Republic of Tunisia
*****
Ministry of Higher Education and Scientific Research
*****
University of Monastir
*****
Higher Institute of Computer Science and Mathematics of Monastir
Department of Computer Science

Order N°:

End of Studies Project Report

Presented in order to obtain the

National Bachelor's Degree in Computer Science

Specialization:

Software Engineering and Information Systems

By:

Computer Vision Models Fit for NPUs

Presented on in front of the jury composed of:

	President
	Member
Ms. Nada Haj Messaoud	Academic Supervisor
Mr Ilyes Tlili	Professional Supervisor

Academic Year: 2025 / 2026

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1 Introduction

This introductory chapter presents the scope of the end-of-studies project conducted within **in2-Technologies** in the context of the **EYES (EYE-D)** solution. It introduces the hosting organization, clarifies the project context, and formalizes the problem statement and objectives. It also provides a high-level overview of the adopted approach, with particular emphasis on producing computer vision models that are compatible with deployment constraints on NPU-based systems.

2 Host Organization Presentation

2.1 in2-Technologies

in2-Technologies is an IT company that positions itself around digital innovation by delivering solutions that combine software engineering, creative design, and advanced technologies. In particular, the organization emphasizes applied Artificial Intelligence as a practical lever to deliver real-world value (Figure 1.1(a)).

According to the public communications provided for this report, in2-Technologies is actively involved in the innovation ecosystem through participation in events and technology programs. These activities help connect the company with partners and stakeholders, and they support continuous improvement by confronting products with real deployment needs.

In this context, in2-Technologies develops EYES, an AI-driven solution focused on video analytics and on-device intelligence.

2.2 EYES (EYE-D)

EYES is an AI-driven solution developed within in2-Technologies and centered on computer vision and video analytics. It is presented as an approach that transforms existing video surveillance infrastructures into a system capable of producing actionable insights (Figure 1.1(b)).

Based on the provided communication materials, EYES is designed to integrate with existing camera networks and perform on-site processing. This positioning aims to reduce dependence on cloud processing, improve responsiveness, and address operational constraints such as privacy considerations and deployment cost.

EYES is also described as modular and expandable, enabling the activation or integration of AI modules according to the targeted needs. The communication materials highlight use cases related to security monitoring and industrial safety, with real-time alerting and reporting.

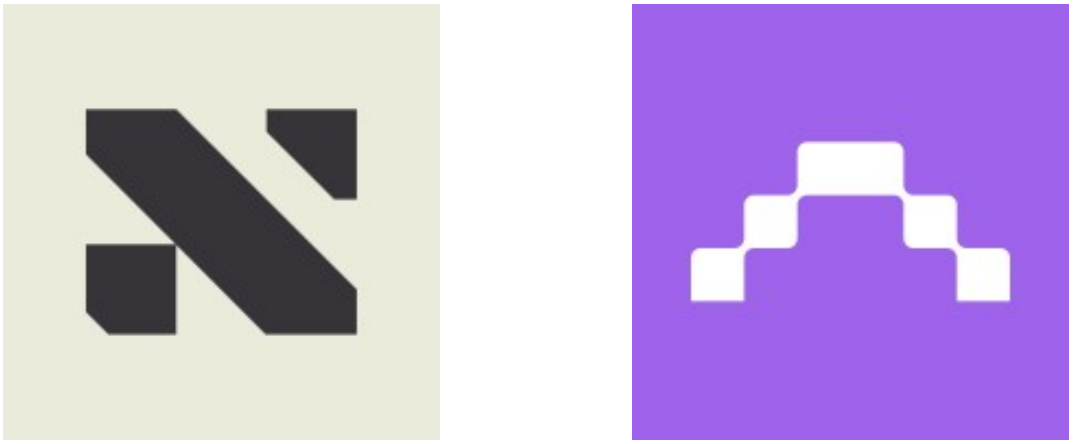
From an operational perspective, the provided materials emphasize the following aspects:

- **Integration**: EYES connects to an existing camera network and can be installed with minimal disruption.
- **On-device processing**: video analytics are executed locally to provide real-time insights and to reduce reliance on external connectivity.
- **Expandable modules**: the solution supports an evolving library of AI modules that can be enabled according to the deployment context.
- **Monitoring and reporting**: the solution provides alerts and reporting capabilities to support operational decision-making.

The brochure-style description provided for EYES points to a set of security and industrial safety features, including (non-exhaustively) access control and intrusion detection, people and vehicle detection, fire and emergency response related detection, and safety compliance monitoring. These capabilities motivate the need for robust, efficient, and deployable computer vision models under edge constraints, which is the focus of this project.

According to the public information provided for this report, EYES has been highlighted through multiple milestones and participations, including:

- Selection as a winner in an AI-focused innovation program (AI Garage Cohort 3 at Novation City).
- First place in the Tunisia IoT & AI Challenge 2025, followed by participation in the Arab IoT & AI Challenge in Dubai.
- Presence in technology events such as Big Tech Africa, STEP Dubai 2025, Security Expo London 2025, and GITEX Expand North Star.



(a) in2-Technologies (b) EYES (EYE-D)

Figure 1.1: in2-Technologies and EYES logos

3 Project Presentation

3.1 Project Context

The project addresses the engineering challenge of developing and deploying computer vision models on resource-constrained platforms equipped with specialized accelerators. The target device and its detailed specifications are **confidential under a signed NDA**. In this report, the target platform is described in a general manner as an embedded system equipped with an **NPU (Neural Processing Unit)**.

Within the EYES context, the project contributes to an applied video analytics solution where on-device inference is required for responsiveness, operational constraints, and deployment practicality. The scope of the work includes a set of computer vision tasks required by the solution, covering multiple model families. The overall goal is not limited to training models, but extends to building complete and reliable pipelines that support deployment and monitoring in realistic operational conditions.

3.2 Problem Statement

Although computer vision models can achieve strong accuracy in research settings, deploying them on NPU-based platforms introduces several constraints and practical difficulties. In particular, the project must address the following issues:

- **Deployment constraints**: limited memory and compute budgets, strict latency requirements, and limited operator support depending on the target accelerator.
- **Model portability**: ensuring that trained models can be exported to an interoperable format (e.g., ONNX) and then converted into optimized inference artifacts.
- **Performance trade-offs**: balancing accuracy, inference speed, and power consumption when selecting architectures and optimization strategies.
- **Pipeline robustness**: handling runtime failures and edge cases through systematic error handling and stable orchestration across multiple vision tasks.

3.3 Requirements and Constraints

In line with the project context and problem statement, the work is driven by a set of requirements and constraints that guide design decisions and validation activities:

- **On-device inference**: the deployed solution must support local execution on the target embedded platform equipped with an NPU.
- **Latency and resource limits**: inference must remain compatible with strict latency targets and limited compute and memory budgets.
- **Deployment compatibility**: trained models must be exportable to an interoperable format and convertible into optimized inference artifacts.
- **Reliability in operation**: inference pipelines must handle runtime failures and edge cases through systematic validation and error handling.
- **Confidentiality constraints**: hardware specifications and internal data collection tooling remain confidential under NDA, requiring careful abstraction in documentation.

3.4 Project Purpose

The main objective of this project is to design an end-to-end workflow that produces **computer vision models fit for NPU deployment** while maintaining reliable inference pipelines for EYES.

The following guiding questions and objectives summarize the expected contribution of the project.

Guiding Question	Project Objective
How can we deploy computer vision models under strict edge constraints?	Build an optimization and deployment pipeline compatible with NPU execution.
How can we maintain acceptable accuracy without exceeding latency and power budgets?	Select and tune suitable architectures, then benchmark trade-offs on representative workloads.
How can we ensure stable integration in a multi-task video analytics product?	Integrate inference into robust pipelines with monitoring, validation, and failure handling.

Table 1.1: Guiding questions and corresponding project objectives

3.5 EYES Web Application

In addition to the on-device inference components, EYES is accompanied by a web-based interface that supports operational use. This interface is used to visualize analytics outputs, consult alerts and reports, and support supervision activities around deployed AI modules. In this report, the web application is considered as the user-facing component that consumes and presents the outputs produced by the embedded inference pipelines.

The 5 W's

In order to clarify the scope of the project and align expectations, the project can be summarized using the **5 W's** framework.

Aspect	Description
Who?	• The project stakeholders include in2-Technologies, the EYES product team, and the operational users of video analytics solutions.
What?	• The project involves the development and integration of an end-to-end workflow for computer vision models within EYES. • The workflow covers training, export to ONNX, optimization, and integration for edge inference.
When?	• EYES is under continuous development and is updated in response to evolving market needs.
Where?	• Within the in2-Technologies environment, targeting an embedded deployment platform equipped with an NPU (details confidential under NDA).
Why?	• Reduce reliance on cloud processing by enabling reliable on-device analytics. • Improve responsiveness by meeting latency, compute, and power constraints on the target platform.

Table 1.2: The Project's 5 W's

4 Project Steps

This project follows an iterative and incremental approach to deliver deployable computer vision components for EYES.

At a high level, the work is organized into the following steps:

- Data collection and preparation using an internal data collector (confidential under NDA).
- Model selection and training using **PyTorch**, focusing on architectures suitable for edge deployment.
- Model export to **ONNX** and validation to ensure correctness after conversion.
- Optimization into deployment-ready inference artifacts suitable for the target platform.
- Integration into multi-task pipelines with systematic validation and failure handling.
- Benchmarking of runtime indicators (latency, memory, and power) under representative inference workloads.

4.1 Existing Solutions

Several solution strategies are commonly adopted to deliver computer vision capabilities under operational constraints:

- **Cloud-centric processing**: cameras stream data to a server or cloud environment that runs inference, then returns results to client applications.
- **Edge computing with accelerators**: inference is executed close to the cameras on embedded devices equipped with GPUs or NPUs, reducing latency and dependency on connectivity.
- **Standardized model exchange and deployment tooling**: workflows often rely on interoperable formats (e.g., ONNX) and on optimization toolchains (quantization, pruning, operator fusion) to obtain deployable artifacts.
- **Modular analytics pipelines**: products typically combine multiple specialized models (detection, tracking, classification) orchestrated to produce higher-level events.

4.2 Critique of Existing Solutions

Although these approaches are effective in many contexts, they present limitations that motivate the focus of this project:

- **Latency and bandwidth trade-offs**: cloud-centric approaches increase dependency on network connectivity and may not satisfy strict real-time constraints.
- **Conversion and portability issues**: model export and optimization may fail due to unsupported operators, numerical differences, or vendor-specific constraints.
- **Accuracy degradation**: deployment-oriented optimizations such as quantization may reduce accuracy if not carefully validated and calibrated.
- **Operational robustness**: multi-model pipelines can be sensitive to runtime failures and edge cases, requiring systematic monitoring and error handling.
- **Vendor lock-in risks**: NPU deployment toolchains may impose specific constraints that limit portability across platforms.

5 Adopted Project Management Methodology

The project is conducted using an iterative approach that allows incremental delivery and continuous validation. This approach is suitable for AI engineering workflows where model performance, deployment compatibility, and resource constraints must be checked throughout the development cycle.

For this project, we adopt the **Scrum** methodology as an Agile project management framework. The work is organized into time-boxed sprints, enabling regular planning, implementation, and review cycles. This organization supports continuous refinement of both model quality and deployment readiness.

Scrum Element	Project Adoption
Product Backlog	Maintain a prioritized list of tasks covering training, optimization, and integration activities.
Sprint Planning	Select sprint goals and define deliverables that can be validated on the target platform.
Sprint	Implement and integrate incremental improvements, with frequent checks on performance constraints.
Sprint Review	Demonstrate the sprint increment (models, pipelines, or benchmarks) and collect feedback.
Sprint Retrospective	Identify process improvements and adapt the next sprint organization accordingly.

Table 1.3: Scrum methodology and its adoption in the project

For project tracking and collaboration, **GitHub** is used as the central platform. It supports version control, task tracking through issues, and coordination of development activities throughout the project lifecycle.

6 Conclusion

This chapter presented the hosting organization (in2-Technologies) and the context of the EYES solution, then formalized the project overview, requirements, objectives, and the adopted project management approach. The next chapter will introduce the technical background required for this work, including computer vision modeling, ONNX export, and deployment-oriented optimization for NPU-based inference.