Pradeep Pujari

Home: 510-894-1468 https://www.kaggle.com/ppujari **Email:** ppujari@hotmail.com
LinkedIn Twitter Github Google Scholar

OBJECTIVE:

A highly innovative and results-oriented AI Engineer with expertise in Deep Neural Network, Machine Learning, Natural Language Processing (NLP), Generative AI (LLMs), Search Science and scalable web architecture. I am seeking a challenging role where I can apply my technical acumen and problem-solving abilities to deliver advanced, customer-focused software solutions.

CORE COMPETENCIES:

- Extensively worked on Machine Learning, **Deep Neural Networks**, **Computer Vision**, Large Language Models (LLM), Reinforcement Learning, Natural Language Understanding (NLU), Graph Neural Networks, Transformer Model, Knowledge Graphs, and Information Retrieval.
- Hands-on experience with Big Data technologies including Hadoop, Spark and Cloud Computing.
- Experience in architecting, designing, and coding scalable e-commerce platforms**.**
- Proficient in **data structures, algorithm design and complexity analysis**
- Hands-on experience in RDBMS, NO-SQL database Cassandra, Mongo, Vector database FAISS
- Good at Object Oriented Design principles, Design Patterns, middleware components.
- Hands-on experience in RAG, prompt engineering, CoT, LangChain, LangGraph, DSPy
- Hands-on experience Data quality control, Versioning, Standardization, Data set generation
- Enjoy Mentoring and fostering team cohesion, work effectively in cross-functional teams!

TECHNICAL SKILLS:

Programming Languages	**Java**, **Python**, C++, Pandas, NumPy
ML/Deep Learning Framework	Scikit-Learn**,** TensorFlow, Keras**, PyTorch,** OpenVINO, **Apache TVM**
Machine Vision Tool Kit	OpenCV, CNN, Faster R-CNN, YOLO, **CLIP, BLIP, GLIDE**
Graph Neural Network	NetworkX, PyTorch Geometric
MLOps	Apache Ray, SageMaker, Kubernetes, FeatureStore – FEAST, ONNX
NLP Tool Kit	**spaCy**, Core NLP**, Chain of Thought, BART, LangChain, DSPy**
Search Science Technology	Lucene-Solr, Elastic Search, Nutch, **Neural IR,** WordNet, LLMs
Distributed Computing	Redis, Hadoop, PySpark, Hive, ZooKeeper, **Amazon Bedrock**
Networking & Operating System	Edge Caching, UNIX internals, Tomcat, Flask
SQL/NoSQL/Vector Database	Cassandra, Mongo, Hbase, Oracle, PostgreSQL, FAISS, WandB
Distributed Streaming	Kafka, KSQL, Faust, Apache Storm, REST API
GenAI/LLM	Agentic LLM, **AutoGen**, **BERT**, T5, GPT, **Prompt, cursor**

OPEN-SOURCE CONTRIBUTION: Lucene-Solr Project, OpenAI

- Created multi-agent collaboration projects such as training agent pairs for tennis @Udacity.
- Implemented DDPG paper in PyTorch, evaluated deep RL models.
- Designed and coded Neural Machine Translation @Udacity
- Road lane detection, path planning @Udacity Self driving course.
- **LLM - Detect AI Generated Text @Kaggle My model accuracy was 0.851**
- Worked on Few-shot Question Answer LLM Model and Identifying Age-Related Conditions
- **Vision Language Models** – Stable Diffusion, BLIP, Interrogation CLIP@Huggingface,

Technology: DDPG-Actor-Critic, LangChain, LangGraph, **PyTorch, BERT,** LLaMA

WORK EXPERIENCE:

EleutherAI – AI Research Scientist **May2024-Present**

- Collaborated with researchers to develop evaluation metrics, run performance evaluation pipelines, and debug LLM models.

- Researching AI interpretability and alignment to enhance model transparency, ethical AI behavior, and adherence to human values.

Technology: Training LLM, GPT-J, GPT-NeoX, Pythia suite, Wandb, LangGraph

GenAI Architect – KP Digital: **Feb 2022-Dec2024**

- Designed and implemented a conversational AI chatbot for customer service, reducing support call volume by 25% and improving response accuracy. RAG Architecture to reduce hallucination

Natural Language Processing for Literature Mining project:

- Applied NLP techniques for knowledge extraction from scientific literature and databases. Developed text mining tools for literature curation, knowledge discovery, and hypothesis generation. Identified and anonymized PII data.
- Designed and built scalable evaluation systems that automated model assessments, integrating seamlessly with CI/CD pipelines.

Technology: Azure, Kubernetes, LLM, **PyTorch, T5, BART, MedAlpaca,** LlaMA2, RLHF, LangChain

Research Scientist-Meta: **Oct 2022 - Apr2023**

- Benchmarked and optimized **data and model parallelism** for a large-scale **SparseNN Ad Ranking ML model** on cutting-edge AI accelerators and chips, improving performance and efficiency for Meta's high-traffic advertising infrastructure.
- Developed and enhanced a **personalization model** that tailored ad content to individual user preferences, leveraging real-time data and optimizing for both relevance and engagement, which led to improved user experience and ad effectiveness.
- Applied hyperparameter tuning, quantization, and QLoRA to optimize memory usage and accelerate inference while maintaining model accuracy.
- Leveraged LLVM-IR to perform advanced code analysis and optimizations, enabling efficient translation of high-level languages into optimized machine code for multiple architectures.
- Collaborated with cross-functional teams to integrate new hardware, utilizing **Python** and **CUDA** for AI model acceleration and GPU optimization.

Technologies : PyTorch, Quantization Techniques, LLM, QLoRA, Python, CUDA Programming

Principal NLP Engineer – CVSHealth (contract) **Apr 2021 - Aug2021**

- Medical Imaging for Tumor Detection- involves several imaging modalities, advanced image processing techniques, and machine learning algorithms to accurately identify and diagnose tumors.
- Collected Dataset: ChestX-ray14, ISIC, The Cancer Imaging Archive (TCIA), **BRATS**: Brain Tumor Segmentation Challenge dataset. Deployed the model in a clinical setting.

Technology: Python, PyTorch, **OpenCV, scikit-image, CNN,** AWS, MLflow, train LLM models

Principal Machine Learning Engineer – Oto Analytics **Feb 2021 - Jul2021**

- Designed, developed, and implemented **Feature Store -Feast**, ML pipeline.

Technology: Python, Redis, Apache Airflow, Apache Kafka, AWS, MLflow, **Amazon Bedrock**

Machine Learning Architect - ServiceNow (contract) **Aug 2020 - Jan2021**

- Worked on Intent extraction and Question Answering app using **BERT**, NLU
- Architected and Built Chatbot conversational AI system for Help Desk with RASA NLU Lib

Technology: Python, Transformers, AWS, Mongo, Nutch, RASA – NLU, SquAD, BART, T5

Principal Research Scientist (ML) - Kohl's Innovation Lab **Jan 2018 - Jul2020**

- Implemented real-time object detection, segmentation, gesture recognition system non-rigid tracking with pre-trained deep learning models YOLO, SSD
- This project involves FPGA programming, interfacing with camera modules, and optimizing the inference pipeline for real-time performance.
- Store Shopping Intent Analysis models for **people counter, Age and Gender identification.**
- **Optimized and deployed machine learning models on edge devices using Apache TVM,**

enabling efficient inference with minimal latency and resource utilization.
Technology: Python, pyTorch, OpenCV, spaCy, LLM, YOLO, Fast R-CNN, OpenVINO, TVM

Staff Machine Learning Engineer - Walmart Labs **May 2015 - Mar 2017**

- Built ML models for attribute extraction such as detecting brand, color, size etc. with
  - A. Supervised Learning B. Sequence Labelling and CRF C. **CNN** approach.
  - Relation extraction, Semantic Parsing, and knowledge graph
- Implemented Neural Network model for Item Matching System based on title and description.
  - Used pre trained Word embedding – **Glove**-contextual word representation.
- Worked in distributed data pipeline to orchestrate raw item JSON through a series of **micro services** producing a sellable item.
- High-Performance Computing (HPC) workloads: Large-Scale Machine Learning Training: Built distributed machine learning frameworks to train deep neural networks on massive datasets.

Technology: Python, Scikit Learn, **TensorFlow**, **Keras**, ZooKeeper, Kafka, **Storm**, Cassandra, Mongo

Senior Architect - Angie's List **Mar 2012 - Mar 2014**

- Designed and programmed Query understanding and rewriting with a heuristic Ranking module.
- Developed **Geospatial Search**, boosting revenue and retention by 25%.
- Developed and implemented a data discovery pipeline to identify, classify, and analyze structured and unstructured data, enabling improved data management and actionable insights.
- Designed, implemented a cold start solution to optimize recommendations and user experience by leveraging data analysis, machine learning models, and feature engineering for new users and items
- Implemented **Multi Objective Session based Recommender System** for Deals, Coupons

Technology: Java/J2EE, **Lucene-Solr**, Python, Multitask Ranking, Kea, Ling Pipe, NER, Elastic Search

Staff Software Engineer - Walmart Labs Search and Platform Team **Jun 2010 - Mar 2012**

- Designed Sentiment Analysis Network: Product Reviews data set – sentiment aware tokenizer.
- Built Meta-Search Engine with Apache Carrot, Enhanced content with annotation engine-UIMA.
- Object Tracking in video, Handwritten digit recognition, Hand gesture recognition, pose detection

Technology: Scikit Learn, SentiWordNet, SGD, Sentiment Treebank, Word2Vec, OpenCV, Flask, REST

Tech Leader – Search Science - Macys.com **May 2001 - May 2010**

- Implemented automatic IR Evaluation System. **MS-MARCO** document ranking dataset, Real-Time Indexing, Search Ranking Algorithms, Auto Suggest component, Sponsored Search
- Implemented web crawler-NUTCH secure and static web pages, Faceted Search and Browse
- New Ranking algorithm for Search Relevancy Tuning, Query Understanding and rewrite,
- Implemented Deep Personalization ranking mechanism based on a user's search and click history.

Technology: Java J2EE, Tomcat, Lucene-Solr, Vector Space Model, **BM25**, Learn to Rank, Elastic Search

EDUCATION

MS in Computer Science and Applications - National Institute of Technology, India Jun 1988
B. Sc (Physics Hons) - Berhampur University, India Jun 1984
Data Science Specialization – Johns Hopkins University May 2015
AI Specialization 3 Semester course certification– Stanford University July 2009

PROFESSIONAL DEVELOPMENT

Udacity Project: Home Service Robot **May–Dec2020**
Programmed a Home service Robot that maps environment and navigates to pick up and deliver objects.
Technology: C++, ROS, Monte Carlo Localization, Path Planning

Udacity Project: Self Driving Car Nano degree **Oct – Dec2017**
Completed: a. Advanced Lane Lines detection b. Vehicle Detection c. Extended Kalman Filter
d. Traffic Sign Classifier e. behavioral cloning f. controls MPC e. PID control, Simulation and Testing
Technology: C++, PyTorch, OpenCV, PCL (Point Cloud Library), Mask R-CNN

PUBLICATIONS AND PRESENTATIONS:

Published Paper "Detecting Cyber bullying instances" https://arxiv.org/abs/2411.05958 Apr 2022
Coauthor of Book "Practical Convolutional Neural Network" – Packt Publisher Feb 2018
Presented "Sentiment Analysis with Solr" at Sentiment Symposium, San Francisco May 2013