**ARAVIND B** **AI/ML ENGINEER**

A blue square with white letters

Description automatically generated[LinkedIn](https://www.linkedin.com/in/aravind-b-bb7197231/) | A blue and white cell phone

Description automatically generated 813-568-1511 | A white and red envelope with a red line

Description automatically generated [baravind9406@gmail.com](mailto:baravind9406@gmail.com)

**Professional Summary:**

* **AI/ML Engineer** with 10 years of expertise in **Data Science**, **Machine Learning**, **Deep Learning**, and **Large Language Models (LLMs)**, specializing in building, fine-tuning, and deploying scalable AI-driven solutions.
* Expertise in developing, fine-tuning, and deploying **LLMs** and **Generative AI** using **GPT-4**, **Gemini AI**, **LLaMA**, **Falcon**, and **BLOOM**, with proficiency in **diffusion models**, **GANs**, **VAEs**, and **parameter-efficient fine-tuning** (LoRA, QLoRA, PEFT).
* Designed and deployed **personalized treatment recommendation systems** using **reinforcement learning** and patient history analysis to optimize medication plans and improve healthcare outcomes.
* Implemented **Retrieval-Augmented Generation (RAG)** using vector databases (**FAISS**, **Pinecone**, **Weaviate**, **ChromaDB**) for enterprise AI solutions; hands-on experience with **Python**, **TensorFlow**, **PyTorch**, and **Hugging Face Transformers**.
* Developed and optimized **NLP models** (e.g., **BERT**, **GPT**, **T5**, **RoBERTa**) with advanced tokenization techniques (**Word2Vec**, **FastText**, **BPE**, **Sentence Transformers**) for **chatbots**, **document summarization**, and **text analytics** using **NLTK**, **SpaCy**, and **OpenAI APIs**.
* Deployed **AI/ML models** on **AWS**, **Azure**, and **GCP** using **SageMaker**, **Vertex AI**, and **Azure ML**; skilled in **serverless architectures** (**AWS Lambda**, **Cloud Run**) and distributed model training with **PySpark** and **Databricks**.
* Built scalable **AI-driven data pipelines** using **Apache Airflow**, **Kafka**, and **Snowflake**, ensuring efficient data processing and real-time inference with integrated **feature stores** (**Feast**, **Databricks Feature Store**).
* Proficient in processing large-scale structured and unstructured data with **Python**, **PySpark**, **Snowflake**, **BigQuery**, and **Redshift**, enhancing real-time model inference and retraining workflows.
* Hands-on experience with **containerized deployments** using **Docker**, **Kubernetes**, and **Helm**, optimizing model serving with **KServe**, **TensorFlow Serving**, and **Triton Inference Server**.
* Utilized **MLflow**, **Kubeflow**, and **TensorFlow Extended (TFX)** for automated **model versioning**, **tracking**, and lifecycle management, ensuring continuous improvement and governance.
* Integrated **AWS SageMaker Model Monitor**, **GCP Vertex AI**, and **Azure ML** for **model drift detection**, **performance evaluation**, and **anomaly detection**, maintaining reliable AI performance.
* Built **HIPAA**- and **FHIR-compliant AI/ML models** for secure healthcare data processing and **EHR** integration, enhancing **predictive analytics** for patient outcomes.
* Proficient in **API development** using **FastAPI**, **Flask**, and **Uvicorn**, delivering asynchronous **AI inference services** and scalable API solutions for AI-powered applications.
* **Skilled in evaluation metrics** like **precision**, **recall**, **F1-score**, and **AUC-ROC**, ensuring **95%+ accuracy** and **90% precision** in **classification** and **regression** tasks.

**Skills:**

|  |  |
| --- | --- |
| Programing & Scripting Languages | Python, Pyspark, Java, SQL, Bash, Perl, YAML, Groovy |
| Machine Learning | Supervised Learning, Unsupervised Learning, Feature Engineering, Model Optimization, Time Series Analysis, Recommendation Systems, Sentiment Analysis, Evaluation Metrics. |
| Deep Learning | TensorFlow, PyTorch, Keras, CNNs, RNNs, LSTMs, GANs, Transformer Models, ANN, Transfer Learning, Ensemble Models, Computer Vision. |
| Large Language Models (LLMs) | GPT, BERT, LLaMA, Falcon, Hugging Face Transformers, LoRA, DeepSpeed. |
| Retrieval-Augmented Generation (RAG) | FAISS, ChromaDB, Pinecone, Weaviate, Llama Index. |
| Generative AI | Fine-tuning LLM’s, Langchain, Llama index, RAG, AI Agents, CrewAI, Auto Gen, Prompt Engineering, Hugging face, Phi, OpenAI, Llama, FIASS. |
| Big Data & Distributed Computing | HDFS, MapReduce, PySpark |
| Model Serving | FastAPI, Flask, Streamlit, Kserve, TensorFlow Serving, TorchServe, Cloud Run, Vertex AI Endpoints, SageMaker Endpoints, Azure ML endpoints. |
| Cloud Platforms | **AWS** (SageMaker, Bedrock, Lambda, Lex, CloudWatch, CloudTrail, Redshift ML, DynamoDB, CodeBuild, CodeDeploy, S3, EC2, IAM, AMIs). **GCP** (Vertex AI, AutoML, Cloud Vision API, Dialogflow, NVIDIA GPUs, BigQuery ML, VM Instance, VPC).  **Azure** (Azure ML, Azure AI & OpenAI, Blob Storage, Azure Functions, Azure Cognitive Services) |
| MLOps & CI/CD | GitHub Actions, GitLab CI/CD, Bitbucket Pipelines, Jenkins, CircleCI, Cloud Build, CodePipeline, Azure Pipelines, MLflow, Kubeflow, DVC, DagsHub |
| Monitoring & Logging | AWS CloudWatch, GCP Cloud Monitoring, Azure Monitor, CloudTrail Logs, Prometheus, Grafana, MLflow, Weights & Biases, TensorBoard, Evidently AI, Vertex AI Model Monitoring, Cloud Logging |
| MLOps & Workflow Orchestration | Kubeflow, MLflow, Airflow, TFX Pipelines, Vertex AI Pipelines, Supervised/Unsupervised Algorithms, ANN, CNN, NLP, Computer Vision, GAN’s, LSTM, Feature Engineering, Transfer Learning, ensemble models, Time series, Recommendation systems, Sentiment Analysis, Evaluation Metrics.  . |
| Frame Works | PySpark, Tensorflow, Keras, PyTorch, Scikit-Learn, openCV, NLTK, Pandas, Transformers, Flask, Celery, FastAPI, Streamlit, Gradio, Pickle, Pydantic, Anaconda, Jupyter Notebook. |
| Version Control | GIT, GitHub, Azure Repos, AWS Code Commit. |
| Data Warehousing | Snowflake, Data Lake Storage, Google Big Query. |
| Container Orchestration and Infrastructure as Code | Docker, Kubernetes, Terraform. |
| Operating Systems | Windows, Ubuntu, Linux. |

**Certifications:**

* Azure Certification - [Link](https://learn.microsoft.com/en-us/users/bejawadaaravind-1668/transcript/d8222cx08kqw23w)

**Professional Experience:**

**Client: Motive MI, SFO, CA(Remote) Jun 2023 - Jan 2025**

**Role: Senior AI/ML GenAI Engineer**

**Responsibilities:**

* Developed **Generative AI-powered chatbots** using **AWS Bedrock** and fine-tuned **LLMs** (e.g., **GPT-4**, **Falcon**, **T5**) for real-time patient symptom analysis and personalized treatment recommendations.
* Implemented **Retrieval-Augmented Generation (RAG)** with **FAISS** and **AWS OpenSearch** to enhance chatbot accuracy by integrating real-time **EHR** and clinical trial data.
* Leveraged **Deep Learning** and **NLP** models (e.g., **BERT**, **T5**) with **SpaCy** for clinical text processing, enabling the chatbot to extract vital health indicators from unstructured medical documents.
* Designed and deployed **LLM-powered chatbot APIs** using **TensorFlow Serving**, **PyTorch**, **FastAPI**, and **Google Vertex AI**, ensuring low-latency inference and scalable deployment for real-time medical queries.
* Developed feature engineering pipelines using **PySpark**, **Pandas**, **NumPy**, **SciPy**, and **scikit-learn** for efficient data ingestion from **AWS S3**, **Google Vertex AI**, and **Snowflake** to optimize training data.
* Built secure, scalable, and compliant chatbot solutions with **SageMaker endpoints**, **Google Vertex AI**, **Docker**, and **Kubernetes**, ensuring HIPAA-compliant inferencing.
* Integrated **AWS CloudWatch**, **Prometheus**, and **Grafana** with **Google Vertex AI** for real-time chatbot monitoring, model drift detection, and automated retraining.
* Deployed fine-tuned **LLMs** on **AWS SageMaker**, **AWS Bedrock**, and **Google Vertex AI** for high-performance inference, powering the chatbot’s natural language processing capabilities.
* Applied **TF-IDF**, **BERT embeddings**, and **PCA** for text feature extraction and dimensionality reduction, improving chatbot accuracy in understanding medical queries.
* Developed **AI/ML models** for secure healthcare data processing and predictive analytics using **Python**, **TensorFlow**, **PyTorch**, and **NLP libraries**, ensuring compliance with healthcare regulations like **HIPAA** and **GDPR.**
* Preprocessed **medical images** using **OpenCV** and wavelet transforms, enabling multi-modal capabilities in the chatbot to process and analyze medical visuals, all through **Python**.
* Implemented **Named Entity Recognition (NER)** using **SpaCy** and **BioBERT** in **Python**, enabling the chatbot to extract important medical entities such as symptoms and treatments from unstructured clinical texts.
* Fine-tuned **GPT-4**, **LLaMA**, and **BioBERT** models for **clinical text analysis** and **AI-driven diagnosis** within the chatbot, optimizing performance using **LoRA**, **quantization**, **TensorRT** with **TensorFlow** and **PyTorch** to reduce inference latency and improve conversational accuracy.

**Client: OPTUM, Albany, NY Dec 2021 – May 2023**

**Role: AI/ML Engineer**

**Responsibilities:**

* **Developed a predictive healthcare analytics system** to assess **patient readmission risks** using **Azure ML, Snowflake, and Python**.
* **Designed and implemented machine learning models** (**Logistic Regression, Random Forest, XGBoost, and LSTMs**) using **TensorFlow and PyTorch** for time-series patient history analysis.
* **Processed and cleaned structured & unstructured medical data** from **Azure Data Lake & Snowflake**, handling missing values, feature engineering, and text-to-numeric transformations.
* **Built an NLP pipeline** using **TensorFlow and PyTorch** for analyzing doctor’s notes and extracting clinical insights with **Named Entity Recognition (NER)**.
* **Deployed ML models as RESTful APIs** using **FastAPI and TensorFlow Serving**, enabling real-time predictions for hospitals and medical staff.
* **Optimized model inference performance** with **ONNX and TensorRT**, reducing latency by 40% for real-time decision-making.
* **Implemented dimensionality reduction techniques (PCA, t-SNE)** to enhance feature selection and improve model efficiency on large-scale patient datasets.
* **Integrated real-time monitoring** with **Azure Monitor, Prometheus, and MLflow**, tracking model drift and performance metrics in production.
* **Ensured data security and compliance** by implementing **RBAC (Role-Based Access Control), HIPAA, and GDPR standards** with **Azure Key Vault** for encrypted storage.
* **Developed a self-learning pipeline** for continuous model retraining using new patient data, automating updates through **Azure ML Pipelines and Databricks Notebooks**.
* **Designed and maintained an end-to-end CI/CD pipeline** for ML model deployment using **GitHub Actions, Docker, and Kubernetes** in **Azure Kubernetes Service (AKS)**.
* **Collaborated with cross-functional teams** including data scientists, cloud engineers, and healthcare professionals to improve **patient risk prediction accuracy by 30%**.

**Client: British American Tobacco, Albany, NY Mar 2020 - Nov 2021**

**Role: ML Engineer**

**Responsibilities:**

* **Designed and implemented a predictive analytics solution** using **Azure Machine Learning and Databricks** to forecast both **cost changes and future sales volumes** of tobacco products.
* **Built ML models using Linear Regression, Gradient Boosting, SVM, ElasticNet, Random Forest, and K-Means clustering** to analyze cost trends, demand patterns, and optimal pricing strategies.
* **Developed large-scale data pipelines in PySpark within Databricks Notebooks** to process **past 3 years of historical sales data** from **Snowflake** and predict **sales amounts and costs for the next 2 years**.
* **Implemented feature engineering and data transformation** techniques to extract meaningful insights from sales data, customer trends, and external market factors.
* **Integrated Snowflake as a scalable data source** for both **batch and real-time ingestion**, enabling automated data pipelines using **Azure Data Factory**.
* **Trained and optimized ML models** to minimize forecasting errors and improve **cost efficiency and sales planning accuracy** by leveraging **Azure ML AutoML and Hyperparameter tuning**.
* **Deployed predictive models as endpoints in Azure ML** and exposed them via **FastAPI and Azure Functions** for real-time inference and integration with business applications.
* **Implemented CI/CD pipelines with Azure DevOps, GitHub Actions, and MLflow** to automate model retraining, deployment, and versioning for continuous improvement.
* **Monitored model performance, accuracy, and drift** using **Azure Monitor, MLflow, Prometheus, and Grafana**, ensuring reliability in sales and cost predictions.
* **Developed dashboards and visualizations in Power BI** to provide stakeholders with insights into **future sales forecasts, pricing strategies, and demand planning**.
* **Reduced forecasting errors by 25% and improved sales planning efficiency**, enabling optimized production and pricing decisions for British American Tobacco.

**Maveric Systems, Chennai, India Apr 2018 - Feb 2020**

**Role: Data Scientist & DevOps Engineer**

**Responsibilities:**

* Developed **predictive models** using **Python (pandas, NumPy, SciPy, scikit-learn)** for credit risk and fraud detection.
* Designed **feature engineering pipelines**, handling missing values, outliers, and categorical encoding, improving model accuracy by **XX%**.
* Designed and implemented end-to-end **machine learning pipelines** using **PySpark** and **Databricks Notebooks** for sales forecasting.
* Built **time-series forecasting models (ARIMA, Holt-Winters, Exponential Smoothing)** to predict customer default risks.
* Developed **ETL workflows** using **AWS Glue and PySpark**, processing structured and semi-structured data.
* Deployed trained models as **Flask APIs** on **AWS EC2**, integrating them into real-time risk analysis systems.
* Automated **CI/CD pipelines** with **Jenkins, AWS CodePipeline, and Docker**, ensuring seamless model deployment.
* Configured **AWS S3 for data storage** with lifecycle rules for cost optimization.
* Monitored and logged model performance using **AWS CloudWatch and Prometheus**.
* Integrated models into business intelligence tools like **Tableau and Power BI** for reporting and analytics.

**Client Name: ACE Technologies, Hyderabad, India Jul 2015 - Mar 2018**

**Role: DevOps Engineer (Data Scientist)**

**Responsibilities:**

* Conducted **customer segmentation** using **K-Means Clustering, Hierarchical Clustering**, and **Principal Component Analysis (PCA)**.
* Developed **Python-based data pipelines** using **pandas, NumPy, and PySpark**, processing large financial datasets for risk modeling.
* Designed **Python-based data preprocessing modules**, handling missing values, feature scaling, and categorical encoding.
* Built **demand forecasting models** using **Linear Regression, Decision Trees, and Random Forest** with **Python**.
* Created **ETL pipelines with Azure Data Factory**, automating data ingestion and transformation from multiple sources.
* Managed data in **Azure SQL Database and Azure Blob Storage**, optimizing performance with indexing and partitioning.
* Deployed **REST APIs with FastAPI and Flask** on **Azure App Services** for real-time analytics.
* Automated model retraining and deployment using **Azure DevOps pipelines**, reducing deployment time by **XX%**.
* Set up **monitoring and logging** with **Azure Monitor and Application Insights**, tracking API usage and errors.
* Developed interactive **Power BI dashboards**, providing stakeholders with insights on customer behavior and sales trends.
* Secured data access with **Azure RBAC, Managed Identities, and encryption**, ensuring compliance with data security policies.

**Education**:

* Bachelors in Computer Science – JNTUK(2011-2015)