# PREDISTINE - A Privacy-Preserving and Resilient Framework for Distributed Modular Neural Networks on the Tiny Edge

*A Data Management Plan created using DMPonline.be*

**Creator:** Gregory De Ruyter

**Affiliation:** KU Leuven (KUL)

**Funder:** Fonds voor Wetenschappelijk Onderzoek - Research Foundation Flanders (FWO)

**Template:** FWO DMP (Flemish Standard DMP)

**Principal Investigator:** Gregory De Ruyter

**Grant number** / **URL:** 1SH9Y24N

**ID:** 205854

**Start date:** 01-11-2023

**End date:** 31-10-2027

**Project abstract:**

Due to the large amount and diversity of data generated by Internet of Things (IoT) devices, Artificial Intelligence (AI) has seen a breakthrough in IoT applications in recent years. Traditionally this implies that data must be sent to the cloud, which increases communication costs, causes delays in system response, and makes data vulnerable to privacy breaches. A solution is to migrate AI models to constraint devices. However, these are often limited in their ability to implement such models. Besides that, the system also needs to be flexible when devices fail or reconnect.

Therefore, this proposal will focus on a new framework for distributed modular networks on the tiny edge consisting of microcontrollers. The first research objective is to study how a model can be efficiently distributed between edge devices and server and what the implications are considering energy consumption, memory footprint, and bandwidth usage. The second objective is to explore ways to make neural networks adaptable to various combinations of input devices. This makes the inference more resilient in case of failing devices. The final objective is to investigate the integration of distributed learning within the proposed system architecture to enhance model performance and personalize predictions. Finally, the research results will be implemented and validated on two real-world use cases, from which a software framework will be developed and made publicly available for use in other applications.

**Last modified:** 03-04-2024

# PREDISTINE - A Privacy-Preserving and Resilient Framework for Distributed Modular Neural Networks on the Tiny Edge
## FWO DMP (Flemish Standard DMP)

**1. Research Data Summary**

**List and describe all datasets or research materials that you plan to generate/collect or reuse during your research project. For each dataset or data type (observational, experimental etc.), provide a short name & description (sufficient for yourself to know what data it is about), indicate whether the data are newly generated/collected or reused, digital or physical, also indicate the type of the data (the kind of content), its technical format (file extension), and an estimate of the upper limit of the volume of the data.**

| | | | | Only for digital data | Only for digital data | Only for digital data | Only for physical data |
|---|---|---|---|---|---|---|---|
| Dataset Name | Description | New or reused | Digital or Physical | Digital Data Type | Digital Data format | Digital data volume (MB/GB/TB) | Physical volume |
| ProtoNAS Framework | Software framework for prototyping, and evaluating evolutionary NAS algorithms | Generate new & reuse existing data | Digital | Software | .py | < 100MB | NA |
| NAS Algorithms for Distributed Neural Networks | Set of NAS algorithms in software code that can be used to create distributed neural networks for different prediction tasks | Generate new & reuse existing data | Digital | Software | .py | < 100MB | NA |
| Modular Neural Network Framework | Software framework enabling neural network to dynamically adapt to the input data being fed (to create more resiliency in case of missing data) | Generate new & reuse existing data | Digital | Software | .py, .c, .h, .cpp, .hpp | < 100MB | NA |
| Distributed Learning Framework | Software framework facilitating distributed learning on edge devices | Generate new & reuse existing data | Digital | Software | .py, .c, .h, .cpp, .hpp | < 100MB | NA |
| Models found by NAS Algorithm (data formatted & generated by ProtoNAS framework) | Found models (solutions) that can be validated and used in the prediction task they were trained on. | Generate new data | Digital | Other (results generated by framework) | .onnx, .sqlite, .json | < 100GB | NA |
| Food Intake / Human Activity Recognition Dataset | Dataset captured within the health care use case and used to validate the developed framework. This consists of load cells measurements from an in-house smart plate, and gyroscope & accelerometer data from the wearables. | Generate new data | Digital | Observational | .csv, mp4 (only for labelling) | < 300GB | NA |
| Product Quality Monitoring Dataset | Dataset captured within the industry use case and used to validate the developed framework. This dataset will be mainly captured from a manufacturing process on the Ultimate Factory, a lab-scale industrial factory. | Generate new data | Digital | Observational | .csv | < 100GB | NA |
| Literature datasets | Literature datasets used as benchmarks to compare our techniques with the state of the art | Reuse existing data | Digital | Observational & experimental | .csv | < 100GB | NA |

**If you reuse existing data, please specify the source, preferably by using a persistent identifier (e.g. DOI, Handle, URL etc.) per dataset or data type:**

**Literature datasets:**

- UCI-HAR: 10.24432/C54S4K
- CWRU: https://engineering.case.edu/bearingdatacenter/apparatus-and-procedures

More literature datasets might be added in the future.

**Are there any ethical issues concerning the creation and/or use of the data (e.g. experiments on humans or animals, dual use)? Describe these issues in the comment section. Please refer to specific datasets or data types when appropriate.**

- No

**Will you process personal data? If so, briefly describe the kind of personal data you will use in the comment section. Please refer to specific datasets or data types when appropriate.**

- Yes

Food Intake / Human Activity Recognition
To establish ground truth labels for the activities performed by participants, a video is recorder while the subjects are asked to perform activities. This video is subsequently utilized to annotate the measured sensor data. Once the labels are generated, the videos are deleted, ensuring that the traceability of the sensor data to the participant is removed

**Does your work have potential for commercial valorization (e.g. tech transfer, for example spin-offs, commercial exploitation, …)? If so, please comment per dataset or data type where appropriate.**

- Yes

The software frameworks (NAS algorithms, ProtoNAS framework, Modular Neural Network Framework, Distributed Learning Framework), combined or separate, can potentially be used by companies to improve their machine learning applications on the edge.

**Do existing 3rd party agreements restrict exploitation or dissemination of the data you (re)use (e.g. Material/Data transfer agreements/ research collaboration agreements)? If so, please explain in the comment section to what data they relate and what restrictions are in place.**

- No

**Are there any other legal issues, such as intellectual property rights and ownership, to be managed related to the data you (re)use? If so, please explain in the comment section to what data they relate and which restrictions will be asserted.**

- No

**2. Documentation and Metadata**

**Clearly describe what approach will be followed to capture the accompanying information necessary to keep data understandable and usable, for yourself and others, now and in the future (e.g., in terms of documentation levels and types required, procedures used, Electronic Lab Notebooks, README.txt files, Codebook.tsv etc. where this information is recorded).**

The datasets (Food Intake / Human Activity Recognition Dataset, Product Quality Monitoring Dataset) will be accompanied by linked guidance documents.
The software frameworks (NAS algorithms, ProtoNAS framework, Modular Neural Network Framework, Distributed Learning Framework) will always be accompanied by (Jupyter) notebooks, demo scripts, API documentation, and readme files. These resources serve to demonstrate the usage of the software or provide guidance for users to build the software themselves."

**Will a metadata standard be used to make it easier to find and reuse the data? If so, please specify (where appropriate per dataset or data type) which metadata standard will be used. If not, please specify (where appropriate per dataset or data type) which metadata will be created to make the data easier to find and reuse.**

- Yes

As the datasets will be stored on KU Leuven RDR for open access, DataCite will be used as a metadata standard as it is required by the platform.

**3. Data storage & back-up during the research project**

**Where will the data be stored?**

- The software frameworks will be stored on GitLab (hosted by KU Leuven), under the research group's GitLab group (M-Group). This guarantees access to the source code by anyone within the research group, even after the maintainer (me) has left KU Leuven.
- The datasets will be stored, and managed on KU Leuven's Active Data Management Platform (ManGO) while being worked on (active data). Once the datasets are ready, they will be published on KU Leuven RDR for open access, and long-term storage.
- All other data will be stored on a personal OneDrive for Business cloud storage provided by KU Leuven.

**How will the data be backed up?**

The source code will be daily pushed to the GitLab repository.
Local data and files in general are backed up using OneDrive for business' file synchronisation.

**Is there currently sufficient storage & backup capacity during the project? If yes, specify concisely.**
**If no or insufficient storage or backup capacities are available, then explain how this will be taken care of.**

- Yes

GitLab provides 10GB of storage per repository which is more than sufficient to host the source code on.
KU Leuven ManGO provides 1TB of data storage which should be enough to store all (active) datasets on.
KU Leuven RDR provides 50GB of storage, but this can be extended if needed.

**How will you ensure that the data are securely stored and not accessed or modified by unauthorized persons?**

Any acces to the files on OneDrive for business will be given through the usage links with strict access only to the intended receiver. The GitLab repositories, and KU Leuven ManGO storage are moderated, thus any acces has to be given, which can be done with various rights to any additional user as needed (read-only, read-write, admin,...).

**What are the expected costs for data storage and backup during the research project? How will these costs be covered?**

GitLab storage: free
KU Leuven ManGO: €35,00 / year / 1TB

All costs will be covered by the FWO project's bench fee.

#### 4. Data preservation after the end of the research project

**Which data will be retained for at least five years (or longer, in agreement with other retention policies that are applicable) after the end of the project? In case some data cannot be preserved, clearly state the reasons for this (e.g. legal or contractual restrictions, storage/budget issues, institutional policies...).**

All data will be stored in the KU Leuven Research Data Repository (RDR) for a minimum of ten years in accordance with the KU Leuven RDM Policy. Data will be linked to corresponding publications where applicable.

**Where will these data be archived (stored and curated for the long-term)?**

Data will be stored in the KU Leuven Research Data Repository (RDR) for a minimum of ten years. Data will be linked to corresponding publications where applicable.
Source code will remain on the repositories under the research group's GitLab team and a mirror will be created to a personal GitHub repository where it will be kept public under GNU General Public License v3.0 or later licensing.

**What are the expected costs for data preservation during the expected retention period? How will these costs be covered?**

KU Leuven RDR: 50GB / year for free
GitLab storage: free
GitHub storage: free (as it will be published on an open repository)

#### 5. Data sharing and reuse

**Will the data (or part of the data) be made available for reuse after/during the project? In the comment section please explain per dataset or data type which data will be made available.**

- Yes, in an Open Access repository
- Yes, in a restricted access repository (after approval, institutional access only, …)

The datasets will be available on ManGO (active storage, restricted), and in a Open Access repository (KU Leuven RDR) upon publication.
The software framework will be kept on a GitLab repository hosted by KU Leuven (restricted), and a personal open GitHub repository (upon publication) under GNU General Public License v3.0 or later licensing.

**If access is restricted, please specify who will be able to access the data and under what conditions.**

The datasets within ManGO remain private unless other researchers are willing to cooperate on data gathering / dataset creation. In such cases, co-ownership access can be granted.
The source code on the GitLab repository is only accessible by members of the research group (M-Group), however, access can be provided to students within the context of student projects / theses about this project.

**Are there any factors that restrict or prevent the sharing of (some of) the data (e.g. as defined in an agreement with a 3rd party, legal restrictions)? Please explain in the comment section per dataset or data type where appropriate.**

- No

Not applicable

**Where will the data be made available? If already known, please provide a repository per dataset or data type.**

Not applicable

**When will the data be made available?**

Data (both datasets, and software frameworks) will be available upon publication of research results.

**Which data usage licenses are you going to provide? If none, please explain why.**

Datasets: Public Domain (PD)
Software frameworks: GNU General Public License v3.0 or later

**Do you intend to add a PID/DOI/accession number to your dataset(s)? If already available, you have the option to provide it in the comment section.**

- No

**What are the expected costs for data sharing? How will these costs be covered?**

GitLab storage: free
KU Leuven ManGO (includes data sharing): €35,00 / year / 1TB

All costs will be covered by the FWO project's bench fee.

**6. Responsibilities**

**Who will manage data documentation and metadata during the research project?**

Gregory De Ruyter. Those who upload new data are responsible for providing data documentation of their data.

**Who will manage data storage and backup during the research project?**

Gregory De Ruyter

**Who will manage data preservation and sharing?**

Gregory De Ruyter

**Who will update and implement this DMP?**

Gregory De Ruyter