



31 March 2025

Project Timeline: DeepXplain Summer 2025

Students: Røskva Bjørgfinsdóttir (rosktb@gmail.com) and Brage Eilertsen (bragee2002@gmail.com)

Phase 1: Preparation & Onboarding (July 1 - July 4)

- Introduction to the scope and objectives of the project.
- Familiarization with deep learning frameworks. (PyTorch, TensorFlow) and interpretability libraries.
- Reading materials and discussion on explainability techniques (SHAP, LIME, attention mechanisms).
- Setting up the development environment and datasets

Phase 2: Initial Research & Benchmarking (July 7 - July 11)

- Main concepts and definitions of explainability methods.
- Benchmarking existing explainability tools on selected deep learning models.
- First round of experiments: evaluating bias detection, factual consistency, and model transparency.
- Team discussions and evaluation.

Phase 3: Implementation & Experimentation (July 14 - July 18)

- Implementing selected explainability methods for deep learning models
- Testing methods in real-world AI applications (e.g., bias mitigation, factual consistency).
- Iterative improvements based on experimental results.
- Weekly progress updates and peer feedback.

Phase 4: Evaluation (July 21 - July 25)

- Systematic assessment of implemented explainability techniques.
- Comparing performance with existing state-of-the-art methods.
- Conducting user studies (if applicable) to measure interpretability and usability.
- Writing preliminary conclusions and discussing limitations.

Phase 5: Open-Source Contribution & Finalization (July 28 - July 31).

- Discussion on future directions for explainability research
- Closing event and certificate distribution