Summary of the Thesis

High-fidelity Computational Fluid Dynamics (CFD) simulations provide detailed descriptions of fluid flows across a wide range of scientific and engineering domains. However, especially in realistic settings, the resulting flow fields tend to be extremely high-dimensional, strongly dependent on geometry, and computationally expensive to generate, which limits their direct usability within Machine Learning (ML) pipelines. In practical applications, these challenges translate into two coupled limitations: the scarcity of labelled simulations (*Small-n*) and the very large number of degrees of freedom per sample (*Large-p*), which render direct end-to-end learning from CFD data ill-posed and unable to generalise across geometries under realistic data budgets. This thesis addresses ML-based inference from CFD precisely in this regime, where the goal is not to reproduce physical quantities already computable from governing equations, but to infer high-level, non-computable properties such as geometric defects or pathological conditions.

The methodological contribution of this work is structured along two complementary directions, each targeting one of the two limitations aforementioned. The first direction addresses (*Small-n*) through a geometry-based data augmentation framework, with particular emphasis on diagnostic scenarios involving pathological conditions. Instead of attempting to augment flow fields directly, an operation that can easily violate physical constraints or alter semantic labels, the proposed approach acts on the computational domain. Starting from a single reference geometry on which specific pathologies are explicitly defined, computational geometry and shape correspondence techniques are used to transfer these deformations onto anatomies extracted from healthy patients, generating synthetic pathological variants in a controlled and physically consistent manner. CFD simulations performed on the resulting augmented geometries yield an enlarged labelled dataset while preserving physical admissibility and label consistency. This strategy is designed to increase variability in a principled way and to reduce the dependence on costly expert annotation and simulation campaigns.

The second direction addresses (*Large-p*) by developing feature extraction strategies that compress CFD outputs into compact, informative, and learnable representations. Two strategies are investigated. The first is a physics-based clustering approach, where the computational domain is segmented into meaningful regions by clustering quantities derived from the governing equations (e.g., contributions associated with advection, diffusion, pressure gradients, and turbulence-related terms). Features are then computed as regional averages and geometric descriptors, enabling adaptive region definition without manual definition. The second is a morphing-based approach, in which heterogeneous simulations are aligned onto a common reference domain through smooth deformation models; this alignment enables expert-defined regions to be specified once on the reference geometry and then reused consistently across samples. Together, these strategies balance adaptability (physics-based region definition) and transferability (cross-sample region consistency), providing scalable feature definition in the presence of strong geometric variability.

The proposed framework is validated on application scenarios of increasing complexity, including two-dimensional aerodynamic flows around airfoils with controlled geometric variations and three-dimensional simulations of airflow in patient-specific upper airways for pathology classification. Across these settings, the combination of geometry-based augmentation and physically grounded feature extraction enables robust inference, improving scalability and generalisation relative to fully manual, case-dependent feature engineering. As an additional contribution, this thesis also produces and releases curated datasets of geometries, CFD simulations, and derived features, intended to support reproducible research and further developments in ML-based inference from CFD.