and **FastAPI**, deployed via **Azure App Service** and containerized environments, enabling low-latency, real-time inference for healthcare analytics systems.
- Utilized **MLflow (2.x)** integrated with **Azure Machine Learning** for experiment tracking, model versioning, and lifecycle management, ensuring reproducibility, auditability, and governance across ML systems.
- Leveraged **Apache Spark (3.x)** on **Azure Databricks** for distributed data processing and large-scale model training, reducing training time by approximately **35%** and improving system efficiency.
- Applied advanced hyperparameter optimization techniques, including **Grid Search, Random Search, and Bayesian Optimization**, improving model generalization and performance across diverse healthcare datasets.

- Developed NLP-based solutions using **spaCy (3.x)** and **Hugging Face Transformers**, enabling automated **clinical text classification, medical document summarization, and patient feedback analysis**.
- Implemented robust model monitoring, logging, and alerting systems using **Azure Monitor**, **Application Insights**, and custom Python-based solutions, detecting performance degradation, data drift, and anomalies in production environments.
- Collaborated with cross-functional teams including **data engineering, clinical analysts, and product stakeholders**, translating healthcare requirements into scalable AI/ML solutions aligned with organizational goals.
- Built interactive dashboards and reporting tools using **Power BI** and **Tableau**, enabling visualization of model outputs, patient risk scores, and key healthcare KPIs.
- Ensured compliance with healthcare data regulations such as **HIPAA**, implementing secure data handling, access control, and audit-ready ML systems.
- Automated end-to-end machine learning workflows using **Azure Machine Learning Pipelines**, **Azure DevOps**, **Jenkins**, and **Git**, improving deployment efficiency, reliability, and reducing manual intervention.

Client: GEICO – Lakeland, FL
Role: Machine Learning Engineer
Duration: Sep 2019 – Dec 2021 | USA

- Developed machine learning models for **insurance claim fraud detection** using **Python (3.7/3.8)** and **Scikit-learn**, improving fraud detection accuracy by over **25%** and reducing financial losses across high-volume claims processing systems.
- Designed and implemented scalable data processing pipelines using **Apache Spark (2.x/early 3.x)** and **SQL**, integrated with **Azure Data Lake Storage (ADLS)** for efficient storage and processing of large-scale insurance datasets, reducing data preparation time by approximately **30%**.
- Built predictive models for **claims severity estimation and policy risk scoring** using **TensorFlow (2.x)** and **Keras**, supporting underwriting strategies and improving loss ratio performance.
- Developed NLP-based solutions using **NLTK** and **early spaCy**, enabling automated extraction of key information from **claim reports, adjuster notes, and customer communications**, improving operational efficiency.
- Engineered and deployed production-ready ML services using **Flask-based REST APIs** and **Docker**, hosted on **Azure Virtual Machines** and **Azure App Service**, enabling near real-time inference and integration with internal systems.
- Utilized **Azure cloud services (Azure Blob Storage, Virtual Machines, and Azure Functions)** for data storage, batch processing, and lightweight model deployment workflows in a scalable cloud environment.
- Performed comprehensive feature engineering and data preprocessing using **Pandas, NumPy, and SQL**, including handling missing values, encoding categorical variables, normalization, and outlier detection to improve model performance.
- Applied hyperparameter tuning techniques such as **Grid Search and Random Search**, improving model precision, recall, and F1-score for fraud detection and risk modeling tasks.
- Implemented **basic model monitoring and logging mechanisms** using **Azure Monitor** and custom logging solutions, tracking model performance and detecting early signs of data drift.
- Collaborated closely with data engineering teams to integrate ML workflows into existing **ETL pipelines**, leveraging **Azure Data Factory** for data ingestion and transformation.
- Built dashboards and reporting solutions using **Power BI** and **Tableau**, enabling stakeholders to monitor fraud trends, claims insights, and key performance metrics.
- Ensured compliance with **insurance regulations and data privacy standards**, maintaining secure handling of sensitive policyholder and claims data.

- Partnered with business analysts and underwriting teams to translate insurance requirements into scalable ML solutions aligned with business goals.
- Automated ML workflows including **data ingestion, model training, validation, and deployment** using Python scripting and **Jenkins-based CI/CD pipelines**, improving operational efficiency.
- Supported and maintained **production ML systems in a 24/7 environment**, ensuring high availability, rapid issue resolution, and minimal disruption to insurance operations.

Client: Verizon – Sandy, UT
Role: Data Scientist / ML Engineer
Duration: Jul 2017 – Aug 2019 | USA

- Developed machine learning models using **Python (3.6/3.7)**, **Scikit-learn**, and **TensorFlow (1.x / early 2.0)** for telecom use cases such as **network anomaly detection, customer churn prediction, and service quality optimization**, improving prediction accuracy and reducing customer attrition.
- Processed and analyzed large-scale telecom datasets (network logs, call detail records, customer usage data) using **Apache Spark (2.x)** on **AWS EMR**, enabling efficient distributed data processing and reducing data latency for analytics workflows.
- Built anomaly detection models using statistical methods and **Scikit-learn**, identifying network performance issues and service disruptions, leading to improved network reliability and reduced downtime.
- Designed and orchestrated data pipelines using **Python, SQL**, and **Apache Airflow (1.x)**, ensuring reliable ingestion, transformation, and scheduling of high-volume telecom data for downstream analytics and ML applications.
- Developed time-series forecasting models using **LSTM (TensorFlow/Keras)** to predict **network traffic patterns and demand fluctuations**, improving capacity planning and resource allocation.
- Engineered and deployed machine learning services using **Flask-based REST APIs** and **Docker**, enabling scalable inference for internal telecom analytics platforms.
- Leveraged **AWS services including EC2, S3, Lambda, and EMR** for distributed data storage, batch processing, and model deployment, supporting scalable ML workflows in a cloud environment.
- Performed feature engineering and preprocessing using **Pandas and NumPy**, including handling missing data, normalization, and time-based feature extraction to improve model performance.
- Applied statistical analysis and hypothesis testing to validate models and support data-driven decision-making for telecom operations and customer analytics.
- Built dashboards and visualization tools using **Tableau** and **Power BI**, enabling stakeholders to monitor **network performance, customer trends, and key KPIs**.
- Implemented model performance tracking and logging mechanisms, ensuring consistent accuracy and reliability of ML models in production.
- Collaborated with cross-functional teams including **network engineers, operations teams, and business stakeholders**, aligning ML solutions with telecom infrastructure and business objectives.
- Automated data workflows and model retraining pipelines using **Python scripts and Jenkins-based CI/CD**, improving efficiency and reducing manual effort.
- Ensured adherence to **data governance, security, and telecom compliance standards**, maintaining integrity and confidentiality of customer and network data.
- Supported and maintained **production ML systems in a 24/7 environment**, ensuring high availability, rapid issue resolution, and minimal disruption to critical telecom services.

Client: Ruksun Software Technologies Ltd – Pune
Role: Junior Data Scientist / ML Engineer
Duration: Jun 2016 – May 2017 | India

- Developed machine learning models using **Python (Scikit-learn)** for client-focused analytics use cases such as **customer segmentation and sales forecasting**, improving prediction accuracy by **15–20%** across multiple projects.
- Performed end-to-end data preprocessing, cleaning, and transformation using **Pandas and NumPy**, improving dataset quality and enabling more efficient model development.
- Designed and executed **SQL-based data extraction and transformation workflows**, preparing structured datasets for analytics and reporting across client engagements.
- Built and evaluated **classification and regression models**, applying appropriate algorithms and validating performance using metrics such as **accuracy, precision, recall, and F1-score**.
- Implemented foundational NLP solutions using **NLTK** and **early spaCy**, enabling **text classification and keyword extraction** from unstructured data sources.
- Created data visualizations and dashboards using **Tableau** and **Matplotlib**, delivering actionable insights to business stakeholders and supporting data-driven decision-making.
- Developed ETL workflows using **Python**, automating data ingestion from structured and semi-structured sources (CSV, JSON, relational databases).
- Collaborated with senior data scientists and cross-functional teams to improve model performance through **feature engineering and basic hyperparameter tuning**.
- Gained hands-on exposure to **AWS services (EC2, S3)** for data storage and experimentation, supporting scalable data processing and model execution.
- Assisted in deploying machine learning models via **Flask-based APIs**, gaining initial experience in integrating models into application workflows.
- Participated in **Agile development processes**, contributing to sprint planning, task execution, and timely delivery of analytics and ML solutions.