

Contact

www.linkedin.com/in/milosbozic
(LinkedIn)

Top Skills

Large Language Models (LLM)
AI Agents
Retrieval-Augmented Generation (RAG)

Honors-Awards

Speaker at Data Science
Conference Belgrade

Publications

Adaptive Least Squares Support
Vector Machines Method for Short-
term Load Forecasting Based
on Mutual Information for Inputs
Selection

A New Two-Stage Approach to Short
Term Electrical Load Forecasting

A methodology for training set
instance selection using mutual
information in time series prediction

Mutual Information-Based Inputs
Selection for Electric Load Time
Series Forecasting

Miloš Božić

Lead ML/AI Engineer | LLMs, Agents, MLOps , AI Security | Built AI
teams & products | PhD in ML | 10+ yrs shipping ML
Belgrade Metropolitan Area

Summary

Lead ML/AI engineer with a PhD in Machine Learning and 10+ years taking models from research to production. Most recently I founded Pyxero.ai, a bootstrapped AI startup that we funded through ML/AI consulting and engineering services. For clients, I delivered a broad range of projects, from fine-tuning LLMs and building agentic systems and RAG pipelines to training and deploying computer vision models. In parallel, I built our own product, Agent Gate, hands-on: a governance layer that lets AI agents call enterprise tools safely, with MCP-based connectors, IdP-scoped permissions, DLP inspection of every tool call, immutable audit logs, and per-agent latency/error observability. When fundraising for the product didn't come through, I decided to close that chapter and return to what I do best: full-time, hands-on ML engineering. What I bring:- Production LLMs & GenAI: fine-tuning, quantization and inference optimization, RAG pipelines, agentic systems, LLM-based human-robot interaction, and AI security (prompt-injection defense, guardrails, MCP governance)- Computer vision, 3D & robotics: nearly a decade shipping vision systems across industries: recognition, OCR and visual search for car marketplaces (Vast), few-shot detection and synthetic data for retail (Everseen), GAN-based seismic inpainting and drone imagery (NIS), FPGA-accelerated point-cloud segmentation (Anari AI), and vision-guided robotic pick-and-place and visual inspection (Alascom)- Time series & industrial ML: PhD research on forecasting published in Neurocomputing, Entropy and Energies; predictive maintenance for pumps in oil & gas production (NIS)- Leadership: built and led AI teams at Vast, Everseen, NIS and Anari AI; ran NIS's Digital Data Lab; hiring, mentoring, research strategy and roadmaps from concept to production I'm interested in a full-time Staff/Lead ML or AI Engineer role where I can own hard technical problems end to end and grow a team for the long term. Stack: Python, PyTorch, Docker/K8s, Azure, AWS, LangChain, LangGraph, LlamaIndex, HuggingFace, Weights&Biases,

Next.js, MongoDB, PostgreSQLSpeaker: DSC Europe, Belgrade AI Week, MLConference, ...

Experience

PYXERO.ai

Founder & Lead AI Engineer

January 2024 - Present (2 years 10 months)

Belgrade

Designed and built Agent Gate, AI governance tool, a policy-enforcement gateway between AI agents and enterprise tools: MCP server for Claude/Cursor/ChatGPT clients, 30+ maintained connectors, credential vaulting with short-lived scoped tokens.

- Implemented DLP guardrails that inspect tool-call inputs/outputs to block, mask or redact sensitive data.
- Built prompt-injection and AI-security research into the product (talk: Belgrade AI Week 2025).
- Built observability: immutable audit trail streamed to SIEM, per-agent/tool/tenant latency and error metrics.

Alascom

AI Solutions Architect

May 2023 - September 2024 (1 year 5 months)

Belgrade

Designed and led end-to-end AI solutions for industrial automation and robotics clients, from concept to deployment.

Architecture & leadership:

- Led the design of end-to-end AI systems: architecture blueprints, technical specifications and integration plans that ensured scalability, security and performance in production.
- Took 3+ AI projects from concept to deployment, defining scope, objectives and deliverables with product teams and client stakeholders.
- Guided the development team technically and evaluated and selected AI tools and frameworks for the company's stack.

Engineering:

- Built vision-guided robotic pick-and-place for objects of varying shapes and structures.

- Developed camera-robot calibration methods.
- Built visual inspection and anomaly detection systems for manufacturing quality control.
- Developed LLM-based text and voice interfaces for natural-language human-robot interaction.
- Built a synthetic data pipeline that combines realistic 3D rendering with generative-AI post-processing, reducing the need for real labeled data.

Anari AI

AI Lead

December 2021 - May 2023 (1 year 6 months)

Belgrade

Deep-tech startup building AI hardware and a compiler stack for efficient deep learning models inference on FPGA.

Engineering

- Led the architecture and hands-on development of transformer-based 3D point cloud segmentation models, for autonomous driving and mapping use case.
- Optimized deep learning models for hardware acceleration using quantization, pruning and operator fusion, achieving 10× faster inference with less than 3% accuracy loss on FPGA targets.
- Contributed to the ML compiler that maps neural networks onto custom FPGA architectures, enabling deployment of models on the hardware.
- Designed training and evaluation data pipelines for large-scale 3D point cloud datasets.

Leadership

- Built and led a team of 5 AI scientists and engineers, covering hiring, mentoring, goal-setting, and code and research reviews.
- Owned the AI technical roadmap: selected, scoped, and prioritized AI projects to balance research ambition against product and hardware constraints.
- Partnered with product and hardware teams to turn research results into product capabilities and customer demos.
- Shaped the company's AI research agenda.

NIS a.d. Novi Sad

Chief Data Scientist

August 2019 - December 2021 (2 years 5 months)

Belgrade

I led the Digital Data Lab, the company's central ML hub, which delivered data products across retail, geology, production, refining, and seismic operations at one of Southeast Europe's largest energy companies.

Leadership & strategy

- Defined data and ML strategy across 5 business domains, turning business challenges into a prioritized portfolio of data products.
- Built and ran the Digital Data Lab as the company's in-house center for ML delivery, leading a team of 10+ data scientists and engineers through 20+ projects from concept to production.
- Worked directly with business stakeholders to translate requirements into ML solutions and find new opportunities, which grew the portfolio to more than 20 projects.
- Directed partnerships with external vendors and academic institutions to speed up R&D.

Delivered ML portfolio

- Retail: customer churn prediction, segmentation, and product recommendation.
- Geology: lithology facies classification and grain size distribution prediction from well data.
- Production & refining: predictive maintenance for electrical submersible pumps (ESPs) and multi-stage centrifugal pumps.
- Seismic & remote sensing: GAN-based seismic data in-painting (EdgeConnect), plus computer vision for understanding aerial drone terrain imagery.
- Designed and set up the lab's complete on-premise compute infrastructure.

Everseen

1 year 5 months

Division Research Manager

January 2019 - August 2019 (8 months)

Belgrade

Senior Research Scientist

April 2018 - January 2019 (10 months)

Serbia

Everseen builds computer vision AI for retail, deployed in 250+ super-stores to detect checkout errors and losses in real time.

Research leadership:

- Led the research teams at the Belgrade R&D center (15+ scientists and engineers), covering hiring, mentoring, and project reviews.
- Defined the center's research strategy and roadmap, choosing research directions that balanced technical ambition against cost and delivery timelines.
- Turned research results into production models for the core product, working with engineering teams in Ireland and Romania.

Computer vision research & engineering:

- Developed one-shot and few-shot learning methods (CNN-based) for object recognition and detection, letting the system recognize new retail products from less than 20 images per product, instead of large labeled datasets.
- Applied weakly supervised learning to object detection, cutting manual annotation effort by 20% from bounding boxes to image-level labels.
- Built synthetic data generation pipelines to augment training data for rare products and edge cases.

Vast.com

Data Scientist

July 2015 - July 2018 (3 years 1 month)

Serbia, Belgrade

Vast.com powers vehicle search and marketplace platforms. I built and led the company's deep learning computer vision solution within the data science department.

Leadership:

- Built and led the deep learning computer vision team (3 engineers), owning the roadmap from problem selection to production deployment.
- Introduced deep learning to the company's data science practice, moving image analysis from manual or rule-based processes to production CNN models.

Computer vision:

- Built a car make / model / generation recognition system covering 1k vehicle classes, used to auto-tag and validate 100M listings.
- Developed models for car orientation detection and transmission type recognition from listing photos, enabling automated listing quality checks and enrichment.

- Built salvage car detection to flag damaged vehicles in listings, improving marketplace trust and data quality.
- Implemented VIN recognition (OCR) from photos, automating vehicle identification.
- Designed a visual search system that lets users find cars by image similarity.

Pricing & anomaly detection:

- Built car price prediction models and anomaly detection for flagging mispriced or fraudulent listings.

Mathematical Institute of the Serbian Academy of Sciences and Arts Research Associate

October 2012 - July 2015 (2 years 10 months)

Serbia, Belgrade

- Developed mutual-information-based feature and instance selection algorithms for time-series forecasting, applied to electrical load forecasting, stock market trend prediction and heart disease prediction.
- Published in peer-reviewed journals including Neurocomputing, Entropy and Energies (200+ citations across my research), and contributed to 2 national research projects.

Faculty of Electronic Engineering, University of Nis

4 years

Research Associate

2010 - 2012 (2 years)

Niš

Young Researcher

2008 - 2010 (2 years)

Niš

Education

Faculty of Electronic Engineering, University of Niš

Doctor of Philosophy (Ph.D.), Machine Learning, Artificial Intelligence, Data Science · (2010 - 2015)

Faculty of Electronic Engineering, University of Niš

Doctor of Philosophy (Ph.D.), Energy savings in electrical lighting consumption based on consumers behaviour profiling · (2008 - 2010)

Faculty of Electronic Engineering

Master's Degree, Electrical Power Engineering · (2002 - 2008)