Binay Kumar Pradhan

AI Engineer | Founding Engineer | Software Engineer

PERSONAL DETAILS

Address	Cottbus, Germany
Email	pradhan.binay.de@gmail.com
LinkedIn	linkedin.com/in/i-binay
GitHub	github.com/vinayin
Website	vinayin.github.io
Availability	Immediately

PROFESSIONAL SUMMARY

Master's student in Artificial Intelligence with a focus in model compression (pruning, quantization, knowledge distillation), reinforcement learning, ML solution design, and LLMs. Proven track record of over four years of experience in MLOps and software engineering, including backend development, data science and deployment of scalable end-to-end machine learning pipelines on cloud platforms of AWS and GCP.

Seeking AI Engineer related role to deliver production-grade solutions with improved performance, cost-efficiency, and robustness, contributing to sustainable development through ethical AI applications.

WORK EXPERIENCE

Research Assistant

06.2023 – 03.2025

Brandenburg University of Technology (BTU) Cottbus-Senftenberg | Cottbus, Germany

- Engineered a cross-language data visualisation tool using PyQt6 by integrating Python, Java, and MATLAB subroutines and automating MongoDB data extraction pipelines, accelerating research workflows.
- Architected distributed reinforcement learning simulation pipelines for supply chain optimisation (stockyard management) by formulating the problem as a time-cost scheduling task and applying constraint programming, deriving efficient and sustainable logistics strategies scalable to large-scale distribution scenarios.
- Accelerated participant session throughput by 50 % by engineering targeted EEG calibration paradigms and automating data analysis pipelines, eliminating manual bottlenecks in data-intensive research environments.
- Engineered and deployed *pyETA-toolbox* to PyPI, a production-ready open-source toolkit for real-time fixation detection and eye-tracker validation in passive BCI research, demonstrating end-to-end delivery for biomedical signal processing (pypi.org/project/pyETA-toolbox).

AI Data Engineer II

06.2022 – 09.2022

Wolkus Technology Solutions Pvt Ltd | Bangalore, India

- Architected and deployed ETL pipelines on AWS SageMaker by orchestrating a full GCP-to-AWS data mart migration and enhancing end-to-end observability via Elasticsearch integration, supporting scalable ML training workflows.
- Reduced p95 API latency and AWS Lambda cold-start overhead by 20 % by refactoring the serverless codebase, optimising cloud compute efficiency for AI-driven production workloads.
- Engineered a logger and published *pylogger-slack* to PyPI, an open-source structured logging library delivering real-time team observability with Slack integration for production ML pipelines (pypi.org/project/pylogger-slack).

Software Engineer II
10.2021 – 03.2022
Myelin Foundry Pvt Ltd | Bangalore, India

- Developed in-car dashboard AI features for facial emotion recognition, hand gesture detection, and object identification using the MediaPipe computer vision framework.
- Deployed production-grade deep learning models on automotive edge devices by applying model compression techniques (quantization and pruning), achieving low-latency inference on ARM architecture and establishing foundational expertise in cost-efficient implementations.

Data Science Engineer
07.2019 – 10.2021
Utopia India Pvt Ltd | Bangalore, India

- Increased client engagement by 30 % in Q1–Q3 2021 by developing REST API modules, resolving architecture bottlenecks, and delivering enhanced data visualisations for the 4C project (Classify, Collect/Scrape, Correlate/Map, Clean/Standardize).
- Developed a proof-of-concept reinforcement learning model for automated text labelling with real-time human feedback.
- Implemented another proof-of-concept for industrial text generation using BERT and GPT, demonstrating applicability to free-form text.
- Achieved 85 % extraction accuracy and 15× efficiency gains over manual processes by engineering a regex-based information extraction parser with consensus algorithms.
- Maintained CI/CD pipelines, structured logging observability, and cost-optimised AWS SPOT instance load distribution, enabling seamless distributed ML workflows and production-grade delivery for the ML team.
- Designed an image-matching algorithm for automated PDF text segmentation.
- Spearheaded requirements gathering and cross-functional coordination for a 10-member ML team on the 4C data-cleaning and standardisation project, ensuring seamless integration of custom datasets.

Machine Learning Intern
01.2019 – 07.2019
Utopia India Pvt Ltd | Bangalore, India

- Designed and evaluated text-matching algorithms using pre-trained word vectors, custom embeddings, fuzzy logic, and set-relation techniques to improve entity-matching accuracy.
- Developed data-driven NLP algorithms for automated standardisation and structuring of free-form text labels and values.
- Built a batch data-cleansing pipeline using Makefile-driven workflows, enabling scalable large-scale file processing.

Research Intern
05.2018 – 07.2018
Institute of Development and Research in Banking Technology (IDRBT) | Hyderabad, India

- Curated and annotated Hindi–Bengali code-mixed datasets via the Twitter API, contributing to POS tagging research published in ACM TALLIP (2020) with F1-score improvements of 5–6 %, building foundational expertise in linguistic data processing.

PROJECTS

TechEU Editor
07.2025
Tech Europe Hackathon | Berlin, Germany

- Architected a multi-agent orchestration pipeline using CrewAI and OpenAI by deploying specialised agents to autonomously scrape sources, generate structured articles, manage citations, and apply targeted text modifications, enabling end-to-end editorial automation for scholars (github.com/VinayIN/tech-europe-hackathon).

HR Assist System — (Erasmus+ Entrepreneurship Labs 2025)
03.2025 – 07.2025
University of Catania | Catania, Italy

- Developed and prototyped an AI-driven B2B SaaS product by applying unsupervised clustering algorithms to employee performance reviews, skill profiles, and availability, automatically generating optimally balanced team configurations for team building activities.

- Pitched the product concept as part of the Erasmus+ Entrepreneurship Labs 2025 programme to investors and mentors, advancing its translation into a commercially viable student start-up.

EDUCATION

M.Sc. Artificial Intelligence 10.2022 – expected 2026
Brandenburg University of Technology (BTU) Cottbus-Senftenberg | Cottbus, Germany

- Grade: 2.6
- Thesis (in progress)

B.Tech. Information Technology 08.2015 – 06.2019
International Institute of Information Technology Bhubaneswar (IIIT-B) | Bhubaneswar, India

- Grade: 2.8
- Thesis: *Unstructured Text to Structured Text Mapping*, NLP-based techniques for transforming free-form texts and structuring them based on a template.

TECHNICAL SKILLS

Programming Languages: Python, Java, MATLAB, Bash
ML / AI Frameworks: PyTorch, TensorFlow, HuggingFace, Scikit-learn, ONNX, MediaPipe
Generative AI & LLMs: CrewAI, Transformers, Retrieval Augmented Generation (RAG), Agentic AI
AI / ML Specialisations: Deep Learning, Natural Language Processing (NLP), Computer Vision, Model Compression (Pruning, Quantization, Knowledge Distillation), Reinforcement Learning, Speech Recognition, Brain-Computer Interface (BCI)
MLOps & Cloud: MLflow, Docker, CI/CD, Git, REST APIs, Makefile, AWS (SageMaker, Lambda, SPOT Instances), GCP (Firebase)
Databases: MongoDB, Elasticsearch, PostgreSQL, weaviate, supabase
Data Visualisation: Streamlit, Gradio, PyQt6

LANGUAGES

English	C2 — Fluent
German	A2 — Elementary

PUBLICATIONS

Bhattu, S. N., Nunna, S. K., Somayajulu, D. V. L. N., and **Pradhan, B.** (2020). *Improving Code-mixed POS Tagging Using Code-mixed Embeddings*. *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*, Vol. 19, No. 4, Article 50, pp. 1–31. Published: 29 March 2020.
DOI: 10.1145/3380967
Proposed a hierarchical deep recurrent neural network with a CRF layer to improve POS tagging of code-mixed social media text (Bengali-English, Hindi-English, Telugu-English). Developed custom code-mixed word embeddings; achieved F1-score improvements of 5.81 %, 5.69 %, and 6.3 % over CRF baselines across the three language pairs.