Politecnico di Torino

ScuDo
Scuola di Dottorato ~ Doctoral School
WHAT YOU ARE, TAKES YOU FAR

Doctoral Dissertation
Doctoral Program in Computer and Control Engineering (37^{th}cycle)

Latent Space Dynamics and Security Implications in Machine Learning for Industrial Systems

By

Doaa Almhaithawi

******

Supervisor(s):
Prof. Tania Cerquitelli, Academic Supervisor
Dr. Alessandro Bellini, Industrial Supervisor (Mathema srl, Firenze, Italy)

Doctoral Examination Committee:
VARGAS SOLAR GENOVEVA CNRS, France
ZUMPANO ESTER Università della Calabria, Italy
GARZA PAOLO Politecnico di Torino, Italy
QUINTARELLI ELISA Università degli Studi di Verona, Italy
SCHIFANELLA ROSSANO Università degli Studi di Torino, Italy

Politecnico di Torino
2025

Declaration

I hereby declare that, the contents and organization of this dissertation constitute my own original work and does not compromise in any way the rights of third parties, including those relating to the security of personal data.

Doaa Almhaithawi
2025

* This dissertation is presented in partial fulfillment of the requirements for **Ph.D. degree** in the Graduate School of Politecnico di Torino (ScuDo).

Latent Space Dynamics and Security Implications in Machine Learning for Industrial Systems

Doaa Almhaithawi

The rapid advancement of Machine Learning has revolutionised numerous areas. However, Machine Learning models have long been perceived as black boxes — powerful but opaque systems that lack interpretability. In recent years, significant breakthroughs have begun to unveil the internal mechanisms of these models and foster a deeper understanding of their decisions. This shift has driven research in key areas such as explainability, fairness and trustworthiness, with the aim of bridging the gap between human interpretability and artificial intelligence. However, as the understanding of Machine Learning models grows, so do the challenges associated with security and privacy. Risks such as the leakage of sensitive training data and adversarial attacks pose a serious threat and require the development of robust defences to preserve the integrity of Machine Learning systems.

At its core, Machine Learning can be defined as the approximation of functions that map input data to output predictions. This often involves solving tasks that are complex for both humans and traditional algorithms. Deep learning in particular has enabled the investigation of **latent spaces**, the multidimensional vector representations embedded in neural networks. These latent spaces serve as a powerful framework for structuring and synthesising data across different modalities, such as images, text, speech and signals, filtering out noise and irrelevant factors. Despite their empirical success, many fundamental aspects of the structure of latent spaces have not yet been sufficiently explored from a theoretical, methodological and applied perspective.

This dissertation attempts to establish a deeper theoretical foundation for latent spaces while enhancing their practical applications. Motivated by the increasing use of Machine Learning in real industrial applications, this research examines why Machine Learning models behave the way they do and how the representations of latent spaces can be leveraged more effectively. The study unfolds across four key directions:

1. Mathematical foundations – A rigorous mathematical review is introduced that refines the terminology, definitions and theoretical constructs of latent space

representations. The study provides a structured overview of existing methods and algorithms and highlights their strengths and limitations across various applications.

2. Creative applications based on Generative Artificial intelligence – The first case study investigates the use of latent spaces in artistic applications, particularly in the generation of artworks in the style of Leonardo da Vinci. The research identifies potential biases in such models and proposes mitigation strategies to ensure fair and unbiased generative outcomes.

3. Industrial applications for quality control – The second case study applies latent space techniques to real-world manufacturing challenges and focuses on the automatic detection of paint defects in vehicle production. The system supports human operators and reduces cognitive and physical fatigue. It demonstrates how latent space representations can enhance precision and efficiency in industrial environments.

4. Latent space for model security and privacy – Finally, a novel perspective on leveraging latent space structures for Machine Learning security is presented, offering new possibilities for defending models against adversarial attacks and protecting sensitive information. This emerging direction lays the foundation for future research at the intersection of Machine Learning robustness, adversarial resilience and privacy preservation.

By bridging theoretical insights with practical implementations, this research contributes to a more structured, interpretable and secure approach to analysing latent spaces. The findings of this dissertation could be an important reference for researchers and practitioners who want to enhance the transparency, reliability and security of Machine Learning models in both creative and industrial domains.