Hassen Said Ali

Bologna, Italy | hassensaid32@gmail.com | github.com/hassen8 | hasssen.xyz | linkedin.com/in/hassen7

PROFILE

ML Research Engineer with a focus on representation learning and its intersection with probablistic space modelling, efficient inference, and retrieval systems. At Datalogic, identified that hyperspherical embedding spaces produced by ArcFace losses are fundamentally incompatible with Euclidean Gaussian assumptions, formulated a von Mises-Fisher Mixture Model as a principled geometric replacement, and validated it in a production edge vision pipeline, achieving **98% Top-4 Macro recall** and a **1–5% improvement over the GMM baseline across 6 embedder architectures** on 500K+ images. Also designed Matryoshka-ICD, a contrastive bi-encoder system for clinical text using adaptive Matryoshka representations, and AstraGraph, a GraphRAG system combining graph-structured and dense retrieval. MSc AI, University of Bologna (Mar - 2026) with two IEEE publications.

EXPERIENCE

ML Research Intern · *Datalogic USA, Inc. Eugene, Oregon · (On site)* Feb 2025 – Aug 2025

- **Probabilistic Modeling & Latent Space Geometry**: Architected a custom von Mises-Fisher Mixture Model (vMFmm) from scratch via Expectation-Maximization to natively cluster 128-dimensional hyperspherical embeddings extracted via ArcFace and CCE losses, achieving a Top-4 Macro recall of 98% and Top-1 Macro of 85% (a 1-5% improvement over Euclidean GMM baselines).

- **Data Pipeline Engineering, Model Benchmarking & Incremental Learning**: Engineered Python scripts to simulate on-premise continual learning, (a 200-day continuous learning environment) to evaluate 6 alternative ML architectures (XGBoost, SVM, Random Forest, etc.), generating key business insights on model convergence that proved peak predictive accuracy could be achieved within a 50-day window, directly informing the company's deployment timeline and data collection strategies.

- **Hyperparameter Ablation & Statistical Evaluation**: Executed comprehensive ablation studies across 380,000+ training samples to optimize mixture components, statistically demonstrating that a 5-component configuration provided the optimal bias-variance tradeoff across highly imbalanced, long-tail class distributions.

- **Robustness Testing & OOD Generalization**: Stress-tested the hierarchical vision pipeline by injecting targeted label corruption and multi-dimensional Gaussian feature noise, validating zero-shot generalization on out-of-distribution datasets to guarantee real-world inference performance under sensor data degradation.

PROJECTS

AstraGraph: GraphRAG System for Code Intelligence · github.com/hassen8/astragraph · Live Demo: hasssen.xyz April 2026

- Built a full-stack GraphRAG system ingesting Python codebases via a two-pass tree-sitter AST pipeline into a Neo4j property graph and Qdrant vector store simultaneously, modelling the full entity hierarchy (Repository → Package → Module → Class/Function → Attribute/Parameter) with CALLS, INHERITS, and IMPORTS cross-cutting edges; all writes idempotent via deterministic MD5 UUIDs, with unresolved edges surfaced to audit nodes rather than silently dropped.

- Engineered a 5-node LangGraph StateGraph agent routing queries across graph (Cypher), vector (sentence-transformers), and hybrid modes; hybrid runs both retrievers at 2× top-k and fuses results with Reciprocal Rank Fusion. Storage layer built on typed GraphStore and VectorStore Protocols, making backends swappable without touching the pipeline or agent.

- Evaluated via a 22-query structural and semantic dataset. Benchmarked on the FastAPI repo. Containerised with Docker Compose, deployed to a live VPS with FastAPI backend, Cytoscape.js UI, and pluggable LLM providers (Anthropic, Groq, Ollama).

Matryoshka-ICD: Automated Medical Coding with MRL · github.com/hassen8/Matryoshka-ICD Feb 2026 · University of Bologna

- Built a multi-label ICD coding system for MIMIC-CXR clinical radiology reports using BioClinical-ModernBERT with a custom MRL loss over nesting dimensions [64, 128, 256, 768d], enabling adaptive inference at reduced dimensions without retraining; dataset pipeline expands multi-label rows into anchor-positive pairs matching clinical text queries to ICD semantic descriptions for contrastive bi-encoder training.

- Implemented Label-Aware Attention (LAA) for per-label clinical evidence extraction; conducted a 3-variant ablation study (LAA, Standard Attention, Retrieval-Based Bi-Encoder) evaluated across Micro-F1, ROC-AUC, and Precision@5, with experiment tracking via Weights & Biases.

EDUCATION

MSc in Artificial Intelligence · *University of Bologna, Italy* 2023 – 2026
MSc in Computer Science · *Kalinga Institute of Industrial Technology, India* 2021 – 2023
Bachelor's in Computer Applications · *I.K. Gujral Punjab Technical University, India* 2018 – 2021

SKILLS

AI / ML	Deep Learning · Computer Vision · NLP · LLM Integration · GraphRAG · RAG · Metric Learning · Contrastive Learning · Probabilistic Modelling · Mixture Models · Continual Learning · Embedding Geometry · CLIP · BERT · Active Learning· Classical ML
Frameworks	PyTorch · TensorFlow · HuggingFace Transformers · LangGraph · LangChain · OpenCV · scikit-learn
Infrastructure	Docker · FastAPI · Neo4j · Qdrant · GNU/Linux · Git · Slurm
MLOps & Testing	MLflow · Weights & Biases · pytest · GitLab CI · CI/CD
Programming	Python (NumPy · Pandas · SciPy · scipy.special) · JavaScript · Java · C++ (basic)
Web / DB	React · Node.js · MongoDB · SQL · HTML/CSS
Languages	English (Proficient) · Arabic (Native) · Amharic (Native) · Italian (Basic)

PUBLICATIONS

- **Amharic ATS: Graph-Based vs. Statistical Extractive Summarization** · IEEE Xplore, 2023. · Read

- **Automatic Extractive Text Summarization for Ho Language** · IEEE Xplore, 2023. · Read