

ASWIN JOSE

Lead ML Engineer · Applied AI / LLM Systems · Bengaluru, India
mathew.aswin.jose@gmail.com · linkedin.com/in/aswin-jose · github.com/meagoodboy

SUMMARY

ML engineer who likes owning the whole problem, not a slice of it. At Sonus Health, built the ML audio engine behind the company's first public product and took it live in under 8 months, working alongside the backend and app engineers. Before that, cut search latency from 4s to 300ms on a retail platform with a million daily users. Looking for a small team that ships fast.

EXPERIENCE

Lead ML Engineer, Sonus Health / Decorte Future Industries

Mar 2025 – Present
London, UK (Remote)

Health AI · smartphone-based cardiac screening

- Took the company's first public-facing product, smartphone-based cardiac screening, from zero to live in under 8 months alongside its backend and app engineers: data acquisition, annotation, model design, deployment, and monitoring. Own the ML audio-analysis engine end to end.
- Customised the vitals-extraction pipeline for a leading automotive manufacturer's in-vehicle cardiac-analysis pilot, working directly with customer stakeholders; extracted interpretable vitals from in-cabin microphone audio in 85% of cases, a previously untested setting.
- Shipped an end-to-end LLM-in-the-loop pipeline for clinical-level reporting: LLM calls validate model output and generate reports, with human review before release; built a harness and safety gating this needs at clinical stakes.
- Drove research on cardiac capture from bare smartphone microphones, no added hardware, to state-of-the-art results among market offerings.

Software Engineer, Couture.ai

Jun 2024 – Mar 2025
Bengaluru, India

Personalisation & search at e-commerce scale

- Built a RAG-based vector-search pipeline 0 to 1, then scaled it to production for a major retail platform with 1M+ products and 1M daily active users.
- Cut search latency from 4s to 300ms by redesigning the service architecture and optimising the database and retrieval layers: asynchronous high-throughput services, caching, and tuned vector retrieval.
- Deployed and tuned the pipeline hands-on with customer engineering teams against their data, traffic patterns, and business requirements.

Undergraduate Researcher, CCNSB (IIIT Hyderabad)

May 2021 – Jun 2024
Hyderabad, India

Graph ML for cancer genomics

- Researched graph-based ML pipelines (graph auto-encoders, VAEs) for predicting cancer patient outcomes from multi-omics data using cheap, clinically feasible inputs, benchmarked against state of the art; built the end-to-end survival-prediction pipeline: **DeepGraphMut** (Briefings in Bioinformatics, 2025).
- Developed a network-diffusion approach to renal cell cancer stratification, identifying biologically relevant subtypes and the genetic factors driving survival (Mol Genet Genomics, 2023).

Machine Learning Intern, VoltaML

Jul 2022 – Oct 2022
California, US (Remote)

Inference optimisation

- Cut XGBoost inference time using the lileaves compiler, outperforming state-of-the-art approaches like treelite by at least $2.5\times$; benchmarked regression and diffusion models for inference speed and deployability.

PROJECTS

- **Assistbot**: an always-on personal agent platform. A fleet of role-specific subagents which share persistent file-based memory, custom skills, and launchd-scheduled autonomous runs.
- **hackbait** (Python): Honeypot environments that measure reward hacking in LLM coding agents; structural verdicts, a control arm, and benchmark runs against frontier Anthropic and Gemini models.
- **rl-envdoctor** (Python, on PyPI): Adversarial QA for LLM-RL training environments; model-free grader audits at zero API cost.
- **score2midi** (tau-labs.com): Sheet music to MIDI end to end: optical music recognition pipeline deployed on Cloud Run, live in production.
- **PianoLearn** (Kotlin/Android): fully on-device app for learning piano from sheet music: MusicXML scores, MIDI and mic input, and a practice engine.
- **Trading data infrastructure** (Python): market-data platform covering 2,284 NSE/BSE stocks and 7.1M+ OHLCV records over 26 years, with news, fundamentals, and economic-data ingestion and correlation analytics on top.
- **NooooSH** (C): Unix shell written from scratch over raw Linux syscalls: piping, redirection, background job control, and command history.

TECHNICAL SKILLS

ML / AI PyTorch, LLM pipelines & prompt orchestration, RAG, embeddings & vector search, model calibration, GNNs, XGBoost, audio & signal ML, ONNX
Engineering Python, C++, C, JavaScript, Shell, SQL, FastAPI, Docker, Kubernetes, AWS, Kafka, Redis, PostgreSQL, MongoDB, Qdrant, Git

EDUCATION

B.Tech in Computer Science + MS in Computational Natural Sciences, IIIT Hyderabad

2019 – 2024

Winner, Institute Research Award (AY 2022–23).

PUBLICATIONS

- Sonus Health: Calibrated Heart-Murmur Detection from Smartphone-Based Veterinary Auscultation. Jose A, Decorte RPJE, Locquet L. arXiv, 2026. arxiv.org/abs/2606.24885
- Instant Preliminary Cardiac Analysis from Smartphone Auscultation: A Real-World Canine Heart Sound Dataset and Evaluation. Jose A, Decorte RPJE, Locquet L. arXiv, 2026. arxiv.org/abs/2601.13593
- DeepGraphMut: A graph-based deep learning method for cancer prognosis using somatic mutation profile. Jose A, Srivastava A, Ansari A, Vinod PK. Briefings in Bioinformatics 26(4), bbaf409, 2025. doi.org/10.1093/bib/bbaf409
- Network diffusion-based approach for survival prediction and biomarker identification in papillary renal cell carcinoma. Shetty KS, Jose A, Bani M, Vinod PK. Mol Genet Genomics 298, 871–882, 2023. doi.org/10.1007/s00438-023-02022-4