OLLSCOIL TEICNEOLAÍOCHTA BHAILE ÁTHA CLIATH
DUBLIN
TECHNOLOGICAL UNIVERSITY DUBLIN

Erasmus+

Automating Interpretation and explainability of Machine learning Models

hcaim human centred artificial intelligence masters

Real AI

Aim : To develop an AI-driven automation framework for interpreting and explaining machine learning models.

ABSTRACT

- Machine learning models often lack transparency, making it difficult to understand their decisions.
- Existing explainability tools like SHAP and LIME require high computational power and manual effort.
- This research proposes an AI-driven automation framework to enhance interpretability.
- The approach integrates feature attribution, counterfactual explanations, and rule-based approximations.

INTRODUCTION

- Machine learning is widely used but often functions as a "black box," raising concerns about trust and ethics.
- Explainability is crucial for regulatory compliance, bias detection, and informed decision-making.
- Current methods require manual effort and computational resources, limiting adoption.
- This research explores AI-driven automation to make explainability scalable, efficient, and user-friendly.

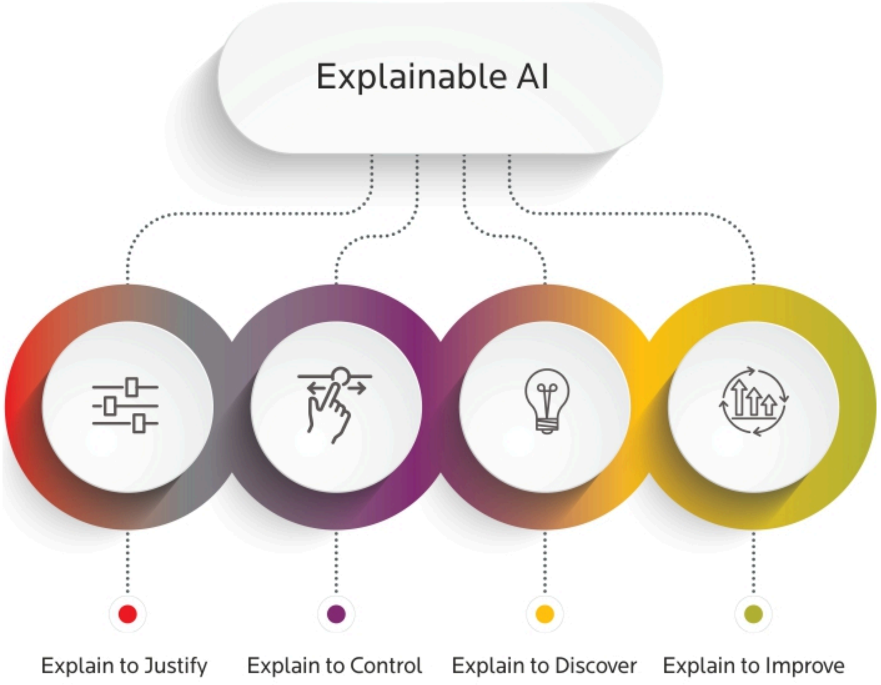
LITERARTURE REVIEW

- **Need for Explainability:** Transparency in AI improves trust, fairness, and regulatory compliance.
- **Existing Methods:** SHAP, LIME, and counterfactual explanations provide insights but have limitations.
- **Challenges:** High computational costs, lack of standardization, and difficulty in interpretation.
- **Automation in Explainability:** Recent studies explore AI-driven solutions but lack practical implementation.
- **Research Gap**: The need for a scalable, automated explainability framework with minimal computational overhead

RESEARCH METHODOLOGY

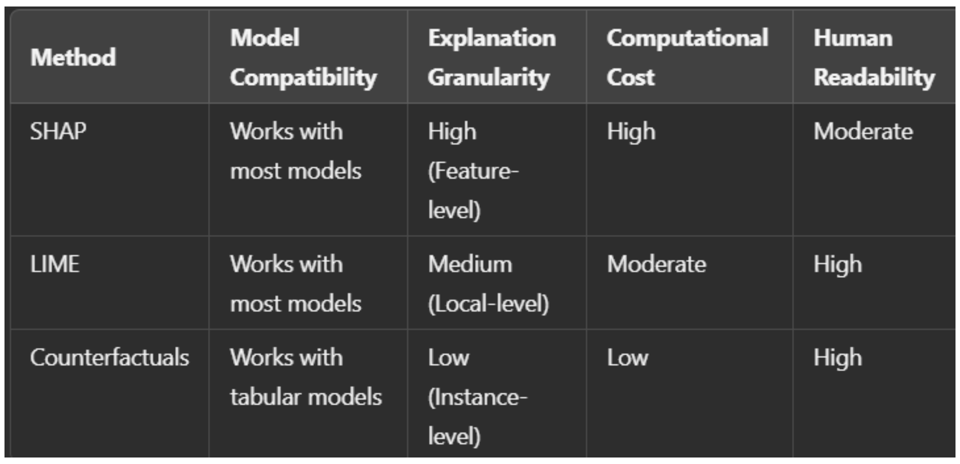
- Model Selection: Train a specific ML model (e.g., decision trees, random forests, deep learning).
- Framework Development: Automate explainability using SHAP, LIME, and AI-based summarization.
- Optimization: Reduce computational load through parallel processing and model simplification.
- Evaluation Metrics: Measure interpretability, efficiency, and usability through case studies.
- Tools: Python, TensorFlow, PyTorch, Alibi, Captum, and visualization tools like Streamlit

REFERENCES
RUDIN, C. (2019). "STOP EXPLAINING BLACK BOX MACHINE LEARNING MODELS FOR HIGH STAKES DECISIONS AND USE INTERPRETABLE MODELS INSTEAD." NATURE MACHINE INTELLIGENCE, 1(5), 206-215.
DOSHI-VELEZ, F., & KIM, B. (2017). "TOWARDS A RIGOROUS SCIENCE OF INTERPRETABLE MACHINE LEARNING." ARXIV PREPRINT ARXIV:1702.08608.
LUNDBERG, S. M., & LEE, S. I. (2017). "A UNIFIED APPROACH TO INTERPRETING MODEL PREDICTIONS." ADVANCES IN NEURAL INFORMATION PROCESSING SYSTEMS (NEURIPS).
RIBEIRO, M. T., SINGH, S., & GUESTRIN, C. (2016). "WHY SHOULD I TRUST YOU? EXPLAINING THE PREDICTIONS OF ANY CLASSIFIER." PROCEEDINGS OF THE 22ND ACM SIGKDD INTERNATIONAL CONFERENCE ON KNOWLEDGE DISCOVERY AND DATA MINING.
WACHTER, S., MITTELSTADT, B., & RUSSELL, C. (2017). "COUNTERFACTUAL EXPLANATIONS WITHOUT OPENING THE BLACK BOX: AUTOMATED DECISIONS AND THE GDPR." HARVARD JOURNAL OF LAW & TECHNOLOGY, 31(2), 841-887.
MOLNAR, C. (2022). INTERPRETABLE MACHINE LEARNING. LEANPUB.
GHORBANI, A., ABID, A., & ZOU, J. (2019). "INTERPRETATION OF NEURAL NETWORKS IS FRAGILE." PROCEEDINGS OF THE AAAI CONFERENCE ON ARTIFICIAL INTELLIGENCE.
HIND, M., HOUDE, S., KOVALEVSKY, A., ET AL. (2020). "EXPERIENCES WITH IMPROVING THE TRANSPARENCY OF AI MODELS AND SERVICES." PROCEEDINGS OF THE 2020 CONFERENCE ON FAIRNESS, ACCOUNTABILITY, AND TRANSPARENCY (FACCT).



PRELIMINARY CONSIDERATIONS

- Existing Challenges: High manual effort, computational inefficiency, and difficulty in integration.
- Stakeholder Needs: Different users require varying levels of explanation complexity.
- Ethical Considerations: Ensuring fairness, reducing bias, and aligning with regulations.
- Scalability: Making explainability solutions cost-effective and enterprise-ready.
- Usability: Creating intuitive interfaces for both technical and non-technical users.



Method	Model Compatibility	Explanation Granularity	Computational Cost	Human Readability
SHAP	Works with most models	High (Feature-level)	High	Moderate
LIME	Works with most models	Medium (Local-level)	Moderate	High
Counterfactuals	Works with tabular models	Low (Instance-level)	Low	High

CONCLUSIONS

This thesis aims to bridge the gap between automated ML explainability and human-centered AI. By automating interpretation techniques while prioritizing usability, the research contributes to more transparent, trustworthy AI systems. Future work includes expanding the framework to support multiple ML models and integrating it into MLOps pipelines for real-world deployment. This work ultimately promotes responsible AI adoption by ensuring explanations are both technically accurate and comprehensible to diverse stakeholders.

ACKNOWLEDGEMENTS: THE HCAIM (THE HUMAN-CENTRED AI MASTER'S PROGRAMME) PROJECT IS CO-FINANCED BY THE CONNECTING EUROPE FACILITY OF THE EUROPEAN UNION UNDER GRANT N°CEF-TC-2020-1 DIGITAL SKILLS 2020-EU-IA-0068. THIS POSTER WAS CREATED AS PART OF THE BLENDED INTENSIVE PROGRAMME ORGANIZED UNDER THE ERASMUS + PROGRAMME OF THE EUROPEAN UNION.